Texture Traits with Uniform-Quantization in Handwriting Documents for Digital Forensics Investigation

IOP Conf. Series: Materials Science and Engineering **645** (2019) 012002 doi:10.1088/1757-899X/645/1/012002

Texture Traits with Uniform-Quantization in Handwriting Documents for Digital Forensics Investigation

D Pratiwi[1,*], Syaifudin[1], T Rahardiansyah[2], A Hilman[1], W Anggriani[1], N Chairunnisa[1]

[1] Department of Informatics Engineering, Trisakti University, Indonesia
[2] Department of Law, Trisakti University, Indonesia

[*] Email: dian.pratiwi@trisakti.ac.id

Abstract. In crime and falsification related to documents are often difficult to verify orauthenticated by the authorities. This is the basis of research to develop a system in assisting digital forensics to investigate and seek the truth of the evidence in the form of digital handwritten documents. The steps that being taken by researchers are collecting and digitizing the documents into image form, converting color from RGB to greyscale, separate the object through thresholding, color histogram, uniform quantization from 256 to 128 of greylevel, texture feature extraction ie variance, skew, relative smoothness, entropy and mean, normalize the feature value and similarity measures through Euclidean distance calculations. From the results of testing of 10 data that has been matched with 20 training data, 6 documents successfully recognized correctly the authenticity of the document owner. Thus, the system in this study produces an accuracy of 60% and can be used to assist digital forensics in analyzing the authenticity of handwritten documents.

1. Introduction

Currently, technology has an important role of various area of life. Technology can be used not only for connect people and build relationship around the world, but also to facilitate in completing a job or various problems. For example in public service area, technology has been using in Akshaya Program [2] to obtain a driver's license or keep the government accountable. Other field, for example, in the healthcare industry, medical technology has been used by many healthcare practitioners to improve their practice from better diagnosis, surgical procedures and improved patient care. And in this study, researcher focus on develop some kind technology in law area, especially for digital forensics to help them solve the problems and various crime cases, especially those involving digital handwritten document fraud crimes in Indonesia.

A Research which related to handwritten document, ever done also by other researchers. As in the study of Plamondon and Srihari [5], which developed a technique of recognition of every word in handwriting either from online or off line through the Hidden Markov method with good results. Then in Gluchev, G [6] various methods of feature extraction approach were introduced to recognize handwriting features as a way of investigating the authenticity of the owner's handwriting. Such as predominant slope feature, quantity of movement, direction of movement, and many others. Research on handwriting analysis has also been done by researchers in 2007 [7] where the researcher succeeded in processing handwritten documents to know the personality type of the owner through Region-

basedShape Feature Extraction with accuracy reaching 81.6%. Then in 2016 [8], researchers have also succeeded in developing a prototype of handwritten document security through Inner Product method with accuracy result obtained from 100 testing document by 72%. From this result, the researcher tries to develop again with the same research theme, where the approach is to use Statistical TextureFeature Extraction combined with Uniform Quantization, and similarity measure technique through Euclidean Distance. Selection of texture-based feature extraction method is due to Chandy's research [9], the result of accuracy is very good that is 85.1% when using texture as input value of irregular, asymmetric, and ill-defined data. Just like everybody's handwriting, which has a characteristic difference in every stroke. This is what then researchers try to apply to handwritten document to produce unique features so that it can distinguish well the handwriting owner to one another. These features are the fineness of strokes, variations of writing, slope and irregularity strokes.

2. Materials and Methods

In this research, we used several methods to develop a mobile system that can match the needs of investigations in the field of digital forensics.

2.1. RGB Color Convertion

RGB color convertion in this study is the stage used to change the true color (24-bits) in handwriting form into greyscale color (8-bits), by calculating the mean value of *R (Red), G (Green),* and *B (Blue)* at each pixels so that it only has a colour interval between 0 and 255 [7].

2.2. Thresholding

Thresholding is one of segmentation methods in image processing which used to group pixels within certain intensity limits. In addition, thresholding also serves to separate the image that matches the object (foreground) and background, and convert the image into binary image to facilitate the next process. In this study, thresholding is used to separate the written object from the background, and the selected value is the middle value of color interval, which the value is 128.

2.3. Color Histogram

A histogram is a graphical representation that shows the gray color distribution in an image [10]. The histogram is calculated by the formula :

$$h_i = \frac{n_i}{n}$$

where $n_i$is the number pixel that have a gray value i ($i = 0 \ldots L-1$), L is maximal color interval, n is amount of pixels in image andh_i is a probability of i gray value [11]. In this research, grey level which used will begrouped again into a certain color scale through the quantization stage.

2.4. Quantization

Quantization is a part of the image digitization stage that classifies the grey level value of continuous image to a certain level. Quantization can divide the large quantity into a discrete number of small parts, often assumed to be integral multiple of a common quantity. There are two types of quantizer, *Non-uniform Quantizer* and *Uniform Quantizer* [12]. Uniform quantizer has the same gray level grouping interval (for example, intensity from 1 to 10 is rated 1, intensity from 11 to 20 is rated 2, and so on). While the non-uniform quantizer is a finer quantization required especially in the image portion that describes the detail or texture or boundary of an object region, and the more rough quantization is applied to the same region on the part of the object. For example, on a satellite photograph of the earth's surface, where densely populated areas are finely quantized, while sea areas are roughly quantized. In this research, the quantization technique which used is *Uniform Quantizer* because handwriting data tend to have simple pattern and detail (not many objects). The formula is $G=2^m$, where G is grey level and m is number of bit.

Quantization will determine the brilliance resolution of an image. If the scale used is too small, the image resolution will be smaller and it may cause blurry, broken or unclear to the image.

2.5. Feature Extraction

Feature extraction (or sometime called by indexing) is a basic step in conducting an image interpretation and classification. There are many ways to extract the feature, depends on the data. In image data, features that can be extracted are color, shape, and texture [17]. And in this study, researcher uses texture. Texture is an intrinsic character of image which related with roughness level, granulation, and regularity of structural arrangement of pixels. Texture is defined as spatial distribution of greylevel in a set of neighbouring pixels. Image feature extraction based on texture in orde one can use statistical method, ie by looking at the greylevel distribution statistic on the image histogram [13]. From the histogram values, it can calculate feature parameters :

a) Variance (σ^2)
   Indicates element variations on the histogram of an image. Variance value is used by researchers to assess the extent of word variaton which present in each handwritten document. The greater the value of variance, the more varied the existing pattern in writing. The formula is :

$$\sigma^2 = \sum_n (f_n - \mu)^2 p(f_n)$$

   Where is an average value of pixels on image, is a value of grey intensity, ($p(f_n)$ is a value of histogram (intensity occurrence probability in document image) and is σ^2a value of variance

b) Skewness (α_3)
   Skewness will indicate the relative historic level of the histogram curve of an image. In this research, value of skewness will be used to assess the inclination of the stroke direction of a post on each document. If the value is close to 0, the direction of the skew is symmetric. The formula is:

$$\alpha_3 = \frac{1}{\sigma^3} \sum_n (f_n - \mu)^3 p(f_n)$$

c) Entropy (*H*)
   Entropy will indicate irregularity level of a pattern, and this is also will be used in research on handwritten document. The higher value of entropy, the variations and information contained in the writing pattern more and more irregularity.

$$H = -\sum_n p(f_n).Log_2 p(f_n)$$

d) Relative Smoothness (*R*) [14]
   Relative smoothness is a value that will indicate the relative degree of smoothness of shape and pattern of an image, and the researcher would also use it to determine how fine or rough the handwriting strokes of each document are. The higher the value of relative smoothness, the smoother the strokes pattern and the faster the writing movement.

$$R = 1 - \frac{1}{(1 + \sigma^2)}$$

   In this research, another feature added is the mean feature, which will indicate the exact feature value to represent the entire representation of handwriting patterns.

2.6. Normalization

Normalization is a step to create a range that has different value to the same scale and smaller as well. There are many ways to normalize value. One of the best method is Min-Max Normalization [16] :

$$D`(t) = \frac{D(t) - \min(D) \cdot (U - L) + L}{\max(D) - \min(D)}$$

where D is the natural data, U and L are the upper and lower normalization bound.

2.7. Similarity Measure

Similarity measures are techniques that used to measure the level of similarity between objects to one another, and usually implemented on pattern recognition. This similarity level is determined by distance. From the scientific and mathematical point of view, distance is defined as a quantitative degree of how far apart two objects are [15]. There are several similarity measures techniques, ie *Eulidean, City Block, Minkowski, Chebyshev, Sorensen, Gower, Soergel, Kulezynsky, Canberra*, andothers. In this study, the technique chosen is *Euclidean* because *Euclidean Distance* is the only metric that is the same in all direction. Its fits very nicely with the general qualities of our universe, which is also rotation invariant.

$$d_{Euc} = \sqrt{\sum_{i=1}^{d} |P_i - Q_i|^2}$$

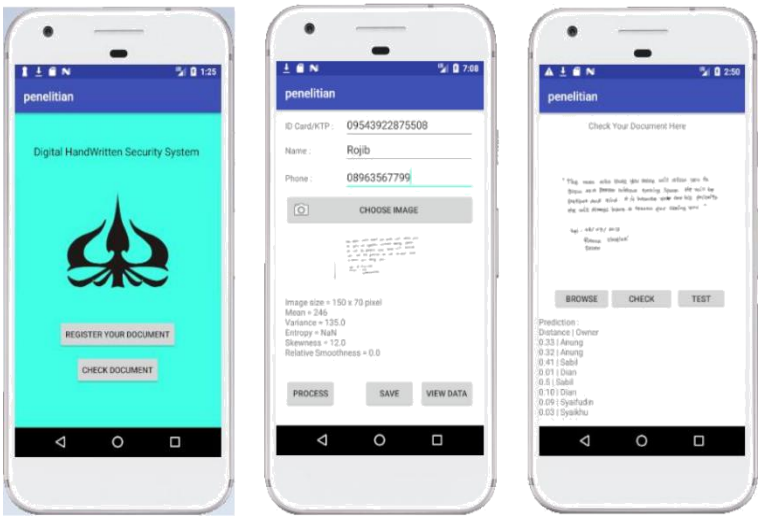
2.8. Research Methods

First of all, researcher collects the data in the form of handwritten document (printed) from several source, where the writing should be written on a white background paper without any lines or ornaments). All of documents which collected amounted to 30 from 10 different authors. From 30 data will be divided into 2 categories, namely 20 training data and 10 testing data. All of data is then will be processed into digital form through scanner, and saved as .jpg with the size of 200x150 pixels. After that, the data will be processed into pre-processing step, started from converting RGB color to greyscale, and then separate the writing area from the background through thresholding techniques. The given threshold value is 120, where this value is taken based on trial and error testing of the best results through the researcher's vision (subjective). Then, pixel's value of each document, will be counted by the color probability of histogram, ranging from grey level 0 to 255. After that, grey level will be quantized into 2^7 or 128 (0 to 127) where every 2 levels of gray will be represented by 1 level. This is to be done to reduce the computational load in finding the texture value at the feature extraction stage. Then from the result of quantization will be proceed to the texture feature extraction to get the value of *variance, skewness, entropy,* and *relative smoothness*. After that, *mean* value will also be used as an additional feature.

This process is applied to all handwriting data, both training data and testing data (digital handwriting query), which is then matched through the similarity measure technique after normalization into [0.0 – 1.0] interval. The smallest distance obtained from the results of this calculation will determine the similarity of documents between the authors one to another. The results from this stage then will be calculated the percentage of accuracy to determine how well the system developed by researchers.

3. Results and Discussion

The system which developed in this research is using Android Studio based on Java programming language. Figure 1 is the view of system. All training data is processed to take its features through the system, and stored in SQLite database which will be used as a reference to determine the owner of document against testing data. The difference in feature intervals will be automatically changed by the system into 0 to 1 value. If there is a feature whose value cannot be defined or NaN (as shown in Figure 1b), the default value of the feature is 0. Table 1 shows the result of measuring the distance between testing and training data using euclidean distance.



(a) (b) (c)

Figure 1. Mobile System (a) Main Menu; (b) Register Menu; (c) Checker Menu

Table 1. Similarity Measure of Handwritten Documents.

No	Distance Value	Target									
		B	B	D	D	H	H	I	I	J	J
1	A	0.73	0.75	3.41	3.59	3.32	3.2	4.52	2.22	3.21	3.18
2	A	1.28	1.19	3.28	3.05	3.39	3.34	4.59	2.33	3.25	3.16
3	B	0.33	0.35	2.44	2.33	2.92	2.71	3.09	1.38	2.51	2.4
4	B	0.37	0.32	2.46	2.4	2.97	2.6	3.07	1.41	2.42	2.37
5	C	4.12	4.08	2.02	2.1	2.41	2.31	1.03	3.1	2.1	2.07
6	C	4.05	4.11	2.08	2.09	2.43	2.32	1.2	1.2	2.12	1.98
7	D	1.75	1.8	0.51	0.01	0.8	0.10	1.18	1.3	0.41	0.35
8	D	1.46	1.76	0.98	0.88	1.12	1.6	2.6	0.4	1.41	1.23
9	E	3.45	3.3	1.12	1.17	1.14	1.62	0.72	2.4	1.31	1.3
10	E	0.86	0.78	1.15	1.01	1.80	0.98	2.01	0.31	1.4	1.35
11	F	0.68	0.77	3.02	3.6	1.89	3.51	4.4	2.33	3.24	3.07
12	F	1.19	1.20	1.02	1.14	1.4	1.3	2.4	0.14	1.03	1.04
13	G	3.45	3.42	1.05	1.71	1.42	1.6	0.69	2.4	1.3	1.34
14	G	1.03	1.06	1.23	1.08	1.4	1.02	2.08	0.33	1.44	1.27
15	H	2.07	2.11	0.41	0.31	0.5	0.21	1.3	1.04	0.15	0.23
16	H	4.27	4.22	2.23	2.51	2.3	2.47	1.5	3.12	2.13	2.01
17	I	1.42	1.45	3.1	3.7	3.1	3.62	4.7	2.43	3.3	3.2
18	I	2.31	2.37	1.4	1.11	1.8	1.2	0.09	2.4	1.51	1.34
19	J	1.02	1.05	1.12	1.31	1.5	1.22	2.3	0.03	1.13	0.14
20	J	1.52	1.66	1.09	1.82	1.01	1.79	2.81	0.54	1.41	0.09

Note *:A: Agung, B : Anung, C: Binti, D: Dian, E: Is, F: Najih, G: Ratna, H: Sabil, I: Syaifudin, J: Syaikhu*

Based on the test result in table 1 above, the error of handwritten document recognition still occur, for example in experiments number 7, 15, and 19. This can be due to the features extracted between

documents D with H, document J with H, document I with J having no distant value difference. There was a written similarities in terms of speed, direction of stroke, irregularities and variations. But in term of accuracy, this study is still better at recognition handwritten patterns when compared with previous studies. In the previous study [8], the accuracy achieved was only 55% (11 out of 20 data), whereas in this study, 6 out of 10 testing documents were identified correctly or achieved accuracy of 60%. This accuracy may increase if the data used as reference (in SQLite) more, because the value of features to be matched will be more varied to allow details of the different features of the posts will be more visible and the system can more easly recognize the pattern. So that, the system built by researchers can be utilized for the purposes of digital forensic in helping investigate the evidence in the form of authenticity of handwritten document ownnership with a fairly good level of accuracy.

4. Conclusion

Implementation of feature extraction based on texture with uniform quantization in the system development for digital forensics has a success rate that is good enough for 60%, to prevent the forgery of handwritten documents and investigate the origins of writing, because of the testing results of 10 documents written from 5 different authors, 6 documents were correctly identified by the authenticity of the owner. From the five texture features which have been used to investigate the authenticity of handwritten, *the relative smoothness* is the most difficult feature applied to differentiate document ownership because it has the same value results for all tested data. This can be due to *the relative smoothness* value associated with flexibility or speed, where handwriting is essentially created automatically by individual capabilities or without fabrications, so that all documents have maximum relative smoothness value of 1. Texture features that can play a good role in distinguishing characteristics of handwriting features are *skewness*, *entropy*, and *variance*. While the mean value may change depending on the number of words contained in the document and the size of writing.

Acknowledgments

This research was fully funded by Ministry of Technology Research DIKTI, Republic of Indonesia, as a Grant for lecturer, and also highly supported by Trisakti University and dedicated to author's beloved parents, especially mrs Sri Mulyani (deceased).

References

[1] Marquez, S. et al. 2008. A Simple and Effective Method of Color Image Quantization. *LectureNotes in Computer Science (LNCS)*, Vol.5197, pp. 749-757.

[2] Kumar, R. Five ways technology is improving public service. 2014. *The World Bank blogs.* [Online]. Available: http://blogs.worldbank.org/governance/five-ways-technology-improving-public-services.

[3] Ramadhani, Y. 2018. Lion Air Laporkan 9 Pilot dan 1 Karyawan Atas Pemalsuan Dokumen. Tito.id. Available: https://tirto.id/lion-air-laporkan-9-pilot-dan-1-karyawan-atas-pemalsuan-dokumen-cKWS.

[4] Santoso, A. 2015. Perjalanan Kasus Kematian Akseyna, Setahun Berlalu Masih Misteri. *Liputan6*. [Online]. Available:http://news.liputan6.com/read/2472824/perjalanan-kasus-kematian-akseyna-setahun-berlalu-masih-misteri.

[5] Plamondon, R. and Srihari, S.N. 2000. On-Line and Off-Line Handwriting Recognition : A Comprehensive Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.22, No.1, pp. 63-84.

[6] Gluhchev, G. 2004. Handwriting in Forensic Investigations. *International Journal "InformationTheories & Applications",* Vol.11, pp. 42-46.

[7] Pratiwi, D., Santoso, G.B., Saputri, F.H. 2017. The Application of Graphology and Enneagram Techniques in Determining Personality Type Based on Handwriting Features. *Journal ofComputer Sciences and Information*, Vol.10, No.1.

[8] Syaifudin and Pratiwi, D. 2016. Security Handwritten Documents by Using Inner Product. *Lecture Notes in Electrical Engineering* 365, Springer Science+Bussiness Media Singapore,DOI 10.1007/978-981-287-988-2_56.

[9] Chandy, D.A., Johnson, J.S., Selvan, S.E. 2014. Texture Feature Extraction using Gray Level Statistical Matrix for Content-based Mammogram Retrieval. *Springer Sciences+BusinessMedia New York : Multimedia Tools and Application*, Vol. 72, Issue 2, pp. 2011-2024.

[10] Purnomo, A., Puspitodjati, S. 2010. Aplikasi Pemrograman C# untuk Analisis Tekstur Kayu Parquet dengan Menggunakan Metode Grey Level Co-occurrence Matrix (GLCM), [Online]. http://www.gunadarma.ac.id/library/articles/ graduate/industrial-technology/2009/Artikel_50405312.pdf.

[11] Najeeb, H. 2017. Histogram of Color Image. Mathworks : Matlab Answers. [Online]. Available:https://www.mathworks.com/matlabcentral/answers/324646-histogram-of-color-image.

[12] Shodhganga. 2017. Quantization Techniques. [Online]. Available: http://shodhganga.inflibnet.ac.in/bitstream/10603/25341/8/08_chapter%203.pdf.

[13] Wong, J.S.J., Zrimec, T. 2006. Classification of Lung Disease Pattern Using Seeded Region Growing, *Lecture Notes in Artificial Intelligence* : *Proceedings 19thAustralian Joint Conferenceon Artificial Intelligence*, pp.233-242, Berlin, Germany : Springer-Verlag.

[14] Singh, B.Kr., Mazumdar, B. 2010. Content Retrieval from X-ray Images Using Color & Texture Features, *International Journal of Electronics Engineering*, 2(1), 25-28.

[15] Cha, S.H. 2007. Comprehensive Survey on Distance/Similariy Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in AppliedSciences*, Issue 4, Vol. 1, pp. 300-307.

[16] Pratiwi, D., Santika, D.D., and Pardamean, B. 2011. An Application of Backpropagation Artificial Neural Network Method for Measuring The Severity of Osteoarthritis. *InternationalJournal of Engineering & Technology IJET-IJENS*, Vol.11, No:03, ISSN : 117303-8585.

[17] Davin, R.P., Pratiwi, D., Syaifudin, Trubus. R., Rizky, D.L.P., 2017. Implementation of Inner Product to Analyze Digital Handwriting based on Texture Traits. Proceedings of The 2017 International Conference on Computer Science and Artificial Intelligence (CSAI), pp. 114-118, ACM, ISBN: 978-1-4503-5392-2.