I Can Hear You

Arabic and English Lip-Reading

Lip-reading

Eman Ibrahim Yusuf Sam
Computer science. Ain shams university
Ain shams university
Qalupia, Egyept
eman.yusuf305@gmail.com

Mahmoud Salama Mohamed
Computer science. Ain shams university
Ain shams university
Qalupia, Egyept
mahmoud.salama.gado@gmail.com

Mahmoud Saeed Sayed
Computer science. Ain shams university
Ain shams university
Qalupia, Egyept
mahmoud.said5456@gmail.comline

Aalaa Ahmad Mohamed
Computer science. Ain shams university
Ain shams university
Cairo, Egyept
aalaaahmad218@gmail.com

Noura Ahmad Hafez
Computer science. Ain shams university
Ain shams university
Cairo, Egyept
nourahafez063@gmail.com

Dr. Mohamed Aabrouk
mohamed.mabrouk@cis.asu.edu.eg

TA. Hazem Yousef
hazem_yousef@cis.asu.edu.eg

***Abstract*— Due to the limited number of people who know sign language or can interpret lip movements, deaf individuals often face difficulties communicating with others. To address this issue, this project aims to develop an application that tracks the lip movements of a person speaking in a video and predicts the corresponding text to make communication easier for the deaf.**

The project consists of two different models:

- **The first model is for Arabic language input data, which takes 3D input data, applies convolutional and pooling operations to extract spatial features, processes temporal information using bidirectional GRU layers, and produces a probability distribution over the classes using a dense layer followed by SoftMax activation. The accuracy of this model is 80%.**
- **The second model is for English language input data, which takes 3D input data, applies convolutional and pooling operations to extract spatial features, processes temporal information using bidirectional LSTM layers, and produces a probability distribution over the classes using a dense layer followed by SoftMax activation. The accuracy of this model is 97%.**

***Keywords*— Speechreading, Lipreading technology, Lip movement analysis, landmarks, Deep learning, Neural networks loss function, Convolutional neural networks (CNNs)Recurrent neural networks (RNNs), Long short-term memory (LSTM), Gated recurrent unit (GRU), Artificial intelligence (AI),Machine learning, Deaf and hard-of-hearing**

I. INTRODUCTION

People with hearing problems as deaf people face difficulties in communicating with other people, some of whom may not know sign language. Lip reading is one communication skill that can help them communicate better and understand what is being said. movements and translating them into text, and Assisting individuals with hearing difficulties in understanding speech is crucial. Since human lipreading performance is often limited, unlike machine lipreaders that offer significant practical potential, converting videos of people speaking into text can greatly enhance communication for deaf individuals. This technology enables them to easily interact and engage with society, we can do this by Lip-reading.

Lip-reading, also known as speechreading, is a technique of understanding speech by visually interpreting the movements of the lips, face, and tongue when normal sound is not available, also it plays a vital role in human communication and speechunderstanding. So, visual speech recognition “lipreading” is a field of growing attention. It is a natural complement to audio-based speech recognition that can facilitate dictation in noisy environments and enable silent dictation in offices and public spaces. It is also useful in applications related to improved hearing aids and biometric authentication Assisting individuals with hearing difficulties in understanding speech is crucial. Since human lipreading performance is often limited, unlike machine lipreaders that offer significant practical potential, converting videos of people speaking into text can greatly enhance communication for deaf individuals. This technology enables them to easily interact and engage with society.

Lip-reading, on the other hand, is commonly used by deaf individuals who do not know sign language, particularly those who were born with hearing abilities but gradually or suddenly lost their hearing during their lifetime. The World Congress of Educators in Milan made a significant decision in 1880, stating that all deaf children should be taught lip reading. Lip reading instruction focuses on observing the movements of the lips, tongue, and jaw, as well as understanding the stress, rhythm, and expression of spoken language. Students learn the lip readers' alphabet and identify groups of sounds that share similar lip shapes (visemes), such as “ د – ط – ت”,” س – ص – ز ”, “ي – ش – ج” "p," "b," "m," or "f," "v." The goal is to grasp the overall message, enabling individuals to confidently engage in conversations and prevent the social isolation that often accompanies hearing loss

Our project involves several tasks aimed at developing and applying deep learning models for lip-reading. First, we will conduct a survey of different machine learning architectures from previous research papers that have worked on our problem. This will help us determine the strengths and weaknesses of different architectures for our specific application. We will also build an Arabic language model and train it on our Arabic dataset, as well as build an English language model and train it on the GRID dataset. Finally, we will develop a web application that enables users to upload or capture a video, then get the text.

II. Related Works

The paper [1] focuses on developing a model capable of classifying digits and phrases in the Arabic language using visual datasets. The authors employ keyframe extraction techniques and concatenated stretch images (CFIs) to represent the sequence of input videos. Various approaches were explored, and the model's performance was evaluated on a collected dataset. The results indicate that the CNN network with VGG19 architecture, coupled with batch normalization for stabilizing the training process, achieves high accuracy compared to state-of-the-art methods that use concatenated frame images as input. The model achieved a test accuracy of 94% for digit recognition, 97% for phrase recognition, and 93% for both digit and phrase recognition experiments. Moving forward, the authors express their intention to further investigate the model by exploring different lip landmark localization techniques and expanding the dataset to improve the accuracy of visual speech recognition. However, it is worth noting that working on lip-reading solely as a classification problem with a limited dataset of four sentences and ten numbers may not be the recommended approach.

Wen et al. [2] have stated that a model with few parameters will decrease performance complexity. They have applied LSTM for extracting the sequence information between keyframes. The collected dataset represents the pronunciation of 10 English digit words (from 0 to 9). The accuracy of the model is 86.5%, which indicates that the model saves computing resources and memory space.

Gergen et al. [3] use speaker-dependent training on an LDA-transformed version. of the Discrete Cosine Transforms of the mouth regions in an HMM/GMM system. This work holds the previous state-of-the-art on the GRID corpus with a speaker-dependent accuracy of 86.4%. Generalization across speakers and extraction of motion features is considered an open problem, as noted in (Zhou et al., 2014). Lip-Net addresses both issues.

Garg et al. [4] apply a VGG pre-trained on faces to classifying words and phrases from the MIRACL-VC1 dataset, which has only 10 words and 10 phrases However, their best recurrent model is trained by freezing the VGG-Net parameters and then training the RNN, rather than training them jointly. Their best model achieves only 56.0%-word classification accuracy, and 44.5% phrase classification accuracy, despite both of these being 10-class classification tasks.

Eldirawy and Ashour [5] built a system to lip-read the numbers from one to ten when spoken in the Arabic language. Three recognition methods were used: K-mean, fuzzy k-mean, and k nearest neighbors (K-NN) classifiers with maximum recognition accuracy of 55.8%. Morade and Patnaik proposed a lipreading algorithm based on localized active counter model (ACM) and a hidden Markov model (HMM). The algorithm was tested by recording videos of English digits from zero to nine from 16 speakers (8 male and 8 female) and the Cuave database. The recognition accuracy varied between 77.8% and 79.6%.

Assael et al. [6] proposed Lip-Net; An end-to-end sentence-level model is first proposed by which combines spatiotemporal convolutions, LSTMs and the connectionist temporal classification (CTC) loss to compute the labeling. This approach achieves 95.2% accuracy on GRID dataset. A deeper learning architecture is raised by, which puts forward a network including spatiotemporal convolutional, residual and Bi-LSTM networks. The goal of first two sub-networks is to extract more powerful visual features. This network attains a word accuracy of 83% on LRW. An extended version of this architecture is applied for audiovisual speech recognition.

Encoder–decoder architecture and CTC approaches are initially relying on recurrent networks. For example, [7] proposes a LCA-Net using a stacked 3D convolutional neural network (CNN), highway network, bidirectional GRU network to encode input images and a cascaded attention-CTC decoder to predict the character probabilities. This approach achieves 97.1%-word accuracy on GRID dataset. Recently, some research find that simple CNNs may perform better for sequence modeling [8]. Fully convolutional networks with CTC have been proposed for automatic speech recognition [9,10]. And also, for machine translation, [11] replaces the encoder and [12] replaces the whole pipeline with a fully convolutional network. At the same time, self-attention mechanism [13] is also found to replace recurrent networks for sequential tasks. In [14], two transformer architectures are raised, which consisted of a stack of multi-head self-attention layers with CTC and seq2seq model, respectively, and the transformer-seq2seq model combining external language model achieves the best accuracy of 50% on the LRS-BBC dataset, only using visual information.

For recognizing full words, Petridis et al. [15] trains an LSTM classifier on a discrete cosine transform (DCT) and deep bottleneck features (DBF). Similarly, Wand et al. [16] uses an LSTM with HOG input features to recognize short phrases. The shortage of training data in lip reading presumably contributes to the continued use of shallow features. Existing datasets consist of videos with only a small number of subjects, and also a very limited vocabulary (<60 words), which is also an obstacle to progress. The recent paper of Chung and Zisserman [17] tackles the small-lexicon problem by using faces in television broadcasts to assemble a dataset for 500 words. However, as with any word-level classification task, the setting is still distant from the real world, given that the word boundaries must be known beforehand. A very recent work [18] uses a CNN and LSTM based network and Connectionist Temporal Classification (CTC) [19] to compute the labelling. This reports strong speaker-independent performance on the constrained grammar and 51-word vocabulary of the GRID dataset [20]. However, the method, suitably modified, should be applicable to longer, more general sentences.

Faisal et al. [21] intended to combine both models to enhance speech recognition in load environments; however, they were unable to merge both networks to confirm the results of audio-visual recognition. They have applied two different deep-learning models for lip-reading spatiotemporal convolution neural network and Connectionist Temporal Classification Loss (CTC), and for audio, they have used the MFCC features in a layer of LSTM cells and output the sequence for the Urdu language.

Saitoh et al. [22] have built a model that depends on the concept of CFI for the required pre-processing of video frames with two types of dataset augmentation, with CNN applied for features extraction. For classification, the

SoftMax layer with cross-entropy loss has been utilized. They have employed the OuluVS2 dataset (frontal view) to evaluate the method using different pre- trained models. The accuracy is 61.7% with the Nin model and 89.4% with the Google Net model.

III. SYSTEM ARCHITECTURE

A. Datasets

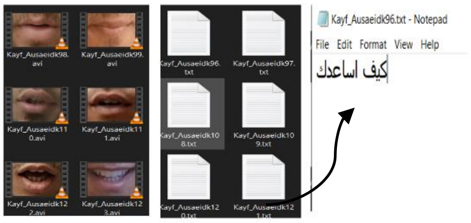
In our project we used two different datasets, the first one was our Arabic dataset, and the second one was GRID.

- *Our Arabic Dataset*

Initially, our search for comprehensive and extensive datasets yielded limited results, with only a single dataset available. Unfortunately, this dataset was relatively small and only contained videos of 14 individuals uttering numbers from 0 to 9, along with four short sentences: "السلام عليكم," "اتصل بالشرطة," "اتصل بالاسعاف," and "احتاج للمساعدة." [1] However, due to the limited size of this dataset, it was deemed insufficient to form the foundation of our project.

Given the insufficient nature of the available dataset, we made the decision to generate our own data following a similar style as the small dataset. In doing so, we carefully selected 24 significant words commonly utilized by Arab citizens in various everyday situations. These words were chosen based on their relevance and frequency.

كيف حالك	وجدت حقيبة	صداع مزمن	قائمة الطعام
صباح الخير	سرقة حقيبة	ضغط عالي	تفضل الطلب
مساء الخير	رخصة قيادة	أريد قهوة	أين المحاضرة
أين المترو	مما تعاني	اريد الحساب	موعد المحاضرة
كيف اساعدك	سرطان الحنجرة	اريد الفاتورة	مبنى الامتحان
تقديم بلاغ	ضعف سمعي	اريد المدير	موعد النتيجة



كيف اساعدك

So, we have dataset which contains from ten digits (0,1,2….,9) and 28 sentences, in total 4400 videos and aligns files.

- *English Dataset*

The second dataset is GRID, which contains 1000 sentences spoken by 34 talkers (18 males and 16 females). Each sentence consists of a six-word sequence of the form: command, color, preposition, letter, digit, and adverb, such as "Place green at H 7 now." GRID contains a total of 51 different words and 165,000-word instances. Videos in this dataset have a fixed duration of 3 seconds at a frame rate of 25 FPS, with a resolution of 720×576, resulting in sequences comprising 75 frames. Each video has a corresponding text file that includes the spoken word and its duration.

B. PreProcessing

Preprocessing Arabic lipreading presented greater challenges than English lipreading due to the individuals speaking often being in motion and the background being blurry. Conversely, English lipreading was relatively easier because the background was fixed and the individuals speaking were static. These differences in video characteristics required tailored preprocessing approaches for each language to account for their specific challenges.

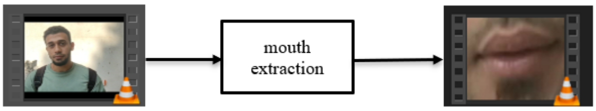
- *Labeling*

Each video in the Arabic dataset was accompanied by an align file, like the English data. This file contained the corresponding text spoken in the video. We generated align files for each Arabic video, containing the transcribed Arabic text spoken in the respective video. This alignment process enabled us to connect the visual cues of the lip movements in the videos with the corresponding textual representation of the spoken words. Aligning the lip movements with the accompanying Arabic text facilitated the training and evaluation of lipreading models for accurate interpretation of Arabic speech.

- *Mouth extraction*

a) Arabic Mouth extraction

We achieved improved accuracy in isolating the lips from the speaker's face by employing the Landmarks algorithm. This enhanced lip extraction process laid the foundation for subsequent analysis and interpretation of lip movements in our lipreading system.



b) English Mouth Extraction

To extract the lip area from the English videos, we divided each frame into smaller segments or slices, focusing solely on the region encompassing the lips. As the speaker's face remained relatively stationary, this slicing technique allowed us to quickly extract the lip area without the need for complex algorithms or extensive computational resources.

C. Model Architecture

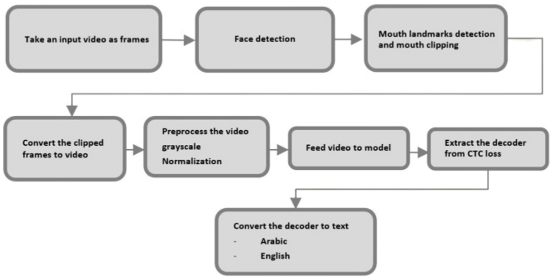
- *Arabic model*

The model described here uses a CTC loss function and expects input data in the shape of (60, 150, 160, 1), representing a 5D tensor with dimensions (batch-size, time-steps, height, width, channels) and a data type of float32. The convolutional layers utilize ZeroPadding3D to preserve spatial dimensions by adding padding of (1, 2, 2), followed by a Conv3D operation with 32 filters, each with a kernel size of (3, 5, 5) and a stride of (1, 2, 2) to reduce the spatial dimensions. Batch-Normalization normalizes activations, ReLU applies activation functions, SpatialDropout3D performs dropout regularization, and MaxPooling3D down-samples feature maps. Recurrent layers use Time-Distributed to apply the layers to each time step of the input tensor independently, followed by Flatten to convert 3D feature maps to 2D, and two layers of bidirectional GRU in both forward and backward directions with 256 units each. The output of the second bidirectional GRU layer is a sequence of hidden states for each time step. The output layers consist of a fully connected Dense layer with 37 units (Arabic chars and symbols) and a SoftMax activation function to obtain a probability distribution over the classes. A Keras Model object is constructed with the input and output tensors.

- *English model*

The model initializes using the Sequential API and uses a CTC loss function. The convolutional layers consist of a Conv3D layer with 128 filters, a kernel size of 3, and 'same' padding. They use ReLU activation functions and MaxPool3D with a pool size of (1, 2, 2) to down-sample the feature maps. Two additional Conv3D layers follow the same pattern, with 256 and 75 filters respectively. The temporal processing layer uses Time-Distributed to apply the layers to each time step independently, followed by Flatten to convert 3D feature maps into a 1D representation. The recurrent layers consist of two layers of bidirectional LSTM with 128 units each, using dropout and recurrent dropout of 0.2 to prevent overfitting. The output layer is a fully connected Dense layer with 41 units, assuming 41 (English chars and symbols) different classes, and a SoftMax activation function to obtain a probability distribution over the classes.

- *System rchitecture summary*


Take an input video as frames
Face detection
Mouth landmarks detection and mouth clipping
Convert the clipped frames to video
Preprocess the video
grayscale
Normalization
Feed video to model
Extract the decoder from CTC loss
Convert the decoder to text
- Arabic
- English

IV. Results

Our results for the English language were 97%, displayed within 6 seconds, which distinguishes our work from others. The Arabic language results were 80%, which was a significant improvement from previous attempts. The average time to display the Arabic text was 30 seconds, which is considered acceptable.

A. English results

Dataset	Epochs	Accuracy
Grid	30	30%
Grid	50	53%
Grid	70	69.4%
Grid	100	86.2%
Grid	130	91.3%
Grid	150	97.15%

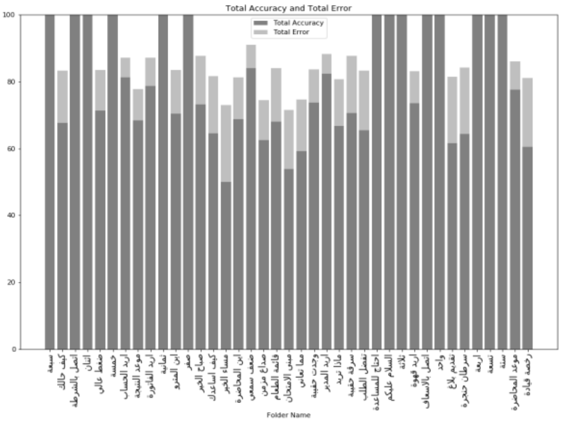
B. Arabic results

Dataset	Checkpoint	Accuracy
Our private data	5	**40%**
Our private data	15	**43%**
Our private data	25	**56%**
Our private data	35	**70.2%**
Our private data	45	**76.3%**
Our private data	65	**79.33%**
Our private data	75	**79.27%**
Our private data	80	**79.83%**
Our private data	83	**80.06%**

Note: each checkpoint is 7 epochs

C. Accuracy

The results for the Arabic dataset, which included 10 digits ranging from 0 to 9 and 28 sentences, showed that the highest accuracy achieved was 100%, while the lowest was 59.21%. The details of the accuracy for each number and sentence can be found in the following table:



D. Prediction time

To obtain the average time to display the text, I analyzed 60 random videos from 6 randomly selected folders, choosing 10 videos at random from each folder. The average time for each folder is shown in the following table:

Number	Sentence	Avg Time
1	أين المترو	29.922 s
2	كيف حالك	28.774 s
3	ماذا تريد	28.186 s
4	صباح الخير	26.915 s
5	سرطان الحنجرة_	30.463 s
6	تفضل الطلب	29.231 s

V. future work

In the future, we plan to focus on two areas to improve our Arabic language processing capabilities. Firstly, we aim to develop our video equipment by purchasing dedicated cameras instead of using mobile phones. This will enable us to capture more data of higher quality, which in turn will improve the accuracy of our models. Secondly, we plan to increase the training of our models on videos to further enhance their accuracy. We also intend to develop our application to operate in real-life scenarios. By working on both these aspects, we aim to achieve even better results in Arabic language processing.

VI. Conclusion

Our work achieved notable results for both English and Arabic languages. Specifically, we achieved an accuracy of 97% for English, with a display time of 6 seconds. This sets our work apart from other similar projects.

For Arabic, we achieved an accuracy of 80%, which represents a significant improvement over previous attempt. The average time to display the Arabic text was 30 seconds, which is considered acceptable.

Overall, our work demonstrated promising results for both languages, with competitive accuracy rates and reasonable display times.

VII. Acknowledgment

At the beginning we would like to express our gratitude & thanks to Allah and to all who helped us completing this project. Especially our supervisors.

Dr. Mohamed Mabrouk for his understanding, patience and encouraging us all the time to do our best, he has always been there for us.

TA. Yousef Hazem for his supervision and great effort.

We are proud that we have been surrounded by this amount of care and beautiful spirit that pushed us to do our best till the last time.

Finally, we would like to thank our family for supporting us and friends for sharing their experience and knowledge

References

[1] Alsulami, N.H., Jamal, A.T., & Elrefaei, L.A. (2021). Deep Learning-Based Approach for Arabic Visual Speech Recognition. In Proceedings

[2] J. Wen and Y. Lu, "Automatic lip reading system based on a fusion lightweight neural network with Raspberry Pi," Applied Sciences, vol. 24, no. 9, pp. 5432, 2019

[3] Gergen, S. Zeiler, A. H. Abdelaziz, R. Nickel, and D. Kolossa. Dynamic stream weighting for turbodecoding-based audiovisual ASR. In Interspace, pp. 2135–2139, 2016.

[4] Garg, J. Noyola, and S. Bagadia. Lip reading using CNN and LSTM. Technical report, Stanford University, CS231n project report, 2016.

[5] Eldirawy I, Ashour W. Visual speech recognition: The digits from one to ten in the Arabic language. LAP LAMBERT Academic Publishing; 2011.

[6] Assael YM, Shillingford B, Whiteson S, de Freitas N. Lipnet: End-to-end sentence-level lipreading; 2016. arXiv:1611.01599

[7] . Xu, K., Li, D., Cassimatis, N., Wang, X.: LCANet: End-to-end lipreading with cascaded attention-CTC. In: 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018), pp. 548–555 (2018)

[8] Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)

[9] Wang, Y., Deng, X., Pu, S., Huang, Z.: Residual convolutional CTC networks for automatic speech recognition. arXiv preprint arXiv:1702.07793 (2017)

[10] Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Laurent, C., Bengio, Y., Courville, A.C.: Towards end-to-end speech recognition with deep convolutional neural networks. In: Conference of the International Speech Communication Association, pp. 410–414 (2016)

[11] Gehring, J., Auli, M., Grangier, D., Dauphin, Y.: A Convolutional encoder model for neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 123–135 (2017)

[12] Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: Proceedings of the 34th International Conference on Machine Learning-Vol. 70, pp. 1243–1252 (2017)

[13] . Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–

[14] Karita, S., Chen, J., Hori, T., Lee, T. Y., & Chang, W. Y. (2019). LRS3-TED: A large-scale multimodal dataset for language-related video understanding. In Proceedings of the International Conference on Computer Vision (ICCV) Workshops

[15] Petridis, S., Stavropoulos, T. G., Zhang, Y., & Cipolla, R. (2018). End-to-end multimodal speech recognition using 3D convolutional neural networks and attention-based CTC. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

[16] Wand, M., Kment, M., & Matas, J. (2016). Lipreading with LSTM. In Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP).

[17] Chung, J. S., & Zisserman, A. (2016). Out of time: Automated lip sync in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[18] Stafylakis, T., & Tzimiropoulos, G. (2017). Combining residual networks with LSTMs for lipreading. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[19] Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the International Conference on Machine Learning (ICML).

[20] Cooke, M., Barker, J., & Cunningham, S. (2006). Shaping the agenda: A new perspective on speech research in the UK. Speech Communication, 48(9), 1041-1055.

[21] Faisal, K., Riaz, N., & Niyaz, Q. (2018). Audio-visual speech recognition using deep learning for Urdu language. In Proceedings of the International Conference on Intelligent Computing and Optimization (ICO).

[22] Saitoh, T., Umezawa, T., & Fujimoto, T. (2018). Lipreading using convolutional neural networks with color-filtering-intensity (CFI) based pre-processing. In Proceedings of the International Conference on Pattern Recognition (ICPR).