Toward Rapid Stroke Diagnosis with Multimodal Deep Learning*

Mingli Yu[1], Tongan Cai[1]**, Xiaolei Huang[1], Kelvin Wong[2],
John Volpi[3], James Z. Wang[1], Stephen T.C. Wong[2]

[1] The Pennsylvania State University, University Park, Pennsylvania, USA
[2] TT and WF Chao Center for BRAIN & Houston Methodist Cancer Center,
Houston Methodist Hospital, Houston, Texas, USA
[3] Eddy Scurlock Comprehensive Stroke Center, Department of Neurology,
Houston Methodist Hospital, Houston, Texas, USA

Abstract. Stroke is a challenging disease to diagnose in an emergency room (ER) setting. While an MRI scan is very useful in detecting ischemic stroke, it is usually not available due to space constraint and high cost in the ER. Clinical tests like the Cincinnati Pre-hospital Stroke Scale (CPSS) and the Face Arm Speech Test (FAST) are helpful tools used by neurologists, but there may not be neurologists immediately available to conduct the tests. We emulate CPSS and FAST and propose a novel multimodal deep learning framework to achieve computer-aided stroke presence assessment over facial motion weaknesses and speech inability for patients with suspicion of stroke showing facial paralysis and speech disorders in an acute setting. Experiments on our video dataset collected on actual ER patients performing specific speech tests show that the proposed approach achieves diagnostic performance comparable to that of ER doctors, attaining a 93.12% sensitivity rate while maintaining 79.27% accuracy. Meanwhile, each assessment can be completed in less than four minutes. This demonstrates the high clinical value of the framework. In addition, the work, when deployed on a smartphone, will enable self-assessment by at-risk patients at the time when stroke-like symptoms emerge.

Keywords: Stroke · Emergency medicine · Computer vision · Facial video analysis · Machine learning

1 Introduction

Stroke is a common but fatal vascular disease. It is the second leading cause of death and the third leading cause of disability [14]. With acute ischemic stroke,

* M. Yu, T. Cai, X. Huang, and J.Z. Wang are supported by Penn State University. S.T.C. Wong and K. Wong are supported by the T.T. and W.F. Chao Foundation and the John S. Dunn Research Foundation.

** M. Yu and T. Cai made equal contributions. All correspondence should be addressed to T. Cai and J. Z. Wang. {cta,jwang}@psu.edu where parts of the brain tissue suffer from restrictions in blood supply to tissues, the shortage of oxygen needed for cellular metabolism quickly causes long-lasting damage to the areas of brain cells. The sooner a diagnosis is made, the earlier the treatment can begin and the better the outcome is expected for patients.

There is no rapid assessment approach for stroke. The gold standard test for stroke is the diffusion-weighted MRI scan that detects brain lesions, yet it is usually not accessible in the ER setting. There are two commonly adopted clinical tests for stroke in the ER, the Cincinnati Pre-hospital Stroke Scale (CPSS) [16] and the Face Arm Speech Test (FAST) [8]. Both methods assess the presence of any unilateral facial droop, arm drift, and speech disorder. The patient is requested to repeat a specific sentence (CPSS) or have a conversation with the doctor (FAST), and abnormality arises when the patient slurs, fails to organize his speech, or is unable to speak. However, the scarcity of neurologists [17] prevents such tests to be effectively conducted in all stroke emergency situations.

Recently, researchers have been striving for more accurate detection and evaluation methods for neurological disorders with the help of machine intelligence. Many researchers focused on the detection of facial paralysis. Studies have adopted 2D image analysis by facial landmarks with a criterion-based decision boundary [3,10,13]. Some used either videos [4,19], 3D information [1,15], or optical flow [20,24] as input features. Either rule-based [6,11] or learning-based [12,18] methods were applied to perform the analysis. However, existing approaches either evaluate their methods between subjects that are normal versus those with clear signs of a stroke, deal with only synthetic data, fail in capturing the spatiotemporal details of facial muscular motions, or rely on experimental settings with hard constraints on the head orientation, hindering their adoption in real clinical practice or for patient self-assessment.

We propose a new deep learning method to analyze the presence of stroke in patients with suspicion or identified with a high risk of stroke. We formulate this task as a binary classification problem, *i.e.*, stroke vs. non-stroke. With patients recorded while performing a set of vocal tests, videos are processed to extract facial-motions-only frames, and the audio is transcribed. A deep neural network is set up for the collaborative classification of the videos and transcripts.

The **main contributions** of this work are summarized in three aspects:

1. To our knowledge, this is the first work to analyze the presence of stroke among actual ER patients with suspicion of stroke using computational facial motion analysis, and is also the first attempt to adopt a natural language processing (NLP) method for the speech ability test on at-risk stroke patients. We expect this work to serve as a launchpad for more research in this area.
2. A multi-modal fusion of video and audio deep learning models is introduced with 93.12% sensitivity and 79.27% accuracy in correctly identifying patients with stroke, which is comparable to the clinical impression given by ER physicians. The framework can be deployed on a mobile platform to enable self-assessment for patients right after symptoms emerge.

3. The proposed temporal proposal of human facial videos can be adopted in general facial expression recognition tasks. The proposed multi-modal method can potentially be extended to other clinical tests, especially for neurological disorders that result in muscular motion abnormalities, expressive disorder, and cognitive impairments.

2 Emergency Room Patient Dataset

The clinical dataset for this study was acquired in the emergency rooms (ERs) of Houston Methodist Hospital by the physicians and caregivers from the Eddy Scurlock Stroke Center at the Hospital under an IRB-approved[4] study. We took months to recruit a sufficiently large pool of patients in certain emergency situations. The subjects chosen are patients with suspicion of stroke while visiting the ER. 47 males and 37 females have been recruited in a race-nonspecific way.

Each patient is asked to perform two speech tasks: 1) to repeat the sentence “it is nice to see people from my hometown” and 2) to describe a “cookie theft” picture. The ability of speech is an important indicator to the presence of stroke; if the subject slurs, mumbles, or even fails to repeat the sentence, they have a very high chance of stroke [8,16]. The “cookie theft” task has been making great success in identifying patients with Alzheimer’s-related dementia, aphasia, and some other cognitive-communication impairments [5].

The subjects are video recorded as they perform the two tasks with an iPhone X’s camera. Each video has metadata information on both clinical impressions by the ER physician (indicating the doctor’s initial judgement on whether the patient has a stroke or not from his/her speech and facial muscular conditions) and ground truth from the diffusion-weighted MRI (including the presence of acute ischemic stroke, transient ischemic attack (TIA), etc.). Among the 84 individuals, 57 are patients diagnosed with stroke using the MRI, 27 are patients who do not have a stroke but are diagnosed with other clinical conditions. In this work, we construct a binary classification task and only attempt to identify stroke cases from non-stroke cases, regardless of the stroke subtypes.

Our dataset is unique, as compared to existing ones [7,10], because our subjects are actual patients visiting the ERs and the videos are collected in an unconstrained, or “in-the-wild”, fashion. In most existing work, the images or videos were taken under experimental settings, where good alignment and stable face pose can be assumed. In our dataset, the patients can be in bed, sitting, or standing, where the background and illumination are usually not under ideal control conditions. Apart from this, we only asked patients to focus on the picture we showed to them, without rigidly restricting their motions. The acquisition of facial data in natural settings makes our work robust and practical for real-world clinical use, and ultimately empowers our method for remote diagnosis of stroke and self-assessment in any setting.

[4] This study received IRB approval: Houston Methodist IRB protocol No. Pro00020577, Penn State IRB site No. SITE00000562.

3 Methodology

We propose a computer-aided diagnosis method to assess the presence of stroke in a patient visiting ER. This section introduces our information extraction methods, separate classification modules for video and audio, and the overall network fusion mechanism.

3.1 Information Extraction

For each raw video, we propose a spatiotemporal proposal of frames and conduct a machine speech transcription for the raw audio.

Spatiotemporal proposal of facial action video: We develop a pipeline to extract frame sequences with near-frontal facial pose and minimum non-facial motions. First, we detect and track the location of the patient's face as a square bounding box. During the same process, we detect and track the facial landmarks of the patient, and estimate the pose. Frame sequences 1) with large roll, yaw or pitch, 2) showing continuously changing pose metrics, or 3) having excessive head translation estimated with optical flow magnitude are excluded. A stabilizer with sliding window over the trajectory of between-frame affine transformations smooths out pixel-level vibrations on the sequences before classification.

Speech transcription: We record the patient's speech and transcribe the recorded speech audio file using Google Cloud Speech-to-Text service. Each audio segment is turned into a paragraph of text in linear time, together with a confidence score for each word, ready for subsequent classification.

3.2 Deep Neural Network for Video Classification

Facial motion abnormality detection is essential to stroke diagnosis [10], but challenges remain in several approaches to this problem. First, the limited number of videos prevent us from training 3D networks (treating time as the 3rd dimension) such as C3D [21] and R3D [22], because these networks have a large number of parameters and their training can be difficult to converge with a small dataset. Second, although optical flow has been proven useful in capturing temporal changes in gesture or action videos, it is ineffective in identifying subtle facial motions due to noise [23] and can be expensive to compute.

Network architecture: In this work, we propose the deep neural network shown in Fig. 1 for binary classification of a video to *stroke* vs. *non-stroke*. For a video consisting of N frames, we take the kth and k+1th frames and calculate the difference between their embedding features right after the first 3×3 convolutional layer. Next, a ResNet-34 [9] model outputs the class probability vector p_k for each frame based on the calculated feature differences. An overall temporal loss $\mathcal{L}$ is calculated by combining a negative log-likelihood (NLL) loss term and smoothness regularization loss terms based on the class probability vectors of three consecutive frames. From the frame-level class probabilities, a video-level class probability vector is obtained by averaging over all frames' probabilities, and the predicted video label will be the class with higher probability.

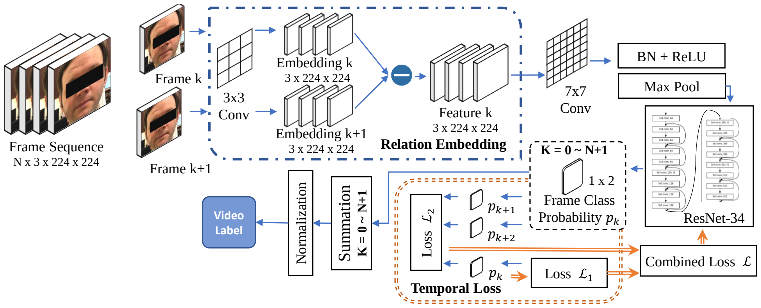

Fig. 1. Flow diagram of the video classification module.

Relation embedding: One novelty of our proposed video module is in the classification using feature differences between consecutive frames instead of using directly the frame features. The rationale behind this choice is in that we expect the network to learn and classify based on motion features. Features from single frames contain a lot of information about the appearance of face which are useful for face identification but not useful in characterizing facial motion abnormality.

Temporal loss: Denote i as the frame index, y_i as the class label for frame i obtained based on the ground truth video label, and p_i as the predicted class probability. The combined loss $\mathcal{L}$ for frame i is defined with three terms: $\mathcal{L}^{(i)} = \mathcal{L}_1^{(i)} + \alpha(\mathcal{L}_{2_{(1)}}^{(i)} + \beta\mathcal{L}_{2_{(2)}}^{(i)})$, where α and β are tunable weighting coefficients. The first loss term, $\mathcal{L}_1^{(i)} = -(y_i \log p_i + (1-y_i)\log(1-p_i))$, is the NLL loss. Note that we sample a subset of the frames to mitigate the overfitting on shape identity. By assuming consecutive frames have continuous and similar class probabilities, we develop a $\mathcal{L}_{2_{(1)}}^{(i)}$ loss defined on the frame class probabilities for three adjacent frames where λ is a small threshold to restrict the inconsistency by penalizing those frames with large class probability differences. We also design another loss, $\mathcal{L}_{2_{(2)}}^{(i)}$, to encourage random walk around the convergence point to mitigate overfitting. Specifically,

$$\mathcal{L}_{2_{(1)}}^{(i)} = \sum_{j\in[0,1,2]} \max\{(|p_{i+j} - p_{i+1+j}| - \lambda), 0\} \ , \tag{1}$$

$$\mathcal{L}_{2_{(2)}}^{(i)} = \sum_{j\in[0,1,2]} \mathbb{1}_{p_{i+j}==p_{i+1+j}} \ . \tag{2}$$

In practice, we adopt a batch training method, and all frames in a batch are weighted equally for loss calculation.

3.3 Transcript Classification for Speech Ability Assessment

We formulate the speech ability assessment as a text classification problem. Subjects without speech disorder complete the speech task with organized sentences and maintain a good vocabulary set size, whereas patients with speech impairments either put up a few words illogically or provide mostly unrecognizable speech. Hence, we concurrently formulate a binary classification on the speech given by the subjects to determine if stroke exists.

Preprocessing: For each speech case $\mathcal{T} := \{t_i, ..., t_N\}$ extracted from the obtained transcripts where t_i is a single word and N is the number of words in the case, we first define the encoding of the words $\mathcal{E}$ over the training set by their order of appearance, $\mathcal{E}(t_i) := d_i$, $d_i \in \mathbb{I}$; $\mathcal{E}(\mathcal{T}) := \mathcal{D}$ and $\mathcal{D} = \{d_i, ..., d_N\} \in \mathbb{I}^n$. We denote the vocabulary size obtained as v. Due to the length difference between cases, we pad the sequences to the max length m of the dataset, so that $\mathcal{D}' = \{d_i, ..., d_N, p_1, ..., p_{m-n}\} \in \mathbb{I}^m$ where p_i denotes a constant padding value. We further embed the padded feature vectors to an embedding dimension E so that the final feature vector has $\mathcal{X} := \{x_i, .., x_m\}$ and $x_i \in \mathbb{R}^{E\times v}$.

Text classification with Long Short-Term Memory (LSTM): We construct a basic bidirectional LSTM model to classify the texts. For the input $\mathcal{X} = \{x_i, .., x_m\}$, the LSTM model generates a series of hidden states $\mathcal{H} := \{h_1, .., h_m\}$ where $h_i \in \mathbb{R}^t$. We take the output from the last hidden state h_t, apply a fully-connected (FC) layer before output (class probabilities/logits) $\hat{y}_i \in \mathbb{R}^2$. For our task, we leave out the last FC layer for model fusion.

3.4 Two-stream Network Fusion

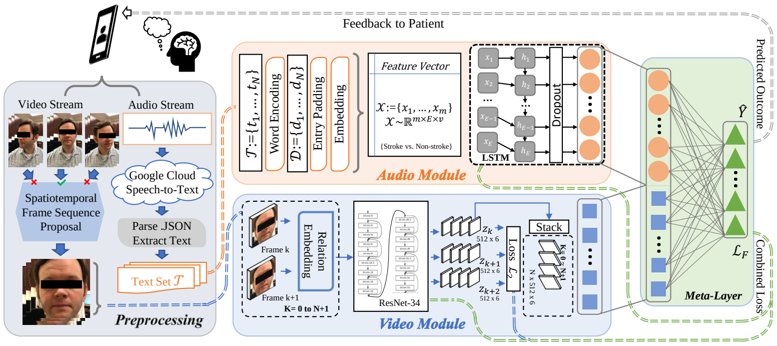

Fig. 2. Flow diagram of the two-stream network fusion process. z_k is the feature output from the ResNet-34 before the final FC layer.

Overall structure of the model: Fig. 2 shows the data pipeline of the proposed fusion model. Videos that are collected following the protocol are uploaded to a database, while audios are extracted and forwarded to Google Cloud Speech-to-Text which generates the transcript. Meanwhile, videos are sent to the spatiotemporal proposal module to perform face detection, tracking, cropping, and stabilization. The preprocessed data are further handled per-case-wise and loaded into the audio and video modules. Finally, a "meta-layer" combines the outputs of the two modules and gives the final prediction on each case.

Fusion scheme: We take a simple fusion scheme of the two models. For both the video and text/audio modules, we remove the last fully-connected layer before the output and concatenate the feature vectors. We construct a fully-connected "meta-layer" for the output of class probabilities. For all the N frames in a video, the frame-level class probabilities from the model are concatenated into $\hat{Y} = \{\hat{p_1}, ..., \hat{p_N}\}$. The fusion loss $\mathcal{L}_F$ is defined in a similar way as the temporal loss; instead of using only video-predicted probabilities, the fusion loss combines both video- and text-predicted probabilities. Note again that the fusion model operates at the frame level, and a case-level prediction is obtained by summing and normalizing class probabilities over all frames.

4 Experiments and Results

Implementation and training: The whole framework[5] is running on Python 3.7 with Pytorch 1.4, OpenCV 3.4, CUDA 9.0, and Dlib 19. The model starts with a pretrained model on ImageNet. The entire pipeline runs on a computer with a quad-core CPU and one GTX 1070 GPU. To accommodate for the existing class imbalance inside the dataset and ER setting, a higher class weight (1.25) is assigned to the non-stroke class. For evaluation, we report the accuracy, specificity, sensitivity and area under the ROC curve (AUC) from 5-fold cross validation results. The loss curves for one of the folds are presented in Fig. 3. The learning rate is tuned to 0.0001 and we early stop at epoch 8 due to the quick convergence and overfitting issue. It is worth mentioning that the early stop strategy and the balanced weight are applied to the baselines below.

Baselines and comparison: To evaluate our proposed method, we construct a number of baseline models for both video and audio tasks. The ground truth for comparison is the binary diagnosis result for each video/audio. General video classification models for video tagging or action recognition are not suitable for our task since they require all frames throughout a video clip to have the same label. In our task, since stroke patients may have many normal motions, the frame-wise labels may not be equal to the video label all the time. For single frame models such as ResNet-18, we use the same preprocessed frames as input

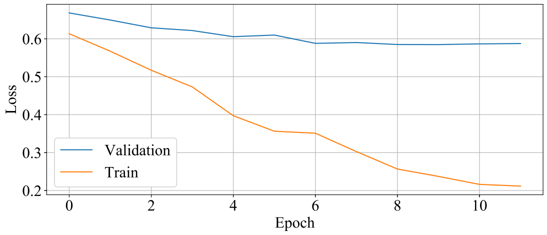

Fig. 3. Loss curves for one of the folds.

[5] Codes are available in https://github.com/0CTA0/MICCAI20_MMDL_PUBLIC .

and derive a binary label for each frame. The label with more frames is then assigned as the video-level label. For multiple frame models, we simply input the same preprocessed video. We also compare with a traditional method based on identifying facial landmarks and analyzing facial asymmetry (“Landmark + Asymmetry”), which detects the mid-line of a patient’s face and checks for bilateral pixel-wise differences on between-frame optical-flow vectors. The binary decision is given by statistical values including the number of peaks and average asymmetry index. We further tested our video module separately with and without using feature differences between consecutive frames. We compare the result of our audio module to that of sound wave classification with pattern recognition on spectrogram [2].

Table 1. Experimental results. Best values are in **bold**.

Metrics Method	Accuracy	Specificity	Sensitivity	AUC
Landmark+Asymmetry	62.35%	66.67%	60.34%	0.6350
Video Module w/o difference	70.28%	11.11%	**98.24%**	0.5467
Video Module w/ difference	76.67%	62.21%	96.42%	0.7281
ResNet-18	58.20%	42.75%	70.22%	0.5648
VGG16	52.30%	27.98%	71.22%	0.4960
Audio Module Only	70.24%	40.74%	84.21%	0.6248
Soundwave/Spectrogram	68.67%	59.26%	77.58%	0.6279
Proposed Method	**79.27%**	66.07%	93.12%	**0.7831**
Clinical Impression	72.94%	**77.78%**	70.68%	0.7423

As shown in Table 1, the proposed method outperforms all the strong baselines by achieving a 93.12% sensitivity and a 79.27% accuracy. The improvements of our proposed method from the baselines on accuracy are ranging from 10% to 30%. It is noticeable that proven image classification baselines (ResNet, VGG) are not ideal for our “in-the-wild” data. Comparing to the clinical impressions given by ER doctors, the proposed method achieves even higher accuracy and greatly improves the sensitivity, indicating that more stroke cases are correctly identified by our proposed approach. ER doctors tend to rely more on the speech abilities of the patients and may overlook subtle facial motion weaknesses. Our objective is to identify real stroke and fake stroke cases among incoming patients, who are already identified with high risk of stroke. If the patterns are subtle or challenging to observe by humans, ER doctors may have difficulty on those cases. We infer that the video module in our framework can detect those subtle facial motions that doctors can neglect and complement the diagnosis based on speech/audio. On the other hand, the ER doctors have access to emergency imaging reports and other information in the Electronic Health Records (EHR). With more information incorporated, we believe the performance of the framework can get further improved. It is also important to note that by using feature difference between consecutive frames for classification, the performance of the video module is greatly improved, validating our hypothesis about modeling based on motion.

Through the experiments, all the methods are experiencing low specificity (*i.e.*, identifying non-stroke cases), which is reasonable because our subjects are patients with suspicion of stroke rather than the general public. False negatives would be dangerous and could lead to hazardous outcome. We also took a closer look at the false negative and false positive cases. The false negatives are due to the labeling of cases using final diagnosis given based on diffusion-weighted MRI (DWI). DWI can detect very small lesions that may not cause noticeable abnormalities in facial motion or speech ability. Such cases coincide with the failures in clinical impression. The false positives typically result from background noise in audio, varying shapes of beard, or changing illumination conditions. A future direction is to improve specificity with more robust methods on both audio and video processing.

We also evaluate the efficiency of our approach. The recording runs for a minute, the extraction and upload of audio takes half a minute, the transcribing takes an extra minute, and the video processing is completed in two minutes. The prediction with the deep models can be achieved within seconds with GTX 1070. Therefore, the entire process takes no more than four minutes per case. If the approach is deployed onto a smartphone, we can rely on Google Speech-to-Text's real-time streaming method and perform the spatiotemporal frame proposal on the phone. Cloud computing can be leveraged to perform the prediction in no more than a minute, after the frames are uploaded. In such a case, the total time for one assessment should not exceed three minutes. This is ideal for performing stroke assessment in an emergency setting and the patients can make self-assessments even before the ambulance arrives.

5 Conclusions

We proposed a multi-modal deep learning framework for on-site clinical detection of stroke in an ER setting. Our framework is able to identify stroke based on abnormality in the patient's speech ability and facial muscular movements. We construct a deep neural network for classifying patient facial video, and fuse the network with a text classification model for speech ability assessment. Experimental studies demonstrate that the performance of the proposed approach is comparable to clinical impressions given by ER doctors, with a 93.12% sensitivity and a 79.27% accuracy. The approach is also efficient, taking less than four minutes for assessing one patient case. We expect that our proposed approach will be clinically relevant and can be deployed effectively on smartphones for fast and accurate assessment of stroke by ER doctors, at-risk patients, or caregivers.

References

1. Claes, P., Walters, M., Vandermeulen, D., Clement, J.G.: Spatially-dense 3D facial asymmetry assessment in both typical and disordered growth. Journal of Anatomy

219(4), 444–455 (2011)

2. Dennis, J., Tran, H.D., Li, H.: Spectrogram image feature for sound event classification in mismatched conditions. IEEE Signal Processing Letters **18**(2), 130–133 (2010)
3. Dong, J., Lin, Y., an Liu, L., Ma, L., Wang, S.: An approach to evaluation of degree of facial paralysis based on image processing and pattern recognition. Journal of Information & Computational Science **5**(2), 639–646 (2008)
4. Frey, M., Michaelidou, M., Tzou, C.H., Pona, I., Mittlböck, M., Gerber, H., Stüssi, E.: Three-dimensional video analysis of the paralyzed face reanimated by cross-face nerve grafting and free gracilis muscle transplantation: Quantification of the functional outcome. Plastic and Reconstructive Surgery **122**(6), 1709–1722 (2008)
5. Giles, E., Patterson, K., Hodges, J.R.: Performance on the Boston Cookie Theft picture description task in patients with early dementia of the Alzheimer's type: missing information. Aphasiology **10**(4), 395–408 (1996)
6. Guo, Z., Dan, G., Xiang, J., Wang, J., Yang, W., Ding, H., Deussen, O., Zhou, Y.: An unobtrusive computerized assessment framework for unilateral peripheral facial paralysis. IEEE Journal of Biomedical and Health Informatics **22**(3), 835–841 (2017)
7. Guo, Z., Shen, M., Duan, L., Zhou, Y., Xiang, J., Ding, H., Chen, S., Deussen, O., Dan, G.: Deep assessment process: Objective assessment process for unilateral peripheral facial paralysis via deep convolutional neural network. In: Proceedings of the IEEE International Symposium on Biomedical Imaging. pp. 135–138 (2017)
8. Harbison, J., Hossain, O., Jenkinson, D., Davis, J., Louw, S.J., Ford, G.A.: Diagnostic accuracy of stroke referrals from primary care, emergency room physicians, and ambulance staff using the Face Arm Speech Test. Stroke **34**(1), 71–76 (2003)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
10. He, S., Soraghan, J.J., O'Reilly, B.F.: Automatic motion feature extraction with application to quantitative assessment of facial paralysis. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. vol. 1, pp. 441–444 (2007)
11. Horta, R., Aguiar, P., Monteiro, D., Silva, A., Amarante, J.M.: A facegram for spatial–temporal analysis of facial excursion: Applicability in the microsurgical reanimation of long-standing paralysis and pretransplantation. Journal of Cranio-Maxillofacial Surgery **42**(7), 1250–1259 (2014)
12. Hsu, G.S.J., Chang, M.H.: Deep hybrid network for automatic quantitative analysis of facial paralysis. In: Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance. pp. 1–7 (2018)
13. Hu, Y., Chen, L., Zhou, Y., Zhang, H.: Estimating face pose by facial asymmetry and geometry. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (F&G). pp. 651–656 (2004)
14. Johnson, W., Onuma, O., Owolabi, M., Sachdev, S.: Stroke: A global response is needed. Bulletin of the World Health Organization **94**(9), 634 (2016)
15. Khairunnisaa, A., Basah, S.N., Yazid, H., Basri, H.H., Yaacob, S., Chin, L.C.: Facial-paralysis diagnostic system based on 3D reconstruction. In: AIP Conference Proceedings. vol. 1660, p. 070026. AIP Publishing (2015)
16. Kothari, R.U., Pancioli, A., Liu, T., Brott, T., Broderick, J.: Cincinnati Prehospital Stroke Scale: Reproducibility and validity. Annals of Emergency Medicine **33**(4), 373–378 (1999)

17. Leira, E.C., Kaskie, B., Froehler, M.T., Adams Jr, H.P.: The growing shortage of vascular neurologists in the era of health reform: Planning is brain! Stroke **44**(3), 822–827 (2013)
18. Li, P., Tan, S., Zhou, X., Yan, S., Zhao, Q., Zhang, J., Lv, Z.: A two-stage method for assessing facial paralysis severity by fusing multiple classifiers. In: Proceedings of the Chinese Conference on Biometric Recognition. pp. 231–239. Springer (2019)
19. Song, A., Xu, G., Ding, X., Song, J., Xu, G., Zhang, W.: Assessment for facial nerve paralysis based on facial asymmetry. Australasian Physical & Engineering Sciences in Medicine **40**(4), 851–860 (2017)
20. Soraghan, J.J., O'Reilly, B.F., McGrenary, S., He, S.: Automatic facial analysis for objective assessment of facial paralysis. In: Proceedings of the 1st International Conference on Computer Science from Algorithms to Applications. Cairo, Egypt (2009)
21. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4489–4497 (2015)
22. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
23. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: Proceedings of European Conference on Computer Vision. pp. 20–36 (2016)
24. Wang, S., Li, H., Qi, F., Zhao, Y.: Objective facial paralysis grading based on p_{face} and EigenFlow. Medical & Biological Engineering & Computing **42**(5), 598–603 (2004)