Automated forensic ink determination in handwritten documents by clustering

Michael Kalbitz[1,2]
[1]Otto-von-Guericke-University Magdeburg
Faculty of Computer Science
Universitätsplatz 2
39106 Magdeburg, Germany
Email: michael.kalbitz@ovgu.de

Claus Vielhauer[2]
[2]University of Applied Sciences Brandenburg
Department of Informatics and Media
Magdeburger Str. 50
14770 Brandenburg an der Havel, Germany
Email: kalbitz@th-brandenburg.de
vielhauer@th-brandenburg.de

***Abstract*—Even in today's highly digitalized world, the use of handwriting is still widely in use for legal documents such as testaments, contracts, bank cheques or professional certificates. Thus, forgery analysis of handwriting still poses challenges for criminalistics forensic document examiners and one of the investigation questions is, if a questioned document has been written with more than one ink. If so, this may indicate a forgery by manipulations of a possible counterfeiter after production of the genuine original. By means of chemical analysis, it is possible today to identify an ink with almost 100% certainty. However, this process is manual, tedious and needs an initial suspicion by a human expert, that more than one ink was applied on a specific area on the document. Further, most chemical approaches are destructive and limited to very small areas. To improve and to automate the initial investigation on the use of multiple inks, this work proposes a pattern recognition approach based on signal processing, feature extraction and classification by data clustering, which is based on spectral imaging, acquired in almost non-destructive and contact-less manner. The goal is to support forensic examiners by an automated digital detection of regions, which have been written using different ink, which they then can further examine. For experimental evaluation, a benchmark is introduced to evaluate the accuracy of detection results on a specifically created test set, which is also presented. Test results indicate that the best clustering in our investigation has been achieved by the expectation maximisation (EM) approach, with a correct ink cover rate of above 80% for the first and 73% for the second ink in average. Even more relevant for forensic experts is the observation, that false detections occurred in less than 1% of the cases in average. Future work will include extension of data sets and automatic analysis and parameter adjustments in the clustering process.**

***Index Terms*—Forensics, Pattern clustering, Handwriting recognition, Forgery, Spectral analysis, spectroscopy, ink**

I. Introduction

Signal processing and pattern recognition are becoming relevant for a drastically increasing number of real-world applications. This should also have strong impacts to the domain of computer-based crime scene analysis. In this field physical traces on crime scenes, such as finger- and footprints, ammunition, fibre traces and also traces of handwritten documents, are analysed. This analysis based on digital imagery and subsequent signal processing to support forensic experts.

However, to date, the vast majority of forensic investigations are still being performed manual and analogue, especially in the context of handwriting [1] [2], which remains important considering that the use of handwriting is still widely in use for legal documents such as testaments, contracts, bank cheques or professional certificates. Thus, forgery analysis of handwriting still poses challenges for criminalistics and research is underway as how image processing and machine learning approaches can support forensic document examiners (FDEs) to improve the investigation results and/or to decrease their investigation time. Typical goals of a FDE investigation include the determination, if a questioned document is genuine or forged, whether or not it is a modification of an original writing by falsified amendments, or who was the original writer of the document. Thus, amongst the different possibilities to forge a document, there exists the scenario, where a forger adds smaller written elements such as a stroke (e.g. change the symbol "-" to "+") or single digits (e.g. adding a digit to a number, for example in written amounts of money in testaments) to previously written genuine documents. This scenario is the motivation of the work presented in this paper, since by best knowledge of authors, this work is today still done manually by FDEs. Our goal is to support FDEs work by suggesting automated tools, which are able to determine pre-selections of region of interest (ROI), which have been produced with different writing tools, prior to any chemical analysis (which partly destroys the original trace) and in an automated, non-destructive manner. The idea is, that based on this ROI, the FDE can do further investigations such as a chemical identification of the possible different inks to further prove the observation. Basis of the approach proposed here is pattern recognition based on signal processing of imagery, acquired by a spectroscope which acquires spectral responses from UV–VIS–NIR spectroscope (UV-VIS-NIR), with subsequent feature extraction and data clustering. Further, due to the lack of benchmarking databases in the research domain, a new benchmark framework for experimental evaluation is proposed, including newly acquired data sets. Based on this benchmarking setup, first experimental evaluation results are presented.

The remaining part of the paper is organized as follows: Section II presents a short overview about related work. In section III the methodology is presented while section IV

This work has been funded by the German Federal Ministry of Education and Research (BMBF, contract no. FKZ 03FH028IX5).

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

describes experimental setup, experimental evaluation and results. Discussion of the results, conclusion and future work are presented in section V.

II. Related work

To determine if a document is genuine or forged, one approach is to identify the written ink(s), because the occurrence of more than one ink could indicate post-original amendments, as described in the introduction. For this purpose, Denman et al. describe in [3] an organic and inorganic discrimination of ballpoint pen inks. Based on a Time-of-Flight Secondary Ion Mass Spectrometry, they identify the organic and inorganic components of the ink. From the contribution of these components they identify the ink with 91% accuracy. The experiments have been applied on 24 blue ballpoint pens. In [4] Silva et al. investigate 10 different types of blue and black pens. They report classification accuracy between 91.3% and 100%. For the classification, they propose infrared spectroscopy and linear discriminant analysis. In [5] Nunkoo et al. compare five different methods (thin layer chromatography, Fourier transform infrared, visible spectroscopy, filtered light examination and Raman spectroscopy) for ink analysis. For the experiments, they apply 78 different pens (with black, blue, red and green colour) and report the best result for filtered light examination (over 94% accuracy).

However, most of the work identified is based on a manual pre-selection of ROI by a FDE, thus there remains the requirement to determine ROI automatically, for example by pattern recognition and machine learning approaches. As commonly known, there exist the two concepts of supervised and unsupervised learning in context of machine learning. Especially in cases, where a great or even huge number of training samples exist, methods such as deep learning have recently shown impressive recognition results in various application domains. In application scenarios however, where the number of reference samples is quite low, as is the case in the forensic analysis of handwritten documents, the benefit of supervised learning appears rather limited to date. Consequently, authors have decided to focus on unsupervised learning, i.e. clustering methods for this work. The weka framework [6] provides a powerful tool for this and based on the aim of the paper to identify ROI, different clustering approaches from this toolbox have been studied. In clustering, the goal is to divide all data points of a given sample/observation into similar groups, whereby two different types of clustering concepts exist. In the first type, the resulting total number (N) of groups is resulting only on the actually given data (data-driven). The other type divides the data into a priori given number (N) of groups. In first investigations for this work, the potential of data-driven clustering in weka (canopy [7], cobweb [8] [9] and EM [10] algorithms) has been studied, which have shown that the data-driven types did not deliver suitable results. Therefore, the clustering algorithms for our proposed scheme have been chosen as a priori types provided in the weka toolkit: farthest first (FF) [11] [12], simple k means (SKM) [13] and EM (note EM has the special property that it supports both data-driven and a priori number of groups). The EM approach determines a probability for each data point which indicates the probability of it belonging to a cluster. FF calculates the best possible heuristic for the k-center problem and the common SKM divides the data into k groups (where k is actually a priori number of groups N). For each of the groups the approach tries to minimize the distance to the center for each data entry.

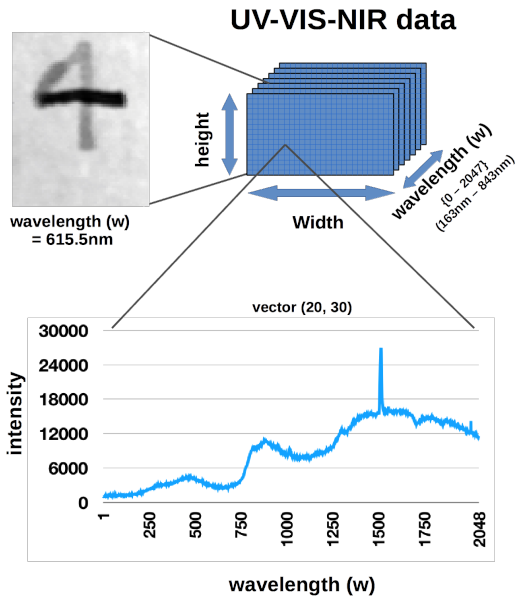

Fig. 1. The data acquisition provides a stack of images (as shown on top right). Each image represents the intensity response for a specific spectrum sub-band/slot of $1/3nm$ of a total range of $160nm$ (ultra violet) to $844nm$ wavelength for entire lateral measurement (x/y) plane. Measurement points are acquired in lateral dot distances of $100\mu m$ vertically and horizontally. For one single lateral measurement point, the spectral intensity across all sub-bands can be modelled as a vector of intensity values (mp), indexed by slot numbers $(1, \ldots, 2048)$. Values of one exemplary vector is shown in the graph on the bottom of the figure. Note that the intensity values on the ordinate (y-axis) do not refer to a specific physical entity as they are solely interpreted as relative values.

III. Methodology

Our proposed ink determination approach follows a four step signal processing pipeline: 1. Data Acquisition, 2. Preprocessing, 3. Feature Extraction and 4. Classification (i.e. Ink Determination).

A. Data acquisition

The data acquisition is done with a UV-VIS-NIR Surface Scanner [14]. This device acquires spectral response for each of the lateral measurement points on the document surface. The spectral values are in the range from $160nm$ (ultra violet) to $844nm$ (near infrared) and are delivered in 2048 data slots. Each slot represents the measured intensity response within the almost linearly distributed spectral sub-bands of width of approximately $1/3nm$. Laterally, the document is digitalized by point measurements with a dot distance of $100\mu m$ in the $2D$ plane (i.e. 10×10 measure points per $1mm^2$ of scan area) and each measure point requires an acquisition time of $500ms$. A detailed description of the structure of the acquired data is given in fig. 1.

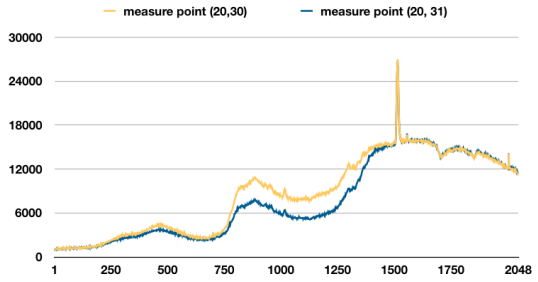

Fig. 2. Intensity response of two neighboured measurement points produced by the same ink. It can observe, that while in the middle wavelengths the absolute values of intensity vary; however, the slopes looks similar. This observation result in the proposed use of derivative features.

B. Pre-processing

The pre-processing is composed of the following steps: first, a low-pass filter removes any responses of wavelength below $200nm$, because the air absorbs UV radiation in this wavelength sector. After that, each measurement point is smoothed by a linear regression filter. The linear regression is applied on the values for the wavelength on each measure point with a window size of 5 in order to remove high-frequency artefacts in the signal.

C. Feature extraction

In initial studies, it has been observed that the use of the spectral intensity response for each wavelength sub-band directly as features for classification did not provide sufficient discrimination between ink classes and background.

However, it has been observed, that the curve shape of graphs of spectral energy across the wavelength is similar for writing areas produced with the same ink (example in fig. 2). Therefore, gradients are determined on the spectral distribution of each measure points and utilised as features in our approach. The determination is approximated by a simplified differential quotient, i.e. given a vector of spectral responses denote as measurement points (mp):
$mp = (w_1, \ldots, w_{2048})$ we calculate the simplified derived vector $mp' = (w'_2, \ldots, w'_{2048})$ with: $w'(i) = w(i) - w(i-1)$ for $i = 2, \ldots, 2048$.

D. Classification: ink determination by clustering

The ink determination is composed of two steps. In the first step the written area (foreground) is determined and separates from the background (plain, unwritten paper) by generating a foreground mask. The second step determines how many inks probably exist in the foreground.

As introduced earlier in the paper, for both classification tasks, clustering approaches are applied. For step 2 (ink determination), all three cluster approaches (FF, SKM and EM) are utilized in different parametrizations, as detailed in the coming section.

IV. Experimental Investigation

This experimental investigation proposes the generation of a new suitable test set. Furthermore, it proposes a benchmark method to evaluate the ink determination accuracy.

A. Data generation

With the best knowledge of the authors, there is no public test database available, containing spectral information of different inks. Thus, an appropriate test set layout was created. Due to the long acquisition durations for each of the samples, the test set layout is composed of rather small writing samples and contains simple stroke crossovers and modifications of numbers. Frauds on numbers are often done by either modifying an original digit or extending the number by an additional digit.

The sensor area is limited to an area of $200x200mm^2$. Therefore, the test set layout contains 126 samples on one page. The samples one to 14 contain four vertical strokes with one ink (ink 1) and one horizontal stroke with another ink (ink 2). The fraud samples 15 to 112 containing modification of digits with the aim to change the value of the former number (14 samples change a 1 to a 4, 14 samples change a 1 to a 7, 14 samples change a 1 to a 9, 14 samples change a 3 to an 8, 14 samples change a 5 to an 8, 14 samples change a 9 to an 8 and 14 samples change a 2 to a 3). The last 14 samples are composed of one or two digits, which are added to a former written number. The former digit / number is written with ink 1. The modification / additional digit is written with ink 2

Based on the above layout, three sets of paper sheets have been generated and scanned to produce the experimental data. These three sets $S1 - S3$ further represent three different degrees of difficulty regarding the ink classification task:

- the simplest set $S1$ contains modifications performed in different high-contrast colours (e.g. red-green)
- the second one in colours, which are closer in spectrum (e.g. blue-black) and
- the third one with different pens/inks but in the same colour (e.g. blue-blue).

For the FDE obviously $S3$ is the most challenging, because the other difficulty degrees can be easily detected by human eyes. Overall the test set contains 36 pens by 9 different types with the colour blue, red, green/lime and black each. The acquisition time varies between 40 and 94 minutes for each sample.

B. Clustering approaches for ink determination

The evaluation is applied on all three clustering approaches (EM, FF and SKM) with default parameters as provided by the weka toolkit and for a set value of $N = 2$, 3 and 4 for the number of different clusters expected. The number of two clusters can be trivially expected, because a forgery in this context is mainly done with one different ink. However, the investigation includes also $N = 3$ and $N = 4$ to study if the clustering approach will find additional cluster which is not based on different ink (false positives).

C. Proposed benchmark for evaluation

After the clustering is done, there is the question of how to measure the quality. Here, there are two main challenges. The first is based on the clustering. There is a priori unknown which cluster will get which label. The second based on the unsharp edge from background to foreground (edge

challenge). For the recognition of a stroke by FDE the precise width/boundary of the stroke is hardly important, because the structural shape of strokes produced by different inks is the main investigation goal here. Therefore, a new benchmark is proposed.

A labeled image is created for the ground truth. The label image contains coded annotations of the position of each ink, overlapped areas of different inks and the background. This mask has the same size as the scan. Therefore, it is possible to determine for each measure point if this represents one or more inks or the background. The results of the clustering processes provide labelled images where each measure point is assigned to a cluster group. Now a labelled image from the cluster exist and a ground truth. But these labels do not represent the same ink. To overcome these, based on the label points in the ground truth mask, ink label of the clustering result will be labeled by the ink from the ground truth mask, having the highest Pearson correlation and will be assigned to this. The number of measure points are summarised to an additional error rate (not assigned cluster), if more number of clusters result from the classification than provided in the ground truth.

A tolerance area was defined at the edges to overcome the edge challenge. Based on the width of the strokes the standard deviations ($\sigma_{stroke\ width}$) for each sample is determined. The area $+/-3\ \sigma_{stroke\ width}$ from the edge is defined as tolerance area and are not considered for calculation of false clustering rates, as detailed below.

False negative rate (FNR), false discovery rate (FDR) and cover rate (CR) are determined for the numeric evaluation. While FNR and FDR are common error measurements in classification evaluation, CR is proposed as an alternative to true positive for benchmarking this specific application scenario. FNR is defined as $\sum false\ \ negative\ \ /\ \ \sum condition\ \ positive$. Where $\sum false\ \ negative$ denotes the number of measure points which are clustered as background, but they are actually assigned as ink in the ground truth mask (except the tolerance area). $\sum condition\ \ positive$ denotes the number of measure points defined as ink from the ground truth mask (again except the tolerance area). FDR is defined as $\sum false\ \ positive\ \ /\ \ \sum condition\ \ positive$. $\sum false\ positive$ are the number of measure points which are clustered as ink but are placed in background area with regard to the tolerance area. FNR and FDR will be determined separately for each ink. CR is proposed as follows. For the CR a skeleton is created from the label of each ink (separately) from the ground truth by reducing the width of the ink to 1 (skeletonize [15] [16]). The CR is then defined as $\sum covered\ measure\ points\ /\ \sum skeleton\ measure\ points$. $\sum covered\ \ measure\ \ points$ are the number of measure points which are clustered as the specific ink and placed in the skeleton from the ground truth for that ink. $sumskeleton\ \ measure\ \ points$ is the number measure point which are placed in the specific skeleton from the ground truth.

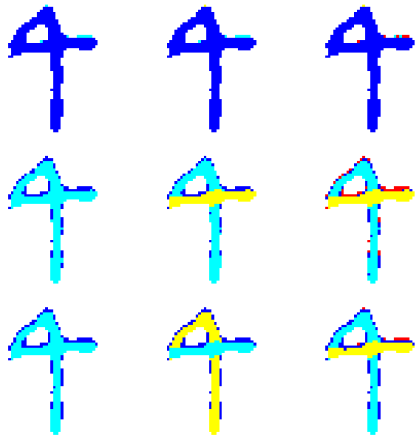

Fig. 3. Exemplary clustering results on S3. Classifiers are as follows: top row FF, middle row SKM and bottom row EM. From left to right right column with N = 2, 3 and 4 as number of clusters. The dark blue and red measure point in the middle and bottom will be count as not assigned cluster.

D. Results

The following results are achieved based on the proposed benchmark. Table I show the results for all difficulty levels ($S1 - S3$). Overall FNR is 1.43%. The first row presents the mean value of each cluster approach. The following rows show the results for the different N $(2, 3, 4)$. It can be observed that with greater N the FDR decreases but the number of not assigned cluster increases. Not assigned clusters are clusters which can not assigned to an ink label in the ground truth. An exemplary result of the clustering results for $N = 2, 3, 4$ is shown in fig. 3. For $N = 2$ no clustering approach detects the second ink. For $N = 3$ and $N = 4$ EM and SKM detect the second ink, but they determine a third cluster on the edge of the ink.

V. Discussion, Conclusion & Future Work

The separation of background and foreground works well; we observe that the average FNR for this as 1.43%. FF achieves a CR of nearly 50% for $S1$ and $S2$. For $S3$ they achieve only 30% and therefore FF seems to be not suitable for that clustering task. SKM and EM achieve at least 65% over all difficulty levels. Also, SKM and EM achieve similar results over all difficulty levels, while the performance of FF decreases with higher level.

The observation, that the best classifications are performed for $N = 3$ and 4 - despite the intuitive expectation of 2 ink classes - needs further studies. For example, an automated decision-making of proper N (number of cluster) for the specific sample should be implemented in further research. Another optimisation option is to investigate other feature extraction or selection methods. For example, at moment the proposed approach applies on all wavelengths above $200nm$. The clustering works maybe better if the wavelengths are limited on wavelengths with high variation between different inks.

TABLE I
OVERVIEW MEAN EVALUATION RESULTS FOR THE DIFFICULTY LEVELS S1 - S3(# not assigned cluster is the number of measure point which can not assigned to an ground truth ink. CR and FDR are shown as percentage. Ink 1 / 2 is the original / modified writing trace as described in section IV-A.)

Difficulty level	Clustering method	# cluster	# not assigned cluster	ink 1 CR	ink 1 FDR	ink 2 CR	ink 2 FDR
S1	FF	mean	22.10	89.96	14.76	46.00	11.59
		2	0.00	86.42	19.95	37.62	19.99
		3	27.50	89.30	14.62	46.90	8.73
		4	38.81	87.76	9.71	53.49	6.05
	SKM	mean	77.63	83.67	2.07	74.59	5.04
		2	0.00	92.32	3.93	78.51	8.52
		3	89.46	82.94	1.25	75.18	3.92
		4	143.43	75.73	1.04	70.08	2.67
	EM	mean	91.53	82.64	1.41	75.19	3.96
		2	0.00	93.05	2.29	82.74	6.39
		3	108.10	81.12	1.15	74.38	3.12
		4	166.49	73.74	0.80	68.44	2.38
S2	FF	mean	18.63	84.62	17.20	41.71	9.92
		2	0.00	85.68	21.65	32.50	18.66
		3	19.48	85.21	16.22	43.50	5.83
		4	36.41	82.96	13.72	49.13	5.27
	SKM	mean	69.60	81.65	1.33	76.93	3.78
		2	0.00	88.21	3.06	82.74	5.92
		3	74.83	81.83	0.64	76.95	2.77
		4	133.95	74.89	0.30	71.09	2.66
	EM	mean	80.24	81.01	1.12	76.24	2.97
		2	0.00	88.35	2.51	84.47	5.13
		3	95.06	80.88	0.61	75.27	2.03
		4	145.67	73.80	0.24	68.98	1.76
S3	FF	mean	12.42	81.58	22.09	29.56	11.64
		2	0.00	83.27	27.65	19.65	13.95
		3	15.35	81.68	20.59	33.12	12.26
		4	21.91	79.80	18.04	35.90	8.71
	SKM	mean	71.87	76.82	6.67	67.87	13.34
		2	0.00	83.62	10.65	68.87	19.33
		3	80.48	77.28	5.52	68.39	11.94
		4	135.12	69.57	3.85	66.34	8.74
	EM	mean	92.90	75.94	5.63	68.86	12.26
		2	0.00	82.74	12.19	70.19	22.05
		3	109.46	75.44	3.31	71.03	9.12
		4	169.24	69.66	1.40	65.37	5.59

ACKNOWLEDGMENT

The authors would like to thank the students of the University of Applied Sciences Brandenburg, who were involved in the creation of the test set layout and the digitalization of the samples as well as the funding agency, BMBF.

REFERENCES

[1] N. Köller, K. Nissen, M. Rieß, and E. Sadorf, Probability Conclusions in Expert Opinions on Handwriting - Substantiation and Standardization of Probability in Expert Opinions, H. Schielke, Ed. Luchterhand Verlag GmbH, 2004.

[2] ENFSI, "Best practice manual for the forensic examination of handwriting," Website, Jun. 2018, eNFSI - BPM - FHX - 01 Version 02. [Online]. Available: http://enfsi.eu/documents/best-practice-manuals/

[3] J. A. Denman, W. M. Skinner, K. P. Kirkbride, and I. M. Kempson, "Organic and inorganic discrimination of ballpoint pen inks by ToF-SIMS and multivariate statistics," Applied Surface Science, vol. 256, no. 7, pp. 2155–2163, jan 2010.

[4] C. S. Silva, F. de Souza Lins Borba, M. F. Pimentel, M. J. C. Pontes, R. S. Honorato, and C. Pasquini, "Classification of blue pen ink using infrared spectroscopy and linear discriminant analysis," Microchemical Journal, vol. 109, pp. 122–127, jul 2013.

[5] M. I. Nunkoo, M. B. Saib-Sunassy, H. L. K. Wah, and S. J. Laulloo, "Forensic analysis of black, blue, red, and green ballpoint pen inks," in Crystallizing Ideas – The Role of Chemistry. Springer International Publishing, 2016, pp. 323–339.

[6] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining. Elsevier Science, 2016. [Online]. Available: https://www.ebook.de/de/product/27213132/ian_h_witten_eibe_frank_mark_a_hall_christopher_j_pal_data_mining.html

[7] A. McCallum, K. Nigam, and L. Ungar, "Efficient clustering of high dimensional data sets with application to reference matching," in Proceedings of the sixth ACM SIGKDD internation conference on knowledge discovery and data mining ACM-SIAM symposium on Discrete algorithms, 2000, pp. 169–178.

[8] D. Fisher, "Knowledge acquisition via incremental conceptual clustering," Machine Learning, vol. 2, no. 2, pp. 139–172, 1987.

[9] J. H. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," Artificial Intelligence, vol. 40, pp. 11–61, 1990.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," Journal of the royal statistical society. Series B (methodological), pp. 1–38, 1977.

[11] Hochbaum and Shmoys, "A best possible heuristic for the k-center problem," Mathematics of Operations Research, vol. 10, no. 2, pp. 180–184, 1985.

[12] S. Dasgupta, "Performance guarantees for hierarchical clustering," in 15th Annual Conference on Computational Learning Theory. Springer, 2002, pp. 351–363.

[13] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of carefull seeding," in Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 2007, pp. 1027–1035.

[14] FRT GmbH, "FRT GmbH – THE ART OF METROLOGY," Website, 2019, available online at https://frtmetrology.com/en, visited on February 15th 2019.

[15] T.-C. Lee, R. L. Kashyap, and C.-N. Chu, "Building skeleton models via 3-d medial surface axis thinning algorithms," CVGIP: Graphical Models and Image Processing, vol. 56, no. 6, pp. 462–478, 1994.

[16] M. Kerschnitzki, P. Kollmannsberger, M. Burghammer, G. N. Duda, R. Weinkamer, W. Wagermaier, and P. Fratzl, "Architecture of the osteocyte network correlates with bone material quality," Journal of bone and mineral research, vol. 28, no. 8, pp. 1837–1845, 2013.