2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Upper Airway Classification in Sleep Endoscopy Examinations using Convolutional Recurrent Neural Networks*

Umaer Hanif[1,3,4], *Member, IEEE*, Eric Kezirian[2,5], Eva Kirkegaard Kiær[3,5], Emmanuel Mignot[4,5], Helge B. D. Sorensen[1,5], *Senior Member, IEEE*, and Poul Jennum[3,5]

***Abstract*— Assessing the upper airway (UA) of obstructive sleep apnea patients using drug-induced sleep endoscopy (DISE) before potential surgery is standard practice in clinics to determine the location of UA collapse. According to the VOTE classification system, UA collapse can occur at the velum (V), oropharynx (O), tongue (T), and/or epiglottis (E). Analyzing DISE videos is not trivial due to anatomical variation, simultaneous UA collapse in several locations, and video distortion caused by mucus or saliva. The first step towards automated analysis of DISE videos is to determine which UA region the endoscope is in at any time throughout the video: V (velum) or OTE (oropharynx, tongue, or epiglottis). An additional class denoted X is introduced for times when the video is distorted to an extent where it is impossible to determine the region. This paper is a proof of concept for classifying UA regions using 24 annotated DISE videos. We propose a convolutional recurrent neural network using a ResNet18 architecture combined with a two-layer bidirectional long short-term memory network. The classifications were performed on a sequence of 5 seconds of video at a time. The network achieved an overall accuracy of 82% and F1-score of 79% for the three-class problem, showing potential for recognition of regions across patients despite anatomical variation. Results indicate that large-scale training on videos can be used to further predict the location(s), type(s), and degree(s) of UA collapse, showing potential for derivation of automatic diagnoses from DISE videos eventually.**

I. INTRODUCTION

Obstructive sleep apnea (OSA) is a sleep disorder during which the upper airway (UA) collapses throughout the night, causing events with partial or complete cessation of breathing during sleep [1]. The development of OSA can be physiologically caused (loop gain, arousal threshold, poor recruitment of dilator muscles) [2] or anatomically caused (craniofacial abnormalities, obesity, narrow UA) [3] and treatment varies depending on the underlying cause. If the pathology of OSA has an anatomical component, surgery may be necessary for treatment [4]. Prior to a potential surgical procedure, it is critical to examine the location(s) of collapse in the UA, which according to the VOTE classification system [5] can occur on four different levels: velum, oropharynx, tongue, and/or epiglottis. The examination is commonly performed using drug-induced sleep endoscopy (DISE) during which the surgeon navigates the endoscope from the velum to the epiglottis to determine the location(s), type(s), and degree(s) of collapse occuring in the UA during OSA events [6].

Analyzing DISE videos to determine the appropriate type of surgery is not a trivial task. First, there is a huge anatomical variation in the UA across subjects. Additionally, movements in the UA stemming from several structures collapsing simultaneously push the endoscope back and forth, while mucus or saliva covering the endoscope distorts the video and reduces quality significantly. These challenges are reflected in a relatively high interscorer variability when different surgeons analyze DISE videos [7]. Due to these limitations, surgeons will benefit from an algorithm capable of analyzing DISE videos automatically to assist in determining the locations(s), type(s), and degree(s) of collapse.

The first step towards such a goal is being able to estimate which region of the UA the endoscope is in at any given time. The clinically meaningful distinction is between the velum (V) and anything below the velum (OTE). Thus, the aim of this study is to classify whether the endoscope is in the V or OTE region at any given time in a DISE video. Furthermore, we introduce a third class (X) for any time the video is so distorted that it is impossible to determine where the endoscope is. For this problem, we propose a convolutional recurrent neural network (CRNN) which is trained, validated and tested on a small dataset of annotated DISE videos. This study is the first attempt to apply a data-driven approach to identify regions in the UA during a DISE procedure.

II. DATA DESCRIPTION

We included a total of 24 DISE videos collected at Copenhagen University Hospital, which were performed in accordance with the DISE procedure guideline described by Kiaer et al. [8]. The Institution's Ethical Review Board approved all experimental procedures involving human subjects.

The videos were approximately 2-5 minutes in duration with a frame rate of 25 frames per second and a resolution of 864×540 pixels. All videos were anonymized by removing parts of recordings where the endoscope was not inside the

*Research has been supported by the Klarman Family Foundation, Stanford University, Technical University of Denmark, and Rigshospitalet with supporting grants from Danmark-Amerika Fondet, Vera og Carl Johan Michaelsens Legat, Reinholdt W. Jorck og Hustrus Fond, Torben og Alice Frimodts Fond, Christian og Ottilia Brorsons Rejselegat, Marie og M.B. Richters Fond, Oberstløjtnant Max Nørgaard og hustru Magda Nørgaards Legat, William Demant Fonden, Augustinus Fonden, Rudolph Als Fondet, Knud Højgaards Fond, Otto Mønsteds Fond, Julie von Müllens Fond, and Direktør Einar Hansen og hustru fru Vera Hansens Fond

[1]Department of Health Technology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark, umaerhanif@hotmail.com

[2]USC Caruso Department of Otolaryngology - Head & Neck Surgery, Keck School of Medicine of USC, Los Angeles, CA 90033, USA

[3]Danish Center for Sleep Medicine, Rigshospitalet, 2600 Glostrup, Denmark

[4]Stanford Center for Sleep Sciences and Medicine, Stanford University, Palo Alto, CA 94304, USA

[5]Shared last authors

subject. Each video was initially labeled by the surgeon who collected them as a single line summary of where, how, and to what degree the UA collapsed. However, for machine learning purposes, labels were required that detailed each time the endoscope transitioned either from one region to another (i.e. V to OTE or OTE to V) or from visible video to distorted video or vice versa (i.e. V to X, OTE to X, X to V or X to OTE). Videos were labeled in this manner by consulting with the surgeon who initially labeled the videos and another expert surgeon who introduced the VOTE classification in 2011 [5]. Fig. 1 visualizes different examples of the three classes, while Table I shows an example of the structure of labels created for this study. Finally, Table II outlines the distribution of the three classes within the dataset.

Fig. 1. Three examples of each class representing a region in the upper airway, i.e. velum (V) in the first column, oropharynx, tongue or epiglottis (OTE) in the second column, and distortion in video (X) in the third column.

TABLE I

EXAMPLE OF LABELS CREATED FOR PART OF A DISE VIDEO USING THE THREE CLASSES, I.E. VELUM (V), OROPHARYNX, TONGUE OR EPIGLOTTIS (OTE), AND DISTORTION IN VIDEO (X).

Time (s)	7-15	16-28	29-35	36-40	40-45
Region	V	X	OTE	X	V

TABLE II

DISTRIBUTION OF THE THREE CLASSES IN THE DATASET: VELUM (V), OROPHARYNX, TONGUE OR EPIGLOTTIS (OTE), AND DISTORTION IN VIDEO (X).

Class	Total duration (s)	N Frames
V	1,543	7,715
OTE	2,041	10,205
X	376	1,880
Total	3,960	19,800

III. METHODS

A. Preprocessing

Initially, all frames were extracted from each video, yielding 25 frames per second. Subsequently, every 5th frame was selected, yielding 5 frames per second, because no visual difference was observed between consecutive frames during inspection. Assuming the network would extract features primarily related to anatomical structures and not color differences, all frames were converted to gray scale to reduce computational cost of training the subsequent network. All frames were rescaled to 224×224 pixels, which was found to be appropriate for reducing computational cost while still preserving discriminatory information between UA structures. Finally, the dataset was split into a training set (18 videos amounting to 15,275 frames), a validation set (3 videos amounting to 2,375 frames), and a test set (3 videos amounting to 2,150 frames).

B. Convolutional Recurrent Neural Network

The proposed network architecture for learning was a combination of a ResNet18 [9] convolutional neural network (CNN) and a two-layer bidirectional long short-term memory (LSTM) neural network [10] as shown in Fig. 2. The input layer of the ResNet18 model was modified to take a 1-channel input instead of RGB images with 3 channels, since the frames were grayscale. The input consisted of 25 frames amounting to 5 seconds at 5 frames per second. Each frame was individually input to the CNN and resulting outputs were subsequently concatenated, forming a 25×512 dimensional feature matrix, i.e. 25-time steps each with 512 features. This sequence of features was then input to the bidirectional LSTM to learn context in both forward and backward directions. Both LSTM layers had 128 hidden neurons in both directions followed by a softmax activation function with three outputs such that each class had an output probability. The optimal number of time-steps and hidden neurons were found using hyperparameter tuning.

Optimization of the network was performed using a batch size of 2 with cross entropy as loss function and Adam [11] as optimizer. Weights were applied in the loss function for the V and X classes, since the dataset was heavily imbalanced as witnessed in Table II. The weights were calculated as the ratio between the majority class (OTE) and a given other class. The learning rate was set to $1 \cdot 10^{-5}$ with a weight decay of $5 \cdot 10^{-4}$. Early stopping was applied when the validation loss did not decrease for 3 consecutive epochs. The network was implemented in Pytorch and all experiments were carried out on a GeForce RTX 2080 graphics card. The model took approximately one hour to train on this dataset.

C. Performance

Model performance was evaluated on the three videos in the test set. Accuracy, F1-score, and the confusion matrix were computed by summing correct classifications on a frame-by-frame basis and averaged over individual videos as well as over the entire test set, respectively.

Operation	Out dim [C, H, W]
Conv (7x7, 64, s = 2, p = 3), BatchNorm, ReLU	[64, 112, 112]
MaxPool (3x3, s = 2)	[64, 56, 56]
Conv (3x3, 64), BatchNorm, ReLU	[64, 56, 56]
Conv (3x3), BatchNorm	[64, 56, 56]
Conv (3x3), BatchNorm, ReLU	[64, 56, 56]
Conv (3x3), BatchNorm	[64, 56, 56]
Block (1)	[128, 28, 28]
Block (2)	[256, 14, 14]
Block (4)	[512, 7, 7]
AvgPool, Flatten	[512]
Operation	**Out dim [seq_len, C]**
Concatenate outputs from 25 frames	[25, 512]
BiLSTM (nI = 25, nH = 256)	[25, 256]
FC (in_dim = nH · 2, out_dim = 128)	[25, 128]
BiLSTM (nI = 25, nH = 256)	[25, 256]
FC (in_dim = nH · 2, out_dim = 3)	[25, 3]
Softmax	[25, 3]

Block (k)

Operation	Out dim [C, H, W]
Conv (3x3, 128 · k, s = 2), BatchNorm, ReLU	[128 · k, 56/(2 · k), 56/(2 · k)]
Conv (3x3, 128 · k), BatchNorm	[128 · k, 56/(2 · k), 56/(2 · k)]
Conv (3x3, 128 · k), BatchNorm, ReLU	[128 · k, 56/(2 · k), 56/(2 · k)]
Conv (3x3, 128 · k), BatchNorm	[128 · k, 56/(2 · k), 56/(2 · k)]
Bottleneck Conv (1x1, 128 · k, s = 2), BatchNorm	[128 · k, 56/(2 · k), 56/(2 · k)]

Fig. 2. Architecture for the proposed network for classifying UA regions. The input is a 5-second video consisting of 25 frames. The frames are input individually to the CNN and the outputs are concatenated before the recurrent part of the network. The parameters in the convolution operations (Conv) are kernel size, number of output channels, stride (s), and padding (p), and the output dimensions are specified by number of channels (C), height (H), and width (W). The parameters in the bidirectional LSTM (BiLSTM) are number of input features (nI) and number of hidden neurons in each direction (nH). The parameters in the fully connected layers (FC) are input features (in_dim) and output features (out_dim).

IV. RESULTS

The best performing model converged after 2 epochs of training. An overall accuracy of 82% and F1-score of 79% was obtained over the entire test set. Furthermore, F1-scores for V, OTE, and X were 68%, 80%, and 88%, respectively. Fig. 3 shows the confusion matrix for the classification, while Fig. 4 depicts examples of misclassified frames for each class. Table III summarizes the performance for each of the 3 individual videos in the test set, respectively.

V. DISCUSSION

This is the first attempt to use a data-driven approach to identify UA regions during the DISE procedure and the overall accuracy and F1-score obtained using the proposed model was 82% and 79%, respectively. In contrast, if the network had simply predicted all frames to be the majority class in the test set (i.e. OTE), the overall accuracy and F1-score would be 52% and 23%, respectively. In this context, the model performs much better than random guessing. In terms of class F1 scores, the model performed best for the X class, then OTE, and finally V.

The X class is intuitively the easiest to recognize since it means that the video is too distorted to derive anything and

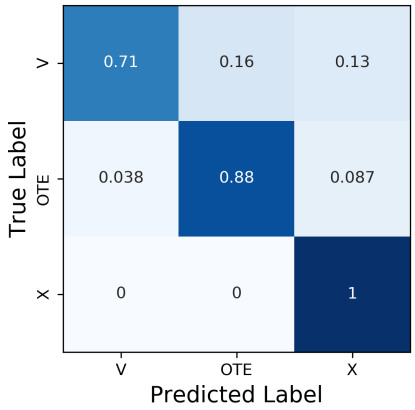

Fig. 3. Confusion matrix for classifying regions in the upper airway with three different classes: velum (V), oropharynx, tongue or epiglottis (OTE), and distorted video (X).

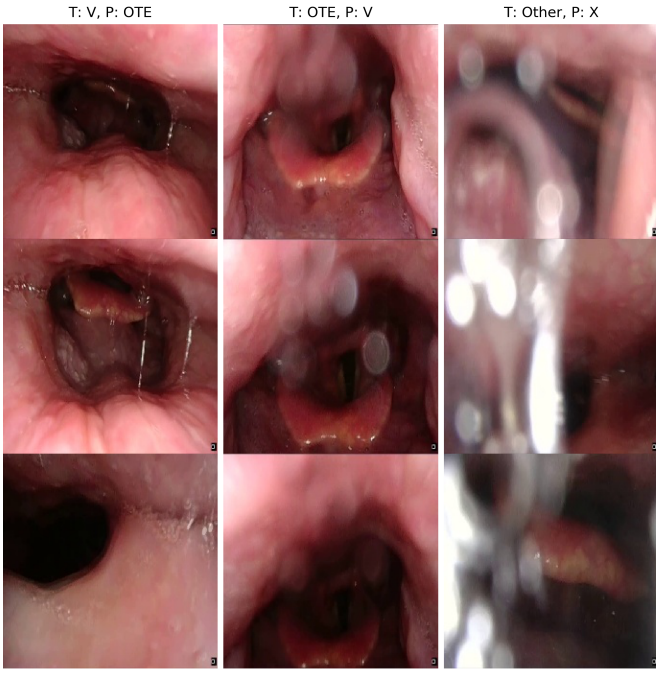

Fig. 4. Examples of misclassifications for each class, i.e. velum (V), oropharynx, tongue or epiglottis (OTE), and distorted video (X), where T is the true class and P is the predicted class.

TABLE III

PERFORMANCE FOR THE THREE VIDEOS IN THE TEST SET FOR CLASSIFYING REGIONS IN THE UPPER AIRWAY WITH THREE DIFFERENT CLASSES: VELUM (V), OROPHARYNX, TONGUE OR EPIGLOTTIS (OTE), AND DISTORTED VIDEO (X).

Video	Accuracy	F1	Class	Class F1	N Frames
1	93%	93%	V OTE X	94% 91% -	386 264 0
2	86%	75%	V OTE X	67% 93% 65%	166 741 93
3	63%	62%	V OTE X	61% 54% 70%	252 123 125

it would be a trivial task to recognize this class even for a person unfamiliar with DISE videos. This is also reflected by the fact that even with the limited number of frames with class X in the dataset (Table II), the model was easily able to learn to recognize this class. Looking at Fig. 3, it is noted that the sensitivity is 100%, meaning that none of the frames labeled X are misclassified. However, both the V and OTE classes are occasionally misclassified as X, which Fig. 4 shows examples of. It is noted that the model classifies a frame as X any time there is mucus or saliva on the endoscope even if some structures are still visible to some degree. When annotating the data, a frame was only labeled as X if there was no way to estimate the region based on the video or context from previous frames, while the model has learned the relation that any mucus or saliva on the endoscope equals a classification of X.

The model also performed well for the OTE class, reflected by a high sensitivity and F1 score. It is noted from Fig. 3 that when OTE is misclassified, it is mostly as X, which is again explained by the fact that the model is sensitive to mucus and saliva on the camera, even if it is possible to derive the UA region. Scenarios where OTE is misclassified as V is illustrated in Fig. 4, where it is observed that this occurs when the endoscope is at the border between the V and OTE regions. Even experts analyzing these frames could have scored them as V instead of OTE, and it appears that the last frame at the bottom has been wrongly annotated as OTE even though the endoscope is in the V region.

For the V class, the model did not perform as well as for the two other classes. Fig. 3 shows that the misclassifications are almost equally split between OTE and X. The frames misclassified as X are due to the same reason as for OTE. Examples of V being misclassified as OTE are shown in Fig. 4. In this case it appears that the misclassifications do not necessarily occur when the endoscope is close to the OTE region, but rather when the OTE region is visible from the V region so that the model can recognize structures such as the tongue and epiglottis. It makes sense that with the limited amount of data the model has seen, it is not able to derive distance-based decisions to estimate the region as well as it recognizes structures associated with a given region. Furthermore, the large amount of noisy frames in the video (approximately 25%) most likely causes noise in the context of the bidirectional LSTM, which contributes to the poor performance for video 3.

Table III outlines performance for each individual video in the test set. It is observed that the best performance is obtained for video 1, which has no frames with distorted video and also few misclassifications for the V and OTE regions. During video 2, the endoscope is by far the most in the OTE region and the F1-score is high for that class. The F1-score for both V and X is modest because V is misclassified as both OTE and X, whereas OTE is misclassified a few times as X as well. During video 3, most time is spent in the V region but the F1 score is modest for all classes. In this case, V is still misclassified as both OTE and X, but OTE is also sometimes misclassified as V and not only X, which is most likely due to wrong annotations, similar to the bottom frame in the middle column of Fig. 4, where V is labeled as OTE.

There are two main limitations of this study: the quantity of data is extremely low, and the problem posed is simplistic with respect to utilizing this in clinical practice. However, the results serve as an important proof of concept, which shows that it is possible to apply deep learning techniques on DISE videos, even though they depict large variations in terms of both anatomical structure and angles/positions in the UA across videos. Considering this, it is quite impressive that the proposed model obtains such a high performance on so little data and that it actually manages to learn meaningful mappings between the classes and the series of frames that are used as input. For a future study, we will obtain a much larger quantity of data (1000 videos) and expand the problem for classification of where the UA collapses, how it collapses, and what the degree of collapse is.

VI. CONCLUSION

This study shows potential for large-scale learning on DISE videos in order to automatically recognize regions in the UA and thereby derive where the collapse occurs during OSA events, which is critical before any potential surgery to treat OSA. The study was performed on a very limited dataset and serves as a proof of concept for a future study, where a larger quantity of data will be utilized and several variables will be predicted. The presented method has potential application for use in clinical medicine to identify UA collapse.

REFERENCES

[1] P. Lévy, M. Kohler, W. T. McNicholas, F. Barbé, R. D. McEvoy, V. K. Somers, L. Lavie, and J. Pépin, Obstructive sleep apnoea syndrome, Nature Reviews Disease Primers, vol. 1, no. 1, pp. 1-21, 2015.

[2] A. M. Osman, S. G. Carter, J. C. Carberry, and D. J. Eckert, Obstructive sleep apnea: current perspectives, Nature and Science of Sleep, vol. 10, pp. 21–34, 2018.

[3] R. W. W. Lee, K. Sutherland, and P. A. Cistulli, Craniofacial Morphology in Obstructive Sleep Apnea: A Review, Clinical Pulmonary Medicine, vol. 17, no. 4, pp. 189–195, 2010.

[4] K. K. Green et al., Drug-Induced Sleep Endoscopy and Surgical Outcomes: A Multicenter Cohort Study, Laryngoscope, vol 129, pp. 761–770, 2019.

[5] E. J. Kezirian, W. Hohenhorst, and N. de Vries, Drug-induced sleep endoscopy: the VOTE classification, European Archives of Oto-Rhino-Laryngology, vol. 268, pp. 1233–1236, 2011.

[6] W. Hohenhorst, M. J. L. Ravesloot, E. J. Kezirian, and N. de Vries, Operative Techniques in Otolaryngology, vol. 23, no. 1, pp. 3-10, 2012.

[7] E. J. Kezirian, D. P. White, A. Malhotra, W. Ma, C. E. McCulloch, and A. N. Goldberg, Interrater Reliability of Drug-Induced Sleep Endoscopy, Archives of Otolaryngology - Head & Neck Surgery, vol. 136, no. 4, pp. 393-397, 2010.

[8] E. K. Kiaer, P. Tonnesen, H. B. Sorensen, N. Rubek, A. Hammering, C. Moller, A.M. Hildebrandt - P.J. Jennum - C. von Buchwald, Propofol sedation in Drug Induced Sedation Endoscopy without an anaesthesiologist – a study of safety and feasibility, Rhinology, vol. 57, no. 2, pp. 125-131, 2019.

[9] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, Deep Residual Learning for Image Recognition, The IEEE Conference on Computer Vision and Pattern Recognition, Nevada, 2016, pp. 770-778.

[10] M. Schuster and K. K. Paliwal, Bidirectional Recurrent Neural Networks, IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, 1997.

[11] D. P. Kingma and J. L. Ba, Adam: A Method for Stochastic Optimization, 3rd International Conference on Learning Representations, San Diego, 2015.