Article

Recognition of Micro-Expressions using the Spatiotemporal Capsule Network

Sana Asif [1,*], Muhammad Mugees Asif [2], Rabia Hussain [2] and Imtnan Khalid [2]

[1] Management Department, Air University, Islamabad 44230, Pakistan
[2] Department of Computer Science, Lahore Garrison University, Lahore 94777, Pakistan
* Correspondence: mesanaasif@gmail.com (S.A.)

Abstract: Micro-expression (ME) is one of the key psychological stress reactions. It is a modest, spontaneous facial mechanism. ME has significant applicability in a variety of psychologically-related sectors because to its precision and unpredictability with regard to psychological manifestations. Nevertheless, the current Micro-expression recognition (MER) algorithms have poor accuracy and a limited quantity of ME data, and this study issue has not been thoroughly investigated. Therefore, we present an approach for deep learning based on a Spatio-temporal capsule network (STCP-Net). STCP-Net has four components: a jitter reduction module, a differential feature extraction module, an STCP module, and a fully linked layer. The first two modules are aimed to extract diversifying differential features more precisely and to limit the influence of head jitter. The STCP module is used to extract Spatio-temporal features layer by layer, taking the temporal and geographical connection between features into account. This research runs sufficient trials using the Leave One Subject Out (LOSO) methodology for cross-validation using the CASMEII dataset. The conclusion and analysis demonstrate that the algorithm is innovative and efficient.

Keywords: Capsule network; differential features; deep learning; micro-expression recognition; spatiotemporal features

1. Introduction

Facial expressions are one of the most significant means of expressing emotions and a direct method of conveying feelings to others. According to time and spatial scales, facial expressions may be classified into macro-expressions and micro-expressions (MEs). The length of macro-expressions ranges between 0.75 and 2 seconds, and the amplitude of facial muscle movements changes with the intensity of the emotion. MEs have a duration between 0.04 and 0.2 seconds, and facial muscular movements are often rather minute [1, 2].

In contrast to the controllability of macro-expressions, MEs feature a significant spontaneity that enables the expressor to consciously convey his or her concealed actual feelings to the outer world. Due to this property, MER has attracted great interest in health care, criminal investigation, and national security [3, 4]. However, it is difficult for humans to detect and identify MEs without specialised training, thus Ekman created ME training tools [5] to assist individuals learn to identify MEs. However, despite minimal training, the accuracy and efficiency of recognition remain inadequate.

The advancement of computer technology has ushered researchers into the age of using algorithms to identify MEs. In the early stages of study, local binary patterns from three orthogonal planes (LBP-TOP) [6] proposed by

 doi:10.20944/preprints202211.0488.v1

Zhao et al. and 3D gradient descriptors [7] presented by Polikovsky et al. obtained success.

Researchers have regularly created standard datasets such as CASME [8], SMIC [9], and SAMM [10] for the creation of MER tasks. In addition, the IEEE International Conference on Automatic Face & Gesture Recognition led to the organisation of a number of MER contests [11]–[14], which not only highlighted the development objectives of ME tasks but also considerably accelerated the rate of MER development.

MER papers may be divided into three types according on the feature extraction techniques used: handmade, handcrafted-deep learning, and deep learning.

Handcrafted-based approaches include LBP-based [6], [15]–[20], optical-flow-based [21]–[24], and more techniques (e.g., colour space, histogram.). Using deep learning to extract additional handcrafted features [25]–[27] or fusing handcrafted features with deep-learning features, respectively [28]. Deep-learning-based may be separated into three categories depending on the input data: input onset frame and apex frame [31], which extract the face texture and light-shadow. ME's geometric change features are extracted from the input frame sequence [32]– [35] using the geometric change features. Compared to single-frame and double-frame inputs, the input multi-frame MER approach is able to extract more Spatio-temporal information from MEs and has a superior recognition result.

To derive Spatio-temporal information from MEs, scholars have examined the phenomenon from several angles. According to the start and apex frames, Su et al. [25] retrieved first-order motion, second-order motion, and segmentation probability maps. Yang et al. [26] and bai et al. [32] retrieved spatial features using VGGNet-16 and VGGFace, respectively, and then extracted temporal features using LSTM. Wang et al. [27] extracted features from frame sequences and optical flow sequences, respectively, using STAN-A and STMN-A, and then conducted fusion classification. Lei et al. [31] extracted characteristics between facial action units using graph convolution (AUs). Bai et al. [36] used video motion amplification to amplify MEs on a spatiotemporal scale. Wu et al. [37] and Song et al. [38] both used three-stream convolutional networks, although with different configurations. Kosiorek [29] and Liu [39] classified single and multiple frames, respectively, using capsule networks. These deep learning approaches surpassed conventional algorithms in terms of recognition efficiency and precision in MER tests.

Observations of MEs indicate that the ME process is quick, duplicate information still exists [31], [40]. Besides frame number redundancy, there is also due to the Mild facial muscular contractions [1] and [2]. To get the Spatio-better management of temporal information in ME, we suggest STCP-Net. STCP-Net employs a differential technique to eliminate duplicate packets. The static information of the face from which the more efficiently transmitting dynamic information between frames, and then extract the geographic data using the STCP module the temporal information inside frames and between frames. In addition, to diminish the impact of head shaking. We develop a head network for the differential feature section.

In conclusion, our principal contributions are as follows:

1) We propose the STCP module for the first time after analysing the ME dataset and considering relevant literature. Due to the unique qualities of the capsule network, we expand the conventional capsule network into a multi-layer capsule network capable of learning spatio-temporal data. Experiments demon-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0488.v1

strate that the module is superior than the conventional capsule network.

2) STCP-Net is built with a network capable of aligning the identical features of distinct frames in preparation for the next phase, differential features.

3) The STCP-Net model operating on ME sequences is built, tests are conducted on the CASMEII dataset, and the results are compared to other state-of-the-art approaches. The final findings indicate that the procedure outperforms other current cutting-edge techniques.

2. STCP-NET

In this part, we outline the overall architecture of STCP-Net before introducing the specific modules. Our suggested technique consists of two major components: the extraction of differentiating differential characteristics and the extraction of spatial-temporal information. As an important prerequisite for the second half, the first part must guarantee that accurate face dynamic information is retrieved; therefore we present the Debouncer and Differential Feature Extraction module's framework structure in the first part. In the second section, we present the composition structure and operation of this innovative Spatio-temporal feature extraction method—STCP module.

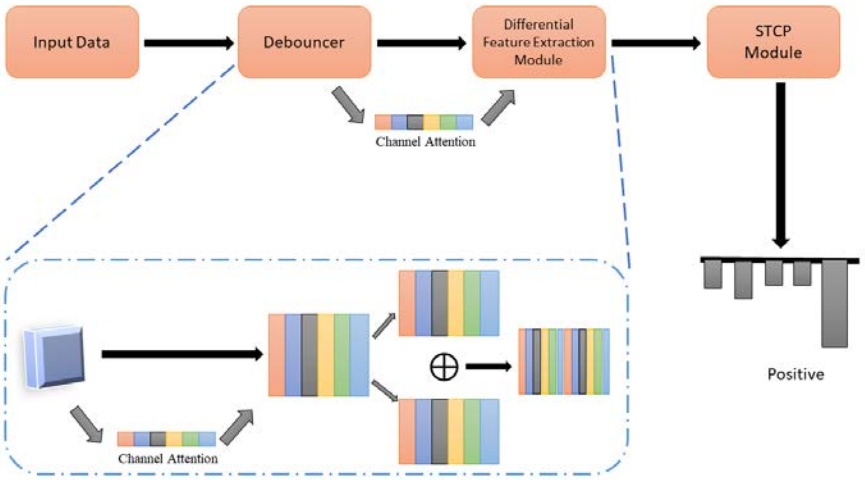

Figure 1. *General framework of STCP-Net*

As seen in Figure 1, we input a series of 3X3 ME frames, get alignment characteristics through the Debouncer, and add a channel attention mechanism to the alignment features. Then, using the Differential Feature Extraction module, we extract and fuse the differentiating differential features. Finally, utilize the STCP module to examine and categories the differential characteristics.

During an episode of ME, the portion of the face that displays movement represents dynamic information, whilst the portion that does not display movement represents static information. The LBP-TOP algorithm, the optical flow approach, and the facial dynamic map (FDM) [22] algorithm forgo the static information of the face in favor of its dynamic information. Numerous deep learning algorithms use Optical Flow techniques to extract dynamic data [21], [24], [41]. Most trials demonstrate that the algorithm that inputs many frames is more accurate than the algorithm that inputs a single frame. Algorithms that employ multiple frames as input data attempt to ex-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0488.v1

tract and discriminate this dynamic information more efficiently. In light of this, we forsake the static information in ME sequences and apply deep learning and the difference method (DM) to extract the dynamic information in ME sequences.

Table 1. *Network architecture of debouncer blocks*

Layer	Kernel	Stride	Padding	Output Size
Input	N/A	N/A	N/A	74 X 74 X 3
Conv1	3 X 3	1 X 1	3 X 3	74 X 74 X 64
Conv2+Relu	3 X 3	2 X 2	2 X 2	78 X 78 X 64
Conv3	3 X 3	2 X 2	N/A	40 X 40 X 64
Conv4	3 X 3	1 X 1	N/A	19 X 19 X 64
MaxPool	3 X 3	1 X 1	1 X 1	17 X 17 X 64

As input, we use the ME series to extract dynamic information; the start frame is the first frame in the frame sequence, and the peak frame is the last. Song et al. [38] proved that block-based segmentation of the face might enhance feature extraction. Before extracting features using the Debouncer, we divided the 224-by-224-pixel frame into nine 74-by-74-pixel blocks using a 3x3 grid. Tab1 depicts the precise network configuration used to maximise the alignment effect of Debouncer.

Finally, we execute the difference operation between the collected alignment features and the corresponding alignment features from the first frame in order to extract the dynamic information from the features. However, the information between the two frames represents a complicated shift in texture. To identify the rise or decrease of texture, the Relu operation is used on the resulting differential feature to produce the differential feature F d and the inverse.

After extracting the alignment features, it is difficult to see the variations between each frame's alignment characteristics with the naked eye. However, when F d and F antid are extracted, we can see that the face is constantly in motion. Group (c) displays the dynamic data of the mouth and eye, while Group (d) isolates the motion of the glabella and left chin areas. The various difference approach may successfully eliminate duplicate information and enhance the model's expressiveness, highlighting the significance of the reverse difference characteristics. To extract useful Spatio-temporal characteristics from dynamic data, we proposed a multi-layer capsule structure for MER based on the capsule network. Sabour et al. [42] presented a capsule network as a solution for CNNs' lack of spatial information.

The capsule network employs vector groups to represent the instantiation parameters of features, and the weights of the low-level and higher-level capsules are acquired through dynamic routing in order to rate-code the likelihood of objects in space and are stored as vectors in the top capsules. Mazzia et al. [43] abandoned the original routing method and presented a capsule network based on a self-attentive routing mechanism, which allowed the capsule network to operate effectively with just 2% of the original parameters. This module's starting point is the capacity of the capsule network to extract spatial information properties.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0488.v1

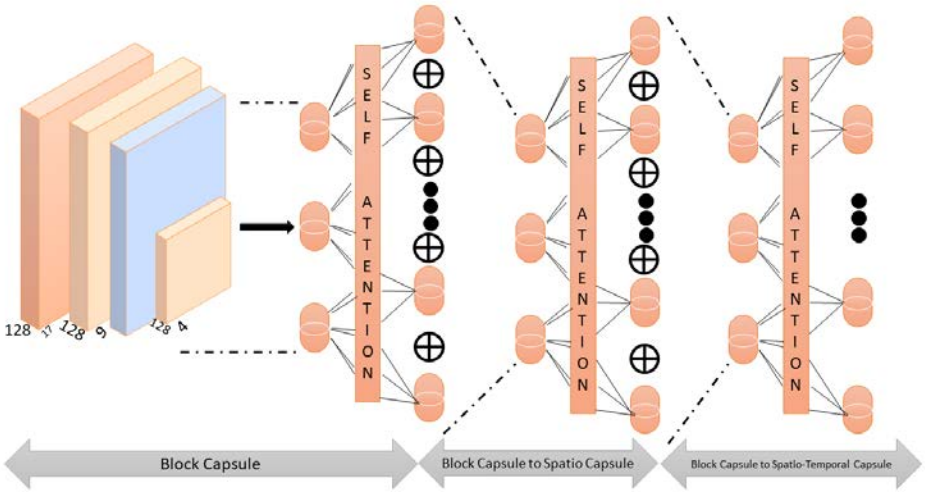

***Figure 2**. General framework of STCP module.*

Capsules are taken from block characteristics and expanded progressively to spatial and temporal characteristics. Figure 2 depicts the whole STCP module architecture. Initially, we use algorithms like as convolution to extract low-level capsules from block characteristics. The Efficient- Capsnet approach provided by Mazzia et al. [43] is then used to create higher-level capsules based on block characteristics. Until now, we have extracted many higher-level capsules on each frame block, and these capsules hold different entity properties. Our data may be represented as a matrix of s k n0 d0 at this point. Where s is the number of frames in the ME sequence, k is the number of blocks a single frame is split into, n0 is the number of higher-level capsules in a single block, and d0 is the capsule dimension.

We combine all higher-level capsules retrieved from a single block in a single frame into a single capsule, such that a single block capsule (BC) holds all of this block's properties. By continuing to route the k BCs in a frame, many capsules carrying the spatial information of a single frame may be obtained. Currently, the data may be represented by the matrix s n1 d1. Where n1 is the number of capsules of a higher level packed in a single frame and d1 is the capsule's diameter. Using this strategy, continue fusing all higher-level capsules in a single frame into a single capsule to create s space capsules (SC), each of which contains all spatial information in a single frame.

Multiple SCs are then routed to produce a capsule holding Spatial-temporal information (STC), where the data may be described as n2 d2, where n2 and d2 represent the number and dimensions of STC, respectively.

3. Results

This section describes the dataset utilized, the implementation specifics, and the experimental outcomes comparison. To test our proposed STCN technique, we perform exhaustive MER experiments on the CASMEII dataset. Yan et al. [8] obtained the CAS- MEII dataset at the Institute of Psychology, Chinese Academy of Sciences. CASMEII contains 255 laboratory-collected samples from 26 people.

***Table 2**. Sample distribution of CASMEII datasheet*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0488.v1

Category	Emotion	Total
Five	Positive (32) Negative (69) Surprise (28) Repression (27) Others (99)	255
Four	Positive (32) Negative (96) Surprise (28) Others (99)	255
Three	Positive (32) Negative (96) Surprise (28)	156

Table 2 displays the CASMEII dataset's specifics. All experiments employ the LOSO cross-validation methodology. In other words, 20% of the complete dataset is utilized as the test set, while the remaining 80% is used for training. For the training data, one subject every round was chosen as the validation set, while the other subjects were used as the training set. The F1 score and precision is the mean of the 26 training outcomes on the test set. The LOSO cross-validation methodology may decrease the error resulting from over fitting and assure the dependability of experimental outcomes. It is a regularly used approach for MER cross-validation

Experiments choose Adam as the optimizer and set the learning rate to 1e-4, which declines as the number of repetitions grows. When the validation loss curve is typically steady, a dynamic, iterative procedure is utilized to finish the current round of tests and store the optimum model. The following is the formula for the cross-entropy, which is selected as the loss function:

Table 3. *Hardware and software environment configuration for experiments*

Category	Version/Model
GPU	Nvidia GTX 166I0Ti 6G
CPU	Macbook Pro i5
Python	3.11
Pytorch	1.10.1+cu113
RAM	8GB

Probability expectation p is the expected output, while probability distribution q is the actual output. Table 3 displays additional configurations. We ran three-category, four-category, and ablation experiments on the CASMEII dataset in this part.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0488.v1

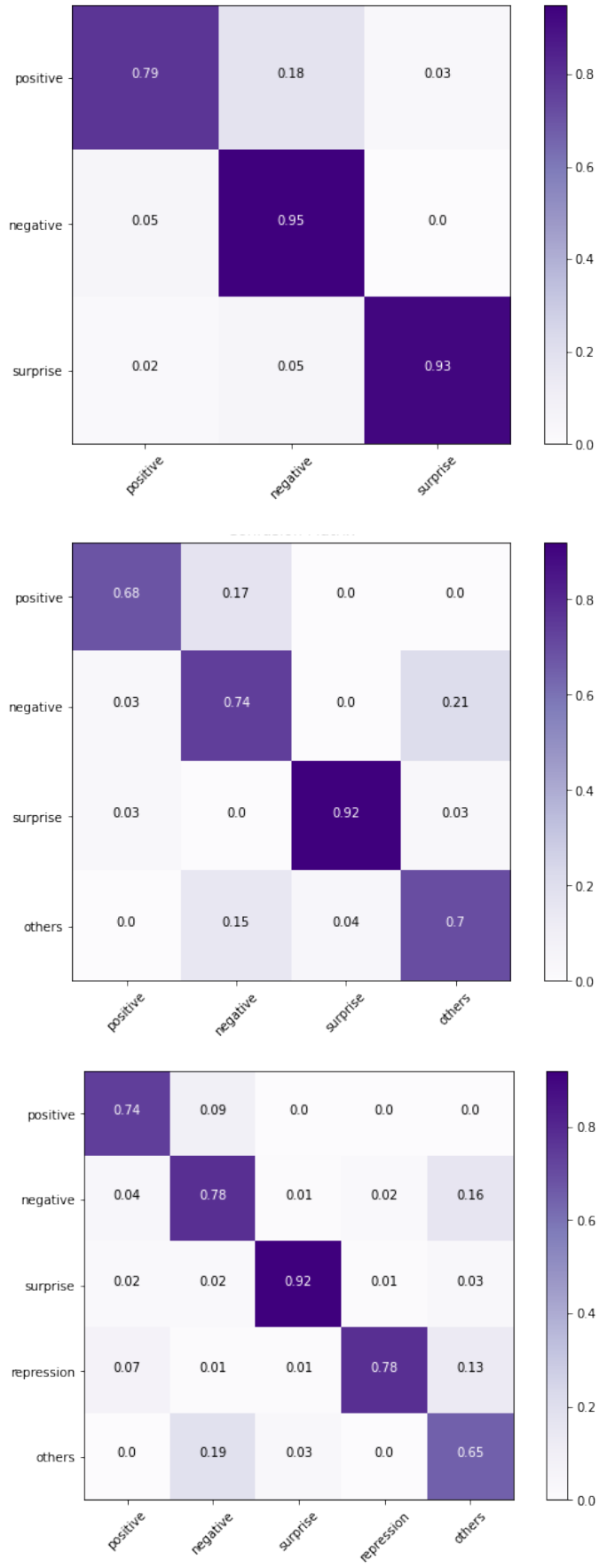

Figure 4. *Confusion Matrices*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0488.v1

Figure 4 depicts the main confusion matrices. STCP-Net is compared to other cutting-edge techniques, all of which employ the LOSO protocol. The following equations are used to compare UAR, accuracy, and F1 scores in Tables 4, 5, and 6.

where K represents the total number of categories, M represents the confusion matrix, Recalli represents a category's recall, and P reisioni represents a category's prediction.

Table 4. *Comparison with state-of-the-art methods on three-category*

Year	Abbreviation	UAR	ACC (%)	UFI
2007	LBP-TOP [6]	0.7429	N/A	0.7026
2018	Bi-WOOF [23]	0.8026	N/A	0.7805
2019	CapsuleNet [29]	0.7018	69.34	0.7068
2019	EMR [44]	0.8209	N/A	0.8293
2019	STST-Net [45]	0.8686	N/A	0.8382
2019	Dual-Inception	0.8560	N/A	0.8621
2020	AU-GACN [33]	N/A	71.20	0.3550
2021	AU-GCN [31]	0.8710	N/A	0.8798
2022	FR [47]	0.8915	88.73	N/A
2022	Ours	0.9316	94.03	0.9316

As demonstrated in Table 4, we describe the performance of STCP-Net across three categories and compare it to other approaches using the LOSO protocol. Our STCP-Net has an accuracy of 91.03 percent, a UAR of 0.8906, and an F1 score of 0.88 in three-category classification. Compared to the cutting-edge approaches EMR [44], STST-Net [45], and Dual-Inception [46], our UF1 improved by 0.0697, 0.0220, 0.0346, and 0.0296, respectively.

Compared to the FR algorithm, our suggested STCP-Net technique in three-category enhanced recognition accuracy by 2.3%.

Table 5. *Comparison with state-of-the-art methods on four-category*

Year	Abbreviation	ACC(%)	UFI
2011	LBP-TOP[15]	40.9	0.369
2014	LBP-SIP[19]	45.7	0.425
2015	STLBP-IP[20]	45.1	0.497
2018	Bi-WOOF[23]	58.9	0.41
2019	STRCN[35]	**80.3**	0.747
2020	Graph-ten[49]	73.6	N/A
2021	DSTAN[27]	75.2	0.728
2022	FR[47]	68.38	N/A
2022	ours	85.36	0.8536

According to Table 5, our STCP-Net has a precision of 74.31% and a UF1 of 0.7541. In comparison to the state-of-the-art methods Bi-WOOF [23], DSTAN [27], and FR [47], our accuracy increases by 15.41%, 0.711%, and 5.93%, respectively. The precision is inferior to DSTAN [27] and STRCN [35],

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0488.v1

but UF1 is improved by 0.0261 and 0.0071, respectively. The comparison demonstrates that our proposed STCP-Net method is highly competitive across all four categories.

***Table 6**. Comparison with state-of-the-art methods on five-category*

Year	Abbreviation	ACC(%)	UFI
2007	LBP-TOP[6]	39.68	0.3589
2016	STCLQP[50]	58.39	0.5835
2018	Bi-WOOF[23]	58.85	0.61
2019	3DCNNs[51]	65.9	N/A
2019	TSCNN[38]	74.05	0.7327
2020	AU-GACN[33]	56.1	0.394
2020	CNNCapsNe[39]	64.63	0.5894
2021	METRA[26]	60.54	N/A
2021	KFC-MER[25]	72.76	0.7375
2021	TSNN-IF[37	73.81	0.60601
2021	TSNN-LF[37]	75.49	0.6142
2022	FR[47]	62.85	N/A
2022	Ours	85.36	0.8536

In the five-category trial, our STCP-Net accuracy was 73.32 percent with a UF1 of 0.7576, as shown in Table 6. Our STCP-Net outperforms AU-GACN [33], METRA [26], CNNCapsNe [39], KFC-MER [25], 3DCNNs [48], and FR [47] in general. Compared to the TSNN [37] and TSCNN [38] algorithms, the UF1 score is enhanced by 0.1434 and 0.024, respectively.

***Table 7.** Ablation study on CASMEII dataset*

	ACC(%)	UAR	UFI
Without-Difference	63.27	0.598	0.5747
Without-Relu	84.12	0.8056	0.8188
Without-Anti	76.85	07157	0.7337
TSCP-Net	86.3	0.8647	0.8452
STCP-Net	**92.05**	0.9205	0.9188

To test the efficacy of differential features and STCP-Net, we conduct a comprehensive ablation analysis on Table 7 of the CASMEII dataset. We compared this method to five other versions: (Without-Difference); using differ- ential features in Debouncer but no Relu operation (Without-Relu); using differ- ential features in Debouncer and performing Relu operation but no extraction F anti. (Without-Anti); after extracting the BC, the extracted temporal information operation is executed first, then the extracted spatial information operation (TSCP-Net).

***Table 8.** Comparison of average time for classification of single sample by MER models.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0488.v1

Year	Approaches	Method Type	Accuracy (%)	FI	Classification(s)	Total(s)
2011	LBP-TOP (CPU)	Handcraft	40.9	N/A	0.584	18.873
2014	LBP-TOP (CPU)	Handcraft	45.7	0.369	0.208	16.088
2017	LBP-TOP (CPU)	Handcraft	N/A	N/A	0.584	1.584
2018	Residual Network (CPU)	Deeplearning	74.7	0.64	0.95	0.95
2020	MA (CPU)	Deeplearning	76.3	0.668	1.1	1.1
2022	Ours (GPU)	Deeplearning	85.43	0.8543	0.013	0.013
2022	Ours (CPU)	Deeplearning	85.43	0.8543	0.28	0.28

Additionally, we have analysed the performance of different methods and included them in Table 8. Operating time in this work is 0.38 seconds per sample on the CPU and 0.015 seconds per sample on the GPU; the running environment is shown in Figure 3.2. It can be noted that the deep learning-based MER technique significantly increased running speed and accuracy rate compared to conventional methods. STCP-Net reduces the running time of Residual Network and MA by 0.57s and 0.72s, respectively, compared to the state-of-the-art deep learning-based algorithms Residual Network and MA.

5. Conclusions

This article presents a Spatio-temporal capsule network (STCP-Net) for MER. Extensive tests are conducted and compared on the public spontaneous dataset CASMEII to assess the suggested technique. The experimental findings demonstrate that our strategy raises the recognition speed and accuracy considerably.

In addition, this work investigates a topic that has not garnered a great deal of attention in recent studies. ME is distinguished from macro-expressions by its subtle action amplitude, which renders the majority of macro-expression recognition algorithms inappropriate for ME. Fundamentally, the amplitude of movements in MEs is significantly less than in macro-expressions, resulting in the majority of face information in ME sequences being duplicated, and an excessive amount of redundant information causing algorithm performance to degrade. Our suggested technique drastically minimises duplicate information in ME sequences, maintains the efficacy and simplicity of the features, and decreases the number of model parameters to 66 million by including differential features.

This approach is faster than the optical flow method, but it has stricter data entry requirements. The input ME sequence must not have a head with considerable jitter. In other words, the approach may attain its optimum performance when the head jitter is decreased or when the heads of several frames are aligned precisely. In the future, we also want to investigate how to eliminate head tremors more efficiently. In conclusion, while the technique described in this research has shown positive outcomes, there are still room for improvement. For instance, continue developing Debouncer to correct head jitter across a broader range; refine the differential feature extraction approach to increase the feature variety.

Author Contributions: Conceptualization, M.M.A. and S.A.; methodology, R.H.; software, I.K.; validation, M.M.A. and S.A.; formal analysis, I.K.; investigation, R.H.; resources, M.M.A.; data curation, M.M.A.; writing—original draft preparation, M.M.A.; writing—review and editing, I.K.; visualization, R.H.; supervision, S.A.; project administration, M.M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.
Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception,"Psychiatry, vol. 32, no. 1, pp. 88-106, 1969.
2. P. Ekman, Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition). WW Norton & Company, 2009.
3. M. G. Frank and P. Ekman, "The ability to detect deceit generalizes across different types of high-stake lies," Journal of personality and social psychology, vol. 72, no. 6, p. 1429, 1997.
4. T. A. Russell, E. Chu, and M. L. Phillips, "A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool," British journal of clinical psychology,vol. 45, no. 4, pp. 579-583, 2006.
5. M. G. Frank, C. J. Maccario, and V. Govindaraju, "Behavior and security,"Protecting airline passengers in the age of terrorism, pp. 86-106, 2009.
6. G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 6, pp. 915-928, 2007.
7. S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor," 2009.
8. W.-J. Yan et al., "CASMEII: An improved spontaneous micro-expression database and the baseline evaluation," PloS one, vol. 9, no. 1, p. e86041, 2014.
9. X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in 2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg), 2013: IEEE, pp. 1-6.
10. A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Samm: A spontaneous micro-facial movement dataset," IEEE transactions on affective computing, vol. 9, no. 1, pp. 116-129, 2016.
11. W. Merghani, A. Davison, and M. Yap, "Facial Micro-expressions Grand Challenge 2018: evaluating spatio-temporal features for classification of objective classes," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018: IEEE, pp. 662-666.
12. J. Li, S.-J. Wang, M. H. Yap, J. See, X. Hong, and X. Li, "MEGC2020-The Third Facial Micro-Expression Grand Challenge," in 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG2020)(FG), 2020: IEEE Computer Society, pp. 234-237.
13. J. See, M. H. Yap, J. Li, X. Hong, and S.-J. Wang, "Megc 2019-the second facial micro-expressions grand challenge," in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG2019), 2019: IEEE, pp. 1-5.
14. L. Zhou, X. Y. Shao, and Q. R. Mao, "A survey of micro-expression recognition," (in English), Image and Vision Computing, vol. 105, Jan 2021.
15. T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in 2011 international conference on computer vision, 2011: IEEE, pp. 1449-1456.
16. S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, and X. Fu, "Micro-expression recognition using dynamic textures on tensor independent color space," in 2014 22nd international conference on pattern recognition, 2014: IEEE, pp. 4678-4683.
17. S.-J. Wang et al., "Micro-expression recognition using color spaces," IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 6034-6047, 2015.
18. Y. Guo, Y. Tian, X. Gao, and X. Zhang, "Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method," in 2014 international joint conference on neural networks (IJCNN), 2014: IEEE, pp. 3473-3479.

19. Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition," in Asian conference on computer vision, 2014: Springer, pp. 525-537.
20. X. Huang, S.-J. Wang, G. Zhao, and M. Piteikainen, "Facial microexpression recognition using spatiotemporal local binary pattern with integral projection," in Proceedings of the IEEE international conference on computer vision workshops, 2015, pp. 1-9.
21. Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," IEEE Transactions on Affective Computing, vol. 7, no. 4, pp. 299-310, 2015.
22. F. Xu, J. Zhang, and J. Z. Wang, "Microexpression identification and categorization using a facial dynamics map," IEEE Transactions on Affective Computing, vol. 8, no. 2, pp. 254-267, 2017.
23. S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Less is more: Micro-expression recognition from video using apex frame," Signal Processing: Image Communication, vol. 62, pp. 82-92, 2018.
24. S. Happy and A. Routray, "Fuzzy histogram of optical flow orientations for micro-expression recognition," IEEE Transactions on Affective Computing, vol. 10, no. 3, pp. 394-406, 2017.
25. Y. Su, J. Zhang, J. Liu, and G. Zhai, "Key Facial Components Guided Micro-Expression Recognition Based on First & Second-Order Motion," in 2021 IEEE International Conference on Multimedia and Expo (ICME), 2021: IEEE, pp. 1-6.
26. B. Yang, J. Cheng, Y. Yang, B. Zhang, and J. Li, "MERTA: micro-expression recognition with ternary attentions," Multimedia Tools and Applications, vol. 80, no. 11, pp. 1-16, 2021/05/01 2021.
27. Y. Wang et al., "Micro Expression Recognition via Dual-Stream Spatiotemporal Attention Network," Journal of Healthcare Engineering, vol. 2021, 2021.
28. M. A. Takalkar, M. Xu, and Z. Chaczko, "Manifold feature integration for micro-expression recognition," Multimedia Systems, vol. 26, no. 5, pp. 535-551, 2020.
29. N. Van Quang, J. Chun, and T. Tokuyama, "Capsulenet for microexpression recognition," in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 2019: IEEE, pp. 1-7.
30. Y. Zhao and J. Xu, "A convolutional neural network for compound micro-expression recognition," Sensors, vol. 19, no. 24, p. 5553, 2019.
31. L. Lei, T. Chen, S. Li, and J. Li, "Micro-Expression Recognition Based on Facial Graph Representation Learning and Facial Action Unit Fusion," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1571-1580.
32. M. Bai and R. Goecke, "Investigating LSTM for Micro-Expression Recognition," in Companion Publication of the 2020 International Conference on Multimodal Interaction, 2020, pp. 7-11.
33. H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "Au-assisted graph attention convolutional network for micro-expression recognition," in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2871-2880.
34. L. Lo, H.-X. Xie, H.-H. Shuai, and W.-H. Cheng, "MER-GCN: Micro-expression recognition based on relation modeling with graph convolutional networks," in 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2020: IEEE, pp. 79-84.
35. Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," IEEE Transactions on Multimedia, vol. 22, no. 3, pp. 626-640, 2019.
36. M. Bai, R. Goecke, and D. Herath, "Micro-Expression Recognition Based On Video Motion Magnification And Pre-Trained Neural Network," in 2021 IEEE International Conference on Image Processing (ICIP), 2021: IEEE, pp. 549-553.
37. C. Wu and F. Guo, "TSNN: Three-Stream Combining 2D and 3D Convolutional Neural Network for Micro-Expression Recognition," IEEJ Transactions on Electrical and Electronic Engineering, vol. 16, no. 1, pp. 98-107, 2021.
38. B. Song et al., "Recognizing spontaneous micro-expression using a three-stream convolutional neural network," IEEE Access, vol. 7, pp. 184537-184551, 2019.
39. N. Liu, X. Liu, Z. Zhang, X. Xu, and T. Chen, "Offset or Onset Frame: A Multi-Stream Convolutional Neural Network with CapsuleNet Module for Micro-expression Recognition," in 2020 5th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), 2020: IEEE, pp. 236-240.
40. X. Nie, M. A. Takalkar, M. Duan, H. Zhang, and M. Xu, "GEME: Dual-stream multi-task GEnder-based micro-expression recognition," Neuro-computing, vol. 427, pp. 13-28, 2021.
41. J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3D flow convolutional neural network," Pattern Analysis and Applications, vol. 22, no. 4, pp. 1331-1339, 2019.
42. S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," Advances in neural information processing systems, vol. 30, 2017.
43. V. Mazzia, F. Salvetti, and M. Chiaberge, "Efficient-capsnet: Capsule network with self-attention routing," Scientific reports, vol. 11, no. 1, pp. 1-13, 2021.

 doi:10.20944/preprints202211.0488.v1

44. Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," in 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019), 2019: IEEE, pp. 1-4.
45. S.-T. Liong, Y. S. Gan, J. See, H.-Q. Khor, and Y.-C. Huang, "Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition," in 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019), 2019: IEEE, pp. 1-5.
46. L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 2019: IEEE, pp. 1-5.
47. L. Zhou, Q. Mao, X. Huang, F. Zhang, and Z. Zhang, "Feature refinement: An expression-specific feature learning and fusion method for microexpression recognition," Pattern Recognition, vol. 122, p. 108275, 2022.
48. Y. Wang, H. Ma, X. Xing, and Z. Pan, "Eulerian motion based 3dcnn architecture for facial micro-expression recognition," in International Conference on Multimedia Modeling, 2020: Springer, pp. 266-277.
49. L. Lei, J. Li, T. Chen, and S. Li, "A novel graph-tcn with a graph structured representation for micro-expression recognition," in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2237-2245.
50. X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," Neurocomputing, vol. 175, pp. 564-578, 2016.
51. R. Zhi, H. Xu, M. Wan, and T. Li, "Combining 3D convolutional neural networks with transfer learning by supervised pre-training for facial micro-expression recognition," IEICE Transactions on Information and Systems, vol. 102, no. 5, pp. 1054-1064, 2019.