Facial Image Analysis for Fully-Automatic Prediction of Difficult Endotracheal Intubation

Gabriel L. Cuendet, *Student Member, IEEE,* Patrick Schoettker, Anıl Yüce *Student Member, IEEE,* Matteo Sorci, Hua Gao, Christophe Perruchoud, Jean-Philippe Thiran, *Senior Member, IEEE*

***Abstract*—*Goal:* Difficult tracheal intubation is a major cause of anesthesia related injuries with potential life threatening complications. Detection and anticipation of difficult airway in the preoperative period is thus crucial for the patients' safety. We propose an automatic face analysis approach to detect morphological traits related to difficult intubation and improve its prediction. *Methods:* For this purpose, we have collected a database of 970 patients including photos, videos and ground truth data. Specific statistical face models have been learned using the faces in our database providing an automated parametrization of the facial morphology. The most discriminative morphological features are selected through the importance ranking provided by the random forest algorithm. The random forest approach has also been used to train a classifier on these selected features. We compare a threshold tuning method based on class prior with two methods which learn an optimal threshold on a training set for tackling the inherent imbalanced nature of the database. *Results:* Our fully-automated method achieves an AUC of 81.0% in a simplified experimental setup where only easy and difficult patients are considered. A further validation on the entire database has proven that our method is applicable for real-world difficult intubation prediction, with AUC = 77.9%. *Conclusion:* The system performance is in line with the state-of-the-art medical diagnosis, based on ratings provided by trained anesthesiologists, whose assessment is guided by an extensive set of criteria. *Significance:* We present the first completely automatic and non-invasive difficult intubation detection system that is suitable for use in clinical settings.**

***Index Terms*—Anesthesia, Difficult intubation prediction, Pattern recognition, Facial image analysis.**

I. Introduction

THE priority of the anesthesiologist, after having induced general anesthesia is to ventilate the patient and secure his airways. As the patient is under the influence of drugs, whose main effects are the loss of consciousness, analgesia and muscular paralysis, mechanical ventilation is mandatory. Despite all the advancements in anesthesiology, difficult airway management still represents a major cause of anesthesia-related injuries with potential life threatening complications [1]. Recent analysis of airway management related claims in the UK [2] and in the USA [3] show that respiratory events, most of them being difficult intubation or inadequate ventilation, come first in the proportion of cases with poor clinical outcomes (severe harm, brain damage or death). The worst case scenario in airway management is the *"Can't intubate, can't ventilate"* situation in which the patient is impossible to be ventilated by face mask and intubated with an endotracheal tube. The estimated incidence of such a situation is estimated between 0.01 and 3 in 10'000 cases [4]. Nowadays, up to one third of all deaths attributed to anesthesia are consecutive to the inability to either ventilate or intubate [5]. Numerous technical advances have allowed facilitation of intubation by improving the view at laryngoscopy [6]–[8] or monitoring the placement of the endotracheal tube [9], [10] but difficult intubation still remains an area of concern [2].

Detection and anticipation of difficult airway in the preoperative period is crucial for patients' safety. In cases of suspected difficulty, specific equipment and personnel will be called upon to increase safety and the chances of successful intubation. In daily practice, anesthesiologists predict the difficulty of tracheal intubation with bedside tests, which correlate poorly with the ground truth. Experienced anesthesiologists associate, in addition to the available bedside tests, a global clinical judgment, probably based on a larger number of morphological parameters than those contained in the available bedside tests (see section I-A). Nevertheless some patients with a difficult airway remain undetected despite the most careful preoperative airway evaluation.

The usage of computer vision methods, and more specifically face analysis methods, is on the rise in areas such as marketing and emotion analysis [11], [12], face-tracking systems to increase safety in cars [13], [14] as well as in medicine [15]–[17] to name just a few. Improvements in facial landmarks detection and tracking [18], [19] allow for fast and robust face trackers [20]. Those can detect and interpret specific features of the face, based on landmark positions, making them suitable for facial morphology analysis.

In this study we describe a clinical application of face analysis to detect morphological traits related to difficult intubation, hypothesizing that advanced face analysis methods could improve the prediction of difficult intubation and iden-

G. L. Cuendet, A. Yüce and H. Gao are with the Signal Processing Lab (LTS5), Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, e-mail:(see http://people.epfl.ch/name.surname)

J.-P. Thiran is with the Signal Processing Lab (LTS5), Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland and the University Hospital Center (CHUV) and University of Lausanne (UNIL), Switzerland, e-mail:(see http://lts5www.epfl.ch/thiran.html)

P. Schoettker, M. Sorci and C. Perruchoud are with the Department of Anaesthesiology, University Hospital Center (CHUV), Lausanne, Switzerland, e-mail: name.surname@chuv.ch. C. Perruchoud is also with the Department of Anesthesiology and Pain Management, Ensemble Hospitalier de la Côte (EHC), Morges, Switzerland

This project was funded by the swiss commission for technology and innovation (CTI), under the project number 12636.1 and the title *Prediction of difficult tracheal intubation with automatic face analysis and artificial intelligence*.

tify relevant characteristics helping the prediction.

A. Related work

For the last 30 years, numerous definitions have been proposed and used by anesthesiologists, but no unique definition of difficult intubation exists. The vast majority of endotracheal intubations are performed using a laryngoscope which allows the visualization of the larynx and the placing of the endotracheal tube between the vocal cords, into the trachea. Cormack and Lehane proposed a classification of the laryngoscopic view using four grades based on the visibility of laryngeal structures or glottic exposure [21]. This classification was later modified by Yentis and Lee who proposed to divide the original grade 2 into grade 2a and grade 2b [22]. The later classification is used to define the difficult laryngoscopy as a view corresponding to grade 3 or grade 4. Nevertheless, it has recently been pointed out by Krage et al. that the reproducibility of this classification is limited [23]. Moreover, a poor view of the vocal cords can increase the difficulty of the intubation but other factors, such as the position of the head of the patient or the experience of the anesthesiologist also have influence on the success of the intubation. Despite the need for a standard classification of the difficult intubation in the medical community, no such uniform definition has been widely adopted. Thus, the incidence and the factors associated with difficult intubation vary from one institution to another and are virtually impossible to compare directly. The incidence of difficult laryngoscopy in the operating room has been reported to range from 0.3% to 13% [24]. In an attempt to provide a definition of the difficult intubation, Adnet et al. proposed the *Intubation Difficulty Scale (IDS)* [25], taking into account the number of attempts, the number of operators directly attempting the intubation, the use of alternative devices or techniques, the glottic exposure or the lifting force applied during laryngoscopy.

Prediction of difficult endotracheal intubation has been largely explored in the past twenty-five years by anesthesiologists. Several physical and morphological characteristics have been identified as predictors of difficult laryngoscopy or difficult intubation. Those include: obesity, poor mobility of the head and neck, poor mobility of the jaw, receding mandible, long upper incisors, decreased mouth opening (or small interincisor gap with the mouth fully open), shortened thyromental distance, short neck and small neck circumference. Several difficult intubation bedside screening tests exist.

The thyromental distance (TMD), or Patil-Aldreti test, is the distance from the upper edge of the thyroid cartilage to the chin, measured with the head fully extended. A short thyromental distance equates to an anterior lying larynx that is at a more acute angle and also results in less space for the tongue to be compressed by the laryngoscope blade. A thyromental distance greater than 7 cm is usually associated with easy intubation whereas a thyromental distance smaller than 6 cm may predict a difficult intubation. However, with a sensitivity of 48% and a specificity of 79% in predicting difficult intubation [26], this distance is not a good predictor by itself and is often used in combination with other predictors. The ratio of height to thyromental distance (RHTMD) improves the accuracy of predicting difficult laryngoscopy compared to TMD alone (sensitivity and specificity of 77% and 54% respectively) [27].

Originally described by Mallampati et al. [28] and modified by Samsoon and Young [29], the Mallampati score assesses the airway according to the visibility of oropharyngeal structures observed on a sitting patient with the mouth wide open and the tongue out. The hypothesis of the author is that the larger the base of the tongue, the more it overshadows the larynx, resulting in a poor laryngoscopic view and a potentially difficult laryngoscopy. The volume of the tongue is thus an important, yet difficult to assess, parameter when assessing the difficulty of endotracheal intubation. Since it is not possible to determine the volume of the tongue relative to the capacity of the oropharyngeal cavity, it is logical to infer that the base of tongue is disproportionately large when it is able to mask the visibility of the faucial pillars and uvula. The score ranges from class 1 to class 4, class 1 indicating full visibility of the oropharyngeal structure and class 4 none. Various meta-analysis reported different sensitivity and specificity for the Mallampati and modified Mallampati tests. In [30], the authors reported a sensitivity and a specificity of 35% and 91% respectively. In [31], the authors included 55 studies and 177088 patients and reported a sensitivity of 0% to 100% and a specificity of 44% to 100%. They computed a ROC curve and the area under the curve (AUC) was 0.753 which categorize the diagnostic test as good. In [32], the reported AUC for the Mallampati and modified Mallampati tests are respectively 0.58 and 0.83. In those studies, the authors agree that the clinical value of the Mallampati test is limited as it has poor to moderate discriminative power when used alone.

The upper lip bite test, proposed by Khan et al. [33] evaluates the ability of the patient to cover his upper lip with the lower incisors by moving forward the lower jaw (in a movement of *prognathism*). The results range from grade I to grade III where grade I and II predicts easy laryngoscopy whereas grade III predicts difficult laryngoscopy. The authors initially observed a sensitivity of 76.5% and a specificity of 88.7%. Those results were confirmed in a recent study (78.95% and 91.96% respectively) [34].

Eberhart et al. conducted a comparison between Mallampati score and upper lip bite test on 1107 patients [35] and concluded that both tests are poor predictors for difficult laryngoscopy when used as single preoperative bedside screening tests. None of those simple tests have been shown to be accurate in predicting airway management problems. Their sensitivity and predictive positive values are generally low, precluding an accurate prediction of difficult endotracheal intubation. Thus, several studies have been proposed to derive a score from multivariate analysis.

The Wilson risk sum score [36] scores five of the aforementioned factors from 0 to 2: the weight, the vertical head and neck movement, the jaw movement (prognathism), the receding mandible and buck teeth. By varying the threshold values on the sum of those scores, the true positive rate and false positive rate of difficult laryngoscopy assessment are varied. The authors initially proposed a threshold value of 4, i.e. a score greater or equal to 4 predicts a difficult

TABLE I
COMPARISON OF THREE MULTIVARIATE TESTS [38]

Model	Sens.	Spec.	PPV	NPV	AUC	Acc.
Wilson model [36]	40.2	92.8	25.6	96.2	79.0	66.5
Arné model [37]	54.6	94.9	39.7	97.1	87.0	74.7
Naguib model [24]	81.4	72.2	15.3	98.4	82.0	76.8

endotracheal intubation.

Arné et al. proposed a simplified score model [37]. In addition to the morphological criteria such as interincisor gap, ability to prognate, thyromental distance and range of head and neck movement, it also considers the medical history of the patient and the Mallampati score.

Naguib et al. performed a clinical, radiologic and 3D computer imaging study [24] on 57 patients among which 25 had an unanticipated difficult intubation. A multivariate discriminant analysis was performed on the clinical measurements and identified four risk factors that correlated with the difficult laryngoscopy and intubation: thyrosternal distance, thyromental distance, neck circumference and Mallampati classification.

Table I shows the predictive performances of those three multivariate models, as reported in [38]. The authors recruited 194 patients (97 with a difficult airway and 97 controls) over a period of 5 years. For the purpose of their study, unanticipated difficult intubation was defined as difficult laryngoscopy (corresponding to a grade 3 or 4 Cormack and Lehane laryngoscopic view) and difficult tracheal intubation (2 or more attempts at placing the endotracheal tube) or the use of an alternative device (laryngeal mask airway or bougie) when using optimal head and neck positioning (the sniffing position). Positive predictive value (PPV) and negative predictive value (NPV) were calculated based on a prevalence of difficult intubation of 5.8%, as reported in a recent meta-analysis [39]. Note that the sensitivity, specificity and AUC are the most appropriate measures to compare performances between datasets, mainly due to the class imbalance problem.

Recently, Fritscherova et al. [40] conducted a case-control study on 148 patients and concluded that the three statistically higher predictors were the interincisors distance, the thyromental distance and a decreased temporomandibular joint movement.

As none of those tests fulfill the high sensitivity and high positive predictive value criteria, anesthesiologists themselves do not agree on the usefulness of such a prediction [41], [42].

New technological approaches aimed at craniofacial phenotyping, using still photographs, x-ray technologies or laser scanning with an automated three-dimensional rendering, have been recently applied to the detection of difficult airways. Suzuki et al. calculated five ratios and angles from measurements derived from placement of anatomic markers on patients photographs [43] demonstrating that the submandibular angle seemed to be associated with difficult tracheal intubation. They also used morphing software to construct average easy and difficult to intubate faces. The improved availability of cone-beam computed tomography, 3D imaging and computer simulation has been used by Schendel and Hatcher for evaluation of the airway [44]. In the recent years, some studies took advantage of machine learning [45] or statistical face models [46] in order to provide better prediction and defend the usefulness of preoperative difficult tracheal intubation prediction. However, these newer methods require either x-ray or computed tomographic imaging methods with issues such as availability, cost and radiation dose to the patient.

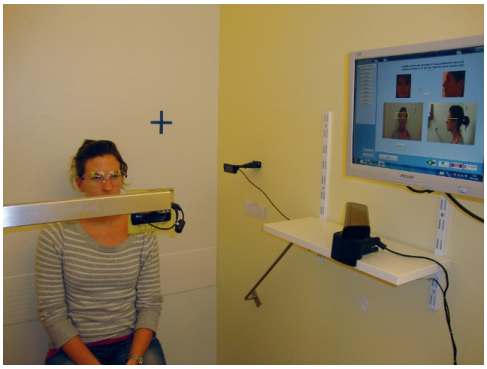


Fig. 1. Photo booth at CHUV

The method proposed in [46], even though conceptually similar to our proposed approach, presents several important differences. Their method is not fully automatic but semi-automatic as it requires manual placement of fiducial markers and manual measurement of the thyromental distance by an anesthesiologist. To limit any potential confounding effects of gender and racial group, they recruited only male Caucasians. Moreover, the definition of the difficult intubation used does not include all patients but only very easy and difficult patients such that those who were neither easy nor difficult to intubate according to their criteria were not included. This significantly reduces the variability in the data due to other factors than the difficulty of intubation and renders the resulting model inapplicable in a real-world clinical environment. This work presents biased results as the authors do not clearly separate the data into training and test sets and use the test set to select the model.

Finally, the number of patients considered to validate those newer approaches is often low. For instance, in [46] the authors reported results on a validation set of only 20 difficult and 20 easy patients thus not demonstrating the generalizability of the proposed method.

More recently, Cattano et al. proposed a new assessment form on airway prediction but showed that it did not improve resident ability to predict a difficult airway [47].

Our proposed method has been developed and validated using more than nine hundred patients. It does not require any medical history or measurement on the patient other than frontal and profile photographs, making it practical even for untrained personnel. The processing of the photographs is completely automatic and does not require any manual initialization. The processing time is of the order of the second, making the proposed method directly applicable in a clinical setting. Specifically, for one out of the four images, the face detection requires approximately 0.9s, the image alignement, the features extraction and classification run in real-time, i.e.

TABLE II
PATIENTS POPULATION METADATA

	Mean [min,max]
Age	53 [17, 92]
Height [cm]	169.5 [142,205]
Weight [Kg]	76.8 [40,160]
Gender [M/F]	488/482
Total	970

in approximately 30-40ms. In order to assess its performance in a real-world scenario, we present results including all levels of difficulty and not only very easy and difficult patients. We demonstrate that the proposed method performs as well as state-of-the-art multifactorial tests performed manually by experienced anesthesiologists.

An outline of the paper is as follows: the data collection process and setup is described in Section II. In Section III, we describe the face models training and fitting processes as well as the learning process. The results obtained are presented in Section IV and compared to diagnosis based prediction results. Finally, conclusions and a discussion of future research topics are given in Section V.

II. DATA COLLECTION

Since March 2012, adult patients at the University Hospital in Lausanne (CHUV) undergoing general anesthesia requiring tracheal intubation and related to any type of elective surgical procedures except obstetric and cardiac surgery have been preoperativelly recruited. The study has been approved by the Human Research Ethics Committee (Ethical approval number 183/09, Chairperson Prof R. Darioli) from the Ethical Committee of the Canton of Vaud, Switzerland. Each patient gets appropriate information about the research by the anesthesiologist during the preoperative consultation and gives his or her written consent to participate in the study.

A. Setup

We developed and set up a *photo booth-like* equipment (see Fig. 1) in the surgical pre-hospitalization center to collect multi-modal data on recruited patients. These data include frontal and profile photos and videos taken with two HD webcams, one in front and one on the left side of the patient at approximately 40 cm. We also record the voice of the patient and capture depth maps with a Microsoft Kinect® for future analysis.

While sitting in the photo booth, the patient is asked to perform different facial motions as well as head motions. Those include: neutral expression, opening the mouth, sticking the tongue out, lateral rotation and vertical extension of the head. A graphical user interface, developed on Matlab, allows an operator to guide the patient through the different poses he has to take and to capture the data at the appropriate moment.

We also collect patient demographics such as age, sex, weight, height and presence of denture during the preoperative anesthesia consultation. Details of peroperative airway management by the in-charge anesthetist are introduced in a dedicated database containing information on ease of face-mask ventilation, laryngoscopic grade [22] with an appropriate size MacIntosh blade, years of training of intubator (minimum of 2 years training in anesthesia is mandatory), lifting force necessary for intubation (normal or increased), usage of accessory means such as external laryngeal manipulation, intubation bougie, stylet or video-laryngoscopic equipment and injuries related to airway management. Number of airway providers and number of intubation trials are also recorded. The Intubation Difficulty Scale [25] is routinely calculated. This information allows obtaining a ground truth for the intubation difficulty.

In the two years period from March 2012 to March 2014, we have recorded 2725 patients. The ground truth is available for 970 of those (see section III-C1). Table II shows the metadata of the patients' population used in this work.

III. METHODS

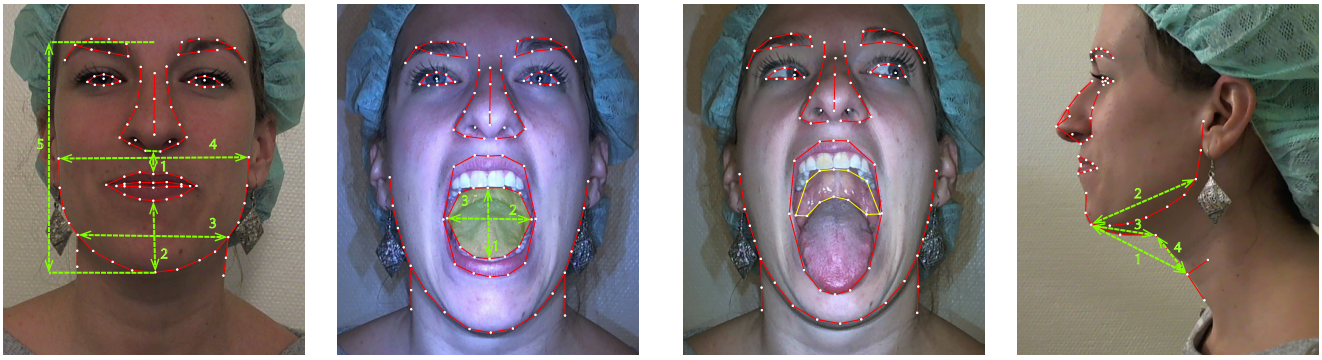
Given a set of images for each patient, we make use of face analysis methods in order to extract meaningful features from the face and neck. These features include simple distances between selected landmarks as well as information on the global shape or texture variation of the head. In a second step, the statistical relevance of those features is computed in order to discover which of them are relevant in the scope of *prediction of difficult intubation*. The most relevant features are then fed to a classifier. The classifier *learns* how to discriminate between easy, intermediate and difficult to intubate patients.

A. Detecting the face and tracking the landmarks

Facial image analysis methods often include two main parts: first we need to determine automatically the rough location of a face in the image using a *face detector*, then precise locations of each landmark are found by accurately *fitting a model* of the face on the image. Finally, features are computed using individual landmark positions as well as their global configuration.

1) Face detector: In order to initialize the fitting of the face model, both the rough location of the face in the image, as well as its scale, need to be determined.

We use Yang's Parts Based Detector [48] in order to detect the face in the images. This method is a general, flexible mixture of parts model able to capture contextual co-occurrence relations between parts, augmenting standard spring models that encode spatial relations. It has been shown to perform very well on face detection [18] and to be particularly reliable for extreme head poses. The good flexibility of the method allows us to train a single detector for all frontal images, even though the patients are performing very different facial motions, such as opening the mouth widely or sticking out the tongue. An additional detector is trained for profile images as many parts of the frontal images are not visible in the profile images. We use a manually annotated subset of our data to train both detectors. For the frontal detector, the training set consists of 406 annotated images including neutral face, mouth open and tongue out images. Both the original image and the horizontal flip of the image are used. For the profile detector, the training set consists of 134 annotated images.



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

Fig. 2. Details of the four templates, each corresponding to a facial motion: (a) frontal, neutral, 99 points (b) frontal, mouth open, 99 points (c) frontal, tongue out, 99 points (d) profile, neutral, 52 points. In green, the anatomical and morphological features described in section III-B.

The frontal face detector performs very well and detects 100% of the frontal faces in the 2910 images of the 970 patients performing all facial motions. This set includes 2553 unseen images, i.e. not used for traininig the face detector. The profile face detector, on the other hand, fails to detect the face of only 4 patients, which are removed from the final analysis, reaching a detection rate of 99.56% on unseen images.

The detection of the face provided by the face detector is then used to initialize the fitting process of the face model.

2) Face model for the image alignment problem: Finding the precise location of each pre-defined landmark in a new, unseen image is considered as an *image alignment problem*. Image alignment is the process consisting of rigidly moving and non-rigidly deforming a *template* to minimize its *distance* to a query image. Image alignment process is characterized by three elements: *template representation*, *distance metric* and *optimization scheme*.

In this work, we follow the image alignment method described in [20]. The template is non-parametric and consists of scale-invariant feature transform (SIFT) features [49] extracted from patches around each landmark. This non-parametric shape model is able to better generalize than other parameterized appearance models (PAMs) in unseen situations and this representation is robust against changes in illumination. The squared difference between the SIFT features values computed in the aligned image and in the template is used as the distance metric. This results in the following minimization problem over $\Delta\mathbf{x}$:

$$f(\mathbf{x_0} + \Delta\mathbf{x}) = \|\mathbf{h}\left(\mathbf{d}(\mathbf{x_0} + \Delta\mathbf{x})\right) - \boldsymbol{\phi}_*\|_2^2 \ , \qquad (1)$$

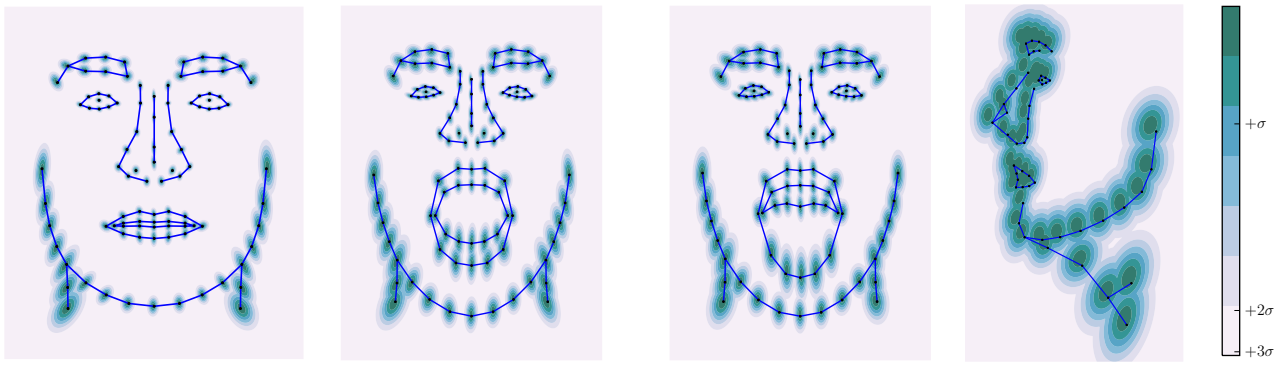
where $\mathbf{x_0}$ is the mean shape, $\Delta\mathbf{x}$ is the update of the shape, $\mathbf{d}$ is the image, $\mathbf{h}$ is a non-linear feature extraction function (in our case the SIFT features) and $\boldsymbol{\phi}_* = \mathbf{h}(\mathbf{d}(\mathbf{x}_*))$ represents the SIFT values in the manually labeled landmarks.

The *supervised descent method* (SDM) optimization scheme, thoroughly described in [20], learns a series of descent directions and re-scaling factors (done by the Hessian in the case of Newton's method) such that it produces a sequence of updates ($x_{k+1} = x_k + \Delta x_k$) starting from x_0 that converges to x_* in the training data. x_0 is the initial configuration of the landmarks provided by the face detector which corresponds to an average shape, scaled and translated, and x_* is the correct configuration of the landmarks, generally obtained by manual annotations of the images.

a) Definition of the templates: In the scope of this work, we define one template per facial motion, necessary to get accurate landmark positions on photos with different facial motions. In order to train those models, we have defined one neutral and frontal template with 99 points, two different frontal 99 points templates with large facial motions (one with the mouth open and the second with the mouth open and the tongue out) and one profile template consisting of 52 points. We then manually annotated images for each of those templates to train the face model described above.

The template corresponding to a neutral position and neutral expression contains landmarks for each eyebrow, eye, the nose, the mouth and the chin; it has 99 points in total (see Fig. 2a). It also includes points on the neck in order to assess neck characteristics, such as the width. The two templates corresponding to images with extreme facial motions (mouth open and tongue out) have the same points as the neutral 99 points template (see Fig. 2b and 2c). The landmarks defining the internal perimeter of mouth opening follow teeth or lips, depending on what is present in the image. The same set of landmarks was used for assessing the tongue out movement with a segmentation of the oral cavity, allowing grading of an automated modified Mallampati classification [50]. The segmentation of the oral cavity is shown in yellow on Fig. 2c. For profile images, a template of 52 points was defined (see Fig. 2d). The points on the jaw and the neck allow assessing jaw movement while performing mandibular movement.

b) Validation of the face model: In order to validate the face model, we use K-fold cross-validation. For each model, the images from one fold are kept for testing the model while the images from all other folds are used to train the model. The greater the number of folds, the more training images are used at each run. The obtained model is then fitted on the annotated images in the excluded fold and the obtained landmark positions are compared to the manual annotations. This procedure is repeated for each fold. This way, the model is tested on each available annotated image. Note that the face detector is first run on the images in order to initialize the face



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

Fig. 3. Distribution of the errors (differences between the landmark positions obtained automatically and the manual annotations) on each landmark for the four templates: (a) frontal, neutral, 99 points (b) frontal, mouth open, 99 points (c) frontal, tongue out, 99 points (d) profile, neutral, 52 points

model. We thus test the whole pipeline at once. In order to quantify the evolution of the error with respect to the number of training images, we run this K-fold cross-validation scheme for each model with 2, 3, 4, 5 and 10 folds. These correspond to 50%, 66.6%, 75%, 80% and 90% of the annotations used for training. The total number of annotated images is 150 for each of the frontal models and 92 images for the profile model.

Fig. 3 shows the distributions of the errors for each landmark and each model when trained and tested using 10 folds cross-validation (90% of the annotations for training). During the testing step, the error between each landmark and the corresponding annotation is computed for each test image. We then report these errors on the mean shape of each model and fit a Gaussian function for better visualization.

The quality of the model varies from one model to the other. The profile model is the least accurate (see Fig. 3d) but is also trained on fewer images. Moreover, the annotations might be less consistent from one training image to the other, due to the increased difficulty of annotating the profile face. The points on the chin and the neck from the profile model do not correspond to any salient landmarks on the images, therefore increasing the annotation difficulty as well as decreasing the face tracker ability to precisely locate those landmarks.

Fig. 4 shows the mean point-to-point error normalized by the distance between the eyes for the three frontal models. Amongst those, the two models with the mouth open and the tongue out exhibits a larger normalized point-to-point error than the neutral one. Again, the points on the chin and the neck are the less accurate (see Fig 3). It should be noted that the points around the mouth are reasonably accurate and those are also the most interesting for our application. The points around the eyes are the most accurate, thus making them good candidates for normalization. It can be seen that removing the landmarks from the chin and the neck from the mean computation improves the mean point-to-point error by 15% to 25% depending on the model. Indeed, those landmarks are significantly less accurate than the rest of the model, as discussed earlier. In the final application, all available annotated images will be used for training. Thus, the actual performances of the models will be better as they will have been trained with more annotated images.

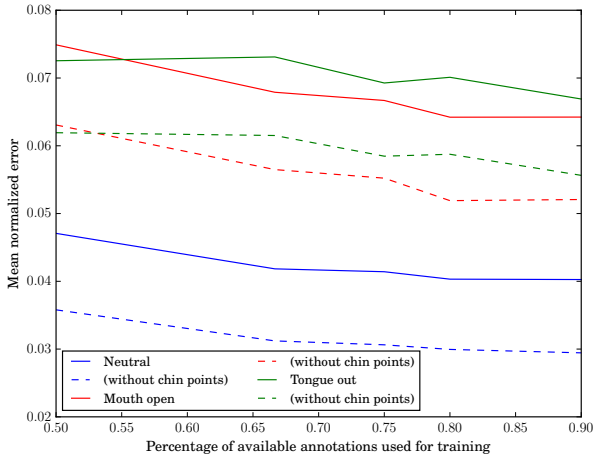


Fig. 4. Mean point-to-point error (distance between the landmark positions obtained automatically and the manual annotations) normalized by the distance between the eyes

B. Computing the features

Most of the anatomical and morphological features of interest consist of distances between landmarks on the face and the neck. The aligned template gives the positions of those landmarks after fitting the face model on the subject image. Specifically, those distances are: the vertical distance between the upper lip and the nose, the vertical distance between the lower lip and the tip of the chin, the width of the neck, the width of the face and the height of the face, all five computed on the frontal neutral image (see lines 1-5 respectively on Fig. 2a). They are the thyromental distance in neutral position, the distance between the angle of the mandible and the tip of the chin, the distance between the hyoid bone and the chin and the distance between the hyoid bone and the thyroid cartilage, all four computed on the profile neutral image (see lines 1-4 respectively on Fig. 2d). Finally they are the height of the mouth opening, the width of the mouth opening and the area of the mouth opening, all three computed from the frontal image

with the mouth open (see lines 1-2 and surface 3 respectively on Fig. 2b). In addition, we compute the distance between the eyes on all frontal images. This distance is used to normalize the features listed above allowing us to be more robust against moderate head pose variations, and to be able to compare them between patients. Indeed, the fact that all patients do not sit at the exact same distance to the camera and do not have the same head pose introduces an important bias in the features. After normalization, all distances are divided by the distance between the eyes. This one exhibits small variations between subjects, is most likely not correlated with difficult intubation and can be computed reliably from the landmarks around the eyes as they are very accurate.

In addition to the distances between landmarks, we also consider coefficients from a PCA on the shape and coefficients from a PCA on the texture for the inside of the mouth on the frontal, tongue out, model as features. Specifically we compute those coefficients in the following manner:

To compute the PCA-coefficients on the shape, we consider the set of face images used for training, each image having a set of v 2D landmarks returned by the face tracker, $[x_i, y_i], i = 1, 2, ..., v$. The collection of landmarks of one image is treated as one observation from the random process defined by the shape model $\mathbf{s} = [x_1, y_1, x_2, y_2, ..., x_v, y_v]^T$. Eigenanalysis is applied to the observation set, keeping 98% of energy, and the resultant model represents a shape as

$$\mathbf{s}(\mathbf{p}) = \mathbf{s}_0 + \sum_{i=0}^{n} p_i \mathbf{s}_i \ , \tag{2}$$

where $\mathbf{s}_0$ is the mean shape, $\mathbf{s}_i$ is the i^{th} shape basis and $\mathbf{p} = [p_1, p_2, ..., p_n]$ are the shape parameters.

Those parameters $\mathbf{p}$ provide information on the global variation of the shape. They are ranked by the value of their corresponding eigenvalue, in a decreasing order, or, similarly by the amount of total variance of the training data that they explain. The first modes of variation explain the bigger amount of total variance and are thus likely to explain the variance of the data due to head pose, gender or other factors that are not significant in the prediction of the difficult intubation. On the other hand, the last ones only explain a small amount of the total variance and merely model the effect of noise in the annotations. Even though not all coefficients are relevant for classification, each of them has the advantage of encoding a variation mode affecting the relative configuration of several landmarks by itself. Thus, by selecting a few, relevant coefficients, we can potentially get information about global configurations of landmarks (or global morphology of the face) correlated with difficult intubation.

To compute the PCA-coefficients on the texture, we first compute a piecewise affine transform between the landmarks segmenting the oral cavity on each image (see the yellow contour on Fig. 2c) and the same landmarks on the mean shape. The texture inside those landmarks is then warped onto the mean shape and normalized to zero mean and unit standard deviation. At training time, the warped and normalized texture from the images in the training set are used to compute a PCA basis. Similarly to the PCA on the shape, the eigenvectors corresponding to the biggest ordered eigenvalues and explaining 75% of the texture variance are kept while the others are discarded. At testing time, the warped and normalized texture from the images in the test set is then projected on that basis, resulting in a vector of coefficients used as features. For more details, the reader is referred to [50] where the same method is used for automatic Mallampati classification.

Section III-C3 provides more details about the *feature selection* techniques that have been used to find those relevant coefficients.

C. Classification

1) Class definitions: In order to train and test the system each patient is assigned one of the following labels, considered as ground truth and related to their difficulty of intubation: *easy*, *intermediate* or *difficult*. As no precise definition of the *difficult intubation* has been unanimously accepted, this classification is obtained by combining two complementary definitions, namely the widely accepted definition of the *difficult laryngoscopy*, which considers a laryngoscopy as difficult if the Cormack-Lehane view of the larynx is graded III or IV [21] and the definition based on the *intubation difficulty score (IDS)* proposed by Adnet [25], which considers an intubation as difficult if the IDS is greater than 5. We use this broader definition of the difficult intubation in order to remove, as much as possible, the subjectivity of using only the laryngoscopic grade, while still assigning laryngoscopic grades III and IV to the difficult class. More specifically, the class labels are defined as follows:

- *easy* $IDS = 0$, this implies a laryngoscopic grade of I and a successful intubation at the first attempt;
- *intermediate* $0 < IDS \leq 5$ and laryngoscopic grade smaller than III;
- *difficult* $IDS > 5$, or laryngoscopic grade of III or IV.

Out of the 2725 patients who have been recorded, information allowing to compute the IDS is available for 34.4% and laryngoscopic grade for 51.4% at the time of writing this article.

Table IIIa shows the distribution of patients according to the laryngoscopic view for all recorded patients and for the subset of patients with available groundtruth and face detection. The laryngoscopic view was observed by the anesthesiologist at the intubation time. It should be noted that the classes are largely unbalanced, higher laryngoscopic grades being rarely observed which makes the classification task more challenging. Table IIIb shows the classification of the recruited patients according to their IDS score. The same remark applies regarding high IDS scores.

Table IIIc shows the distribution of each class according to the classification described above for the 966 patients used in total. The *easy*, *intermediate* and *difficult* labels are used as ground truth. Note that this does not directly correspond to the IDS because 8 patients with $IDS \leq 5$ have a laryngoscopic grade greater than II and are labelled as *difficult* and 29 other patients with a laryngoscopic grade greater than II have missing IDS score.

TABLE III
(A) PATIENTS LARYNGOSCOPIC GRADE (LG) DISTRIBUTION AS OBSERVED BY THE ANESTHESIOLOGIST AT INTUBATION TIME (B) PATIENTS IDS SCORE DISTRIBUTION (C) FINAL GROUND TRUTH LABELS DISTRIBUTION.

	recorded patients		966 used patients	
LG		[%]		[%]
1	1083	77.30	708	73.29
2a	208	14.85	158	16.36
2b	57	4.07	47	4.86
3	40	2.85	40	4.14
4	13	0.93	13	1.35

(a)

IDS score	Difficulty		[%]
0	Easy	561	59.87
$0 < \text{IDS} \leq 5$	Slight Difficulty	353	37.67
$5 < \text{IDS}$	Moderate to Major	23	2.46

(b)

Difficulty		[%]
Easy	561	58.07
Intermediate	345	35.72
Difficult	60	6.21

(c)

2) Data partition for training and testing and class imbalance problem: The feature selection, the choice of the hyperparameters and the training of the classifier are performed on a subset of patients: the *training* set. A distinct subset of patients is then used to test the classifier and compute the different metrics assessing its performance: the *test* set. The partition of the original data into those two subsets is random but the original distribution of classes is maintained (*stratified* partitioning). In order to compute proper statistics for the results, those training and test sets are generated several times, each time with different random partitions of the patients.

Note that both the training and the test set follow the same class distribution as the original dataset. As previously discussed, the occurrence of difficult laryngoscopy has been reported to range from 0.3% to 13% [24]. More recently the occurrence of difficult intubation has been reported between 4.5% and 7.5% in the overall population [39]. In the present dataset, 6.21% of the patients fall in the *difficult* class. From a machine learning point of view, skewed distributions of classes make the learning of concepts more difficult. This is known as the *class imbalance problem*. Even a relatively small imbalance ratio of the order of 10:1, as in our case, is sufficient to hinder the learning process.

Artificially balancing the classes is possible using *sampling methods*. However, those methods present some significant drawbacks [51]–[53]. Undersampling from the majority class(es) allows reducing the imbalance ratio or even totally compensating for the class imbalance. But removing samples from class(es) may result in loss of information, thus potentially penalizing the classifier's performance. In the other case, oversampling from the minority class(es) allow for the same reduction of class imbalance but presents a different drawback. Replicating samples tends to lead to overfitting. Even though more complex techniques exist, several problems prevent from finding a good approximation of the original class density function, for example small disjuncts or class overlapping.

In this work, we consider binary classifiers. To oversome the class imbalance problem, we use the fact that for each sample, *probabilistic classifiers* compute confidence values of belonging to each class. The classifier then usually assigns the most probable label to each sample by maximizing $P(j|x)$, the posterior probability of classifying a sample x as j. Nevertheless, in cost-sensitive learning, given a cost matrix defined as $C(i, j)$ the misclassification cost of classifying an instance from its actual class j into the predicted class i, the minimum expected loss can be determined as:

$$\mathcal{R}(i|x) = \sum_{j \in \{0,1\}} P(j|x) \cdot C(i, j) \ , \tag{3}$$

where $\mathcal{R}$ is the Bayes risk and $P(j|x)$ is the posterior probability. Elkan [54] showed that modifying the classifier's threshold, that is choosing the positive class if its confidence value is greater than a threshold but not necessarily greater than the confidence value of the other class, has the same effect as sampling in terms of bias but without the drawbacks mentioned above. Thus, defining a threshold θ for the classifier allows compensating for the bias towards the majority class. Specifically, in cost-sensitive learning the optimal threshold θ^* of a classifier with respect to a given cost matrix is defined as:

$$\theta^* = \frac{C(1, 0)}{C(1, 0) + C(0, 1)} \ , \tag{4}$$

In binary classification, $C(1, 0)$ represents false positive (FP) and $C(0, 1)$ represents false negative (FN). The prior probabilities of the negative and positive samples ($p(0)$ and $p(1)$ respectively) are proportional to the number of samples in the original training set. As doubling FN or halving FP has the same effect as doubling $p(1)$, we train the classifier on the complete (unbalanced) training set and when testing it on the test set, the threshold θ is set to the imbalance ratio between the classes:

$$\theta = \frac{FP}{FP + FN \cdot \frac{p(0)}{p(1)}} = \frac{1}{1 + \frac{p(0)}{p(1)}} \approx \frac{p(1)}{p(0)} \ , \tag{5}$$

where $\frac{p(0)}{p(1)}$ is bigger than 1 as the positive class, with the label *difficult*, is the class for which we have less samples.

As modifying the threshold of the classifier is equivalent to sampling, we compare three methods of choosing this threshold:

- the class imbalance ratio method as described above (see Eq. (5)).
- minimizing the distance between the corresponding point on the ROC curve and the (0,1) point (upper left corner)
- maximizing the Youden index, i.e. the vertical distance between the corresponding point on the ROC curve and the line of no-discrimination.

The latter two methods use four fold cross-validation on the training set to learn the optimal threshold. In order not to hinder the learning process when training the classifier on an unbalanced set, we use the area under the ROC curve (AUC)

TABLE IV
COMPARISON OF OUR RESULTS ON THE EASY VS DIFFICULT PROBLEM WITH THREE MULTIVARIATE TESTS [38] AND A SEMI-AUTOMATIC METHOD [46]

Model	Sens. [95% CI]	Spec. [95% CI]	AUC
Wilson model [36]	40.2 [30.0, 50.0]	92.8 [88.0, 98.0]	79.0
Arné model [37]	54.6 [45.0, 65.0]	94.9 [90.0, 99.0]	87.0
Naguib model [24]	81.4 [74.0, 89.0]	72.2 [63.0, 81.0]	82.0
Connor [46]	90.0	80.0	84.0
Ours			81.0
class imbalance	79.7 [77.4, 81.9]	67.4 [66.4, 68.4]	
distance to (0,1)	77.1 [74.8, 79.4]	70.6 [69.4, 71.8]	
Youden index	78.9 [76.5, 81.3]	66.7 [64.7, 68.6]	

as criterion. The ROC curve is generated by plotting the false positive rate (FPR) against the true positive rate (TPR) for all values of the classifier threshold. Independently of what kind of classifier is used, we train it such that the ROC curve generated from the output confidence values maximizes the AUC, since AUC is insensitive to the class imbalance. As a post-processing step, we then compute the threshold to apply on the confidence values in order to obtain the final classification of each sample.

3) Feature selection and classification: Feature selection is performed on the training set. The goal is to determine which features are the most relevant for difficult intubation prediction. Amongst the complete set of features, only those most relevant features are then used to train the classifier. Reducing the dimensionality of the data, as well as removing noisy, irrelevant features from the data helps improving the classification performance.

Random Forest classifiers provide a feature importance measure which allows for feature ranking and selection [55]. The feature importance is measured by randomly permuting the feature in the out-of-bag samples and calculating the percent increase in misclassification rate as compared to the out-of-bag rate with all variables intact. From the ranking of the features according to their importance, we only keep the k best and discard all the rest. The parameter k is considered a hyper-parameter and its best value is found using grid-search and K-fold cross-validation on the training set at the same time as the classifier hyper-parameters.

For the final classification, a second Random Forest classifier is used. Random Forest classifiers are known to be less prown to overfitting, due to their use of bagging. Indeed, the training algorithm for Random Forest aims at constructing a forest of trees where for each tree it randomly samples with replacement in the training set and trains the tree by considering only a random subset of the features at each splitting node. The hyper-parameters of the classifier are selected using four fold cross-validation on the training set. Specifically, those hyper-parameters are the following: the number of the k best features to keep (in the range 20-180 by step of 10) and the percentage of features to consider at each node when looking for the best split (in the range $0.5\sqrt{N}$ - $2\sqrt{N}$, where N is the total number of features). We use *entropy* as the splitting criterion, as it is less sensitive to class imbalance than the usual accuracy [52]. Our implementation uses Scikit-learn [56], a python machine learning library.

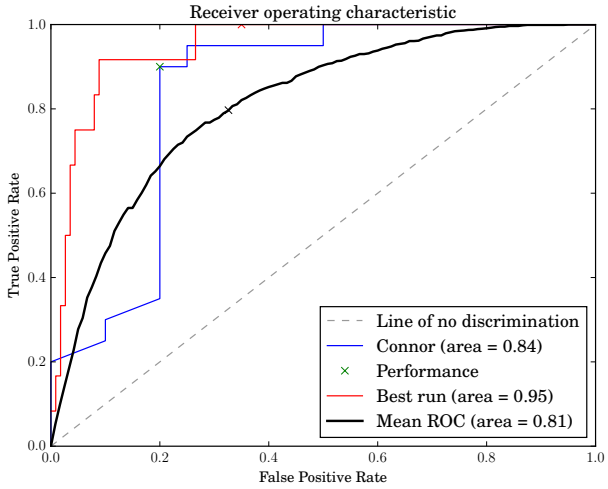


Fig. 5. Mean ROC curve for the easy *vs* difficult classification and the best ROC curve out of the 100 runs, with performance obtained using the class imbalance threshold method compared to the ROC curve obtained on the validation set in [46]

IV. RESULTS

We present two scenarios: an *easy* versus *difficult* classification considering easy control patients and difficult ones, as well as a more realistic difficult intubation prediction scenario where all patients are considered. The second one would correspond to a real-world scenario where each and every incoming patient gets a prediction.

A. Easy vs difficult classification

In this scenario, we followed the same protocol as Naguib did in his comparative study of three multi-variate difficult tracheal intubation models [38] in which for each *difficult* patient, an *easy* one (*control* patient) is selected. In our case, we do not enforce a one to one correspondence, but keep the imbalance between the classes. Removing the *intermediate* patients, we end up with two disjoint classes: the *easy* and the *difficult* patients.

We use 80% of the patients for training and 20% for testing. The partition is repeated 100 times randomly and the results are averaged over those different partitions. This results in 496 training patients (448 *easy* and 48 *difficult*) and 125 test patients (113 *easy* and 12 *difficult*).

The performances of the classifier are reported in table IV, along with the results reported in the literature for three manual tests [38] and a previous attempt for semi-automatic difficult intubation prediction from [46]. We report the mean values of the sensitivity and specificity with their 95% confidence interval (CI).

As can be seen in table IV our fully automatic system achieves comparable performance on the easy vs difficult intubation classification as compared to manual assessment using state-of-the-art multifactorial tests. In this binary example, the only metric that can be compared directly is the area under the ROC curve (AUC). All other metrics reported can be tuned by varying the threshold of the classifier, depending on the importance given to sensitivity or specificity. This can be

TABLE V
COMPARISON OF OUR RESULTS ON THE REAL-WORLD PROBLEM

Model	Sens. [95% CI]	Spec. [95% CI]	AUC
Real-world			77.9
class imbalance	77.7 [75.7, 79.7]	64.1 [63.2, 65.0]	
distance to (0,1)	72.9 [70.3, 75.5]	68.4 [67.2, 69.5]	
Youden index	74.8 [72.0, 77.5]	65.5 [63.5, 67.4]	

seen by comparing the three methods to compute an optimal threshold. The class imbalance method provides the higher sensitivity, which in this application is an important metric as it is critical to detect as many difficult intubations as possible, even at the cost of more false positive.

Fig. 5 presents the averaged ROC curve over the 100 partitions. In red, it also shows the ROC curve corresponding to the best run out of 100, i.e. the one with the highest AUC. In blue, we regenerated the ROC curve corresponding to the validation set in [46]. We used the values of each samples in the validation set provided in [46] to compute TPR and FPR for all thresholds. The highlighted performance points on the best run and mean ROC curve have been obtained by setting the threshold of the classifier to the class imbalance ratio. On the mean ROC curve, it corresponds to the results reported in Table IV whereas the sensitivity and specificity for the best run are respectively 1.0 and 0.65.

As for comparison with the results reported in [46], we would like to emphasize that such a comparison would not be a fair one, as already mentioned in I-A. Firstly, the authors trained and tested their system only on male caucasian patients, while we report our results on a much more representative population (see Table II). Secondly, the authors in [46] state that they perform model selection such that they get the best product of AUCs on the training and validation sets. In addition, they do not perform any kind of cross-validation and demonstrate results on a single partitioning. Methodologically, there is no evidence in their work that similar results would be obtained on an independent test set or a different partitioning of the data. In this work, on the other hand, we present our results on multiple runs, each of them on randomly created independent test sets. Although in average our AUC score (0.81) is lower than the AUC calculated on the validation set in [46] (0.84), our results are better validated in a more generalized way. Fig. 5 shows indeed that our fully-automatic system can achieve better results on the best run compared to the semi-automatic system in [46]. This shows that a single run is not representative of the merit of a system.

B. Real-world difficult intubation prediction

In the real-world *difficult intubation prediction* problem, the goal is to identify *difficult* to intubate patients from all the others. Considering this task the problem remains a two-class classification problem. Thus we first group together the *easy* and *intermediate* classes and relabel the new class as *easy*, which *de facto* represents the non-difficult to intubate patients. When a patient is diagnosed as *difficult*, it sends a strong signal to the anesthesiologists on the potential difficulty of that patient, which is high. Thus, we do not consider only very easy patients as control patients versus difficult ones, but instead we take into account all patients, ranging from very easy to impossible to intubate without gap.

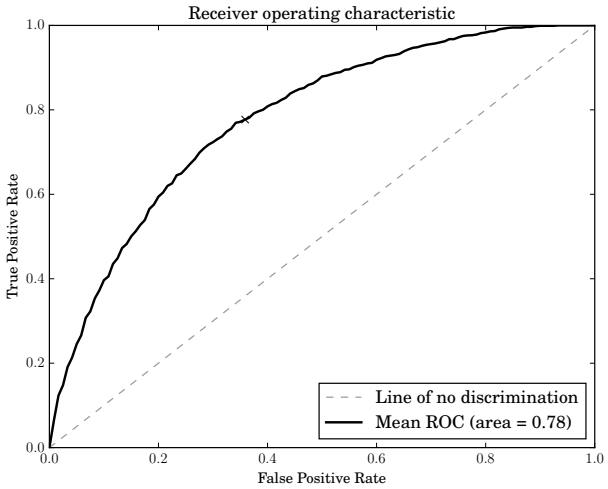


Fig. 6. Mean ROC curve for the real-world difficult intubation prediction

We use 80% of the patients for training and 20% for testing. The partitioning is repeated 100 times randomly and the results are averaged over those different partitions. This results in 772 training patients (724 *easy* and 48 *difficult*) and 194 test patients (182 *easy* and 12 *difficult*). Note that in this case, the class imbalance is more severe, creating an additional challenge to the fact that there is more variation among the samples as compared to the previous scenario. The performances of the classifier are reported in table V. Fig. 6 presents the averaged ROC curve over the 100 partitions.

As can be seen in table V, the performances of the system drop slightly when considering all patients, without gap between the classes. We observe a 3.1% decrease on the AUC and between -1.2% and -4.2% on the sensitivity and specificity. By considering all patients, the variance of the data is larger. Thus the learning of concepts is hindered as this larger variance can be seen as noise. Moreover, the absence of gap between the classes potentially decreases the class separability, again hindering the learning of concepts. Indeed, the classes become less distinct and when testing on a different dataset than that used for training, the chances are higher that the classes overlap. Note that the definition of the ground truth also has an importance in the performance of the system. More specifically, the subjectivity and poor reproducibility of the Cormack-Lehane grade make the ground truth label less reliable.

V. CONCLUSION

In this work, we presented a completely automatic, morphology based method allowing predicting a patient's difficulty of intubation with performances comparable to state-of-the-art medical diagnosis based predictions by experienced doctors. Our method has been validated on more than nine hundred patients, both in a research oriented scenario with only easy and difficult patients and in a real-world oriented scenario where all patients are considered.

The database used in this work is, to the best of our knowledge, the largest database of images, videos and ground truth data related to endotracheal intubation.

The open question of how to quantify a difficult intubation remains a penalizing factor for our results. Indeed, the recognized subjectivity as well as the large variability of the factors taken into account in order to quantify the difficulty of intubation of a patient creates an additional confound for the system. This raises the question of the direct clinical usefulness of such an automatic tool. Yet we demonstrate that it can achieve close to human performance even with such existing limitations. It is thus encouraging to further investigate the usage of facial image analysis in the scope of difficult endotracheal intubation prediction.

Due to the rarity of patients difficult to intubate, obtaining a reasonable number of them is a long term procedure. Thus, current and future development include the collection of more data. Another future research axis is to use other modalities that may be indicative of intubation difficulty. For this purpose, we also record the voice of the patient and the depth of the mouth cavity using the kinect. Further analysis of the data will include the use of these two modalities.

REFERENCES

[1] G. N. Peterson *et al.*, "Management of the difficult airway: A closed claims analysis," *Anesthesiology*, vol. 103, no. 1, pp. 33–39, 2005.

[2] T. M. Cook and S. R. Macdougall-Davis, "Complications and failure of airway management," *British J. of anaesthesia*, vol. 109, pp. i68–i85, 2012.

[3] J. Metzner *et al.*, "Closed claims' analysis," *Best Practice and Research: Clinical Anaesthesiology*, vol. 25, no. 2, pp. 263–276, 2011.

[4] A. M. B. Heard, R. J. Green, and P. Eakins, "The formulation and introduction of a 'can't intubate, can't ventilate' algorithm into clinical practice," *Anaesthesia*, vol. 64, no. 6, pp. 601–608, 2009.

[5] L. D. Hove *et al.*, "Analysis of deaths related to anesthesia in the period 1996-2004 from closed claims registered by the danish patient insurance association," *Anesthesiology*, vol. 106, no. 4, pp. 675–680, 2007.

[6] M. F. Aziz *et al.*, "Routine clinical practice effectiveness of the glidescope in difficult airway management: An analysis of 2,004 glidescope intubations, complications, and failures from two institutions," *Anesthesiology*, vol. 114, no. 1, pp. 34–41, 2011.

[7] W. H. L. Teoh *et al.*, "Comparison of three videolaryngoscopes: Pentax airway scope, c-mac™, glidescope® vs the macintosh laryngoscope for tracheal intubation," *Anaesthesia*, vol. 65, no. 11, pp. 1126–1132, 2010.

[8] G. Serocki *et al.*, "Management of the predicted difficult airway: A comparison of conventional blade laryngoscopy with video-assisted blade laryngoscopy and the glidescope," *European J. of anaesthesiology*, vol. 27, no. 1, pp. 24–30, 2010.

[9] E. J. Juan, J. P. Mansfield, and G. R. Wodicka, "Miniature acoustic guidance system for endotracheal tubes," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 6, pp. 584–596, 2002.

[10] J. O. Räsänen, G. Rosenhouse, and N. Gavriely, "Effects of diameter, length, and circuit pressure on sound conductance through endotracheal tubes," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 7, pp. 1255–1264, 2006.

[11] H.-I. A. and R. W. Picard, "Measuring affective-cognitive experience and predicting market success," *IEEE Trans. Affective Computing*, 2014.

[12] F. Ringeval *et al.*, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Lett.*, 2014.

[13] Y. Dong *et al.*, "Driver inattention monitoring system for intelligent vehicles: A review," in *Proc. IEEE Intelligent Vehicles Symp.*, 2009, pp. 875–880.

[14] H. Gao, A. Yuce, and J.-P. Thiran, "Detecting emotional stress from facial expressions for driving safety," in *Proc. Int. Conf. on Image Process.*, 2014.

[15] G. Baynam *et al.*, "Intersections of epigenetics, twinning and developmental asymmetries: Insights into monogenic and complex diseases and a role for 3d facial analysis," *Twin Research and Human Genetics*, vol. 14, no. 4, pp. 305–315, 2011.

[16] P. Claes *et al.*, "Dysmorphometrics: The modelling of morphological abnormalities," *Theoretical Biology and Medical Modelling*, vol. 9, no. 1, 2012.

[17] Q. Zhao *et al.*, "Automated down syndrome detection using facial photographs," in *Proc. Annu. Int. Conf. IEEE Eng. Medicine and Biology Soc.*, 2013.

[18] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Comput. Soc. Conf. on Comput. Vision and Pattern Recognition*, 2012, pp. 2879–2886.

[19] H. Cevikalp, B. Triggs, and V. Franc, "Face and landmark detection by using cascade of classifiers," in *10nth IEEE Int. Conf. Automat. Face and Gesture Recognition*, 2013.

[20] X. Xiong and F. De La Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Comput. Soc. Conf. on Comput. Vision and Pattern Recognition*, 2013, pp. 532–539.

[21] R. S. Cormack and J. Lehane, "Difficult tracheal intubation in obstetrics," *Anaesthesia*, vol. 39, no. 11, pp. 1105–1111, 1984.

[22] S. M. Yentis and D. J. H. Lee, "Evaluation of an improved scoring system for the grading of direct laryngoscopy," *Anaesthesia*, vol. 53, no. 11, pp. 1041–1044, 1998.

[23] R. Krage *et al.*, "Cormack-lehane classification revisited," *British J. of anaesthesia*, vol. 105, no. 2, pp. 220–227, 2010.

[24] M. Naguib *et al.*, "Predictive models for difficult laryngoscopy and intubation. a clinical, radiologic and three-dimensional computer imaging study," *Canadian J. of Anaesthesia*, vol. 46, no. 8, pp. 748–759, 1999.

[25] F. Adnet *et al.*, "The intubation difficulty scale (ids): Proposal and evaluation of a new score characterizing the complexity of endotracheal intubation," *Anesthesiology*, vol. 87, no. 6, pp. 1290–1297, 1997.

[26] P. A. Baker, A. Depuydt, and J. M. D. Thompson, "Thyromental distance measurement - fingers don't rule," *Anaesthesia*, vol. 64, no. 8, pp. 878–882, 2009.

[27] B. Krobbuaban *et al.*, "The predictive value of the height ratio and thyromental distance: Four predictive tests for difficult laryngoscopy," *Anesthesia and Analgesia*, vol. 101, no. 5, pp. 1542–1545, 2005.

[28] S. R. Mallampati *et al.*, "A clinical sign to predict difficult tracheal intubation; a prospective study," *Canadian Anaesthetists' Soc. J.*, vol. 32, no. 4, pp. 429–434, 1985.

[29] G. L. T. Samsoon and J. R. B. Young, "Difficult tracheal intubation: A retrospective study," *Anaesthesia*, vol. 42, no. 5, pp. 487–490, 1987.

[30] D. Cattano *et al.*, "Risk factors assessment of the difficult airway: An italian survey of 1956 patients," *Anesthesia and Analgesia*, vol. 99, no. 6, pp. 1774–1779, 2004.

[31] L. H. Lundstrøm *et al.*, "Poor prognostic value of the modified mallampati score: A meta-analysis involving 177 088 patients," *British J. of anaesthesia*, vol. 107, no. 5, pp. 659–667, 2011.

[32] A. Lee *et al.*, "A systematic review (meta-analysis) of the accuracy of the mallampati tests to predict the difficult airway," *Anesthesia and Analgesia*, vol. 102, no. 6, pp. 1867–1878, 2006.

[33] Z. H. Khan, A. Kashfi, and E. Ebrahimkhani, "A comparison of the upper lip bite test (a simple new technique) with modified mallampati classification in predicting difficulty in endotracheal intubation: A prospective blinded study," *Anesthesia and Analgesia*, vol. 96, no. 2, pp. 595–599, 2003.

[34] Z. H. Khan *et al.*, "The diagnostic value of the upper lip bite test combined with sternomental distance, thyromental distance, and interincisor distance for prediction of easy laryngoscopy and intubation: A prospective study," *Anesthesia and Analgesia*, vol. 109, no. 3, pp. 822–824, 2009.

[35] L. H. J. Eberhart *et al.*, "The reliability and validity of the upper lip bite test compared with the mallampati classification to predict difficult laryngoscopy: An external prospective evaluation," *Anesthesia and Analgesia*, vol. 101, no. 1, pp. 284–289, 2005.

[36] M. E. Wilson *et al.*, "Predicting difficult intubation," *British J. of anaesthesia*, vol. 61, no. 2, pp. 211–216, 1988.

[37] J. Arné *et al.*, "Preoperative assessment for difficult intubation in general and ent surgery: Predictive value of a clinical multivariate risk index," *British J. of anaesthesia*, vol. 80, no. 2, pp. 140–146, 1998.

[38] M. Naguib *et al.*, "Predictive performance of three multivariate difficult tracheal intubation models: A double-blind, case-controlled study," *Anesthesia and Analgesia*, vol. 102, no. 3, pp. 818–824, 2006.

[39] T. Shiga *et al.*, "Predicting difficult intubation in apparently normal patients: A meta-analysis of bedside screening test performance," *Anesthesiology*, vol. 103, no. 2, pp. 429–437, 2005.

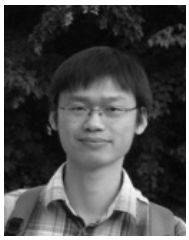
[40] S. Fritscherova *et al.*, "Can difficult intubation be easily and rapidly predicted?" *Biomedical Papers*, vol. 155, no. 2, pp. 165–172, 2011.
[41] S. M. Yentis, "Predicting difficult intubation - worthwhile exercise or pointless ritual?" *Anaesthesia*, vol. 57, no. 2, pp. 105–109, 2002.
[42] É. Orozco-Díaz *et al.*, "Predictive factors of difficult airway with known assessment scales," *Cirugia y cirujanos*, vol. 78, no. 5, pp. 393–399, 2010.
[43] N. Suzuki *et al.*, "Submandible angle in nonobese patients with difficult tracheal intubation," *Anesthesiology*, vol. 106, no. 5, pp. 916–923, 2007.
[44] S. A. Schendel and D. Hatcher, "Automated 3-dimensional airway analysis from cone-beam computed tomography data," *J. of Oral and Maxillofacial Surgery*, vol. 68, no. 3, pp. 696–701, 2010.
[45] O. Langeron *et al.*, "Prediction of difficult tracheal intubation: Time for a paradigm change," *Anesthesiology*, vol. 117, no. 6, pp. 1223–1233, 2012.
[46] C. W. Connor and S. Segal, "Accurate classification of difficult intubation by computerized facial analysis," *Anesthesia and Analgesia*, vol. 112, no. 1, pp. 84–93, 2011.
[47] D. Cattano *et al.*, "Anticipation of the difficult airway: Preoperative airway assessment, an educational and quality improvement tool," *British J. of anaesthesia*, vol. 111, no. 2, pp. 276–285, 2013.
[48] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. IEEE Comput. Soc. Conf. on Comput. Vision and Pattern Recognition*, 2011, pp. 1385–1392.
[49] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. of Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[50] G. L. Cuendet *et al.*, "Automatic Mallampati Classification Using Active Appearance Models," in *Proc. of Int. Workshop on Pattern Recognition for Healthcare Analytics*, 2012.
[51] V. López *et al.*, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inform. Sci.*, vol. 250, pp. 113–141, 2013.
[52] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.
[53] M. Galar *et al.*, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern. C*, vol. 42, no. 4, pp. 463–484, 2012.
[54] C. Elkan, "The foundations of cost-sensitive learning," in *IJCAI Int. Joint Conf. on Artificial Intelligence*, 2001, pp. 973–978.
[55] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[56] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *J. of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

Gabriel L. Cuendet received his B.Sc. and M.Sc. degrees in electrical engineering with specialization in biomedical engineering from the Ecole Polytechnique Fédérale de Lausanne, Switzerland, in 2012, where he is currently working toward the Ph.D. degree in developing facial image analysis for medical diagnosis applications. The research is focused on computer vision methods for 2D and 3D facial landmarks detection and tracking.

Patrick Schoettker studied medicine in Lausanne, Switzerland, specialized in Anesthesia and Emergency Medicine in 2001 while pursuing his clinical activity in Switzerland and Australia. He is responsible for the difficult airway curriculum in the University Hospital Lausanne CHUV where he is the head of the Neuro, ENT and Trauma Anesthesia. Intense research activity on the difficult airway subject.

Anıl Yüce received his B.Sc. from Middle East Technical University, Turkey in 2008 and M.Sc. from Ecole Polytechnique Fédérale de Lausanne, Switzerland (EPFL) in 2010 in electrical engineering. Since then he is pursuing a PhD degree at the Signal Processing Laboratory (LTS5) at EPFL. His main research interest is facial image analysis for various applications, particularly analysis of facial expressions and their dynamics. He is a student member of the IEEE since 2011.

Matteo Sorci received his B.Sc. and M.Sc. degrees from the Faculty of Telecommunication Engineering, University of Siena, Italy in 2001, and the doctoral degree in 2009 from EPFL (Swiss Federal Institute of Technology), in the Signal Processing Laboratory under the supervision of Prof. Jean-Philippe Thiran. His main research interests are behavioural modelling, dimensionality reduction, machine learning and computer vision. Matteo is currently CTO and co-founder at nViso SA.



Hua Gao received the Dipl.-Inf. and Ph.D. degrees in computer science from the Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2008 and 2013, respectively.

He is currently a post-doc at EPFL, Lausanne, Switzerland. His research interests include the fields in facial image processing, e.g. face tracking, 3D face reconstruction, facial expression recognition and face recognition.

Christophe Perruchoud received his medical degree from the University of Lausanne in Switzerland. He then trained in anesthesiology and pain management and is currently medical chief officer at the Hospital of Morges (EHC) and consultant at the University Hospital of Lausanne (CHUV). His main research topics include evaluation and management of difficult airways in the perioperative period.

Jean-Philippe Thiran is Associate Professor of Image Processing and director of the Signal Processing Laboratory (LTS5) at the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. He also holds an Associate Professor position with the Department of Radiology of the University Hospital Center (CHUV) and University of Lausanne (UNIL). His research field is image analysis and multimodal signal/image processing, with applications in many domains including medical image analysis, human-computer interaction, remote sensing of the Earth, and surveillance. Dr Thiran is author of co-author of more than 130 journal papers, 9 book chapters, more than 185 papers in peer-reviewed proceedings of international conferences, and holds 4 international patents. He is currently an associate editor of the IEEE Transactions on Image Processing and a reviewer for many journals and conferences. He is a senior member of the IEEE.