A New Protocol for On-Line User Identification Based on Hand-Writing Characters

Ryota Hanyu, Qiangfu Zhao, Yuya Kaneda
The University of Aizu
Aizu-wakamatsu, Fukushima, Japan, 965-8580
Email: {m5191121, qf-zhao, d8161108}@u-aizu.ac.jp

***Abstract*—Biometric authentication (BA) is becoming more and more popular. Usually, we expect that BA can make various service systems more secure, but in fact it can be more dangerous. For example, fingerprint is one of the popular biometrics for authentication. We say it is dangerous because we cannot change our fingerprints even if they are collected and duplicated by some malicious third parties. This kind of "life-long" biometrics, once they are stolen, can never be used as an authentication factor in the future. To solve the problem, we may use "changeable" biometrics. Examples include face, voice, and hand-writing characters. In this study, we use hand-writing characters. Hand-writing characters can change naturally in the aging process, they can also be changed intentionally through training. This paper investigates the feasibility of on-line user identification using hand-writing non-alphanumeric characters. Our main purpose is to develop some core technologies that can improve the security of service systems in some Asia countries that use Chinese characters.**

I. Introduction

Biometric authentication (BA) has become a popular method for authentications. Fingerprint is one of the most popular factors used for BA. It is widely used in our daily lives (entrance control, banking, mobile device login, etc.) and many people trust it. In fact, there is a trend that most users trust BA too much. Of course, fingerprint-based BA is more secure than weak human-memorable pass-phases or PIN codes. However, BA also has several not negligible problems. One problem is that the number of biometric factors we can use is limited. Generally, we have 2 hands, 10 fingers, and 2 eyes. We may have vein, fingerprint, iris, and so on, and these are the main factors for BA. Even if we use one of them to authenticate the user for each device or place, we can provide only 14 different features. This number is definitely not enough in our life time (around 80 years in average).

Another problem is the easiness to collect and duplicate. In the case of fingerprint, some third parties can get our fingerprints easily because we touch many things everyday. In other words, we may spread our own authentication factors everyday, and everywhere. It is worse that even if our biometrics are stolen, we cannot change them in our whole lives. Thus, using "life-long" biometrics for BA can be very dangerous.

To avoid the above problems, it is natural to use changeable factors such as face, voice, and hand-writing (or written) characters. Since face and voice can be collected and duplicated easily by the third party, we focus on hand-writing characters in this study. Hand-writing characters can change naturally in the ageing process, they can also be changed intentionally through training. In addition, writing environment (e.g. tablet terminal or smart phone) may also change the features of hand-writing characters. For example, although the same person can write very similar characters on the same smart phone, it is difficult for the third party to duplicate the characters with a different device. Further, in recent years, we usually do not really "write", but "type". Therefore, it is difficult for the third party to collect a set of writing characters via the internet or paper-based notebooks. Thus, hand-writing characters can be a secure factor for BA.

In this paper, we investigate the feasibility of BA based on non-alphanumeric characters. In Japan, it is natural to ask the ordinary users to write some Chinese characters (Kanji) or Kana characters in the authentication process. This time, we investigate, through some preliminary experiments, the "strength" of different characters (e.g. Kanji, Hiragana and Katakana) for BA.

II. The Proposed Protocol

In this section, we present the detail of the protocol which we propose. Fig. 1 is the flowchart of the our protocol.

A. Client Side

The role of the client is to collect strokes of hand-writing characters and send them to the server. Each character consists of a set of writing points, times, and the ID of the stroke to which each point belongs. We can reproduce hand-writing characters from the strokes.

B. Server Side

On the server side, the server generates a character set randomly or based on some prespecified rule, and sends them to the client, it then receives strokes from the client. Finally, the server identifies the user using the received strokes and send back the result.

C. The flow of authentication

1) User requests to acquire service.
2) The server generates a character string from the character set and send them to the client.
3) The client collects hand-writngs based on the generated random string and send them to the server.
4) The server authenticate the user and sends result to the client.

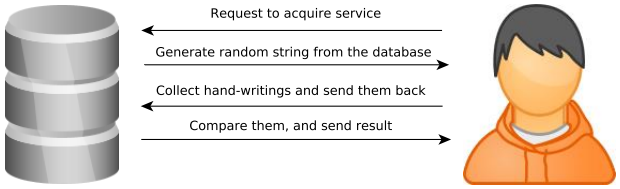

Fig. 1. Flow of the proposed protocol

III. Fundamental knowledge

A. Relationship among Hiragana, Katakana, and Kanji

As we know, Kanji was born in China, and Japanese people have imported and used them for so many years. Japanese people made Hiragana and Katakana from Kanji around 1,100 years ago, but there is clearly difference between Hiragana and Katakana. Hiragana was made from Kanji in cursive style, but Katakana was made from a part of Kanji, therefore some of Katakana are very similar to Kanji. For example, Katakana 二 and Kanji 二, Katakana ハ and Kanji 八 are very similar.

B. Shape Primitives Probability Distribution Function

We choose the shape primitives probability distribution function (SPPDF)[1] as the method to extract features from hand-writing characters. This is a very effective method to extract features from hand-writing Chinese Kanji character sentences to identify writers.

SPPDF treats strokes in the character as simplified direction. The probability distribution of the combinations of these directions is the extracted feature. We can obtain 256 (16 directions by 16 directions) dimensional feature vectors for each character. Fig. 2 and Fig. 3 show the features corresponding to あ and ア written by two different writers.

IV. Dataset

In this section, we provide information about how we collect the data and the environment. We collect hand-writing Japanese characters which are written by young Japanese. As we mentioned in section III-A, there are 3 types of characters in Japanese, Hiragana, Katakana, and Kanji. Hiragana (Table I) and Katakana (Table II) have 48 kinds of characters each. And we choose 80 basic Kanji (Table III) that most be mastered by first-year elementary school students in Japan. We totally collected 880 kinds of characters.

We use the iPad Air to collect characters, because it is one of most popular tablet devices. One of our purposes is to propose a new protocol for user identification without any special equipment, therefore we have to choose a common device.

Each writer wrote each character 5 times using his finger. Fig. 4, 5, and 6 show some examples.

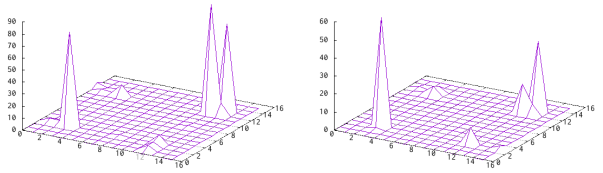

Fig. 2. Features of Hiragana あ and Katakana ア by writer-A

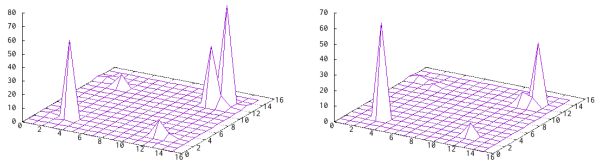

Fig. 3. Features of Hiragana あ and Katakana ア by writer-B

TABLE I
Hiragana characters

あ	い	う	え	お	か	き	く	け	こ	さ
し	す	せ	そ	た	ち	つ	て	と	な	に
ぬ	ね	の	は	ひ	ふ	へ	ほ	ま	み	む
め	も	ら	り	る	れ	ろ	や	ゆ	よ	わ
を	ん									

TABLE II
Katakana characters

ア	イ	ウ	エ	オ	カ	キ	ク	ケ	コ	サ
シ	ス	セ	ソ	タ	チ	ツ	テ	ト	ナ	ニ
ヌ	ネ	ノ	ハ	ヒ	フ	ヘ	ホ	マ	ミ	ム
メ	モ	ラ	リ	ル	レ	ロ	ヤ	ユ	ヨ	ワ
ヲ	ン									

TABLE III
Kanji characters

一	右	雨	円	王	音	下	火	花	貝	学
気	九	休	玉	金	空	月	犬	見	五	口
校	左	三	山	子	四	糸	字	耳	七	車
手	十	出	女	小	上	森	人	水	正	生
青	夕	石	赤	千	川	先	早	草	足	村
大	男	竹	中	虫	町	天	田	土	二	日
入	年	白	八	百	文	木	本	名	目	立
力	林	六								

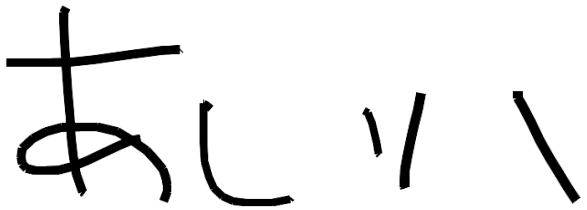

Fig. 4. Characters written by writer-A

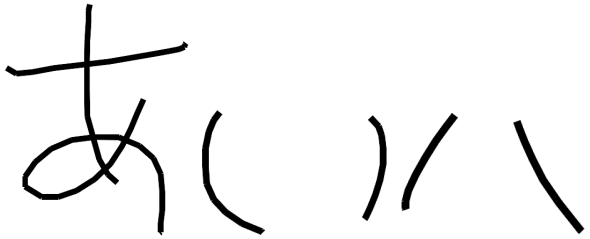

Fig. 5. Characters written by writer-B

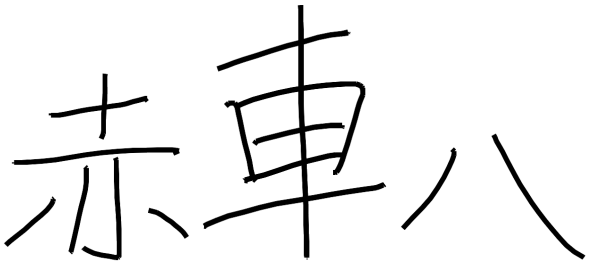

Fig. 6. Characters written by writer-C

V. Experiment I: Confirmation of Usefulness of Characters

A. Method

We calculated Euclidean distances between the feature vectors of each pair of characters. Figs. 2 and 3 are several sample graphs of the extracted features. And we define some distances among hand-writing characters to measure usefulness of characters.

1) Intra distance
   This is the distance between hand-writings of the same character written by the same writer. We focus on 3 types of intra distances, the minimum, the average, and the maximum.
2) Inter distance
   This is the distance between hand-writings of the same character written by different users. And we also focus on 3 types of inter distances, the minimum, the average, and the maximum.
3) Inter-intra distance
   We define inter-intra distance as the difference between the minimum inter distance and the maximum intra-distance.

B. Results

We obtained feature vectors for 432 character sets by 3 writers. We show the result which is sorted by inter-intra distance in ascending order in Table IV, the Table V is same the result but in descending order. Both of them show the top 20 sets.

The intra-class means distance between the same characters written by the same writer, and the inter-class means distances between same character written by different writers.

The column which is named inter-intra shows difference between minimum value of the inter-class distance and maximum value of the intra-class distance, therefore this is one of the index of usefulness of the character. When this index becomes higher value, it means the character written by the writer has unique characteristics, otherwise not so much.

Table IV and Table V show features extracted from Kanji are more unique than Katakana, even if the Kanji is very simple (like 八 or 口). And there is an interesting result. Kanji 八 written by writer-B is in the second place by inter-intra class distance, but Katakana ハ written by the same writer is in the lower second place. Kanji 八 and Katakana ハ are almost the same character, but the score shows opposite results. We will discuss this later.

We can also read from Table V that more than half of worst 20 sets are Katakana characters. And from the all results, around 50 sets are not suitable for our purpose. Most of the 50 sets are also Katakana or Hiragana characters.

C. Discussion

In this time, we obtained results for 432 character sets. And we give two Tables IV, V for their top/worst of 20 characters.

As we mentioned in the previous section, Kanji 八 and Katakana ハ which are written by the same writer show the opposite results. To discuss this problem, we provide some figures and graphs. Fig. 7, 9, 11, 13 are the images of characters, and Fig. 8, 10, 12, 14 are the graphs for their features.

As seen, they two have almost same silhouette, so there is no difference between the features which extracted from them written by writer-A. But in the case of written by writer-B, the feature extracted from 八 is different from other ハ and 八. Writer-B has habit on 八 and it makes difference among other similar characters.

Kanji 八 written by writer-B is unique. It is the cause that makes inter-class distance between writer-A and writer-B on Kanji 八 to be far. On the other hand, their Katakana ハ are not unique, but plain characters, so the interclass distance becomes close.

From these facts, simple Katakana and Kanji characters which make from straight lines are basically not suitable. But some writers have very unique habit on their handwriting, and it could be unique features.

As we mentioned, there is tendency to Katakana characters are not suitable for this usage, because most of them are very simple characters. It shows that we should not only remove simple characters from the candidates, but also need to collect features from not suitable characters and compare to others.

VI. Experiment II: Confirmation of Performance

A. Method

First, we filter characters by inter-intra distance. This means to remove unsuitable characters.

We repeat 1,500 times to randomly select a writer, 5 characters written by another randomly selected writer as a testing set, and apply bellow steps for each character.

Step 1: Define similarity values among input character and characters which in the database without using the testing dataset. In this time, we choose some simple similarity definitions. The minimum intra-class distance-based similarity S_{min} is given in Eq. 2, the average of intra-class distance-based similarity S_{avg} is given in Eq. 3, and the maximum intra-class distance-based similarity S_{max} is given in Eq. 4.

Step 2: Define dynamic thresholds for each characters, and writers. We make groups of characters by writer and character, and obtain threshold from character without the test character by Eq. (1).

Step 3: Compare the similarity and thresholds. In the case of same writer, the similarity should be less than the threshold.

$$T_{user\_char} = \frac{1}{1 + d_{max}} \tag{1}$$

where d_{max} is the maximum intra-class distance.

$$S_{min} = \frac{1}{1 + d_{min}} \tag{2}$$

where d_{min} is the minimum intra-class distance.

$$S_{avg} = \frac{1}{1 + d_{avg}} \tag{3}$$

where d_{avg} is the average of intra-class distance.

$$S_{max} = \frac{1}{1 + d_{max}} \tag{4}$$

where d_{max} is the maximum intra-class distance.

Following steps are the flow for training, testing.

Step 1: Initialization
1. Select the true user: i
2. Select the login user: j
3. Select M characters: $m = 1, 2, \cdots, M$
4. Number of success: $N_S = 0$
5. Counter for test: $k = 1$

Step 2: $m = 1.success = 0$

Step 3: Select test data
1. $TestData = SelectTestData(\Omega(m, j))$
2. $TrainData = \Omega(m, i)$
   where $\Omega(character\_id, writer)$ is a function which returns a randomly selected character filtered by $character\_id$ and $writer$.
3. If $i == j$ then
   $TrainData = TrainData - \{TestData\}$

Step 4: $S = Similarity(TrainData, TestData)$

Step 5: Check results
1. If $i == j \ AND \ S > T(i, m)$ then $success++$
2. If $i! = j \ AND \ S < T(i, m)$ then $success++$

Step 6: $m++$; If $m < M$ then return to Step. 3

Step 7: If $success > T$ then N_S++

Step 8: $k++$; If $k < N_{test}$ then return to Step. 2

Step 9: $N_S/N_{test} \rightarrow successrate$

B. Results

In this time, we filtered characters which have inter-intra distance less than 0.0 and used 398 characters for the experiments.

Here, we show results of experiments used similarity Eqs. (2), (3), and (4).

But our results are arrays of Boolean values, so we need to define how to evaluate them. In this time, we simply count up the number of true in the result and if it is more than the threshold t then treat it as true, otherwise false. We choose 3, 4, and 5 as the threshold and Table VI, VII, VIII are results with 3, Table IX, X, XI are $t = 4$, and Table XII, XIII, XIV are $t = 5$ setting.

TABLE VI
Accuracy with $S_{min}(t = 3)$: 65.6%

		Prediction	
		True	False
Actual	True	500	516
	False	0	484

TABLE VII
Accuracy with $S_{avg}(t = 3)$: 88.2%

		Prediction	
		True	False
Actual	True	494	170
	False	6	830

TABLE VIII
Accuracy with $S_{max}(t = 3)$: 88.3%

		Prediction	
		True	False
Actual	True	343	18
	False	157	982

TABLE IV
RESULT WHICH SORTED BY INTER-INTRA CLASS DISTANCE IN ASCENDING ORDER

				intraclass distance			interclass distance		
Character	Kind	Writer	inter-intra	min	average	max	min	average	max
七	Kanji	B	1.3774	0.01726	0.0285	0.0414	1.4188	1.7772	2.0911
八	Kanji	B	1.2755	0.12171	0.1572	0.2061	1.4817	2.0238	2.6188
五	Kanji	A	1.1826	0.02201	0.0310	0.0385	1.2211	1.5303	1.8347
八	Kanji	A	1.0832	0.09932	0.1361	0.2262	1.3094	2.1362	3.3982
シ	Katakana	B	1.0609	0.06341	0.1000	0.1517	1.2126	1.6633	2.3982
下	Kanji	C	1.0374	0.04558	0.0623	0.0934	1.1309	1.6839	2.2556
上	Kanji	C	1.0149	0.02957	0.0437	0.0617	1.0767	1.7898	2.5465
気	Kanji	C	1.0003	0.00868	0.0112	0.0161	1.0165	1.1072	1.2168
五	Kanji	B	0.9479	0.02684	0.0337	0.0389	0.9869	1.1229	1.2640
九	Kanji	B	0.9273	0.02926	0.0442	0.0656	0.9929	1.2765	1.4434
口	Kanji	A	0.9170	0.02248	0.0292	0.0375	0.9546	1.1931	1.3975
水	Kanji	C	0.8875	0.03026	0.0443	0.0661	0.9536	1.5556	2.0847
小	Kanji	C	0.8692	0.02432	0.0377	0.0493	0.9185	1.3124	1.8073
九	Kanji	C	0.8670	0.01775	0.0337	0.0554	0.9224	1.2003	1.4418
五	Kanji	C	0.8629	0.01823	0.0243	0.0377	0.9006	1.1781	1.3703
上	Kanji	B	0.8609	0.05344	0.0731	0.1035	0.9644	1.2249	1.7748
目	Kanji	B	0.8421	0.03415	0.0536	0.0897	0.9319	1.0963	1.2838
オ	Katakana	B	0.8344	0.02146	0.0349	0.0542	0.8886	1.1928	1.4979
一	Kanji	C	0.8090	0.08092	0.1024	0.1182	0.9272	1.1907	1.5854
口	Kanji	C	0.8074	0.05000	0.0769	0.1482	0.9556	1.0929	1.3742

TABLE V
RESULT WHICH SORTED BY INTER-INTRA CLASS DISTANCE IN DESCENDING ORDER

				intraclass distance			interclass distance		
Character	Kind	Writer	inter-intra	min	average	max	min	average	max
ハ	Katakana	B	-0.5553	0.3919	0.6226	1.0579	0.5025	1.1585	1.8558
く	Hiragana	B	-0.4173	0.1571	0.2962	0.6489	0.2316	0.5358	1.4878
へ	Hiragana	B	-0.3899	0.1962	0.4519	0.9019	0.5119	1.1465	3.0017
し	Hiragana	A	-0.2869	0.2163	0.3349	0.6537	0.3667	0.5814	1.1191
ツ	Katakana	B	-0.2465	0.2154	0.3683	0.5689	0.3224	0.7749	1.2402
フ	Katakana	A	-0.2209	0.1213	0.2122	0.4467	0.2258	0.4931	1.3606
さ	Hiragana	B	-0.1280	0.0872	0.1723	0.2780	0.1499	0.3201	0.4841
林	Kanji	C	-0.1228	0.1071	0.2015	0.4108	0.2879	0.6237	0.8708
ト	Katakana	A	-0.1029	0.2455	0.3821	0.6347	0.5318	0.8090	1.1012
チ	Katakana	B	-0.0884	0.1005	0.1527	0.3248	0.2363	0.3371	0.5777
ま	Hiragana	A	-0.0709	0.0636	0.1133	0.1711	0.1001	0.2132	0.3763
イ	Katakana	A	-0.0703	0.0902	0.1803	0.3522	0.2819	0.5544	1.2672
ケ	Katakana	A	-0.0664	0.0455	0.0801	0.1700	0.1035	0.2048	0.5042
リ	Katakana	B	-0.0634	0.1275	0.2296	0.3429	0.2794	0.4553	0.6682
ヤ	Katakana	B	-0.0548	0.1491	0.2172	0.3048	0.2500	0.4757	0.7306
メ	Katakana	A	-0.0438	0.1637	0.2139	0.2781	0.2342	0.4729	0.7608
コ	Katakana	B	-0.0392	0.1483	0.2395	0.5096	0.4703	0.6131	0.7473
の	Hiragana	B	-0.0377	0.0492	0.0831	0.1270	0.0892	0.1492	0.2782
レ	Katakana	B	-0.0257	0.1228	0.1937	0.2786	0.2529	0.5415	1.0873
ア	Katakana	A	-0.0242	0.0824	0.1220	0.1984	0.1741	0.2662	0.4004

C. Discussion

In this time, we prepare a small dataset for our experiments. We have only 5 data for each characters/writers, so we can use only 4 of them for training. But the results of experiment B show that we can obtain 3.1% (Table X) of false-positive rate and totally 92.0% of accuracy by using simple features and methods.

In the previous section, we presented 9 tables of the results. These results show that the size of dataset is not enough, in other words, the effect from outliers is large. So the accuracy could be better with larger dataset.

These results mean that the feasibility of on-line user identification using hand-writing non-alphanumeric characters is high.

VII. CONCLUSION

In the paper, we discussed feasibility of on-line user identification based on hand-writing Japanese characters. According to results given in this paper, we can identify users by their hand-writing characters.

However, the results show that there are some not suitable characters. Table V shows more than half of worst 20 characters are Katakana characters, but as mentioned in previous section, we should not only remove simple characters from the candidates, but also need to collect features from not suitable characters and compare to

TABLE IX
ACCURACY WITH $S_{min}(t = 4)$: 85.0%

		Prediction	
		True	False
Actual	True	488	213
	False	12	787

TABLE X
ACCURACY WITH $S_{avg}(t = 4)$: 92.6%

		Prediction	
		True	False
Actual	True	421	31
	False	79	969

TABLE XI
ACCURACY WITH $S_{max}(t = 4)$: 78.0%

		Prediction	
		True	False
Actual	True	172	1
	False	328	999

TABLE XII
ACCURACY WITH $S_{min}(t = 5)$: 90.5%

		Prediction	
		True	False
Actual	True	390	32
	False	110	968

TABLE XIII
ACCURACY WITH $S_{avg}(t = 5)$: 81.8%

		Prediction	
		True	False
Actual	True	229	2
	False	271	998

TABLE XIV
ACCURACY WITH $S_{max}(t = 5)$: 68.8%

		Prediction	
		True	False
Actual	True	32	0
	False	468	1000

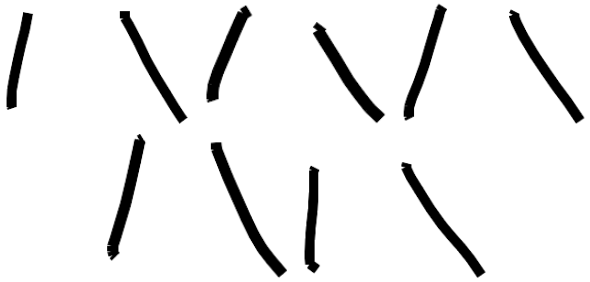

Fig. 7. Five Kanji 八 written by writer-A

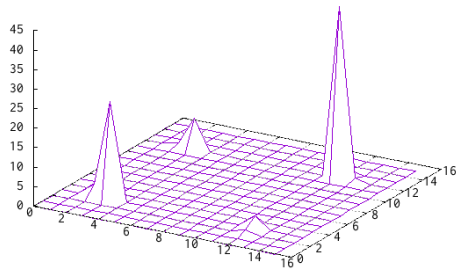

Fig. 8. Features of Kanji 八 written by writer-A

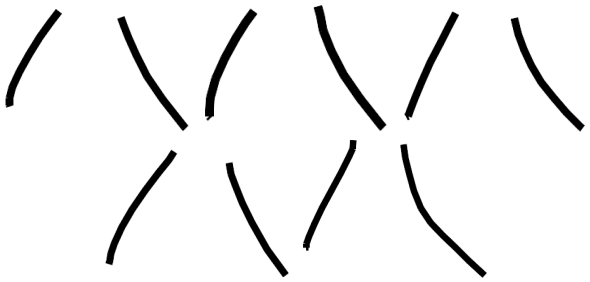

Fig. 9. Five Kanji 八 written by writer-B

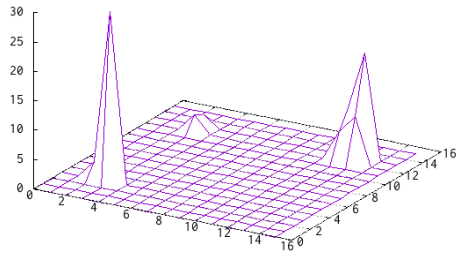

Fig. 10. Features of Kanji 八 written by writer-B

other. After that, measure usefulness of them and decide to use them or not.

SPPDF was originally designed for identifying writers from hand-writing Chinese sentences. But the results shows that it can be used for not only Chinese but also Japanese characters under this situation and has enough performance.

In this time, we do not use some features which are size of characters and writing speed. They could be effective features which represented by previous researches.

In the feature, we will try to use these features and compare them. And we will also try to collect larger dataset to verify these results and improve its performance.

Using classifier is another future work. Making mod-

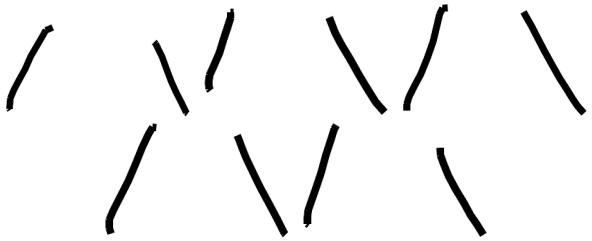

Fig. 11. Five Katakana /\ written by writer-A

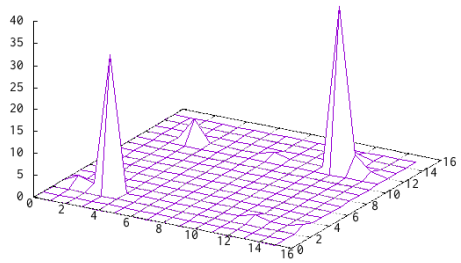

Fig. 12. Features of Katakana /\ written by writer-A

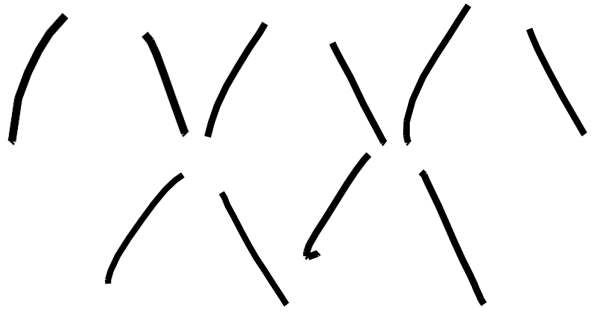

Fig. 13. Five Katakana /\ written by writer-B

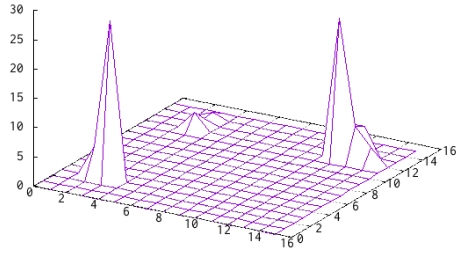

Fig. 14. Features of Katakana /\ written by writer-B

els for each writer and character probably makes better results. We will consider to calculate cost and its performance.

References

[1] Li, B. and Sun, Z. and Tan, T., Hierarchical shape primitive features for online text-independent writer identification, Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2009.

[2] Shin ONODA, Yuuki GOUBARA, and Yasushi YAMAZAKI, A Study on the Performance of a Mobile Terminal-Based Signature Verification System under Different Writing Environments, Journal of IEICE A Vol.J98-A No.12 pp.664-pp7, 2015

[3] Joulia Chapran, Biometric Writer Identification: Feature Analysys and Classification, International Jornal of Pattern Recognition and Artificial Intelligence Vol.20, No.4 (2006) 483-503

[4] Donato Impedovo and Giuseppe Pirlo, Automatic Signature Verification: The State of the Art, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). VOL. 38.NO. 5, 2008