Dynamic Facial Dataset Capture and Processing for Visual Speech Recognition using an RGB-D Sensor

Naveed Ahmed, Mohammed Lataifeh, and Imran Junejo

Abstract— **This work presents a new RGB-D acquisition system to capture a comprehensive dynamic facial dataset that can be used for visual speech recognition. The RGB-D facial dataset acquisition system uses a Kinect to record detailed facial features of a person. The dynamic facial dataset is comprised of the facial data of 20 individuals saying 20 common English words or phrases. The acquisition system employs Kinect facial tracking, which records a large number of dynamic facial features. These features include: facial points, facial outline, RGB data, depth data, mapping between RGB and depth data, facial animation units, facial shape units, and finally 2D and 3D face representations of the face along with the 3D head orientation. The effectiveness of acquired RGB-D dynamic facial dataset is demonstrated by presenting a new visual speech recognition method that employs three-dimensional spatiotemporal data of different facial feature points. A number of visual speech recognition methods from the literature are also tested on the new dataset and they obtain a comparable or favorable visual speech recognition results. The results demonstrate the effectiveness of the proposed RGB-D dynamic facial dataset and show that it can be effectively employed in a visual speech recognition system.**

***Index Terms*—RGB-D, Kinect, Facial Dataset, Visual Speech Recognition, Facial Tracking.**

I. Introduction

SPEECH recognition from dynamic image data is an active area of interest in Computer Vision and Human Computer Interaction. It is particularly useful for the scenarios where lip reading must be employed, and the audio data is not available. It has several applications, ranging from surveillance to interacting with the computing devices in noisy environments. Given the diversity and usefulness of its applications, it is important to explore and improve the existing visual speech recognition algorithms.

In general, a speech recognition system should be able to identify spoken words from image or speech data. Zheng et al. [1] presented a thorough review of lip-reading recognition methods and their review demonstrates that is a very active area of research. Neumeyer et al. [2] presented an automatic scoring method to rank the pronunciation quality. They created a database of various speeches and used statistical models for the scoring. Turk et al. [3] used continuous-mixture hidden Markov models for the speech recognition tasks.

For the visual speech recognition, traditionally, the data is obtained using an RGB (color) video camera. Image processing techniques are used to segment the face, and mouth areas. Afterward, the important feature points from the face or mouth area are used to classify speech segments using different techniques, e.g., histogram of oriented gradients (HOG)[4][5], hidden Markov models [6], spatio-temporal descriptors [7], optimization [8], random forest [9], artificial neural network [10], k-nearest neighbors (KNN)[11][12], optical flow[13], and learning algorithms [14][15][16][17].

Zhao et al. [7] presented a local spatio-temporal descriptor to represent isolated phrases of speech only using the visual data. The spatio-temporal binary local descriptors were then used for speaker-dependent and speaker-independent visual speech recognition. Chen et al. [8] explored an eigenspace-based fast speaker adaptation approach. The approach was used to improve the maximum likelihood linear regression technique for the adaptations. Their method works especially well for speeches of more than 10 seconds in length.

Pei et al. [9] used unsupervised random forest manifold alignment to estimate affinity of patch trajectories in the video data. They employed manifolds alignment technique for matching between query and reference video clips. In contrast, Bagai et al. [10] employed a neural network and train it on the horizontal and vertical distance between the lips for each word. They created a database of the lips data and their recognition accuracy was 52%.

Pao et al. [11], used weighted KNN-based classifiers to recognize 10 digits, from 0 to 9 in Mandarin. They employed traditional KNN, weighted KNN, and weighted D-KNN. They also created a new audio-visual database in English and Mandarin. Their method resulted in an accuracy of around 72% correct recognition. Shaikh et al. [13] computed optical flow of the lips motion and extracted feature vectors to train support vector machines. They also only targeted individual words from the video data for the visual speech recognition.

Noda et al. [14] presented a deep learning-based visual speech recognition system. They apply a convolution neural network for visual feature extraction. The training was done on the dynamic data of the mouth area that was associated with a phoneme label. Similar to our method, they focused

Manuscript received September 23, 2019; revised August 19, 2020.

N. Ahmed and M. Lataifeh are with the Department of Computer Science, University of Sharjah, Sharjah, 27272, UAE. E-mail: nahmed@sharjah.ac.ae and mlataifeh@sharjah.ac.ae. I. Junejo is with College of Technological Innovation, Zayed University, UAE. Email: Imran.Junejo@zu.ac.ae.

on the recognition of isolated words. Assael et al. [15] presented LipNet that could decode the text from the movement of speaker's mouth. They proposed an end-to-end network that used spatio-temporal convolutions, a recurrent network, and employed a connectionist temporal classification loss function.

Recently, Lu et al. [16] used a deep convolution neural network, combined with a recurrent neural network for automatic lip reading. They employed a visual geometry group network to extract lip image features. Petridis et al. [17] presented an end-to-end deep learning based approach for visual speech recognition. They tested their method on a number of datasets, and it resulted in a high classification accuracy for the OuluVS2 dataset. The OuluVS2 dataset is closer to our proposed dataset but it is only in the RGB domain. In their work, they first started by creating a difference image between two consecutive frames. The model consists of two streams of extracted features from the original images and the difference images. Each stream consists of two parts: an encoder and a Bidirectional Long Short-Term Memory (BLSTM) neural network. Finally, a softmax layer is used as the output layer. The entire system is trained end-to-end, so the feature extraction and classification layers are trained jointly.

Visual speech recognition has a number of applications in addition to the basic recognition of simple words. It has been employed to distinguish words from different languages [18], or in automatic translation from speech to sign language [19]. Furthermore, based on a number of different visual features extraction techniques [20], new areas are benefiting from facial datasets. These include: gesture based interactions [21], cultural tourism [22], and emotion recognitions [23][24]. Another promising area of application is speech therapy [25][26], in which geometric features of the spoken language are employed to correctly guide a patient to not only speak properly but also move the mouth muscles in a correct manner.

Additional medical applications have also been reported in identifying facial structure deformation associated with some disorders as in the case of Obstructive Sleep Apnea Syndrome [27]. These approaches are also accelerating recent advances in autism disorder screening utilizing gaze [28] or facial expressions [29]. Nonetheless, the mere availability of similar datasets as described here is of critical importance for benchmarking amongst similar work, and for new adopters to have it contextualized into new domains traditionally linked to language and human behaviors [30] as far as creating new form of lie-detectors [31].

Therefore, a wider application continued to be witnessed with a wave of interest in using RGB-D cameras to capture the facial data [32][33][34][35][36]. Nowadays, the most commonly employed RGB-D camera is Kinect from Microsoft. Kinect is an off-the-shelf, low cost sensor, which not only captures high resolution color information, but also the depth information [37][38]. The depth data can be used generate 3D point clouds and is used in many innovative applications [39][40][41]. In addition, Kinect also provides a body posture and face tracking SDK [42] that can be directly used to localize the face and the important facial feature points [35]. Schatz et al. [43] presented detailed performance of the sensor with the provided SDK.

Kinect face tracking has been employed in a number of applications for visual speech recognition. Rekik et al. [32][33] employed RGB-D facial data processing for mouth region tracking and extraction. Afterward, they employed spatial and temporal descriptors, histogram of oriented gradients (HOG) and motion boundary histograms (MBH). Finally, Support Vector Machine (SVM) was employed for the visual speech classification.

Yargic et al. [36] employed Kinect face tracking SDK to record the facial data and its feature points. They computed the angles between the lip points in the 3D coordinates as features to classify the words. Afterward, KNN classifier was used to classify the words with Manhattan and Euclidean distances to find the best feature vectors. In addition, they did not capture a comprehensive dataset that can be used for training and validation of their method.

In this work, we present a comprehensive RGB-D facial dataset comprising of 20 individuals, with unique 20 English language words or phrases. The dataset is acquired using Kinect facial tracking SDK, and a large number of dynamic facial features are acquired. These features include: facial points, facial outline, RGB data, depth data, mapping between RGB and depth data, facial animation units, facial shape units, and finally 2D and 3D face representations of the face along with the 3D head orientation. To the best of our knowledge, there is no other dynamic RGB-D facial dataset available that provides such a comprehensive RGB-D data, which can be used to train algorithms in a number of areas. The proposed dataset can be employed in a number of applications including visual speech recognition. We demonstrate the effectiveness of our dataset by applying a number of different algorithms for visual speech recognition and compare our results with the earlier methods.

In the first step, we apply a modified KNN algorithm based on the work of Yargic et al. [36] for the visual speech recognition. As explained earlier, they created feature vectors using different distance measures (Manhattan and Euclidean) between the lips feature points to represent a word. Their work only used the angles between the lip points and does not consider the temporal trajectory of the mouth data that we employ to improve the feature vectors. In addition, they only considered four feature points, whereas we consider all the 18 points and their temporal trajectory to improve the feature vectors. The results of the modified method and the comparison with this work will be discussed in Section IV.

Secondly, we apply the modified Pao et al. [11] weighted D-KNN method and compare our results. In addition, we also apply an optical flow-based mechanism similar to Shaikh et al. [13] and compare our results. For the next comparison, we train a neural network, similar to Bagai et al [10] and compared the results with the originally proposed method. For these three methods, the main difference to these works and ours is that we were able to utilize the depth data as an additional measure in the tracking algorithm, whereas these three methods relied on the RGB data. The depth data provides us with a notion of three space position and consequently we can obtain the motion vectors in three dimensions. As will be shown in the next section that Kinect provides a mapping between the depth and RGB data. This mapping allows us to link the tracking information between both data streams and obtain a comprehensive representation of the visual speech data.

We also compare our results to Rekik et al. [32][33] and show that they outperform these methods in the Speaker Independent (SI) setting by more than 2%. We also tested the new end-to-end deep learning approach presented by Petridis et al. [17] on our dataset. As shown in the results that the results quality was high, and the classification accuracy was 87.1%. The comparative results along with a detailed discussion will be presented in Section IV.

Our results show that not only we are able to acquire a highly effective and comprehensive dynamic RGB-D facial dataset, but also validates its effectiveness by employing it in an important application of visual speech recognition. The acquired dataset is versatile, and a number of visual speech recognition algorithms can be applied on the data.

In the following sections, we will first discuss the data acquisition setup in Section II. Afterward, we will present all the details of the dataset in Section III. Different algorithms applied for the visual speech recognition are presented in Section IV. Finally, the paper concludes in Section V.

II. Data Acquisition

We acquire the RGB-D dynamic facial dataset using Microsoft Kinect. We used Kinect face tracking SDK [42] that can track the facial outline and its feature points at 15 frames per second. In addition, Kinect also captures both the RGB (color) and depth information of the human actor at the same frame rate. An example of the real-time visualization of the RGB, depth and facial tracking data from Kinect can be seen in Fig. 1a, 1b, and 1c.

The resolution of RGB and depth data from Kinect is 640x480 pixels. The maximum supported frame rate is 30 frames per second. The overhead of face tracking SDK reduces the frame rate to 15 frames per second. This is not ideal but is still real-time. For the acquisition, we connect Kinect to a PC comprising of Intel i7 running Windows 10 with 16 GB of RAM. Kinect implicitly provides the mapping between RGB and depth data, along with the mapping of the depth data to a three-space coordinate system. As the face is tracked in the RGB space, mapping the mouth area to the RGB data is trivial as shown in Fig. 1d. The RGB to depth mapping also allows to map the tracked mouth region to the depth data, as shown in Fig. 1e. Finally, these mappings can also be used to generate a 3D point cloud with RGB mapping. An example of the 3D point cloud can be seen in Fig. 1f.

For each recording session, we record a person saying 20 different English words or phrases. The details of the dataset can be seen in the next section. On average, each recording takes around 80 seconds, resulting in 1200 frames. At each frame, we record facial points, facial outline, RGB data, depth data, mapping between RGB and depth data, facial animation units, facial shape units, and finally 2D and 3D face representations of the face along with the 3D head orientation. To minimize the overhead of writing the data on the disk, we dynamically store all the data in the memory and write it to the disk at the end of the acquisition. The details of the dataset are discussed in the next section.

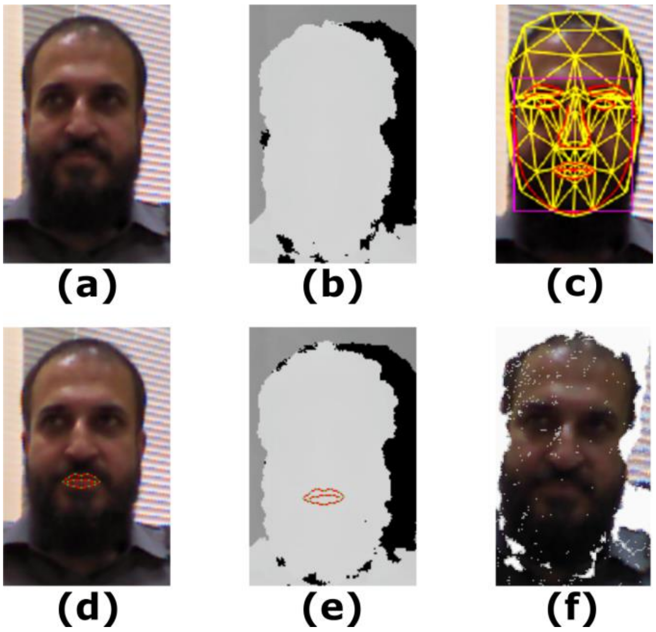

Fig. 1. (a) RGB image, (b) Depth image, (c) Tracked face with facial features, (d) tracked mouth area overlapping the RGB image, (e) tracked mouth area overlapping the depth image, and (f) 3D point cloud.

III. Dataset

As explained in the previous section, we acquire all the data using the Kinect face tracking SDK. The details of RGB and depth data are given in the previous section, below we describe the details of the facial parts of the dataset. The dataset is comprised of 20 individuals, each saying 20 words or phrases of English language in a single session. The 20 words or phrases are: Hello, Excuse me, I am sorry, Thank you, Goodbye, See you, Nice to meet you, You are welcome, How are you?, Have a good time, Who is calling?, Time is up, I agree, I love this game, So far so good, Anything else, What's up?, So do I, Be careful, and Bottoms up.

We record the 20 people saying these words or phrases using the Kinect face tracking SDK and capture the following facial features at each frame:

1) One hundred feature points, shown in Fig. 2a. 87 points are shown in the figure, while 13 are not shown.

2) 3D Head pose angles: Yaw, Pitch and Roll, as shown in Fig. 2b.

3) A 3D mesh representation of the face, as shown in Fig. 1c. Eleven shape units that capture head height, eyebrows vertical position, eyes vertical position, eyes width, eyes height, eye separation distance, nose vertical position, mouth vertical position, mouth width, eyes vertical difference, and chin width.

4) Animation units that capture animation states of various facial muscles. These include upper lip raised or lowered, jaw raised or lowered, lip stretchiness, inner brow raised or lowered, lip corner depression, outer brow raised or lowered.

As can be seen from this description that our dataset is extremely comprehensive and provides the maximum possible RGB-D dynamic facial information required for a number of algorithms. It is to be noted that all the facial features are tracked in the RGB space, but since we have the RGB to depth mapping, we can easily find the three-space position of each facial feature point (Fig. 1e).

The RGB, depth, and facial data can easily be employed for the facial recognition in both RGB space and three-space. The shape and animation units combined with the facial features and 3D head pose angles can be used for emotion recognition. Finally, the facial features in two-space or three-space can be used to localize the mouth region and can be used for the visual speech recognition. In the following section we present a new method for visual speech recognition that is adapted from Yargic et al. [36], in addition we also test a number of visual speech recognition methods to show the versatility of our data set. Our results show the effectiveness of the captured dataset and robustness of the modified method.

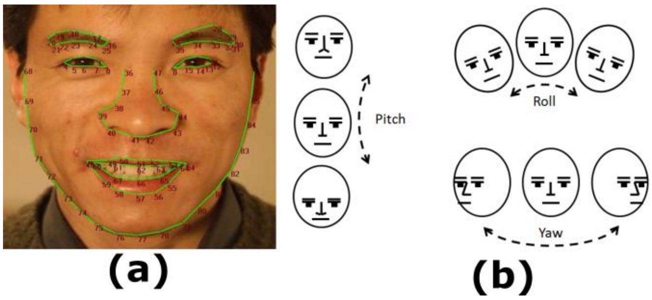

Fig. 2. (a) Facial features point tracked using Kinect face tracking SDK, and (b) 3D head orientation.

IV. Visual Speech Recognition

For the visual speech recognition, we adapt the KNN-based algorithm of Yargic et al. [36]. Kinect face tracking SDK captures 18 points in the mouth region. 10 points for the outer lips and 8 points for the inner lips (Fig. 2a). We segment the 20 words during the acquisition process, therefore there is no need for a post-acquisition segmentation. To this end, we know the number of frames T_w for each word segment S_w, where $w = 1 \ldots 20$.

For each word segment S_w, let the outer lip three-space points to be defined as O_i^t and the inner lip points as I_j^t. Where $i = 1 \ldots 10$, $j = 1 \ldots 8$, and $t = 1 \ldots T_w$. In their work, Yargic et al. [36], only considered four corner feature points on the outer lips to define two angles that are used to detect how much the mouth is opened. Their feature vector is comprised of 0s and 1s for each frame representing if the lips are passive or active at that frame. In our work, we do not discard any of the mouth feature points, rather we consider the three-dimensional trajectory of each O_i^t and I_j^t. For two consecutive frames, starting from the second frame, we compute 18 motion vectors in three-space using the 10 outer points: $V_w^t = O_i^t$-O_i^{t-1} and eight inner points: $V_w^t = I_j^t$-I_j^t. These 18 three-space vectors are projected on the XY, XZ and YZ planes. On each plane, the projected vectors have a specific orientation within a unit circle. The orientation with the maximum angle is assigned to each vector, and based on the maximum orientation we assign it a specific quadrant number from 1 to 8 (Fig. 3), or 0 if the motion is very small and the length of the vector is close to zero. Thus, at each frame we obtain an 18-dimensioanl feature vector V_w^t, where each coordinate of that vector is a value from 0 to 8, e.g. [2 1 4 1 3 6 8 1 2 4 3 0 5 1 7 7 0 1].

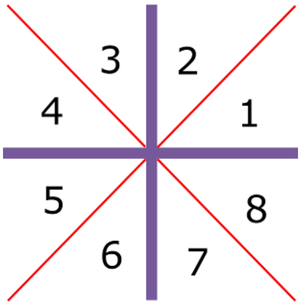

Fig. 3. Vector classification based on its projection in the coordinate frame.

For each word segment S_w, we obtain V_w^t, using the data from 10 subjects. For the other 10 subjects, V_w^t are used with KNN to classify the words. Compared to the accuracy results of 73.42% from Yargic et al. [36] using Manhattan distances, our method results in the 76.54% of correct classifications. In addition, we also compared the methods using Geodesic distances, and compared to the accuracy of 74.12% from Yargic et al. [36] our method results in the 78.42% of correct classifications.

The above results show that not only the modified algorithm is more effective compared to the original work, but in addition, our dataset can be used to for visual speech recognition. Therefore, to further validate the effectiveness of the dataset, we tested it using five more visual speech recognition algorithms.

The first method we directly applied was from Pao et al. [11]. They studied three different KNN methods. In our work, we directly employed weighted D-KNN and were able to obtain 82.0% accuracy compared to 79.3% with the D-KNN classifier weighted by the Fibonacci sequence.

The second direct application was for the optical flow-based method by Shaikh et al. [13]. We identified the motion of the feature points using the optical flow in both RGB and depth images. This resulted in a modified algorithm compared to the original work that only relied on the color data. Thus, for each feature point we ended up with two motion trajectories. Both of these motion vectors were then used for the training and recognition. Compared to the original work, our method slightly exceeds in the recognition of 1.5% at around 86%.

For the next comparison, we also used a neural network for the visual speech recognition, similar to Bagai et al [10]. Again, we made use of both RGB and depth data. Their recognition rate was on a lower side at 52%. Compared to it, our results are moderately higher at around 65%.

We also compared the methods proposed by Rekik et al. [32][33]. We only opted for the Speaker Independent (SI) setting to make sure that there is no bias due to the inclusion of the same speaker for both training and testing. In their work, they obtained the best performance with the combination of three $HOG_c + HOG_d + MBH$ descriptors, giving the accuracy of 64.2%. In contrast, on our dataset the accuracy of this method was 66.3%.

Finally, we tested the very recent work by Petridis et al. [17] that employed end-to-end deep learning for visual speech recognition. As a comparison, we tested their method on 20 subjects from OuluVS2 database. The network was trained on 10 subjects and tested on the other 10. Similarly, we also tested their method on our dataset with the same structure of 10 subjects for training and 10 for testing. Their method resulted in high classification accuracy. For the OuluVS2 dataset their method resulted in 86.3% accuracy, and for our dataset the accuracy was 87.1%. This shows that our dataset is well suited to be employed in a deep learning architecture. A comparison of all tested method on their original dataset compared to our dataset can be seen in Fig. 4.

Our method has a couple of limitations. The number of participants in our study are medium. In order to expand the data, we need varied participants and a large number of participants. As this is the initial work and recording 20 different words with Kinect face tracking is a time-consuming task, we had to rely on 20 volunteers. In future, we would like to rectify this by increasing the number of participants. We would also like to keep a good balance between the male and female participants. In addition, we would like to increase the diversity in terms of the age, so that all the age groups are covered.

Similarly, we are currently relying on Kinect v1 for the capture of facial data. Kinect v2 has advantages in terms of the quality of the data and the higher capture rate. Therefore, in future, we would like to employ Kinect v2 to capture even more comprehensive dataset and augment the current data with higher quality features and data captured at higher frame rate.

Despite the limitations, we have shown the an RGB-D sensor can be used to capture a comprehensive facial dataset. The dataset can be employed successfully for visual speech recognition, and having depth data allows for better recognition results compared to the earlier methods.

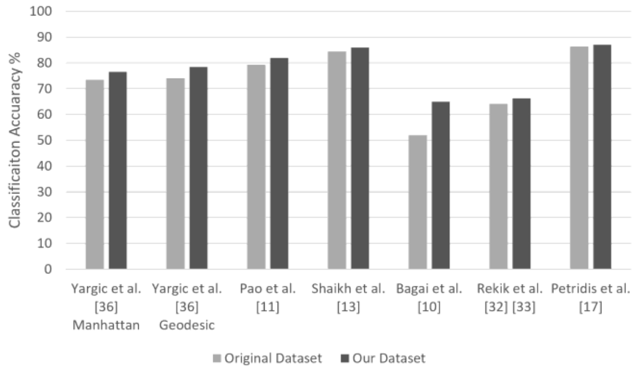

Fig. 4. Comparative results from applying a number of visual speech recognition methods from the literature. The methods are tested on their original dataset and the proposed dataset. The results show that the proposed dataset can be used for visual speech recognition and due to the higher quality of data, all the results show improvements.

V. Conclusion

This work presented a new RGB-D acquisition system to capture a comprehensive dynamic facial dataset that can be used for visual speech recognition. The RGB-D facial dataset acquisition system uses a Kinect to record detailed facial features of a person. The dynamic facial dataset is comprised of the facial data of 20 individuals saying 20 common English words or phrases. The acquisition system employs Kinect facial tracking, which records a large number of dynamic facial features. These features include: facial points, facial outline, RGB data, depth data, mapping between RGB and depth data, facial animation units, facial shape units, and finally 2D and 3D face representations of the face along with the 3D head orientation.

The comprehensive dynamic facial dataset can be employed in a number of applications including visual speech recognition. The effectiveness of acquired RGB-D dynamic facial dataset is demonstrated by presenting a new visual speech recognition method that employs three-dimensional spatiotemporal data of different facial feature points in the mouth region. The feature points are tracked over the word segment and their motion trajectory is used to create a novel representation of lips motion resulting in an 18 dimensional feature vector for each frame. We used KNN to classify the words and demonstrate the effectiveness of our method by 76.54% of correct classification. In addition, we also tested five more visual speech recognition methods on our dataset. These methods used weighted D-KNN, optical flow, neural network, HOG_c + HOG_d + MBH descriptors, and deep learning for visual speech recognition. We obtain comparable or better results when these methods were tested on our dataset.

The results demonstrate our comprehensive RGB-D dynamic facial dataset can be effectively employed in a visual speech recognition system. It also shows that having additional depth and three-space information greatly helps the visual speech recognition algorithms. In future, we would like to extend our work and employ the new dataset for facial and emotion recognition.

References

[1] G. L. Zheng, M. Zhu, and L. Feng, “Review of lip-reading recognition,” Proc. - 2014 7th Int. Symp. Comput. Intell. Des. Isc. 2014, vol. 1, pp. 293–298, 2015.

[2] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, “Automatic scoring of pronunciation quality,” Speech Commun., vol. 30, no. 2, pp. 83–93, 2000.

[3] O. Turk and L. Arslan, “Speech Recognition Methods for Speech Theraphy,” in Signal Processing and Communications - IEEE 12th, 2004, pp. 410–413.

[4] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), 2005, pp. 1–15.

[5] N. Dalal, B. Triggs, and C. Schmid, “Human detection using oriented histograms of flow and appearance,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 3952 LNCS, pp. 428–441, 2006.

[6] L. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” in Proceeding of the IEEE, 1989, vol. 77, no. 2, pp. 257–285.

[7] G. Zhao, M. Barnard, and M. Pietikäinen, “Lipreading with local spatiotemporal descriptors,” IEEE Trans. Multimed., vol. 11, no. 7, pp. 1254–1265, 2009.

[8] K. T. Chen, W. W. Liau, H. M. Wang, and L. S. Lee, “Fast speaker adaptation using eigenspace-based maximum likelihood linear

regression," 6th Int. Conf. Spok. Lang. Process. ICSLP 2000, pp. 2–5, 2000.

[9] Y. Pei, T. K. Kim, and H. Zha, "Unsupervised random forest manifold alignment for lipreading," Proc. IEEE Int. Conf. Comput. Vis., pp. 129–136, 2013.

[10] D. Kalbande, A. A. Mishra, S. Patil, S. Nirgudkar, and P. Patel, "Lip reading using neural networks," Int. Conf. Graph. Image Process. (ICGIP 2011), vol. 8285, no. 4, p. 828519, 2011.

[11] T. L. Pao, W. Y. Liao, and Y. Te Chen, "Audio-visual speech recognition with weighted KNN-based classification in Mandarin database," Proc. - 3rd Int. Conf. Intell. Inf. Hiding Multimed. Signal Process. IIHMSP 2007., vol. 1, pp. 39–42, 2007.

[12] Lin Chen, Dong Wei, and Jie-Sheng Wang, "Research on Magnetic Resonance Imaging Segmentation Algorithm," Engineering Letters, vol. 27, no.3, pp559-567, 2019.

[13] A. Shaikh, D. K. Kumar, W. C. Yau, M. Z. C. Azemin, and J. Gubbi, "Lip reading using optical flow and support vector machines," Proc. - 2010 3rd Int. Congr. Image Signal Process. CISP 2010, vol. 1, pp. 327–330, 2010.

[14] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Lipreading using convolutional neural network," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, no. September, pp. 1149–1153, 2014.

[15] Assael, Yannis & Shillingford, Brendan & Whiteson, Shimon & Freitas, Nando. (2016). LipNet: Sentence-level Lipreading.

[16] Y. Lu and H. Li, "Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory," Appl. Sci., vol. 9, no. 8, 2019.

[17] Stavros Petridis, Yujiang Wang, Pingchuan Ma, Zuwei Li, Maja Pantic, End-to-end visual speech recognition for small-scale datasets, Pattern Recognition Letters, Volume 131, 2020.

[18] J. L. Newman and S. J. Cox, "Language identification using visual features," IEEE Trans. Audio, Speech Lang. Process., vol. 20, no. 7, pp. 1936–1947, 2012.

[19] A.-G. A.-R. Abdel-Samie, F. A. Elmisery, A. M.Brisha, and A. H. Khalil, "Arabic Sign Language Recognition Using Kinect Sensor," Res. J. Appl. Sci. Eng. Technol., vol. 15, no. 2, pp. 57–67, 2018.

[20] S. Bhaskar, T. M. Thasleema, and R. Rajesh, "A Survey on Different Visual Speech Recognition Techniques," in Data Analytics and Learning, Lecture Notes in Networks and Systems, vol. 43, P. Nagabhushan, Ed. Springer Singapore, 2019, pp. 413–419.

[21] R. Shehzad, N. Ahmad, M. W. Iqbal, and I. . Feroz, "Gestural User Interfaces for Hearing and Speech Impaired People Using KINECT," in International Conference on Engineering and Emerging Technologies (ICEET), 2019.

[22] J. Yu, Q. Wang, and H. Chen, "Application of kinect-based motion recognition algorithm in cultural tourism," ACM Int. Conf. Proceeding Ser., pp. 307–311, 2018.

[23] Newen, A. Welpinghus, and G. Juckel, "Emotion recognition as pattern recognition: The relevance of perception," Mind Lang., vol. 30, no. 2, pp. 187–208, 2015.

[24] Wendemuth, B. Vlasenko, I. Siegert, R. Böck, F. Schwenker, and G. Palm, "Emotion Recognition from Speech," in Companion Technology, Cognitive Technologies, A. W. S. Biundo, Ed. Springer, 2017, pp. 409–428.

[25] K. Uhler, A. Biever, and R. H. Gifford, "Method of speech stimulus presentation impacts pediatric speech recognition monitored live voice versus recorded speech," Otol. Neurotol., vol. 37, no. 2, pp. e70–e74, 2016.

[26] J. Svec, "Designing Advanced Geometric Features for Automatic Russian Visual Speech Recognition," in International Conference on Speech and Computer, 2018, vol. 1, pp. 245–254.

[27] J. Volak, D. Koniar, L. Hargas, F. Jabloncik, N. Sekel'Ova, and P. Durdik, "RGB-D imaging used for OSAS diagnostics," 12th Int. Conf. ELEKTRO, pp. 1–5, 2018.

[28] Bin Dris, A. Alsalman, A. Al-Wabil, and M. Aldosari, "Intelligent Gaze-Based Screening System for Autism," 2019 2nd Int. Conf. Comput. Appl. Inf. Secur., pp. 1–5, 2019.

[29] D. A. Trevisan, M. Hoskyn, and E. Birmingham, "Facial Expression Production in Autism: A Meta-Analysis," Autism Res., vol. 11, no. 12, pp. 1586–1601, 2018.

[30] Simon and W. M. Baum, "Allocation of speech in conversation," J. Exp. Anal. Behav., vol. 107, no. 2, pp. 258–278, 2017.

[31] S. van der Zee, R. Poppe, P. J. Taylor, and R. Anderson, "To freeze or not to freeze: A culture-sensitive motion capture approach to detecting deceit," PLoS One, vol. 14, no. 4, 2019.

[32] Rekik A., Ben-Hamadou A., Mahdi W. (2014) A New Visual Speech Recognition Approach for RGB-D Cameras. In: Campilho A., Kamel M. (eds) Image Analysis and Recognition. ICIAR 2014. Lecture Notes in Computer Science, vol 8815. Springer, Cham

[33] Rekik, A. Ben-Hamadou, and W. Mahdi, "An adaptive approach for lip-reading using image and depth data," Multimed. Tools Appl., vol. 75, no. 14, pp. 8609–8636, 2016.

[34] Parekh, A. Gupta, S. Chhatpar, A. Y. Kumar, and M. Kulkarni, "Lip Reading Using Convolutional Auto Encoders as Feature Extractor," 2018.

[35] Lee, J. Lee, and K.-E. Kim, "Multi-view Automatic Lip-Reading Using Neural Network," in ACCV 2016 Workshops, vol. 1, C. Chen, J. Lu, and K. Ma, Eds. 2017, pp. 277–289.

[36] Yargic and M. Dogan, "A lip reading application on MS Kinect camera," 2013 IEEE Int. Symp. Innov. Intell. Syst. Appl. IEEE INISTA 2013, 2013.

[37] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," IEEE Trans. Cybern., vol. 43, no. 5, pp. 1318–1334, 2013.

[38] T. Fujita and T. Yoshida, "3d terrain sensing by laser range finder with 4-dof sensor movable unit based on frontier-based strategies," Engineering Letters, vol. 24, no. 2, pp. 164–171, 2016.

[39] G. Sanchez, E. Leal, and N. Leal, "A linear programming approach for 3d point cloud simplification," IAENG International Journal of Computer Science, vol. 44, no. 1, pp. 60–67, 2017.

[40] N. Ahmed, "Spatio-Temporally Coherent 3D Animation Reconstruction from Multi-view RGB-D Images using Landmark Sampling," In Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1, Mar. 13, 2013.

[41] N. Ahmed, and M. Lataifeh, "A Comparative Analysis of Time Coherent 3D Animation Reconstruction Methods from RGB-D Video Data," IAENG International Journal of Computer Science, vol. 45, no.4, pp592-600, 2018

[42] Microsoft, "Kinect for Windows SDK 1.8 Face Tracking." https://developer.microsoft.com/en-us/windows/kinect/, 2012

[43] M. Schatz, A. Prochazka, O. Tupa, O. Vysataa, and V. Sedlak, "Face Movement Analysis with MS Kinect," in International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)., 2016.

Naveed Ahmed is an associate professor at the Department of Computer Science, University of Sharjah. He received his PhD in computer science from the University of Saarland (Max-Planck-Institute for Informatics), Germany, in 2009. He worked as a research and development engineer at Autodesk in Cambridge, United Kingdom, for 2 years. He is currently working as an assistant professor at the Department of Computer Science, University of Sharjah. His research interests include 3-D animation, dynamic scene reconstruction, vision-based computer graphics, and multi-view video-based modeling and rendering.

Mohammed Lataifeh is an assistant professor at the Department of Computer Science, University of Sharjah. He received his PhD in design and Information Technology from the De Montfort University (Institute of Creative Technologies), UK, in 2015. He worked for several institutions as solution consultant (ERP systems, E-commerce, Graphics and Multimedia). He is currently working as an assistant professor at the Department of Computer Science, University of Sharjah. His research interests include enterprise systems, graphics, and HCI within mixed reality environments.

Imran Junejo is an associate professor at the College of Technical Innovation, Zayed University. He received his PhD in computer science from the University of Central Florida, USA, in 2007. After a post doc at INRIA-Rennes, he joined the Department of Computer Sciences, University of Sharjah, where he worked for 8 years. His current focus of research is human action recognition from arbitrary views. Other areas of research interests include camera calibration, metrology, path modeling, video surveillance, scene understanding, and event detection.