Metallographic Specimen Imaging Classification: A Machine Learning Approach

Fabrizzio Alphonsus A. M. N. Soares*, Manoel I. Queiroz Junior†, Rogerio Salvini*
*Instituto de Informática, Universidade Federal de Goiás
Goiânia, GO, Brazil 74.690–900
†Instituto Federal de Goiás, Campus Goiânia
Goiânia, GO, Brazil, 74.050–110

***Abstract*—Metallography is a field of study focused on metal analysis of microstructure, defects, etc, and material identification. ASTM International provides E112 protocol [1] to support material observation based on average grain size. This method requires to count total of grain cut on a circular area of $645\,mm^2$ or $1\,inch^2$ and following directions to identify the material. However, this process demands high accuracy and knowledge, it is very handwork and subject to human errors. Moreover, previous knowledge about the material is required to choose the most suitable protocol. In this work we present an approach for metallographic specimen identification based on imaging classification with classic machine learning algorithms. We prepared specimens following ASTM [2] for six different materials and collected sample images on a microscope. We compared K-Nearest Neighbor, Decision Tree and Linear Discriminant Analysis algorithms, using flatten raw pixels, gray histogram and GLCM features as input data. Our experiments were performed with 1,200 patch samples with different pixel set size reaching an average accuracy of 96.8%. Thus, the proposed approach presents a path toward automated metallographic studies.**

***Index Terms*—Metallography; Image Classification; Machine Learning.**

I. Introduction

Metallography studies constitution, texture and structure of metals and their product alloys. Moreover, it studies mechanical, physical and chemical properties of materials and their manufacture process. Since material properties are intrinsically related to their microstructure nature, observing material microstructure allows to have a qualitative view of different micro-constituents and defects and also, to identify their most likely characteristics and properties.

First metallic structure observation dates from 1863 when Henry Clifton Sorby [3] spent almost three years making a huge number of preparations of steel in order to get true structure without artifacts on a microscope. Nowadays, ASTM International[1] provides ASTM E3-11 protocol [2] to direct material scientists on all steps for metallographic studies similar to Sorby's approach.

Material alloys, or even pure materials, present different structure, since they can have specific grain boundaries, phase boundaries, inclusion distribution, and so forth. Thus, during microscope observation, material engineer has to focus on many details to get a better identification.

[1]Formerly known as American Society for Testing and Materials.

A prior supposition about material helps on choosing the most suitable protocol to gather the best specimen characterization. There are a lot of protocols to direct recognize material structure, for instance ASTM E112 [1], which provides directions to determine average grain size for metals. However, observing material structure is hard even for the most trained engineer, since different materials have different characteristics and specific protocol. Learning the whole set of protocols and procedures for all materials, is not just almost impossible, but also useless since, protocols are reviewed frequently.

Since metallography studies relies on imaging observation and decision making by observers, this work presents an approach for metallography of commercially available materials by image classification with machine learning algorithms. Our main contribution is to provide a simple and efficient method to identify materials without any assumptions about feature present on them.

The text is organized as follows: Section II describes some relevant approaches available in the literature related machine learning and metallography. Section III presents metallographic specimen and dataset preparation for method validation in our experiments, described in Section IV. Section V describes and analyzes the experimental results and discussion about them. Finally, Section VI concludes our work and presents directions for future work.

II. Literature Review

Although metallographic study requires knowledge and experience, it also requires patience, since material specimen preparation process demands careful steps since material selection until observation. Disclosing microstructure through an microscope comprises on getting a very plane and well polished specimen and highlighting its microstructure with a preferential attacking method [1]. Then, once images are gathered on microscope, they are studied using visually observations with methods based on planar defect characteristics, such as comparative evaluation, grain counting and intercepts method. Grain counting, for instance, consists in counting number of cut grains by a circle with $645\,mm^2$ ($1\,inch^2$) on a $100\times$ magnified image, which is pretty hard work, although very accurate. With this information, one can verify grain size and then identify the material basted on table available on protocols [4].

978-1-5386-2410-4/18/$31.00 ©2018 IEEE

(a) Cut (b) Sanded to #80 (c) Sanded to #320 (d) Sanded to #600 (e) Sanded to #1,200 (f) Polished (g) Acid etch

Fig. 1. Microscope image of 1020 Steel (magnified 100x).

Last years, researchers have been working on computational methods for material classification. Sundararaghavan et al (2004) [5] proposes a statistical method which combines histogram for intercept length distribution, PCA and Support Vector Machines to classify microstructure images on the basis of extracted grain shape and size features. They applied on 375 raw images of single phase polyhedral microstructures with 256 gray levels and divided on 11 classes based on grain shapes. Authors reports between 92.53% to 95.82% of accuracy. Local variation of grain size was used by Letho et al (2016) [6] in order to investigate how it influences the mechanical hardness property in two weld joints materials. A bag of visual features was used by Decost and Holm (2015) [7] to create generic microstructural signatures of materials such as, cast iron, brass, and so forth and reported results greater than 80%. DeCost et al (2017) [8] used Support Vector Machines supported by two image representation approaches: Convolutional Neural Networks and a mid-level image feature based on Bag of Visual Words. As dataset, it was used about 961 scanning electron microscopy of commercial Ultrahigh Carbon Steel classification and they had accuracy from 45% to 98.3% based on different methods and datasets groups.

III. Metallographic Specimen and Dataset Preparation

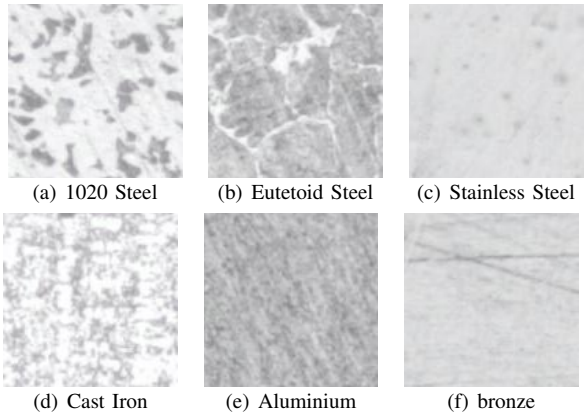
A. Specimen Preparation

Sample materials were prepared following ASTM E3-11 standard protocol [2]. We made image samples from 1020 steel (carburized steel), eutetoid steel, stainless steel, commercial bronze (coper-zinc alloy), aluminum and cast iron.

Firstly, we cut the materials in small samples, about 1 cm height, performed a small leveling and removal of burrs in an emery and then inlaid them in bakelite. Next, specimens were sanded with semi automatic sander with water sandpaper from medium to super fine grains (#80, #320, #600, #1,200) and polished with semi automatic polisher moistened on suspended alumina (aluminum oxide) $1\mu m$. Finally, we perform acid etching with nital $125ml$. Figure 1 presents image samples of 1020 steel $100\times$ magnified from cut to acid etching. Presented images are only illustrative and are off-scale.

B. Dataset Preparation

In all phases, images were taken in gray-scale in 50, 100 and 1000 times magnification lenses for analysis with Olympus BX51M microscope and PL-A662 Pixellink camera. However, only images after acid etching stage are used, since, before it, material microstructure is not revealed. Also, only microscope images magnified at $100\times$ are used to comply ASTM E3-112 protocol [1] for microstructure analysis. Grayscale images



(a) 1020 Steel (b) Eutetoid Steel (c) Stainless Steel

(d) Cast Iron (e) Aluminium (f) bronze

Fig. 2. Sample patches for classification.

were collected in many lighting condition adjustment and in different parts of material and light conditions.

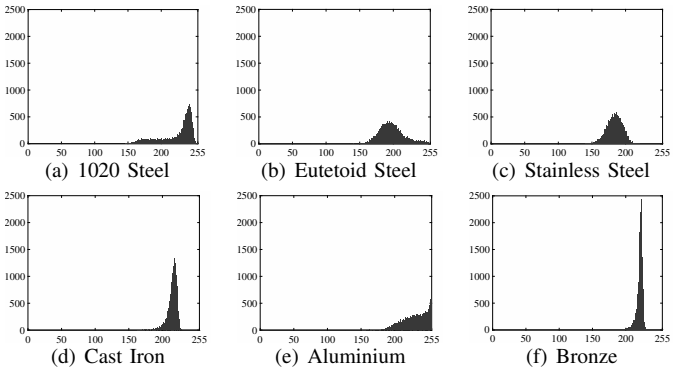
We prepared 4 datasets with patches of 32×32, 64×64, 128×128 and 256×256 pixels. Thus, original images were splitted into 200 random selected patches of each specimen dataset. Figure 2 shows sample patches of 128×128 pixels of specimens, after acid etch, used for classification.

IV. Proposed Method

A. Feature Extraction

As input of classifiers, three types of vector data are used, flatten images, gray level histogram [9] and Grey-Level Co-Occurrence Matrix (GLCM) [10].

1) Raw Pixels: Raw pixel of image patches are used as input. Thus, we flattened all patches creating input vectors for each dataset with 1,024 values ($32 \cdot 32$), 4,096 values ($64 \cdot 64$), and so forth.



(a) 1020 Steel (b) Eutetoid Steel (c) Stainless Steel

(d) Cast Iron (e) Aluminium (f) Bronze

Fig. 3. Histogram of sample images.

2) histogram: As a second type, gray level histograms of all patches are used. We highlight that in this approach, vector size of each sample input are same, so histograms were built

with 256 bins. Figures 3(a) to 3(f) present histogram sample for a patch of each material.

3) GLCM features: Gray Level Co-ocurrence Matrix is used for extracting statistical texture features like contrast, correlation, energy and homogeneity. Statistical features are calculated with Equations 1, 2, 3 and 4,

$$Contrast = \sum_{i,j} |i-j|^2 p(i,j) \tag{1}$$

$$Correlation = \sum_{i,j} \frac{(i-\mu i)(j-\mu j)p(i,j)}{\sigma_i \sigma_j} \tag{2}$$

$$Energy = \sum_{i,j} p(i,j)^2 \tag{3}$$

$$Homogeneity = \sum_{i,j} \frac{p(i,j)}{1+|i-j|} \tag{4}$$

where, $p(i,j)$ is value of each pixel of each patch.

B. Classification

In this work, we compared K-Nearest Neighbors (KNN) [9], Decision Tree (DT) and Linear Discriminant Analysis (LDA) [11] algorithms on metallographic specimen classification. Euclidean distance is used as metric for K-Nearest Neighbors, as on Equation 5,

$$d(S,T) = \sqrt{\sum_{i=1}^{n} (S_i - T_i)^2} \tag{5}$$

where, d is the distance between S and T input feature vectors.

For model classification evaluation a 5-fold cross validation was applied, and performance parameters such as accuracy, sensitivity, recall and F1-score were calculated. Three sets were used, 3/5 for algorithms training, 1/5 for KNN k-value tuning, and 1/5 for final testing. Since LDA and DT has no parameter to be adjusted, they were tested with same samples of KNN, but tuning data set was ignored. Score of each classifier varies in each run, thus algorithms were run 50 times on each dataset in order to obtain an average score.

Experimentation code to perform classification was developed with python 3.6 and Scikit-Learn 0.19, which includes all necessary code to build KNN, LDA and DT algorithms and train and validate them [12].

V. Results and Discussion

Our experiments were performed 50 times to get a average accuracy of classifiers for all sort of feature dataset, raw pixels, histogram and GLCM, and 32×32, 64×64, 128×128 and 256×256 patch sizes.

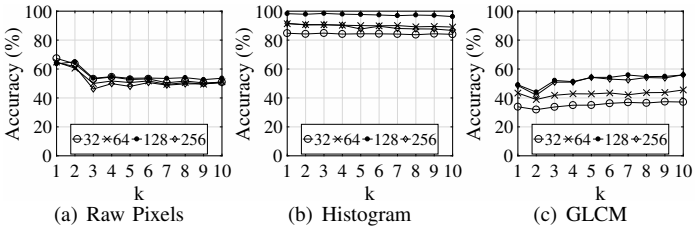
All classifiers, KNN, DT and LDA were trained with 3/5 of dataset with raw pixels, histograms and GLCM datasets. In order to get the best k-value for KNN, 1/5 of each dataset was used to tune it.

In all k-value evaluation experiments, accuracy changing was very small and most of the time same k-value was selected on test. Table I and Figure 4 present best k-value and its accuracy for one single trial.

Table I
Classification Test

Method	32×32		64×64		128×128		256×256	
	k	Accuracy (%)	k	Accuracy (%)	k	Accuracy (%)	k	Accuracy (%)
Raw	1	67.34	1	64.82	2	65.02	1	64.52
Histogram	1	84.75	1	91.32	3	98.45	1	91.61
GLCM	9	37.40	10	45.54	10	55.89	10	56.07

In experiments, for raw pixels best k-value was between 1 and 2, for histogram was between 1 and 3, although all k-values provided very high accuracy, while for GLCM dataset the best k-value was between 9 to 10.



(a) Raw Pixels (b) Histogram (c) GLCM

Fig. 4. K performance evaluation for KKN.

After k-value selection, algorithms were tested with remainder test set (1/5), for each feature type and patch size. Table II presents average classification accuracy for all algorithms, all feature type and patch size.

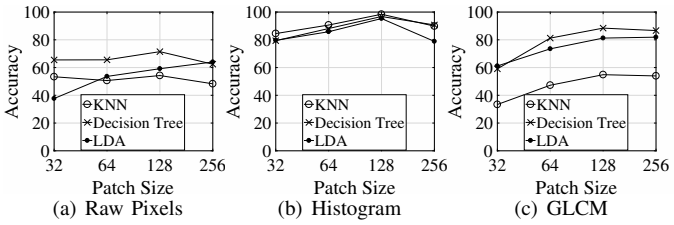
When raw pixel is considered, best results are between 64.0% to 71.5% of accuracy LDA-64 and DT-128, respectively, with DT-32, DT-64 and DT-256 presented values in this range. For histogram, KNN shown best performance on patches between 32 and 128 with accuracy 84.5% to 98.4%, while DT performed better for 256 with 90.7% of accuracy. When GLCM is considered, KNN presented regular performance with accuracy between 33.4% to 54.9% for 32×32 and 128×128, respectively, while DT and LDA presented similar performance with accuracy between 59.1% to 88.4% also for 32×32 and 128×128, respectively.

Table II
Average Classification Accuracy

Feature Type	Patch Size	KNN		DT		LDA	
		Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)
Raw Pixel	32×32	53.4	2.1	65.5	2.6	37.7	2.1
	64×64	50.6	2.4	65.6	3.2	53.6	3.2
	128×128	54.3	1.5	71.5	2.7	59.2	4.0
	256×256	48.3	3.0	62.4	4.5	64.0	4.0
Histogram	32×32	84.5	1.2	79.5	1.0	79.6	1.7
	64×64	90.6	0.7	88.1	1.6	85.8	1.1
	128×128	98.4	0.8	96.8	1.0	95.3	1.7
	256×256	89.9	2.1	90.7	2.7	78.9	4.0
GLCM	32×32	33.4	2.2	59.1	3.6	61.2	2.4
	64×64	47.3	3.0	81.2	1.7	73.5	1.6
	128×128	54.9	2.8	88.4	2.1	81.2	2.0
	256×256	54.0	5.2	86.7	4.0	81.9	5.6

Figure 5 presents average classifier accuracy plot for all datasets. Comparing all feature types, histogram presented the best accuracy for all algorithms and 128×128 patch presented

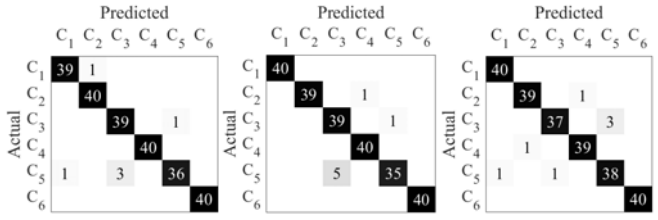
as best patch size for classification. Thus, for material classification, in experiments better feature type and patch size are histogram-128.



(a) Raw Pixels (b) Histogram (c) GLCM

Fig. 5. Classifiers Performance.

Figure 6 presents confusion matrix of KNN, DT and LDA for histogram-128 feature-patch pair. For test data (1/5 - 40 samples of each material), main diagonal shows very high true positives, almost 100% for most of materials, and few confusions in worst case for stainless steel in KNN and DT and bronze for LDA. LDA presented some very small confusion for stainless steel, aluminum and eutetoid steel.



(a) KNN (b) DT (c) LDA

Fig. 6. Classifier confusion matrix for 128×128 pixels. C_1.1020 steel, C_2.eutetoid steel, C_3.bronze, C_4.aluminium, C_5.stainless steel, C_6.cast iron.

Table III presents performance measures of classifications for all classes with dataset pair histogram-128.

Table III
CLASSIFICATION PERFORMANCE MEASURES

Class		Precision (%)			Recall (%)			F1-score (%)		
		KNN	DT	LDA	KNN	DT	LDA	KNN	DT	LDA
C_1	1020 steel	100	98	86	97	100	95	99	99	98
C_2	Eutetoid steel	98	100	94	100	95	80	99	97	86
C_3	Bronze	100	100	100	100	97	100	100	99	100
C_4	Aluminium	87	95	82	97	97	93	92	96	87
C_5	Stainless steel	100	100	98	100	100	100	100	100	99
C_6	Cast Iron	97	95	92	85	97	82	91	96	87
	Average / total	97	98	92	97	98	82	97	98	87

In the table, average precision for KNN, DT and LDA are 97%, 98% and 92%, respectively. Although DT presented greater value, KNN is quite similar, while LDA is also very close. Average Recall and F-Score are greater for LDA and KNN with 98% and 97%, respectively, while LDA presented 82% and 87% for average recall and F1-score.

Finally, we performed a 2-way ANOVA for KNN, DT and LDA for histogram-128 dataset pair for all classes with a p-value 0.05 of significance. The ANOVA presented a p-value of 0.0001 which suggest to reject null hypothesis, thus there is no statistical difference among all algorithms for histogram feature and 128×128 patch size.

VI. CONCLUSIONS

In this paper, we presented an approach for metalographic specimen identification by imaging classification. We made samples following ASTM protocols of 1020, eutetoid and stainless steel, cast iron, aluminum and bronze alloy. Grayscale images were collected to prepare datasets with patches from 32×32, 64×64, 128×129 and 256×256 pixels. A comparing of classification algorithms and feature type dataset was performed in order to find best algorithm. Thus, K-Nearest Neighbor, Decision Tree and Linear Discriminant Analysis were tested for classification, and raw image, gray histogram and GLCM were used as input dataset. In our experiments, histogram format and 128×128 patch size presented as best input type for all algorithms with average accuracy of 98.4% in best case, while KNN, DT and LDA presented excellent performance on this dataset. Furthermore, an 2-way ANOVA was conducted to compare classifiers and they presented no statistical difference between them for histogram-128 dataset. We are planning to experiment on classification of specimens of same material such as, 1020, 1030, 1045 steel, and so forth, which can present itself an even more challenging problem. However, our experiments presented very promising results, showing that approach has very potential to classify the selected material with right feature format and patch size.

ACKNOWLEDGMENT

The authors would like to thank FAPEG (Fundação de Apoio à Pesquisa de Goiás), process number 01/2018, for the financial support.

REFERENCES

[1] A. Standard, "E112: Standard test methods for determining average grain size," *ASTM International, West Conshohocken, PA*, 1996.
[2] ——, "E3-11: Standard guide for preparation of metallographic specimens," *ASTM International, West Conshohocken, PA*, 2012.
[3] K. Geels, "The true microstructure of materials," *Structure: Struers Journal of Materialography*, no. 35, pp. 5–13, 2000.
[4] J. F. Shackelford and M. K. Muralidhara, "Introduction to materials science for engineers," pp. 125–130, 2005.
[5] V. Sundararaghavan and N. Zabaras, "Representation and classification of microstructures using statistical learning techniques," in *AIP Conference Proceedings*, vol. 712, no. 1. AIP, 2004, pp. 98–102.
[6] P. Lehto, J. Romanoff, H. Remes, and T. Sarikka, "Characterisation of local grain size variation of welded structural steel," *Welding in the World*, vol. 60, no. 4, pp. 673–688, 2016.
[7] B. L. DeCost and E. A. Holm, "A computer vision approach for automated analysis and classification of microstructural image data," *Computational Materials Science*, vol. 110, pp. 126–133, 2015.
[8] B. L. DeCost, T. Francis, and E. A. Holm, "Exploring the microstructure manifold: image texture representations applied to ultrahigh carbon steel microstructures," *Acta Materialia*, vol. 133, pp. 30–40, 2017.
[9] J. KIM[1], B.-S. Kim, and S. Savarese, "Comparing image classification methods: K-nearest-neighbor and support-vector-machines," *Ann Arbor*, vol. 1001, pp. 48 109–2122, 2012.
[10] R. M. Haralick, K. Shanmugam *et al.*, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.
[11] G. M. Foody and A. Mathur, "A relative evaluation of multiclass image classification by support vector machines," *IEEE Transactions on geoscience and remote sensing*, vol. 42, no. 6, pp. 1335–1343, 2004.
[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.