Timeseries Anomaly Detection Using SAX and Matrix Profiles Based Longest Common Subsequence

Thi Phuong Quyen Nguyen[1[*]], Trung Nghia Tran[2], Hoang Ton Nu Huong Giang[3], Thanh Tung Nguyen[4]

[1] Faculty of Project Management, The University of Danang- University of Science and Technology, 54 Nguyen Luong Bang, Danang, Vietnam
[2] Faculty of Applied Science, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam
[3] Department of Information Systems & Data Analytics, School of Computing, National University of Singapore, Singapore
[4] Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam

Abstract. Similarity search is one of the most popular techniques for time series anomaly detection. This study proposes SAX-MP, a novel similarity search approach that combines Symbolic Aggregate ApproXimation (SAX) and matrix profile (MP). The proposed SAX-MP method consists of two phases. The SAX method is used in the first phase to extract all of the subsequences of a time series, convert them to symbolic strings, and store these strings in an array. In the second phase, the proposed method calculates the MP based on the symbolic strings that are represented for all subsequences extracted in the first phase. Since a subsequence is represented by a symbolic string, the MP is calculated using a distance-based longest common subsequence rather than the z-normalized Euclidean distance. Top-*k* discords are detected based on the similarity MP. The proposed SAX-MP is implemented on several time series datasets. Experimental results reveal that the SAX-MP method is particularly effective at detecting anomalies when compared to HOT SAX and MP-based methods.

Keywords: Anomaly detection, Matrix profile, SAX, HOT SAX, LCS method.

1 Introduction

Anomaly detection is the process of identifying unusual states that occur in the dataset. Anomalies in time series can be caused by a variation in the amplitude of data or an alteration in the shape of the data [1]. Many wasteful costs and damages can be avoided if anomalies are detected early. For instance, real-time transaction data detection aids in the detection of fraudulent Internet transaction activities. Or anomaly detection in industrial manufacturing prevents incidents that may stop the production line.

Time series similarity search or discord search is one of the most popular techniques for anomaly detection, which is widely used in a multitude of disciplines, including healthcare systems, energy consumption, industrial process, and so on [2]. Symbolic

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_21

Aggregate ApproXimation (SAX) [3] is a representative method of similarity search-based approaches. The original SAX method aims to reduce the data dimensionality by converting time series data to a symbolic string. SAX is superior to other representations for detecting anomalies and discovering motifs in data mining [3]. Keogh *et al.* proposed a HOT SAX that employed a heuristic framework that included an outer loop and inner loop to order the SAX sequence for discord discovery [4]. Several extensions of SAX, such as HOT aSAX [5], ISAX [6], HOTiSAX [7], and SAX-ARM [8], were proposed to improve both effectiveness and efficiency concerning SAX and HOT SAX.

Besides, matrix profile (MP) [9, 10] is a measure of similarity between all subsequences of a single time series. MP is also a widely used technique of similarity search-based approach for discord discovery. The MP, which is based on an all-pair-similarity-search on the time series subsequences, is regarded as one of the most efficient methods for comprehensive time series characterization. Assume that there is a time series, *T*, and a subsequence length, *m*. The MP presents the *z*-normalized distance between each subsequence in *T* and its nearest neighbor. As a result, the MP is used to discover motifs, shapelets, discords, and so on. However, the main disadvantage of the MP is that its complexity rises quadratically with time series length.

Regarding the aforementioned analysis, this study proposes a novel similarity search approach that combines SAX and MP (denoted as SAX-MP) for anomaly detection. The proposed SAX-MP can capitalize on the benefits of both the SAX and the MP while minimizing their drawbacks. First, the MP requires quadratic space concerning the length of time series data, whereas SAX can reduce time series dimensionality. Two conversions are made by SAX: 1) Piecewise Aggregate Approximation (PAA) is employed to reduce the dimensionality of time series from *n* dimensions to *w* dimensions, and 2) the subsequences are finally transformed to symbolic representation. Thus, the proposed SAX-MP uses SAX to firstly segment and convert time series data into SAX symbols. Thereafter, the concept of MP is used to calculate the distance between each subsequence and its nearest neighbors. Time series discords are discovered based on the similarity-based MP, with extreme values. In addition, this study employs Longest Common Subsequence (LCS) method [11], which is an effective similarity measure for two strings, to calculate the similarity MP in the proposed SAX-MP.

The paper is structured as follows. Section 2 provides some related works. Section 3 describes the procedure of the proposed SAX-MP method. The numerical results are shown in Section 4. Section 5 comes with the conclusion and future research direction.

2 Related works

This section provides an overview of related studies on time series anomaly detection. The key definitions linked to this work are presented at the beginning of this section.

Definition 1. A *time series T* is a set of real-valued numbers, $T = t_1, t_2, \dots, t_n$ where *n* is the length of *T*.

Definition 2. A *subsequence* $T_{i,m}$ is a subset that contains *m* continuous variables from *T* beginning at location *i*. $T_{i,m} = t_i, t_{i+1}, \dots, t_{i+m-1}$, where $1 \leq i \leq n - m + 1$.

Definition 3. A *distance profile* D_i is a vector of Euclidean for fix *m* between a given

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_21

$T_{i,m}$ and each subsequence in the set of all subsequences. $D_i = [d_{i,1}, d_{i,2}, \dots, d_{i,n-m+1}]$, where $d_{i,j}$ is the distance between $T_{i,m}$ and $T_{j,m}$, $1 \le j \le n - m + 1$.

Definition 4: A *matrix profile MP* is a vector of z-normalized Euclidean distances between each $T_{i,m}$ with its nearest neighbors $T_{j,m}$. MP = [min(D_1), min(D_2),..., min(D_{n-m+1}), where D_i is a distance profile that is determined in Definition 3.

Definition 5: A *matrix profile index I* is a vector of integers corresponding to matrix profile MP: $I = I_1, I_2, \dots, I_{n-m+1}$, where $I_i = j$ if $\min(D_i) = d_{i,j}$.

Definition 6. *Time Series Discord*: Given a time series *T* and its subsequence $T_{i,m}$, $T_{i,m}$ is defined as a discord of *T* if it has the largest distance to its nearest neighbors. Regarding the matrix profile's definition, a subsequence $T_{i,m}$, is a k^{th}discord if its matrix profile is the k^{th} the largest distance in the vector of the MP.

2.1 Review of anomaly detection-based SAX approaches

Anomaly detection, which is the problem of discovering anomalous states in a given dataset, has been a significant issue in the field of data signals research. The detection of anomalies is regarded as the first and most important stage in data analysis to notify the user of unexpected points or trends in a collected dataset. The current approaches for anomaly detection on sensor data can be grouped into several classes such as statistical and probabilistic approaches, pattern-matching approaches, distance-based methods, clustering approaches, predictive methods, and ensemble methods [12]. This study focuses on reviewing some approaches based on the SAX method.

SAX is a symbolic representation of time series that gives an approximation. SAX employs PAA [13], a non-data adaptive representation method, to reduce dimensionality. The PPA technique partitions the data into subsequences of equal length. Each subsequence's mean is calculated and used as a subsequence representation. Then, SAX will convert each subsequence into a symbolic representation. SAX is an efficient and useful method in querying time series subsequences and can also be used in a variety of time series data mining activities, including pattern clustering and classification [2].

Regarding the SAX method, several applications focus on simple and effective signal conversion. Lin et al. [3] employed SAX to compare the differences between data patterns by converting data to strings. Keogh et al. [4] proposed a HOT SAX to search and compare abnormal signals for time series data. Three main advantages of the SAX method are also determined [4]. First, SAX can perform the data dimension reduction which reduces the complexity of subsequent analysis and reduces memory usage. Second, SAX converts data into strings to make data pattern comparison more convenient and easy to interpret. Third, SAX gains an advantage in the definition of the minimum distance formula, which is used to determine the difference between strings. If the distance between two strings is smaller, the shape is closer. For the two closer strings, the greater distance represents the more difference in morphology. SAX simply compares different strings by computing the distance between the strings and using the distance to find the highly repetitive patterns or the most different patterns. The related research employed the SAX method for anomaly detection can be seen in [3-5, 8, 13].

2.2 Review of the Longest Common Subsequence (LCS) method

Suppose that there are two strings $X = (X_1, X_2, \dots, X_n)$, $Y = (Y_1, Y_2, \dots, Y_n)$, and their

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_21

common subsequence is denoted as *CS(X, Y)*. The longest common subsequence of *X* and *Y*, LCS(*X, Y*) is defined as a CS(*X, Y*) with maximum length [14]. The concept of LSC is to match two sequences by allowing them to stretch without altering the order of each value in the sequence. Unmatched values discovered throughout the LCS searching and matching process would be excluded (e.g., outliers) without impacting the final LCS outcome. The LCS might reflect the similarity of the two subsequences in this research. The recurrence relations of the LCS method are described as follows:

$$LCS(X_i, Y_j) = \begin{cases} 0 & if\ i = 0\ or\ j = 0 \\ LCS(X_{i-1}, Y_{j-1}) + 1 & if\ i, j > 0\ and\ X_i = Y_j \\ \max\{LCS(X_i, Y_{j-1}), (X_{i-1}, Y_j)\} & if\ i, j > 0\ and\ X_i \neq Y_j \end{cases} \quad (1)$$

where $LCS(X_i, Y_j)$ represents the set of LCS of prefixes X_i and Y_j.

3 Proposed method

This section presents the proposed SAX-MP method that combines SAX and MP for time series anomaly detection. The SAX-MP method consists of two phases. In the first phase, time series data is converted to alphabet strings using the SAX method. The second phase is to calculate the MP based on the LCS technique. Time series discords are found based on the MP's outcome. The SAX-MP procedure is illustrated in Fig. 1.

The proposed method is described detailed as follows. A given time series *T* is firstly normalized in the proposed SAX-MP method. Then, SAX is employed to convert it into alphabet strings. Regarding the SAX method, two parameters, i.e., alphabet size "*a"*, and word size "*w"*, need to be defined. Referring to the work of HOT-SAX [4], the SAX alphabet size "*a"* was set as 3 after conducting an empirical experiment on more than 50 datasets. The SAX word size "*w"* highly depends on time series characteristics. A smaller value of *w* is preferred for datasets that are typically smooth and slowly altering, whereas a greater value of *w* is preferred for more complicated time series. Given a length of subsequence (*m)*, a set of all subsequences is extracted by moving the *m*-length window across the time series. A SAX representation is created based on the selected parameters of "*a"* and "*w"*. Each subsequence is then converted to alphabet strings and stored in an array A. This whole process is quite similar to the Outer Loop of HOT-SAX. Fig.2 illustrates this converting process.

In the next stage, the MP is calculated based on the set of all subsequences resulting from the first stage. The distance profile D_i is firstly computed based on Definition 3. The distance between two subsequences is calculated by the LCS method. Herein, each subsequence is represented by a symbolic string. Thus, the distance profile D_i based LCS is defined as $D_i = [LCS_{i,1}, LCS_{i,2}, \ldots, LCS_{i,n-m+1}]$, where $LCS_{i,j}$ is the distance between A_i and A_j, $1 \leq j \leq n - m + 1$, A_i and A_j belong to array A. The $LCS_{i,j} = LCS(A_i, A_j)$ is calculated by Eq.(1). Thereafter, the *MP*-based LCS and index *I* are obtained using Definitions 4 and 5. Several algorithms were used to compute the *MP* such as STAMP [9], STOMP, SCRIMP++, and AAMP, with their time complexity being $O(n^2 \log n)$, $O(n^2)$, $O(n^2 \log n/m)$, and $O(n(n-m))$, respectively.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_21

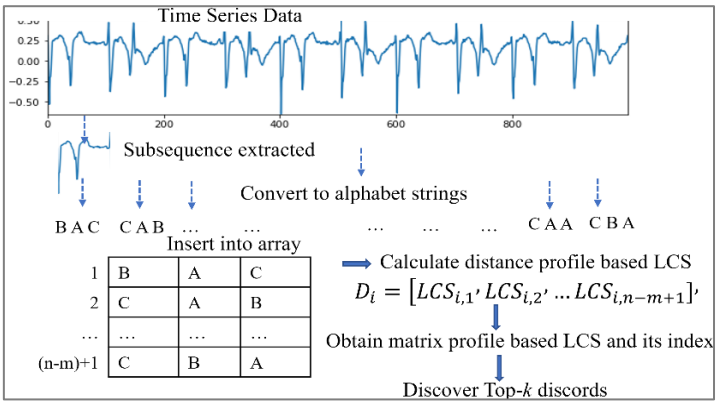

Fig. 1. An illustration of the proposed SAX-MP.

This study inherits the process to perform a pairwise minimum operation in STAMP, STOMP, and SCRIMP algorithms to obtain the matrix profile *MP* and its index *I*.

Fig. 2 describes the procedure to obtain MP-based LCS. Array A is obtained from the first stage of the proposed SAX-MP. Lines 1 and 2 initialized the matrix profile *MP* and matrix profile index *I*. Besides, *idxes* is initialised as the number of subsequences.. Line 4 calculates the distance profile-based LCS using Eq.(1). Line 5 computes the *MP*-based LCS and matrix profile index *I* using the pairwise minimum operation [9]. Finally, the vector *MP* and corresponding *I* are obtained.

Time series discords are found based on the obtained matrix profile *MP*. The maximum value in the MP shows the first discord while its corresponding index *I* reflects the location of the first discord. According to Definition 6, the top-*k* discords can be detected where is a user-defined parameter. An anomalous pattern is identified starting at the discord *k* index and lasting till the end of subsequence length *m*.

Procedure to find *MP* based LCS

Input: array *A* (contains n-m+1 subsequences in symbolic representation)
Output: matrix profile *MP* based *LCS* , matrix profile index *I*

```
1   Initialization:
2       MP ← infs, I ← zeros, idxes ← n-m+1
3   for each idx in idxes do
4       D ← LCS (A(idx), A) // Calculate distance profile based LCS
5       MP, I ←Element Wise Min (MP, I, D, idx)
6   end for
7   Return MP, I
```

Fig. 2. Algorithm procedure to calculate the MP in the proposed SAX-MP method

4 Experimental results

This section provides the empirical results to evaluate the performance of the proposed SAX-MP method on different datasets. The proposed SAX-MP is compared with the MP method [9], and HOT SAX [4]. The result of HOT SAX is obtained from its original paper. All the datasets in the experiment are collected from a supported page provided

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_21

by Keogh. E. [18]. The algorithm is evaluated based on two performance metrics: effectiveness, and efficiency. Regarding effectiveness, F_1 score is employed. F_1 score is calculated as $F_1 = ((2 * PR)/(P + R)) * 100\%$, where $P = (TP/(TP + FP)) * 100\%$, and $R = (TP/(TP + FN)) * 100\%$ are precision and recall, respectively. TP, FP, and FN are denoted for true positive, false positive, and false-negative results respectively. F1 scores range from 1 to 0, with 1 being the best and 0 being the worst.

4.1 Anomaly detection on the tested datasets

(I) Space Telemetry dataset

This dataset contains 5000 data points which are divided into five energize cycles. The fifth cycle is remarked as the Poppet pushed considerably out of the solenoid before energizing, indicating the discord. The parameter is set up as m = 128, w =4, and a =3. The SAX-MP detects the first discord at the beginning location of 4221, which is quite similar to the result of the HOT SAX method as well as the experts' annotations. The result of the MP method is a little different from HOT SAX and the SAX-MP since the MP can detect the top 3 discords at the beginning locations of 3870, 2872, and 3700, respectively. Thus, the SAX-MP and HOT SAX can effectively detect the anomalies in this dataset while the MP method has some deviation from the real discord.

(II) Electrocardiograms data

The first ECG dataset is stdb_308 which contains two time-series. The length of the sequence is set as m =300 while "w" and "a" are set as 4 and 3, respectively. The proposed SAX-MP finds out the discord at I_{MP}=2286 and I_{MP}=2266 for the datasets stdb308_a and stdb308_b, respectively. This result is consistent with HOT SAX and the ground fact. However, the first discord found by the MP method is quite different from the beginning position of 2680 for the dataset stdb308_a. The second discord of the MP method (at position 2282) is relatively close to the I_{MP}.

The second ECG dataset is mitdb/x_mitdb/x_108, which is more complicated with different types of anomalies. The length of the subsequence is selected as m=600. The proposed SAX-MP method points out the top three discords at 10868, 10022, and 4020, respectively while the top three discords are located at 10871, 10014, and 4017, respectively by the HOT SAX method. The MP method also finds out the top three discords at positions 10060,11134, and 4368, respectively. The result proves that the proposed SAX-MP method is quite effective in anomaly detection on time series data.

(III) Video surveillance dataset

A video surveillance dataset was compiled from a video of an actor performing various acts with and without a replica gun. The first three discords are founded by the SAX-MP method at locations 2250, 2850, and 290, respectively. The parameter setting of the SAX-MP method is as follows: m=250, w =5, and a =3. The HOT SAX detects 3 potential discords at locations 2250, 2650, and 260, respectively, while the discords discovered by MP are quite similar to the SAX-MP, located at 2192, 2634, and 1942. Note that three anomalous patterns detected by the SAX-MP are covered by the top two discords discovered by the HOT SAX. Besides, compared with the real discords of this dataset in [19], the proposed SAX-MP can detect the majority of anomalies.

Table 1 summarizes the discord discovery result of the SAX-MP, MP, and HOT SAX methods. Fig. 3 displays the result on mitdb/x_mitdb/x_108 and video surveillance datasets for illustrations.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_21

Table 1. Result summary of discord discovery

Dataset	Discord length	Top-*k* discords	Location of discord		
			SAX-MP	MP	HOT SAX
Shuttle	128	1	4221	3780	4228
stdb308_a	300	1	2286	2680	2286
stdb308_b	300	1	2226	2710	2226
mitdb/x_mitdb/x_108	600	1	10868	10060	10871
		2	10022	11134	10014
		3	4020	4368	4017
Video	250	1	2320	2192	2250
		2	2630	2634	2650
		3	2090	1942	260

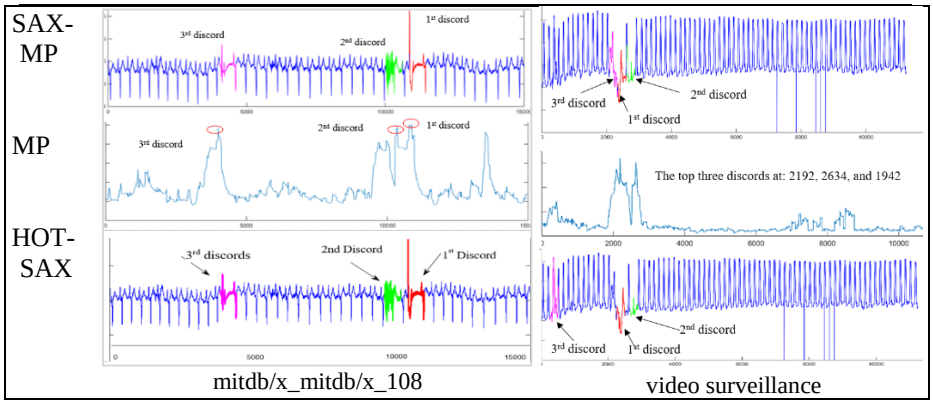

Fig. 3. Experimental result in ECG dataset (mitdb/x_mitdb/x_108)

4.2 Performance evaluation

To evaluate the performance of the proposed method, the F_1 score and running time are used. Since the abnormal sequences were labeled by domain experts, we considered the results provided in [11] to be correct and these results are used for comparison. The detected results of the proposed SAX-MP are entirely consistent with the experts' discords on Shuttle and the two ECG datasets. Thus, these datasets obtain the F_1 score of 1. For the video dataset, the SAX-MP only achieves an F_1 score of 0.89 while HOT SAX and MP methods adopt the F_1 score as 0.81 and 0.72, respectively. This result proves the effectiveness and efficiency of the proposed method.

5 Conclusion

SAX is a well-known method for reducing time series data dimensions by converting a time series to symbolic strings. The proposed SAX-MP approach, which combines SAX and MP, can take advantage of SAX in reducing the dimensionality of data while lowering the MP's drawback in time complexity. A time series data is firstly segmented using the PPA technique and then converted to symbolic strings. Thereafter, the MP is employed on the converted strings. Instead of using the Euclidean Distance, the proposed SAX-MP utilizes the LCS technique to compute the similarity MP for the strings

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_21

extracted from the SAX procedure. The similarity MP is used to detect the anomalies in time series data. The experimental results in five time series datasets show that the proposed SAX-MP is extremely effective in discovering the top three discords of the tested datasets.

The length of the subsequence, the SAX word size, and the alphabet size all have an impact on the results of the proposed SAX-MP. Thus, these features can be optimized using some meta-heuristic approaches such as sine-cosine algorithm in future research.

Acknowledgment
This work was funded by Vingroup Joint Stock Company (Vingroup JSC) and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2020.DA19. The support is much appreciated.

References

1. Izakian, H. and W. Pedrycz: Anomaly detection in time series data using a fuzzy c-means clustering. in 2013 Joint IFSA world congress and NAFIPS annual meeting (IFSA/NAFIPS), (2013).
2. Hsiao, K.-J., et al., Multicriteria similarity-based anomaly detection using Pareto depth analysis. IEEE transactions on neural networks and learning systems, 27,1307-1321, (2015).
3. Lin, J., et al. A symbolic representation of time series, with implications for streaming algorithms. in Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, (2003).
4. Keogh, E., J. Lin, and A. Fu. Hot sax: Efficiently finding the most unusual time series subsequence. in Fifth IEEE International Conference on Data Mining (ICDM'05), (2005).
5. Pham, N.D., Q.L. Le, and T.K. Dang. HOT aSAX: A novel adaptive symbolic representation for time series discords discovery. in Asian Conference on Intelligent Information and Database Systems, Springer, (2010).
6. Sun, Y., et al., An improvement of symbolic aggregate approximation distance measure for time series. Neurocomputing, 138, 189-198, (2014).
7. Buu, H.T.Q. and D.T. Anh. Time series discord discovery based on iSAX symbolic representation. in Third International Conference on Knowledge and Systems Engineering, (2011).
8. Park, H. and J.-Y. Jung, SAX-ARM: Deviant event pattern discovery from multivariate time series using symbolic aggregate approximation and association rule mining. Expert Systems with Applications, 141, 112950, (2020).
9. Yeh, C.-C.M., et al. Matrix profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. in IEEE 16th international conference on data mining (ICDM), (2016).
10. Yeh, C.-C.M., et al., Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile. Data Mining and Knowledge Discovery, 32(1), 83-123, (2018).
11. Hunt, J.W. and T.G. Szymanski, A fast algorithm for computing longest common subsequences. Communications of the ACM, 20(5), 350-353, (1977).
12. Cook, A.A., G. Mısırlı, and Z. Fan, Anomaly detection for IoT time-series data: A survey. IEEE Internet of Things Journal, 7(7), 6481-6494, (2019).
13. Keogh, E., et al., Dimensionality reduction for fast similarity search in large time series databases. Knowledge and information Systems, 3(3), 263-286, (2001).
14. Iliopoulos, C.S. and M.S. Rahman, Algorithms for computing variants of the longest common subsequence problem. Theoretical Computer Science, 395(2-3), 255-267, (2008).
15. E., K. www.cs.ucr.edu/~eamonn/discords/, (2005).
16. Hu, M., et al., A novel computational approach for discord search with local recurrence rates in multivariate time series. Information Sciences, 477, 220-233, (2019).

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_21