Analysis on compact data formats for the performance of Handwritten Signature Biometrics

Oscar Miguel-Hurtado
Dpto. Electronic Technology
Universidad Carlos III de Madrid
Madrid, Spain
omiguel@ing.uc3m.es

Luis Mengibar-Pozo
Dpto. Electronic Technology
Universidad Carlos III de Madrid
Madrid, Spain
mengibar @ing.uc3m.es

Inmaculada Tomeo-Reyes
Dpto. Electronic Technology
Universidad Carlos III de Madrid
Madrid, Spain
itomeo @ing.uc3m.es

Judith Liu-Jimenez
Dpto. Electronic Technology
Universidad Carlos III de Madrid
Madrid, Spain
jliu @ing.uc3m.es

Abstract— **This paper deals with the Signature Data Formats proposed by ISO 19794 project: 19794 part 7 Full Format and Compact Format (published in 2007) and the new 19794 part 11, which is under development. It will be shown how these formats handle the raw data coming from a Signature Input Device, and what the size of a Biometric Information Record is for each one. Another compression method, using LZ77 compression algorithm, is proposed and tested. The paper will also show the impact of using these compact formats on the performance of two different algorithms: Dynamic Time Warping and Gaussian Mixture Models. MCyT and SVC2004 signature databases have been used to carry out all tests.**

Keywords- Signature, Compression, ISO/IEC Data Interchange Format, Compact Format

I. INTRODUCTION

Automatic Verification Systems based on handwritten signature have the advantage of being a very familiar way of identification for most citizens, as it is used in our daily life as the normal and customary way for identity verification. Therefore, this modality faces no rejection and seems very friendly for users, who usually complain about privacy in other biometrics modalities.

A huge work is being carried out from both public and private organizations in order to achieve reliable Signature Automatic Verification Systems (SAVS). They are not only interested in the scientific challenges but also in the valuable applications that this field offers.

SAVS applications are continuously growing along with the development of more sophisticated and easy-to-use input devices for online handwriting acquisition. For instance, SAVS can be a valuable contribution to document management and health care applications, as well as passport, ID-card and driving license related systems. It can also be a really valuable application for online banking and retail point-of-sales (POS).

Therefore, it is worth highlighting the idea of using smart cards in order to securely store the user's biometrics reference, and, whenever possible, to perform the biometric comparison inside the card. In both cases, the size of the signature sample becomes a really important issue.

As the number of input devices and techniques for handwriting acquisition increases, device interoperability has become an area of greater relevance. Currently, standardization of data interchange formats is being done. In the case of dynamic signature, data captured in the form of a time series arise as the way to provide high interoperability between different systems. The aim of these efforts is to make easier the integration of signature verification technologies into other standard equipment to develop complete solutions for a wide range of commercial applications as the ones aforementioned.

On that way, ISO/IEC has published a standard within the ISO/IEC 19794, part 7 [1]: Biometric data interchange formats – Signature/sign time series data. This standard defines two different sub-formats for raw signature data. One is named Full Format, and the second one is named Compact Format, which tries to minimize the size of a signature biometric data block (BDB).

There is also a proposal for a new data interchange format for processed signature data within 19794 project. This is the new part 11 [2], which is based on the definition of signature singular points, named "turning points". This part also tries to minimize the size of the BDB by means of storing only information of those singular points, instead of storing every single sample points as in part 7.

This paper analyzes the size of the above mentioned three standardized 19794 data formats and also proposes a different lossless compression method to create a more viable compact format. This lossless compression method is based on the deflation algorithm, as a variation of LZ77 [3].

Furthermore, as compact formats imply some loss of information, the impact over the performance of signature algorithms has been studied. The signature algorithms that have been used to analyze the impact of the lossy compact formats are two of the most well known: Gaussian Mixture Models [4] and Dynamic Time Warping [5].

MCyT [6] (occidental users) and SVC2004 [7] (oriental and occidental) signature databases have been used to carry out all tests.

Following sections II and III will briefly explain the way of the three ISO Data Interchange Formats handle and store the raw data provided from signature capture devices. In section IV the deflation algorithm used as a lossless compression method will be introduced. Some modifications of 19794-7 Full Format will be also proposed to improve its compression ratio. The databases used in order to find out the average BDB size for these formats are described in Section V, while the results of the BDB size for all the formats will be shown in Section VI. Once the study of size and compression is completed, Section VII will briefly introduce the signature algorithms used to evaluate the impact of the lossy compact formats and disclose the outcomes of this evaluation. Section VIII will draw some conclusions from this work. Acknowledges and References will conclude this paper.

II. 19794 PART 7: SIGNATURE/SIGN TIME SERIES DATA

Part 7 of 19794 project defines how data captured by a signature input device, in the form of time-series raw data, has to be stored in order to get interoperability between different biometric systems and/or applications.

These raw data are divided in different channels. Allowed channels are: x and y position (X, Y), velocity (VX, VY) and acceleration (AX, AY), z position (Z), time (T) and time difference (DT), pen tip force (F), switch state (S) and pen orientation (TX, TY, Az, El, R).

Two different formats are defined within this part. The first one is the Full Format for general use. Second one is the Compact Format, which is used in applications where the size of the biometric record is an important issue, as occurs with smart cards and other tokens. Both will be explained in detail in the following subclauses.

A. Full Format

19794-7 Full Format is made up of a BDB header and a BDB body.

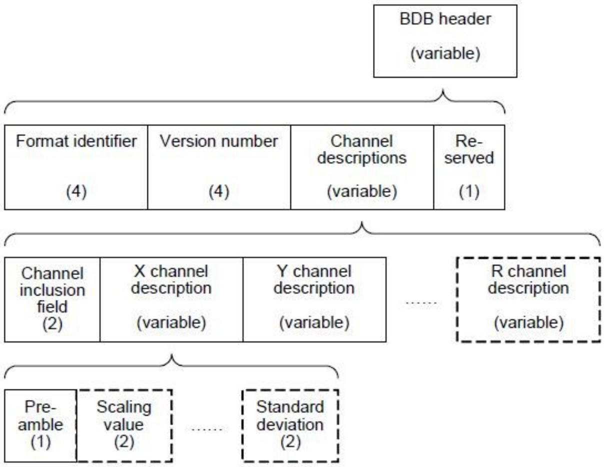

Figure 1. BDB Header of ISO/IEC 19794-7 Full Format

The BDB header, figure 1, mainly contains information about the channels which are included and their characteristics (maximum and minimum values, scaling values, etc…). Most of the information depends on the input devices used for capturing the signature.

The BDB Body, figure 2, contains a sequence of sample points where the values of the different channels captured by the input device in each sample time are stored. These values shall be encoded in 2 octets, except switch state, which is encoded in 1 octet.

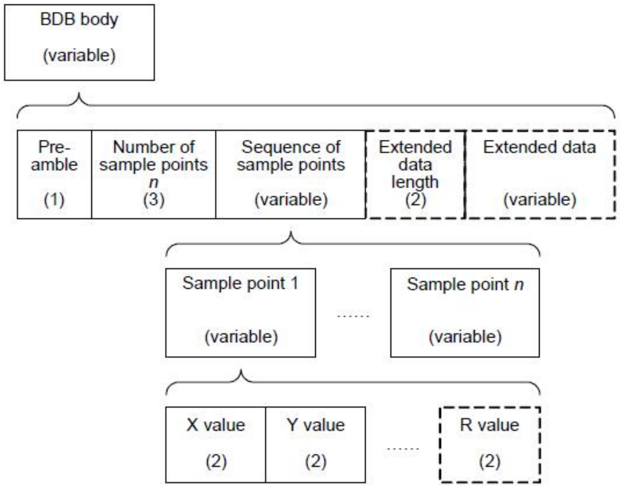

Figure 2. Sequence of Sample Points of ISO/IEC 19794-7 Full Format

This format also allows the inclusion of extended data, but it does not define its structure.

B. Compact Format

Unlike Full Format, Compact Format has only a BDB Body and it does not allow extended data. The BDB Body, as in Full Format, contains a sequence of sample points for all the channels captured by the signature input device, but, instead of encoding them in 2 octets, those values shall be encoded in just 1 octet, figure 3.

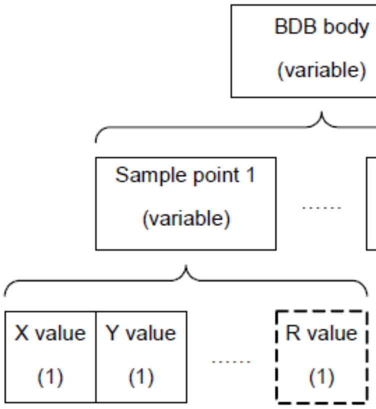

Figure 3. BDB body of ISO/IEC 19794-7 Compact Format

III. 19794 PART 11: SIGNATURE/SIGN PROCESSED DYNAMIC DATA

Part 11 of 19794 [2] project is currently under development, being in Committee Draft version.

This part describes itself as a compression of part 7, compact enough to be stored in smart cards and others tokens. This compression is based on the segmentation of the signature into components of pen-strokes and pressure-strokes. Instead of storing all the sample values of every stroke, a summary of each stroke will be stored. Strokes summarize all information using some values (min, max, mean, etc.) of x and y axes, velocity, acceleration, pressure and time.

19794-11 defines two different segmentations, one for x and y channels, and other for pressure channel.

A BDB conformance with part 11 is made up of a BDB Header and a BDB Body.

The BDB Header, figure 4, stores the information about the data interchange format and its version as well as the length, in bytes, of the BDB, the number of representations (number of views) stored within this BDB, and information about the capture device.

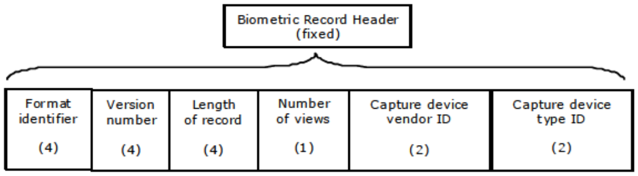

Figure 4. BDB Header of ISO/IEC 19794-11

The BDB Body, figure 5, stores information about the capture device (as scaling values and sample resolution), and it also allows the inclusion of Extended data, again without defining its structure.

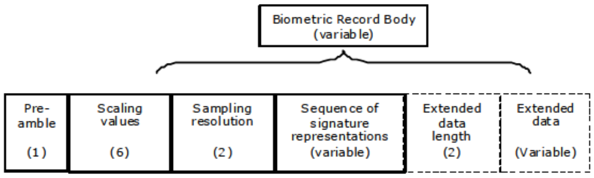

Figure 5. BDB Body of ISO/IEC 19794-11

The BDB Body also includes the representations of the signatures. These representations are made of a sequence of pen and pressure strokes together with some overall features.

A. Pen-Strokes

X and y segmentation is based on pen-strokes. A pen-stroke is defined as the movement of a pen between two singular points. These singular points can be a pen-down, a pen-up or a turning point. A turning point is defined as a sample point where either x, y or both axes values change from increasing to decreasing. Therefore, 4 types of strokes are defined:

1. Pen-down to turning point
2. Turning point to turning point
3. Turning point to pen-up
4. Pen-down to pen-up

For each pen-stroke, it will be stored its start and end (x-plane, y-plane and t values), and attributes such as the velocity (maximum, minimum and mean of vx and vy values), acceleration (maximum, minimum and mean of ax and ay values) and pressure (maximum, minimum and mean of pressure values) during the pen-stroke. Pen-stroke length and vector direction is also recorded.

B. Pressure-Strokes

Similar as occurs with pen-strokes, pressure segmentation is based on pressure-strokes, which are defined as the movement of a pen between two singular points. These singular points can be a pen-down, a pen-up or a turning point as well. A turning point is defined as a sample point where the pressure value changes from increasing to decreasing or vice versa. Again, 4 types of pressure-strokes are defined:

5. Pen-down to turning point
6. Turning point to turning point
7. Turning point to pen-up
8. Pen-down to pen-up

For each pressure-stroke, it will be stored the start and end of the pen movement (x-plane, y-plane and t values), and pressure data (end, start, maximum, minimum and mean values).

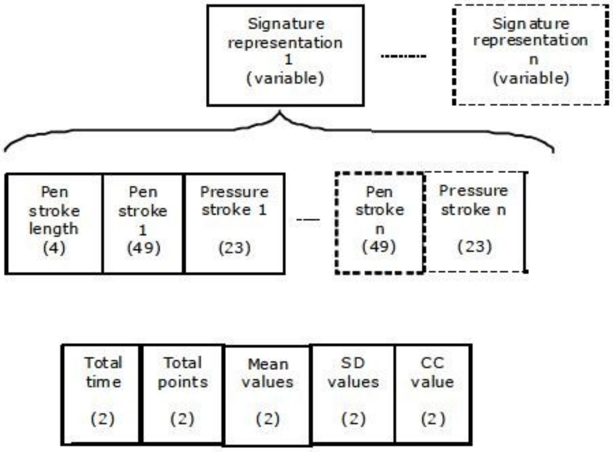

Figure 6. Signature Representation of ISO/IEC 19794-11

C. Overall Data

Together with the pen-stroke and pressure-stroke data, additional data is recorded representing overall features of the signature representation, figure 6. These data are:

1. Total time
2. Total number of sample points
3. Mean values of x, y and p channels.
4. Standard deviation of x, y and p channels.
5. Correlation coefficient.

D. Implementation 19794-11

In order to obtain the format 19794-11 from the data stored in 19794-7 Full Format, some solutions have been taken.

These solutions are:

a) Velocity and Acceleration have been calculated from X and Y channels as:

$$v_{Xi} = \frac{x_{i+1} - x_i}{t_{i+1} - t_i}$$

with $v_{X0} = 0$

$$a_{Xi} = \frac{v_{i+1} - v_i}{t_{i+1} - t_i}$$

with $a_{X0,1} = 0$

(Same for y channel)

b) Pen-strokes and pressure-strokes between turning points are force to be longer than 2 points. This simplification avoid to record strokes prompted by noise or small vibrations during the signing process.

Figure 7 shows an example of pen-strokes:

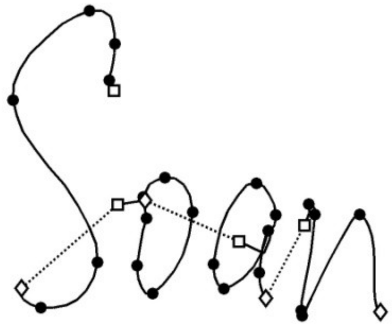

Figure 7. Example of singular points in a Signature Sample

Where dot lines indicate pen-up movement, square markers indicate pen-down event, diamond markers indicate pen-up events and round markers indicate turning points.

IV. LOSSLESS COMPRESSION METHOD

Both signature compact formats from ISO/IEC 19794, part 7 Compact and part 11, reduce the amount of data provided by the capture device. Part 7 Compact forces the signals to be in a range from 0 to 255, whereas part 11 only stores information about singular points.

We propose a new method of compression without losing any data. This method is based on the deflation algorithm as a variation of LZ77 [3], used by compression applications as zip and gzip.

LZ77 algorithm exploits the repetitions of a string, replacing such repetitions by a pointer to the previous string. In order to improve the compression ratio achieved when applying LZ77 to part 7 Full Format, three different options have been tested, named C0, C1 and C2.

As it has been told before, a BDB conformance with 19794 part 7 Full Format is made out of a BDB Header and a BDB Body. LZ77 algorithm will be applied only to the sequence of sample points, keeping BDB Header and the other fields of the BDB Body without compression.

A. C0

LZ77 algorithm will be applied to the sequence of sample points as it is defined in the standard, this is, as, a sequence of points, each point containing the values from all the channels included.

B. C1

Instead of storing the sequence of samples in the way it is defined in part 7 Full Format, it could be a better option to store each channel separately, link all channels together and finally compress the resulting data structure. This could improve the LZ77 algorithm performance.

C. C2

As a last option, it is proposed to store separately not the values of each channel but the difference between consecutive samples of the same channel. This way, the first value of a channel will be stored as it is, and the following values will be the difference with the previous one. This values will have to be stored as a signed integer instead an unsigned. After calculating the differences, the results of each channel will be linked together and then compressed.

V. SIGNATURE DATABASES

Evaluation of the size of a BDB for these signature formats has been carried out using both MCyT-Signature-Database Corpus [6] and SVC2004 Database [7]. Both of them are public available.

A. MCyT Signature Database

MCyT database is constituted by 100 different users. Each user produced 25 genuine signatures. 25 skilled forgeries were also captured. These skilled forgeries were produced by the 5 subsequent users, who practiced it until they felt confident. To capture the signatures of the database, a Wacom Intous A6 USB graphic tablet was used.

Users captured in this database are mainly Spanish writers.

B. SVC2004 Signature Database

SVC2004 database is constituted by 40 different users. Each user produced 20 genuine signatures. 20 skilled forgeries were also captured. Users where asked not to use their real signatures. They made out a new signature and practiced it until they got confident on it. To capture the signatures of the database, a Wacom Intous A6 graphic tablet was used.

Users captured in this database are occidental or oriental. From these 40 different users, 24 are occidental users, whereas 16 are oriental users. Unlike Spanish signatures, English users are used to write their name or initials without pictorial strokes. Chinese signatures are completely different than occidental signatures and are based on shorts strokes.

VI. ANALYSIS OF BDB SIZE

For all data formats analyzed, in order to find out the average BDB size and other characteristics, as mean number of points or pen and pressure strokes, only genuine signatures from both databases have been used.

Due to ISO/IEC 19794-11 data format contains only information about x and y axis and pressure, both Full and Compact formats included in ISO/IEC 19794-7 have been made with these 3 channels plus time and switch channels. Data from tilts as azimuth and elevation, which are included in both databases, has been omitted.

Figure 8 shows the average BDB size of all data formats studied, whereas Figure 9 shows the compression ratio.

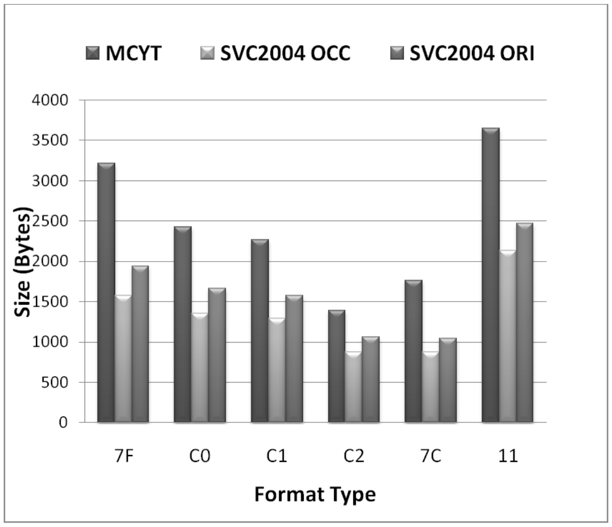

Figure 8. BDB Size of all data formats studied

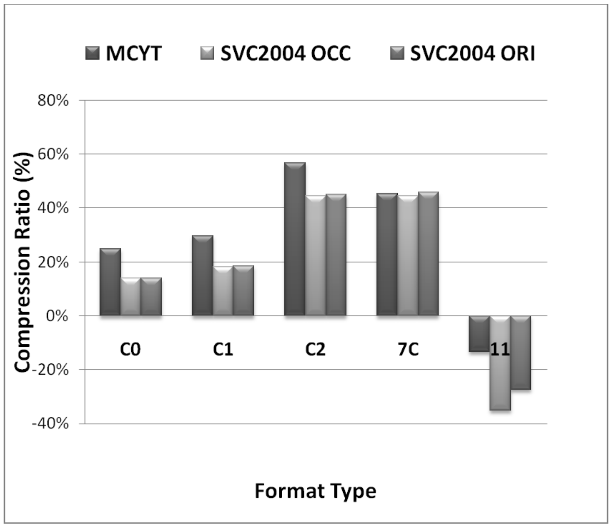

Figure 9. Compression Ratio of compact data formats

MCyT signatures have an average BDB size around 3.1 Kbytes for ISO/IEC 19794-7 Full Format, whereas ISO/IEC 19794-11 has an average BDB size around 3.5 Kbytes. This shows that ISO/IEC 19794-11 does not compress part 7 Full Format as it was expected, getting a higher average BDB size of 3.5 Kbyte, which means a compression ratio around -13% for MCyT database. The reason of this seems to be that there is too much information recorded on each singular point (as 6 velocities, 6 accelerations, etc.). Some information is also duplicated for consecutive strokes, as each stroke records the initial and end point information. As consecutive strokes share information about their starting and ending points, this duplicated information should be removed. This size raise could be avoided using a low pass filter before velocity is calculated and d defining turning points as a zero-crossing velocity event. The use of this low pass filter could mean a lower number of singular points, and therefore, a lower BDB size. If the inclusion of that filter wants to be avoided, the use of a more complex velocity calculation formula, not affected by small variations between sample points, could be useful.

ISO/IEC 19795-7 Compact Format has a compression ratio around 45%, due to the transformation of the data value size from 2 bytes to only 1 byte. The average BDB Size for this compact data format is 1.7 Kbytes.

Looking at the proposed lossless compression, it can be noticed than each approximation has a better compression ratio, achieving a ratio of 57% for C2, being the associated average BDB size equal 1.3 Kbytes.

With SVC2004 database, similar results are obtained, although the lossless compression method proposed gets a lower compression ratio, 45% for C2.

TABLE I. SAMPLE POINTS AND SIGNATURE TIME FOR THE THREE KIND OF USERS

	Number of Sample Points	Signature Time (s)
MCyT	351	3,51
SVC2004 OCC	168	2,07
SVC2004 ORI	208	2,85

Table 1 shows the average number of sample points and the signature total time for each kind of user (MCyT, SVC2004 occidental and SVC2004 oriental users) using ISO/IEC 19794-7 formats. MCyT users have longer signature than occidental and oriental SVC2004 users, having occidental SCV2004 users the lowest signature length. MCyT signatures have more sample points and take longer total time than SVC2004 signatures (both occidental and oriental), as they are used to write their names adding pictorial strokes. This bigger number of sample points it is also due to the fact that MCyT database include pen-up movement. Oriental SVC2004 signatures also have more sample points and take longer total time than occidental, as they have more strokes and longer signature than occidental SVC2004 signers.

Such a big difference between occidental users from MCyT and SVC2004 databases, in both number of sample points and signature total time, could also be caused by the fact that SVC2004 database has not real user signatures. As users had to create a new signature, they may have chosen easy and simple ones in order to get used to do them quickly. This fact can make the resulting signatures less stable than real ones, something that would entail worse algorithm performance as it will be shown in next section.

The results obtained using ISO/IEC 19794-11 are shown in Figure 10. The figure shows the mean number of the different singular point types defined in ISO/IEC 19794-11. There are some differences worthy of being highlighted between kinds of user. Oriental SVC2004 users have the biggest number of pen-up and pen-down events (a mean of 7.6 pen-ups and pen downs), being the difference with MCyT users (a mean of 6.5 pen-ups and pen downs) not too significant. Occidental SVC2004 users have only a mean of 4.5 pen-ups and pen downs

Regarding pen and pressure turning points, the difference between MCyT users and both occidental and oriental SVC2004 users is really significant, meaning that MCyT signatures are more complex in terms of x and y axis variations as well as pressure variations. Again, this big difference between occidental users from MCyT and SVC2004 databases may be caused by the fact that SVC2004 has no real signatures.

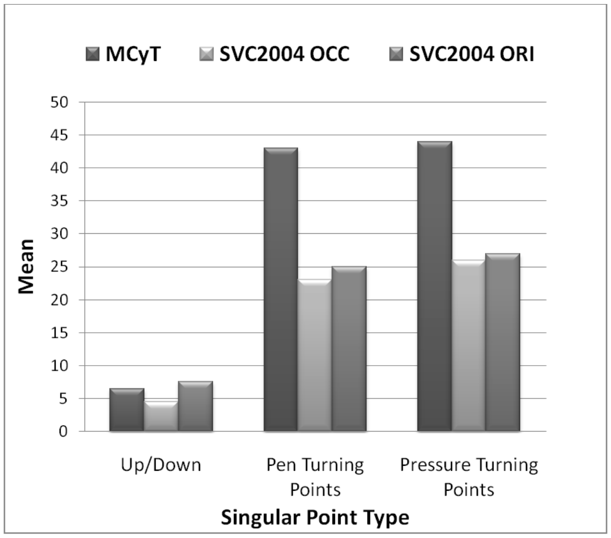

Figure 10. Number of singular point type for the three kinds of user

VII. PERFORMANCE OF THE THREE SIGNATURES FORMATS

In order to evaluate the impact on the performance of the compact formats under test, two different algorithms have been used. First of them is Gaussian Mixture Modelling (GMM) [4], which is based on the extracted features from signature signals. Second one is Dynamic Time Warping (DTW) [5][8], based on time alignment between the biometric sample and the biometric reference template.

Better performance than those which are shown in this paper has been obtained using both algorithms together [8]. However, for the purpose of this study, in order to analyze the impact of the compact data formats in two specific kinds of algorithms, DTW and GMM will be used separately.

This section analyzes whether the information lost on both ISO/IEC compact formats (Part 7 Compact Format and Part 11) has an impact on the performance of the Signature Verification Algorithms or not.

In order to use the standard ISO/IEC 19794-7 Compact Format on the algorithms used to perform the test, no additional preprocessing has been needed.

Unlike previous case, ISO/IEC Part 11 requires the addition of a new step within the preprocessing. From the information stored in each singular point (x and y position, pressure, time and init vector direction), the corresponding four temporal signals have been interpolated.

This interpolation has been carried out using Mathlab Interpolation ToolBox [9] [10]. Cubic spline interpolation has been used on X and Y channels whereas Piecewise Cubic Hermite interpolation has been used for Pressure Channel.

A. *GMM*

Figure 11 shows GMM algorithm results for both databases, the three ISO/IEC data formats under study, and the three different kind of users.

MCyT users and occidental SVC2004 users got worse algorithm performance of compact data formats ISO/IEC 19794-7 Compact Format and ISO/IEC 19794-11, therefore, some relevant information has been removed during the data compression process.

On the contrary, oriental SVC2004 users got better performance of the compact data formats. It seems that the information lost during the data compression process removes irrelevant information, improving the GMM algorithm performance.

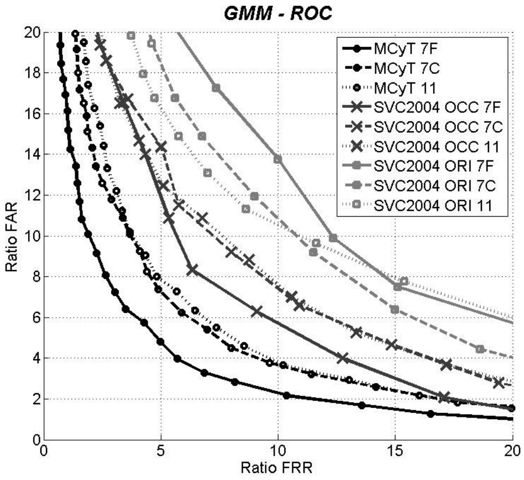

Figure 11. GMM algorithm results

On the other hand, the overall performance obtained with SVC2004 database is significantly worse than the one obtained with MCyT database. This fact can be due to MCyT database is made up of real signatures, whereas SVC database is made up of new-created signatures. As it has been noticed before, MCyT users have more complex signatures, and therefore, it is more difficult to forge them. Also, MCyT database has information about pen-up movements and more users captured. Further research should be done on these matters.

For SVC2004 database, occidental users got also better performance than oriental ones. Due to the specific characteristics of oriental users, the GMM algorithm, which was tuned for MCyT users, fit better the occidental SVC2004 signatures.

B. *DTW*

DTW algorithm results are shown in Figure 12. There is not a significant difference between ISO/IEC 19794-7 Full Format and Compact Format on MCyT Database, showing that the conversion from 2 bytes to 1 byte does not mean a lost of relevant information for this DTW algorithm.

On the other hand, ISO/IEC 19794-11 has a worse performance, due to the re-construction by means of interpolation. The lack of pen-up movements within ISO/IEC 19794-11 could explain this lower performance. Again, further research should be done.

Regarding occidental SVC2004 users, the performance is again worse than the one obtained using MCyT database. The reason again could be that SVC2004 database is made up of non-real signatures.

Performance in the case of oriental SVC2004 users is again the worst, meaning that DTW algorithm does not fit with oriental signature characteristics. In this case, data compression does not improve the performance of the algorithm.

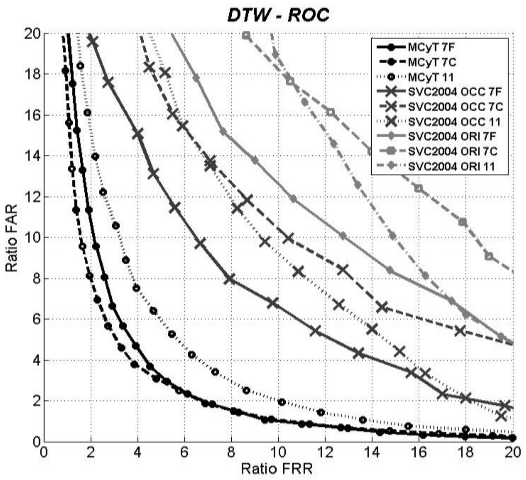

Figure 12. DTW algorithm results

VIII. CONCLUSIONS

Due to the growing of Signature Automatic Verification Systems possible applications, and the increase of the number of different types of capture devices, data formats standardization has become a real need to guarantee interoperability.

In this paper, three signature data formats proposed by ISO/IEC have been studied.

As a result of this study, it has been shown that ISO/IEC 19794-7 Compact Format reduces 45% BDB data size, whereas ISO/IEC 19794-11 does not imply any compression; otherwise, it increases the BDB data size. The use of a low pass filter or more complex velocity calculations has been proposed as an option to improve the results.

It has been also proved that data lossless compression, as LZ77 algorithm with the proposed ISO/IEC specification changes (C2), is a good alternative to get 57% compression ratio without losing any data information.

An analysis of three different kinds of user (Spanish, English and Chinese writers) has been done as well. Nationality affects some signature characteristics as number of strokes, total time writing and complexity. These characteristics have an impact on the size of the BDB as well as on the complexity of the signatures. The occidental signatures collected in SVC2004 are the simplest ones. This can be explained considering that this database has, instead of real signatures, new signatures created just for the purpose of collecting this database.

Regarding data information being lost within the compact formats, it has been proved that the reduction from 2 bytes to 1 byte specified in ISO/19794-7 Compact Format, has an impact on the performance of GMM and DTW algorithms for MCyT database. In the case of ISO/19794-11 data format, it has been also shown that with the information stored within the strokes (x and y position, pressure and time), it is possible to recreate the original signals captured by the input devices. However, lower performance on the signature algorithms tested is obtained.

For SVC2004 database, the overall performance of both algorithms is worse than the one of MCyT database. It could be due to SVC2004 has no real signatures and pen-up movements are not captured. MCyT users have more complex signatures, and therefore, it is more difficult to forge them. Further research should be done to confirm all these hypotheses. It should be notice that the number of occidental and oriental users in SVC2004 are too small, therefore, the results obtained using these database subgroups have no significant statistical values.

Finally, it has also been shown that algorithms can be sensitive to user nationalities and also to the database quality. User nationalities determine some signature characteristics, which can change the algorithm performance. The fact that SVC2004 is made up of no real signatures, makes the comparison between both databases non relevant.

ACKNOWLEDGMENT

Authors would like to thank J. Ortega-Garcia and J. Fierrez-Aguilar for the provision of the MCyT Signature Database.

This work has been funded by the Spanish Ministry of Science and Education (TEC2006-12365).

REFERENCES

[1] Information Technology—Biometric Data Interchange Formats—Part 7: Signature/Sign Time Series Data, ISO Standard ISO/IEC FCD 19794-7, 2006.

[2] Information Technology—Biometric Data Interchange Formats—Part 11: Signature/Sign Processed Dynamic Data, ISO Standard ISO/IEC CD 19794-11, 2009.

[3] Ziv J., Lempel A., "*A Universal Algorithm for Sequential Data Compression*", IEEE Transactions on Information Theory", Vol. 23, No. 3, pp. 337-343.

[4] J. Richiardi, A. Drygajlo "*Gaussian Mixture Models for On-line Signature Verification*". Proc. ACM Multimedia, 2003.

[5] Y. Sato and K. Kogure, "*On-Line Signature Verification Based on Shape, Motion, and Writing Pressure*" IEEE Proc. Sixth Int'l Conf. on Pattern Recognition, pp. 823-826, 1982.

[6] J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J.-J. Igarza, C. Vivaracho, D. Escudero and Q.-I. Moro, "*MCYT baseline corpus: a bimodal biometric database*", IEEE Proc.-Vis. Image Signal Process., Vol. 150, No. 6, December 2003.

[7] D.-Y. Yeung, H. Chang, Y. Xiong, S. George, R. Kashi, T. Matsumoto, and G. Rigoll, "*SVC2004: First international signature verification competition*" (Lecture Notes in Computer Science 3072), in ICBA 2004, D. Zhang and A. K. Jain, Eds. Berlin, Germany: Springer-Verlag, 2004, pp. 16–22.

[8] Miguel-Hurtado, O.; Mengibar-Pozo, L.; Pacut, A., "*A new algorithm for signature verification system based on DTW and GMM*" Security Technology, 2008. ICCST 2008. 42nd Annual IEEE International Carnahan Conference on , vol., no., pp.206-213, 13-16 Oct. 2008

[9] Interpolation ToolBox, Mathlab, www.mathworks.com

[10] de Boor, C., *A Practical Guide to Splines*, Springer-Verlag, 1978.