Volume 10 Issue 1 February 2017 ISSN 2088-7051

Jurnal Ilmu Komputer dan Informasi

Journal of Computer Science and Information

Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information). 10/1 (2017), 1-10
DOI: http://dx.doi.org/10.21609/jiki.v10i1.440

DE-IDENTIFICATION TECHNIQUE FOR IOT WIRELESS SENSOR NETWORK PRIVACY PROTECTION

Yennun Huang[1], Szu-Chuang Li[1], Bo-Chen Tai[1], Chieh-Ming Chang[1], Dmitrii I. Kaplun[2], and Denis N. Butusov[2]
1CITI, Academia Sinica, 128 Academia Road, Section 2, Nankang,, Taipei City, 115, Taiwan
2Saint Petersburg Electrotechnical University "LETI", ul. Professora Popova 5, 197376 Saint Petersburg, Russian Federation

E-mail: yennunhuang@citi.sinica.edu.tw

Abstract

As the IoT ecosystem becoming more and more mature, hardware and software vendors are trying create new value by connecting all kinds of devices together via IoT. IoT devices are usually equipped with sensors to collect data, and the data collected are transmitted over the air via different kinds of wireless connection. To extract the value of the data collected, the data owner may choose to seek for third-party help on data analysis, or even of the data to the public for more insight. In this scenario it is important to protect the released data from privacy leakage. Here we propose that differential privacy, as a de identification technique, can be a useful approach to add privacy protection to the data released, as well as to prevent the collected from intercepted and decoded during over-the-air transmission. A way to increase the accuracy of the count queries performed on the edge cases in a synthetic database is also presented in this research.

Keywords: *differential privacy, internet of things, sensor network*

Abstrak

Sebagai ekosistem IOT menjadi lebih dan lebih dewasa, vendor hardware dan software berusaha menciptakan nilai baru dengan menghubungkan semua jenis perangkat bersama melalui IOT. Perangkat IOT biasanya dilengkapi dengan sensor untuk mengumpulkan data, dan data yang dikumpulkan ditransmisikan melalui udara melalui berbagai jenis koneksi nirkabel. Untuk mengekstrak nilai data yang dikumpulkan, pemilik data dapat memilih untuk meminta bantuan dari pihak ketiga dalam analisis data, atau bahkan data kepada publik untuk wawasan yang lebih dalam. Dalam skenario ini penting untuk melindungi data yang dirilis dari kebocoran privasi. Di sini kami mengusulkan bahwa privasi diferensial, sebagai teknik identifikasi de, dapat menjadi pendekatan yang berguna untuk menambah perlindungan privasi data yang dirilis, serta untuk mencegah diambil dan diterjemahkan selama transmisi over-the-air. Sebuah cara untuk meningkatkan akurasi query count dilakukan pada kasus tepi dalam database sintetis juga disajikan dalam penelitian ini.

Kata Kunci: *privasi differensial, internet of things, jaringan sensor*

1. Introduction

As the IoT ecosystem becomes more and more mature in recent years, hardware and software vendors are trying to create new value by connecting all kinds of devices together via IoT. One of the primary functions of an IoT device is to collect and transfer data using equipped sensors. Rapid and enormous data collection has been happening in the past years on PC and mobile phones. According to IBM during the last few years 2.5 billion gigabytes of high-velocity data, such as social media posts, information gathered from sensors and medical devices, videos and transaction records, are created in a variety of forms every day, and the rise of the IoT devices in numbers will cause the quantity of data collected each day to skyrocket. Gartner[1] predicts that in 2016 there're already 6.4 billion IoT devices, and the number will be tripled in 2020, making it 20.8 billion.

IoT devices possess very different qualities than a PC or mobile phone. First, they're often deployed in large number: in the future we might have several wearable devices per person, as well as multiple IoT-enabled electronics in a household. Second, a lot of IoT devices will be deployed outdoors, and those devices will be vulnerable to physical hacking, and the transmitted data

[1] http://www.gartner.com/newsroom/id/3165317, retrieved on Jan. 26th, 2017

might be intercepted, causing every kind of possibility of privacy leakage. Last, IoT devices usually possess very limited storage and computing resource, making it difficult to use advanced encryption schemes to protect data storage and transmission.

De-identification techniques can be an effective alternative to deal with privacy preserving data transmission and analysis in for IoT. Existing de-identification methods such as K-anonymity and its derivatives, and differential privacy-compliant mechanisms consumes relatively small resource while providing data privacy. In this paper we'll first describe a field test we've done at a local theme park, utilizing a custom-built Bluetooth network and proximity tags to collect spatio-temporal data of the visitors, and we'll discuss how we can remove the sensitive attributes from the data while preserving its statistical utility, so that we can release the data to a third-party for further analysis without revealing privacy information. After coping with the problem of privacy preserved IoT data release, we'll take a brief look at a current option to propose how we can use de-identification techniques to protect data transmission.

2. Methods

Collection of Spatio-Temporal Data from a Custom Bluetooth Sensor Network

Ways to collect spatio-temporal data

With the emergence of wearable devices and sensor technology, there have been plenty attempts to collect and analyze spatio-temporal data. The most common used technologies to retrieve positional information are still GPS and Wifi [1-4]. Recently Bluetooth has become a viable choice to provide positioning service, especially in an indoor scenario. Typically the Bluetooth beacons are configured to send out simple ID information. When installed its physical location will be recorded to a database on a central server or a small local database that's attached to an mobile APP. Whenever a mobile device gets near the Bluetooth beacon and receives the ID information broadcasted by the Bluetooth beacon, it will match the ID information against the data stored in a server or local database on the mobile APP and react accordingly. Recently researchers have been trying to get more precise positional information out of Bluetooth beacons by taking Bluetooth signal strength into account and/or combine information from multiple beacons [5]. Another approach, though, is to get positional information via "crowd sensing". Jamil et. Al. [3] had an attempt to combine mobile phones with Bluetooth proximity tags to rebuild the traces of visitors.

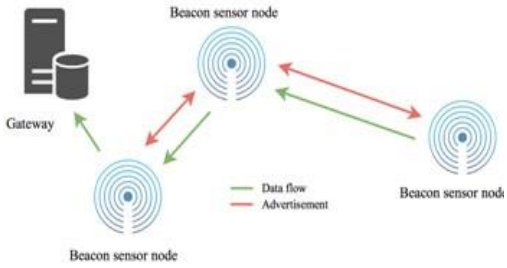

Figure 1. Relaying data through a series of Bluetooth beacons

A custom solution to collect data in a wireless Internet-less environment

As mentioned, the most common usage of Bluetooth beacons is to use them as broadcasting stations. But since Bluetooth specification actually allows a beacon to work in scan and broadcasting mode, it is possible to relay limited information between Bluetooth beacons, while scanning for Bluetooth proximity tags nearby back and forth. This way the beacons can collect the ID information sent by Bluetooth proximity tags and relay them through a series of beacons. At the end of the beacon chain we can setup a PC as a Bluetooth network-to-Internet gateway to relay collected information to a remote cloud server for data storage and analysis, as in Figure 1.

The "Bluetooth Gateway" is a PC bor Server Jonnected to the Internet with a Bluetooth Interface, and each Bluetooth beacon should be placed within the broadcasting range of the next and previous Bluetooth beacons. The Bluetooth beacons are programmed to carry custom payload, enabling them to do two-way communication in the following fashion:

Upstream communication

Figure 2. Custom made Bluetooth beacon

Figure 3. Bluetooth bracelet from Xiaomi Technology

In the connected Bluetooth network it's possible for a beacon to send data to the gateway PC and even to the Internet when needed. The use of custom payload enables arbitrary data to be relaed all the way to the Bluetooth gateway for further processing. The data are Bluetooth mostly device IDs, but it's possible to send control codes too when needed.

Downstream communication
Information regarding all Bluetooth beacons was aggregated at the Bluetooth gateway, making it possible to send control codes downstream to a particular beacon. For example, the gateway PC can send a command to change the scan interval to a particular Bluetooth beacon to change its behavior. It is also possible to send application related information such as a short message or a URL pointer to all the Bluetooth or mobile phones near a particular beacon.

Power efficiency of Bluetooth beacons
To enable easy deployment and allow Bluetooth beacons to run on batteries for extended period of time, scanning interval of the custom-built Bluetooth beacons were configured to rest for 15 seconds after 5 seconds of scanning and Broadcasting. Coupled with the clocked switch which only turns on Bluetooth beacons during the work hours, a Bluetooth beacon can run for 72 days without batteries changed with 2x 3000mAh batteries installed. Please note that the two-way network is not suitable for real time communication. The beacons are configured to scan peridically. Buffering and confirmation mechanism has been designed very carefully to ensure the reliability of data transmission, and the time required for the packets to travel to the destination is long and may vary. In our experiments the transfer time can be as long as 1 minute when the beacon chain is long.

In past researches Internet connections are required to send the collected positional infomation to a remote server. For example, if we want to collect spatio-temporal information of visitors in a theme park for optimizing the visiting experience, the theme park will have to make visitors install mobile APPs and configure properly and provide wireless Internet access for them if they do not have it themselves. It could be expensive and uh-realistic for a theme park do create those infrastructures or to expect every visitor to have an Internet connection subscription. By using a custom Bluetooth beacon network described here we'll give proximity tags to visitors (Bluetooth bracelets or stickers), and setup beacons along the popular paths. As in Figure 1 the beacons can then repay detected ID information all the way to the Internet. This is made possible by the utilization of the CC2541 SoC's programmable chip from TI, which is used to create a custom protocol to relay information through a series of Bluetooth beacons. A local theme park called "Little Ding Dong science theme park" agreed to let the research team setup more than 50 beacons around the theme park. The devices we used to setup this experiment includes:

Custom-made Bluetooth beacons
Inside the beacon container Three are four somponents: (1) A programmable SoC from Texas Instruments with 8051 ALU and integrated Bluetooth functions, (2) An antena, (3) A waterproof case for reliable operation indoor/outdoor, (4) A pair of batteries that allows the beacon to work for several weeks when fully charged

Utilizing the SoC's programmability, we were able to implement some of the key features of the system: (1) Change signal scanning / transmitting interval to increase power efficiency: to increase power efficiency, the interval of scanning time of Bluetooth beacons can be tuned. Extensive experiments were performed for us to learn about the optimal parameters that balance energy and data transferring efficiency. Based on the experiment results we configure the beacons to scan or broadcast for 5 seconds and sleep for 15 seconds. The beacons will also be configured to

Figure 4. Physical placement of Bluetooth beacons

run for 8 hours a day. A beacon equipped with 2 3000mAh batteries can run for 72 days nonstop using this setting. This enables fast deployment and easy maintenance for the Bluetooth beacon Networks, (2) Enabling two-way communication: the beacons are programmed to relay "upstream" and "downstream" data. For instance, identification information of Bluetooth bracelets collected by the beacons will be sent "upstream" to the gateway PC (described later), and will be relay to cloud server thereafter. The gateway PC can send commands "downstream" to a particular beacon through a predefined path. Please note that, to accommodate the energy efficiency arrangements, the two-way communication will not be real-time and will inevitably introduce latency in data transmission.

Sending out identification info: Bluetooth bracelet / proximity tag to send
Bluetooth bracelets from Xiaomi technology are affordable and serve the purpose well. Around 50 units were given to the visitors when they enter the theme park, and the bracelets were returned when they leave the theme park in exchange for coupons that offer a discount when they visit the theme park next time.

Gateway PC with Internet connection and Bluetooth connectivity
There'll be a PC with Bluetooth connectivity and Internet connection at the end of the Bluetooth beacon chain. It will act as a gateway to enable the Bluetooth network to exchange information with the Internet.

Remote cloud server
To analyze the data collected effectively, a remote cloud server with adequate processing power and storage will serve as a storage and data analysis platform. The web server will provide HTTP REST-based API to process data storage requests and attraction recommendation information to users. Route prediction algorithm will also be implemented on the web server.

Setting up the beacons
More than 50 beacons were setup in the theme park to collect spatio-temporal data of the visitors. Since we want to deploy as few beacons as possible, the beacons were tested and it is confirmed that their range of transmission is 15-20 meters. A person will be detected by nearby beacons, and since we're not utilizing signal strength data at this time, placing beacons farther apart will help to reduce redundant detection of visitors from the same beacons. Also since the beacons are placed mostly outdoor, it is important that there're clear path between beacons for Bluetooth signal to be transmitted reliably (no walls present to reflect the signals). In Figure 4, it is shown that the beacons often have to be placed higher above the ground to ensure that there're clear paths between the beacons.

It is worth noting that there're Bluetooth beacons on the market that can run for years on battery, but this is not the case in our study. The custom-built beacons do not just sending out ID information, instead they keeps switching Petween scanning and broadcasting mode, and have to buffer data before relaying them to the other beacons. By carefully tuning the switching interval they still manage to last 8 to 9 weeks before the batteries have to be replaced.

3. Results and Analysis

Following the BLE specification a Bluetooth packet can only be 32-byte in length, and we have to design the transmission data format around this restriction. To ease power consumption, the beacons will detect at most 28 visitors' proximity tag at each round of scanning, and the data will be squeezed into a single packet and transmitted to the next beacon in line. As illustrated in Fig 1. the data will be transmitted along the chain of Bluetooth beacons, all the way to the gateway and eventually to a remote server in the cloud. A MySQL server is installed on the cloud server to store the collected data. We setup a data schema to store such data as n Table 1.

TABLE 1
COLLECTED DATA ATTRIBUTES

Data	Note
Beacon ID	Which beacon detected this bracelet
Bracelet ID	Which bracelet was detected
Timestamp	The time that this data is written to database

From this data we can perform some analysis on the users' visiting behavior. For example, we can reconstruct the route of a particular visitor using the data (Figure 5.), or draw a histogram to show which attractions in the theme park is most visited.

More analysis can be performed on the raw data to gain more insight regarding how the visitors visit the theme park. However, sometimes the data collector doesn't necessarily have the ability do make the most out of the data, hence the need to share those data with a third-party or even release it to the public for further analysis. In this case adequate privacy must be ensured, or the release of such data can violate privacy regulations. We'll discuss how we can protect raw data before release in the following paragraph.

Figure 5. Reconstruction of route for a particular visitor

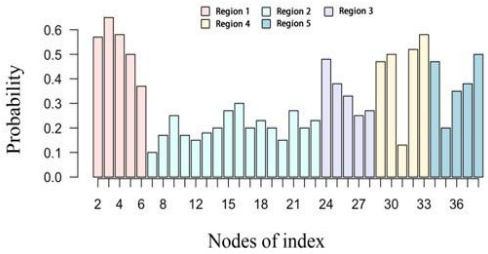

Figure 6. Reconstruction of route for a particular visitor

Ensuring Privacy When Releasing Data to a Third Party or the General Public

It is expected the number of IoT devices will grow rapidly in the coming years. IoT devices not only possess processing power and storage capability, but are also equipped with sensors and actuators. Massive amount of data will be collected by the sensors, and then transferred and stored. Eventually they have to be analyzed to generated value. To ensure privacy of released data, there have been some developed methodology trying to achieve this goal, and those techniques are often labeled as "data de-identification". The more mentioned ones includes K-anonymity [6] and its derivatives[7,8], differential privacy [16], and other attempts from statistical discipline [9]. Due to its deployment by major companies such as Apple[2] and Google, here we'll discuses differential privacy as a potential solution to ensure privacy on IoT data release. It is worth noting that all kinds of data de-indentification techniques so far have to face the problem of privacy-utility tradeoff. The more a data set is processed extensively to hide all the sensitive information, the more decrease in data utility can be expected.

Differential Privacy

Differential Privacy is first proposed by [16], with a provable definition of privacy. The idea is that when one perform a query on a data set (e.g. count number of the entries that fits a set of criterions), the result will be randomized so that the result would not be significantly different whether a particular record presents in the data set or not. The most widely known definition is as below:

Definition 1 [16]. *A randomized κ function gives ε-differential privacy if for all data sets D1 and D2 differing on a t most one element, and all $S \subseteq Range(\kappa)$,*

$$\mathbf{Pr}[\boldsymbol{k}(\boldsymbol{D_1}) \in \boldsymbol{S}] \leq \mathbf{exp}(\boldsymbol{\epsilon})\, \boldsymbol{x}\ \mathbf{Pr}[\boldsymbol{k}(\boldsymbol{D_1}) \in \boldsymbol{S}] \ldots \quad (1)$$

The probability is taken is over the coin tosses of k.

The single record that is different in D1 and D2, can cause privacy leak if the value is vastly different from the other values in the data set. For example, if there's a millionaire in the area, by

[2] Andy Greenberg, Apple's 'Differential Privacy' is about collecting your data, butyou're *your* data, https://www.wired.com/2016/06/apples-differential-privacy-collecting-data/, retrived on Feb. 6th, 2017.

looking at the average income of a data set it could be easy to tell if this person's income is present in the data set or not. So when we decide how much "noise" we want to add to the query result we must take this into account.

Definition 2 [16]. *For f: $D \rightarrow R^k$, the sensitivity of is*

$$\Delta \boldsymbol{f} = \max_{\boldsymbol{D1},\boldsymbol{D2}} \|\boldsymbol{f}(\boldsymbol{D_1}) - \boldsymbol{f}(\boldsymbol{D_2})\|_1 \ldots \quad (2)$$

By withdrawing each record from the data set and calculate the query result on the remaining data entries, we can identify the maximum possible difference the absence of an data entry with extreme value can produce, and take it into account when we decide how much "noise" we should add to a query result to ensure differential privacy. There are some "randomized functions" that fits this definition, but the most commonly used one is Laplace mechanism.

Theorem 1. *For f: $D \rightarrow R^k$, the mechanism K_f that adds independently generated noise with distribution Lap(Δf /ϵ) to each of the j output terms enjoys ϵ–differential privacy [16].*

According to theorem 1, on query function f the privacy mechanism K responds with equation(3).

$$\boldsymbol{f}(\boldsymbol{X}) + (\boldsymbol{Lap}(\Delta \boldsymbol{f}/\boldsymbol{\epsilon}))^{\boldsymbol{k}} \quad (3)$$

will make the query results returned satisfy ϵ–differential privacy.

By adding "noise" to query results, we hope to prevent an advisory from identifying a person by conducting similar queries on a data set. However it is worth noting that by making the same query over and over again the advisory may still learn the real value of a query overtime, so differential privacy it is still needed to limit the query number of a particular person. This is often referred to as "privacy budget." Also one can always choose a larger ϵ to make the noise smaller, but this will result in higher disclosure risk.

Differential Privacy-Compliant Synthetic Database

As we are trying to deal with the problem of data release to a third party, the query-based version of differential privacy does not really suit our needs. [16] also addressed the issue of "non-interactive differential privacy" and proposed that a synthetic dataset can be generated from the results of a series of counting queries performed on the source data. Essentially, one can first identify all the possible value combinations of the attributes in a data set, and count the occurrence of each instance. According to Definition 2. the sensitivity of count queries is a fixed "1", as when we remove or add a data entry to a data set, the result of counting query will be at most "1". This makes the calculation of sensitivity extremely simple. There are several ways suggested by [16] to generate synthetic data set from counting query results, and below we will describe two of the three approaches she recommended.

The first approach is to simply add Laplace noise to each of these counting results, and rebuild a data set from those counting information. Since

TABLE 2
THE ORIGINAL DATA SET

Age	Height	Weight	Income	TRV	HTN	DGF
64	159	66	39	11	0	0
53	178	78	39	13	0	0
53	168	61	35	9	0	0
57	172	78	50	12	0	1
64	173	53	35	8	0	0

TABLE 3
THE SYNTHETIC DATA SET

Age	Height	Weight	Income	TRV	HTN	DGF
66.5	165.5	71.5	27.83	4.5	0	1
47.5	171.5	77.5	78.44	36.5	0	0
55.5	168.5	79.5	54.34	51.5	0	0
54.5	142.5	87.5	90.49	17.5	1	1
61.5	169.5	96.5	91.7	34.5	0	1

the synthetic data set is built from a series of counting query results that is protected by differential privacy, the data set should preserve privacy well. However, the number of count queries that needs to be done using approach can be excessive large, thus if the source data set is large with multiple attributes and value variation, the calculation time needed will be excessively large. Also although the noise added to each cell of this "contingency table" is relatively small, any query for a marginal (aggregate counting queries that fits certain conditions) can be too large for the result to be useful.

The second approach proposed by [16] is to produce some subset of the "contingency table", which are called "marginal tables", and to connect them together via probabilistic inference mechanism. Some of the attempts of this approach are PrivBayes [10] and DPTable [11], and in this research we use the latter and improve it with ways to improve accuracy without sacrificing privacy, which we'll describe later. Here we'll first describe a the steps involved in DPTable to generate a synthetic data set that can preserve most of the statistical properties of the original data set [12]: (1) Calculating the pair wise mutual information value between attributes. When mutual information value exceeds a certain preset threshold the relationship between the attributes will be preserved in the following process. Noise will be added to the mutual information calculated. (2) Based on step 1. Dependency graph will be constructed. The graph will also be "triangulated" for further processing. (3) The dependency graph will be converted to a junction tree, upon which marginal tables will be built. (4) Noise will be added to the marginal in the marginal tables. (5) The marginal tables as a whole will act as a joint distribution from which new dataset can be synthesized. (6) The data user will then be able to sample arbitrary number of data rows from the joint distribution.

To test DPTable, we made an artificial data set with columns age, height, weight, income, travel, high blood pressure (binary flag) and diabetes (binary flag) attributes. The data set has 100,000 rows. For reference the first 5 rows of the original data set is as Table 2, and the first 5 rows of the synthetic data set is as Table 3.

Please note that Table 3 was not "converted" from Table 2. As described in the step-by-step of DPTable, the DPTable mechanism uses the information in Table 2. To build a joint distribution, and then samples data from the joint distribution to build Table 3. To compare the statistical properties of the original data set and the synthetic data set, we calculate the average and standard deviation of each attributes in the table for a rough comparison. Please note that the attribute "HTN" and "DGF" are binary attributes, so in the "average" column we show the counts of positive ("1") value in those attributes.

In Table 4 we can see that the difference between the average value of INCOME and TRV is larger at around 8% and 31% respectively. For other attributes the difference in average value is quite small. For the binary attribute counts, the synthetic data produces 26% error for HTN and 4% error for DGF respectively. Overall the average values of different attributes are preserved quite well in the synthetic data set. For standard deviation the error for most attributes are significantly higher. Please note, though, this synthetic data set is generated using a small ϵ parameter at 0.01, which means that privacy is very well-protected. If one wishes to favor precision over privacy protection, he or she can always select a larger ϵ.

Use "K-aggregation" to improve the privacy-utility tradeoff in differential privacy compliant synthetic data

Besides tuning the ϵ parameter, researchers are actually trying to find ways to improve the techniques to improve privacy without sacrificing utility or vice versa. For example [13] states that by pre-

TABLE 4
A COMPARISON BETWEEN ORIGINAL AND SYNTHETIC DATA

	Average		Standard Deviation	
	Original	Synthetic	Original	Synthetic
Age	53.32771	52.99481	7.804086	7.670179
Height	168.8197	165.9513	7.972777	13.30099
Weight	77.05943	77.77396	7.718009	10.98742
Income	71.91315	78.20495	25.18623	31.03803
TRV	26.43248	34.69463	11.68337	16.89114
HTN(+)	22187		28072	
DGF(+)	28536		29900	

VALUE	1	2	3	4	5	6	7	8	9	10	11	12
COUNT	1	2	2	4	5	8	10	9	7	3	1	1

↓ ↓ ↓ ↓

VALUE	1~3	4	5	6	7	8	9	10~12
COUNT	5	4	5	8	10	9	7	5

Figure 7. Procudure of K-aggregtion when k = 4

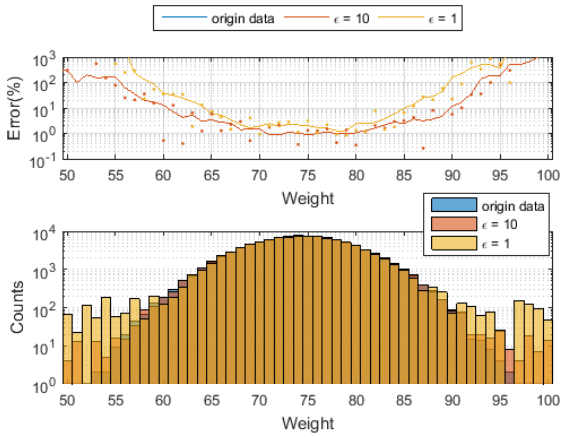

Figure 8. Error % is larger at the edge of a normally distributed dataset due to fewer data counts.

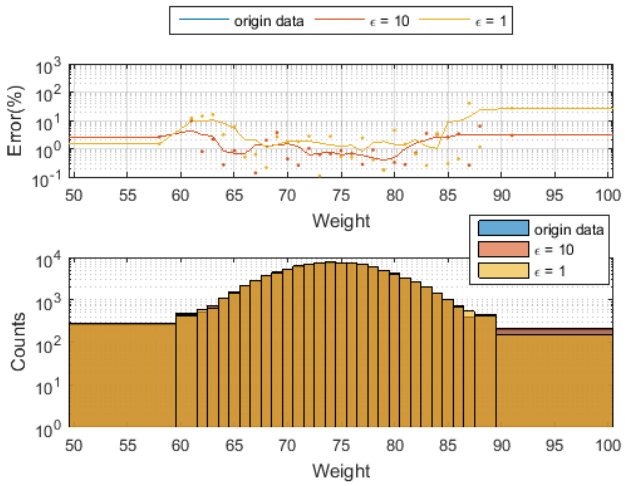

Figure 9. Error percentage variation for the attribute "Weight" with K-aggregation, k=200.

process the dataset using k-anonymity, the amount of noise addition can be reduced to achieve the same privacy in differential privacy, improving accuracy without sacrificing privacy. We examined the procedure and results of DPTable carefully, and here we'll discuss about the ways to improve data utility -- K-aggregation [12].

For data that is normally distributed, there are always fewer counts in extreme cases. For example, people that are extremely tall or short are tend to be small in number and people with more average height would be large in number in a normal distributed data set. As specified previously to apply differential privacy to a tabular data set we first convert to a series of "marginal tables", and then start to add fixed amount of noise to each of the counting query results, and this make it obvious that, proportionally, the marginal (count queries) with fewer count will be influenced by the noise added much more than the marginal with larger counts. Since the marginal at the edge of the data set contains so much noise it

becomes much less precise and, with low utility.

To cope with this problem the research team came up with a method to preprocess data called "K-aggregation". The steps of this method are as the following: Step 1: Two parameters will have to be set in advance. First a threshold will need to be chosen to examine the maximum acceptable error percentage between original dataset and synthetic dataset. Parameter k can be calculated from the maximum acceptable error percentage as stated in previous paragraph. Step 2: After the parameter has been chosen, the original dataset will be put through the DPTable process, from which the synthetic dataset will be generated. Step 3: Synthetic dataset will be compared to the original. If the maximum error across all possible attribute values between the counts in original and synthetic dataset is larger than the error threshold defined in step 1, we will proceed to step 3. Otherwise the synthetic dataset is accepted as usable. Step 4: Since the error is larger than the threshold, we assume that the data value count at the "edge" of the dataset needs to be aggregated to increase due to normal distribution. We will scan the database from the largest and smallest data value and aggregating the counts until the accumulate count exceeds k. In the original table those data value will be replaced with a new value calculated from the weighted mean from the data value. Step 5: If there are multiple attributes presents, step 1~4 can be iterated through all the attributes.

Please take Figure 7 as an example. The algorithm start to scan data from the two sides of the data set, and if the count of a certain value is below the threshold set, it will be combined with the count of the next value. After the threshold was reached the older values will be combined as a weighted new value. After K-aggregation the extremely low counts were combined and more precise counts are possible. To get an idea about the effect of K-aggregation, we also use the artificial data set as an example. In Figure 8, the top chart gives us an idea about the higher error % that the edge cases produce, and it is clear that the cases at the center of the chart produce much lower error %. The chart at the bottom represent shows the data entry count for each weight value.

We process the attribute "weight" in this data set with K-aggregation and have the threshold set to k = 200. In Figure 9. We can see that when the data attribute is pre-processed with K-aggregation, the error % of the counts toward the edge of the data set remains at a much lower level. And in the chart at the bottom we can see that at the edge of the chart the counts are aggregated and given a new value from weighted average of the original values.

To sum it up, K-aggregation can be used to reduce the error % at the edge of a DPTable processed data set, and this also applies to tabular-formatted IoT data sets.

Differential privacy as an option to transfer IoT data securely

Besides releasing sensitive data with privacy protection, differential privacy can also be used to transfer data securely. IoT devices collect information from all kinds of information and send them through Wifi information to remote servers, so it is always possible that someone intercepts those information. If the purpose of data transmission is for further aggregated analysis, differential privacy can come in handy.

Google RAPPOR [14] use differential privacy as a provable mechanism to protect the privacy of transmitted data. When a value is to be transmitted by RAPPOR, its true value will first be converted to binary format, and then passed through a bloom filter. After that the value will then be randomized but "memorized", so that when the value is sent again in the future, this particular randomized value will always represent the same value. And lastly before the values were sent to a remote server the value is randomized again. The remote server will aggregate all those data received and perform statistical estimation regarding how many times a particular string is received. Following this process, one can send carefully randomized information to a remote server for statistical analysis without worrying someone intercepts the data sent. As there are no encryption or decryption involved, there is no risk of leaking a key to an advisory. There're also following up works on RAPPOR to eliminate the need of having to build a dictionary first before data transmission and decoding [15].

4. Conclusion

During the past 10 years research of data anonymity/de-identification has been progress steadily. K-anonymity and differential privacy have been examined extensively to gauge their usefulness in a real world scenario, and the latter has started to be used in some main stream consumer products. In this research we introduced how de-identification techniques can be used for privacy preserve data release and data transmission in an IoT setting. Those techniques can also be used for non IoT purposes, but de-identification techniques, due to its lower requirement for processing power than some of the more sophisticated encryption/decryption schemes, are especially suitable for IoT applications.

Acknowledgement

This research is supported by the Ministry of Science and Technology, Taiwan, R.O.C. under Grant no. MOST 103-2221-E-001-028-MY3 and MOST 106-2923-E-001-001.

References

[1] J. Zhu, K. Zeng, K. Kim, P. Mohapatra, "Improving crowd-sourced Wi-Fi localization systems using Bluetooth beacons." In Proceedings of Sensor, Mesh and Ad Hoc Communications and Networks (SECON), 2012 9th Annual IEEE Communications Society Conference on, 290-298. 2012.

[2] S. S. Chawathe, "Low-latency indoor localization using bluetooth beacons." In Preceedings of 2009 12th International IEEE Conference on Intelligent Transportation Systems, 1-7. 2009.

[3] S. Jamil, A. Basalamah, A. Lbath, "Crowd-sensing traces using bluetooth low energy (BLE) proximity tags." In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, 71-74. 2014.

[4] H. Koyuncu, S. H. Yang, "A Survey of Indoor Positioning and Object Locating Systems." International Journal of Computer Science and Network Security (IJCSNS) 10, 5: 121-128. 2010.

[5] S. S. Chawathe, "Beacon Placement for Indoor Localization using Bluetooth." In Preceedings of 2008 11th International IEEE Conference on Intelligent Transportation Systems, 980-985. 2008.

[6] L. Sweeney, "k-anonymity: a model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5): 557-570, 2002.

[7] Li, N., Li, T., "t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity," Proceedings of the 23nd International Conference on Data Engineering, 2007.

[8] Machanavajjhala, A., Kifer, D., Gehrke J., Venkitasubramania, M., "I-Diversity: Privacy Beyond k-Anonymity," Proceedings of the 22nd International Conference on Data Engineering.(ICDE) , pp.24-35, 2006.

[9] Rubin D. B. Discussion: Statistical Disclosure Limitation, Journal of Official Statistics, Vol. 9, No. 2, pp 461-468, 1993.

[10] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava and X. Xiao, "PrivBayes: private data release via bayesian networks," SIGMOD '14 Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, pp. 1423-1434, 2014.

[11] R. Chen, Q. Xiao, Y. Zhang and J. Xu, "Differentially Private High-Dimensional Data Publication via Sampling-Based Inference," Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,pp. 129-138, 2015.

[12] B. C. Tai, S. C. Li, Y. Huang, "K-aggregation: Improving Accuracy for Differential Privacy Synthetic Dataset by Utilizing K-anonymity Algorithm", to be presented at AINA 2017.

[13] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez and S. Martínez, "Enhancing data utility in differential privacy via micro-aggregation-based k-anonymity," The VLDB Journal, vol. 23, issue 5, pp. 771-794, October 2014.

[14] Ú. Erlingsson, V. Pihur, A. Korolova, "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response." CCS'14, November 3–7, 2014, Scottsdale, Arizona, USA.

[15] G. Fanti, V. Pihur, Ú. Erlingsson, "Building a RAPPOR with the Unknown: Privacy-Preserving Learning of Associations and Data Dictionaries", Proceedings on Privacy Enhancing Technologies, (3):1–21 2016.

[16] Dwork C., "Differential Privacy: A Survey of Results," in Theory and Applications of Models of Computation Volume 4978 of the series Lecture Notes in Computer Science, pp. 1-19, April 2008.G. Smith, "Paper Title" (to be published).

[17] Gartner Website, Gartner Says 6.4 Billion Connected "Things" Will Be in Use in 2016, Up 30 Percent From 2015, http://www.-gartner.com/newsroom/id/3165317, retrieved on Jan. 26th, 2017.

[18] Andy Greenberg, Apple's 'Differential Privacy' is about collecting your data, but you're your data, https://www.wired.com/-2016/06/apples-differential-privacy-collecting-data/, retrieved on Feb. 6th, 2017.

Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information). 10/1 (2017), 11-18
DOI: http://dx.doi.org/10.21609/jiki.v10i1.372

THE APPLICATION OF GRAPHOLOGY AND ENNEAGRAM TECHNIQUES IN DETERMINING PERSONALITY TYPE BASED ON HANDWRITING FEATURES

Dian Pratiwi, Gatot Budi Santoso, Fiqih Hana Saputri

Trisakti University, Jl. Kyai Tapa No.1 Jakarta Barat, Indonesia

E-mail: dian.pratiwi@trisakti.ac.id

Abstract

This research was conducted with the aim of developing previous studies that have successfully applied the science of graphology to analyze digital handwriting and characteristics of his personality through shape based feature extraction, which in the present study will be applied one method of psychological tests commonly used by psychologists to recognize human's personality that is Enneagram. The Enneagram method in principle will classify the personality traits of a person into nine types through a series of questions, which then calculated the amount of the overall weight of the answer. Thickness is what will provide direction personality type, which will then be matched with the personality type of the result of the graphology analysis of the handwriting. Personality type of handwritten analysis results is processed based on the personality traits that are the result of the identification of a combination of four dominant form of handwriting through the software output of previous studies, that Slant (tilt writing), Size (font size), Baseline, and Breaks (respite each word). From the results of this research can be found there is a correlation between personality analysis based on the psychology science to the graphology science, which results matching personality types by 81.6% of 49 respondents data who successfully tested.

Keywords: *Graphology, Enneagram, Psychology, Personality, Handwritten.*

Abstrak

Penelitian ini dilakukan dengan tujuan untuk mengembangkan penelitian sebelumnya yang telah berhasil menerapkan ilmu grafologi untuk menganalisis tulisan tangan digital dan karakteristik kepribadiannya melalui ekstraksi fitur berdasarkan bentuk, yang dalam penelitian ini akan diterapkan salah satu metode tes psikologi yang umum digunakan oleh psikolog untuk mengenali kepribadian manusia yang Enneagram. Enneagram Metode pada prinsipnya akan mengklasifikasikan sifat-sifat kepribadian seseorang menjadi sembilan jenis melalui serangkaian pertanyaan, yang kemudian dihitung jumlah berat keseluruhan jawabannya. Ketebalan inilah yang akan menyediakan jenis arah kepribadian, yang kemudian akan dicocokkan dengan tipe kepribadian dari hasil analisis grafologi dari tulisan tangan. tipe kepribadian dari hasil analisis tulisan tangan diproses berdasarkan ciri-ciri kepribadian yang merupakan hasil dari identifikasi kombinasi dari empat bentuk dominan dari tulisan tangan melalui output software dari penelitian sebelumnya, bahwa Slant (menulis tilt), Size (ukuran font), dasar, dan Breaks (tangguh setiap kata). Dari hasil penelitian ini dapat ditemukan ada korelasi antara analisis kepribadian berdasarkan ilmu psikologi dengan ilmu grafologi, yang menghasilkan tipe kepribadian yang cocok dengan 81,6% dari 49 responden Data yang berhasil diuji.

Kata Kunci: *Grafologi, Enneagram, Psikologi, Kepribadian, tulisan tangan.*

1. Introduction

Generally, the recognation or tests a human's personality is done by implementing a variety of methods in psychology, such as DAP test, MBTI, Wartegg, MAPP, Baum Tree with the intent and specific purpose, such as for the recruitment of new employees to fit the field of the selected job, requirements to go to college for prospective students and others. Each type of test can not be done in a single (only one method) and the result would only be known by psychologists through a series of lengthy analysis and takes time. This is the main reason for researchers to develop further research on the science of personality through handwriting analysis, in order to be able to replace conventional ways, namely through a psychological analysis with a shape feature extraction method based on science of graphology.

From the results of previous studies entitled 'Application of Graphology Science in Developing Handwritten Analyzer Device based on Shape Feature Extraction'[17], researchers have success-

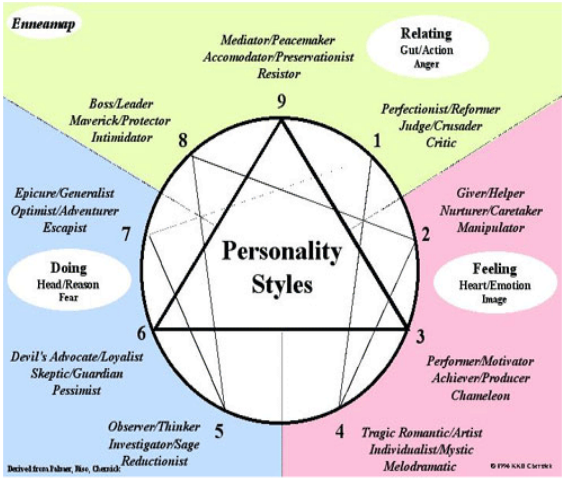

Figure 1. Personality Types based on Enneagram

fully developed a software that can provide results predominant form of a set of handwritten digital tested and the analysis of his personality to the level accuracy of 62.3 – 63.3%. The personality analysis obtained through the implementation of clustering methods, shape-based feature extraction, and the science of graphology to 23 authors (handwriting). However, the result of these studies obtained also some drawbacks such a side validation is only based on the results of the analysis of graphologists named Clifford Howard [8], and the type of personality that represented only through four categories of the dominant form of handwriting that size, slant, breaks, and baseline. Shortage is what researchers want to improve by trying to complete through the Enneagram method, in which the test results were expected to be able to prove the validity of the graphology science in recognizing one's personality and how big the accuracy of the software that was developed after matching with the science of psychology.

Enneagram method is one of psychology method that uses a series of questions to determine a human's personality (Figure 1). In addition to the Enneagram, there are several other types of tests. However, it's categorized as a method Enneagram personality test is most accurate when compared with the method or type of other questionnaire based test, with the percentage of accuracy between 80 – 87% [16][14]. This is because the method of assessing the Enneagram personality types based on the experience of someone who happened since early age. This is the reason researchers uses the Enneagram method in the study of psychology as a basic reference for assessing a human's personality that will be matched with the results of the personality assessment of the graphology science.

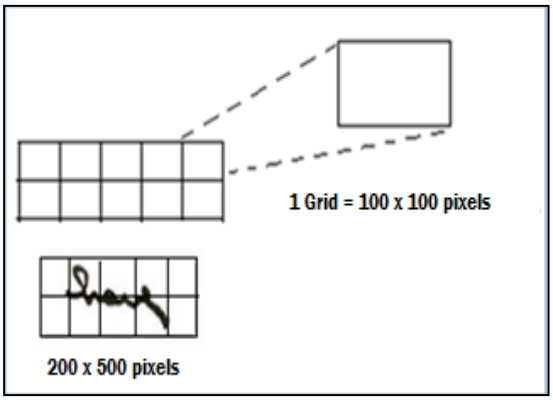

Figure 2. Sample of ROI Process

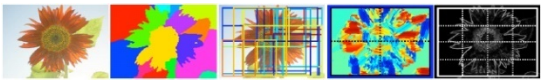

Figure 3. ROI Formation [7]

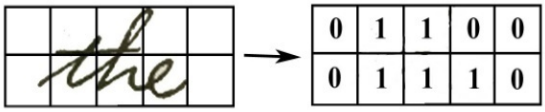

Figure 4. Sample of ROI Process

2. Methods

Preprocessing

An early stage that needs to be done to get the post data in digital form with the size of the pixel and gray level values or the same gray level of a set of handwritten analog that has been digitized through the scanner. This phase consists of the conversion of RGB colors to greyscale and thresholding.

RGB to Greyscale Color Conversion
RGB color conversion to greyscale is the stage for 24-bit color values to 8 bits, so the size of the resulting color will be smaller with the interval between 0 and 255 [2] :

$$RGB = \frac{R + G + B}{3} \qquad (1)$$

where R is the pixel value of red color, G is the pixel value of green, and B is the pixel value of blue color.

Thresholding
Thresholding is a process to separate the object region (foreground) to the background area over a certain threshold value [6]. In this study, the threshold value is also determined by trial and error.

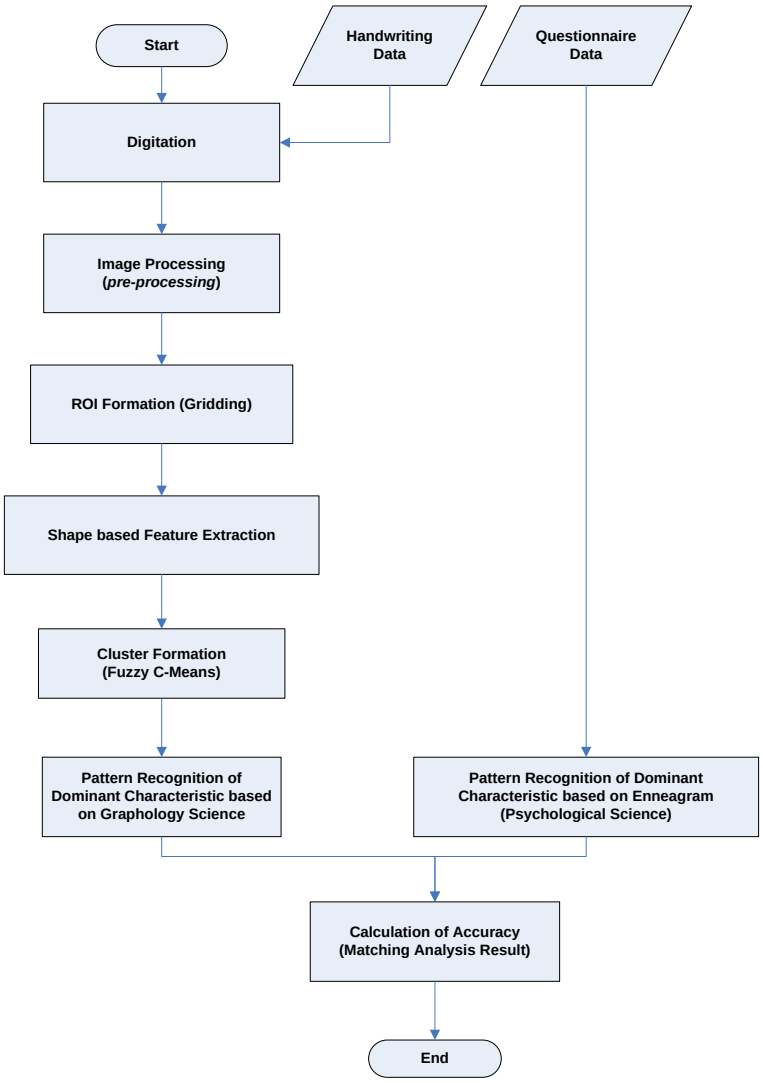

Figure 5. Research Flowchart

Formation of ROI

Formation of ROI (Region of Interest) is a technique that is generally done to help the analysis of the object to be observed, such as fMRI image analysis conducted by researchers from the UCLA – Los Angeles, Russel A. Poldrack in 2007 [5]. This technique can improve the success of the recognition phase, due to the formation of ROI, feature extraction process to be performed is limited to a specific region or area that has been restricted.

In this study, each handwritten documents that have been scanned and automatically crop (the size of 200 x 500 pixels) will be divided into grids with each grid size of 100 x 100 pixels (see Figure 2). The results of this gridding process will provide a total grid as 2 x 5 grid for each document of the handwriting. Each grid of each image document will then be carried shape based feature extraction process. Another example of ROI process is shown in Figure 3.

Feature Extraction

Feature extraction is an important stage of recognition application and pattern analysis. This stage will return the values of the features to be measured or identified as a pattern. With the extraction of features or characteristics, important information of the data (which in this study is in the form of handwritten image data) will be taken and stored in the feature vector [6]. Features that can be extracted in the form of image data including color features, shape, and texture. And in this study, the features will be extracted is based on the representation of the handwritten form.

In Figure 4, the values of the shape based features of the handwritten extraction will be binary values (values ‘0’ or ‘1’) to each grid for

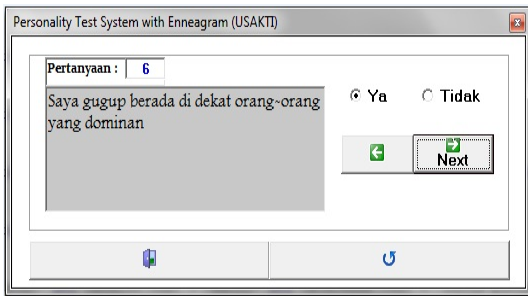

Figure 6. Enneagram Menu

each image, where the value of '0' will be given if the representation of the grid is the background object, and a value of '1' if the representation of the grid is a foreground object with a minimum of 15% of total pixels of each grid is a foreground object [1].

Formation of Cluster

Formation of clusters that are used in a previous study that is to classify the extracted shape features, in accordance with the number of clusters referenced handwritten form on the science of graphology. The method used is Fuzzy C-Means, the development method of Hard K-Means, where the centroid (center of cluster) determined from the acquisition of membership values search repeatedly with minimal distance. Fuzzy C-Means in their utilization can also be used for medical image segmentation, thus simplifying the diagnostic process as in research of Zhou *et al* [9]. Fuzzy C-Means formula is as follows :

$$J_m = \sum_{k=1}^{c}\sum_{i=1}^{N} u_{ki}^{m}\|x_i - v_k\|^2 \quad (2)$$

$$u_{ki} = \frac{1}{\sum_{l=1}^{c}\left(\frac{\|x_i - v_k\|}{\|x_i - v_l\|}\right)^{2/(m-1)}} \quad (3)$$

$$v_k = \frac{\sum_{i=1}^{N} u_{ki}^{m} x_i}{\sum_{i=1}^{N} u_{ki}^{m}} \quad (4)$$

where *Xi* is the image of handwritten to i with a total of *N* feature vectors, *Vk* is the cluster center (centroid) to *k* of the *c* total class, and *J* is the distance obtained quantitatively. To initialize the beginning, the center of cluster (*V*) will be randomly selected from the overall total image of the handwriting used. While *u* is the value of the feature vector of *k* cluster center up to a total *m* cluster center.

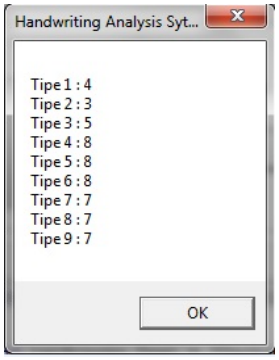

Figure 7. Weight Results of Enneagram

Pattern Recognition

Pattern recognition is one of the techniques of the science of Artificial Intelligence, which aims to recognize the features or the specific characteristics of data set (both text and image documents) and classifies [3]. Pattern recognition can be done in several ways, one of which is by using Similarity Measures.

Similarity measure is a method that can be used to find the similarity of objects to one another, by calculating the distance among them [4]. As the research conducted Anna Huang in 2008, this study also used the technique similarity measures in recognizing handwritten patterns by calculating the distance between the patterns by using the Euclidean distance formula [4]:

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots +} = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (5)$$

where *d* is the total distance between the handwritten image pixel values with each other, *q* and *p* are the pixels image.

Interpretation of Graphology

Graphology is a science which studies reading techniques of human character through handwriting from various viewpoints [8]. Research on the graphology science has been widely used to help figure out a human's personality and character. The experts of graphology (graphologist) states that a handwriting analysis can provide information on: 1) Strengths and weaknesses of one's character; 2) Human's behavior in a given situation; 3) The health condition of psychology, mental, and the ability of a person interactions on the current situation.

In providing the analysis, the science of graphology has some readings angle from handwriting form category, such as the slope of writing, pressure, size, and spacing writing [10]. Each category has a handwriting analysis of personality and character are different also.

Interpretation of Psychology

Psychology is defined as the scientific study of the behavior and mental processes of an organism [15]. By scientific meaningful studies conducted and data collected follows a systematic procedure, using psychological tests.

Psychological testing is a structured technique used to produce the examples chosen behavior. Example of this behavior will be used to make inferences about the psychological attributes of a person [15].

Enneagram is a method that has been proven to be very accurate to describe human personality [13][16]. This is because the method of assessing the personality types based on the experience of someone who happened since early age and see the influence of genetics.

Calculation of Accuracy and Validity

This stage is the final step, in which each of the results of handwriting analysis using graphology and enneagram questionnaire answers matched:

$$\frac{Result\ of\ handwriting\ using\ graphology}{Result\ of\ handwriting\ using\ enneagram}\ x\ 100\%$$

TABLE 1.1
HANDWRITTEN FEATURES MAPPING INTO THE ENNEAGRAM (1)

No	Base-line	Slant	Size	Breaks	Ennea-gram Type
1	Up	Leaning to the Left	Small	Dashed	4
2	Up	Leaning to the Left	Small	Connected	5
3	Up	Leaning to the Left	Large	Dashed	8
4	Up	Leaning to the Left	Large	Connected	7
5	Up	Leaning to the Left	Medium	Dashed	4
6	Up	Leaning to the Left	Medium	Connected	5
7	Up	Vertical	Small	Dashed	6
8	Up	Vertical	Small	Connected	5
9	Up	Vertical	Large	Dashed	7
10	Up	Vertical	Large	Connected	8
11	Up	Vertical	Medium	Dashed	3
12	Up	Vertical	Medium	Connected	5
13	Up	Leaning to the Right	Small	Dashed	7
14	Up	Leaning to the Right	Small	Connected	3
15	Up	Leaning to the Right	Large	Dashed	3
16	Up	Leaning to the Right	Large	Connected	1
17	Up	Leaning to the Right	Medium	Dashed	6
18	Up	Leaning to the Right	Medium	Connected	3
19	Flat	Leaning to the Left	Small	Dashed	4
20	Flat	Leaning to the Left	Small	Connected	6
21	Flat	Leaning to the Left	Large	Dashed	8
22	Flat	Leaning to the Left	Large	Connected	9
23	Flat	Leaning to the Left	Medium	Dashed	6
24	Flat	Leaning to the Left	Medium	Connected	6
25	Flat	Vertical	Small	Dashed	4
26	Flat	Vertical	Small	Connected	5
27	Flat	Vertical	Large	Dashed	3

TABLE 1.2
HANDWRITTEN FEATURES MAPPING INTO THE ENNEAGRAM (2)

No	Base-line	Slant	Size	Breaks	Ennea-gram Type
28	Flat	Vertical	Large	Connected	1
29	Flat	Vertical	Medium	Dashed	6
30	Flat	Vertical	Medium	Connected	7
31	Flat	Leaning to the Right	Small	Dashed	6
32	Flat	Leaning to the Right	Small	Connected	9
33	Flat	Leaning to the Right	Large	Dashed	7
34	Flat	Leaning to the Right	Large	Connected	1
35	Flat	Leaning to the Right	Medium	Dashed	3
36	Flat	Leaning to the Right	Medium	Connected	1
37	Down	Leaning to the Left	Small	Dashed	4
38	Down	Leaning to the Left	Small	Connected	6
39	Down	Leaning to the Left	Large	Dashed	8
40	Down	Leaning to the Left	Large	Connected	6
41	Down	Leaning to the Left	Medium	Dashed	8
42	Down	Leaning to the Left	Medium	Connected	6
43	Down	Vertical	Small	Dashed	2
44	Down	Vertical	Small	Connected	9
45	Down	Vertical	Large	Dashed	7
46	Down	Vertical	Large	Connected	5
47	Down	Vertical	Medium	Dashed	2
48	Down	Vertical	Medium	Connected	9
49	Down	Leaning to the Right	Small	Dashed	4
50	Down	Leaning to the Right	Small	Connected	6
51	Down	Leaning to the Right	Large	Dashed	6
52	Down	Leaning to the Right	Large	Connected	8
53	Down	Leaning to the Right	Medium	Dashed	6
54	Down	Leaning to the Right	Medium	Connected	8

Figure 8. Handwriting Analysis Menu

TABLE 2.1
HANDWRITTEN ANALYSIS RESULTS USING GRAPHOLOGY AND ENNEAGRAM (1)

No.	Respondent	Personality Type	
		Enneagram/ Psychologists	Graphology
1	Marcella	3 / 7 / 9	3
2	ANONYMOUS	3	3
3	Nathania	9	7
4	ANONYMOUS	2/3/5	5
5	ANONYMOUS	5/9	9
6	ANONYMOUS	3/8	3
7	ANONYMOUS	9	4
8	ANONYMOUS	2	2
9	ANONYMOUS	4	4
10	ANONYMOUS	-	-
11	Zilzikridini	4	9
12	Veni Emiriya	3	3
13	Yayang Nafisa	7	7
14	Vitria R Claudia	7/9	7
15	Nadine A.S	7	2
16	Sarah	3	1
17	Titania Raras N	4	4
18	Khumaira A.	6/9	6
19	Tara Ayu A.P	9	9
20	Niken	7	7
21	Nurul Mulya P.	9	9
22	Putri	3	3
23	Labiba	7	7
24	Melinda	9	9
25	Nur Fajriyah	4/7/9	7

The results of these matches will give system accuracy and prove the validity of the method of graphology in recognizing humas's personality.

Data Collection Technique

Collecting data in this research through the istribution of questionnaires directly to the various speakers, which in this study were collected totaled 50 data. The questionnaire contains 90 questions and an empty column to fill this handwriting will be processed in software to produce in accordance with the type of personality psychology and graphology.

TABLE 2.2
HANDWRITTEN ANALYSIS RESULTS USING GRAPHOLOGY AND ENNEAGRAM (2)

No.	Respondent	Personality Type	
		Enneagram/ Psychologists	Graphology
26	Tiara	7	7
27	Ranyta Diani	5	3
28	Sarah Widiyanti	6	6
29	Lady Margaretta	2	4
30	Rania Bahasoean	9	9
31	Sisilya Eva	3/7/8	8
32	Safira	5	5
33	Nissa	1/9	1
34	Ari Satria	8	8
35	Robert Muliawan	5	5
36	TiffanyMarcellie	2/6/7/9	6
37	Tiffany	2/9	9
38	Qotrunnadya	9	1
39	Siti Salediah	2	2
40	Namira	6	6
41	Petra Mario S.	2	2
42	Lea Insani	5/6/9	5
43	Marcellina	7	7
44	Sherli Betris	1	1
45	Viola	2	2
46	Shazlin	1	1
47	Manindya	9	7
48	Miranti V	8	8
49	Stella Febrina	1/7	1
50	RM	3	3

Research Design

In Figure 5 can be seen that in this study the researchers used comparison of the results of the introduction of personality types based on analysis graphology and psychology (through enneagram methods) to determine the accuracy of the system being developed. The personality characteristics of 11 features cluster referenced handwritten form obtained by reference to the science of graphology [8][11][12] and the results of research conducted by researchers previously [17]. Among the clusters 'Baseline', 'Slant', 'Breaks', and 'Size'. Baseline consists of 'Up', 'Flat', 'Down'. Slant consists of 'Leaning to the Left', 'Leaning to the Right' and 'Vertical'. Size consists of 'Large', 'Small', and 'Medium'. Breaks consists of 'Dashed', 'Connected'. Table 1 is a mapping table of handwriting form features into the enneagram through psychologists analysis results.

3. Results and Analysis

From the 50 data collected (Table 2), 1 data is considered not valid because the filling in its questionnaire that do not fit. Thus, the data examined totaled 49, where testing is done using software which has been developed.

The results of the answers in the enneagram menu (Figure 6), then calculated to determine the most weight. The highest weights will determine

the type of personality that is owned by the respondent.

Figure 7 shows that the respondent data were tested had the greatest weight value on types 4, 5, and 6. These results will then be matched with the testing of graphology method, which analyze trends owned personality type from the And-writing to see the shape features.

On the menu at the top (see Figure 8), shows that images of handwritten notes of the respondents will be extracted the shape features with a series of processes, which will then be assessed by a system of personality types and descriptions. Those results will then be matched with the enneagram assessment results :

And from the 49 tested data (1 data is not valid), this study got a match rate of 81.6% between the personality analysis of psychology to the graphology science. Where these percentages obtained from 40 data were tested with similar results, and 9 data have different results for the personality type (at Table 2). Thus, this study was able to prove that there is a correlation between a human's personality by the shape of her And-writing, so that in analyzing the behavior, characteristics, psychological condition of someone can through handwriting and not necessarily through psychological tests.

4. Conclusion

From these results, the researchers can then provide conclusions include :

Implementation of graphology method in analyzing someone's personality type of handwriting can have the same results with the analysis using enneagram method in psychology, with a match rate of 81.6%. So therefore, the science of graphology has been scientifically proven its validity through this research.

Implementation of graphology in giving analysis of human's personality results can be faster than the application of the enneagram, because graphology analysis requires only handwriting samples with simple sentences and without going through a series of questions and take as many psychological tests with enneagram.

References

[1] L. Guojun. *Multimedia Management Database Systems*. Artech House Inc. 1999

[2] D. Pratiwi. *The Use of Self Organizing Map Method and Feature Selection in Image Database Classificaton System*. International Journal of Computer Science Issues (IJCSI). 2012. Vol.9 Issue 3. No.2 ISSN : 1694-0814

[3] Y.A Absultany. *Pattern Recognition using Multilayer Neural Genetic Algorithm*. Neurocomputing. Elseiver Science. 2003. 237-247.

[4] A. Huang. *Similarity Measures for Text Document Clustering. New Zealand Computer Science Research Student Conference*. Christchurch New Zealand. 2008.

[5] R.A Poldrack. *Region of Interest Analysis for fMRI*. Oxford Journal. Los Angeles USA. 2007. Vol.2. Issue 1. Pp 67-70.

[6] D. Pratiwi, D.D Santika, B. Pardamean. *An Application of Backpropagation Artificial Neural Network for Measuring The Severity of Osteoarthritis*. International Journal of Engineering & Technology (IJET-IJENS). 2011. Vol.11. No.3. ISSN : 117303-8585

[7] G. Kim, A. Torralba. *Unsupervised Detection of Region of Interest Using Iterative Link Analysis*. Massachusetts Institute of Technology. 2009

[8] C. Howard. *Graphology*. The Pen Publishing Company. Philadelphia. 1922

[9] H. Zhou, G. Schaefer, C. Shi. *Fuzzy C-Means Techniques for Medical Image Segmentation*. Fuzzy System in Bioinformatics and Computational Biology Studies in Fuzziness and Soft Computing. Springer-Verlag Berlin. 2009. Vol. 242. Pp 257-271. ISSN : 1434-9922

[10] B. Ludvianto. *Grapho for Success : Analisis Tulisan Tangan*. PT Gramedia Pustaka Utama. Jakarta. 2013

[11] R. Coll, A. Fornes, J. Liados. *Graphological Analysis of Handwritten Text Documents for Human Resources Recruitment*. International Conference on Document Analysis and Recognition. IEEE Computer Society. DOI: 10.1109/ICDAR.2009.213 .2009.

[12] V, Kamath et al. *Development of an Automated Handwriting Analysis System*. Asian Research Publishing Network (ARPN) Journal of Engineering and Applied Science. 2011. Vol.6 No.9 ISSN : 1819-6608.

[13] R. Baron, E. Wagele. *Eneagram :Mengenal 9 Tipe Kepribadian Manusia dengan Lebih Asyik*. PT Serambi Ilmu Semesta. Jakarta. 2014.

[14] D. R. Riso, R. Hudson. *The Riso-Hudson Enneagram Type Indicator (RHETI)*. Version 2.5. Enneagram Institute. New York. 2001. Pp. 1-18. ISBN : 9780-9703824-0-5

[15] A. Poizner. *Graphology in Clinical Practice*. Psychologica. Spring Burlington. 2004. Vol. 24 No.1.

[16] S. S. Eric, T. Thomas. *The Enneagram and Brain Chemistry*. The Enneagram Institute.

https://www.enneagraminstitute.com/the-enneagram-and-brain-chemistry/ [diakses 2 Januari 2016]

[17] D. Pratiwi, A.B. Ariwibowo, F. Oktaviyanti. Penerapan Ilmu Grafologi dalam Membangun Piranti Penganalisa Tulisan Tangan melalui Ekstraksi Fitur Bentuk. SNTI IV Universitas Trisakti. Jakarta. 2014. 82-1 – 82-6. ISSN: 2355-925X

Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information). 10/1 (2017), 19-28
DOI: http://dx.doi.org/10.21609/jiki.v10i1.405

IDENTIFYING MEDICINAL PLANT LEAVES USING TEXTURES AND OPTIMAL COLOUR SPACES CHANNEL

C H Arun[1], and D Christopher Durairaj[2]

[1]Department of Computer Science, Nesamony Memorial Christian College, Marthandam-629165, India
[2]Department of Computer Science, Virudhunagar Hindu Nadars Senthikumara Nadar College, Virudhunagar-626001, India

E-mail: dass.arun@gmail.com

Abstract

This paper present an automated medicinal plant leaf identification system. The Colour Texture analysis of the leaves is done using the statistical, the Grey Tone Spatial Dependency Matrix (GTSDM) and the Local Binary Pattern (LBP) based features with 20 different color spaces (RGB, XYZ, CMY, YIQ, YUV, YCbCr, YES, U*V*W*, L*a*b*, L*u*v, lms, l$\alpha\beta$, $I_1I_2I_3$, HSV, HIS, IHLS, HIS, TSL, LSLM, and KLT). Classification of the medicinal plant is carried out with 70% of the dataset in training set and 30% in the test set. The classification performance is analysed with Stochastic Gradient Descent (SGD), k Nearest Neighbour(kNN), Support Vector Machines based on Radial basis function kernel(SVM-RBF), Linear Discriminant Analysis(LDA) and Quadratic Discriminant Analysis(QDA) classifiers. Results of classification on a dataset of 250 leaf images belonging to five different species of plants show the identification rate of 98.7 %. The results certainly show better identification due to the use of YUV, L*a*b* and HSV colour spaces.

Keywords: *colour spaces, texture features, plant identification, pattern recognition*

Abstrak

Makalah ini menyajikan sebuah tanaman obat sistem identifikasi daun otomatis. Analisis Warna Tekstur dari daun dilakukan dengan menggunakan statistik, Grey Tone Spatial Dependency Matrix (GTSDM) dan Pola Binary lokal (LBP) fitur berbasis dengan 20 ruang warna yang berbeda (RGB, XYZ, CMY, YIQ, YUV, YCbCr, YES, u * V * W *, L * a * b *, L * u * v, LMS, l$\alpha\beta$, $I_1I_2I_3$, HSV, HIS, IHLS, HIS, TSL, LSLM, dan KLT). Klasifikasi tanaman obat dilakukan dengan 70% dari dataset di set pelatihan dan 30% dalam tes set. Kinerja klasifikasi dianalisis dengan Stochastic Gradient Descent (SGD), k Tetangga terdekat (kNN), Dukungan Mesin Vector berdasarkan Radial fungsi dasar kernel (SVM-RBF), Linear Discriminant Analysis (LDA) dan kuadrat Analisis Diskriminan (QDa) pengklasifikasi. Hasil klasifikasi pada dataset dari 250 gambar daun milik lima spesies yang berbeda dari tanaman menunjukkan tingkat identifikasi 98,7%. Hasil tentu menunjukkan identifikasi yang lebih baik karena penggunaan YUV, L * a * b * dan ruang warna HSV.

Kata Kunci: *ruang warna, fitur tekstur, identifikasi tanaman, pengenalan pola*

1. Introduction

In the western ghats region of Asia, especially in Kanyakumari district, the southern most district of India, traditional home herbal remedies are readily available in most of the homes. These herbal plants are used in treating common sicknesses like common cold, diarrhoea and headache. The herbal practitioners, herbal medical industries and the even the common lay people have a good knowledge about using these herbal plants. They collect the needed herbal plants from the available sources and then use the plant leaves for the preparation of the medicine.

World has seen a steady interest and usage of traditional herbal remedies and herbal produces [1]. World Health Organization (WHO) reports that about 65-80% of the world's population in developing countries depend on plants for their primary healthcare due to poverty and lack of access to modern medicine [2].

Leaf is one of the major identifying features of a plant. Leaves are far being the only discriminant visual key between species but, due to the shape and size, they have the advantage to be easily observed, captured and described [3]. Though all the medicinal plant parts like roots, flowers and barks are used for medicinal purposes, the most common part with more medicinal value are the fresh leaves. Identification of the

respective medicinal plant leaves are normally done using plant taxonomic field guides by the Botanists. The Siddha medicinal practitioner (the traditional herbal physician) who mostly rely on medicinal plant leaves, and the households, depend on their knowledge gained and their experience in identifying the plant leaves. Though, this has been the regular practice for generations, modernisation has it is impact and most of the younger generations do not have enough knowledge in identifying the respective medicinal plant leaves. Wrong identification of these leaves can lead to more damage while treating the patients. As there is a need for precise and quick identification of the plant leaves, an automatic identification system would prove to be an effective solution.

Five medicinal plant leaves are considered in the present work and some of their medicinal uses are given below. A sample of the medicinal plant leaf dataset from different plant leaves is given in the Figure 1. (0) Desmodium gyrans - Antidote for snake poison, effective for heart diseases, rheumatic complaints, diabetes and skin ailments. (1) Butea monosperma - promotes diuresis & antihelminthic, treats leucorrhoea & diabetes. (2) Malpighia glabra - help lower blood sugar, increases collagen and elastin production, treats diarrhoea, dysentery, and liver problems. (3) Helicteres isora - help treat intestinal complaints, colic pains and flatulence. (4) Gymnema sylvestre - suppresses the sensation of sweet, anti-diabetic.

Automatic plant leaf identification is a difficult problem because there is often high intra-species variability, and low interspecies variation [4]. But many promising approaches have emerged. Image based plant leaf identification is done using morphological characteristics(shape, colour, texture, margin [4-7]). Automatic identification of medicinal plant leaves are done using texture, colour, statistical and Local Binary Pattern(LBP) features [8-10]. The identified plant leaves are used in food processing, medical, botanical gardening and cosmetic industry [11].

Colours are represented using different colour spaces. There is no particular colour space that is proving itself best for all the colour images. The use of colour increases the performances of the standard grey level texture analysis techniques [12,13]. To exploit the advantage of the colour component of the medicinal plant leaves which are mostly green in colour, identifying medicinal plant leaves using colour shall be considered in this research work. Though colour is not stable during the growth period of the leaves, colour space usage will certainly be providing more information in better identification. Selection of

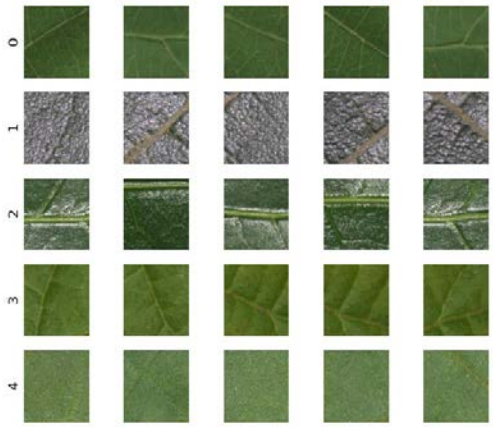

Figure 1. Leaf images of Desmodium gyrans(row 0), Butea monosperma(row 1), Malpighia glabra(row 2), Helicteres isora(row 3)and Gymnema sylvestre(row 4)

the suitable colour space for the particular identification depends on the specific image [14]. Different colour component based image processing produces reliable and accurate results [15] and improves classification performance [16]. Several approaches used in automatic identification of plant leaves use HSV colour space, which results in better performance [17].

The decorrelated colour space L*a*b* is said to perform best in colour transfer algorithms which yields better quantification results in foods with curved surfaces [18]. Y CbCr colour model outperforms other specified models RGB, YIQ, YCbCr, HSV and HSI in terms of objective quality assessment for Colour Image Fusion [19], YUV colour space is used to solve parameter optimisation of blind colour image fusion [20]. The hybrid colour space RCrQ possesses complementary characteristics and enhances discriminative power for face recognition [21].

The objective of the present work is to compute the colour texture features using various colour spaces based on greylevel, the Grey Tone Spatial Dependency Matrix (GTSDM) and the LBP operators and to find out the suitable colour space in identifying the medicinal plant leaf automatically.

Suitable image classification approaches can be used for the identification process. The rest of the paper is organized as follows: Methodology of the work is discussed in section 2 which includes the details of texture feature extraction, colour transform, classification and the description of the dataset. Design and implementation of the work are discussed in section 2. The results and discussion are given in section 3, which is followed by conclusion in section 4.

Algorithm 1:

Overview

1: Get image L from the leaf dataset.
2: Apply the colour transform in L producing the transformed image T.
3: Extract the features fi from the transformed image T (where i = 1 to 12).
4: Store the feature values fi in the database.
5: Repeat the steps 1 to 4 for all training images.
6: Read a test image I from the dataset.
7: Extract the features t_i from the test image I (where i= 1 to 12)
8: Apply the classifiers using the features fi and t_i.
9: Display the classified output.
10: Calculate the classification accuracy.
11: Repeat the stages 6 to 10 for all testing images.

end

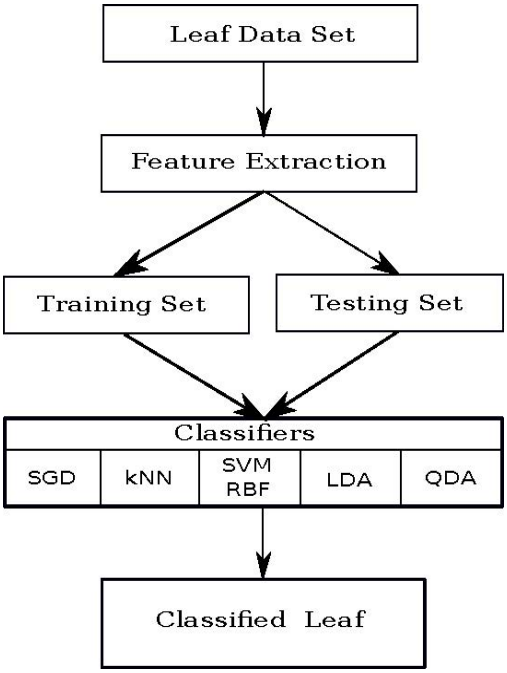

Figure 2. Block diagram for Classification

2. Methods

The general approach adopted for identifying medisinal plant leaves are explained here. The different stages like texture feature extraction, colour transform and classification algorithms are presented in detail. Primary dataset is elaborated. The Algorithm 1 shows the overview of procedure to obtain the classified medicinal plant leaf.

Texture Feature Extraction

Texture features in identifying medicinal plant leaves can be computed using first-order statistical moments (mean, variance and skewness) and second-order statistical moment [Grey Tone Spatial Dependency Matrix-GTSDM (energy, homogeneity, dissimilarity, entropy, correlation and contrast) and Local Binary Pattern-LBP (mean and standard deviation)] operators[9]. First-order grey level statistical features are the simplest texture descriptors. GTSDM or GLCM are the second-order statistical features which are more sensitive and more intuitive than first-order features. Local Binary Pattern (LBP) operators are known for its invariance to local gray scale variations, monotonic photometric changes and its high descriptive power.

Colour Transform

The colour space suitable for medicinal plant leaf identification has to be found using experimental testing, as in most cases, the best established colour space is not suiting to all types of dataset [15]. Out of the various existing colour spaces, twenty colour spaces are considered. The colour spaces considered for our experimental testing are categorized into five colour space families [22]. The transformation equations for all the considered colour spaces are given in their respective citations. The colour spaces are, primary colour spaces: RGB, XYZ [23] and CMY [24]; luminance-Chrominance spaces: YIQ[23], YUV [23], Y CbCr [21], YES [25], U*V *W* [26], L*a*b* [27], L*u*v* [27], lms [28] and l [28]; independent axis space: $I_1I_2I_3$ [29]; the perceptual spaces: HSV [18], HIS [30], IHLS [31], HIS [32] and TSL [33] and the other colour spaces: LSLM [27] and KLT [34].

Classification

Texture based image classification involves deciding the most pertinent texture category of the observed image [35]. Classification here would mean the machine language classification of the different classes and not the linnaean taxonomic system of plant classification [36]. When the prior knowledge of the established classes are available and the texture features are extracted, the given image could be classified to the appropriate class.

The texture class i consisting of a set of n images can be represented as

$$\boldsymbol{T_i} = \{\boldsymbol{t_{i,1}}, \boldsymbol{t_{i,2}}, \dots, \boldsymbol{t_{i,n}}\} \quad (1)$$

Where $t_{i,j}$ is the member image.

Various classifiers used in this work are Stochastic Gradient Descent (SGD) [37,38], k Nearest Neighbour (kNN) [39,40], Support Vector Machines based on Radial basis function kernel(RBF) [41], Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) [42]. SGD is an approach to discriminative learning of linear classifiers under convex loss functions. kNN computes the distance between a test point and all points in the training set. SVM are supervisor learning methods which are similar to SGDs and these are effective in high dimensional feature spaces. Linear Discriminant Analysis (LDA) is a classifier with a linear decision boundary, generated by fitting class conditional densities to the data. Quadratic Discriminant Analysis (QDA) is a classifier with a quadratic decision boundary, generated by fitting class conditional densities to the data. Parameter tuning is of importance while dealing with certain classifiers. The k value in the kNN classifier ranges from 1 to a maximum of the square root of the number of instances [43]. Choosing a lower k value will have a greater influence of noise on the result while a higher k value will increase the computation time. Moreover the value of k is chosen to be a odd value to lessen the computation time [44]. The parameters of SVM with a Gaussian radial basis function (RBF) kernel are C and $\boldsymbol{\gamma}$ the parameter C controls the influence of each individual support vector. A low bias and high variance is obtained from larger C and vice versa. The parameter handles non-linear classification. A higher gamma value will give a high bias, low variance and vice versa. The standard and optimal method to choose the optimal parameters C and $\boldsymbol{\gamma}$ is a Grid Search (GS) [45].

The classifier places all of the N training images in a vector space with N dimensions, based on the features extracted from the image. When a new uncategorized test image is placed in that vector space, the best suited image is found using the pairwise-distance function. The pairwise-distance between two images are calculated as given in the equation (2) where d1 and d2 are the two images. When the distance between d1 and d2 are the smallest, the most suitable image is obtained.

$$\boldsymbol{distance(d1, d2) = 1/L} \qquad (2)$$

Where L is the number of matching images.

The test image is estimated to belong to a specific category, after generation of its features, when the majority of its neighbors are also in the same category.

Algorithm 2:

Functional

1: Let L be the leaf image, where $L = LIj$ and $j = 1,2\cdots p$
2: Form the transformed image T using the colour transform for the leaf image L, where $T = CTj$ and $j = 1,2\cdots p$
3: Calculate all the following features for each channel of the transformed matrix T.
   a) Grey features - mean($\mu 0$), variance($\mu 2$), standard deviation(σ) and skewness($\mu 3$)
   b) Grey Tone Spatial Dependency matrix(GTSDM) - energy(ζ), homogeneity(η), dissimilarity(ζ), entropy(E), correlation(ρ) and contrast(λ)
   c) Local Binary Pattern(LBP) - mean(μL) and standard deviation(σL).
4: Insert calculated features in the feature vector, $f1$ for each channels of the transformed matrix T.
5: Repeat the stages 1 to 4 for all the training set images
6: Read the test image I from the dataset
7: Form the transformed image K for the test image I, where $K = CTj$ and $j = 1,2\cdots p$
8: Calculate all the features listed in step 3 for each channel of the transformed image K.
9: Insert the calculated feature in the feature vector $f2$ for the test image.
10: Apply the classifiers SGD, kNN, SVM-RBF, LDA and QDA separately using the feature set $f1$ and $f2$.
11: Repeat the stages 6 to 10 and display the classified output
12: Calculate the classification accuracy

end

Datasets

The medicinal plant leaves are collected around the Western Ghats region of Kanyakumari forests. The digital images are obtained using the Canon EOS 1000D digital camera on the abaxial portions of the medicinal plant leaves in the uncompressed JPEG format with the dimension, 3420 x 4320 x 3, in a closed environment to maintain constant illumination. As the leaf characteristics, especially in terms of the colour, vary widely from its tender stage to the mature stage, the proposed algorithm is restricted for images of mature leaves of a

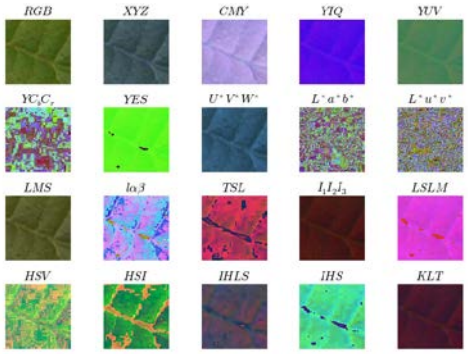

Figure 3. A leaf sample image in all considered Colour Spaces

TABLE 1
FIRST ORDER STATISTICAL SAMPLE FEATURE VALUES OF VARIOUS CLASSES

Mean	Variant	Std.Dev	Skewness	Class
133.1596	545.1827	23.3491	0.47	0
58.1753	525.6315	22.9267	3.1372	1
53.3689	713.9928	26.7206	-0.469	2
68.3939	452.3641	21.2689	0.0202	3
60.5693	221.0215	14.8668	0.2602	4

plant. The working dataset is formed by dividing the images into a size of 50 x 50 x 3 and forming a set total of 250 images, 50 images each from 5 medicinal plant leaves. The training and the testing sets are formed from the image datasets.

Design and Implementation

The implementation of the proposed algorithm is done using Python as the programming language in Linux platform along with the open-source libraries namely numpy, scipy, scikit-learn, colorsys and grapefruit. The various stages of the algorithm developed in identifying the medicinal plant leaves using colour textures are as given in Algorithm 2 and are explained here.

The medicinal leaf image from our dataset is available as an m x n x p colour image L with p colour channels in the form of equation (3)

$$\boldsymbol{LI_k} = [\boldsymbol{c_{ij}}] \tag{3}$$

where i = 1,2,…m, j = 1, 2,… n, k = 1,2…p. p represents the number colour channels of the image.

The different colour space transformation, CTk in equation(4), is calculated from the colour channels LIk using the corresponding colour transform equations given in the respective citations of section 2. A sample of all the considered colour space leaf images are given in the Figure 3

TABLE 3
CLASSIFICATION RESULT BASED ON DIFFERENT COLOUR SPACES

Colour Space	SGD	kNN	SVM RBF	LDA	QDA
RGB-B	94.7	96.0	96.0	93.3	96.0
XYZ-Y	94.7	93.3	94.7	90.7	96.0
CMY-Y	94.7	96.0	96.0	93.3	96.0
YIQ-Y	94.7	94.7	94.7	90.7	96.0
YUV-U	96.0	93.3.	93.3	92.0	98.7
YCbCr-Cb	92.0	92.0	94.7	93.3	96.0
YES-E	94.7	86.7	92.0	89.3	97.3
U*V*W*-V*	94.7	94.7	94.7	90.7	96.0
L*a*b*-a*	97.3	93.3	98.7	93.3	94.7
L*u*v*-L*	96.0	88.0	93.3	85.3	94.7
LMS-S	94.7	96.0	96.0	93.3	96.0
$l\alpha\beta$-l	94.7	92.0	93.3	89.3	97.3
$I_1I_2I_3$-I_1	94.7	93.3	96.0	90.7	96.0
HSV-H	96.0	97.3	98.7	92.0	96.0
HIS-I	94.7	93.3	96.0	90.7	96.0
IHLS-H	92.0	92.0	96.0	93.3	96.0
IHS-I	94.7	93.3	96.0	90.7	96.0
TSL-T	92.0	90.7	96.0	90.7	96.0
LSLM-LM	92.0	92.0	89.3	88.0	97.3
KLT-K	94.7	93.3	96.0	90.7	96.0

$$\boldsymbol{CT_k} = [\boldsymbol{t_{ij}}] \tag{4}$$

where i = 1, 2,… m,j = 1, 2… n, k = 1, 2… p. p represents the number colour channels of the image.

The statistical features using mean ($\boldsymbol{\mu 0}$), variance ($\boldsymbol{\mu 2}$), standard deviation ($\boldsymbol{\sigma}$) and skewness ($\boldsymbol{\mu 3}$) are calculated directly from the transformation CTk: The co-occurrence matrix and the haralick features energy ($\boldsymbol{\zeta}$), homogeneity ($\boldsymbol{\eta}$), dissimilarity ($\boldsymbol{\xi}$), entropy ($\boldsymbol{\epsilon}$), correlation ($\boldsymbol{\rho}$) and contrast ($\boldsymbol{\lambda}$), are calculated from the selected channel of the transformation CTk [35]. The feature values are used for classification.

The image CTk is used to calculate the local binary pattern operator feature using 8 neighbors and a unit radius for computation of the pattern. From the computed LBP histogram image, the mean of the histogram ($\boldsymbol{\mu}$L) and standard deviation ($\boldsymbol{\sigma}$L) are calculated and used as a feature.

The general procedure of classification shown in Figure 2 is followed with the dataset. After extracting the features $\boldsymbol{\mu 0}$, $\boldsymbol{\mu}2$, $\boldsymbol{\sigma}$, $\boldsymbol{\mu}3$, $\boldsymbol{\zeta}$, $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, $\boldsymbol{\epsilon}$, $\boldsymbol{\lambda}$, $\boldsymbol{\rho}$, $\boldsymbol{\mu}$L and $\boldsymbol{\sigma}$L for the data set, which can be divided into two sets, the classification is done in two stages: training and testing. The 70% of dataset is considered as training samples and the remaining 30% as testing images. The medicinal plant leaves can be classified using the trained

TABLE 2
SECOND ORDER GTSDM, LBP SAMPLE FEATURE VALUES OF VARIOUS CLASSES

GTSDM Energy	GTSDM Homogeneity	GTSDM Dissimilarity	GTSDM Entropy	GTSDM Correlation	GTSDM Contrast	LBP Mean	LBP SD	Class
0.077	0.3624	2.2302	5.4448	0.8678	8.7101	119.0731	84.315	0
0.2847	0.6988	0.9385	3.3051	0.8725	4.7482	108.5133	91.9709	1
0.2775	0.7737	0.4689	2.89	0.8868	0.5507	113.5907	88.2488	2
0.2312	0.6958	0.6588	3.2908	0.8587	0.9144	113.0331	86.4206	3
0.4263	0.88	0.2412	2.086	0.8727	0.2476	97.5045	90.3729	4

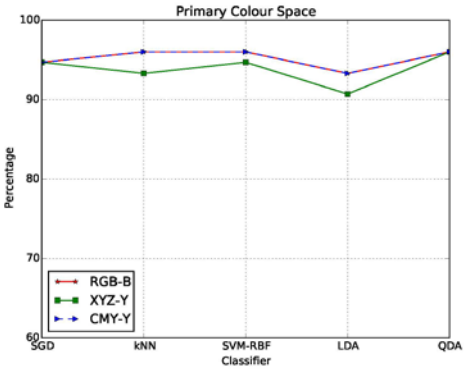

Figure 4. Primary Colour Spaces

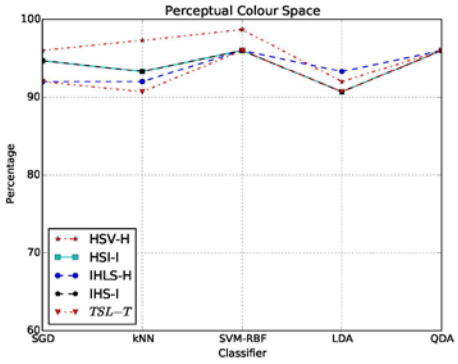

Figure 6. Perceptual Colour Spaces

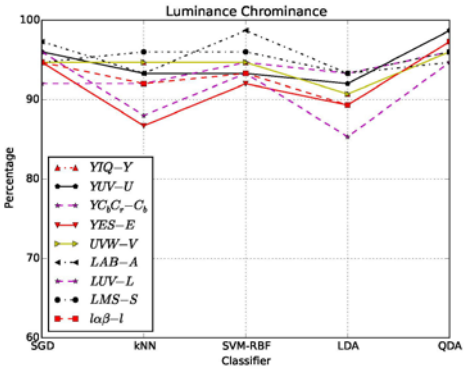

Figure 5. Luminance Chrominance Colour Spaces

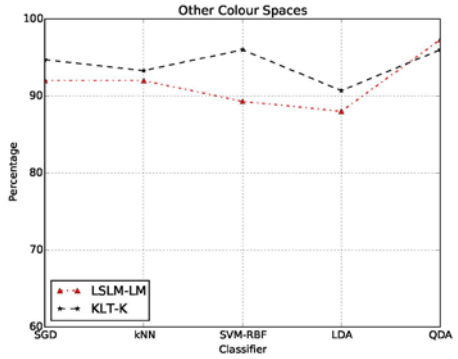

Figure 7. Other Colour Spaces

samples. The classification is performed by the different classifiers SGD, kNN, SVM-RBF, LDA and QDA. The parameter k for the kNN classifier in our work is chosen to be 3 which is an odd k value and is between the specified range. The C and γ parameters in SVM-RBF classifier are selected using Grid Search function as C = 1 and γ = 2. The parameters for SGD classifier considered are α = 0.001 and the number of iterations = 100.

3. Results and Analysis

The different colour spaces play a key role in this process of classification. The leaf images taken from the dataset are transformed into different colour spaces. For each colour space is considered with different colour channels. The feature values are computed for all colour channels. Sample feature values of various classes are given in the Table 1. The classification is done using the feature values of each colour channel.

Performance Analysis

The recognition percentage obtained for different colour spaces varies on the colour channels. The best classification result for different colour spaces shown here is considered based on the best classification result channel. The detailed classification result obtained for the colour channel of various colour spaces are shown in Table 3.

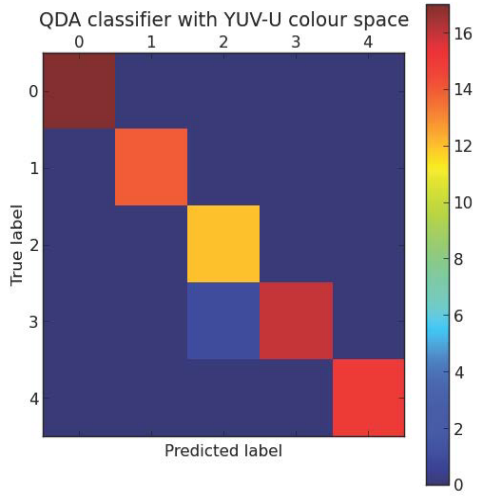

Figure 8. Confusion Matrix for QDA classifier with YUV-U colour space

TABLE 4
CLASSIFICATION REPORT ON THE QDA CLASSIFIER FOR YUV-U COLOUR SPACE

Leaf	Precision	Recall	f1-score	Support
Leaf 0	1.00	1.00	1.00	17
Leaf 1	1.00	1.00	1.00	14
Leaf 2	0.92	1.00	0.96	12
Leaf 3	1.00	0.94	0.97	17
Leaf 4	1.00	1.00	1.00	15
Avg/total	0.99	0.99	0.99	75

Primary Colour Spaces: It is observed that the blue channel produced good result of 96% than the red and green channels of RGB colour space for kNN, SVM-RBF and QDA classifiers. In the XYZ colour space, the Y channel gives 96%, which is the best accuracy with X, Y and Z channels for QDA classifier. The CMY colour space performs best with 96% on the Y channel for the three classifiers kNN, SVM-RBF and QDA in yellow channel. The consolidated primary family colour spaces performances are given in Figure 4 The RGB-B and CMY-Y are both giving the best performance of 96 % with kNN, SVM-RBF and QDA.

Luminance Chrominance Colour Spaces: The YIQ colour space presents the best recognition rate of 96% for QDA classifier in Y channel. For YUV colour space, U channel gives very good performance than all other colour space and channels with 98.7% for QDA classifier. The Y CbCr colour space performs the result of 96% in QDA classifier with Cb channel. In YES colour space E channel presents the 97.3% recognition rate for QDA classifer. The U*V*W* colour space produced 96% in V* channel for QDA classifier. The colour space L*a*b* very well performed

TABLE 5
COMPARISON WITH OTHER METHODS

Work	Method	Dataset (classes)	Accuracy
[47]	Fourier Moments	32 (flavia)	62.00%
[48]	1-NN	60	82.33%
[47]	PNN-PCNN	32 (flavia)	91.00%
[9]	I+GTDSDM+LBP	5	94.70%
[46]	Neighborhood	30	95.83%
[47]	SVM-BDT	32 (flavia)	96.00%
Proposed	I+GTDSDM+LBP+ 20 colour spaces	5	98.70%

with 98.7% in SVM-RBF classifier. The channel a* projected the highest recognition rate of 98.7% in L*a*b*. The L*u*v* colour space manages 96% in L* channel for the SGD classifier. The S channel in LMS colour space proves with the performance of 96% in three classifiers. The L*u*v* colour space provides 97.3% in L* channel for QDA classifier. The consolidated Luminance Chrominance colour spaces performances are given in Figure 5. The YUV-U with QDA, L*a*b*-a* with SVM-RBF gives the best performance of 98.7%.

Independent Colour Space: The $I_1I_2I_3$ colour space gives 96% in I_1 channel for SVM-RBF and QDA classifiers.

Perceptual Colour Spaces: The HSV colour space is the best performing colour space for all classifiers except LDA. For SVM-RBF classifier, HSV-H presents the maximum performance of 98.7%. The colour spaces HSI-I, IHLS-H, IHS-I and TSL-T are best at 96% with SVM-RBF and QDA. The performances of Perceptual family of colour spaces are given in Fig 6. The HSV-H yields the best performance of 98.7% with SVM-RBF.

Other Colour Spaces: LSLM-LM presents the maximum performance of 97.3%. The LSLM-LM colour space performs best at 97.3% with QDA. The KLT-K colour space performs best at 96.0% with SVM-RBF and QDA.The consolidated performances are given in Figure 7.

When the classifier performance is considered, the QDA performs the best in all colour spaces. The SVM-RBF comes next to QDA in the classification rate. For L*a*b* and HSV, the SVM -RBF produced the best result of 98.7%. For YUV, the QDA produced the best result of 98.7%.

Precision Recall

The precision, recall, f1-support and support are calculated for one of the best performing classifiers QDA, which gives a performance of 98.7% in identification of the leaf and is given in the

Table 4. The confusion matrix for the classifier QDA with the YUV-U channel is given in Figure

Comparison with Related Works

The proposed work is done with the texture features and the colour transforms. The leaf shape, venation are not considered here. The work presented here produced best results than the previous [9] work done with the same database. The comparative results of the different methods are listed out in Table 5. The computations were done with gray scale operations. The neighborhood method [46] operated with 30 plants gives 95.83% accuracy. The Fourier moments [47] with Flavia database gives 62.00%. The PNN-PCNN [47] method applied with Flavia database produces 91.00% accuracy. The SVM-BDT [47] applied with Flavia database presents 96.00% performance. The 82.33% accuracy is found for 1-NN method [48] with 60 species. The texture feature method [9] for 5 species of medicinal plants produced the performance accurac of 94.7%. The proposed method applied with the individual components of the 20 colour spaces considered, yielded the better accuracy of 98.7%.

4. Conclusion

In this paper, the method to identify the medicinal plants automatically with the help of colour textures calculating the statistical, GTSDM and LBP texture features is proposed. Though a single shade of colour for the considered mature plant leaf is not stable during the growth period of the leaves, usage of colour space is providing more information that helps for better identification. The limitation of this method in identification is that this method works only with the matured leaves of the plant and that it demands more computational time. The colour space YUV-U with QDA, L*a*b*-a* with SVM-RBF, and HSV-H with SVM-RBF give the finest performance of 98.7% among the twenty colour spaces considered and hence it is concluded that the suitable colour space for identifying medicinal plant leaves are YUV with U channel, L*a*b* with a* channel and HSV with H channel. The QDA classifier performs best in most of the different colour spaces considered.

References

[1] “Homeopathic and herbal remedies-us-report.” Mintel, April 2011.

[2] J. B. Calixto, “Twenty-five years of research on medicinal plants in latin america: a personal view,” Journal of ethnopharmacology, vol. 100, no. 1, pp. 131–134, 2005.

[3] A. Joly, H. Goe¨au, P. Bonnet, V. Bakic´, J. Barbe, S. Selmi, I. Yahiaoui, J. Carre´, E. Mouysset, J.-F. Molino et al., “Interactive plant identification based on social image data,” Ecological Informatics, vol. 23, pp. 22–34, 2014.

[4] T. Beghin, J. S. Cope, P. Remagnino, and S. Barman, “Shape and texture based plant leaf classification,” in International Conference on Advanced Concepts for Intelligent Vision Systems. Springer, 2010, pp. 345–353.

[5] H. Kebapci, B. A. Yanikoglu, and G. B. U¨ nal, “Plant image retrieval using color, shape and texture features,” Comput. J., vol. 54, no. 9, pp. 1475–1490, 2011.

[6] J. S. Cope, P. Remagnino, S. Barman, and P. Wilkin, “Plant texture classification using gabor co-occurrences,” in International Symposium on Visual Computing. Springer, 2010, pp. 669–677.

[7] J. S. Cope, D. P. A. Corney, J. Y. Clark, P. Remagnino, and P. Wilkin, “Plant species identification using digital morphometrics: A review,” Expert Syst. Appl., vol. 39, no. 8, pp. 7562–7573, 2012.

[8] Y. Herdiyeni and N. Wahyuni, “Mobile application for indonesian medicinal plants identification using fuzzy local binary pattern and fuzzy color histogram,” in Advanced Computer Science and Information Systems (ICACSIS), 2012 International Conference on, Dec 2012, pp. 301–306.

[9] C. H. Arun, W. R. Sam Emmanuel, and D. Christopher Durairaj, “Texture feature extraction for identification of medicinal plants and comparison of different classifiers,” International Journal of Computer Applications, vol. 62, no. 12, pp. 1–9, January 2013, published by Foundation of Computer Science, New York, USA.

[10] Y. Herdiyeni, E. Nurfadhilah, E. A. Zuhud, E. K. Damayanti, K. Arai, and H. Okumura, “Article: A computer aided system for tropical leaf medicinal plant identification,” International Journal on Advanced Science, Engineering and Information Technology, vol. 3, no. 1, pp. 23–27, 05 2013.

[11] B. Satyabama, “Content based leaf image retrieval (cblir) using shape, color and texture features,” Indian Journal of Computer Science and Engineering (IJCSE), vol. 2, no. 2, pp. 202–211, Apr - May 2011.

[12] A. Drimbarean and P. F. Whelan, “Colour texture analysis: A compara- tive study,” Vision System Laboratory, School of

Electronic Engineering, Dublin City University, 2000.

[13] P. F. Whelan and O. Ghita, "Colour texture analysis," 2008.

[14] K. Y. Lin, J. H. Wu, and L. H. Xu, "A Survey On Color Image Segmentation Techniques," Journal of Image And Graphics, vol. 10, no. 1, pp. 1–10, 2005.

[15] E. Reinhard and T. Pouli, "Colour spaces for colour transfer," in Proceedings of the Third International Conference on Computational Color Imaging, ser. CCIW'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 1–15.

[16] A. Drimbarean and P. F. Whelan, "Experiments in colour texture analysis," Pattern Recogn. Lett., vol. 22, no. 10, pp. 1161–1167, Aug. 2001.

[17] A. Ehsanirad and Y. Sharath Kumar, "Leaf recognition for plant clas- sification using glcm and pca methods," Oriental Journal of Computer Science and Technology, vol. 3, no. 1, pp. 31–36, 2010.

[18] F. Mendoza, P. Dejmek, and J. Aguilera, "Calibrated color measurements of fruits and vegetables using image analysis," Postharvest Biology and Technology, vol. 41, pp. 285–295, 2006.

[19] W. Rattanapitak and S. Udomhunsakul, "Comparative efficiency of color models for multi-focus color image fusion," Hong Kong, 2010.

[20] Y. Niu and L. Shen, "The optimal multi-objective optimization using pso in blind color image fusion," in 2007 International conference on multimedia and ubiquitous engineering (MUE'07). IEEE, 2007, pp. 970–975.

[21] Z. Liu and C. Liu, "Fusion of color, local spatial and global frequency information for face recognition." Pattern Recognition, vol. 43, no. 8, pp. 2882–2890, 2010.

[22] L. Busin, J. Shi, N. Vandenbroucke, and L. Macaire, "Color space selection for color image segmentation by spectral clustering," in Signal and Image Processing Applications (ICSIPA), 2009 IEEE International Conference on, Nov 2009, pp. 262–267.

[23] P. Shih and C. Liu, "Comparative assessment of content-based face image retrieval in different color spaces." IJPRAI, vol. 19, no. 7, pp. 873–893, 2005.

[24] P. Bourke, "Converting between rgb and cmy, yiq, yuv," Texture Colour, pp. 1–7, Feb 1994.

[25] E. Saber and A. M. Tekalp, "Frontal-view face detection and facial fea- ture extraction using color, shape and symmetry based cost functions," Pattern Recognition Letters, vol. 19, no. 8, pp. 669 – 680, 1998.

[26] J. Ladd and J. Pinney, "Empirical relationships with the munsell value scale," Proceedings of the Institute of Radio Engineers, vol. 43, no. 9, p. 1137, September 1955.

[27] P. Colantoni et al., "Color space transformations," see http://www.radugar-yazan. ru/files/doc/colorspacetrans-form95. pdf, 2004.

[28] C.-M. Wang and Y.-H. Huang, "A novel color transfer algorithm for image sequences," Journal of Information Science and Engineering, vol. 20, no. 6, pp. 1039–1056, 2004.

[29] Y. Ohta, Knowledge-based interpretation of outdoor natural color scenes. Morgan Kaufmann, 1985, vol. 4.

[30] W. ladys law Skarbek and A. Koschan, "Colour image segmentation a survey," IEEE Transactions on circuits and systems for Video Technol- ogy, vol. 14, 1994.

[31] J. Angulo and J. Serra, "Mathematical morphology in color spaces ap- plied to the analysis of cartographic images," Proceedings of GEOPRO, vol. 3, pp. 59–66, 2003.

[32] T.-M. Tu, S.-C. Su, H.-C. Shyu, and P. S. Huang, "A new look at ihs-like image fusion methods," Information Fusion, vol. 2, no. 3, pp. 177–186, 2001.

[33] D. Butler, S. Sridharan, and V. Chandran, "Chromatic colour spaces for skin detection using gmms," in Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, vol. 4. IEEE, 2002, pp. IV–3620.

[34] R. Kountchev and R. Kountcheva, "Image color space transform with enhanced klt," in New Advances in Intelligent Decision Technologies, ser. Studies in Computational Intelligence, K. Nakamatsu, G. Phillips-Wren, L. Jain, and R. Howlett, Eds. Springer Berlin Heidelberg, 2009, vol. 199, pp. 171–182.

[35] R. M. Haralick, K. S. Shanmugam, and I. Dinstein, "Textural features for image classification." IEEE Transactions on Systems, Man, and Cybernetics, vol. 3, no. 6, pp. 610–621, 1973.

[36] J. Comstock, An introduction to the study of botany: including a treatise on vegetable physiology, and descriptions of the most common plants in the middle and northern states, ser. Harvard science and math textbooks preservation microfilm project. Robinson, Pratt & Co., 1837.

[37] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Stochastic gradient descent training for l1-

regularized log-linear models with cumulative penalty," in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1, ser. ACL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 477–485.

[38] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang, "Large-scale image classification: fast feature extraction and svm train- ing," in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011, pp. 1689–1696.

[39] A. Bosch, A. Zisserman, and X. Mun˜oz, "Scene classification using a hybrid generative/discriminative approach," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 4, pp. 712–727, 2008.

[40] G. Amato and F. Falchi, "kNN based image classification relying on local feature similarity," in Proceedings of the Third International Conference on SImilarity Search and APplications. ACM, 2010, pp. 101–108.

[41] Y.-W. Chang, C.-J. Hsieh, K.-W. Chang, M. Ringgaard, and C.-J. Lin, "Training and testing low-degree polynomial data mappings via linear svm," Journal of Machine Learning Research, vol. 11, pp. 1471–1490, 2010.

[42] T. Hastie, R. Tibshirani, and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations. New York: Springer-Verlag, 2008.

[43] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern classification. John Wiley & Sons, 2012.

[44] A. B. Hassanat, M. A. Abbadi, G. A. Altarawneh, and A. A. Alhasanat, "Solving the problem of the k parameter in the knn classifier using an ensemble learning approach," arXiv preprint arXiv:1409.0919, 2014.

[45] M. Habshah, "A hybrid technique for selecting support vector regression parameters based on a practical selection method and grid search procedure," 2016.

[46] L. Jiming, "A new plant leaf classification method based on neigh- borhood rough set," Advances in information Sciences and Service Sciences(AISS), vol. 4, no. 1, pp. 116–123, 1 2012.

[47] S. G. Krishna Singh, Indra Gupta, "SVM-BDT PNN and fourier moment technique for classification of leaf shape," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 3, no. 4, pp. 67–78, December 2010.

[48] C.-L. Lee and S.-Y. Chen, "Classification of leaf images," International Journal of Imaging Systems and Technology, vol. 16, no. 1, pp. 15–23, 2006.

Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information). 10/1 (2017), 29-35
DOI: http://dx.doi.org/10.21609/jiki.v10i1.408

TERM WEIGHTING BASED ON POSITIVE IMPACT FACTOR QUERY FOR ARABIC FIQH DOCUMENT RANKING

Rizka W. Sholikah, Dhian S. Y. Kartika, Agus Zainal Arifin, and Diana Purwitasari

Informatics Department, Faculty of Information and Technology, Institut Teknologi Sepuluh Nopember, Jl. Raya ITS Kampus Sukolilo, Surabaya, 60111, Indonesia

E-mail: rizka.wakhidatus15@mhs.if.its.ac.id, dhian.satria15@mhs.if.its.ac.id, agusza@cs.its.ac.id, diana@if.its.ac.id

Abstract

Query becomes one of the most decisive factor on documents searching. A query contains several words, where one of them will become a key term. Key term is a word that has higher information and value than the others in query. It can be used in any kind of text documents, including Arabic Fiqh documents. Using key term in term weighting process could led to an improvement on result's relevancy. In Arabic Fiqh document searching, not using the proper method in term weighting will relieve important value of key term. In this paper, we propose a new term weighting method based on Positive Impact Factor Query (PIFQ) for Arabic Fiqh documents ranking. PIFQ calculated using key term's frequency on each category (mazhab) on Fiqh. The key term that frequently appear on a certain mazhab will get higher score on that mazhab, and vice versa. After PIFQ values are acquired, TF.IDF calculation will be done to each words. Then, PIFQ weight will be combine with the result from TF.IDF so that the new weight values for each words will be produced. Experimental result performed on a number of queries using 143 Arabic Fiqh documents show that the proposed method is better than traditional TF.IDF, with 77.9%, 83.1%, and 80.1% of precision, recall, and F-measure respectively.

Keywords: *Document Ranking, Arabic, Term Weighting, Query, PIFQ.*

Abstrak

Query menjadi salah satu faktor penentu dalam pencarian dokumen. Dalam sebuah *query* terdiri dari beberapa kata, dimana salah satunya menjadi *key term*. *Key term* adalah kata yang memiliki nilai informasi dan bobot lebih tinggi dibandingkan kata lain. Hal tersebut berlaku untuk semua jenis dokumen teks, termasuk dokumen fiqih berbahasa Arab. Penitik beratan pada *key term* dalam proses pembobotan kata memungkinkan terjadinya peningkatan relevansi pencarian. Di dalam pencarian dokumen fiqih berbahasa Arab, jika metode pembobotan kata yang digunakan tidak tepat, *key term* tidak akan memberikan pengaruh berarti. Oleh karena itu diusulkanlah sebuah metode pembobotan baru pada kata berbasis *Positive Impact Factor Query* (PIFQ) untuk perangkingan dokumen fiqih berbahasa arab. PIFQ dihitung menggunakan frekuensi kemunculan *key term* pada setiap kategori (mazhab) dalam fiqih. Semakin tinggi frekuensi *key term* tersebut pada suatu mazhab semakin tinggi pula nilainya pada mazhab tersebut, begitu pula sebaliknya. Setelah didapat nilai PIFQ, kemudian dilakukan perhitungan TF.IDF untuk setiap kata. Selanjutnya bobot PIFQ akan dikom-binasikan dengan TF.IDF sehingga menghasilkan bobot baru untuk masing-masing kata. Hasil dari pengujian yang dil-akukan pada sejumlah *query* dengan 143 dokumen fiqih berbahasa Arab menunjukan bahwa metode usulan dapat lebih unggul jika dibandingkan metode TF.IDF, dengan nilai *precision*, *recall*, dan *F-measure* masing-masing sebesar 77,9%, 83,1%, dan 80,1%.

Kata Kunci: *Perangkingan Dokumen, Bahasa Arab, Pembobotan Kata, Query, PIFQ.*

1. Introduction

Documents ranking is one of the research topics in information retrieval. One of its implementation is to sort the query results. The top result is considered as the most relevant according to query entered by user.

Many researches about documents ranking and sorting have been done before, such as N-gram method, to find relevant documents by matching the query and the document itself [1]. Another research used TF.IDF weighting and represent it to a Vector Space Model (VSM) [2]. TF (Term Frequency) is a method to set the weight of each term by calculating term's frequency in a document. While IDF (Inverse Document Frequency) consider that the fewer term appear in multiple documents, then the higher the weight of

that term. Those two methods combined and called as TF.IDF term weighting [3]. Basically, TF.IDF only counts terms occurrence and give positif discrimination to rare terms in a document; then Fauzi (2014) [4] proposed an improved method called IBF (Inverse Book Frequency) which consider the rare terms in a book. IBF uses the same principal method as IDF, but with difference scope. IBF also proved to have higher precission and recall than basic TF.IDF. Another research about term weighting in Arabic document also done by Khadijah (2015) [5]. The method assume that the relevancy of query and search results also depend on user's subjectivity, in this case called as preference. Therefore, they proposed IPFα method to accommodate such requirement. This method proved can be applied in documents search and has higher recall than others methods.

In addition to weighting methods, another documents categorizing researches using supervised method has also been developed by Emmanuel (2013) [6]. They proposed Positive Impact Factor (PIF) based on assumption that "positive impact from a certain feature on a certain category could be used to calculate its own negative impact to the other categories". The result shows that PIF can improve accuracy of documents categorizing while compared with the other existing methods.

In a query that contains several words, there must be a word that have higher information and value than the others. That word is called key term. Using key term in term weighting process could led to an improvement on result's relevancy. In Arabic Fiqh document searching, term weighting that use improper method will relieve important value from key term.

In this paper, we propose a new term weighting method based on Positive Impact Factor Query (PIFQ) for Arabic Fiqh documents ranking. This method takes note to key term on each query and will give higher weight to it rather than the others. This method is expected to improve documents relevancy compared to existing methods. It may also can be implemented to rank another language documents.

2. Methodology

In this study, we used Arabic Fiqh data set from Al-Mahtabah As-Shamela that can be download from http://shamela.ws/. The dataset consist of 4 mazhab, Hanafiyah, Malaikiyah, Syafi'iyah, and Hanabilah. Mazhab is kind of methodology that used by Islamic leader to explain the law in Islam so that everyone knows their principal, part, rules, and boundaries [7]. Each mazhab has their own base rule to explain and solve the problems. There are 143 documents that used in this paper, one document is represented by one page. Every documents are took from different books in different mazhab. Figure 1 shows every phases of proposed method include preprocessing.

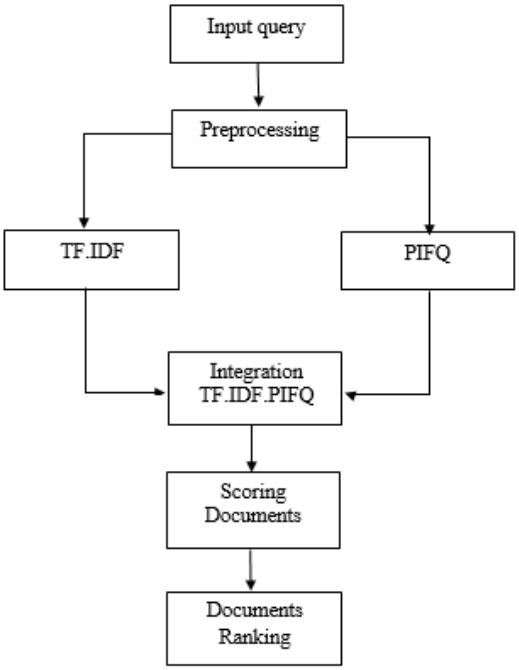

Figure 1. Phases in proposed method

Preprocessing

We used 7 different queries in this study. Both queries and documents are trained through the preprocessing phase. The first stage of preprocessing is tokenizing. Tokenizing was done to eliminate space, punctuation, and numbers so that the document will consist by a set of single word. In this phase, the vowels in every word are removed. The next step is to do stop-word removal by removing the words (terms) that has no valuable information. The terms that constantly appear in

TABLE 2.
KEY TERM FREQUENCY

Mazhab	Frequency
Hanafiyah	15
Malikiyah	10
Syafi'iyah	6
Hanabilah	7

TABLE 2.
KEY TERM PIFQ

Mazhab	PIFQ
Hanafiyah	1,218
Malikiyah	1,133
Syafi'iyah	1,075
Hanabilah	1,088

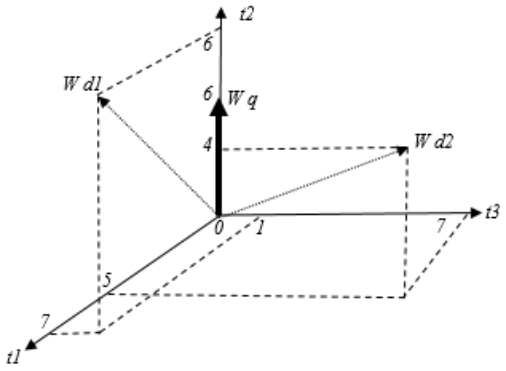

Figure 3. The illustration of VSM

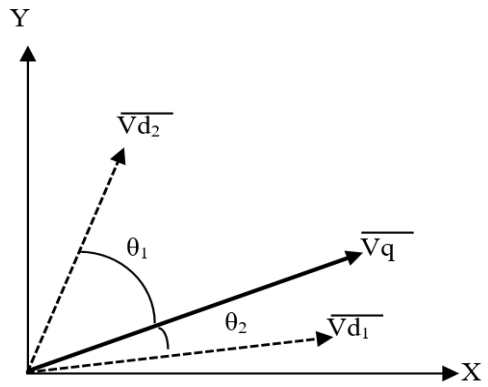

Figure 2. Cosine similarity illustration

every documents with high frequency can be categorizing as stop-word. In this paper, we used Arabic stop-word list that contain of 13.016 words. The stop-word list can be download from http://arabicstemmer.codeplex.com/.

After we got the valuable terms, then each term will going through normalization process. Normalization in Arabic text is important to be done because Arabic text has various way to write the same word. Normalization can be done by following this step [8, 9]: 1) convert term (إ), (أ), (ئ) ,(ؤ) ,(ء) into alif (ا); 2) convert ta marbutoh (ة) into ha (ه); 3) convert ya (ى) into ya (ي).

The last stage is stemming, that used to obtain the root in each words. In this study we used Light Stemmer [10]. Light stemmer is one of the method to find root in Arabic without using dictionary. This method get the root only by removing conjunction for example wa (و), some prefixes (ف, ك, ل, ب, ال, لل, قال, كال, بال, وال) and suffixes (ها, ون, وا, ين, ان, يه, ا, ه) [9]. To perform stemming and normalization in Arabic, we used library from Apache Lucene in Java that can be download from http://lucene.apache.org/.

TF.IDF

The common method for term weighting is TF.IDF. TF (Term Frequency) is one of the method to get weight of terms by calculating the frequency of terms in a document [3, 11]. IDF (Inverse Document Frequency) assume that each term which rarely appear on the multiple document in data set has higher value [3, 11]. TF for every terms t_i in documents d_j can be calculated using equation(1).

$$\boldsymbol{W_{TF}(t_i, d_j) = f(t_i, d_j)} \quad (1)$$

In a corpus consisting of D documents, there are $\boldsymbol{d_{(t_i)}}$ documents that contained terms t_i. The IDF can be calculated as defined in equation(2).

$$\boldsymbol{W_{IDF}(t_i) = 1 + log\left(\frac{D}{d_{(t_i)}}\right)} \quad (2)$$

TF.IDF weight calculation is done by multiplying the equation 1 with 2 resulting in equation(3).

$$\boldsymbol{W_{TF.IDF}(t_i, d_j) = f(t_i, d_j) \times \left(1 + log\left(\frac{D}{d_{(t_i)}}\right)\right)} \quad (3)$$

Proposed Method

PIFQ (Positive Impact Factor Query) is a modification of the PIF (Positive Impact Factor) method used to perform documents categorizing [6]. PIF using the occurrence of each terms in every category to get the weight for creating a classifier model. PIF method can be calculated by using equation(4).

$$\boldsymbol{W_{PIF} = log\left(\frac{F_{Mi}}{\sum_{j=1}^{n} F_{Mj}} + 1\right), j \neq i} \quad (4)$$

Based on PIF, this proposed method using the occurrence of key term in every category (mazhab) to improve the relevancy between query and search result. Key term itself is a word in a query that has higher value and information than the other words in that query. In this study, the first word in the query is considered as a key term. The idea of PIFQ is to increase the similarity between query vector and documents vector by increasing the weight of the documents using key term. PIFQ provides higher value in document that located in high key term frequency mazhab. It is based on assumption that those documents has higher probability as relevant documents than documents in other mazhab.

To calculate PIFQ, it will be seen whether the terms t_i is key term k or not. If t_i is key term k then the PIFQ will be calculated. If not, the value will be assign by 1. In a corpus of Arabic Fiqh documents which consists of four mazhab $M = \{M_1, M_2, M_3, M_4\}$, frequency of the key term that

TABLE 3.
CONFUSION METRICS

	Relevant	Not relevant
Retrieve	TP	FP
Unretrieve	TN	FN

occur in every mazhab F_{Mi} will be calculated. Then the weight of each terms in every documents *d* according to their mazhab M_i will also be calculated. PIFQ weight calculation can be seen in equation(5).

$$\boldsymbol{W_{PIFQ}(t_i)} = \begin{cases} \mathbf{1} + \boldsymbol{log}\left(\dfrac{\boldsymbol{F_{Mi}}}{\sum_{j=1}^{4}\boldsymbol{F_{Mj}}} + \mathbf{1}\right), \boldsymbol{j \neq i}, \boldsymbol{t_i = k} \\ \mathbf{1}, \boldsymbol{t_i \neq k} \end{cases} \quad (5)$$

For the query الجمعة صلاة في المسجد, since the first word is الجمعة, then that term is considered as the key term of the query. After we get the key term, the next step is to calculate the frequency of key term in every mazhab. Table 1 shows an example of the calculation of frequency.

After that, PIFQ value in every mazhab can be calculated using equation(5). The example of calculation if key term is located in mazhab Hanafiyah can be seen below.

$$\boldsymbol{W_{PIFQ}}(\text{الجمعة}) = \mathbf{1} + \boldsymbol{log}\left(\frac{\mathbf{15}}{\mathbf{10+6+7}} + \mathbf{1}\right) = \mathbf{1,218}$$

Table 2 shows the PIFQ in every mazhab. The key term value in every documents can be different depend on where the document is located. While, for the non-key term the value will be assign by 1.

The final step is combining TF.IDF with PIFQ by multiplying the weight of each word from TF.IDF and PIFQ calculation. The formula for determining the weight using the proposed method can be seen in equation(6).

From equation(6) we can see that if term t_i equals key term *k*, the term will get a higher weight than other terms. Thus, if the documents contain more key terms, the weight will be superior to other documents. By using the proposed method, the value of key term in the documents that belong to different madzhab will be different. Since this method use the density of key term in every madzhab by calculating the frequency, then if the document located at madzhab that have high frequency of key term, the weighting of key term will also be high, and vice versa.

TABLE 4.
LIST OF QUERY

ID Query	Query
Q1	الماء الذي ينجس والذي لا ينجس
Q2	المسخن الماء
Q3	الوضوء من النوم
Q4	الجنابة غسل
Q5	عورة الحرة
Q6	الصدقة دراهم
Q7	الزكاة ليست الأرز

$$\boldsymbol{W(t_i)} = \begin{cases} \boldsymbol{W_{TF.IDF}(t_i, d_j)} \times \left(\mathbf{1} + \boldsymbol{log}\left(\dfrac{\boldsymbol{F_{Mi}}}{\sum_{j=1}^{4}\boldsymbol{F_{Mj}}} + \mathbf{1}\right)\right), \boldsymbol{j \neq i}, \boldsymbol{t_i = k} \\ \boldsymbol{W_{TF.IDF}(t_i, d_j)} \times \mathbf{1}, \boldsymbol{t_i \neq k} \end{cases} \quad (6)$$

Vector Space Model (VSM)

After getting the weight for each terms, then the document will be represented in a vector space using VSM. VSM is a model used to measure the similarity between document and a query. Query and documents considered as vectors in n-dimensional space, where *k* is the sum of all terms in the lexicon. Lexicon is a list of all the terms in the index. After that, the cosine angle of the two vectors, namely *Wd* of each document and *Wq* of the query, will be calculated. VSM is usually used when the terms weighting is done with TF.IDF method. This is because TF.IDF method allows the similarity weight on the document. Every term *t* in a document described as one dimension, so that if there are three terms would form a 3 dimension. Figure 2 shows an illustration of the VSM.

Similarity Measurement

One of the method to measure the similarity between documents is cosine similarity [12]. This measure calculates the cosine of the angle between

TABLE 5.
EXPERIMENT RESULT

Query	IDF			TF.IDF			IDF.PIFQ			TFIDFPIFQ		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
Q1	43.8	50.0	46.7	56.3	68.8	61.9	43.8	50.0	46.7	**75.0**	**75.0**	**75.0**
Q2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Q3	**66.7**	**83.3**	**74.1**	58.3	75.0	65.6	**66.7**	75.0	70.6	50.0	66.7	57.1
Q4	50.0	**100.0**	66.7	60.0	90.0	72.0	60.0	**100.0**	75.0	**70.0**	90.0	**78.8**
Q5	**100.0**	100.0	100.0	85.7	100.0	92.3	**100.0**	100.0	100.0	**100.0**	100.0	100.0
Q6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Q7	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Average	72.9	83.3	76.8	72.9	**83.4**	77.4	74.3	82.1	77.5	**77.9**	83.1	**80.1**

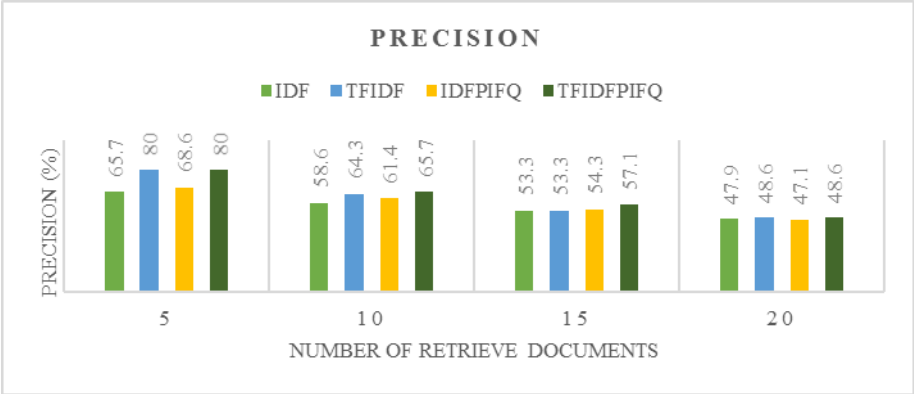

Figure 4. Precision for variation of displayed documents

the two vectors. The use of cosine similarity in text matching has angle limitation between 0◦ and 90◦. This is because in text matching the similarity of documents cannot be negative.

Suppose there are two document vectors d_1, d_2 and query vector q. Cosine similarity will calculate the value of θ of each document to the query q. For each word in the document which has a weight $W(t_i, d_j)$ and each word in a query that has a weight $W(t_i, q)$ calculation cosine similarity can be done by applying the formula in equation(7), while for an illustration of cosine similarity can be seen in Figure 3.

$$\cos(\boldsymbol{\theta}) = \frac{\boldsymbol{d} \,.\, \boldsymbol{q}}{\|\boldsymbol{d}\| \times \|\boldsymbol{q}\|} = \frac{\sum_{i=1}^{N} \boldsymbol{W(ti,q)} \,.\, \boldsymbol{W(ti,dj)}}{\sqrt{\sum_{i=1}^{Nq} |\boldsymbol{Wq}|^2} \,.\, \sqrt{\sum_{i=1}^{N} |\boldsymbol{Wd}|^2}} \quad (7)$$

The result of the similarity between documents and query by using cosine similarity will yield a value between 0 and 1. 0 indicates the documents and the query has absolutely nothing in common and 1 indicates that the document and the query is identical.

3. Result and Analysis

The proposed method is evaluated by using precision, recall, and F-measure. Precision, recall, and F-measure are commonly used to evaluate performance in Information Retrieval (IR). The experiment in other methods [4, 5, 6] also using precision, recall, and F-measure to evaluate the method's performance. Based on Table 3 precision, recall, and F-measure are defined as Equation 7, 8, and 9 each.

Precision can be described as the number of documents relevant (*tp*) from the total of document that have been retrieve (*tp* + *fp*). Precision is used to measure the effectiveness of IR systems (equation 8). Recall is used to measure relevancy of the system (equation 9). This approach can be calculated as the number of doc-ument relevant and retrieve (*tp*) from total number of relevant documents in collections (*tp* + *tn*). The F-measured approach can be created by combining precision and recall as shown in equation(10).

$$\boldsymbol{precision} = \frac{\boldsymbol{tp}}{\boldsymbol{tp} + \boldsymbol{fp}} \quad (8)$$

$$\boldsymbol{recall} = \frac{\boldsymbol{tp}}{\boldsymbol{tp} + \boldsymbol{tn}} \quad (9)$$

$$\boldsymbol{Fmeasure} = \frac{\boldsymbol{2 \times recall \times precision}}{\boldsymbol{precision + recall}} \quad (10)$$

We used 7 queries to test our proposed method. The queries can be seen in Table 4. Precision, recall, and F-measure will be performed for each query.

In this experiment we used TF.IDF.PIFQ as term weighting. Then we will compare the result of proposed method with IDF, IDF.PIFQ, and TF.IDF, so that the relevancy of proposed method can be observed. Result of the experiment can be seen in Table 5. From the table we know that our proposed method has higher precision values than IDF and IDF.PIFQ in Q1 with 75.0% and Q4 with 70.0%. While compared with TF.IDF, the proposed method has higher precision for Q1, Q4, and Q5, with 75.0%, 70.0%, and 100.0% respectively. Only in Q3 the precision value of proposed method is lower than the three other methods.

The recall been calculated by retrieving 20 documents. The result using IDF, IDF.PIFQ, TF.IDF, and TF.IDF.PIFQ produce the same value at 4 query, i.e. Q2, Q5, Q6 and Q7. In Q1 recall of the proposed method has higher value than the three other methods, with 75.0%. Whereas, in Q3

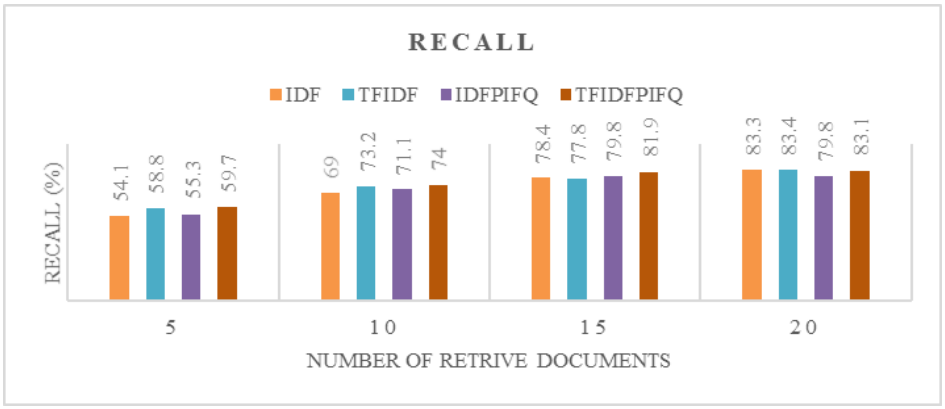

Figure 5. Recall for variation of displayed documents

recall value using IDF method is the highest, with 83.3%. The similarity result in some query caused by the four methods retrieve the same number of relevance documents, but has the different order. The TF.IDF.PIFQ method placed the documents that has high frequency of key term above the lower ones.

The average results showed that the proposed method, TF.IDF.PIFQ has higher value in precision and F-measure, with 77.9% and 80.1%. But the average of recall value is smaller than TF.IDF, the proposed method with 83.1% and TF.IDF with 83,4%.

The experiment also conducted by calculate precision and recall with variation of the number of documents retrieved to user. The first experiment displayed 5 documents relevant, then 10 documents, 15 documents, and 20 documents. For each variety will be calculated the precision and recall for IDF, IDF.PIFQ, TF.IDF and TF.IDF.PIFQ method. The experiment result can be seen in Figure 4 and 5. From Figure 4 we know that the more document that displayed to user, the precision value produced by all methods will decrease. When system retrieve 10 and 15 documents that displayed to user, TF.IDF.PIFQ method got higher precision than other methods, with 65.7% and 57.1% respectively. Whereas, when retrieve 5 and 20 documents the proposed method has the same value with TF.IDF, 80% and 48.6% each.

The result of recall can be seen in Figure 5. For all methods the more documents that displayed to user the result value will increase. The result indicates that our proposed method has higher recall than the others method, with 59.7% for 5 documents, 74% for 10 documents and 81.9% for 15 documents. Only in variety number of documents 20, TF.IDF.PIFQ has 83.1% lower than TF.IDF with 83.4% and IDF 83.3%.

Figure 4 and 5 also shows that adding the PIFQ in IDF method can give positive result in precision and recall. The precision of IDF.PIFQ when retrieve 5, 10 and 15 documents has higher value than IDF method. The recall also show that in 5, 10, and 15 documents using IDF.PIFQ has higher recall than only using IDF. Therefore, we can conclude that adding PIFQ can increase the result of IDF method by considerig the density of key term in every mazhab.

In TF.IDF.PIFQ, high key term frequency mazhab has higher probability contain relevant documents than other mazhab. That condition affects the search result, and can increase the precision and recall. But if there are documents relevant that located in low key term frequency mazhab, the order will be lower than non-relevant documents that contain key term in high key term frequency mazhab. This factor can decrease the precision and recall of the method.

In some queries Q2 and Q6, the query's terms has high occurrence in relevant documents, but does not or just rarely occurre in other documents. So in this condition, implementing all method will led to the same result.

Overall, based on the experiment result by comparing method using precision, recall, and F-measure, we can conclude that TF.IDF.PIFQ method is superior when compared to the method IDF, IDF.PIFQ and TF.IDF. It also prove that the proposed method can be used to improve the relevance of searching result in Arabic Fiqh Documents. However, this method has weakness when the distribution of key terms in each mazhab are equals. So that PIFQ value for key term in every mazhab will be the same. This will reduce the influence of the importance of key terms in the search process.

4. Conclusion

This paper shows a new terms weighting method which observe the impact of key term density on each mazhab. This new method

combines PIFQ and TF.IDF to calculate the weight of each terms. This new method proved to be able to be implemented on terms weighting on Arabic Fiqh document ranking. It can be seen on the experimental result which shows that TF.IDF.PIFQ method has better precision and F-measure than IDF, IDF.PIFQ and TF.IDF, which are 77.9% and 80.1% respectively. Moreover, using various number of documents displayed to the user, TF.IDF.PIFQ is able to get higher precision and recall value compared to the other three methods by the variation of 10 and 15 documents. The experiment also shows that adding PIFQ can increase precision and recall of IDF method. The average result of precision and F-measure using IDF.PIFQ is higher than using IDF, with 74.3% and 77.5% respectively. Besides used on Arabic documents, this method could also be used on documents with other languages.

References

[1] S. H. Mustafa and Q. A. Al-Radaideh, "Using N-Grams for Arabic Text Searching," *Journal of The American Society for Information Science and Technology*, pp. 1002-1007, 2004.

[2] F. Harrag, A. Hamdi-Cherif and E. El-Qawasmeh, "Vector Space Model for Arabic Information Retrieval - Application to Hadith Indexing," *Proceedings of the First IEEE Conference on the Application of Digital Information and Web Technologies*, pp. 107-112, 2008.

[3] G. Salton and B. Christopher, "Term-weighting approach in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513-523, 1988.

[4] M. A. Fauzi, A. Z. Arifin and A. Yuniarti, "Term Weighting Berbasis Indeks Buku dan Kelas untuk Perangkingan Dokumen Berbahasa Arab," *Lontar Komputer*, vol. 5, no. 2, 2014.

[5] K. F. H. Holle, A. Z. Arifin and D. Purwitasari, "Preference based term weighting for arabic fiqh document ranking," *Journal of Computer Science and Information*, vol. 8, no. 1, 2015.

[6] E. M, S. M. Khatri and R. B. D.R., "A Novel Scheme for Term Weighting in Text Categorization : Positive Impact Factor," *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2292-2297, 2013.

[7] "Mazhab," Wikimedia, [Online]. Available: http://id.wikipedia.org/wiki/Mazhab. [Retrieved December 25, 2015].

[8] A. Odeh, A. Abu-Errub, Q. Shambour and N. Turab, "Arabic Text Categorization Algorithm using Vector Evaluation Method," *International Journal of Computer Science & Information Technology (IJCSIT)*, vol. 6, no. 6, 2014.

[9] H. K. H. Chantar, "New Techniques for Arabic Documents Classification", Ph.D Thesis, School of Mathematical and Computer Science, Heriot-Watt University, 2013.

[10] L. S. Larkey, L. Ballesteros and M. E. Connell, "Light Stemming for Arabic Information Retrieval," *Arabic computational morphology Springer*, pp. 221-243, 2007.

[11] G. Salton, Automatic text processing : The Transformation, analysis, and Retrieval of Information by Computer, Addison-Wesley, 1989.

[12] S. Tata and J. M. Patel, "Estimating the selectivity of tf-idf based cosine similarity predicates," *SIGMOD Record*, vol. 36, no. 4, pp. 75-80, 2007.

Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information). 10/1 (2017), 36-42
DOI: http://dx.doi.org/10.21609/jiki.v10i1.410

RBF KERNEL OPTIMIZATION METHOD WITH PARTICLE SWARM OPTIMIZATION ON SVM USING THE ANALYSIS OF INPUT DATA'S MOVEMENT

Rarasmaya Indraswari, Agus Zainal Arifin, Darlis Herumurti

Department of Informatics, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember (ITS), Kampus ITS Sukolilo, Surabaya, 60111, Indonesia

E-mail: rarasmaya15@mhs.if.its.ac.id

Abstract

SVM (Support Vector Machine) with RBF (Radial Basis Function) kernel is a frequently used classification method because usually it provides accurate results. The focus of most SVM optimization research is the optimization of the input data, whereas the parameter of the kernel function (RBF), the sigma, which is used in SVM also has the potential to improve the performance of SVM when optimized. In this research, we proposed a new method of RBF kernel optimization with Particle Swarm Optimization (PSO) on SVM using the analysis of input data's movement. This method performed the optimization of the weight of the input data and RBF kernel's parameter at once based on the analysis of the movement of the input data which was separated from the process of determining the margin on SVM. The steps of this method were the parameter initialization, optimal particle search, kernel's parameter computation, and classification with SVM. In the optimal particle's search, the cost of each particle was computed using RBF function. The value of kernel's parameter was computed based on the particle's movement in PSO. Experimental result on Breast Cancer Wisconsin (Original) dataset showed that this RBF kernel optimization method could improve the accuracy of SVM significantly. This method of RBF kernel optimization had a lower complexity compared to another SVM optimization methods that resulted in a faster running time.

Keywords: *parameter, Particle Swarm Optimization, RBF kernel, sigma, Support Vector Machine*

Abstrak

Metode klasifikasi SVM (*Support Vector Machine*) dengan RBF (*Radial Basis Function*) kernel merupakan metode yang sering digunakan karena memberikan hasil klasifikasi yang cukup akurat. Penelitian mengenai optimasi pada SVM sementara ini masih banyak berfokus pada optimasi dari nilai data masukan padahal parameter fungsi kernel (RBF), yaitu parameter *sigma*, yang digunakan pada SVM juga memiliki potensi untuk meningkatkan performa dari SVM apabila dioptimasi. Pada penelitian ini diajukan metode baru optimasi RBF kernel dengan *Particle Swarm Optimization* (PSO) pada SVM berdasar analisis persebaran data masukan. Metode ini melakukan optimasi terhadap bobot data masukan sekaligus parameter RBF kernel berdasarkan analisis persebaran data masukan sehingga terpisah dari proses penentuan margin pada SVM. Tahapan dari metode ini adalah inisialisasi parameter, pencarian partikel optimal, perhitungan nilai parameter kernel, dan klasifikasi dengan SVM. Pada proses pencarian partikel optimal, nilai *cost* dari tiap partikel dihitung berdasar fungsi RBF. Nilai parameter kernel dihitung berdasar pergerakan partikel data masukan pada PSO. Hasil uji coba pada dataset *Breast Cancer* Wisconsin (Original) menunjukkan bahwa metode optimasi RBF kernel mampu meningkatkan akurasi klasifikasi SVM secara cukup signifikan. Metode optimasi parameter RBF kernel ini memiliki kompleksitas yang lebih rendah dibandingkan dengan metode optimasi SVM lainnya sehingga menghasilkan *running time* yang lebih cepat.

Kata Kunci: *parameter, Particle Swarm Optimization, RBF kernel, sigma, Support Vector Machine*

1. Introduction

SVM (Support Vector Machine) classification is a method proposed by Boser, Guyon, and Vapnik in 1992 and often used in various fields such as pattern recognition, bioinformatics, and text categorization [1]. One of the SVM method that is commonly used is SVM with RBF kernel. A lot of research about SVM with RBF kernel optimization used an optimization methods such as Particle Swarm Optimization (PSO) and Genetic Algorithm (GA). In general, the focus of SVM optimization that was performed is divided in two types, namely the data optimization and parameter optimization.

On the data optimization type, optimization methods are used to determine the weight of the

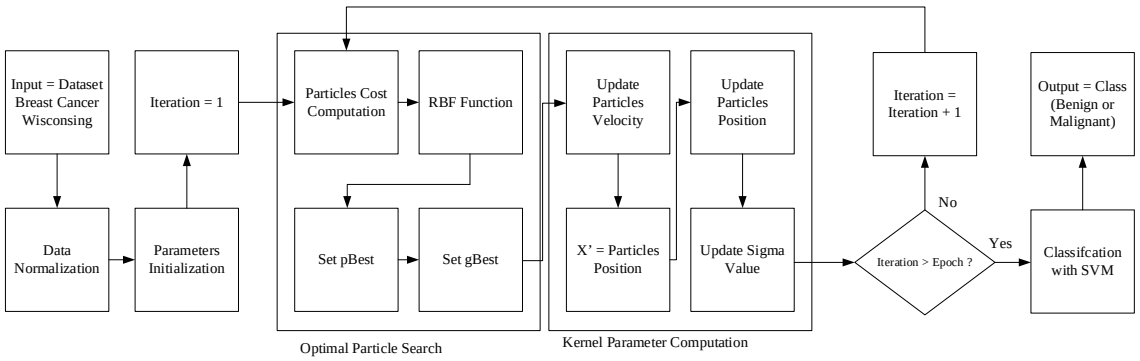

Figure 1. The steps of RBF kernel optimization method with Particle Swarm Optimization on SVM using the analysis of input data's movement.

input data to provide a more optimal result of the SVM classification. In 2014, Devos et. al was using a Genetic Algorithm to perform the optimization of input data which was performed on olive oils data [2]. Fitness value that was used is the error of the k-fold cross validation applied to the SVM.

On the parameter optimization type, researches that were conducted usually about selecting the combination of SVM margin parameter (*C*) with a kernel parameter (*sigma*) that can provide the most optimal result of SVM classification. So that on the usual optimization process, the input used the combination of SVM margin parameter and kernel parameter, and the output is the accuracy of SVM classification with k-fold cross validation method. Grid algorithm is an alternative method to find the optimal combination of parameter *C* and *sigma* on SVM with RBF kernel. However, this method required a long computational time and sometimes it did not go well [3] [4]. On of the research about the optimization of parameter *C* and *sigma* on SVM with RBF kernel was done by Ding and Li in 2009 which showed that the SVM parameter optimization with PSO can improve the accuracy of SVM classification result [5].

Those two types of SVM optimization should have been combined so that in one optimization process, the data optimization and the SVM parameter optimization could be conducted simultaneously. In 2006, Cheng-Lung Huang and Chieh-Jen Wang proposed an SVM parameter *C* and *sigma* optimization method using Genetic Algorithm. This method also performed the input data optimization, which is the selection of the features of the dataset using GA [1]. Features subset from the dataset became the part of the GA chromosom together with the values of parameter *C* and *sigma*. In 2008, Lin et al. optimized the parameters of SVM and did feature selection using PSO [6]. However, the method proposed in both of those researchs called the SVM classification process in many times that resulted on a high complexity.

In this study we proposed a new method of RBF kernel optimization with Particle Swarm Optimization (PSO) on SVM using the analysis of input data's movement. This method performed the optimization of the input data values and the RBF kernel parameter (*sigma*) at once based on the analysis of the input data's movement so that it was separated from the process of determining the margin on SVM. Classification process with SVM was called once after the optimal value of the data and the kernel's parameter was obtained so that the complexity of this RBF kernel optimization method is not increased too much compared with the usual SVM with RBF kernel method. PSO was selected to do the optimization process because PSO method is represent a swarm of data group, while the RBF kernel shape is circular so the PSO is suitable to be applied to the RBF kernel optimization.

Particle Swarm Optimization

PSO was introduced by Kennedy and Eberhart to imitate social behavior of animals such as birds flocking in searching for food [7]. Each particle flies in hyperspace searching for the best solution by adjusting position and velocity based on its own flying experience (*pBest*) and its companions' experience (*gBest*). Each particle has a fitness value or cost which was evaluated using the fitness function to be optimal, position, and velocity that controls the movement of the particles. The inertia weight *w* was later introduced to improve the PSO optimizer [8]. The steps of PSO are as follow :

1) Initialization of PSO parameters.
2) Compute the cost of each particle using the fitness function.
3) Search the *pBest* dan *gBest* values.
4) Update the particle velocity v_i using equation(1)

$$v_i(t) = wv_i(t-1) + c_1 r_1 \left(x_{pBest_i}(t) - x_i(t)\right) + c_2 r_2 \left(x_{gBest}(t) - x_i(t)\right) \quad (1)$$

where the c_1 and c_2 is a variable value namely *correction factor*, r_1 and r_2 is a random variable whice value is between 0 and 1.

5) Update the particle position x_i using equation (2).

$$x_i(t+1) = x_i(t) + v_i(t) \quad (2)$$

RBF Kernel

Kernel function is used to change the manufacturing process SVM models that are linear to non-linear computing without overly complicated. RBF kernel is the kernel that can generally be used for all types of data. It uses a Gaussian kernel function RBF to get the inner product of x and x' using equation (3)

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \quad (3)$$

where $\|x - x'\|$ is the Euclidean Distance of the data values in two different feature space and σ (*sigma*) is a free parameter on RBF kernel which determine the weight of the kernel. In SVM, parameter σ need to be adjusted to provide a more accurate classification result. The default value of σ is 1. In RBF function, we can also used parameter *gamma* which value is $\gamma = \frac{1}{2\sigma^2}$.

2. Methods

RBF kernel optimization methods with Particle Swarm Optimization (PSO) on SVM using the analysis of input data's movement has four main steps, which are parameters initialization, optimal particle search, kernel's parameter calculation, and classification with SVM. The details of the stages of this method is shown in Figure 1.

Parameters Initialization

In this step, we initialize the PSO parameters which are inertia *w*, correction factor *c*, and the maximal number of iteration (*epoch*). In this research, the value of those parametes are *w*=0.4, *c*=0.7, dan *epoch*=100 which are the optimal value based on the experiment that had been conducted. We also set the initial velocity *v*=0.

In this step, we also set the initial value of RBF kernel parameter that will be optimized, σ, as 0.01. The value of data matrix *x* and the value of the data in the new feature space *x'* is initialized to be equal with the value of the input data. The value of *x'* is the particles position in PSO. We also performed the *pBest* value initialization of each particle, where the initial *pBest* should be a big number because the initial *pBest* value must be substituted by the cost that obtained in the first iteration.

Optimal Particle Search

After the parameters were initialized, we calculate the cost of each particle. Fitness function that is used is the RBF kernel function in equation (3) with the value of *x* is always fixed, the value of *x'* is the particle position in PSO, and the σ value is the value of the parameter σ at the current iteration.

After obtained the cost of each particle, we search the new local optimal value (*pBest*) from each particle *i*. The new *pBest* value ($pBest(t)$) was obtained by comparing the value of the previous *pBest* ($pBest(t-1)$) with the cost of the particle at iteration (t). To get the value of the new *pBest* we use equation (4).

$$pBest(t)_i = \min(pBest(t-1)_i, cost_i) \quad (4)$$

From the all *pBest* values that have been obtained, we search the minimum value of *pBest*. This minimum *pBest* value is the value of *gBest*

TABLE 1
EXPERIMENTAL RESULT ON INERTIA VALUE

Epoch	Optimized Kernel Parameter σ	Accuracy (%)
0.1	0.475	94.3
0.2	0.558	92.9
0.3	0.669	92.9
0.4	0.809	95.0
0.5	0.992	94.3

TABLE 2
EXPERIMENTAL RESULT ON CORRECTION FACTOR VALUE

Correction Factor	Optimized Kernel Parameter σ	Accuracy (%)
0.1	0.074	64.3
0.3	0.256	85.0
0.5	0.520	94.3
0.7	0.809	95.0
0.9	1.073	94.3

TABLE 3
EXPERIMENTAL RESULT ON EPOCH VALUE

Epoch	Optimized Kernel Parameter σ	Accuracy (%)
50	0.311	90.0
100	0.809	95.0
150	1.250	94.3
200	1.596	93.6
250	1.785	94.3

TABLE 4
COMPARISON WITH OTHER METHODS

Fold	RBF Kernel Optimization Method		SVM (sigma = 1)		SVM Parameters Optimization [5]	
	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)
2	71.75	0.970	65.35	0.110	69.40	10.490
5	71.74	0.796	67.66	0.076	71.70	14.542
10	71.16	0.720	65.18	0.063	73.27	15.513
Average	**71.55**	**0.830**	**66.06**	**0.080**	**71.46**	**13.520**

TABLE 5
ANALYSIS OF THE EFFECT OF DATA OPTIMIZATION PROCESS ON NUMBER OF FOLD = 2

K-	Optimized σ Value	Accuracy of RBF Kernel Optimization Method (%)	Accuracy of SVM with Optimized σ Value (%)
1	0.470	71.40	71.40
2	1.755	70.50	68.90
Average		**70.95**	**70.15**

(global optimal value). In this RBF kernel optimization method, the fitness function that used is the RBF function so that the cost of each particle that obtained in accordance with the data position to the kernel. The minimum value of the RBF kernel has been chosen as *gBest* because the smaller value of $K(x, x')$, the greater the distance between the initial value of the data and the value of the data in the new feature space so that the swarm of the data in the new dimension is moving by following the particle which move furthest.

Kernel's Parameter Computation

Having obtained the value of pBest and gBest, we calculate the velocity of each particle. In PSO, we use equation (1) to calculate the velocity. However, based on the experiment that had been conducted, if equation (1) is applied directly to this RBF kernel optimization method, then the given accuracy of the SVM classification results will become unstable because of the influence of the random variables r_1 and r_2. Therefore, the RBF kernel optimization method is using equation (5), which is a modification of equation (1), to calculate the value of velocity at which the value of variable c_1 and c_2 in equation (1) are combined into a single variable called the correction factor *c*. Besides, the random variable r_1 and r_2 in equation (1) is set to be the same value, namely *r*.

$$v_i(t) = wv_i(t-1) + cr\left(\left(x_{pBest_i}(t) - x_i(t)\right) + \left(x_{gBest}(t) - x_i(t)\right)\right) \quad (5)$$

After we get the velocity value of each particle, the position of each particle is updated using equation (2). The new position of the particle is the value of data *x'* in the new feature space. The position value of the particle is updated based on the velocity of the particle. Therefore, changes in the σ value is also updated based on the average velocity of the particles using equation (6) where *n* is the number of the particles.

$$\sigma(t) = \sigma(t-1) + \frac{1}{n}\sum_{i=1}^{n} v_i(t) \quad (6)$$

After the position of each particle is updated and the new value of parameter σ is obtained, we do optimal particle search using the new value of the particle's position and σ. This process is done repeatedly until the number of iteration reaches the epoch value.

Classification With SVM

After the epoch value is reached, the matrix of the mapping result from RBF kernel $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}')$ on the last iteration is used to perform the classification process using SVM as described by Cortes and Vapnik in their research in 1995 [9]. Because the kernel used is RBF, then the *hyperplane* that used is circular. The optimal margin value $\left(\frac{\mathbf{1}}{\|\boldsymbol{w}\|}\right)$ calculation using equation (7) is done to determine the support vectors, where the parameter *C* is the parameter of soft margin SVM and $\boldsymbol{\xi}$ is the margin error. Output of this process is a SVM model that will be used to classify the data test.

$$minimize\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i\right) \quad (7)$$

3. Results and Analysis

The dataset used in this study is the Breast Cancer Wisconsin (Original) dataset from UCI machine learning repository [10]. This dataset consists of 699 data with 9 numeric attributes which are divided into two classes : benign and malignant.

TABLE 6
ANALYSIS OF THE EFFECT OF DATA OPTIMIZATION PROCESS ON NUMBER OF FOLD = 5

K-	Optimized σ Value	Accuracy of RBF Kernel Optimization Method (%)	Accuracy of SVM with Optimized σ Value (%)
1	0.539	72.00	72.00
2	0.617	72.00	72.00
3	0.675	72.00	72.00
4	0.579	72.00	72.00
5	0.646	70.80	64.60
Average		**71.76**	**70.52**

The experiment conducted is the experiment to determine the optimal parameters of PSO and the comparison of the accuracy between the RBF kernel optimization method with the existing SVM optimization method. At the experiment on determining the optimal parameters, there are three parameters were tested, which are the inertia parameter, correction factor, and the epoch value.

In the comparison of the accuracy of the method, there are two experiments that were conducted. The first experiment compared the classification accuracy of RBF kernel optimization method proposed, the SVM method with default *sigma* value (which is 1), and the methods of parameters *C* and *sigma* on SVM optimization as proposed in [5]. The second experiment comparing the results of the classification accuracy between RBF kernel optimization method with SVM methods without optimization using *sigma* value which has been optimized with PSO to analyze the influence of the movement of the input data to increase the accuracy of RBF kernel optimization method that was proposed. In these two experiments, we used k-fold cross validation with many fold is 2, 5, and 10 pieces to perform the separation between the training and the test data.

Inertia

Inertia parameter is the percentage of the velocity of the particle at the previous iteration that will be used in the current iteration. The greater the inertia, the greater the velocity of the particles in a iteration, likely. Table 1 shows the results of the experiment on determining the optimal value of inertia that was measured from the accuracy that is achieved. It appears from Table 1 that the optimal accuracy is obtained by the inertia value = 0.4.

Correction Factor

Correction factor is the parameter that determines the magnitude of the effects of optimal swarm particle (*pBest* and *gBest*) value on the particle's velocity. Table 2 shows the results of the experiment on determining the optimal value of correction factor that was measured from the accuracy that is achieved. It appears from Table 2 that the optimal accuracy is obtained by the correction factor value = 0.7.

Epoch

Epoch is the limit of iterations performed while running the PSO algorithm. The greater the value of the epoch, could be that the position of the particles in the swarm is getting closer and centralized, but could also further apart when the central point was missed because the value of epoch is too big. The greater the value of another parameters (inertia and correction factor) the less epoch value needed to reach the central point. Table 3 shows the results of the experiment on determining the optimal value of epoch that was measured from the accuracy that is achieved. It appears from Table 3 that the optimal accuracy is obtained by the epoch value = 100.

Comparison With Another Methods

We conducted an experiment to determine whether the RBF kernel optimization method can provide the optimal value of parameter *sigma*. The optimality of the parameter sigma is measured from the given accuracy of the classification. The accuracy is compared with the usual SVM with RBF kernel using the default sigma value, which is 1. We also made comparison between RBF kernel optimization method with SVM parameter optimization method as in [5]. This experiment is done using k-fold cross validation with the number of folds are 2, 5, and 10 as shown in Table 4.

In Table 4, it appears that the method proposed in this research, RBF kernel optimization method, have a higher average accuracy than the usual SVM with RBF kernel method and SVM parameter optimization method. Seen in Table 4 that the bigger the value of the fold, which means the amount of training data is also bigger, the accuracy of RBF kernel optimization method is decreased. This could have happened if the input data lying scattered so that the value of the sigma parameter obtained is not optimal, and not good as the value obtained by SVM parameter optimization method. Seen in Table 4 that the accuracy obtained RBF kernel optimization method is quite stable and

TABLE 7
ANALYSIS OF THE EFFECT OF DATA OPTIMIZATION PROCESS ON NUMBER OF FOLD = 10

K-	Optimized σ Value	Accuracy of RBF Kernel Optimization Method (%)	Accuracy of SVM with Optimized σ Value (%)
1	0.160	69.20	69.20
2	0.227	69.20	69.20
3	0.224	69.20	69.20
4	0.239	69.20	69.20
5	0.233	69.20	69.20
6	0.216	75.00	75.00
7	0.246	75.00	75.00
8	0.210	75.00	75.00
9	0.219	75.00	75.00
10	0.213	72.70	72.70
Average		**71.90**	**71.90**

better than the SVM method without optimization.

In terms of complexity, RBF kernel optimization method has a lower complexity than SVM parameter optimization method. In general, the complexity of SVM with RBF kernel is *O (nd)* [11]. In SVM parameter optimization with PSO method, SVM is executed repeatedly according to the value epoch (*i*) and the number of particles PSO (*p*) so that the complexity of this SVM parameter optimization method is *O (ipnd)*. In the RBF kernel optimization method, although the function of RBF also executed many times according to the number of PSO epoch, the process of finding the optimal margin value on SVM is only run once so that the complexity of RBF kernel optimization method is lower than the complexity of SVM parameter optimization method. This is proven by the value of the running time of both methods that shown in Table 4.

Analysis of the Effect of Data Optimization Process

We conduct an experiment to analyze the effect of the PSO on the movement of the input data to increase the accuracy of the RBF kernel optimization methods. We do a comparison between RBF kernel optimization method with the usual SVM with RBF kernel method in which the *sigma* values used in the usual SVM method is the optimal *sigma* parameters obtained from RBF kernel optimization method. The experiment is done using the number of folds are 2, 5, and 10 that is shown in Table 5, Table 6 and Table 7 respectively. Average accuracy obtained RBF kernel optimization methods and SVM with optimal parameters *sigma* at this experiment is 71.54% and 70.86% respectively.

From Table 5, Table 6 and Table 7 it appears that the accuracy obtained with the RBF kernel optimization method and the usual SVM method that uses optimized *sigma* value does not vary much. Therefore it can be concluded that the data optimization process on RBF kernel optimization method can improve the accuracy of RBF kernel optimization method, but the effect is not significant.

In the RBF kernel optimization method, the process of determining the optimal *sigma* value have a more significant influence on the improvement of SVM classification accuracy than the input data optimization process. In addition, the value of the optimal parameter *sigma* obtained from the analysis of the movement of the input data can also improve the classification accuracy of usual SVM significantly as shown in Table 4 in which the average of the results of the classification accuracy of usual SVM without optimization of only 66.06%, while the average accuracy results of usual SVM classification which uses optimal parameter sigma is 70.86%.

4. Conclusion

RBF Kernel optimization with PSO (Particle Swarm Optimization) can improve the accuracy of SVM (Support Vector Machine) classification method quite significantly. According to the experimental results using the Breast Cancer Wisconsin (Original) dataset, RBF kernel optimization methods gives the average accuracy of 71.55% while the SVM method without optimization gives the average accuracy of 66.06%. In addition, the RBF kernel optimization methods also provide accuracy results which were not much different from the SVM parameter optimization method with PSO. The advantages of RBF kernel optimization method lies in its lower complexity so that the running time of RBF kernel optimization method is faster than SVM parameter optimization methods.

Particle Swarm Optimization (PSO) which is used in the optimization method RBF kernel affects two things, which are the movement of the input data and the movement of the parameter *sigma* in RBF kernel. According to experiment, it can be concluded that the movement of the input data increases the accuracy of RBF kernel optimization method, but the effect is not significant. While the optimal parameter *sigma* values obtained from the

analysis of the movement of the input data has a significant influence on the accuracy of the classification results when applied to either RBF kernel optimization method and the usual SVM with RBF kernel.

Determination of the proper parameters of PSO reasonably affect the accuracy of the classification of RBF kernel optimization method. So that a further research about the methods for determining the value of PSO parameters automatically based on the analysis of the input data can be done.

References

[1] C.-L. Huang and C.-J. Wang, "A GA-based feature selection and parameters optimization for support vector machines," *Expert System with Applications*, vol. 31, pp. 231-240, 2006.

[2] O. Devos, G. Downey and L. Duponchel, "Simultaneous data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils," *Food Chemistry*, vol. 148, pp. 124-130, 2014.

[3] C. W. Hsu and C. J. Lin, "A simple decomposition method for support vector machine," *Machine Learning*, vol. 46, pp. 219-314, 2002.

[4] S. M. LaVelle and M. S. Branicky, "On the relationship between classical grid search and probabilistic roadmaps," *International Journal of Robotics Research*, vol. 97, pp. 673-692, 2002.

[5] S. Ding and S. Li, "PSO Parameters Optimization Based Support Vector Machines for Hyperspectral Classification," in *International Conference on Information Science and Engineering (ICISE)*, 2009.

[6] S.-W. Lin, K.-C. Ying, S.-C. Chen and Z.-J. Lee, "Particle swarm optimization for parameter determination and feature selection of support vector machines," *Expert Systems with Application*, vol. 35, pp. 1817-1824, 2008.

[7] J. Kennedy and R. Eberhart, "Particle swarm optimization," *Proceedings of IEEE International Conference on Neural Networks*, vol. 4, pp. 1942-1948, 1995.

[8] Y. Shi and R. Eberhart, "A modified particle swarm optimization," *Proceedings of the IEEE Congress on Evolutionary Computation*, pp. 69-73, 1998.

[9] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, pp. 273-297, 1995.

[10] C. Blake and C. Merz, "UCI repository of machine learning databases," 1998. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html.

[11] H. Cao, T. Naito and Y. Ninomiya, "Approximate RBF Kernel SVM and Its Applications in Pedestrian Classification," in *The 1st International Workshop on Machine Learning for Vision-based Motion Analysis - MLVMA*, Marseille, France, Oct 2008.

Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information). 10/1 (2017), 43-49
DOI: http://dx.doi.org/10.21609/jiki.v10i1.428

LEAST SQUARES SUPPORT VECTOR MACHINES PARAMETER OPTIMIZATION BASED ON IMPROVED ANT COLONY ALGORITHM FOR HEPATITIS DIAGNOSIS

Nursuci Putri Husain, Nursanti Novi Arisa, Putri Nur Rahayu, Agus Zainal Arifin, and Darlis Herumurti

Department of Informatics, Faculty of Information Technology,
Institut Teknologi Sepuluh Nopember (ITS)
Kampus ITS, Surabaya, 60111.

E-mail: nursuci.husain15@mhs.if.its.ac.id, nursanti15@mhs.if.its.ac.id, putri15@mhs.if.its.ac.id

Abstract

Many kinds of classification method are able to diagnose a patient who suffered Hepatitis disease. One of classification methods that can be used was Least Squares Support Vector Machines (LSSVM). There are two parameters that very influence to improve the classification accuracy on LSSVM, they are kernel parameter and regularization parameter. Determining the optimal parameters must be considered to obtain a high classification accuracy on LSSVM. This paper proposed an optimization method based on Improved Ant Colony Algorithm (IACA) in determining the optimal parameters of LSSVM for diagnosing Hepatitis disease. IACA create a storage solution to keep the whole route of the ants. The solutions that have been stored were the value of the parameter LSSVM. There are three main stages in this study. Firstly, the dimension of Hepatitis dataset will be reduced by Local Fisher Discriminant Analysis (LFDA). Secondly, search the optimal parameter LSSVM with IACA optimization using the data training, And the last, classify the data testing using optimal parameters of LSSVM. Experimental results have demonstrated that the proposed method produces high accuracy value (93.7%) for the 80-20% training-testing partition.

Keywords: *Classification, Least Squares Support Vector Machines, Improved Ant Colony Algorithm, Local Fisher Discriminant Analysis, Hepatitis Disease.*

Abstrak

Banyak metode klasifikasi yang mampu mendiagnosa seorang pasien mengidap penyakit Hepatitis, salah satunya adalah menggunakan metode klasifikasi Least Squares Support Vector Machines (LSSVM). Terdapat dua parameter yang sangat berpengaruh pada LSSVM yaitu parameter kernel dan parameter regularisasi. Penentuan parameter optimal tersebut harus diperhatikan untuk mendapatkan akurasi klasifikasi yang tinggi pada LSSVM. Penelitian ini mengusulkan metode optimasi Improved Ant Colony Algorithm (IACA) dalam penentuan parameter optimal LSSVM untuk mendiagnosa penyakit Hepatitis. IACA membuat penyimpanan solusi untuk menjaga rute dari keseluruhan semut. Solusi yang disimpan adalah nilai parameter LSSVM. Ada 3 tahapan utama pada penelitian ini yaitu, dimensi dataset Hepatitis direduksi menggunakan metode Local Fisher Discriminant Analysis (LFDA), kemudian parameter optimal LSSVM dicari dengan metode optimasi IACA menggunakan data training, setelah itu data testing diklasifikasikan menggunakan parameter optimal LSSVM. Hasil uji coba menunjukkan bahwa metode yang diusulkan menghasilkan nilai akurasi yang tinggi (93,7%) pada partisi 80-20% training dan testing.

Kata Kunci: *Klasifikasi, Least Squares Support Vector Machines, Improved Ant Colony Algorithm, Local Fisher Discriminant Analysis, Hepatitis.*

1. Introduction

LSSVM classification method is proposed by Suykens, et al. [1], and LSSVM is a development method of the SVM [2]. In the SVM, the optimal hyperplane is obtained by solving quadratic programming problem by minimizing a function with an inequality condition. Different with SVM, LSSVM gives solutions with linear equations, not with the quadratic programming problems [3]. There are two parameters that very influence to improve the classification accuracy on LSSVM, they are kernel parameter (σ^2) and regularization parameter (γ). Determining the optimal parameters must be considered to obtain a high classification accuracy on LSSVM. The parameters can be searched by trial and error, but trial and error is very not efficient and not effective.

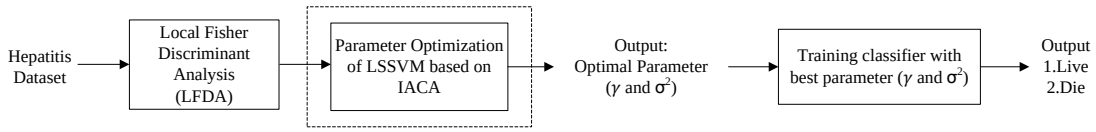

Figure 1. Hepatitis classification system

Therefore, many studies used the Cross Validation (CV) to optimize the parameters LSSVM. However, CV has an disadvantage in computing speed and the accuracy of classification still in average. So that, many researchers proposed a method for optimizing the parameters of the LSSVM.

Zhigang and Chengling [4] proposed a method for predicting the damage depth of coal seam floor using LSSVM. The optimal parameters of LSSVM optimized by Particle Swarm Optimization (PSO). PSO is one of the optimization methods inspired by the behavior of a group of animal movements such as the movement of a group of birds (flock). Each object of animals becomes a particle. A particle in search space has a position that encoded as vector coordinates. This position vector is considered as a state of being occupied by a particle in the search space. Each position in the search space is an alternative solution that can be evaluated using the objective function. Based on experimental results for their testing and training data, PSO-LSSVM can measure the depth of the coal seam floor damage.

Gupta et al [5], proposed a hybrid method of Genetic Algorithm (GA) and LSSVM to increase fault classification of the power transformer. GA generate the initial population randomly, expanding the search space, improve the speed of convergence and search the optimal parameters. The GA-SVM method successfully improve the accuracy of classification errors on the power transformer using DGA dataset (Dissolved gas analysis).

Then, Hegazy et al. [6], proposed a hybrid method Artificial Bee Colony (ABC) and LSSVM on stock price prediction. ABC algorithm is an algorithm inspired by the habits of bees exploration (foraging) to find the optimal solution. ABC chose the best parameters for the LSSVM and avoid the over-fitting problem. That study compared the proposed method with PSO optimization method, where the method of ABC-LSSVM has high convergence speed than PSO-LSSVM. However, the ABC method produces a local minimum parameter [7].

This study proposed an optimization method based on Improved Ant Colony Algorithm (IACA) in determining the optimal parameters in LSSVM to diagnose Hepatitis disease. This algorithm aims to find the optimal path. The search path is based on the behavior of ant colonies in finding the path to a food source [8]. This basic idea then used to solve the problems which illustrated by the behavior of ants. IACA is an optimization method that makes the storage solution to keep the whole route of the ants. The solutions that have been stored were the value of the parameter LSSVM. IACA produces a global minimum parameter at the end of the iteration.

The remainder of this paper is organized as follows. Section 2 describes the methods. Section 3 explained the experimental results. And finally, conclusions and recommendations for future work are summarized in Section 4.

2. Methods

Local Fisher Discriminant Analysis (LFDA)

The dataset that have large feature dimension can be affected to the classification process. The feature dimension can be reduced with dimensionality reduction method. According to [9], dimensionality reduction method is divided into two, they are feature extraction and feature selection. Feature extraction is one of dimensionality reduction method to looking for features that have most relevant information to the original data by transforming the input data into a set of data with the feature that have been reduced [10]. There are several stages in this study (Figure 1).

This study used feature extraction method called Local Fisher Discriminant Analysis (LFDA) to reducing the feature dimension of the dataset. LFDA proposed by Sugiyama [11], LFDA maximize between class separation and defending within class local structure [10]. $\boldsymbol{S}^{(\boldsymbol{bc})}$and $\boldsymbol{S}^{(\boldsymbol{wc})}$are scattered matrix of between class and within class, both of them calculated by equation (1) and (2) [10].

$$\boldsymbol{S}^{(\boldsymbol{bc})} = \frac{\mathbf{1}}{\mathbf{2}} \sum_{\boldsymbol{ij}=\mathbf{1}}^{\boldsymbol{n'}} \boldsymbol{W}_{\boldsymbol{i,j}}^{(\boldsymbol{bc})} \left(\boldsymbol{x_i} - \boldsymbol{x_j}\right)\left(\boldsymbol{x_i} - \boldsymbol{x_j}\right)^{\boldsymbol{T}}, \quad (1)$$

$$\boldsymbol{S}^{(\boldsymbol{wc})} = \frac{\mathbf{1}}{\mathbf{2}} \sum_{\boldsymbol{ij}=\mathbf{1}}^{\boldsymbol{n'}} \boldsymbol{W}_{\boldsymbol{i,j}}^{(\boldsymbol{wc})} \left(\boldsymbol{x_i} - \boldsymbol{x_j}\right)\left(\boldsymbol{x_i} - \boldsymbol{x_j}\right)^{\boldsymbol{T}}, \quad (2)$$

Algorithm 1: Parameter Optimization of LSSVM based on IACA

Input: Number of solutions (*N*), number of ants (*m*), range parameter value (γ, σ^2), size of solutions storage (*k*), termination criterion

Output: Optimal parameter values (γ, σ^2) for LSSVM and classification accuracy

```
Begin
    Initialize N solutions
    Call LSSVM to evaluate N solutions
    //Sort solutions and save them in solutions storage
    A = Sort(S0,S1,....Sn)
    While termination criterion is not do
    //Generate m solutions
    For i = 1 to m do
        //Build solution
        Choose S according to its weight vector
        Save new solution
        Call LSSVM to evaluate new solutions
    End
    // Sort solutions and choose the best N
    A = best(Sorting S0, S1, ..Sn + m), N)
End
```

n is the number of sample in the dataset while $(x_i - x_j)$ is the value based on the local scaling approach [11]. Then the transformation matrix of LFDA $T^{(M)}$ defined in equation (3) [10].

$$T^{(M)} = \underset{T \epsilon \mathbb{R}^{dxr}}{\arg max} \left| tr(T^T S^{(bc)} T (T^T S^{(wc)} T)^{-1} \right|, \qquad (3)$$

d is the number of dataset dimension, and r is the feature dimension that have been reduced. LFDA search the transformation matrix T of the scatter space between class $T^T S^{(bc)} T$ the distribution are maximized and the scatter space within class $T^T S^{(wc)} T$ the distribution are minimized. Dataset dimension that have been reduced is divided into two training – testing partitions, namely 70-30% and 80-20%.

Parameter Optimization of Least Squares Support Vector Machines based on Improved Ant Colony Algorithm

Least Squares Support Vector Machines

Least Squares - Support Vectors Machines (LSSVM) is one of modification of SVM [2] that have been proposed by Suykens and Vandewalle [1]. Besides, the complexity of the calculation is lower, training process LSSVM in large scale is also faster and computation resource is lower than SVM. The same as SVM, LSSVM can be used to classification problem and regression both in the linear case or nonlinear. In the nonlinear case, kernel technique can be applied in the LSSVM. Kernel option that can be used the same as SVM, they are linear, polynomial, RBF, and MLP [1].

$$Q(w, b, \alpha, \xi) = \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i\{y_i[(w.x_i) + w_0] - 1 + \xi_i\} \qquad (4)$$

Every training set is expressed by $(x_i . y_i)$ with $i = 1, 2, \ldots$ N, then $x_i = (x_{i1}, x_{i2}, ..., x_{iq})$ are attribute or feature for the training set i, y_i= (-1, +1) the class label. Then ||w|| is the weight vector of *w*, *C* is a parameter that used to control the trade-off between margin and classification error. Then, *n* is the number of data, and ξ is the slack variable. LSSVM trained by minimizing two function equation using Lagrange Multiplier and become equation (4) [1].

The next difference between SVM and LSSVM are α_i (Lagrange multipliers), the value α_i in LSSVM is positive or negative, whereas in SVM the value must be positive. Besides that, if using the RBF kernel, the number of parameters that must be optimized on the LSSVM lower than SVM, in LSSVM only two parameters that need to determined, they are kernel parameter (σ^2) and regularization parameter (γ). These parameters can be optimized to make the LSSVM hyperplane separating the classes optimally, even in the high dimension space.

Improved Ant Colony Algorithm Optimization

Ant colony algorithm proposed by Marco Dorigo [8] is a probabilistic technique used to solve optimization problems of computing with finding the optimal value to a parameter. The optimal value is the value that obtained through a process and considered to be the best solution of all existing solutions. Ant Colony Algorithm deal with discrete and continuous functions. However, Ant Colony Algorithm that deals with continous functions is considered as a research field [8].

Improved Ant Colony Algorithm (IACA) is modification algorithms based on ant colony algorithm. This algorithm is applied to seen clearly the optimization of continuous functions with increasing several algorithms, such as the objective function below:

$$\min f\ (x_1, \ldots x_n), x_i \in [a_1, b_1], i = 1, 2, .., n \quad (5)$$

Firstly, initializing the number of solution N, then define the range of parameters (γ and σ^2), m shows the population of ants, $x_i^{(0)} = (x_1, x_2, \ldots x_n)$ shows the initial position toward the destination position and $x_i^{(0)}$ is a random point in range variable *i*.

IACA create a storage solution to keep the route of the overall ants. The solution that have been saved is the parameter value (γ and σ^2) LSSVM. The storage of solution need transition probability equation called the weight vector (*w*), *w* will calculate the solution that have been saved in storage solution with the following equation:

$$w_t = \frac{1}{Qk\sqrt{2\pi}} e^{\frac{(t-1)^2}{2Q^2k^2}} \quad (6)$$

Q is the parameter that controls the process of finding the solution, and k is the size of the storage solution. Our proposed algorithm can be seen in Algorithm 1.

Classification accuracy in this algorithm is used to update the storage solution. Then, the transition probability equation is used to choose the solution route of an ant. The solution of the ant will be used as parameters (γ and σ^2) in the kernel RBF from LSSVM classification.

Training LSSVM using the optimal parameter

After obtained the classification model, the next step conducts the prediction process on testing data. In this study, kernel RBF is used to LSSVM classification because its ability to handled the high dimension data [2] and produce a good performance [12].

3. Results and Analysis

In order to evaluate the effectiveness of the proposed method, this study conducts experiments on the Hepatitis dataset. Hepatitis dataset that have been used in this study is from the KEEL Repository [13]. The aim of this dataset is to predict whether a patient Hepatitis disease will die or still live. Hepatitis dataset consists of 19 attributes, 80 instances, and two class labels, they are “die” or “live”. There are 13 class instances labeled "die" and 67 class labeled "live". Hepatitis dataset of KEEL Repository did not contain the missing value. Table 1 is a Table 1nformation 19 attributes of dataset Hepatitis.

Feature extraction method that has been used is LFDA, LFDA algorithm was implemented in Matlab [11]. The number of features extracted using LFDA that we get in the experiment was 5 features. The scatter Plot of class after the dataset dimension reduced can be seen in Figure 2. The sign x indicates the class "live" and o shows the class "die". After getting the reduced dataset dimension, the dataset will be used as input into the next process. The next process is training LSSVM using the Improved ant colony optimization as the system diagram shown in figure 1.

Dataset dimension that have been reduced are divided by two training - testing partitions namely 70-30% and 80-20% as shown in Table 1I. After that, this study makes a classification model and search optimal parameters of LSSVM based on IACA optimization using the training data.

The Input parameters that have been used in IACA process in search the optimal parameters of LSSVM as shown in Algorithm 1 are the number

TABEL 1
HEPATITIS DATASET

No	Attribute	Value
1	Age	10, 20, 30, 40, 50, 60, 70, 80
2	Sex	Male, Female
3	Steroid	No, Yes
4	Antivirals	No, Yes
5	Fatigue	No, Yes
6	Malaise	No, Yes
7	Anorexia	No, Yes
8	Liver Big	No, Yes
9	Liver Firm	No, Yes
10	Spleen Palpable	No, Yes
11	Spiders	No, Yes
12	Ascites	No, Yes
13	Varices	No, Yes
14	Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
15	Alk Phosphate	33, 80, 120, 160, 200, 250
16	Sgot	13, 100, 200, 300, 400, 500
17	Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
18	Protime	10, 20, 30, 40, 50, 60, 70, 80, 90
19	Histology	No, Yes

TABLE 2
TRAINING SET AND TESTING SET

Training – testing partition (%)	Number of instances	
	Training set	Testing set
70-30	56	24
80-20	64	16

TABLE 3
OPTIMAL PARAMETER FOR EACH PARTITION FOUND BY IACA

Partition (%)	σ^2	γ
70-30	17.3	99.8
80-20	18.1	100.4

TABLE 4
CONFUSION MATRIX FOR EACH PARTITION

Actual	Predicted		Partition (%)
	Die	Live	
Die	5	0	70-30 training-testing
Live	2	17	
Die	3	0	80-20 training-testing
Live	1	12	

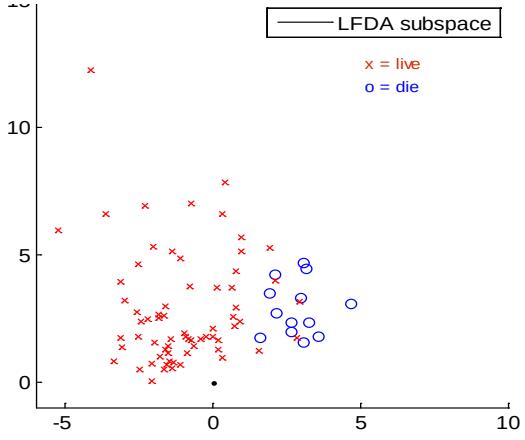

Figure 2. scatter plot class of the reduced dataset found by LFDA

of solution (N) = 50, the number of ants (m) = 10, range $\gamma \in (0, 150)$ and range $\sigma^2 \in (0, 20)$.

The optimal parameters obtained by IACA optimization shows in Table 1II. Each partition generate different value of γ and σ^2, for 70-30% training-testing partition gain γ = 17.3 and σ^2 = 99.8. And 80-30% training-testing partition gain γ = 18.1 and σ^2= 100.4. After classification model obtained, this study conduct prediction process on the testing data.

In order to evaluate the prediction performance of our proposed method, this study computes classification accuracy, sensitivity, and specificity, as shown in confusion matrix for each partition. A classification system is expected to classify all data sets correctly. Generally, the way to measure the performance of a classification using confusion matrix. The confusion matrix is a table that records the result of classification. Based on confusion matrix, it can be seen the amount of data of each class that predicted correctly. By knowing the amount of data that classified correctly, so that it is easy to know the accuracy of the prediction. Another quantity that can be used as a performance classification metric is the sensitivity and specificity. Both of these quantities provide a more relevant performance value. Sensitivity or true positive rate is used to measure the proportions of the original positives correctly predicted as positive. While the specificity used to measure the proportions of the original negatives correctly predicted as negative.

Formula accuracy, sensitivity, and specificity can be seen in equation (7) - (9). In that equation, *TP* (True Positive) is the number of data that is identified properly, *TN* (True Negative) is the number of data that is rejected correctly, *FP* (False Positive) is the number of data that is identified wrongly, and the *FN* (False Negative) is the number of data wrongly rejected.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \; x \; 100\% \quad (7)$$

$$Sensitivity = \frac{TP}{TP + FN} \; x \; 100\% \quad (8)$$

$$Specificity = \frac{TN}{FP + TN} \; x \; 100\% \quad (9)$$

Classification results can be seen using the confusion matrix in Table 1V. We can see in that table the number of false positives more decrease if the size of the training data improved. Then, the classification accuracy using the proposed method can be seen in table V, highest accuracy at 80-20% training - testing partition is 93.7%. The sensitivity that had been achieved for both partitions is 100% while the specificity values for 70-30% training-testing partition was 89% and specificity for 80-20% training - testing partition is 92%.

This study compared the proposed method to classify Hepatitis dataset using LFDA method as a dimensionality reduction method and LSSVM as a classification method without using IACA optimization. Classification accuracy that had been obtained as shown in table VI is 83.3% for the 70-30% training-testing partition and 87.5% for the 80-20% training - testing partition.

This study also classified Hepatitis dataset without using the method of dimension reduction and optimization in determining the optimal parameter and only used the LSSVM classification method. We can see in table VII the classification accuracy that obtained was 79.1% for the 70-30% training-testing partition and 81.25% for the 80-20% training-testing partition.

We also compared the proposed method with previous studies that proposed a classification method for diagnosed Hepatitis disease. Can be seen in table VIII, Genetic Algorithm (GA) and SVM or GA_SVM method [14] proposed by Tan et al., achieve accuracy value as much as 90%. GA is a heuristic search algorithm based on the biological evolution mechanisms. In that study, the GA method used to select the best attributes of

TABLE 5
ACCURACY, SENSITIVITY, SPECIFICITY USING IACA-LSSVM

Metrics	70-30% training – testing	80-20% training - testing
Accuracy	91.6	93.7
Sensitivity	100	100
Specificity	89	92

TABLE 6
CLASSIFICATION ACCURACY USING DIMENSION REDUCTION (LFDA) AND LSSVM WITHOUT IACA OPTIMIZATION

Partition (%)	Accuracy (%)
70-30	83.3
80-20	87.5

TABLE 7
CLASSIFICATION ACCURACY ONLY USING LSSVM

Partition (%)	Accuracy (%)
70-30	79.1
80-20	81.2

TABEL 8
CLASSIFICATION ACCURACIES OBTAINED WITH OUR METHOD AND OTHER METHODS

Method	Accuracy (%)
CSFNN	90
LDA	86,4
GA-SVM	89,6
Our Method	93,7

Hepatitis dataset, then the dataset that have been selected were classified using SVM with 20 fold cross validation.

While the Local Discriminant Analysis (LDA) method [15] proposed by Stern and got 86,4% of accuracy. In that study, Linear Discriminant Analysis (LDA) modified by Bayes algorithm function called maximum likelihood. LDA is a classification method that tries to find a linear subspace and maximize the separation of two classes based on Fisher Criterion.

And the method proposed by Ozyilmaz and Yildirim namely Conic Section Function Neural Network (CSFNN) [16] achieve an accuracy values as much as 77,4%. The CSFNN is NN algorithm that combined Multilayer Perceptron (MLP) and Radial Basis Function (RBF) to improved the Back Propagation performance.

4. Conclusion

This study proposed an optimization method based on Improved Ant Colony Algorithm (IACA) for LSSVM in determining the optimal parameters for diagnosing Hepatitis disease. IACA Algorithm gives optimal parameter LSSVM in each iteration. This study has three main steps: 1) the dimension of Hepatitis dataset reduced by LFDA, 2) search the optimal parameter LSSVM with IACA optimization using the data training, and 3) classify the data testing using optimal parameters of LSSVM. The experimental results show that the proposed method is able to improve the accuracy classification of Hepatitis disease.

This study compared the performance of our method with three other methods, they are LDA, CSFNN, and GA-SVM. Our proposed method achieved high accuracy for the 80-20% training-testing partition (93.7%).

Future investigation will pay attention about the influence of range value γ and σ^2 that we used in search the optimal parameter of LSSVM. Then, Analyzing the input parameter of IACA should be our future work.

References

[1] Suykens, J. A. K., & Vandewalle, J. “Least squares support vector machine classifiers”, *Neural Processing Letters*, vol. 9 (3), pp. 293–300, 1999.

[2] Vapnik, V. “The nature of statistical learning theory”. New York: Springer, 1995.

[3] Tsujinishi, D., & Abe, S. “Fuzzy least squares support vector machines for multi-class problems”, *Neural Networks Field*, vol. 16, pp. 785–792, 2003.

[4] Zhigang, Yan., Chengling, Cui., “An intelligent model for predicting the damage depth of coal seam floor based on LSSVM optimized by PSO”, *Jurnal of applied sciences* 13 (11), pp. 1954-1959, 2013.

[5] Gupta, Aparna R. Et al., “LSSVM Parameter Optimization Using Genetic Algorithm To Improve Fault Classification Of Power Transformer, Engineering Research and Applications”, IJERA, Vol.2, Issue 4, pp.1806-1809, July-August 2012.

[6] Hegazy, Osman., Omar S. Soliman, and Mustafa Abdul Salam, “LSSVM-ABC Algorithm for Stock Price prediction”, *International Journal of Computer Trends and Technology* (IJCTT) – vol: 7 number 2, Jan 2014.

[7] Gao, W., & Liu, S. (2012). “A modified artificial bee colony”, *Computers & Operations Research*, vol: 39, pp. 687-697, 2012.

[8] Dorigo, M. and Stutzle, T., “Ant Colony Optimization”, *The Massachusets Institut of Technology Press*, Cambridge, 2004.

[9] Pudil, P.; Novovicová, J. "Novel Methods for Feature Subset Selection with Respect to Problem Knowledge". In Liu, Huan; Motoda, Hiroshi. *Feature Extraction, Construction, and Selection*. p. 101, 1998.

[10] Chen, Hui-Ling. “A new hybrid method based on local fisher discriminant analysis and support vector machine for Hepatitis disease diagnosis”, *Internasional Journal of Engineering and science*, vol: 38, pp. 11796-11803, 2011.

[11] Sugiyama, M. “Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis”, *Journal of Machine Learning Research*, vol. 8, pp.1027-1061, 2007.

[12] Zhang, H. et al., “Three-Class Classification Models of LogS and LogP Derived by Using GA – CG – SVM Approach”, *Molecular Diversity*, Springer, vol. 13, no. 2, pp. 261-268, 2009.

[13] J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera. KEEL Data-Mining Software Tool: "Data Set Repository, Integration of Algorithms and Experimental Analysis Framework". *Journal of Multiple-Valued Logic and Soft Computing* 17: 2-3, pp. 255-287, 2011.

[14] K.C. Tan, E.J. Teoh, Q. Yu, K.C. Goh, "A hybrid evolutionary algorithm for attribute selection in data mining", *Expert Systems with Applications*, vol: 36 (4), pp. 8616–8630, 2009.

[15] B. Ster, A. Dobnikar, "Neural Networks in Medical Diagnosis: Comparison with Other Methods", 1996.

[16] L. Ozyilmaz, T. Yildirim, "Artificial neural networks for diagnosis of Hepatitis disease", in: *Proceedings of the International Joint Conference on Neural Networks*, 2003, vol. 1, pp. 586–589, 2003.

Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information). 10/1 (2017), 50-58
DOI: http://dx.doi.org/10.21609/jiki.v10i1.445

COMPARISON OF IMAGE ENHANCEMENT METHODS FOR CHROMOSOME KARYOTYPE IMAGE ENHANCEMENT

Dewa Made Sri Arsa[1], Grafika Jati[1], Agung Santoso[2], Rafli Filano[2], Nurul Hanifah[2], and Muhammad Febrian Rachmadi[3]

[1] Faculty of Computer Science, Universitas Indonesia, Kampus UI, Depok, 16424, Indonesia
[2] Study Program of Biomedic Technology, Magister Program, Universitas Indonesia
[3]School of Informatics, The University of Edinburgh, 11 Crichton Street, Edinburgh EH8 9LE, United Kingdom

E-mail: dewa.made51@ui.ac.id[1], agung_i_s@yahoo.com[2], s1467961@sms.ed.ac.uk[3]

Abstract

The chromosome is a set of DNA structure that carry information about our life. The information can be obtained through Karyotyping. The process requires a clear image so the chromosome can be evaluate well. Preprocessing have to be done on chromosome images that is image enhancement. The process starts with image background removing. The image will be cleaned background color. The next step is image enhancement. This paper compares several methods for image enhancement. We evaluate some method in image enhancement like Histogram Equalization (HE), Contrast-limiting Adaptive Histogram Equalization (CLAHE), Histogram Equalization with 3D Block Matching (HE+BM3D), and basic image enhancement, unsharp masking. We examine and discuss the best method for enhancing chromosome image. Therefore, to evaluate the methods, the original image was manipulated by the addition of some noise and blur. Peak Signal-to-noise Ratio (PSNR) and Structural Similarity Index (SSIM) are used to examine method performance. The output of enhancement method will be compared with result of Professional software for karyotyping analysis named Ikaros MetasystemT M . Based on experimental results, HE+BM3D method gets a stable result on both scenario noised and blur image.

Keywords: *Chromosome, DNA, Image Enhancement, MetaSystem Ikaros, Image Processing*

Abstrak

Kromosom adalah kumpulan struktur DNA yang membawa informasi makhluk hidup. Informasi yang dapat diperoleh dengan proses Kariotyping. Proses ini membutuhkan citra yang jelas sehingga kromosom dapat dievaluasi dengan baik. Preprocessing harus dilakukan pada citra kromosom melalui penajaman citra. Proses ini dimulai dengan menghapus latar belakang citra. Langkah berikutnya ialah penajaman citra menggunakan metode image enhancement. Makalah ini membandingkan beberapa metode untuk peningkatan citra. Kami mengevaluasi beberapa metode dalam peningkatan gambar seperti Histogram Equalization (HE), Contrast-limiting Adaptive Histogram Equalization (CLAHE), Histogram Equalization with 3D Block Matching (HE+BM3D), dan unsharp masking. Penulis mengevaluasi dan membahas metode terbaik untuk meningkatkan citra kromosom. Oleh karena itu, untuk mengevaluasi metode, gambar asli dimanipulasi dengan penambahan beberapa kebisingan dan blur. Peak Signal-to-noise Ratio (PSNR) and Structural Similarity Index (SSIM) digunakan untuk mengukur kinerja metode. Hasil penajaman dari metode-metode yang dievaluasi akan dibandingkan dengan hasil software profesional untuk analisis kariotipe bernama Ikaros Metasystem T M . Berdasarkan eksperimen diperoleh hasil bahwa HE + BM3D merupakan metode yang paling stabil pada kedua skenario baik citra mengandung noise maupun citra yang kabur.

Kata Kunci: *Kromosom, DNA, Peningkata Citra, Sistem Meta Ikaros, Pengolahan Citra*

1. Introduction

Chromosomes can not be observed with naked eyes. The existence of chromosomes is not known until the development of microscopes which can magnify an object image up to 1000 times. Chromosomes are surrounded by a wall and a thick membrane that is called as the nucleus. It is transparent, but pale, so difficult to distinguish from the environment. Then chromosome is colored into purple, red, green using extract Gentiana flower. In 1888, a German cytological, W. Waldeyer bring out chromosome, which comes

from the Greek word chroma is meaning color and some is meant body [1].

A chromosome is a collection of DNA arrangements that carry information of our life. The human gen coding only use (3%) of the total DNA, then the rest of it is called as nonsense DNA. Gilbert in 2000 said that humans had 150,000 genes, But in 2001 Bork and Copley clarify that basa in the human genome is 39.114 genes (Celera) and 31.780 (The Public Sequence). And then in 2003, Pennisi informs that human genes was decreased again to 20,000. A single gene can produce hundreds of different proteins. These proteins can be grouped into 1,000 families based on their similarity. Therefore, at least 1,000 primary gene responsible for the protein [1,2].

Chromosomes contain of information about human being. Chromosomes can be used to diagnose diseases and disorders suffered by the individual. Several diseases and disorders can be detected from the different chromosomes between normal and abnormal chromosomes. Some disorders are diagnosed based on the chromosome information like Down Syndrome, Patau Syndrome, Edward's Syndrome, Turner Syndrome, Klinefelter Syndrome, Cri-du-chat (cat's cry) syndrome, Prader-Willi Syndrome (PWS), Angelman Syndrome (AS) and so on.

The chromosomes information can be extract using a karyotype technique. Karyotype is started by isolated chromosome from cell then observed through a microscope and organized into somatic chromosomes 1 to 22 and one sex chromosome. We utilize microscope to produce chromosome image as primary diagnostic tool. Chromosome image analysis methodology is divided into two step pre- processing algorithm and classification or analysis. That process still be done manually. Pre-processing algorithm are perform bounding box each chromosome, textural correction, and geometric correction [3]. The textural correction is essential cause original image contain noises, blur, even distortion. They come from poor sample preparation, low contrast band patterns, digital quantization and imaging [4-7]. These noise and distortion cause the karyotyping and identification of chromosome become more difficult and inaccurate. Therefore, the image needs to improve the quality. Moreover, enhance image quality was produced by microscope still conduct manually. The result still depend on human who enhance the image. To solve this problem, computer processing and image processing algorithm are needed and open problem. Some methods were developed such as Histogram Equalization [8], Contrast-limiting Adaptive Histogram Equalization [9], Histogram Equalization with 3D Block Matching (HE+BM3D) [10], and well-know unsharp masking [11]. This paper is going to evaluate the performance of image enhancement methods for chromo- some image analysis. We have to get best methods that obtain good result on both noised or blurred chromosome image. The appropriate methods will improve the image quality through image enhancement. So the abnormality of chromosome can be identified well [12].

This paper has been written in 7 sections. The first section is the introduction which is this section. The second section, which is the next section, is describing some related works for chromosome theory and previous researches which tried to improve image quality. The third section describing methods that we used in the experiment. Then, the fourth section showed our scenario of experiments. Section fifth showed experiment results. The sixth section is the discussion of the results of the experiment. The last section is the references.

2. Methods

Chromosomes Theory

The chromosomes can be analyzed microscopically. Karyotyping will identify individual Chomosome based on size and shape. In humans, there are 22 pairs of somatic chromosomes and one pair of sex chromosome. Each pair of chromosomes has various sizes and structure. In general, there are four types of structures of different chromosomes in humans, namely telocentric, acrocentric, metacentric and submetacentric. Photomicrograph metaphase Chomosomes of cells are arranged in homologous pairs from longest chromosomes until shortest Chomosomes. Then sex chromosome is grouped steparately (Figure 1) [2].

The structure and amount of chromosomes is not always normal and there is always the possibility of a chromosomal mutation that cause function of certain parts of the body became abnormal. Therefore, it is important to get the clear picture karyotype to identify an abnormality or disease. Many abnormalities that can viewed of chromosomal mutations, such as Down Syndrome where their aneuploidy trisomy of the 21st chromosome (Figure 4); Cri-du-chat syndrome, also known as 5p-syndrome that occurs deletion on chromosome 5 (Figure 5); and Philadelphia chromosome t (9; 22) (Q34; Q11), namely the occurrence of chromosomal translocation, Nowell in 1960 found the tiny chromosomes unusual (Figure 6) see in leukocytes in patients with chronic myelogenous leukemia [12].

(a) (b)

Figure 1. (a) Fetus Amnion chromosome (MetaSystem), (b) Enhanced image produced by MetaSystem

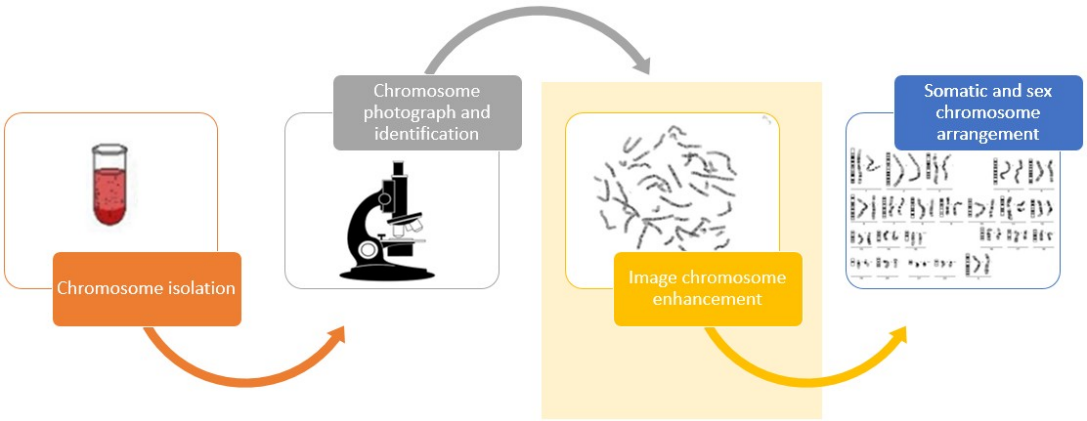

Figure 2. Research position

Previous Works to Improved Quality Images

Imaging has been progressing with the advancement of technology and science. The reason why the images can be digitized is to be converted into a computer can be stored in memory or other forms of storage media such as CD-ROM or hard disk. The digital image is a representation of a two-dimensional image as a finite set of digital values known as pixels or picture elements. Digitization procedure can be done through the scanner or video camera. After the image is digital, it can be done using digital image Fr-ocessing by placing emphasis on two main tasks: improvement of pictorial information for the mur-poses of human interpretation and processing of image data for storage, representation, and trans-mitting the perception of independent machines. Digital image processing (Digital Image Pro-cessing) is a discipline that studies the techniques of image processing. Image is meant here is of still images (photos) and moving images (derived from the webcam). Meanwhile, here are intentions that digital image processing /image is done digitally using a computer [13,14].

The purpose of digital image processing is to find the delegate intensity distribution of any image and transform 3D to 2D images that can be used for quantitative morphological descriptions and representations. Digital image processing often involved many procedures such as for-matting and improve data, digital Increased to increase the enhanced visual interpretation, or automatic target classification and features henti-rely by computer. There are several methods or different techniques of digital image processing procedures. As preprocessing, Image Enhance-ment, Image Transformation and Image Classifi-cation and Analysis are commons image pro-cessing functions provided in the image analysis. Preprocessing is often required before the primary analysis of data and information extraction. It is also need broadly classified as geometric or radiometric correction. Image enhancement invol-

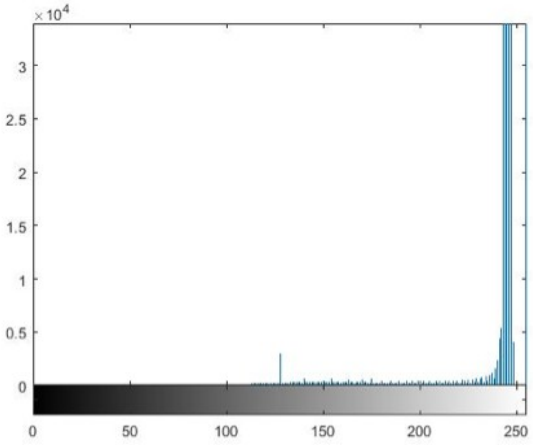

Figure 3. Histogram of chrommosome image before background was removed

Figure 4. Fetus

(a) (b) (c)

Figure 5. (a) Fetus Amnion chromosome (MetaSystem), (b) Enhanced image produced by MetaSystem

ves the modification of one or more components of the image to enhance the clarity and details through visual analysis and interpretation [13].

Some researches have been done to provide a robust method for improved the quality of images. The old one is histogram equalization [8]. This method has been optimized to improved enhancement performance, such as contrast-limited adaptive histogram equalization [9]. Beside of that, neural network is also used to improve image quality. In [15], autoencoder method is used to enhance image quality based on reconstruction method. The network first strained using some images with noise or blur, and the good images as the target. But the neural network based method needs a lot of data to perform well. Moreover, [16,17] developed methods based on entropy theorem. These previous methods are developed

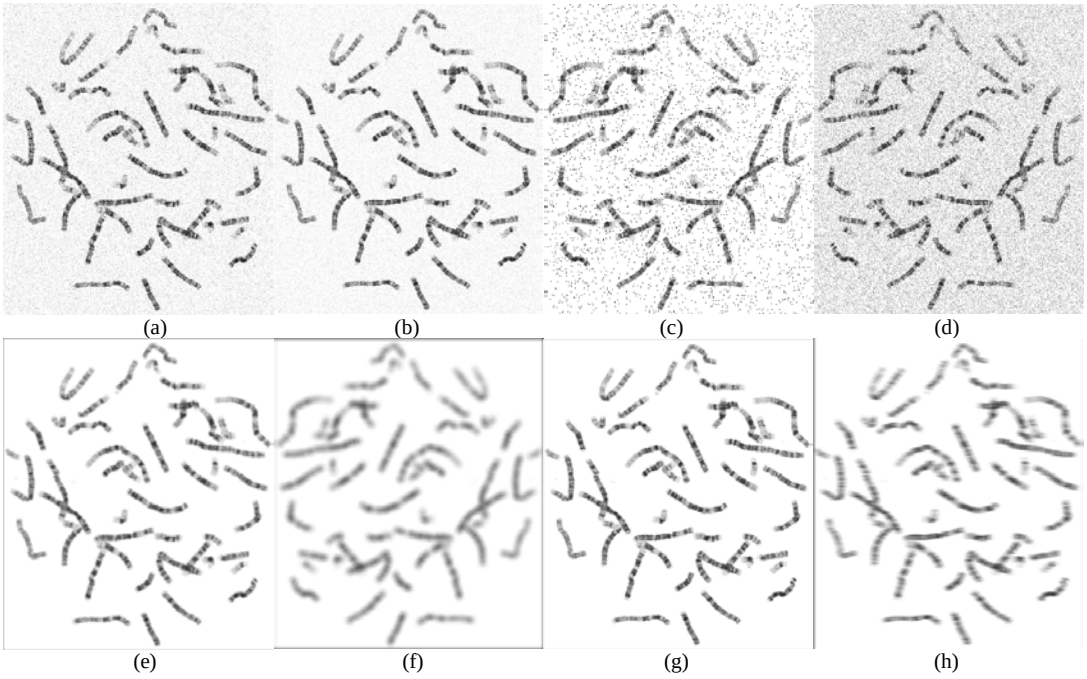

Figure 6. (a-d)Noised images (gaussian, poisson, salt & pepper, speckle), (e-h) Blurred images (average, disk, gaussian, motion)

for natural images. For enhance chromosome image quality, [4-7] have been conducting resarc-hes. [4] proposed multi scale differential operator method. Then, [5-7] used image enhancement methods based on morphological operator. These methods need to set up some parameters manually to achieve the best performance.

Karyotyping Image Processing

Karyotyping image processing is required to clarify the readings, it is important that there are no errors in the readings in this case there are several steps required so that images can be obtained with good. The process of taking pictures in order to obtain a good image also requires its own skills and should be supported by Micro-scope, camera and software sufficient [18].

In this case the chromosome image are a chromosome from MetaSystem image demo. MetaSystem is a company engaged in the field of chromosomes image processing software along with hardware and equipment. One of the exam-ple for chromosome image is chromosome image from the amniotic fluid. The inspection process of karyotyping performed when the parents of these infants were about to find out whether the baby have congenital defects or not. Pictures of the chromosome can be seen in figure 1a. The process will be done by program which build in Matlab software. Figure 2 is showed how chromosome extracted until the chromosome classify. Then the chromosome is observed under microscope and chromosome photograph. After that, this image enhanced manually using karyotype machine.

Image Enhancement Various Methods

There are some methods to enhance image quality. They are histogram equalization, contrast-limiting adaptive histogram equalization, gamma adjust-ment, and histogram equalization with 3D block matching. The methods are described below in detail.

Histogram Equalization (HE)
Histogram equalization is a technique to adjust the in- tensity of image for a better quality [8]. For an image, I, the histogram of equalized image g can be defined by equation(1):

$$g_{i,j} = floor((K-1)\sum_{n=0}^{f_{i,j}} p_n) \quad (1)$$

Where p_n is defined by equation(2):

$$p_n = \frac{pixels\ number\ with\ intensity\ n}{total\ pixels\ number} r \quad (2)$$

Where n has values from 0 to K-1.d

TABLE 1
EXPERIMENT RESULTS USING FIRST SCENARIO

No	Images	Method	PSNR	SSIM
		Unsharp	18.5185	0.9059
1	Fig. 5(a)	HE	12.0534	0.7424
		CLAHE	**24.5843**	0.9703
		HE+BM3	24.5418	**0.9634**
		Unsharp	16.2509	0.8636
2	Fig. 5(b)	HE	12.0534	0.7424
		CLAHE	**21.6709**	**0.9451**
		HE+BM3D	20.9779	0.9317
		Unsharp	12.0496	0.7441
3	Fig. 5(c)	HE	12.0556	0.7425
		CLAHE	**17.1304**	**0.8681**
		HE+BM3D	15.8016	0.8359

TABLE 2
EXPERIMENT RESULTS USING THE SECOND SCENARIO

No	Images	Method	PSNR	SSIM
		Unsharp	14.8434	0.1786
1	Fig. 6(a)	HE	6.3547	0.0685
		CLAHE	13.5149	0.1871
		HE+BM3	**27.5375**	**0.9773**
		Unsharp	14.8434	0.1786
2	Fig. 6(b)	HE	6.3547	0.0685
		CLAHE	13.5149	0.1870
		HE+BM3D	**30.6567**	**0.9846**
		Unsharp	14.8434	0.1786
3	Fig. 6(c)	HE	6.3547	0.0685
		CLAHE	13.5149	0.1870
		HE+BM3D	**18.2848**	**0.4071**
		Unsharp	14.8434	0.1786
4	Fig. 6(d)	HE	6.3547	0.0685
		CLAHE	13.5149	0.1870
		HE+BM3D	**20.6997**	**0.7049**

TABLE 3
EXPERIMENT RESULTS USING THE THIRD SCENARIO

No	Images	Method	PSNR	SSIM
		Unsharp	22.3642	**0.9301**
1	Fig. 6(e)	HE	9.9749	0.6012
		CLAHE	25.2435	0.9238
		HE+BM3	**26.6189**	0.9235
		Unsharp	20.2934	**0.8151**
2	Fig. 6(f)	HE	7.4021	0.4638
		CLAHE	17.5649	0.7344
		HE+BM3D	**20.5542**	0.7883
		Unsharp	23.3867	0.9249
3	Fig. 6(g)	HE	10.3996	0.6232
		CLAHE	29.9063	**0.9598**
		HE+BM3D	**32.3481**	0.9594
		Unsharp	21.3768	0.8736
4	Fig. 6(h)	HE	8.8323	0.5711
		CLAHE	20.7138	0.8492
		HE+BM3D	**22.9089**	**0.8759**

Unsharp Masking (UM)

Unsharp masking is a method that used for sharpening image quality [11] [19]. The first step is blurring. The blurring process is conducted to original image. Then, the original image is subtracted with the blurred image. After that for the last step is adding some deferences to the original image. Unsharp masking is appropriate to sharpen the images. This technique has two noteworthy downsides, for example, the complexity in the darker region is improved substantially more profound than the lighter zone. The next issue is the technique additionally improves the noise and digitization impacts. Due to these issues, the pictures looses its inventiveness much of the time [11]. Among all image enhancement method, the Unsharp Masking approach is the most effortless and less difficult. Equation(3) is UM formulation.

$$f(m,n) = g(m,n) + \lambda Hg(m,n) \qquad (3)$$

Where, f is the unsharpened image. g is the input image, H is the high pass filter image, and λ is gain.

Matlab implement UM in imsharpen function. This function have three parameters Radius, Amount, and Threshold. Radius is the standard deviation of the Gaussian lowpass filter. The Amount is enhance the effect of sharpening and Threshold is the minimum contrast needed for a pixel that is associated with the limit pixels [20]. These three functions influence each other to obtain good sharpness. In the default condition for the Radius value is 1, Amount is 0.8 and the Threshold is 0.

Contrast-limiting Adaptive Histogram Equalization (CLAHE)

CLAHE is a method which is developed based on adaptive histogram equalization (AHE) [9]. In AHE, the intensity of the pixel is transformed into the display range proportional to the pixel intensity's rank in the local intensity histogram. In some cases, AHE can be over-enhancement the noise and may reduce the effect of edge-shadowing. To prevent this event, CLAHE uses strategy to enhance image quality by reduced for very uniform image areas. The calculation is a refinement method which is modified using miposing a user-specified maximum of the local histogram height, and thus on maximum factor of contrast enhancement.

Histogram Equalization with 3D Block Matching (HE+BM3D)

BM3D is an algorithm for noise removal which is proposed by [10]. This method utilizes Wiener filter as a collaborative form. The filter is used to filter dimensional patches block by clustering similar blocks from 2D to 3D array of data and afterward denoising the gathered fixes mutually. Then, the denoised patches are connected back to the first pictures by a voting instrument which expels noise from the considered area.

In general, this method consists of two steps, basic estimate step and final estimate step. These two steps have similar method to produce output. The differences are on the input and grouping phase. Basic estimate step uses a noisy image as the input. Then the input is grouped by block matching. In final estimate step, there are two inputs, the noisy image and the estimate image which is basic estimate step result. On the grouping phase, final estimate step will produce two groups (3-D arrays). They come from basic estimate step result and the noisy image.

$$d(Z_{XR}, Z_X) = \frac{\|Y'(T_{2D}^{HT}(Z_{XR})) - Y'(T_{2D}^{HT}(Z_X))\|_2^2}{(N_1^{ht})^2} \quad (4)$$

Experiment Setup

In this section, we were described about data which we used in the experiment, scenarios of experiments, and performance measurement methods.

Data

To conduct experiments, we used chromosome image which is produced after karyotyping process. We did two transformations before the image ready to enhance. First, we removed the background image by converted into a white color. As figure 1a shows that the background region has gray color. The gray pixels are spun above 225 as shown in figure 3. Then, the intensity of pixels above 225 are converted into 255, so the background changes into white color. Figure 1b was the result of background removing step.

Scenarios of Experiments

We arranged our experiments into three scenarios. In the first scenario, the image that is enhanced by a meta system are embezzled. This process is conducted manually using Photoshop Software. We made three different dark images which have a different dark level. Moreover, the second scenario is enhanced image which has added some noises. There are several noise types we used. They are Gaussian noise, poison noise, salt and pepper noise, and speckle noise. The last scenario is enhanced blurred image. We made 4 images. Each image was filtered using different filter. We chose an average filter, disk filter, Gaussian filter, and motion filter. They are built in Matlab. Some filters (log, Sobel, Prewitt) are not used because those filters used to detect object edges.

Performance Measurement Methods

To measure the methods, we used two measure0-ment methods, peak signal-to-noise ratio and structural similarity index. These methods are described below.

Peak Signal-to-Noise Ratio (PSNR). PSNR is an image comparison method which is usually used to measure reconstruction quality for lossy compression. [21] showed that PSNR is a valid method to measure image quality. On the other hand, PSNR also has low complexity. It makes the implementation easier. As an example of PSNR method with image I, which have mxn and free of noise, the mean square error (MSE) of I and reconstructed version L is expressed by equation(5):

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[(I(i,j) - L(i,j))]^2 \quad (5)$$

Then, the PSNR in decibels (dB) is defined by equation(6):

$$PSNR = 10\,\mathbf{log}_{10}\left\{\frac{\max(I)^2}{MSE}\right\} \quad (6)$$

Structural Similarity Index (SSIM). SSIM is used to capture the sense digital images and video quality. SSIM is formed as a metric which may perceive digital images and video quality [22] [23]. It is used to get the similarity on two images. PSNR and MSE are known as methods which quantify absolute error for the reference and result image. A measurement method like PSNR and MSE may not measure a complete similarity. To give a good measurement, SSIM tries to quantify the change in structure of the image. SSIM can be expressed by equation(7):

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (7)$$

3. Results and Analysis

In this section, we described the results of the experiments based on experiment scenarios. First, we described experiment results from using dark images. After that, results by using noise images as input images. Then, the results of experiments using the third scenario.

Table 1 shows the result of experiments which is conducted using the first scenario. As described above, we used 3 chromosome images (figure 5). CLAHE shows that has higher PSNR for the three images. Beside of that, HE+BM3D has PSNR value not far enough from CLAHE. The SSIM measurement also shows that these two methods performed well.

The next experiment is using the second scenario. Based on the result of a experiment, HE+BM3D method out performed the other method. This result is not surprising that HE+BM3D outperformed the other methods for 4 types of noise images. It is caused because the other method tried to enhance the intensity without noise removing process.

Table 3 showed the result for the third scenario. It shows that on PSNR measurement, HE+BM3D give the highest value for every image. Different results are shown on the SSIM measurement. HE+BM3D did not give the highest value, but it is always on the two highest SSIM value.

4. Conclusion

In this study, some methods, which is developed to enhance image quality, are compared. These methods are used to enhance the image that resulted by karyotyping process. This process is a part of preprocessing in classifying chromosome. By the experiments and analysis, HE+BM3D gave stable results for the three scenarios, event hough, in the first and third scenario didn't give the highest PSNR and SSIM values. It is also shown that HE+BM3D can adapt on dark image, noised image, and blurred image.

References

[1] A. Lima-de Faria, Praise of Chromosome" folly": Confessions of an Untamed Molecular Structure. World Scientific, 2008.

[2] J. S. Hwang, "Topic 2a: Chromosome," class notes for Human Molec- ular Genetics, USCSI University, Kuala Lumpur, Malaysia, 2013.

[3] A. Khmelinskii, R. Ventura, and J. Sanches, "Chromosome pairing for karyotyping purposes using mutual information," in 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, May 2008, pp. 484–487.

[4] Y.-P. Wang, Q. Wu, K. R. Castleman, and Z. Xiong, "Chromosome image enhancement using multiscale differential operators," IEEE Transactions on Medical Imaging, vol. 22, no. 5, pp. 685–693, 2003.

[5] W. Yan, "Mathematical morphology based enhancement for chromo- some images," in Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009. 3rd International Conference on. IEEE, 2009, pp. 1–3.

[6] P. Wayalun, P. Chomphuwiset, N. Laopracha, and P. Wanchanthuek, "Images enhancement of g-band chromosome using histogram equal- ization, otsu thresholding, morphological dilation and flood fill tech- niques," in Computing and Networking Technology (ICCNT), 2012 8th International Conference on. IEEE, 2012, pp. 163–168.

[7] W. Yan, "Enhancement methods for chromosome images," in Electri- cal and Control Engineering (ICECE), 2011 International Conference on. IEEE, 2011, pp. 3024–3026.

[8] Y.-T. Kim, "Contrast enhancement using brightness preserving bi- histogram equalization," IEEE transactions on Consumer Electronics, vol. 43, no. 1, pp. 1–8, 1997.

[9] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in Graphics gems IV. Academic Press Professional, Inc., 1994, pp. 474–485.

[10] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," IEEE Trans- actions on image processing, vol. 16, no. 8, pp. 2080–2095, 2007.

[11] T. Luft, C. Colditz, and O. Deussen, Image enhancement by unsharp masking the depth buffer. ACM, 2006, vol. 25, no. 3.

[12] J. S. Hwang, "Topic 2b: Chromosomal mutation," class notes for Human Molecular Genetics, USCSI University, Kuala Lumpur, Malaysia, 2013.

[13] M. S. Alkoffash, M. J. Bawaneh, H. Muaidi, S. Alqrainy, and M. Alzghool, "A survey of digital image processing techniques in character recognition," International Journal of Computer Science and Network Security (IJCSNS), vol. 14, no. 3, p. 65, 2014.

[14] R. Kusumanto and A. N. Tompunu, "Pengolahan citra digital untuk mendeteksi obyek menggunakan pengolahan warna model normal- isasi rgb," Semantik, vol. 1, no. 1, 2011.

[15] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," arXiv preprint arXiv:1511.03995, 2015.

[16] Y. Niu, X. Wu, and G. Shi, "Image enhancement by entropy max- imization and quantization resolution upconversion," in 2014 IEEE International Conference on Image Processing (ICIP). IEEE, 2014, pp. 4047–4051.

[17] C. Li, Y. Yang, L. Xiao, Y. Li, Y. Zhou, and J. Zhao, "A novel image enhancement method using fuzzy sure entropy," Neurocomputing, vol. 215, pp. 196–211, 2016.

[18] H. Kapitza and S. Lichtenberg, Microscopy from the very beginning. Zeiss, 1997.

[19] J. N. Archana and P. Aiswarya, "A review on the image sharpening algorthms using unsharp masking," International Journal of Engi- neering Science and Computing, vol. 6, no. 7, 2016.

[20] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, Digital Image Processing Using MATLAB. Upper Saddle River, NJ, USA: Prentice- Hall, Inc., 2003.

[21] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," Electronics Letters, vol. 44, no. 13, pp. 800–801, June 2008.

[22] A. Łoza, D. R. Bull, P. R. Hill, and A. M. Achim, "Automatic contrast enhancement of low-light images based on local statistics of wavelet coefficients," Digital Signal Processing, vol. 23, no. 6, pp. 1856–1866, 2013.

[23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, vol. 13, no. 4, pp. 600–612, 2004.