# Web-based Dynamic Similarity Distance Tool

Nur Atikah Arbain[1], Mohd Sanusi Azmi[1], Azah Kamilah Muda[1], Amirul Ramzani Radzid[1], Noor Azilah Muda[1] and Radzuan Nordin[2]
*[1]Faculty of Information and Communication Technology,*
*[2]Centre for Languages and Human Development,*
*Universiti Teknikal Malaysia Melaka, MALAYSIA.*
*sanusi@utem.edu.my*

***Abstract*— Similarity or distance measures is a well-known method and commonly used for calculating the distance between two samples of a dataset. Basically, the distance between the dataset samples is an important theory in multivariate analysis research. This paper proposes a tool that provides seven common distance methods that can be used in various research area. This tool is a web-based application which can be accessed through the internet browser. The objective of this tool is to introduce a web-based similarity distance application for many analysis and research purposes. Besides, a ranking method based on the Mean Average Precision is also implemented in this tool in order to increase the classification accuracies. This tool can process features that contain numerical values from any type of dataset.**

***Index Terms*— Accuracy; Similarity Distance; Ranking; Web-based.**

## I. Introduction

The similarity distance methods play such important roles in solving many pattern recognition problems such as data clustering, data classification and data retrieval. The concept of the distance is the most significant basis for classification. In an unsupervised learning work, there is no validation mechanism by resources of objects with known clusters. But, in supervised learning standard distances, it rarely can lead to generating significant results [1]. According to [2], the binary similarity and the distance measure approaches have been introduced in many fields such as biology [3], geology [4], image retrieval [5], chemistry [6] and taxonomy [7]. Besides that, similarity distances have also been applied actively in biometric research such as fingerprint and handwriting recognitions. The similarity distance has been implemented in [8] to handle different types of queries for content-based image retrieval. Meanwhile, the similarity distance such as Euclidean has been applied in [9] for image recognition-related problems and a ranking method using Mean Average Precision [10] was applied to increase the classification accuracies. [10] has concluded that the similarity distance is vital for the ranking method that is based on the Mean Average Precision.

In this paper, several common distance methods that are Euclidean, Manhattan, Chebyshev, Canberra, Sorenson, Angular Separation and Minkowski are emphasized. These similarity distance methods are widely used by the researchers to calculate the distance between two samples of datasets. Therefore, this paper introduces a web-based similarity distance tool for diverse analysis and research purposes. This paper is organized as follows. In Section 2, the related work is discussed. The problems in existing software for similarity distance are discussed in Section 3. Then, the proposed method is analyzed in Section 4. Next, the proposed web-based tool interfaces are presented in Section 5. Finally, Section 6 concludes this paper.

## II. Related Work

### *A. Dynamic Similarity Distance with Mean Average Precision Tool*

SD Tool is an initial work that is developed by combining the similarity distance and the ranking method [11]. This tool provides seven similarity distance methods which are commonly used the previous research. The objective of this tool is to provide distance measurements capabilities in the analysis and recognition field. It is developed to assist the researcher in selecting the most suitable tool for each dataset they used. Following are the guidelines of the SD Tool.

- SD Tool is only applicable to process Comma Separated Value (CSV) files.
- SD Tool requires two types of data namely the test and the train (model) data.
- SD Tool only accepts numerical values.
- SD Tool can support a dynamic number of features where the column size of test and train must be equal. Otherwise, the calculation between two samples of the dataset cannot be processed due to unbalance size of features column.

### *B. Other Works*

The similarity distance algorithms have been applied in various fields. A new Euclidean distance called IMage Euclidean Distance (IMED) was proposed by [12] in 2005. IMED is slightly different compared to the traditional Euclidean distance. IMED focuses on the spatial relationships of pixels which makes it robust to the small disturbance of images [12]. The IMED can be embedded into any image recognition algorithm easily because most of the algorithms are developed based on the Euclidean distance such as Radial Basis Function Support Vector Machines (RBF SVMs) [13], Principal Component Analysis (PCA) and Bayesian similarity. The Euclidean similarity distance has been applied in [14] by embedding the generalized similarity measure (GSM) into the Euclidean distance as a catalyst and is applied in existing approximate nearest neighbour search (ANNS) methods. The author has also stated that GSM is a similarity measure which has been presented by a linear combination of the Mahalanobis distance and the bilinear similarity to obtain the recognition accuracies for face recognitions [13]. The mean average precision technique has been applied by [15] to increase the result of classification accuracies.

## III. Problems In Existing Software For Similarity Distance

The similarity method can be used to calculate the distance between two samples of datasets. Nowadays, the existing software such as Excel or MATLAB can be used to calculate the similarity distance as well. However, the input for similarity distance calculation is manually defined. The process of preparing the input requires knowledge to convert the mathematical formula to the distance algorithm. The software normally needs the researchers to insert the distance formula manually and transforms the formula to mathematical procedure before it can be used. However, unlike Excel or MATLAB, the Dynamic Similarity Distance with Mean Average Precision Tool (SD Tool) is developed to focus on the distance method and the ranking calculation that is based on the Mean Average Precision (MAP) [10] only. The user can choose to calculate the distance using any distance methods from the selection menu. However, the existing tool like SD Tool, for example, was built on a standalone platform which requires the users to download it first before it can be used. Among the problems identified with the existing software for similarity distance method is as follow:

- Spreadsheet such as Excel and mathematic software such as MATLAB requires the user to write or program the formula (similarity distance and mean average precision). The error might occur.
- Software developed from previous research [11] required the user to download. The dedicated computer also required Java Runtime Environment (JRE).

## IV. Proposed WebSDT

The Web-based Dynamic Similarity Distance Tools (WebSDT) is developed on a web-based platform extending the previous work of Dynamic Similarity Distance with Mean Average Precision Tool (SD Tool) [11]. The proposed web-based tool introduces simple and user-friendly interfaces that can aid the users to easily understand and use the web-based tool. By retaining the same functions as in the SD Tool, the proposed web-based tool is equipped with additional features that are beneficial to the users. The most suitable distance method for a dataset is depending on the dataset itself such as the characteristic of the dataset.

The similarity distance algorithms are commonly used to match real time images with model images. Currently, there are a few well-known similarity distance algorithms such as i.Euclidean, ii.Manhattan, iii.Chebyshev, iv.Canberra, v.Sorenson, vi.Angular Separation and vii.Minkowski. However, these algorithms require knowledge to convert the mathematical equations into datasheets like Excel or mathematical software such as MATLAB or Mathematica. The WebSDT is developed to assist the users in applying any similarity distance algorithms to process CSV formatted input data. The WebSDT provides different mechanisms for processing the input data based on the types of similarity distance algorithms. It will calculate the distance from the input data file then lists the similarity results from the smallest value to the biggest value or vice versa. Furthermore, the WebSDT has an additional function that is the Mean Average Precision (MAP) method to rank the similarity results by weightage. Therefore, the matching or recognition results will be improved and significantly accepted.

The WebSDT is built based on the web-based platform and Java Server Faces is used as the language programming. Previously, the SD Tool was built on the standalone platform which only applicable to be used with desktops. Since the current software was not developed for analysis and research purpose thus, the proposed web-based tool is developed to overcome the current limitations of the SD Tool. Besides that, the proposed web-based tool also offers hassle-free application for classification, recognition and ranking as the users can do the tasks online.

### A. The Workflows of the Proposed WebSDT

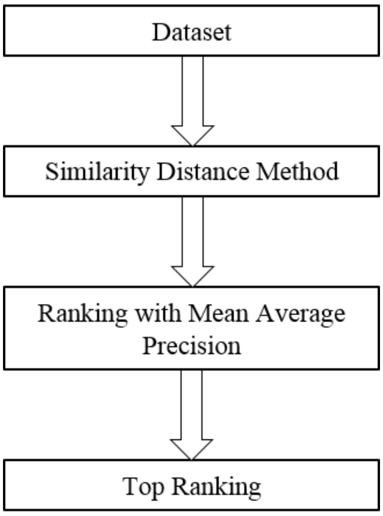

Dataset
Similarity Distance Method
Ranking with Mean Average Precision
Top Ranking

Figure 1: The process of proposed WebSDT

Figure 1 shows the workflow process of the proposed WebSDT. In general, two samples of datasets which are the test and the train data will be measured using the similarity distance method. The measured distances between the test and the train data will then be used in the ranking method. The ranking method is used to find the most suitable distance measurement based on the weight to determine whether the objects are naturally relevant to be in the same groups or clusters. This is relevant to the statement in [16] as the author has stated that the distance measurement plays an important part to obtain good results.

### B. Process of Ranking Method

The Mean Average Precision (MAP) which has been introduced by [10] is used in the WebSDT to rank the result from the similarity distance measurements. The processes of the ranking method in the WebSDT are described as follows:

1. The handwriting digit of BANGLA dataset [17] is used. There are 10 folders for training images which contained 19,392 images while another 10 folders for testing images contained 4,000 images. Each of the folders was represented as digit class of BANGLA where BANGLA dataset had been classified based on folder name from 000 until 009. In this experiment, we select randomly 20 test images of BANGLA dataset and compared them with 200 training images of BANGLA dataset.
2. The distance between the two samples dataset is

calculated using one of the distance method (the Euclidean method is demonstrated in this for example). The Euclidean distance formula can be referred in [18].

3. The distances between the data are compared and will be arranged in ascending order or descending order. The smallest distance between test and training data is matching for each calculation of Euclidean distance method. In this example, the Euclidean results is arranged in ascending order.
4. However, the accuracy of matched data between the test and the train data shows non-promising results as shown in Figure 2. The figure illustrates the example image from class bangla1. However, the result shows the nearest class is from class bangla9 (position 1) after the Euclidean distance method is applied.
5. To increase the accuracy result, the ranking method is applied by using the Mean Average Precision (MAP) algorithm. The MAP algorithm is used to get the rank positions, for example, 5, 10 and 15. The detail description of the MAP formula can be found in [10]. Figure 3 illustrates the ranking calculations using MAP. Before the MAP algorithm is applied, the accuracy result shows the image from class bangla9. However, after applying the algorithm, the accuracy result is pointed to the image from class bangla1 as shown in Figure 3. In the MAP calculation, the ranking is based on the highest weight. In this example, we use 10 ranks as the ranking.

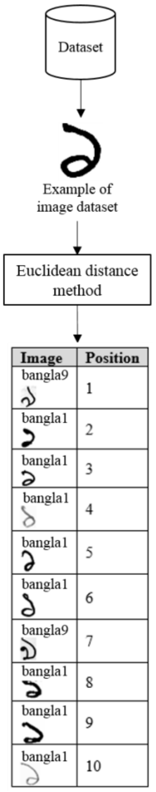


Figure 2: Result using the Euclidean distance method

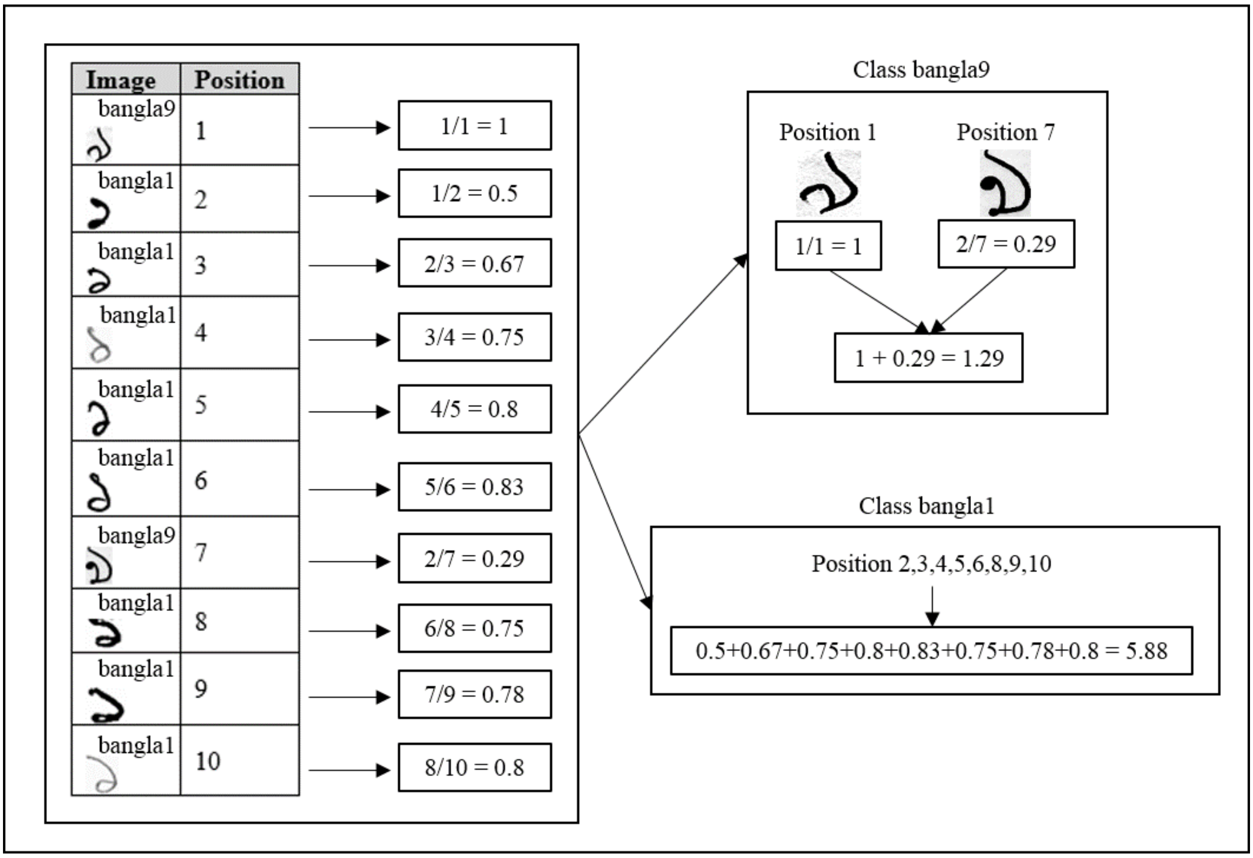
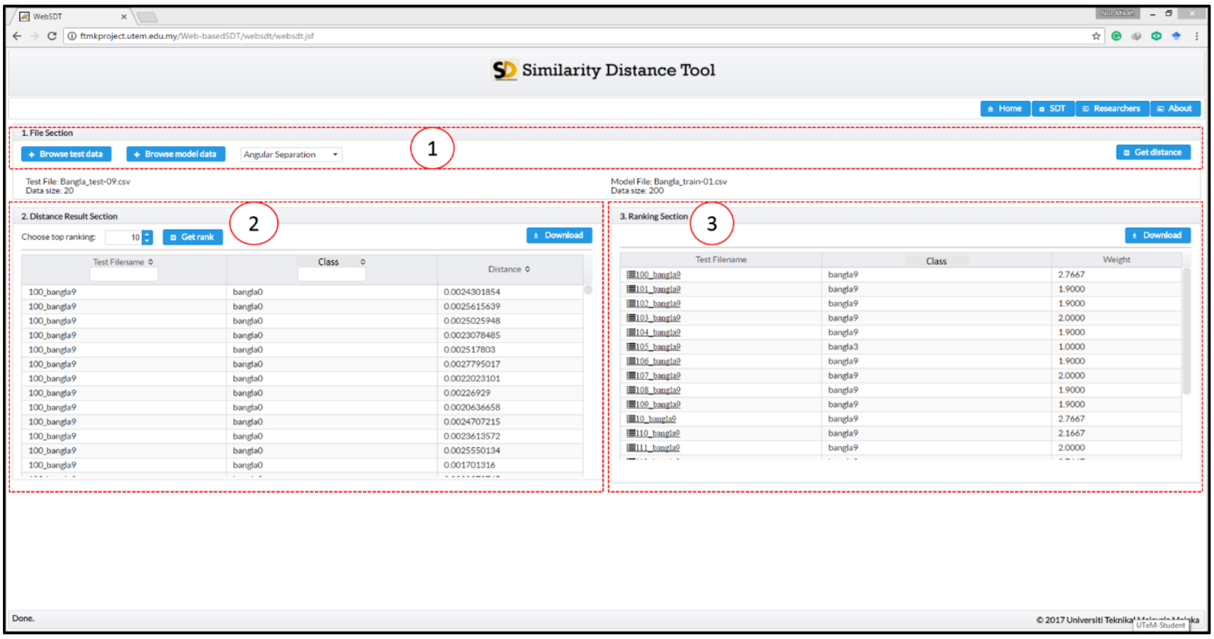


Figure 3: MAP calculation for top 10 ranks

## V. PROPOSED WEBSDT INTERFACE



1. File Section

Test File: Bangla_test-09.csv
Data size: 20

Model File: Bangla_train-01.csv
Data size: 200

2. Distance Result Section

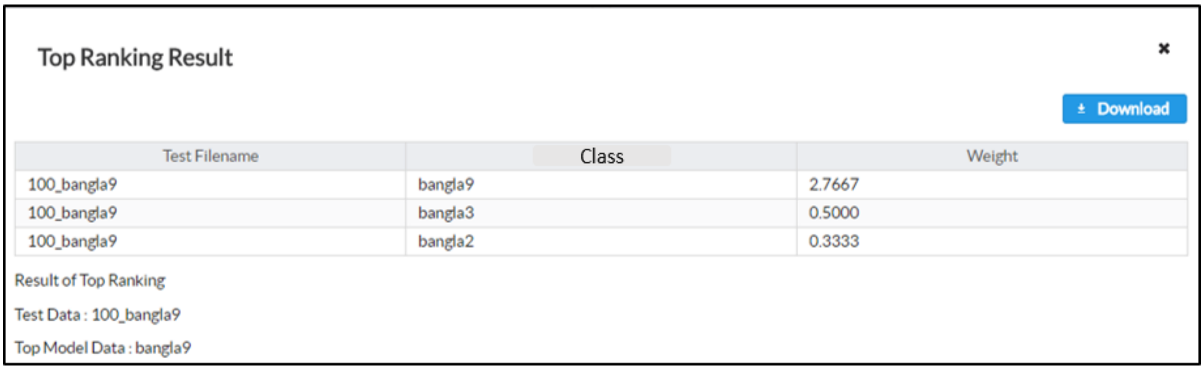
Choose top ranking: 10

| Test Filename | Class | Distance |
|---|---|---|
| 100_bangla9 | bangla0 | 0.0024301854 |
| 100_bangla9 | bangla0 | 0.0025615639 |
| 100_bangla9 | bangla0 | 0.0025025948 |
| 100_bangla9 | bangla0 | 0.0023078485 |
| 100_bangla9 | bangla0 | 0.002517803 |
| 100_bangla9 | bangla0 | 0.0027795017 |
| 100_bangla9 | bangla0 | 0.0022023101 |
| 100_bangla9 | bangla0 | 0.00226929 |
| 100_bangla9 | bangla0 | 0.0020636658 |
| 100_bangla9 | bangla0 | 0.0024707215 |
| 100_bangla9 | bangla0 | 0.0023613572 |
| 100_bangla9 | bangla0 | 0.0025550134 |
| 100_bangla9 | bangla0 | 0.001701316 |

3. Ranking Section

| Test Filename | Class | Weight |
|---|---|---|
| 100_bangla9 | bangla9 | 2.7667 |
| 101_bangla9 | bangla9 | 1.9000 |
| 102_bangla9 | bangla9 | 1.9000 |
| 103_bangla9 | bangla9 | 2.0000 |
| 104_bangla9 | bangla9 | 1.9000 |
| 105_bangla9 | bangla3 | 1.0000 |
| 106_bangla9 | bangla9 | 1.9000 |
| 107_bangla9 | bangla9 | 2.0000 |
| 108_bangla9 | bangla9 | 1.9000 |
| 109_bangla9 | bangla9 | 1.9000 |
| 10_bangla9 | bangla9 | 2.7667 |
| 110_bangla9 | bangla9 | 2.1667 |
| 111_bangla9 | bangla9 | 2.0000 |

Figure 4: Web-based Similarity Distance Tool (WebSDT)



Top Ranking Result

| Test Filename | Class | Weight |
|---|---|---|
| 100_bangla9 | bangla9 | 2.7667 |
| 100_bangla9 | bangla3 | 0.5000 |
| 100_bangla9 | bangla2 | 0.3333 |

Result of Top Ranking

Test Data : 100_bangla9

Top Model Data : bangla9

Figure 5: Selected ranking result based on MAP method

Figure 4 illustrates the web-based similarity distance tool (WebSDT) interface. There are three sections in WebSDT interface which are File Section (Figure 4 (1)), Distance Result Section (Figure 4 (2)) and Ranking Section (Figure 4 (3)). In File Section, the user needs to input test and train data. Then, select any distance method in drop-down list menu where there are seven types of similarity distance are provided. Next, click “Get Distance” button for calculating the distance between test and train data. The distance result of the selected distance method will be displayed in Distance Result Section. To apply ranking method, the user can choose any number of top ranking, for example, 10 as top ranking. Then, click “Get rank” to process the ranking calculation where the Mean Average Precision method is applied to calculate the ranking. The overall ranking results will be displayed in Ranking Section. In Figure 4 (3), the ranking results for each test data will be displayed along with the class of dataset. The class of dataset is determined by the result of weight. In the MAP calculation, the ranking is based on the highest weight. Figure 5 shows the detail of ranking result based on selected test data. The user can click on the test filename (in Figure 4 (3)) for showing the detail of ranking result for selected test file data. The explanation in Figure 5 can be referred in Section 4.2. (Note: the ranking result is not same as in Section 4.2 due to different distance method is used).

## VI. CONCLUSION

This paper introduces the proposed web-based tool which has been extended from the previous work, the Dynamic Similarity Distance with Mean Average Precision Tool. The proposed tool is developed to overcome the current limitation where the researchers need to download the current existing software which has been developed as a stand-alone tool. The current software also provides limited similarity distance method which can be used for certain cases only. The proposed web-based tool however provides various similarity distances method along with the ranking method which is the Mean Average Precision. As it offers various methods of similarity distance, the tool will definitely can be used diversely. This web-based tool is built as a web-based application which provides the users free access online [19]. The WebSDT also provides the Application Programming Interface (API) for users or industries to embed the tool in their products or machines. Further research will be focused on developing this tool in application mobile for ease use.

## ACKNOWLEDGMENT

The authors would like to thank to the Ministry of Education for funding the study through grants:

- PJP/2016/PBPI/HI4/S01476
- PJP/2016/FTMK/HI4/S01477

Thanks are also extended to UTeM Zamalah Scheme for providing the scholarship. Besides, thank you to the Universiti Teknikal Malaysia Melaka and Faculty of Information and Communication Technology for providing the excellent research facilities.

## REFERENCES

[1] C. Weihs and G. Szepannek, "Distances in classification," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5633 LNAI, no. 1, pp. 1–12, 2009.

[2] C. Seung-Seok, C. Sung-Hyuk, and C. C. Tappert, "A Survey of Binary Similarity and Distance Measures.," *J. Syst. Cybern. Informatics*, vol. 8, no. 1, pp. 43–48, 2010.

[3] Z. Hubalek, "Coefficients of Association And Similarity, Based On Binary (Presence-Absence) Data: An Evaluation," *Biol. Rev.*, vol. 57, no. 4, pp. 669–689, 1982.

[4] M. E. Hohn, "Binary coefficients: A theoretical and empirical study," *J. Int. Assoc. Math. Geol.*, vol. 8, no. 2, pp. 137–150, 1976.

[5] A. Mohanan and S. Raju, "A Survey on Different Relevance Feedback Techniques in Content Based Image Retrieval," *Int. Res. J. Eng. Technol.*, vol. 4, no. 2, pp. 582–585, 2017.

[6] P. Willett, J. M. Barnard, and G. M. Downs, "Chemical Similarity Searching," *J. Chem. Inf. Comput. Sci.*, vol. 38, no. 6, pp. 983–996, 1998.

[7] S.-H. Cha, "Taxonomy of Nominal Type Histogram Distance Measures," *Proc. Am. Conf. Appl. Math.*, no. 2, pp. 325–330, 2008.

[8] M. Arevalillo-Herráez, J. Domingo, and F. J. Ferri, "Combining similarity measures in content-based image retrieval," *Pattern Recognit. Lett.*, vol. 29, no. 16, pp. 2174–2181, 2008.

[9] M. S. Azmi, "Fitur Baharu Dari Kombinasi Geometri Segitiga Dan Pengezonan Untuk Paleografi Jawi Digital," Doctoral dissertation, Universiti Kebangsaan Malaysia, 2013.

[10] J. Kalofolias, "Towards Mean Average Precision," in *Symposium on Natural Language Processing*, 2015, pp. 1–6.

[11] "Dynamic Similarity Distance With Mean Average Precision Tool." [Online].Available: https://www.dropbox.com/sh/wm8umplys8xjmtq/AAAaznDfaZt7_1JkwPb0auyja?dl=0.

[12] L. Wang, Y. Zhang, and J. Feng, "On the Euclidean Distances of Images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1334–1339, 2005.

[13] V. Vapnik, "Statistical learning theory." Wiley, New York, 1998.

[14] Y. Utsumi, T. Mizuno, M. Iwamura, and K. Kise, "Fast search based on generalized similarity measure," *IPSJ Trans. Comput. Vis. Appl.*, vol. 9, no. 1, p. 11, 2017.

[15] M. S. Azmi, M. F. Nasrudin, K. Omar, and K. W. M. Ghazali, "Farsi/Arabic Digit Classification Using Triangle Based Model Features with Ranking Measures," *2012 Int. Conf. Image Inf. Process. (ICIIP 2012)*, vol. 46, no. Iciip, pp. 128–133, 2012.

[16] S. S. S. Ahmad, "Feature and Instances Selection for Nearest Neighbor Classification via Cooperative PSO," in *Information and Communication Technologies (WICT), 2014 Fourth World Congress*, 2014, pp. 45–50.

[17] U. Bhattacharya and B. B. Chaudhuri, "Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 444–457, 2009.

[18] M. Greenacre and R. Primicerio, "Measures of distance between samples: Euclidean," *Multivar. Anal. Ecol. Data*, pp. 47–59, 2013.

[19] "Web-Based Dyanmic Similarity Distance Tool." [Online]. Available: http://ftmkproject.utem.edu.my/Web-basedSDT/websdt/websdt.jsf.