Estimating continuous affect with label uncertainty

Niki Maria Foteinopoulou
School of Electronic Engineering and Computer Science
Queen Mary, University of London
London, United Kingdom
n.m.foteinopoulou@qmul.ac.uk

Christos Tzelepis
School of Electronic Engineering and Computer Science
Queen Mary, University of London
London, United Kingdom
c.tzelepis@qmul.ac.uk

Ioannis Patras
School of Electronic Engineering and Computer Science
Queen Mary, University of London
London, United Kingdom
i.patras@qmul.ac.uk

***Abstract*—Continuous affect estimation is a problem where there is an inherent uncertainty and subjectivity in the labels that accompany data samples – typically, datasets use the average of multiple annotations or self-reporting to obtain ground truth labels. In this work, we propose a method for uncertainty-aware continuous affect estimation, that models explicitly the uncertainty of the ground truth label as a uni-variate Gaussian with mean equal to the ground truth label, and unknown variance. For each sample, the proposed neural network estimates not only the value of the target label (valence and arousal in our case), but also the variance. The network is trained with a loss that is defined as the KL-divergence between the estimation (valence/arousal) and the Gaussian around the ground truth. We show that, in two affect recognition problems with real data, the estimated variances are correlated with measures of uncertainty/error in the labels that are extracted either by considering multiple annotations of the data, or by manually cleaning the dataset.**

***Index Terms*—Affect estimation, uncertainty, noisy labels**

I. Introduction

Affect recognition in the wild is a problem that traditionally is using the labels assigned by expert annotators or self-reporting as the ground truth. Even though the labels obtained in that manner are not as noisy as, for example, through social media scraping, there is an inherent element of subjectivity in annotation that can be regarded as noise or bias. This subjectivity in available affect datasets can have an effect on generalisation and interpretability of results.

In the recent years, several works have attempted to address label uncertainty. DivideMix [1] introduces a methodology for training on noisy labels by leveraging a semi-supervised technique. The method is simultaneously training two networks and uses the per-sample training loss to co-divide the data unto a clean- and a noisy-label subset. However, the methodology proposes a hard label correction by assigning pseudo-labels on noisy samples during training and requires co-training of two networks. In the regression framework, He et al. [2] model the difficulty in predicting object boundaries in object detection by estimating the uncertainty in predicting the bounding box in the form of variance and introducing a Kullback-Leibler (KL) based loss term that allows the estimation of the variance for each predicted boundary. However, none of the above have been introduced in the domain of affective computing for continuous arousal and valence estimation.

In this work, we adopt a similar approach and build on the work of He et al. [2] in order to address the problem of label uncertainty in the domain of affective computing. We address the problem of affect estimation as a regression problem predicting a continuous value for arousal and valence. We propose to estimate the uncertainty of the label for each sample in the form of variance, so that the model estimates both the target and the label variance. By contrast to approaches such as DivideMix [1] that model the distribution of the loss over multiple samples, and make a hard decision between which samples are noisy and clean, our measure is continuous and is derived per sample by a branch of the network. Our network is trained on a KL-divergence based loss using standard back-propagation. We evaluate the methodology on two continuous affect datasets, namely AMIGOS [3] for video affect estimation and AffectNet [4] for affect estimation in static images. We show that the derived measure is positively correlated to the variance of annotators in AMIGOS where, multiple annotations are available. In AffectNet, where multiple annotations are not available, we use the rules proposed by [5] to obtain a clean and a noisy validation set and show that the estimated variances in the clean subset are lower than in the noisy one by performing a statistical significance test. Finally, we show that the proposed methodology consistently improves the performance in both datasets against their baselines.

The main contributions of this paper can be summarised as follows:

1) We propose addressing the problem of continuous affect estimation with label uncertainty, by modelling the ground truth label as a uni-variate Gaussian distribution with unknown variance and training a network that

This work is supported by EPSRC DTP studentship and EU H2020 AI4Media No. 951911 project.

learns to predict it. To the best of our knowledge, this is the first work doing so in this domain.

2) We show that the proposed methodology improves the performance upon the adopted baselines on both image and video data affect recognition problems.
3) We quantitatively evaluate the predicted variance metric as a measure of uncertainty and show that it is positively correlated with the variance of multiple human annotators in AMIGOS, and higher in part of AffectNet that were deemed to contain noisy samples.

The paper is organised as follows. Section II discusses related literature, Section III introduces our methodology, Section IV introduces the experimental setup, Section V reports the results, and Section VI concludes the paper.

II. Related Work

In this section, we review previous work addressing label uncertainly and multiple annotators, and review works in continuous arousal and valence estimation.

A. Addressing Data Uncertainty

Significant amount of work has been done on data uncertainty in the form of noisy labels for classification tasks. Methodologies such as MixMatch [6], DivideMix [1], and FixMatch [7] are adopting a semi-supervised approach to address noisy labels and make a decision during training that splits samples into clean and noisy subsets. However, this approach, i.e., of making a hard decision on uncertain samples, does not offer interpretability of the per-sample data uncertainty. In contrast, the proposed method adopts a continuous measure which is derived per sample by a branch of the network.

Bayesian deep learning approaches have gained popularity in dealing with data uncertainty; for instance, for the task of image segmentation, Kendall and Gal [8] proposed a per-pixel regression uncertainty-aware approach. Similarly, modelling data uncertainty in latent space [9], [10] has proven to improve face recognition. Moreover, in domains such as object detection [2] and temporal action localisation [11], data uncertainty is addressed by learning the variance of a continuous prediction value, i.e., the bounding box spatial boundaries of an object in an image or the temporal boundaries of an action in a video, by optimising a modified KL divergence loss function.

In these works, uncertainty is modelled per sample as a set of uni-variate Gaussian distributions of the predicted regression values with both mean values and variances being predicted by the network. In contrast, instead of the predictions, the proposed method models the ground truth values as uni-variate Gaussians, for which the true mean values are given and the variances are optimised using a KL-divergence based loss term. Moreover, while data uncertainty modelling has been implemented in other regression problems, it appears that none of these works address the problem in continuous affect estimation.

B. Multiple Annotators

Several works have addressed the issue of uncertainty when multiple annotations of a given sample are available. Using a Gaussian process classification approach has been proven to outperform other approaches (e.g., majority voting) in multiple domains [12], [13]. These works explicitly handle uncertainty arising from annotators' disagreement. Similarly, ensemble architectures that model each annotator and implement decision level fusion [14] for each sample show improvements against baseline. However, such approaches require a large number of annotations per sample to model the annotation distribution and guarantee it is representative. By explicitly handling the uncertainty in Gaussian processes, the network learns the annotator's disagreement rather than the sample ambiguity. Furthermore, the latter approach of ensembles from individual annotator models does not provide sufficient information on the sample's uncertainty.

C. Continuous Affect Estimation

In the field of emotion and affect estimation, Yannakakis et al. [15] propose comparing samples and ranking them rather than using the absolute labels to address data uncertainty. This is an interesting approach to address label uncertainty, however most datasets are annotated in a categorical or continuous manner and not in rankings. A recent work by Toisoul et al. [5] also evaluates against a clean dataset, where samples are excluded when deemed noisy by a set of predefined rules. Their method performs better on the clean evaluation set, even though noisy sample labels are not corrected or excluded during training. Resigno et al. [16] propose the use of personal models for affect recognition to overcome generalisation issues due to physiological or cultural differences. However, the aforementioned works do not estimate the level of label uncertainty in affect estimation, but rather attempt to clean the dataset of noisy samples. Han et al. [17] propose an uncertainty aware methodology for continuous affect estimation by explicitly training on the inter-annotator disagreement as an additional task. Similarly, Chou and Lee [18] propose an ensemble methodology for speech emotion classification and use annotators' disagreement as a target during training. However, while their methodology improves on the baseline showing the importance of uncertainty aware models, it is dependent on individual annotations being available.

III. Methodology

In tasks where multiple annotations per sample are available (specifically in emotion and affect recognition), majority voting or averaging over the given multiple labels approaches are typically followed in order to obtain a single ground truth label per sample. Such methods, however, neglect the uncertainty that is inherent in such annotations and that are introduced by multiple, usually disagreeing, annotators. Furthermore, multiple annotations per sample are not always available, making methodologies that explicitly handle label uncertainty in the data not applicable. In this section, we present our method for a) modelling the aforementioned uncertainty in the given

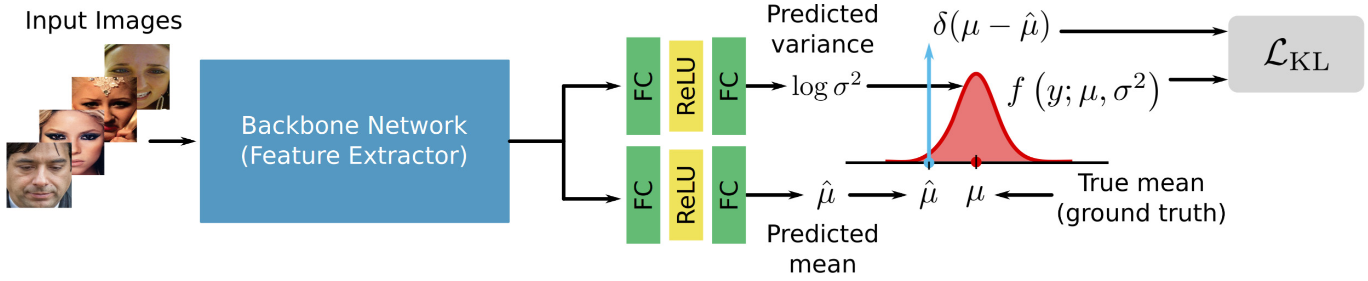

Fig. 1: Proposed method overview: A backbone convolutional neural network is applied to input images in order to extract features which are subsequently used by two MLP heads in order to predict a) the variance σ^2 (top branch) and b) the mean $\hat{\mu}$ (bottom branch) of the annotation $y \sim \mathcal{N}\left(\mu,\sigma^2\right)$ for a given training sample. A KL-divergence loss function is then used to measure the difference between the Gaussian distribution $f(y;\mu,\sigma^2)$ and the Dirac delta distribution $\delta(\mu-\hat{\mu})$.

annotations and b) using it in order to predict both the (ground truth) mean value of the label and its (unknown) variance. By doing so, we expect to estimate an interpretable metric for label uncertainty and improve the performance of affect estimation. An overview of the proposed method is shown in Fig. 1.

A. *Ground truth uncertainty estimation*

We begin by modelling the ground truth annotations as a set of independent uni-variate Gaussian distributions, for which we are given the true mean values (ground truth), and we try to predict both the mean values and the corresponding variances. More specifically, let $y \sim \mathcal{N}\left(\mu,\sigma^2\right)$ denote an annotation label (e.g., the value of arousal for a given sample) with true mean value μ and unknown variance σ^2. For doing so, we jointly optimise a convolutional feature extractor backbone network and two MLP "heads", one predicting the mean and the other predicting the variance of the respective Gaussian, as shown in Fig. 1.

We achieve this by optimising a KL-divergence based loss function, $\mathcal{L}_{\mathrm{KL}}$, which measures the difference between the predicted Gaussian, which is uniquely expressed by its true mean μ and the predicted variance σ^2 and its density is given by $f(y;\mu,\sigma^2)$, and a Dirac delta distribution centred at the predicted mean value $\hat{\mu}$, with density given by $\delta(\mu-\hat{\mu})$ (see Fig. 1).

It is worth noting that, in order to impose positivity on the predicted variance and avoid exploding gradients, we implicitly predict its Napierian logarithm, $s=\log\sigma^2$, and use it as $\exp(s)=\sigma^2$, as we will show below. That is, as shown in Fig. 1, the top MLP predicts the logarithm of σ^2.

We note that KL-divergence is a distribution-wise asymmetric measure, which does not satisfy the triangle inequality, and thus cannot serve as a true metric function. However, it is widely used for measuring dissimilarity between statistical distributions [2], [11]. For instance, He et al. [2] incorporate a similar KL-divergence based loss function for measuring the distance between a uni-variate Gaussian and a Dirac delta distribution.

By following similar arguments as in [2], we introduce a KL-divergence based loss function given by

$$\mathcal{L}_{\mathrm{KL}} = \frac{(\mu-\hat{\mu})^2}{2\sigma^2} + \frac{\log\sigma^2}{2}, \tag{1}$$

when $|\mu-\hat{\mu}| \leq 1$, and by

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{\sigma^2}\left(|\mu-\hat{\mu}| - \frac{1}{2}\right) + \log\sigma^2, \tag{2}$$

when $|\mu-\hat{\mu}| > 1$. That is, in the cases where the predicted mean values are far from their true values (typically during the early training process), we use the latter modified smooth $\mathcal{L}_1$ loss term shown in (2), while after achieving certain convergence we use the former fine-grained and uncertainty-aware loss term (1).

We note that, in contrast to [2] that model their regression predictions as uni-variate Gaussians and optimise their variances, we, instead, predict the variance of the ground truth values for our regression task. This reflects the intuition that affect labelling is prone to noise. The proposed loss takes into account the estimated variances of labels unlike other losses traditionally used for regression problems (eg. Mean Absolute Error or Mean Squared Error); for more ambiguous or noisy samples we expect the model to estimate a higher variance.

B. *Architectures*

As discussed in the previous sections, in this work we address the problem of data uncertainty on continuous affect estimation from both static images and videos. For affect estimation from static images we set the general architecture presented in Fig. 1 so as the backbone feature extractor is implemented by a CNN architecture. More specifically, we have experimented with both VGG16 [19] and ResNet [20] architectures (see Fig. 2), however, the proposed methodology can be implemented on any appropriate network, as described in the previous section.

In the case of continuous affect estimation on untrimmed videos, our basic architecture (Fig. 1) is set so as video features are obtained using a CNN with a trainable NetVLAD [21] layer, as shown in Fig. 3. The NetVLAD architecture [21] is inspired by the Vector of Locally Aggregated Descriptors

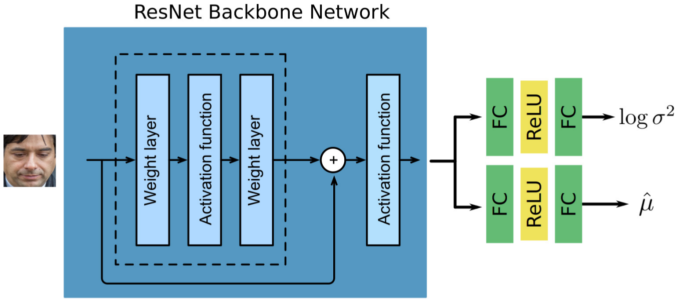

Fig. 2: Residual CNN backbone architecture for extracting features from static images.

(VLAD), which is a pooling method that captures information about the statistics of local descriptors over the image, by storing the sum of residuals from cluster centers.

More specifically, the NetVLAD introduced in [21] can update the cluster centres during training, therefore the layer can be introduced as a pooling layer in a standard convolutional architecture. The original NetVLAD layer is used to generate a $K \times D$ vector from a $W \times H \times D$ convolutional output, where K is the number of centroids to be used in the VLAD vectors, D is the number of channels of the last convolutional layer, and (W, H) are the spatial dimensions of the convolutional output, as shown in Fig. 3.

In this work, we modify the NetVLAD layer architecture to perform pooling along the temporal dimension, instead of the spatial. The input to the network is a set of pre-computed features, obtained during pre-training from each video frame. The network then performs a convolutional and average pooling operations followed by ReLU activation across the temporal dimension and then uses the NetVLAD layer as a pooling layer to standardise the feature vector size. The proposed architecture using NetVLAD offers certain advantages; more specifically, it allows for the use of untrimmed video input and can handle longer sequences. It also offers a good performance versus simplicity trade off.

IV. Experimental Setup

A. Datasets

a) AMIGOS: The AMIGOS dataset [3] consists of audio-visual and physiological responses of participants (either alone or in a group) to a video stimulus. In this work, we use the responses of individuals; 40 participants watched 16 short videos and 4 long ones. The former are defined as videos with length in the 50-150 second range. The responses are broken down to 20-second intervals and annotated by three annotators for *arousal* and *valence* on a scale from -1 to 1. We extract the frames from the video with a framerate of 25 frames/sec and calculate the average score of the three annotators as the ground truth during training for the video segment. During testing, we use the variance of the annotators as an indication of uncertain or ambiguous samples and calculate the Pearson's Correlation Coefficient (PCC) between estimated and annotator's variance.

TABLE I: Correlation Coefficient of Annotators Scores for Arousal and Valence in the AMIGOS dataset

	Arousal			Valence		
	#1	#2	#3	#1	#2	#3
#1	1	0.54	0.62	1	0.7	0.73
#2	0.54	1	0.51	0.7	1	0.63
#3	0.62	0.51	1	0.73	0.63	1

As the individual annotator scores are available, we calculate the correlation matrices for arousal and valence as an indication of Inter-Annotator Agreement (IAA) in continuous affect estimation, as shown in Table I. The correlations in the table indicate that there is disagreement between the annotators particularly for arousal. Higher disagreement of annotators will be introducing higher label uncertainty as it is an indication of the sample's ambiguity. By examining the histogram of variances of the available annotations in Fig. 4, we can see that while most samples will have low disagreement and thus low uncertainty, particularly for arousal there is a significant number of samples with higher variance.

b) AffectNet: AffectNet [4] consists of more that one million facial images collected from the Internet. Approximately 440,000 are annotated manually for categorical emotions, and continuous arousal and valence. In this work we use the manually annotated samples of the eight emotion categories, namely, *Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger*, and *Contempt*, which include over 290,000 samples. Annotations from multiple annotators are not provided in the dataset.

B. Performance Measures

The performance of the proposed methodology and the baselines is assessed using three evaluation metrics, depending on the database. For experiments conducted on the AMIGOS database [3], we report the Mean Square Error (MSE):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(\mu_i - \hat{\mu}_i\right)^2, \tag{3}$$

where n is the number of videos in the database, μ_i is the ground truth and $\hat{\mu}_i$ is the predicted value, as discussed in Sect. III. To better assess the performance of the regression task and to guarantee that results are comparable with other methods that apply transformations on the labels, we use Pearson's Correlation Coefficient (PCC), which for a pair of variables x, y with means $\bar{x}, \bar{y}$ is given by

$$\text{PCC} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}_i\right)\left(y_i - \bar{y}_i\right)}{\sqrt{\sum_{i=1}^{n} \left(x_i - \bar{x}_i\right)^2 \left(y_i - \bar{y}_i\right)^2}}. \tag{4}$$

The above equation is used to evaluate both the performance of the regression when predicting the level of arousal/valence, and the quality of the learnt variance. In addition to PCC, we also evaluate the performance of our method in the regression task in AffectNet using the Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\mu_i - \hat{\mu}_i\right)^2}, \tag{5}$$

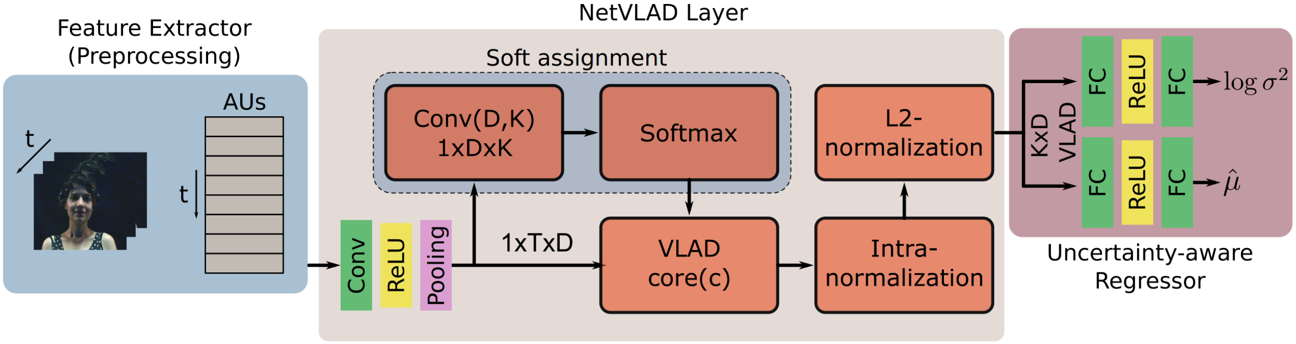

Fig. 3: Video input architecture: Given an untrimmed video with t number of frames, we extract a vector of Action Units (AUs) per frame in the preprocessing phase. The AU time-series is then used to train the NetVLAD architecture along with our uncertainty-aware regressor.

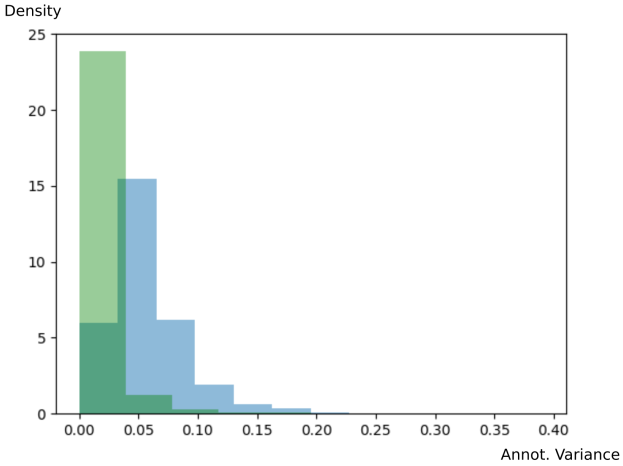

Fig. 4: Histogram of annotators' variance in the AMIGOS dataset for *arousal* and *valence*.

where n, μ_i, and $\hat{\mu}_i$ denote the number of images, the ground truth, and the predicted value for arousal/valence, respectively.

C. Backbone and Implementation Details

a) Affect estimation in videos: We evaluate the proposed method in the task of affect estimation in untrimmed videos using the AMIGOS [3] dataset. For this, we use ResNet50 as a backbone architecture (Fig. 2), which we have pretrained on the CelebA [22] and the EmotioNet [23] datasets for the task of Action Units (AUs) recognition [24]. We use this pre-trained backbone in our preprocessing phase (Fig. 1) in order to extract features. Therefore, the input to the model is a time-series of ten AUs per video segment. The AUs-based features extracted by the backbone are then used to train a simple CNN architecture using a NetVLAD [21] layer to produce a fixed-dimensional feature vector that is then fed forward to the regression and variance estimation fully connected (FC) layers, as shown in Fig. 3. We chose a trainable NetVLAD layer as a baseline since it offers a low simplicity-vs-performance trade off. The 1D convolutional and average pooling layers are set with a kernel size of 7 and stride 5 and the same number of channels according to the input. As we do not down-sample frames in the video sequence, we assume neighbouring frames will have similar values and therefore implement a larger kernel and stride. The NetVLAD layer is initialised with 8 centroids. The training is performed in an end-to-end manner, and we follow a leave-one-subject-out cross validation protocol for each subject in the individual database, until the network converges. The network is trained using an ADAM optimiser with an initial learning rate of 0.01 multiplied by a factor of 0.1 every 100 epochs and a batch size of 512 on two NVIDIA RTX 2080 GPUs.

b) Affect estimation in static images: In the case of affect estimation in static images, we evaluate the proposed method using both the VGG16 [19] and the ResNet50 [20] architectures as a backbone (Fig. 2), in order to assess the effect of the variance prediction and KL divergence loss. We also train a ResNet18 network and initialise convolutional layers with weights pretrained on ImageNet. All networks are trained using Stochastic Gradient Descent (SGD) optimisation, with an initial learning rate of 0.0001 multiplied by 0.8 after 100 epochs, and a batch size of 128 until convergence.

V. Results and Discussion

In order to assess the impact of the learned variances, we compare them with the corresponding variances induced by annotators disagreement – when multiple annotators' scores are available we can estimate uncertainty in the form of variance between annotators' scores. We propose to evaluate the learned variances against the annotator's variances at test time. It is worth noting that, unlike [12], [13], [17], [18], we do not use the annotator's variance in the training phase as a target, but instead we learn each annotation's variance from input and evaluate in the test phase.

TABLE II: PCC of learned variance and annotators variance on AMIGOS dataset

	Arousal	**Valence**
Proposed method	0.34	0.31

In AMIGOS dataset, we use the PCC, given by (4), to calculate the correlation between the learned and the annotators' variances, and we show the results in Table II. We observe a higher PCC for arousal, which also had a lower IAA as seen in Table I. This is an indication of the model's understanding of ambiguity. Examples of clips with low and high predicted variance from the AMIGOS dataset are shown in Fig. 5.

In order to split the evaluation set of AffectNet into a clean and a noisy subset, we follow the rules proposed in [5]. That is, we split the evaluation set based on the categorical and continuous affect labels, since multiple annotations per sample are not available. More specifically, for each sample in the evaluation set, we compare the categorical emotions to their theoretical equivalent in the arousal-valence circumplex and ensure that the assigned label for arousal and valence is in agreement with the arousal and valence of the categorical emotions. For example, a sample with assigned emotion "Happy" in the categorical model, but negative arousal, would be excluded from the clean set. Examples from the two subsets can be seen in Fig. 7. In the top row, we show examples where the categorical emotion is consistent with the continuous arousal and valence, while in the bottom row examples of noisy samples are presented. In total, 141 samples are flagged as noisy.

We then estimate the variance for each sample in the subsets and compare the hypothesised population variances using a student t-test. The resulting average predicted variance for each subset is shown in Table III. The estimated variances are obtained using the ResNet18 architecture initialised with ImageNet weights. Assuming the null hypothesis $H_0 : \sigma_{clean} = \sigma_{noisy}$ and the alternative hypothesis $H_1 : \sigma_{clean} < \sigma_{noisy}$, we perform a one-tailed Student's t-test. We compute t as follows

$$t = \frac{\hat{x}_1 - \hat{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{6}$$

where x_i and s_i represent the means and variances of the two samples, respectively, and n_i is the respective sample size. With the values from Table III, t is estimated at -0.91 and -1.96 for arousal and valence respectively. The calculated p values with 139 degrees of freedom for arousal and valence are 0.18 and 0.025, respectively. Therefore, we can reject the null hypothesis for valence at 95% confidence interval, but not for arousal. As the use of C.I. is dependent on both the problem and how much uncertainty is acceptable for it, we want to note that we can reject the null hypothesis for arousal with a lower C.I. of 80%. While the use of a lower C.I. is atypical for most tests of statistical significance, we want to emphasize that in this case a test with lower confidence successfully shows a relationship between estimated variance and label noise. The weak relationship, shown by accepting the null hypothesis with lower C.I., is also testament to the difficulty of the problem, as well as evidence of other entangled factors affecting label noise. The distributions of the estimated uncertainty for the two subsets are shown in Fig. 7. In the plotted distributions, we can visually confirm the differences between estimated uncertainty for arousal and valence between the sets. While there are some overlapping areas between the distribution of estimated variances of the clean and noisy sets, the mean of the distribution is higher for the noisy set on both targets.

TABLE III: Mean estimated variance for Arousal and Valence on AffectNet subsets

	Samples	**Arousal(std)**	**Valence(std)**
AffectNet clean	3858	0.0775(0.0025)	0.0787(0.004)
AffectNet noisy	141	0.0820(0.003)	0.0872 (0.002)

TABLE IV: Results on AffectNet using VGG16 and ResNet backbones

	Arousal		**Valence**	
	RMSE	**PCC**	**RMSE**	**PCC**
wideResNet	0.3515	0.5394	0.4049	0.5979
wideResNet proposed	0.3483	0.5458	0.4061	0.6136
VGG16	0.351	0.536	0.392	0.616
VGG16 proposed	**0.343**	0.552	0.404	0.624
ResNet18 (pretrained)	0.3444	0.5540	0.3995	0.6217
ResNet18 (pretrained) proposed	0.3449	**0.57**	**0.387**	**0.6321**

In order to evaluate the proposed methodology and the impact of predicting variance to the overall model performance, we compare the architectures against their baseline trained without variance prediction and an MSE loss. The results for AffectNet and AMIGOS are shown in Tables IV and V, respectively. We can see that the improvement in terms of PCC is consistent on estimation from both static image input and time-series input. In the AffectNet (static images), we have experimented with three different backbone architectures, namely VGG16 [19] and two variants of ResNet [20], obtaining consistent improvements in terms of the PCC. The architectures tested are simple uni-modal feed-forward networks as we aim to demonstrate the impact of uncertainty prediction. A higher predicted variance for an uncertain sample allows the network to learn from less ambiguous samples as the optimiser will prioritise lowering the $|\mu - \hat{\mu}|$ term in (1) and (2). Furthermore, by penalising the regression prediction less for uncertain samples, the predicted variance regularises the error.

Finally, for reference we note that the results on AMIGOS are in line with previous work from [25], although not directly comparable as different features and architectures are used. Specifically, in [25] Quantised Local Zernike Moments (QLZM) computed from the per frame facial landmarks were used to train an SVR and an LSTM architecture, while in our case, we used a simple frame-based estimation of a set of Facial Action Units. Moreover, while there are some methodological parallels between the NetVLAD architecture used and Fisher Vectors of QLZM used to train the SVR, recurrent methodologies better capture the temporal dimension of features which is significant in continuous affect. The SVR architecture in [25] achieves a PCC of 0.34 for both arousal and valence, while the LSTM architecture achieves 0.6 and 0.62 respectively.

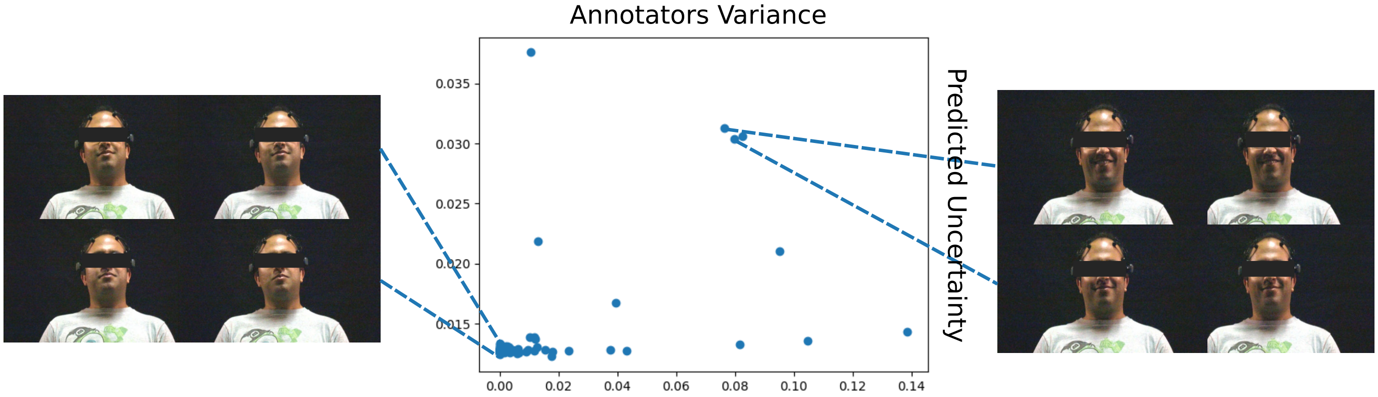

Fig. 5: Examples of clips with low predicted variance (left – annotators assessments: $0.36, 0.12, 0.14$) and high predicted variance (right – annotators assessments: $0.77, 0.21, 0.49$) from a given subject.

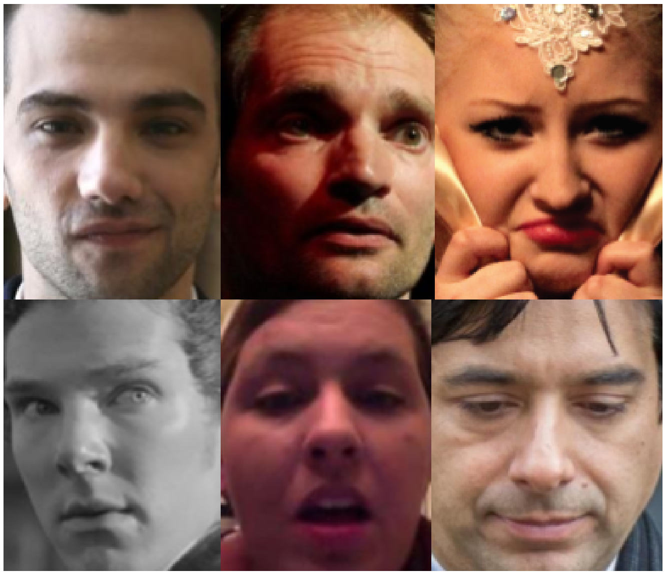

Fig. 6: Examples of samples with clean (top) and noisy (bottom) labels. Top – from left to right the assigned labels are: “Contempt, Arousal: 0.65, Valence:-0.65”, “Fear, Arousal: 0.53, Valence: -0.06”, “Sad, Arousal: -0.24, Valence: -0.66”. Bottom – from left to right the assigned labels are: “Fear, Arousal: -0.32, Valence: -0.08”, “Neutral, Arousal: -0.23, Valence: -0.37”, “Neutral, Arousal: -0.29, Valence: 0.36”.

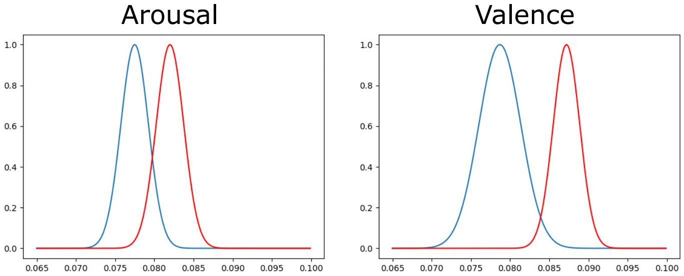

Fig. 7: Distribution of the estimated uncertainty in Arousal (left) and Valence (right) for clean (blue) and noisy (red) labels in AffectNet.

TABLE V: Results on AMIGOS using precomputed per frame Facial Action Units as input and a NetVLAD architecture.

	Arousal		**Valence**	
	MSE (std)	**PCC**	**MSE (std)**	**PCC**
NetVLAD	0.026 ($3e^{-}3$)	0.499	0.018 ($2e^{-}3$)	0.47
NetVLAD proposed	0.0354($6e^{-}3$)	**0.53**	0.018($2e^{-}3$)	**0.52**

VI. Conclusion

Continuous affect estimation is an inherently uncertain problem due to the subjective and ambiguous nature of continuous labels. We have proposed estimating the level of continuous affect along with a certainty metric that represents the true variance in the label distribution of continuous arousal and valence. The methodology is inspired by work on other domains with label uncertainty such as bounding box regression, but to our knowledge this is the first work addressing the problem in affective computing by treating the ground truth as a Gaussian distribution and the predicted level of affect as a Dirac delta function. We evaluate our methodology on two datasets, AMIGOS [3] and AffectNet [4] for affect estimation from video and static images respectively and find that it improves upon the baselines for all architectures tested. We also evaluate the learned uncertainty metric, by comparing the learned variance against the annotators’ variance when multiple annotations per sample are available. We find a positive correlation between the estimated uncertainty and the disagreement between annotators. When multiple annotations are not available, we compare the distribution of the predicted variance on a clean and noisy evaluation subsets and find the estimated uncertainty in the clean set lower using a statistical test. The proposed methodology offers a measure for label uncertainty in continuous affect recognition.

References

[1] Junnan Li, Richard Socher, and Steven C. H. Hoi, "DivideMix: Learning with Noisy Labels as Semi-supervised Learning," *arXiv:2002.07394 [cs]*, Feb. 2020, arXiv: 2002.07394.

[2] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang, "Bounding Box Regression With Uncertainty for Accurate Object Detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, June 2019, pp. 2883–2892, IEEE.

[3] Juan Abdon Miranda Correa, Mojtaba Khomami Abadi, Niculae Sebe, and Ioannis Patras, "AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[4] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, Jan. 2019.

[5] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nature Machine Intelligence*, vol. 3, no. 1, pp. 42–50, Jan. 2021, Number: 1 Publisher: Nature Publishing Group.

[6] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel, "MixMatch: A Holistic Approach to Semi-Supervised Learning," *arXiv:1905.02249 [cs, stat]*, Oct. 2019, arXiv: 1905.02249.

[7] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel, "FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence," *arXiv:2001.07685 [cs, stat]*, Nov. 2020, arXiv: 2001.07685.

[8] Alex Kendall and Yarin Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 5574–5584.

[9] Yichun Shi and Anil K. Jain, "Probabilistic Face Embeddings," 2019, pp. 6902–6911.

[10] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei, "Data uncertainty learning in face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[11] Ting-Ting Xie, Christos Tzelepis, and Ioannis Patras, "Boundary Uncertainty in a Single-Stage Temporal Action Localization Network," *arXiv:2008.11170 [cs]*, Aug. 2020, arXiv: 2008.11170.

[12] Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro, "Gaussian process classification and active learning with multiple annotators," in *Proceedings of the 31st International Conference on Machine Learning*, Eric P. Xing and Tony Jebara, Eds., Bejing, China, 22–24 Jun 2014, vol. 32 of *Proceedings of Machine Learning Research*, pp. 433–441, PMLR.

[13] Chengjiang Long and Gang Hua, "Multi-class multi-annotator active learning with robust gaussian process for visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[14] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton, "Who Said What: Modeling Individual Labelers Improves Classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018, Number: 1.

[15] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Transactions on Affective Computing*, 2018.

[16] Martina Rescigno, Matteo Spezialetti, and Silvia Rossi, "Personalized models for facial emotion recognition through transfer learning," *Multimedia Tools and Applications*, vol. 79, no. 47, pp. 35811–35828, Dec. 2020.

[17] Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller, "From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty," in *Proceedings of the 25th ACM international conference on Multimedia*, New York, NY, USA, Oct. 2017, MM '17, pp. 890–897, Association for Computing Machinery.

[18] H. Chou and C. Lee, "Every Rating Matters: Joint Learning of Subjective Labels and Individual Annotators for Speech Emotion Classification," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5886–5890, ISSN: 2379-190X.

[19] Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Apr. 2015, arXiv: 1409.1556.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition," 2016, pp. 770–778.

[21] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1437–1451, June 2018.

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Large-scale celebfaces attributes (celeba) dataset," *Retrieved August*, vol. 15, no. 2018, pp. 11, 2018.

[23] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, and Aleix M Martinez, "Emotionet challenge: Recognition of facial expressions of emotion in the wild," *arXiv preprint arXiv:1703.01210*, 2017.

[24] Paul Ekman and Wallace V. Friesen, *Facial action coding system: Investigator's guide*, Consulting Psychologists Press, 1978.

[25] Wenxuan Mou, Hatice Gunes, and Ioannis Patras, "Alone versus In-a-group: A Multi-modal Framework for Automatic Affect Recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 2, pp. 1–23, June 2019.