Enriched Long-term Recurrent Convolutional Network for Facial Micro-Expression Recognition

Huai-Qian Khor[1], John See[2], Raphael C.W. Phan[3], Weiyao Lin[4]
[1,2] Faculty of Computing and Informatics, Multimedia University, Malaysia
[3] Faculty of Engineering, Multimedia University, Malaysia
[4] Department of Electronic Engineering, Shanghai Jiao Tong University, China
Emails: [1]hqkhor95@gmail.com, [2]johnsee@mmu.edu.my, [3]raphael@mmu.edu.my, [4]wylin@sjtu.edu.cn

***Abstract*—Facial micro-expression (ME) recognition has posed a huge challenge to researchers for its subtlety in motion and limited databases. Recently, handcrafted techniques have achieved superior performance in micro-expression recognition but at the cost of domain specificity and cumbersome parametric tunings. In this paper, we propose an Enriched Long-term Recurrent Convolutional Network (ELRCN) that first encodes each micro-expression frame into a feature vector through CNN module(s), then predicts the micro-expression by passing the feature vector through a Long Short-term Memory (LSTM) module. The framework contains two different network variants: (1) Channel-wise stacking of input data for spatial enrichment, (2) Feature-wise stacking of features for temporal enrichment. We demonstrate that the proposed approach is able to achieve reasonably good performance, without data augmentation. In addition, we also present ablation studies conducted on the framework and visualizations of what CNN "sees" when predicting the micro-expression classes.**

***Keywords*-Micro-expression recognition; objective classes; LRCN; network enrichment, cross-database evaluation**

I. Introduction

Facial micro-expressions (ME) are brief and involuntary rapid facial emotions that are elicited to hide a certain true emotion [1]. A standard micro-expression lasts between 1/5 to 1/25 of a second and usually occurs in only specific parts of the face [2]. The subtleness and brevity of micro-expressions are a great challenge to the naked eye; hence, a lot of works have been proposed in recent years to utilize computer vision and machine learning algorithms in attempt to achieve automated micro-expression recognition.

The establishment of Facial Action Coding System (FACS) [3] encodes the facial muscle changes to emotion states. The system also establishes a ground truth of the exact begin and end time of each action unit (AU). Different databases [4], [5], [6] may contain different micro-expression classes which are labeled by trained coders based on the presence of AUs. However, a recent discourse by Davison et al. [7] argued that using AUs instead of emotion labels can define micro-expressions more precisely since the training process can learn based on specific facial muscle movement patterns. They further proved that this leads to higher classification accuracy.

In this field of research, several works [8][9][10] have achieved impressive micro-expression recognition performance. These works have proposed carefully crafted descriptors and/or methods that involved a tedious tuning of parameters to attain maximum results. In view of these unwieldy steps, the adoption of *deep learning* techniques or deep neural networks have started to take-off, as seen from several new attempts [11], [12]. However, the usage of deep neural network poses challenges to ME recognition due to the scarcity of samples and class-imbalance in most micro-expression data.

II. Related Works

A. Handcrafted Features

In the last five years, numerous works have been proposed to solve the ME recognition problem. The databases established to advance computational research in spontaneous facial micro-expression analysis i.e. SMIC [4], CASME II [5], SAMM [6], [7], have mainly chosen Local Binary Pattern with Three Orthogonal Planes (LBP-TOP) [13] as their primary baseline feature extractor. The LBP-TOP is a spatio-temporal extension of the classic Local Binary Pattern (LBP) descriptor [14], which characterizes the local textural information by encoding a vector of binary code into histograms. LBP-TOP extracts the said histograms from each of the three planes (XY, XT, YT) and concatenate them into a single feature histogram. The LBP, whilst known for its simplicity in computation, is vastly used because of its robustness towards illumination changes and image transformations.

Wang et al. [15] reduced the redundancies in the LBP-TOP by utilizing only six intersection points in the 3D plane to construct the feature descriptor. Later on, Huang et al. [10] proposed a Spatio-Temporal LBP with Integral Projection (STLBP-IP) that applies the LBP operator on horizontal and vertical projections based on difference images. Their method is shape-preserving and is robust against the white noise and image transformations.

Several works have used LBP-TOP with an accompanying pre-processing technique. Most widely seen is the Temporal

978-1-5386-2335-0/18/$31.00

Interpolation Model [4] which is used to sample uniformly a fixed number of image frames from the constructed data manifold. Recently, [16] proposed Sparsity Promoting Dynamic Mode Decomposition (DMDSP) which acts to select only the significant temporal dynamics when synthesizing a dynamically condensed sequence. A number of other works [17], [18] opt to magnify the video in attempt to accentuate the subtle changes before feature extraction.

Motion information can readily portray the subtle changes exhibited by micro-expressions. Shreve et al. [19] proposed the extraction of a derivative of optical flow called an *optical strain* which was originally used for ME spotting but later adopted as a feature descriptor for ME recognition [20], [21]. Leveraging on the discriminativeness of optical flow, other interesting approaches have come to the fore, among them are Bi-Weighted Oriented Optical Flow (Bi-WOOF) [8] and Facial Dynamics Map [22].

B. Deep Neural Networks

The utilization of deep learning techniques or deep neural networks is fairly new to this field of research despite its widespread popularity in recognition tasks.

One early work [11] to utilize deep learning proposed an expression-state based feature representation. The researchers adopted Convolutional Neural Networks (CNN) to encode different expression states (i.e., onset, onset to apex, apex, apex to offset and offset). Several objective functions are optimized during spatial learning to improve expression class separability. After that, the encoded features are passed to a Long Short-Term Memory (LSTM) network to learn time scale dependent features.

Recently, Peng et al. [12] proposed a two-stream 3-D CNN model called Dual Temporal Scale Convolutional Neural Network (DTSCNN). Different streams of the framework were used to adapt to different frame rates of ME video clips. The authors aggregated both CASME I and II databases, likely to provide sufficient samples for meaningful training to take place. The network was also designed to be shallower to avoid overfitting problem, while optical flow was used to enrich the input data. These two approaches provide the motivation towards the design of our proposed method.

III. Proposed Framework

In this work, we propose an Enriched Long-term Recurrent Convolutional Network (ELRCN) for micro-expression recognition, which adopts the architecture of [23] whilst performing feature enrichment to encode subtle facial changes. The ELRCN model comprises of a deep hierarchical spatial feature extractor and a temporal module that characterizes temporal dynamics. Two variants of the network are introduced: 1) Enrichment of the spatial dimension by input channel stacking, 2) Enrichment of the temporal dimension by deep feature stacking. Figure 1 summarizes the proposed framework with the preprocessing module and both variants of learning module.

A. Preprocessing

The micro-expression videos are first preprocessed using TV-L1 [24] method for optical flow approximation, which has two major advantages: better noise robustness and preservation of flow discontinuities. Optical flow encodes motion of an object in vectorized notations, indicating the direction and intensity of the motion or 'flow' of image pixels. The horizontal and vertical components of the optical flow are defined as follow:

$$\vec{v} = [p = \frac{dx}{dt}, q = \frac{dy}{dt}]^T \tag{1}$$

where dx and dy represent the estimated changes in pixels along the x and y dimension respectively while *dt* represent the change in time. To form a 3-dimensional *flow image*, we concatenate the horizontal and vertical flow images, **p** and **q** and the optical flow magnitude, $\mathbf{m} = |v|$. Normalization of the flow image is not necessary in our case since motion changes are very subtle (not occupying large range of values); this was also proven empirically with negligible drop in performance.

We also obtained the optical strain [19] by computing the derivatives of the optical flow. By employing optical strain, we are able to properly characterize the tiny amount of movement of a deformable object present between two successive frames. This is described by a displacement vector, $\mathbf{u} = [u, v]^T$. The finite strain tensor is defined as:

$$\epsilon = \frac{1}{2}[\nabla u + (\nabla u)^T] \tag{2}$$

or in expanded tensor form:

$$\epsilon = \begin{bmatrix} \epsilon_{xx} = \frac{\delta u}{\delta x} & \epsilon_{xy} = \frac{1}{2}(\frac{\delta u}{\delta y} + \frac{\delta v}{\delta x}) \\ \epsilon_{yx} = \frac{1}{2}(\frac{\delta v}{\delta x} + \frac{\delta u}{\delta y}) & \epsilon_{yy} = \frac{\delta v}{\delta y} \end{bmatrix} \tag{3}$$

where the diagonal strain components, (ϵ_{xx}, ϵ_{yy}), are normal strain components and (ϵ_{xy}, ϵ_{yx}) are the shear strain components. Normal strain measures changes along x and y directions whereas shear strain measures changes in the angles caused by deformation along both axis.

The optical strain magnitude for each pixel can be computed using the sum of squares of the normal and shear strain components:

$$|\epsilon| = \sqrt{\epsilon_{xx}^2 + \epsilon_{yy}^2 + \epsilon_{xy}^2 + \epsilon_{yx}^2} \tag{4}$$

B. Spatial Learning

Recent deep models [25], [26], [27], [28] have proven that the composition of numerous "layers" of non-linear functions can achieve ground-breaking results for various computer vision problems such as object recognition and

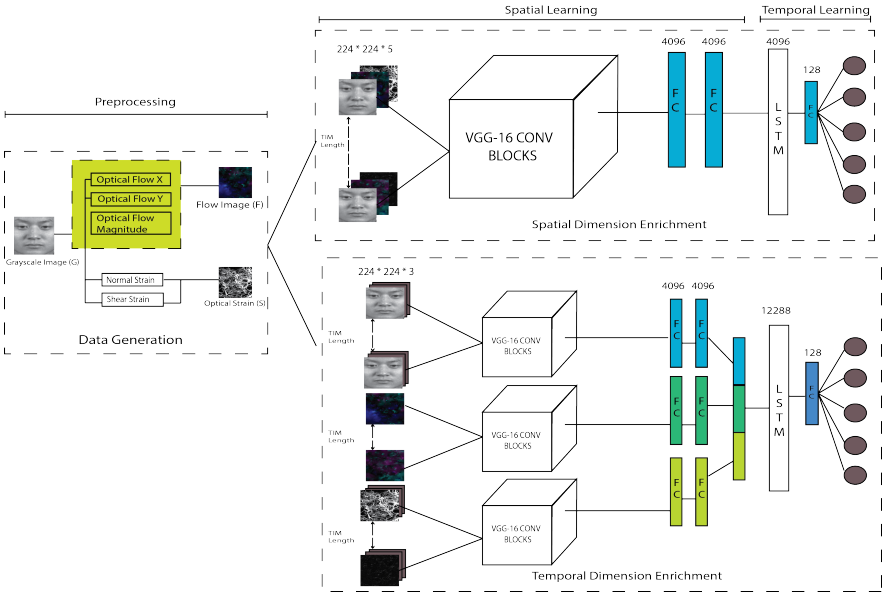


Figure 1. Proposed ELRCN framework

object detection. To leverage the benefit of deep convolutional neural networks (CNN) in a sequential fashion, the input data x is first encoded with a CNN to a fixed-length vector, $\phi(x_t)$ that represents the spatial features at time t. Subsequently, $\phi(x_t)$ is then passed to a recurrent neural network to learn the temporal dynamics.

In this paper, we also hypothesize that by using additional derivative information of the raw input sample, in a process that involves *sample enrichment*, we can minimize underfitting in the learned models, which in turn can result in higher recognition performance. Figure 1 depicts the overall framework of our proposed Enriched Long-term Recurrent Convolutional Network (ELRCN) in two possible variants: Spatial Dimension Enrichment (SE) and Temporal Dimension Enrichment (TE).

The SE model uses a larger input data dimension for spatial learning by stacking an optical flow image ($F \in \mathbb{R}^3$), an optical strain image ($S \in \mathbb{R}^2$) and a gray-scale raw image ($R \in \mathbb{R}^2$) along the input channel, which we denote as $x_t = (F_t, S_t, G_t)$. Hence, the input data is 224 * 224 * 5, which necessitates training the VGG-Very-Deep-16 (VGG-16) [29] model from scratch. The last fully connected (FC) layer encodes the input data into a 4096 fixed-length vector, $\phi(x_t)$.

The TE model utilizes transfer learning [30] with pre-trained weights from VGG-Face model [31] which was trained on a large-scale Labeled Faces in the Wild (LFW) dataset [32] for the purpose of face identification. We fine-tuned the micro-expression data on the pre-trained weights of VGG-Face to allow the model to learn and adapt more effectively. This also facilitates faster convergence because both micro-expression and LFW data involve faces and their components. Since the VGG-Faces model expects a 224 * 224 * 3 input, we duplicated the S and G images ($\mathbb{R}^2 \rightarrow \mathbb{R}^3$) so that they fit the required input dimension (as shown in Figure 1). During the training phase, we fine-tune each input data in separate VGG-16 models with each model yielding a 4096-length feature vector, $\phi(x_t)$ at their last FC layer. This results in a 12288-length feature vector to be passed to the subsequent recurrent network.

C. Temporal Learning

In the current micro-expression domain, several works [33], [21] aimed to preserve the temporal dimension as its dynamics is crucial for recognizing facial movements. We use a popular variant of the recurrent neural network called Long Short-Term Memory (LSTM) [34] to learn the spatially-encoded sequential input, $\phi(x_t)$. The LSTM seeks to learn weights parameters W that maps the input $\phi(x_t)$ at a previous time step hidden state h_{t-1} to an output z_t and updated hidden state h_t. LSTM layers can be stacked serially, followed by a fully connected layer that encodes z_t into a smaller dimension, $\hat{y} = W_z z_t + b_z$. Finally, the prediction $P(y_t)$ is computed with softmax of $\hat{y}_t$:

$$P(y_t = c) = \text{softmax}(\hat{y}_t) = \frac{exp(\hat{y}_t, c)}{\sum_{c' \in C} exp(\hat{y}_t, c')} \qquad (5)$$

where C is a discrete, finite set of outcomes and $y_t \in C$.

D. General Network Configuration

The networks are trained using adaptive epochs or early stopping with a maximum set to 100 epochs. Basically, the training for each fold will stop when the loss score stops improving. We use Adaptive Moment Estimation (ADAM)

[35] as the optimizer, with a learning rate of 10^{-5} and decay of 10^{-6}. The learning rate is tuned to be smaller than typical rates because of the subtleness of micro-expression which poses difficulty for learning. For temporal learning, we fix the number of FC layers after the LSTM layers to one. This is not experimented as our focus is on the number of recurrent layers and units in these layers (see the ablation study in Section IV-E).

IV. Evaluation

A. Databases

CASME II [5] is a comprehensive spontaneous micro-expression database containing 247 video samples, elicited from 26 Asian participants with an average age of 22.03 years old. The videos in this database showed a participant evoked by one of five categories of micro-expressions: Happiness, Disgust, Repression, Surprise, Others.

The Spontaneous Actions and Micro-Movements (SAMM) [6] is a newer database of 159 micro-movements (one video for each) induced spontaneously from a demographically diverse group of 32 participants with a mean age of 33.24 years, and an even male-female gender split. Originally intended for investigating micro-facial movements, the SAMM was induced based on the 7 basic emotions. Eventually, the authors [7] proposed "objective classes" based on the FACS Action Units as categories for micro-expression recognition.

Both the CASME II and SAMM databases have much in common: They are recorded at a high speed frame rate of 200 *fps*, and they have objective classes, as provided in [7].

B. Preprocessing & Settings

The SAMM dataset is preprocessed with Dlib [36] for face alignment while facial landmarks are extracted using Face++ API [37]. Then, each video frame is cropped based on selected facial landmarks at the edge of the face. Meanwhile, CASME II provides pre-cropped video frames which we make use directly. All video frames are resized to $224 * 224$ pixel resolution to match the input spatial dimension to the network. Temporal Interpolation Model (TIM) [4] of length 10 was applied to both databases to fit the sample sequence into the recurrent model that expects a fixed temporal length. The baseline methods that we compared with were implemented using a Support Vector Machine (SVM) with linear kernel and a large regularization parameter of C=10000.

We perform two sets of experiments: (1) Single domain experiment involving only one database (CASME II), (2) Cross domain experiment involving two databases (CASME II and SAMM), specifically, two settings were used – one which holds out one database at each time, another which combines all samples from both databases.

Experiments are measured using F1-Score, Weighted Average Recall (WAR) or Accuracy, and Unweighted Average Recall (UAR). UAR is akin to a "balanced" accuracy (averaging the accuracy scores of each individual class without consideration of class size). We report micro-averaged F1-Score, which provides a balanced metric when considering highly imbalanced data [38].

C. Single Domain Experiment

In this experiment, the CASME II database is our choice of domain for evaluation. Training was performed using Leave-One-Subject-Out (LOSO) cross validation as this protocol prevents subject bias during learning. Table I compares the performance of our proposed methods against the baseline LBP-TOP method (reproduced) and a number of recent and relevant works in literature. The TE variant of the proposed ELRCN method clearly outperforms its SE counterpart, which shows the importance of fine-tuning separate networks for each type of data.

Table I
Performance of proposed methods vs. other methods for micro-expression recognition

Methods	F1-Score	UAR	Accuracy / WAR
LBP-TOP (reproduced)	0.2941	0.3094	0.4595
Adaptive MM + LBP-TOP [39]	N/A	N/A	0.5191
FDM [22]	0.4053	N/A	0.4593
LBP-SIP	0.4480	N/A	0.4656
EVM+HIGO [17]	N/A	N/A	**0.6721**
CNN-LSTM [11]	N/A	N/A	0.6098
ELRCN-SE	0.4547	0.3895	0.4715
ELRCN-TE	**0.5000**	**0.4396**	0.5244

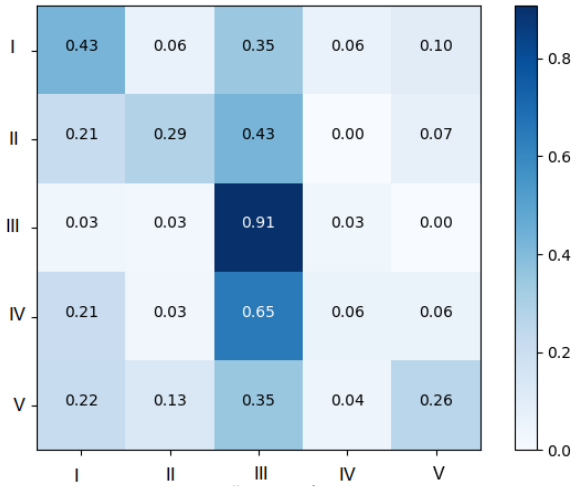


Figure 2. Confusion Matrix of ELRCN-SE on CDE protocol.

D. Cross Domain Experiment

To test the robustness of our deep neural network architecture and its ability to learn salient characteristics from the samples, we use two cross domain protocols introduced

Table II
EXPERIMENTAL RESULTS FOR CDE EVALUATION

Methods	F1-Score	UAR	Accuracy / WAR
LBP-TOP (reproduced)	0.3172	0.3224	0.4229
ELRCN-SE	**0.4107**	**0.3900**	**0.5700**
ELRCN-TE	0.3616	0.3300	0.4700

by the Micro-Expression Grand Challenge (MEGC) 2018[1] – Composite Database Evaluation (CDE) and Holdout-Database Evaluation (HDE). HDE and CDE are Tasks A and B respectively in MEGC 2018. CDE combines both databases (CASME II and SAMM), which totals to 47 subjects after omitting the 6th and 7th objective classes (from [7]) followed by a LOSO evaluation. HDE samples the training and test sets from opposing database (i.e, trains on CASME II and test on SAMM, and vice versa). The results from both folds are then averaged and reported as the overall result.

Table II compares the performance of our two ELRCN variants against the reproduced LBP-TOP baseline on the CDE (Task B) protocol. The proposed methods are clearly superior in generalizing over a large number of subjects as compared to the baseline method. Interestingly, the SE variant posts a much stronger result (WAR 0.57) than the TE variant; this in contrast to results on CASME II alone.

Table III shows the result for the HDE (Task A) protocol. The HOG-3D and HOOF methods were provided by the challenge organizers as other competing baselines. We also reproduced the baseline LBP-TOP method which differed from the results provided by the challenge organizers. This is likely to due to some differences in the face cropping steps or preprocessing steps (such as TIM) which were not disclosed in detail at the time of writing. Similarly, we observe a strong performance from the SE variant of the proposed approach, which surpasses that of the TE variant and the provided baselines.

To better understand what goes on under the hood, we provide the confusion matrix for ELRCN-SE with CDE protocol in 2. Class I and class III have the best results possibly due to larger amount of training samples. Besides, we also provide the confusion matrices for both folds (i.e. train-test pairings of CASME II-SAMM and SAMM-CASME II) in Figures 3 and 4. The CASME II-SAMM fold (F1 0.409, UAR 0.485, WAR 0.382) had noticeably better performance than the SAMM-CASME II fold (F1 0.274, UAR 0.384, WAR 0.322). Class III of CASME II has the most training samples; it performed the best. Likewise, classes that were relatively under-represented in the training set (class II from CASME II, classes IV and V from SAMM) performed very poorly. Hence, it is likely that the small sample size remains a stumbling block for deep learning based approaches.

[1]http://www2.docm.mmu.ac.uk/STAFF/m.yap/FG2018Workshop.htm

Table III
EXPERIMENTAL RESULTS FOR HDE EVALUATION

Methods	F1-Score	UAR	Accuracy / WAR
LBP-TOP (reproduced)	0.2162	0.2179	0.3891
LBP-TOP (provided)	N/A	0.322	0.285
HOG-3D	N/A	0.228	0.363
HOOF	N/A	0.348	0.353
ELRCN-SE	**0.3411**	**0.3522**	**0.4345**
ELRCN-TE	0.2389	0.2221	0.2320

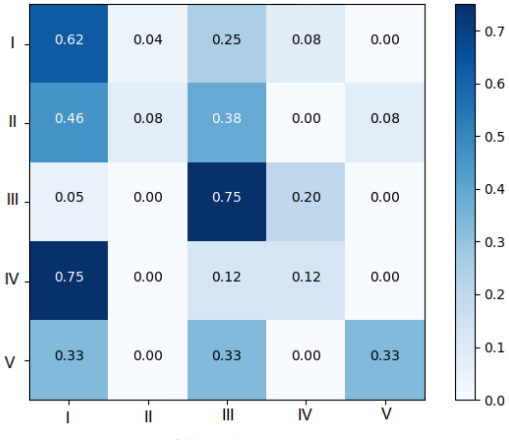


Figure 3. Confusion matrix of ELRCN-SE on HDE protocol, training on CASME II and testing on SAMM database.

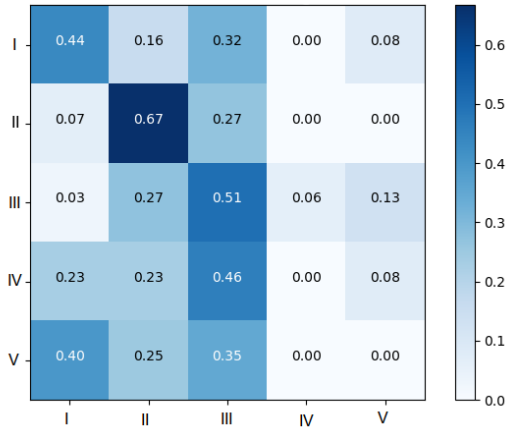


Figure 4. Confusion matrix of ELRCN-SE on HDE protocol, training on SAMM and testing on CASME II database.

E. Ablation Study

For further analysis, we perform an extensive ablation study by removing certain portions of our proposed ELRCN to see how that affects performance. This was carried out using the CASME II database (single domain).

1) Spatial Learning Only: We learn only with the VGG-16 CNN to observe the capability of the spatial module on its own. We regard each video frame as individual images instead of a sequence. Results in Figure 5 on different configurations of the spatial module show that spatial-only

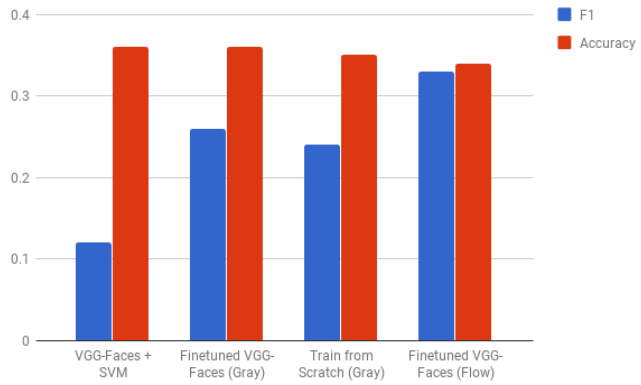


Figure 5. Recognition performance using spatial module only

performances can be poorer than that of the baseline.

2) Temporal Learning Only: The images are resized to 50*50 pixel resolution since recurrent models with large number of recurrent units are computationally demanding. We consider the pixel intensities as the basic representation of the samples as input to the temporal module. A variety of configurations were considered, including both 1 and 2-layer LSTMs. Results in Figure 6 show that the baseline performance can be surpassed by just using pixel intensities as input to 2-layer LSTM networks. With reference to the spatial-only approaches, the importance of temporal dynamics is quite telling, as can be seen here.

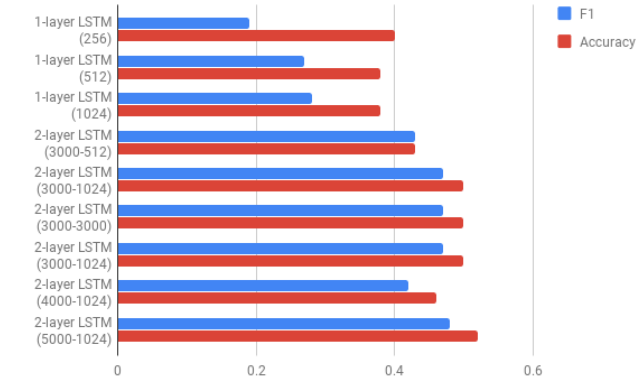


Figure 6. Recognition performance using temporal-only module. The numbers in bracket indicate the number of recurrent units for each LSTM layer.

3) Spatio-Temporal LRCN: From the first two studies, we proceed to gauge the performance of the proposed method (SE variant) by fixing one of the two modules to a reasonably good choice of method and varying the other.

Using Flow data only (the best from spatial-only study), we tested using spatial features from the last and second

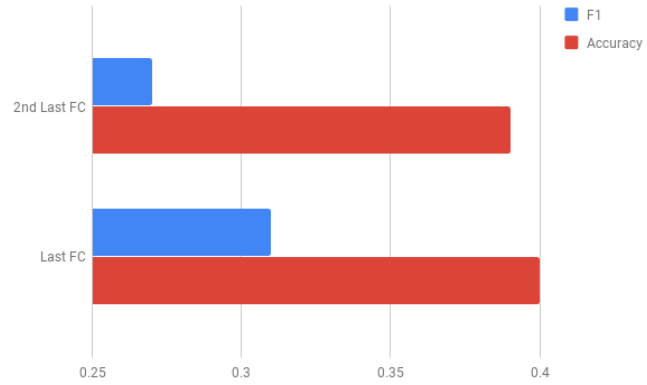


Figure 7. Recognition performance using Flow features encoded at different FC layers (spatial module), on a 2-layer LSTM (3000-1024).

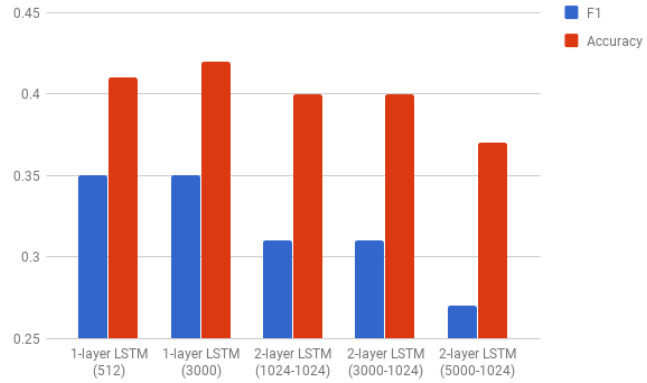


Figure 8. Recognition performance of different temporal recurrent networks using spatial features from the last FC of the VGG-16 CNN.

last fully-connected (FC) layers of the VGG-16 CNN on a 2-layer LSTM (3000-1024), which is the best architecture as of temporal-only study (see Fig. 6). Results in 7 show that the spatial features are the most discriminative when taken from the 4096-length last FC layer. Following this, the opposing study proceeds to test this selected spatial feature against a number of temporal network architectures. Results in Figure 8 shows an interesting case of a single layer LSTM performing better than 2-layer LSTMs in an ELRCN framework, when image-based features are used instead of pixel intensities. Additionally, we note that using more recurrent units also do not necessarily produce better results, but with the sure certainty of an increase in computational cost.

These studies reveal that both spatial and temporal modules have different roles to play within the framework, and they are highly dependent on each other to attain a good level of performance.

V. Discussion

Using more data: The limitations of deep learning techniques is most obvious in the aspect of sample size. Typical deep architectures require a large amount of data to learn well. We experimented with the use of more interpolated frames (higher TIM), but it resulted in poorer results than what was recommended by earlier works [4], [38], i.e. TIM of 10 or 15. However, we do expect some improvement if appropriate data augmentation is used on our proposed network.

Visualizations: To better "see" how the proposed network arrived at its predictions, we utilize the gradient-weighted class activation mapping (Grad-CAM) [40] on the last convolutional layer of the spatial network to provide visual explanations as to which parts of the face are contributing towards the classification decision. The visualizations in Figure 9 are colored based on colors from the visible light spectrum, ranging from blue (not activated) to red (highly activated). The activations correspond to spatial locations that contribute most to the predicted class.

We first show the visualizations from the single domain experiment. The AU 12 (lip corner puller) from the sample in Figure 9(a) correspond quite precisely with the greenish regions near the side of lips. The area around the cheeks of the subject in Figure 9(b) also show relatively strong activations which corresponds to AU 14, the ground truth.

From the cross domain experiment, we also found similar evidence of AU-matching spatial activations from Figures 9(c) and (d). The AUs for Figure 9(c) are 4, 6, 7, 23, which involves movements around eye regions and upper cheek, both of which are reddish strong. Meanwhile, the sample in Figure 9(d) has AU 1 that involves raising eyebrows. Comparing the Grad-CAMs of a same sample on different experiments (shown in Figure 9(e)) generally indicate that models trained on a single domain had more salient locations than that on cross domain.

VI. Conclusion

In this paper, we have proposed two variants of an Enriched LRCN model for micro-expression recognition – one which stacks various input data for spatial enrichment (SE), another which stacks features for temporal enrichment (TE). Empirically, the TE model performs better on a single database while the SE model learns better in cross domain. The Grad-CAM visualization on selected samples demonstrate that the predictions from these models somewhat conform to the AUs marked by experts. Through our ablation study, we also discover that using optical flow information is more beneficial than using raw pixel intensities in providing proper characterization of the input data to the network. In future, we hope to extend our preliminary work with appropriate data augmentation and preprocessing techniques.

Acknowledgments

This work was supported in part by MOHE Grant FRGS/1/2016/ICT02/MMU/02/2 Malaysia and Shanghai 'The Belt and Road' Young Scholar Exchange Grant (17510740100). The authors would like to thank the anonymous reviewers for their helpful and constructive comments. We are also grateful to our lab colleagues for sharing computational resources.

References

[1] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.

[2] ——, "Constants across cultures in the face and emotion." *Journal of Personality and Social Psych.*, vol. 17, no. 2, p. 124, 1971.

[3] ——, "Facial action coding system," 1977.

[4] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *IEEE FG*, 2013, pp. 1–6.

[5] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, no. 1, p. e86041, 2014.

[6] A. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing*, 2016.

[7] A. K. Davison, W. Merghani, and M. H. Yap, "Objective classes for micro-facial expression recognition," *arXiv preprint arXiv:1708.07549*, 2017.

[8] S.-T. Liong, J. See, R. C.-W. Phan, and K. Wong, "Less is more: Micro-expression recognition from video using apex frame," *arXiv preprint arXiv:1606.01721*, 2016.

[9] X. Huang, S. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikainen, "Spontaneous facial micro-expression recognition using discriminative spatiotemporal local binary pattern with an improved integral projection," *arXiv preprint arXiv:1608.02255*, 2016.

[10] X. Huang, S.-J. Wang, G. Zhao, and M. Piteikainen, "Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection," in *IEEE ICCV Workshops*, 2015, pp. 1–9.

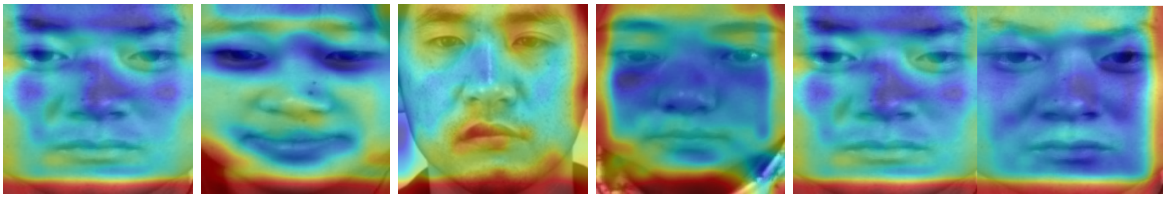
[11] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proc. of the ACM MM*. ACM, 2016, pp. 382–386.

[12] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu, "Dual temporal scale convolutional neural network for micro-expression recognition," *Frontiers in Psychology*, vol. 8, p. 1745, 2017.

[13] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. on PAMI*, vol. 29, no. 6, pp. 915–928, 2007.

[14] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE PAMI*, vol. 24, no. 7, pp. 971–987, 2002.

[15] 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 *ACCV*, 2014, pp. 525–537.



(a) (b) (c) (d) (e)

Figure 9. Grad-CAM visualizations of the ELRCN-TE model: *from left:* (a) Subject 1 (CASME II), Happiness; (b) Subject 2 (CASME II), Others; (c) Subject 13 (SAMM), Class III; (d) Subject 5 (CASME II), Objective class III, CDE protocol (e) Subject 1 (CASME II), Happiness, Comparison between single domain and cross domain experiments.

[16] A. C. Le Ngo, J. See, and R. C.-W. Phan, "Sparsity in dynamics of spontaneous subtle emotions: analysis and application," *IEEE Trans. on Affective Computing*, vol. 8, no. 3, pp. 396–411, 2017.

[17] X. Li, H. Xiaopeng, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikainen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Trans. on Affective Computing*, 2017.

[18] Y. Wang, J. See, Y.-H. Oh, R. C.-W. Phan, Y. Rahulamathavan, H.-C. Ling, S.-W. Tan, and X. Li, "Effective recognition of facial micro-expressions with video motion magnification," *Multimedia Tools and Applications*, vol. 76, no. 20, pp. 21 665–21 690, 2017.

[19] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro- and micro-expression spotting in long videos using spatio-temporal strain," in *IEEE FG*, 2011, pp. 51–56.

[20] S.-T. Liong, J. See, R. C.-W. Phan, A. C. Le Ngo, Y.-H. Oh, and K. Wong, "Subtle expression recognition using optical strain weighted features," in *ACCV*. Springer, 2014, pp. 644–657.

[21] S.-T. Liong, J. See, R. C.-W. Phan, Y.-H. Oh, A. C. Le Ngo, K. Wong, and S.-W. Tan, "Spontaneous subtle expression detection and recognition based on facial strain," *Signal Processing: Image Communication*, vol. 47, pp. 170–182, 2016.

[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] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. of IEEE CVPR*, 2015, pp. 2625–2634.

[24] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l 1 optical flow," *Pattern Recognition*, pp. 214–223, 2007.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[26] ——, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE CVPR*, 2016, pp. 770–778.

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich *et al.*, "Going deeper with convolutions," in *IEEE CVPR*, pp. 1–9.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[30] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 17–36.

[31] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[32] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.

[33] H. Zheng, X. Geng, and Z. Yang, "A relaxed k-svd algorithm for spontaneous micro-expression recognition," in *PRICAI*. Springer, 2016, pp. 692–699.

[34] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," 1999.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[37] M. T. Inc., "Face++ cognitive services." [Online]. Available: https://www.faceplusplus.com/

[38] A. C. Le Ngo, R. C.-W. Phan, and J. See, "Spontaneous subtle expression recognition: Imbalanced databases and solutions," in *Asian conference on computer vision*. Springer, 2014, pp. 33–48.

[39] S. Y. Park, S. H. Lee, and Y. M. Ro, "Subtle facial expression recognition using adaptive magnification of discriminative facial motion," in *Proc. of ACM MM*, 2015, pp. 911–914.

[40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," 2017.