Transactions on Engineering and Computing Sciences - Vol. 11, No. 1
Publication Date: January 25, 2023
DOI:10.14738/tmlai.111.13876.
Wakabayashi, N., & Shimakawa, H. (2023). Standardization of Criteria across Multiple Evaluators to Detect Objects Transactions on Engineering and Computing Sciences, 11(1). 22-36.

SOCIETY FOR SCIENCE AND EDUCATION

Standardization of Criteria across Multiple Evaluators to Detect Objects

Naoya Wakabayashi
Graduate School of Information Science and Engineering,
University of Ritsumeikan, Japan

Hiromitsu Shimakawa
Graduate School of Information Science and Engineering,
University of Ritsumeikan, Japan

ABSTRACT

For a typical object detection task with machine learning technique, there has an absolute correct label on which a mechanical model is constructed. However, there are many tasks in which labels vary with evaluators because they have different criteria for discrimination. It happens when the detection criteria are vague and undefined among evaluators. This problem is avoided by constructing an individual model, but it is not recommended from a long-term perspective, because the model depends on a specific evaluator. This paper proposes a method to standardize the evaluator's criteria. The method is verified in the image detection task of asteroid powders in paint material as an example of ambiguity in the criteria. The performance and variance of detection with the proposed method are compared with conventional ones to evaluate whether it allows us to standardize the evaluators' criteria. It turns out a clear reduction in the variance of evaluators' detection results without significant degradation of the performance for all evaluators. It is confirmed this method standardizes the criteria across multiple evaluators. In addition, the paper discusses how to obtain manuals formally expressing the standardized criteria with text mining.

Keywords: Object detection; Inconsistent labels; Supervised machine learning; Criteria standardization; Label synthesis; Reform based on feedback.

INTRODUCTION

It is highly demanded to detect specific objects in machine vision applications with machine learning. Generally, correct labels given to construct a model should not vary with evaluators. A uniform labeling standard among many evaluators would enable us to detect precisely target objects from a huge amount of image data. However, evaluators often assign different labels despite the same image. It happens when the detection criteria are vague and undefined among evaluators. An individual model constructed from each evaluator could bring a certain level of performance, but detection results would depend on the evaluator.

The paper proposes a method to standardize the evaluator's criteria by creating a generalized labels through the uniformization of the labels obtained for each evaluator.

Celis, et. al [1] have discussed implicit diversity in image search. The work proposes a method to achieve diversification without adding new features. Deselaers, et. al [2] have proposed a method to optimize relevance and diversity in image retrieval. It has ensured a diversity of search approaches using techniques inspired by dynamic programming algorithms. Karako, et. al [3] have proposed a method that presents fairness-conscious variations. The work has verified the trade-off between relevance and fairness to realize a calculation method to ensure fairness. These methods can provide highly relevant objects while considering the diversity of detected objects. Their outputs are standardized to reduce the scattering of data while ensuring diversity. However, all the works fail to address the problem that labels with which each object is categorized differ from one evaluator to another.

Liu, et.al [4] and Gao, et.al [5] both pursued sample diversity to create large amounts of similar training data using GANs. The generated images should be related but diverse. However, this method can only be applied when the criteria for determining labels are well-defined. None of them addresses differences in criteria for labeling.

This study uses the detection of asteroid powders task as an example of ambiguous criteria to detect objects. Diverse color design in recent years has accelerated the development of many paints using metallic pigments [6,7]. The metallic pigments are fabricated through milling processes of metal powder having several morphologies [8,9]. In the process, larger powders happen to be manufactured with smaller ones randomly attached. They are called asteroid powders. Asteroid powders degrade the luminosity of paint to lower the quality of painted products. It is crucial to measure the density and the distribution of the asteroid powders to maintain the quality of painted products.

Wakabayashi, et.al [10] have addressed the task with an object detection method. They have shown that YOLO [11] is useful as an object detection method. However, since the work founds on models depending on evaluators, it has not achieved the standardization of the detection criteria.

This paper challenges standardization. It proposes a method to achieve standardization according to the Delphi method [12,13], a method to lead group opinion to uniformity. It evaluates whether the proposed method standardizes the evaluator's criteria with the comparison of its detection results with those without the method.

ASTEROID POWDERS

Figure 1 shows an image example of powder aggregation observed by microscopy. An enlarged cutout of a part of the image is shown in Figure 2. The type of powder aggregation used in this paper is the same as the one in the previous study [10]. Conventionally, the skilled workers manually mark the asteroid powders from the microscope images as it is shown in figure 1. Automatic marking is desirable to improve productivity.

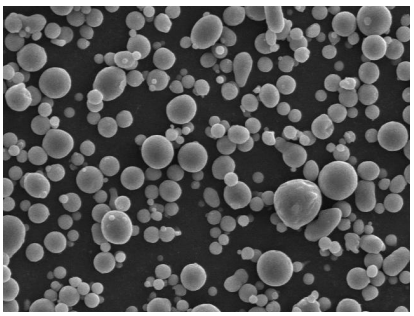


Figure 1. Powder aggregation by microscopy.

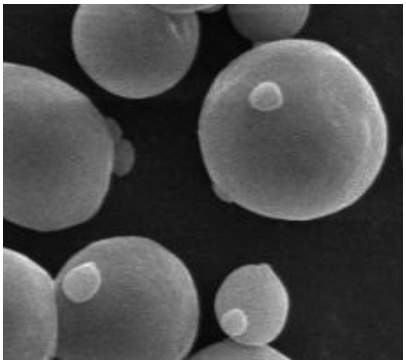


Figure 2. Enlarged cutout of Figure 1.

Figure 3 depicts area examples of asteroid powders each evaluator marks for the microscope images given in Figure 2. The areas enclosed by the rectangles indicate asteroid powders to be detected for each. Since each rater has a different criterion, the ground truth labels are different even for the same image.

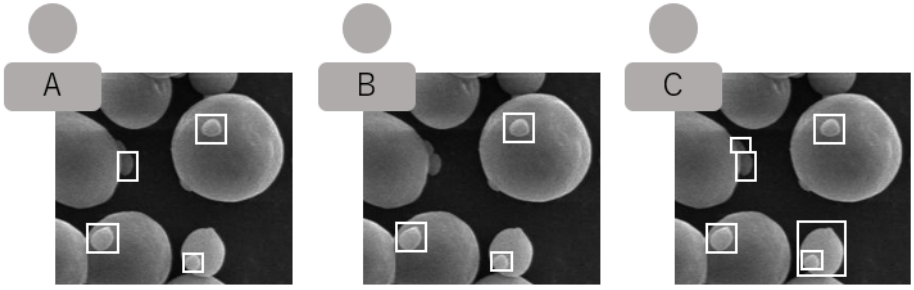


Figure 3. Area examples of asteroid powders each evaluator marks for the microscope images.

A METHOD LEADING CRITERIA TO COINCIDENCE

Diagram of the Method

The paper proposes a method to unify criteria varying with evaluators to lead their labels into coincidence. The method plays a vital role to detect products inferior in their quality. Previous works such as [10] have constructed individual models to detect inferior products because the labels themselves differ for each evaluator. Though product quality in the works depends on each model, it is desirable to unify the evaluator's criteria to stabilize product quality. The proposed method contributes to standardizing product quality. Figure 4 shows a diagram of

the method used in this study. It is evaluated whether the method leads to criteria standardization, taking a quality check task of paint material as an example.



Figure 4. Diagram of the method.

The paper explains the method using an example of the quality check task of paint which detects asteroid powders. Generally, a part of a dataset is labeled to be used for training. In the example, evaluators assign their own labels to the microscope image data indicating the occurrence of asteroid powders.

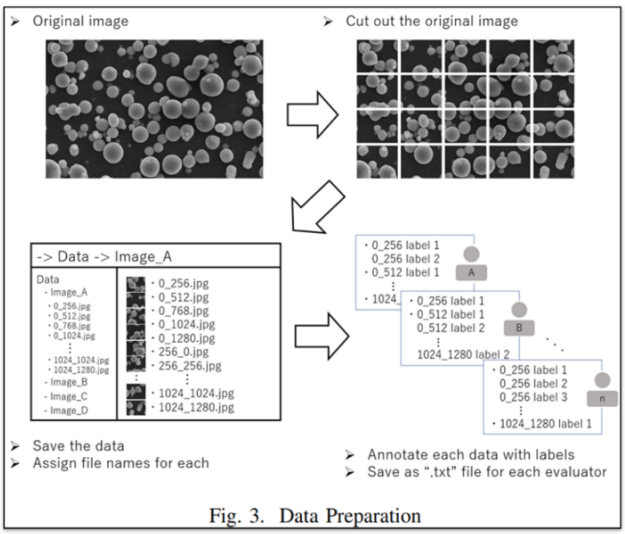
In the proposed method, to evaluate whether the criteria are standardized, the whole dataset is split into 2 parts: one for model construction specific to each evaluator and the other for criteria standardization.

In a microscope image in the former, each evaluator labels powders judged as asteroid ones with binary flags of 1 or 0. The method synthesizes the label information for all evaluator datasets. From various evaluator labels for the same image, a single labeled dataset is synthesized with powders judged as asteroid ones by more than a threshold number of evaluators. The paper refers to it as a generalized labeled dataset, which represents criteria agreed upon by many evaluators.

The results of comparing the contents of the generalized labeled dataset with the information on each evaluator' s label are feedback to each evaluator. The feedback makes each evaluator reform his/her criteria to conform to the standardized criteria. After the reform, all evaluators again label datasets left for criteria standardization. Comparing the performance after the reform with that before the reform, the method assesses whether the criteria can be standardized.

Cutout and Labeling

The way of cutout and labeling of images by each evaluator is the same as in the previous work [10]. Figure 5 shows how to create a labeled dataset from an original image.



Source : Wakabayashi and Shimakawa(CSDE2022)

Figure 5. Cutout and labeling refer to Naoya Wakabayashi, and Hiromitsu Shimakawa, *Choosing Suitable Object Detection Method to Examine Pigment Powders of Similar Shape.* CSDE, 2022. [10]

The original image of a microscope is divided into equal sizes. Those are saved in the file with an assigned number. Powder images are cut out into squares, expecting that the model will be trained smoothly. The original images are divided in a way that neither cutout images overlap nor that the powders in the training data are included in the test data.

Every evaluator label all saved images using a tool to save them as a text file. This study uses LabelImg [14] as the tool. If there are asteroid powders in the loaded image, evaluators specify the rectangle areas to enclose the powders. The coordinate positions of the rectangles are saved with text files. All evaluators labeled each image because the evaluators' criteria are different.

Splitting Dataset

This method compares performance using different labels each evaluator assigns to powders in the same image. It is necessary to provide all evaluators with the same images.

In this study, the dataset is divided into train data, validation data, and test data so that they can be used for building a detection model, where train data and validation data are treated as the same. The phase of splitting the dataset is the part shown in Figure 6.

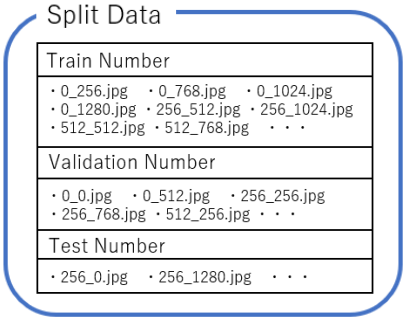


Figure 6. Choose random numbers and split dataset.

In this method, random numbers are prepared in as many images in each dataset. According to the numbers, the method assigns images and accompanying text files commonly to all evaluators. As a result, individual evaluators can assign their own labels to powders in specific images, but the images stored in train data, validation data, and test data are the same for all evaluators.

Label Synthesis

Figure 7 shows how to create a generalized labeled dataset, synthesizing the labels of all evaluators. In this method, train data and validation data are used for feedback, while test data are used for performance evaluation. All image data is used in this phase.

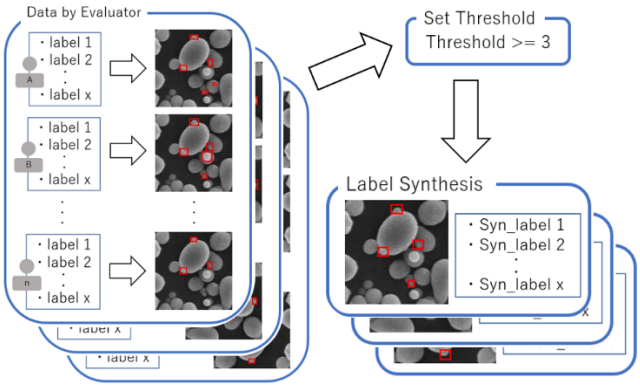


Figure 7. Create a generalized labeled dataset from synthesis labels.

The evaluator assigns labels to each of the multiple images. Each label holds information on the coordinates that specifies a rectangle enclosing a powder. If a specific powder is an asteroid, its rectangle would be pointed out by many evaluators. To synthesize the labels, the method detects the overlap of the rectangles many evaluators point out in the same image. The concept of IoU (Intersection over Union) is used to calculate the overlap. Areas pointed out by more than a certain number of evaluators are regarded as ones many other evaluators would also detect. To represent the amount, the method sets a threshold of the number of coincident evaluators out of the whole.

In the example in Figure 7, it is assumed that asteroid powders are pointed out by three or more people out of six. The threshold setting should be accommodated depending on whether the recall or the precision is more important. The majority is used as the criterion in this study. The labeled dataset obtained through synthesizing labels of all evaluators serves as a criterion to conform to many evaluators.

Feedback and Relabel

Figure 8 shows how each evaluator corrects its criteria based on feedback from the synthesizing labels. This phase uses the training dataset labeled by each evaluator, training data and validation data by the generalized labeled dataset, and test images.

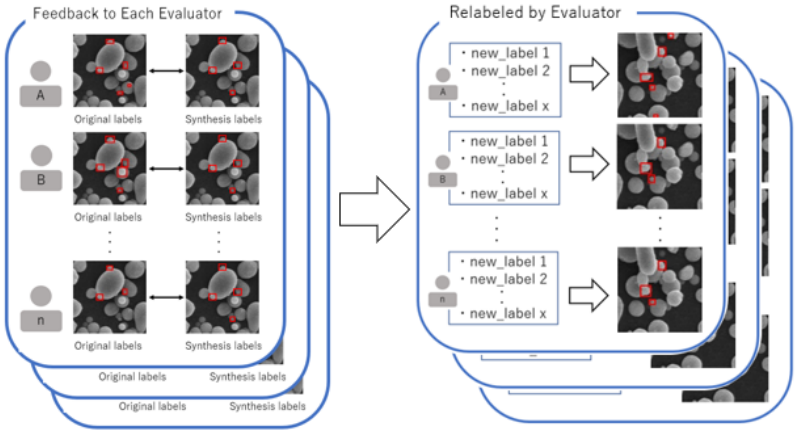


Figure 8. Relabeled by each evaluator who gave feedback and reformed.

The method feedbacks the difference of the training dataset from the generalized labeled dataset to each evaluator. The difference is the coordinate positions of the evaluator labels that do not overlap with the coordinate positions marked in the generalized labeled dataset. All coordinates are visualized in the image so that each evaluator can understand easily when they are feedbacked. Evaluators reform their own labels to match the generalized labeled dataset. They should record the reason for each of the reformed items.

The evaluator labels the images in the test data after the reforming is completed. The tools used for the labeling are the same as mentioned in 3.2. The results are saved as a text file and used for result comparison.

Comparison of Results

Figure 9 shows the method used to evaluate whether all evaluator criteria are standardized in this method.

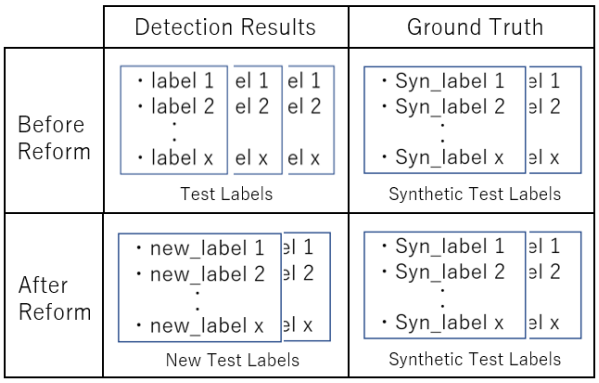


Figure 9. Comparison of results before and after the reform.

Performance and its variance need to be known to compare the individual criteria with the standardized criteria. It is expected that the proposed method would reform criteria that vary with individual evaluators closer to the standardized criteria.

Let the test labels assigned by individual evaluators be detection results representing the individual criteria. On the hand, labels given by models trained with generalized labeled dataset are regarded as ground truth which represents the standardized criterion. The former is referred to as test labels, while the latter as synthetic test labels. The performance of each evaluator model after the reform is compared with that before the reform to confirm that the reform prevents the performance from degrading. On the other hand, to know whether the method contributes to standardizing criteria among evaluators, the variance of performance of each evaluator is examined before and after the reform.

Whole of Field Method

A series of workflow is shown in Figure 10 when this method is applied in the field.

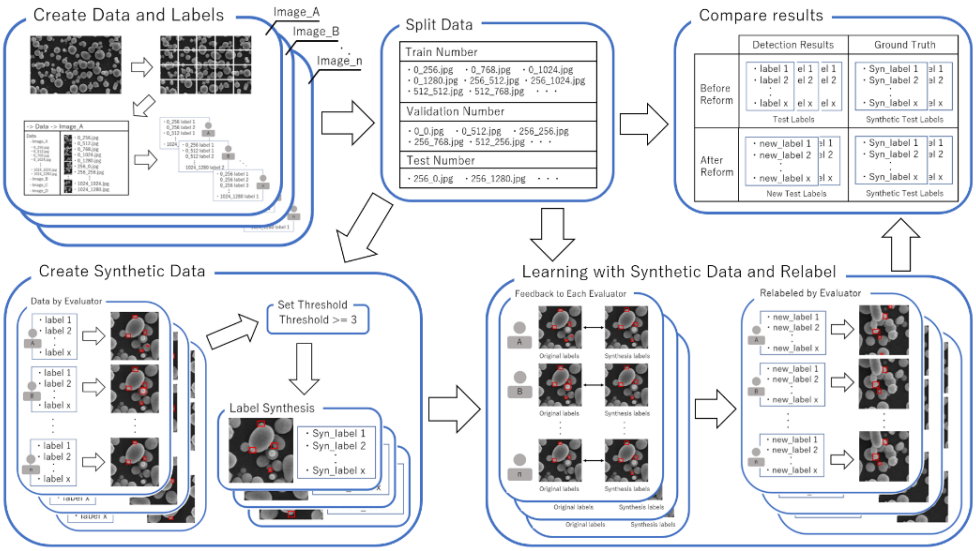


Figure 10. A series of workflow.

Small image data are cut out of the original microscopy image. Each evaluator assigns correct labels to the small image data to create a dataset. The dataset is divided into train data, validation data, and test data to evaluate the standardization. To evaluate the performance fairly, common small image data are provided for all evaluators, each of whom produces the individual labels.

Generalized labels are synthesized from the labels obtained from each evaluator. Values of the generalized labels, which represent the presence or absence of asteroid powders, are determined by most evaluators. Based on the Delphi method, evaluators reform their criteria from the generalized labels. They should describe how they reform their criteria because the description would be analyzed to formalize general rules in the future. Evaluators label the images in the test data according to reformed criteria.

The performance of the detection is examined to assess whether the criteria are generalized with the reform. The reform should not degrade the performance of each evaluator significantly, while the variance of the performance for all evaluators should be smaller after the reform.

DETECTION

Detection Environment and Flow

A method is proposed to standardize discriminant criteria that vary with evaluators. An experiment verifies that the method contributes to converging the different criteria to a certain standard.

In the experiment, six evaluators assigned labels with a total of 75 image data cut out from original micrographs. It consists of 40, 25, and 10 data as training data, validation data, and test data, respectively. The training data are used for the learning of object models, while the verification data are for the adjustment of their hyper-parameters. This study used train data and validation data are treated as the same because the model is not built. The numbers of image data for all evaluators are the same, but the correct labels vary with evaluators.

For every image data, the method synthesizes a generalized label from the ones of all evaluators. For the label synthesis, it is assumed that the same object is located where the IoU of the banding boxes indicating the object's location is greater than 0.5. The IoU is sensitive to the discrepancy between the bounding boxes. The standard value, also used in the PASCAL VOC dataset [15] and elsewhere are 0.5 [16]. The study adopts the threshold as 0.5.

Training data and validation data from the generalized labeled dataset are used for the feedback. Test data from the generalized labeled dataset is regarded as the ground truth for evaluating performance. The synthetic labels of the test data are not disclosed to the evaluators to avoid them from affecting the evaluators on reforming labels in the feedback.

The banding boxes of the generalized labeled dataset for the training data and the validation data visualize the difference from the bounding box the individual evaluators have presented. Figure 11 shows an example of the visualization of the differences in Evaluator1.

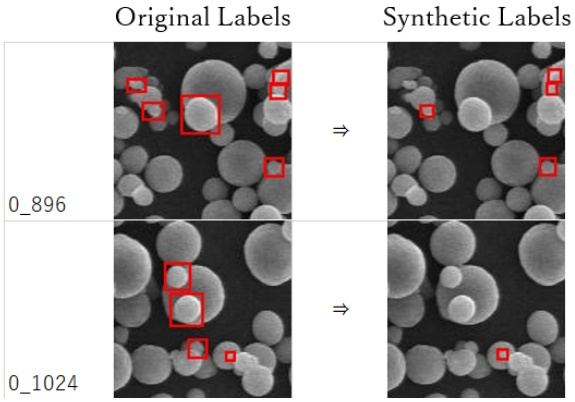


Figure 11. Example of the visualization in Evaluator1.

Each evaluator reforms labels in the training data and the verification data, referring to the differences as shown in Figure 11. The evaluators confirm how they reform their criteria. The evaluator also labels the test data, according to the reformed criteria, to obtain its coordinate information. The above procedure provides 10 relabeled test data for identical images from each of the six evaluators.

The proposed method examines detection results with/without the generalized labeled dataset to evaluate whether the criteria are standardized. To assess the variance in the detection, the number of TP (True Positive), FP (False Positive), and FN (False Negative) is used as the evaluation metrics. The TP refers to the percentage of correctly identified asteroid powder, where the larger the better. The FP refers to the percentage of normal powders misidentified as asteroid powders. The FN refers to the percentage of unrecognized asteroid powders. In the latter 2 metrics, the smaller the better. In addition, the F1-score, which is the harmonic mean of the recall rate and the precision rate, is used to evaluate the accuracy of each evaluator.

The value of the evaluation metrics is calculated by an overlap between rectangles obtained from the evaluators' labels and those from synthetic labels. The verification uses the mAP package [17], which is available on GitHub.

The threshold for the IoU is set at 0.5, which is the default. The package can calculate the mAP (mean Average Precision) from the rectangle coordinates given by the object detection model and those of the synthetic labels. The package provides the TP and the FP simply. The FN is calculated by subtracting the number of TP from the total number of correct labels. The precision, recall, and F1-score are calculated using equation (1), (2), (3).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{F1-score} = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} \quad (3)$$

Detection Results

The proposed method is expected to lower the variance among evaluators after the reform, avoiding the value of the F1-score from changing significantly before and after the reform.

Table 1 shows the results of comparing the labels from the generalized labeled dataset with the labels from each evaluator's criteria.

Table 1. Compare the labels from "generalized labelled dataset" – "each evaluator's criteria".

	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Evaluator 5	Evaluator 6	**Average**	**Variance**
TP	25	27	23	18	23	19	22.5	9.92
FP	16	19	11	0	5	4	9.17	45.81
FN	3	1	5	10	5	9	5.50	9.92
Precision	0.61	0.59	0.68	1.00	0.82	0.83	0.75	0.0208
Recall	0.89	0.96	0.82	0.64	0.82	0.68	0.80	0.0126
F1-score	0.72	0.73	0.74	0.78	0.82	0.75	0.76	0.0012

Evaluator4 and Evaluator2 present the highest precision and the highest recall, respectively. It means objects detected by Evaluator4 are almost always asteroid powders, though many asteroid powders are left undetected. On the other hand, Evaluator2 detects most asteroid powders along with frequent wrong recognition. The highest F1-score is found in Evaluator5, whose precision and recall are ranked in the middle in the order. From the viewpoint of the variance among evaluators, the FP exceeds others largely.

Table 2 compares the labels reformed by the Evaluators referring to the generalized labeled dataset.

	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Evaluator 5	Evaluator 6	**Average**	**Variance**
TP	20	21	23	26	21	19	21.7	5.22
FP	6	10	10	14	6	5	8.50	9.92
FN	8	7	5	2	7	9	6.33	5.22
Precision	0.77	0.68	0.70	0.65	0.78	0.79	0.73	0.0030
Recall	0.71	0.75	0.82	0.93	0.75	0.68	0.77	0.0067
F1-score	0.74	0.71	0.75	0.76	0.76	0.73	0.74	0.0004

The precision is relatively high in Evaluators 1, 5, and 6, while recall is the highest in Evaluator 4. It is notable, Evaluator 4, who has the highest precision and the lowest recall in Table 1, presents the highest recall and the lowest precision in Table 2. It means the evaluator significantly reforms the criteria. The F1 scores get slightly lower, but it is not worth to be noted. The swings of the precision and the recall get smaller. All evaluators show significant changes in all values. The fact indicates that the proposed method has changed the criteria in some way.

To evaluate whether the accuracy has changed significantly, Table 2 is compared with Table 1. The accuracy is assessed by the F1 score. The F1-score in Table 2 is lower than in Table 1, but the difference is only 0.02. It is not a significant change. The labels in the generalized labeled

dataset are determined by majority voting over the individual labels. They correspond to criteria many evaluators approve. Naturally, the precision and the recall get higher than those before the method is applied, which improves the F1 score. These values are likely to decrease after applying the method because the labels biased to each evaluator are compared with the unbiased labels. However, the average of all evaluation metrics has not changed significantly. The small difference before and after reform is a meaningful result.

Next, let us examine Table 1 and Table 2 to evaluate whether the variance gets smaller. The variances of TP, FP, and FN get all smaller. In particular, FP lowers to about one-fifth. It is a large difference. Though the FP of Evaluator 4 gets larger, that of Evaluators 1 and 2 gets clearly smaller which is larger in Table 1. The detection criteria differing from the individual-to-individual result in large swing ranges of the FP in Table 1. However, the swing range is reduced by reformed criteria in Table 2. The variance of the precision, the recall, and the F1 score also get smaller. The variance is small because of small values of FP and FN used to calculate them are small. The values of each evaluator are converged to the average. The F1 score, which is important to evaluate the accuracy, is also averaged out with a small swing, 0.05.

The variance for all evaluation metrics gets smaller, preventing any significant change from occurring in the average accuracy. It suggests that the criteria for each evaluator are standardized through the proposed method.

DISCUSSION

Build a Generalized Model

In the generalized labeled dataset, powders judged to be detected by more evaluators than the threshold are labeled as asteroid powders. The labels are regarded to represent the common wills of many evaluators. The generalized labeled dataset can build detection models based on standardized criteria. This section discusses the performance of the model.

This section constructed YOLOv5m model at Google Colaboratory [18] with a good balance of precision and recall based on previous studies [10,19]. The performance index is the mAP which is commonly used in the evaluation of object detection methods. The best model, which has the lowest loss during training, is used to evaluate performance. Letting the parameters be the same as in the previous studies, codes are taken from the YOLO package [20] and mAP package [17] on GitHub to improve reproducibility.

Table 3 shows the performance of the generalized model.

	Train	Validation	Test
mAP	98.28	54.15	61.12

The mAP for the training data exceeds 98%, which indicates the training is sufficient. The mAP for the validation data has a low value of 54%. This method makes the model overfit because parameters are adjusted with a small amount of image data. It is unavoidable in the detection of similarly shaped objects. Nevertheless, the mAP for test data exceeds 60%, which indicates a certain amount of asteroid powder is detected.

The criteria for everyone are likely to be highly blurred, unlike the individual model whose criteria are consistent. If the blurring of the criteria is large, the performance for the test data might be low. However, it has no significant difference from the performance of individual models shown in the previous studies. The blurring of standardized criteria is likely to be almost identical to the individual models, which suggests a high degree of uniformity.

Making Manual Through Text Analysis

It is expected the text analysis of reform reasons of criteria of each evaluator would lead to formalizing the reasons. The analysis result would enable us to produce a manual explicitly describing criteria to distinguish asteroid powders from others. Text analysis is a technique that examines the frequency of occurrence of words or word sequences. It explores the characteristics of reasons for which the target object is not detected and reasons for misdetection. It facilitates formalizing the criteria to detect target objects.

It also contributes to the generalization of the determination of thresholds that would be implementer dependent in the proposed method. This method records the reason how each evaluator reforms criteria according to the steps explained in 3.5. Evaluators describe the records when comparing generalized labeled datasets with their own data. The reasons for the misrecognition should be described in detail here, not the reasons for the correct recognition. False recognition is defined as the case where a target object is present but not recognized, or where a target object is not present but recognized. The records are subjected to text mining [21]. The evaluator discusses the reasons to achieve the most promising procedures to prevent misinterpretation. It contributes to clarifying the criteria to avoid misinterpretations, which facilitates the creation of manuals. The manual founds on the obvious criteria which is derived through text mining along with discussion among evaluators. Once sufficient records are accumulated, the method does not need to be repeated many times. The procedure is useful from a long-term perspective.

CONCLUSION

This paper proposes a method to standardize evaluation criteria that differ for individual evaluators, according to the Delphi method. It reformulates the evaluators' criteria using a generalized labeled dataset created through the synthesis of their labels. The validation results show that the variance of all the evaluation metrics is reduced without significant change in the accuracy before and after the application of the proposed method. It is found the proposed method can be used to standardize the evaluator's criteria.

The validation uses 70 images, each of which is labeled by 6 evaluators. Significant results are obtained from a small sample size. Expanding the number of evaluators and the number of data, the method would get more effective since the variance among evaluators is expected to converge.

In the future, text analysis methods are applied to records stating reasons for criteria reforms of many evaluators after applying this method. Along with discussion among evaluators, the analysis would bring a manual to avoid misrecognition.

References

[1]. L. Elisa Celis, and Vijay Keswani, *Implicit Diversity in Image Summarization.* Proceedings of the ACM on Human-Computer Interaction, 2020. 4(139): p. 1-28.

[2]. Thomas Deselaers, et al., Jointly optimising *relevance and diversity in image retrieval.* CIVR '09: Proceedings of the ACM International Conference on Image and Video Retrieval, 2009. 39: p. 1-8.

[3]. Chen Karako, and Putra Manggala, *Using Image Fairness Representations in Diversity-Based Re-ranking for Recommendations.* UMAP '18: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, 2018. p. 23-28.

[4]. Zhiming Liu, et al., *ML-CookGAN: Multi-Label Generative Adversarial Network for Food Image Generation.* ACM Transactions on Multimedia Computing, Communications, and Applications, Just Accepted, 2022.

[5]. Lianli Gao, et al., *Lab2Pix: Label-Adaptive Generative Adversarial Network for Unsupervised Image Synthesis.* MM '20: Proceedings of the 28th ACM International Conference on Multimedia, 2020. p. 3734-3742.

[6]. Robert Styrzyński, et al., *Development of paints and varnishes in the face of dynamic changes of end users' needs - challenges for raw material producers.* 2019.

[7]. Frank J. Maile, et al., *Effect pigments - past, present and future.* Progress in Organic Coatings, 2005. vol. 54, 3(1): p. 150-163.

[8]. Wei Liu, et al., *Cold Bonding Method for Metallic Powder Coatings.* Materials, 2018. vol. 11, 11(11).

[9]. Muhammad Sopyan Yahya, and Isti Surjandari, *Classification of Printing Ink Quality based on Pigment, Dispersion Process: A Case Study.* Proceedings of the 7th North American International Conference on Industrial Engineering and Operations Management, 2022. p. 942-951.

[10]. Naoya Wakabayashi, and Hiromitsu Shimakawa, *Choosing Suitable Object Detection Method to Examine Pigment Powders of Similar Shape.* 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2022.

[11]. Joseph Redmon, et al., *You Only Look Once: Unified, Real-Time Object Detection.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. p. 779-788.

[12]. Harold A. Linstone, and Murray Turoff, *The Delphi Method: Techniques and Applications.* Addison-Wesley, 1975.

[13]. Kari K. Lilja, et al., Using the Delphi method. 2011 Proceedings of PICMET '11: Technology Management in the Energy Smart World (PICMET), 2011. p. 1-10.

[14]. Tzutalin Darrenl, *The labelImg package.* [Online]. Available: https://github.com/heartexlabs/labelImg.

[15]. Mark Everingham, et al., *The Pascal Visual Object Classes (VOC) Challenge.* International Journal of Computer Vision, 2010. 88(2), p. 303-338.

[16]. Li Liu, et al., *Deep Learning for Generic Object Detection: A Survey.* 2020. 128(2), p. 261-318.

[17]. Cartucho, *The mAP package.* [Online]. Available: https://github.com/Cartucho/mAP.

[18]. Google, *Google Colaboratory.* [Online]. https://colab.research.google.com.

[19]. Naoya Wakabayashi, and Hiromitsu Shimakawa, *Detection of Powder Granule Attachment Using Object Detection Methods.* FIT2022, In Japanese, 2022.

[20]. Ultralytics, *The yolov5 package.* [Online]. Available: https://github. com/ultralytics/yolov5.

[21]. Michael W. Berry, *Survey of Text Mining: Clustering, Classification, and Retrieval.* Springer, 2004.