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

A Method for Identifying Ground Truth Labels in Regression Problems using Annotator Precision

Benjamin Johnston *Student Member, IEEE* and Philip de Chazal *Senior Member, IEEE*

Abstract— **We propose a novel method for deriving ground truth labels for regression problems that considers the precision of annotators separately for each label. This method ensures that higher performing annotators contribute more to the final landmark position which is in contrast to conventional methods that assume all annotators are equally accurate in completing the set task. In addition to describing the novel method, a set of preliminary experimental results is also provided, comparing the performance of the precision method to that of the global mean.**

Index Terms— **annotation, ground truth, machine learning, facial landmarking, variability, OSA, PAP**

I. INTRODUCTION

There is a commonly known phrase in data analytics and machine learning communities that effectively summarises the need for high quality ground truth samples when training high performant machine learning models: *"rubbish in, rubbish out"*. If the quality of the training data is low one would expect a similar quality of output from the model itself. Conversely, with high quality input data, one can construct simple, yet high performing models that can have provide practical benefit in the physical world. Given the critical nature of using high quality training data one must consider what is required to procure such a dataset. While the specifics of the collection process may vary with the application at hand there is a general process that can be followed. At the preliminary stages one must attempt to understand the problem being solved and the relationship between the desired outcomes and input data as much as possible. This understanding can help to define the data collection process as well as aspects of the dataset such as the source and distribution of labels within the set. In most situations the number of labels for each category will be representative of the population being modelled, while in others under or oversampling of specific labels may be required. One must also provide due consideration to the method by which the labels of the dataset will be generated, as this process is almost exclusively completed manually we must ask: Will each sample be annotated multiple times by multiple annotators? Is special training required to complete the annotation task? Do annotators require some minimum qualifications to complete the task e.g. qualified medical professionals such as radiologists or sleep medicine physicians? How many labels per sample can the project afford? Is it even possible to develop canonical ground truth labels for any or all data samples? To what extent will there be a degree of uncertainty or disagreement amongst the annotators. It is this question in particular that forms the basis of this study. How can canonical ground truth labels be identified for training datasets when there is uncertainty or disagreement amongst the annotators?

We have frequently asked this question within our research group as we have studied the application of modern machine learning techniques to various aspects of the diagnosis [1] and treatment of Obstructive Sleep Apnoea (OSA) [2], [3], [4]. In these applications, high performing models have the potential to directly and positively impact the overall health and clinical outcomes for a patient; though the challenges of producing high quality, ground truth data must first be overcome. When considering models that are capable of assisting in the diagnosis of sleep disordered breathing (SDB), it must first be acknowledged that within the field of sleep medicine the interpretation of sleep studies (the mechanism by which a diagnosis of sleep disoredered breathing is acheived) is subject to varying levels of agreement between individual scores and facilities [5]. In order to determine a canonical ground truth dataset that can be used in training a supervised learning model, this variation between individual scorers must be accounted for. In classification tasks such as this (SDB positive / SDB negative) the foundational work of Dawid and Skene provides the method of maximum likelihood estimation to solve this problem [6], considering the labels provided by and the error rates for each annotator for each observation in assigning the final label.

Our studies have also considered the means by which the treatment of OSA can be improved through the use of machine learning models. the current gold standard for the treatment of OSA is positive airway pressure (PAP) therapy, which involves the application of a pneumatic splint to the upper airway to prevent its collapse. A critical component of this treatment is the mask worn by the patient. This piece of equipment needs to ensure a sufficient seal can be formed with the skin of the patient's face as to maintain therapy, while being comfortable enough to be worn for a complete night's sleep. A number of studies have identified that mask issues are a common cause for patient's to abandon their therapy and thus not receive the health benefits associated with its use [7], [8]. It is hypothesized that improved selection and sizing of the allocated PAP masks can increase compliance and adherence to PAP therapy. Our studies, centered around the usage of frontal facial photographs to determine the final mask selection for a patient, due to the simplicity and

Ben Johnston and Philip de Chazal are with the Sleep Research Group, Charles Perkins Centre, University of Sydney, Sydney, NSW, 2006, Australia, (phone: +612 911 41528; e-mails: {ben.johnston, philip.dechazal}@sydney.edu.au)

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

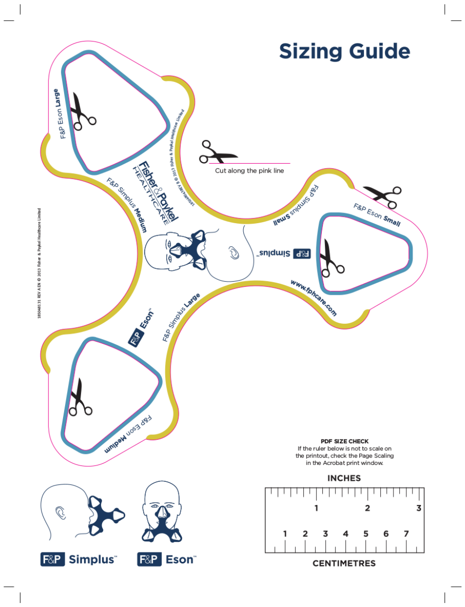

Sizing Guide
F&P Eson Large
Fisher & Paykel HEALTHCARE
185048141 REV B © 2013 Fisher & Paykel Healthcare Limited
Cut along the pink line
F&P Simplus Medium
F&P Simplus Small
F&P Eson Small
F&P Simplus
www.fphcare.com
F&P Simplus Large
F&P Eson
F&P Eson Medium
185048131 REV A EN © 2013 Fisher & Paykel Healthcare Limited
PDF SIZE CHECK
If the ruler below is not to scale on the printout, check the Page Scaling in the Acrobat print window.
INCHES
1
2
3
1
2
3
4
5
6
7
CENTIMETRES
F&P Simplus
F&P Eson

Fig. 1. Sizing Guide for a F&P Eson and Simplus PAP mask

potential telemedicine benefits for the patients. In order to determine a patient's mask size one requires the physical measurement of nose width for nasal PAP masks and face height for full face masks; this can be done using a photograph by identifying the location of the facial landmarks of interest and using a scaling mechanism within the image. Currently in clinical practice these measurements are applied to a fitting guide as shown in Figure 1 to determine the final size selection.

In a previous study we investigated the variability of 12 expert annotators in identifying the facial landmarks associated sizing CPAP masks on images [9] and the potential effect of this variability on the final mask size that could be received by a patient. This study showed while employing expert annotators the variance in landmark selection alone could lead to errors in CPAP mask size estimates between 2.9 and 13.6%. When attempting to accommodate for this error the location of the ground truth landmarks was identified as a potential source for error. However as these labels are not categorical the methods of Dawid and Skene could not be applied to determine their final value. Inspired by the Maximum Likelihood Estimation algorithm [6], in this study we will present a method for identifying canonical ground truth values for regression problems where the values for the labels are provided as continuous variables, opposed to categorical labels. By improving the quality of these selected labels the error rates reported in [9] could be reduced or even removed.

II. Methods

While the method described in this section was developed in the context of identifying ground truth labels for the location of facial landmarks on frontal images, it can also be applied to other regression problems when given continuous values in one or more dimensions. This is also supported by the fact that while the process of identifying facial landmarks can be considered innate to human annotators and does not require high levels of training or education, it can also be said to be a problem for which a single agreed upon canonical value cannot be obtained. The location of the landmarks within the image(s) is a function of the annotator's interpretation of the landmark to be selected, their consistency in selecting the same landmark more than once, as well as the care taken in completing the task at hand. Currently, the most common method for obtaining the single ground truth value for each sample in a supervised learning dataset is to simply take the mean value $\overline{\mathbf{X}}$ across all observations for the label. While being the simplest approach, this method assumes that each landmark, selected by each annotator is of equal quality and thus should contribute equally to the final label. In many cases this is simply not the case; some annotators may take more time and care during the process and thus may provide more considered and consistent values, some landmarks may be harder to identify leading to high degrees of uncertainty amongst the annotators while (particularly in the case of crowd sourced labels) may simply be adversarial, actively providing low quality labels.

A. *Description of Algorithm*

The method proposed in this study, similar to Dawid and Skene provides an interative process that considers the precision of each annotator in the selection process and contributes an amount to the final value according to each annotator's own performance. Say we have a total of A annotators, each providing R replicate values for a single label in D dimensions. The process starts by first seeding the estimate of the ground truth location for a label using the global mean $\overline{\mathbf{X}}$.

$$\overline{\mathbf{X}} = \frac{1}{RA}\Sigma X \tag{1}$$

$$\delta(a) = \sqrt{(X_{RD}(a) - \overline{\mathbf{X}})^2} \tag{2}$$

$$p(a) = \frac{1}{\frac{1}{R}\Sigma\delta(a)} \tag{3}$$

$$w(a) = \frac{1}{P}p(a) \tag{4}$$

$$x_j = \frac{1}{AR}\Sigma_{a=0}^{A} w(a)X(a) \tag{5}$$

$$\tag{6}$$

Where:

- $X \in \Re^{ARD}$ are all selections made across all annotators and replicates for a single label,
- $X_{RD}(a) \in \Re^{RD}$ are all the selections by annotator a in D dimensions,
- $p(a) \in \Re^{D}$ is the precision of annotator a and P is the sum of all values of $p(a)$,
- $w(a)$ is the weight assigned to annotator a,
- x_j is the estimate of the ground truth location at iteration j

The euclidean distance $\delta(a)$ of each replicate by each annotator from the current ground truth location x_j is then

```
Data: Manually annotated labels for a single data
      sample $X \in \Re^{ARD}$. A maximum number of
      iterations $M$ and an acceptable tolerance of
      deviation $T$.
Result: Ground truth labels $X_G \in \Re^{D}$
while $A > 1$ do
    Initialise the current ground truth estimate
      $x_0 = \overline{\mathbf{X}}$;
    for $j : 0 \rightarrow M$ - 1 do
        Compute each annotator's precision $p(a)$;
        Compute each annotator's weights $w(a)$;
        Compute the annotator weighted average
          ground truth estimate $x_{j+1}$;
        if $min(x_{j+1} - x_j) < T$ then
            break;
        end
    end
    Sum the precision of each annotator in all
      dimensions $p_t(a) = \Sigma_{d=0}^{D} p_d(a)$;
    Remove the annotator with the lowest cumulative
      precision $p_t$ from the dataset, such that
      $A \mapsto A - 1$.
end
```

Algorithm 1: Ground truth label derivation for a single label

computed and used to determine a single precision value $p(a)$ for each annotator as per Equation 3. The precision value is then used to determine the contribution of each annotator $w(a)$ (see Equation 4) to the ground truth value. This ensures that higher performing annotators contribute more to the final value, while the input of less precise annotators is not completely ignored. The annotator weights are then used to update the ground truth estimate x_{j+1} (see Equation 5) by computing a new weighted average. This process is repeated until the estimate of the ground truth location stops moving more than some tolerance value T. Finally the least precise annotator and their corresponding labels is removed from the data set and this process is repeated until one annotator remains in the set. If there are N labels to be annotated for a single data sample this process will be repeated independently for each sample in the set. The process is summarised in the pseudocode of Algorithm 1 and a Python implementation can be found at `https://github.com/doc-E-brown/johnstondechazal`.

B. Experimental Facial Landmark Localisation

To determine the efficacy of the proposed method, we completed an experimental study where we asked $N_{expert} = 12$ expert and $N_{crowd} = 100$ crowd sourced annotators to identify a subset of the MULTI-PIE/IBUG configuration as shown in Figure 2 on $I = 2$ different images using Amazon Mechanical Turk . The images were selected from the 300W [10] and AFLW [11] datasets due to the variety in poses, expressions, lighting and environments. To avoid any confusion or unintended bias in landmark idenification process, an example image (see Figure 3) of the landmark to be identified was provided to the annotator, and the order in which the landmarks or images was completely randomised. For more information regarding the data collection process the reader is referred to the Methods section of [2]. Using this data the method described in Algorithm 1 was applied and the final location of the derived ground truth labels were compared to the global mean.

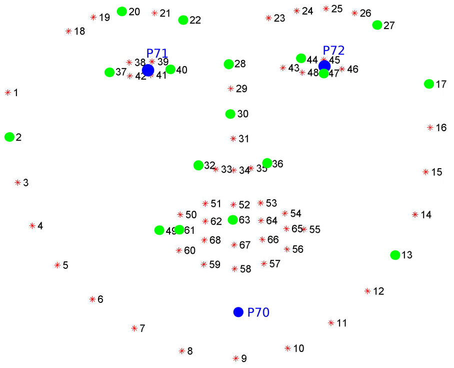
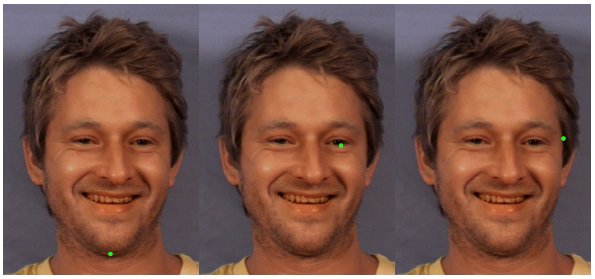


Fig. 2. The landmarks shown by the red asterix were not used but are part of the original 68 MULTI-PIE/IBUG configuration. The landmarks in green are from the MULTI-PIE/IBUG configuration and were included in the study, while the landmarks denoted by the blue circle were also included but are not part of the dataset.



(a) P9 (b) P47 (c) P17

Fig. 3. Example landmark location images.

III. Results

Figures 4 and 5 show the preliminary results of the applying Algorithm 1 to the manually annotated landmarks. The green crosses on each image indicate the position of the mean of all labels made for each landmark, while the red crosses indicate the final position as determined by the method proposed in this study.

IV. Discussion

Referring to both Figures 4 and 5 and the position of landmarks 17, there is a significant difference in the position of the global mean and the final locations. The trajectory of these landmarks over the iterative process shows the movement towards the border of the face as well as positions closer to the outline of the ear. In both images this represents an improvement in the final position, as the landmarks are no longer positioned on the forehead and are closer to the

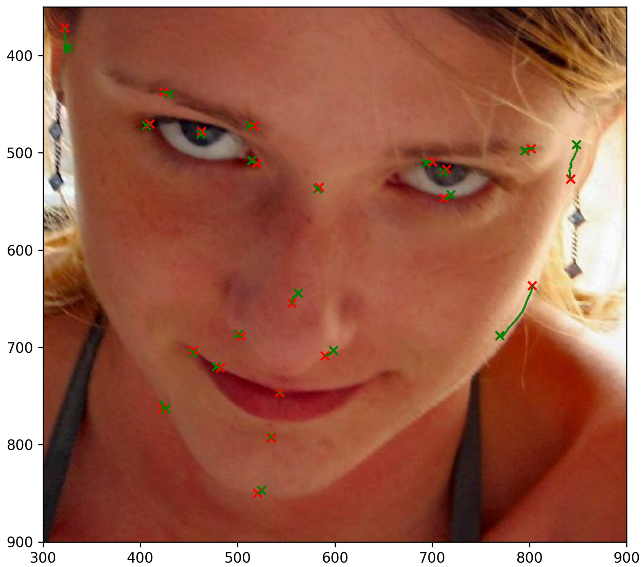


Fig. 4. Annotated Image 1, sourced from [11]. The green crosses mark the initial mean, the red crosses the final location.

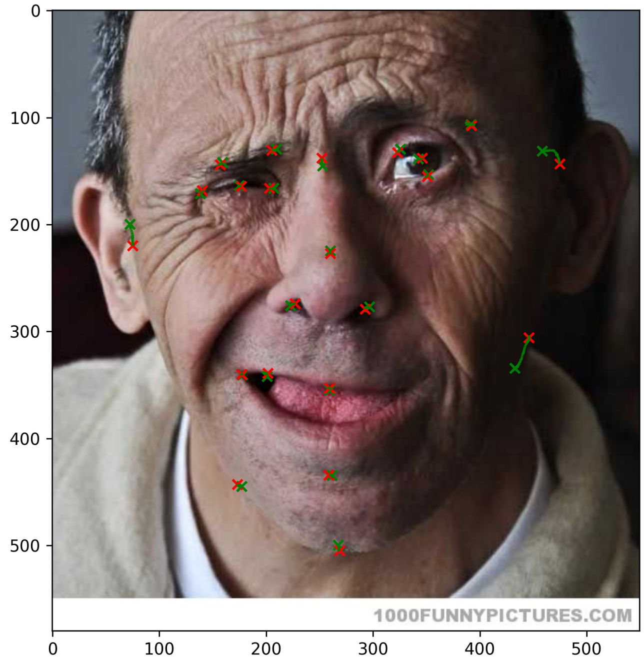


Fig. 5. Annotated Image 2, sourced from [10]. The green crosses mark the initial mean, the red crosses the final location.

designated location indicated by Figure 3c. Other locations of interest include the position of the landmarks designated as P9 and P47 as per Figure 2. Landmark P9 is intended to be identified as the apex of the chin, on the border line of the face. In both result images it can be seen that the final position is closer to the face boundary when compared with the corresponding mean value. The position of landmark P47 also lies closer to the boundary of the eyelid when compared to the starting position. For landmarks where there is a known locating feature, such as the corner of the eyes, the centre of the pupils and corner of the mouth there is little difference between the initially seeded values and the final locations as determined by Algorithm 1. This can be attributed to the fact that the locating nature of the locations leads to less uncertainty amongst the labellers [2]. These results are being reported as preliminary as there is further analysis that is currently being completed, including the effect of modifications to the method itself. As reported in this study, Algorithm 1 eliminates annotators until there is only one remaining in the set. Alternatives to be explored include adjusting the number of annotators in the final pool as well as experimenting with different methods for seeding the process.

V. Conclusion

In conclusion we propose a novel method for determining the final ground truth label from manually annotated datasets for use in supervised regression problems. This method considers the precision of each annotator separately for each label to ensure the most precise annotators contibute the most to the final value. In this study we presented the initial experimental results using this method on a large sample of annotated data.

References

[1] Asghar Tabatabaei Balaei, Kate Sutherland, Peter A. Cistulli, and Philip De Chazal, "Automatic detection of obstructive sleep apnea using facial images," in *Proceedings - International Symposium on Biomedical Imaging*, Melbourne, apr 2017, pp. 215–218, IEEE.

[2] Benjamin Johnston and Philip de Chazal, "Automatic PAP Mask Sizing with an Error Correcting Autoencoder," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. jul 2019, pp. 3677–3680, IEEE.

[3] Benjamin Johnston, Alistair McEwan, and Philip de Chazal, "Semi-automated nasal PAP mask sizing using facial photographs," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. jul 2017, pp. 1214–1217, IEEE.

[4] Benjamin Johnston and Philip de Chazal, "A Fully Automated System for Sizing Nasal PAP Masks Using Facial Photographs," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. jul 2018, pp. 3979–3982, IEEE.

[5] R G Norman, I Pal, C Stewart, J A Walsleben, and D M Rapoport, "Interobserver agreement among sleep scorers from different centers in a large dataset," *Sleep*, vol. 23, no. 7, pp. 901–908, nov 2000.

[6] A. P. Dawid and A. M. Skene, "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 20, 1979.

[7] T E Weaver, N B Kribbs, A I Pack, L R Kline, D K Chugh, G Maislin, P L Smith, A R Schwartz, N M Schubert, K A Gillen, and D F Dinges, "Night-to-night variability in CPAP use over the first three months of treatment.," *Sleep*, vol. 20, no. 4, pp. 278–83, 1997.

[8] R Zozula and R Rosen, "Compliance with continuous positive airway pressure therapy: assessing and improving treatment outcomes.," *Current opinion in pulmonary medicine*, vol. 7, no. 6, pp. 391–8, 2001.

[9] B Johnston and P d. Chazal, "The Effect of Landmark Variability on Automated PAP Mask Sizing," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, jul 2019, pp. 4129–4132.

[10] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic, "300 Faces In-The-Wild Challenge: database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, mar 2016.

[11] Martin Kostinger, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Barcelona, Spain, nov 2011, pp. 2144–2151, IEEE.