Computational Model of Idiosyncratic Perception of Others' Emotions

Shiro Kumano, Ryo Ishii and Kazuhiro Otsuka
NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Japan

***Abstract*—This paper deals with computational modelling for predicting the idiosyncratic perception of others' emotions, namely how individual external observers will score the emotional states of others interacting with each other. We separately model the observer effect (or individual differences of observers), and the conversational-scene effect or the video-clip effect (how interlocutors are interacting), based on Bayes' theorem with the assumption of their conditional independence. The observer term describes the observer's cognitive tendency, including bias, in a probabilistic form, and does not include any clip information. In contrast, the clip term describes how a target clip is recognized by an unspecified observer. The perceived emotion is predicted to be the state that maximizes the conditional probability given the observer and target clip. An experiment with 100 observers and 97 clips demonstrated, in a leave-one-out cross-validation scenario, that 1) there is in fact no statistically and practically significant interaction between observer and clip, and 2) our Bayesian modelling achieves a 97 percent accuracy as a reference of test-retest reliability. Furthermore, when combined with existing observer and clip models that can handle unknown observers and clips, our model yielded an accuracy of around 50 percent in a more challenging leave-one-subject-and-clip-out cross-validation scenario.**

1. Introduction

Face-to-face conversation is the primary way of sharing information, understanding others' emotions, and making decisions in social life. Unfortunately, it is not very easy for people to fully understand what others are feeling, or reach full agreement about a controversial topic. The quality and efficiency of communication can be enhanced by applying information technologies to conversation support systems, such as in real-time computer-mediated visual telecommunication. This requires the automatic understanding of both human behaviour and the interlocutors' emotions. Interestingly, the main target of automatic meeting analysis is now shifting from behaviour to emotion [1], [2].

Emotion has two distinct aspects when considered in social situations: felt emotion, i.e. what the target person is actually feeling, and perceived emotion, i.e. the emotion perceived by observers. Felt emotion has been studied mainly in non-social scenarios, e.g. emotions elicited while watching an image or a video or listening to a music [3], [4], [5]. However, the latter is vital to understanding conversations; the emotion of an interlocutor is perceived by others via his/her behaviours, and the perception evolves over the course of the interaction. A practical application of perceived emotions is to visualize the emotional states of a group meeting for non-meeting members or social psychologists to achieve a deep understanding/analysis of the meeting and its atmosphere. Since not even a meeting participant knows the real emotions of the other participants, perceived-emotion-based descriptions, i.e. third-party objective descriptions, constitute a reasonable approach.

The affective computing research community is keen to infer emotions perceived by ideal or average observers. Thus far, most studies have attempted to make affective perception dependent on a specific stimulus, i.e. the verbal/non-verbal behaviour of the target [2]; e.g. when the target person is smiling, what type of emotion with the observers perceive. To reduce the subjectivity of observers, such as a large number of perception biases, as demonstrated in [6], most previous studies have gathered the perceptions of multiple observers, and targeted their representative value, e.g. the majority/peak [7] or mean [8]. Further, as in [9], observers are often employed who are unacquainted with subjects. Such collective perception or perception of wisdom of crowds (WoC) [10] approach has been widely employed by the affective computing community. However, inferring each observer's idiosyncratic perception remains challenging.

The main aim of this paper is to build a computational model for predicting observers' idiosyncratic perceptions. Our key assumption is the conditional independence of observer and video clip, which enables us to separate the observer effect and the clip effect. This independence assumption means that there is almost no interaction between observer and clip. This may sound very strong to some readers. This paper provides various types of evidence supporting this assumption, including hypothesis testing and prediction performance evaluations. Our modelling approach has two main advantages. The first is that it avoids overfitting from a computer scientific point of view. Secondly, and

S. Kumano and K. Otsuka are, and R. Ishii was, with NTT Communication Science Laboratories, 3-1 Morinosato-Wakamiya, Atsugi, Kanagawa, Japan. kumano@ieee.org

more importantly for the affective computing community, it makes it possible to exploit existing observer and clip models separately as component models.

The remainder of this paper is organised as follows. Related work is introduced in Section 2 to position this study. The proposed modelling is described in Section 3. Experimental data are reported in Section 4, followed by the experimental settings in Section 5. The proposed model is then evaluated and discussed in Sections 6 and 7. Finally, we summarize this study in Section 8.

2. Related Work

For over 50 years, personality and social psychology researchers have studied how humans judge another's internal states or characteristics e.g. emotion, personality and skills. Brunswik's lens model [11] and Kenny's social relations model (SRM) [6] are two well-known interpersonal perception models. Brunswik's lens model mainly views emotional communication as a *process* between an encoder, i.e. a person who is expressing emotion, and a decoder, i.e. a person perceiving the expressed emotion. SRM focuses more on the persons involved in the communication, and further considers the effect of the relationship, in addition to the effects of each individual. Kenny demonstrated that the observer effect is dominant, meaning there is a large interpersonal difference between observers, as also later repeatedly demonstrated in affective computing studies, such as [12]. These previous studies motivated us to predict an observer's idiosyncratic perception.

The affective computing research community is keen to build human-level or more accurate emotion recognition machines [2]. The main target has been felt emotion, i.e. what the target person is actually feeling. However, despite the recent development of self-reporting tools [13], [14], it is not easy to determine the true emotion. This has lead researchers to target representative emotional aspects, i.e. emotions perceived by observers, with or without any explicit distinction between felt and perceived emotion [15]. To reduce observer subjectivity, many researchers have employed multiple observers and targeted a representative value, e.g. the majority/peak [7] or mean [8], implicitly or explicitly assuming that the WoC label is true [10]. Note that some studies instead targeted observers' perceptions as a whole, namely they targeted their rating distribution [9], [16], to preserve all the perception information. Their models can be considered a component model (more specifically a clip model, as explained in 3.1) in the proposed modelling.

Recently, some researchers have tried to obtain a better *ground truth* label for use in model training by considering annotator expertise/bias and/or item properties, including difficulty, rather than by employing simple aggregation, e.g. [17], [18]. This is a variant of truth discovery or crowdsourcing aggregation [19]. Zhang et al. [20] proposed combining self-reports with observers' annotations to enhance the recognition performance of both felt and perceived emotions. Their aim differs from our objective, which is to predict an observer's idiosyncratic perception. Dawid and Skene [21] assumed no interaction between patient and clinician (clip and observer in our case). Their work inspired us to make the conditional independence assumption. However, they avoided any statistical testing for the assumption. Moreover, some crowdsourcing aggregation methods can output idiosyncratic perception as intermediate modelling results, although not as a direct estimation target. Thus, we compare our model with [17], which is a well-known crowdsourcing aggregation method, in 6.2.2.

3. Modelling

We predict an observer's idiosyncratic cognition in a Bayesian inference framework, and thus realize a rational inference under uncertainty [22]. This model predicts the rating score y_{ij} on an M-point scale given by observer $i \in \{1, \cdots, N_i\}$ to video clip $j \in \{1, \cdots, N_j\}$, as the score that maximizes their joint probability, $P(i, j, y)$. The classification is expressed as:

$$\begin{aligned} \hat{y}_{ij} &= \arg\max_y P(i,j,y) \\ &= \arg\max_y P(y)P(i,j|y). \end{aligned} \tag{1}$$

3.1. Proposed modelling

Our key idea is a conditional independence assumption, namely we assume that explanatory variables are independent given a response variable; $P(i, j|y) \approx P(i|y)P(j|y)$. This assumption has been working well in many classification tasks, and is called naïve Bayes modelling by the computer science community [23]. The cognition is expressed as:

$$\hat{y}_{ij} = \arg\max_y P(y)P(i|y)P(j|y). \tag{2}$$

This expression is not easy to handle, because all three terms are in different spaces and most existing models output results in a rating space. To alleviate this problem, we further transform the joint probability by using Bayes' rule as:

$$\begin{aligned} P(y)P(i|y)P(j|y) &= P(y) \cdot \frac{P(y|i)P(i)}{P(y)} \cdot \frac{P(y|j)P(j)}{P(y)} \\ &= P(i)P(j)\frac{P(y|i)P(y|j)}{P(y)} \\ &\propto \frac{P(y|i)P(y|j)}{P(y)}, \end{aligned} \tag{3}$$

where the last transform uses the fact that both probabilities $P(i)$ and $P(j)$ are constant given i and j. All the three terms are now in the rating space.

$P(y|j)$ describes how often each rating y will be given to clip j or, more simply, how clip j will be perceived by a crowd of observers. $P(y|i)$ indicates how likely it is that observer i will give rating y without possessing any clip information. We call this probability distribution cognitive tendency. This is a mixture of factors from the cognition and choice process, but the present study does not decompose

these factors, which would be required for human scientific studies. We discuss this point later in Section 7. Moreover, this paper ignores its temporal structure, and assumes that the cognitive tendency does not change over time.

3.2. Parameter estimation

There are two scenarios as regards estimating the model parameters: direct and indirect estimation scenarios. They are variants of the leave-one-out and leave-one-subject-and-clip-out cross validation schemes, respectively.

3.2.1. Direct parameter estimation scenario. If we want to predict y_{ij} and we have the full $N_i \times N_j$ data matrix, namely the ratings of all observers to all clips, excluding y_{ij}, then we can directly estimate all three probability distributions in Eq. 3, i.e. $P(y)$, $P(y|i)$ and $P(y|j)$, as maximum likelihood estimators, which are equivalent to their normalized histograms [24]. For example, $P(y)$ is the normalized histogram of ratings in the entire training dataset. Having more samples enhances the accuracy.

We use this parameter estimation scenario, to validate our proposed decomposition, Eq. 2, in 6.2.2. However, it should be noted that this scenario cannot handle unknown observers and clips, because they require $y_{i\cdot}$ (the observer's distribution) and $y_{\cdot j}$ (the clip's distribution) to calculate $P(y|i)$ and $P(y|j)$, respectively.

3.2.2. Indirect parameter estimation scenario. If no rating data is available regarding the target observer and/or clip, we need to introduce an observer *model* and/or clip *model*, which approximate these distributions from other information. There are some models, although the number of observer models is small. As an observer model we use that proposed in [25], which predicts an observer's cognitive tendency from his/her gender and personality trait scores. More specifically, it defines $\hat{P}(y|i) = P(y|i^{gender}, i^{persona})$ on the assumption that people who have the same gender and personality trait scores show the same cognitive tendency. For the clip model, this paper adopts the models described in [9], [26], which estimate the probability from the nonverbal behaviours of the target interactants being studied: $\hat{P}(y|j) = P(y|j^{behav})$. See those references for the details of the parameter estimation. Note that the proposal of more sophisticated component models is beyond the scope of this paper, and any other models that output probability distributions are applicable to our modelling.

4. Experimental data

To evaluate the proposed framework, we first generated a dataset that included ratings given by various observers to various stimuli. The stimuli were short video clips in which two people were interacting. The observers were asked to rate the emotional similarity between each pair of interlocutors on a five-point scale.

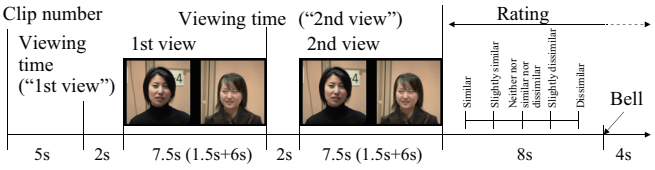

Figure 1. Timeline of emotional rating procedure in each 36-sec clip: Video was played twice after 1.5-sec pose at the initial frame.

4.1. Observers

One hundred observers (50F) participated in the experiment. They were Japanese university students in their early twenties. They had neither met interlocutors in stimulus interactions before the experiment, nor participated in the conversations.

4.2. Emotional rating procedure

For each stimulus the observers were asked to rate the emotional congruence [27] of the interacting pair, more specifically, to judge whether or not their emotions were similar. Emotional congruency or emotional contagion explains a basic aspect of affective/emotional empathy [28], [29]. It was a forced choice on the following 5-point scale: "Similar" (+2), "Slightly similar" (+1), "Neither similar nor dissimilar" (0), "Slightly dissimilar" (-1), and "Dissimilar" (-2). The stimuli consisted of 97 six-sec video clips. The clip order was counter-balanced among the observers. None of the videos included any audio signals to force the observers to focus on the emotions exchanged by visible behaviours.

In each clip, the video was first displayed twice with a short interval, and then each observer was asked to select a score. This process was repeated until the last clip ended with a 5-min interval between the 50th and 51st clips for rest. Figure 1 shows the timeline in each clip. Each observer employed a laptop computer with a 15.6-inch monitor, and used headphones to hear a bell ringing to indicate the end of a clip and to focus more on the task. All the labeling was done in isolation, and all the observers successfully completed the set task.

Furthermore, as inputs for the observer model, described in 3.2.2, the observers were also asked to provide their gender and complete three psychological questionnaires after the labelling task, namely Davis' Interpersonal Reactivity Index (IRI), the Emotional Skills and Competence Questionnaire (ESCQ), and the Tokyo University Egogram (TEG). IRI and ESCQ measure the ability/tendency to understand others' emotions. TEG is a measure of basic personality and tries to explain how people function and express their personality through their behaviour.

4.3. Stimuli

We prepared the 97 short video clips as follows. First, we obtained a face-to-face conversation dataset [9]. We selected

this Japanese dataset, whose interlocutors were of the same ethnicity as the observers, to assist the observers to more correctly recognize the interlocutors' emotions, with reference to social cognition literature [30]. The interlocutors were instructed to hold alternative-type discussions and to build consensus, i.e. agree on a single answer for each discussion topic within 6-8 minutes. The interlocutors were 20 young Japanese women. Their discussions were then annotated continuously in time[1], unlike in the observers' task (clip-by-clip), with the perceived emotional congruency labels by 5 or 9 Japanese women as non-expert coders on the aforementioned five-point scale. Using this annotation data, the present study obtained 97 6-sec clips that yielded a variety of rating histograms[2]; this resulted in balanced rating frequencies, as reported in 6.1.1.

4.4. Test-retest reliability

We consider test-retest reliability to be a good indicator of the upper bound of our prediction accuracy or the goal we hope to achieve. If prediction accuracy is comparable to reliability, it suggests that the model is reasonable, even if neither are very high.

Accordingly, from the original 97 clips, three clips, which yielded different $P(y|j)$ distributions, were pseudo-randomly selected in advance for a reproducibility test, and were re-shown immediately after the original 97 clips. The three clips were the same for all observers. There was no interval between the 97th clip and the first re-shown clip. To exclude samples where the observer was aware of repetition, the observers were debriefed and then asked to provide, independently for the three clips, their awareness levels of the repetition. It was a forced choice among "Clearly aware", "Weakly aware", and "Not aware at all." Only the samples that were denoted "Not aware at all" were used to calculate the test-retest reliability.

5. Evaluation settings

5.1. Hypothesis testing

We tested our core assumption regarding the conditional independence between observer and clip given ratings. The problem here is that each combination of observer and clip has only a single observation. Accordingly, to perform a standard two-way (two-observer-class by two-clip-class) ANOVA, we combined clustering with bootstrapping. We first randomly divided all the observers and clips into two equi-sized groups, and then performed the two-way ANOVA. This was repeated $B = 10,000$ times. The statistical and practical significances were determined by using 95-percent confidence intervals obtained by using B samples.

1. This was done in [9], and required a couple of months.

2. We first determined 97 target rating distributions, and then selected one 6-sec time window that yielded the rating distribution closest to each. In this step, we limited the candidate clips to those where the nonverbal behaviour annotated by 1-3 annotators changed.

5.2. Performance evaluation methods

We evaluated the proposed model using two training strategies. Direct parameter estimation follows a leave-one-out cross-validation scenario, which assumes that full $N_i \times N_j$ data, except for y_{ij}, are available for model training when predicting y_{ij}. Indirect parameter estimation is a more challenging leave-one-subject-and-clip-out cross-validation scenario, where no ratings regarding both target observer i and target clip j are available. $P(y|i)$ and $P(y|j)$ are estimated by using the models described in 3.2.2.

5.3. Performance measure

As a performance measure, this paper reports the accuracy or correct prediction rate. We do not consider that we need to use F-scores, because, as shown in 6.1.1, the rating scores were well balanced. We also use a sign agreement metric (SAGR), a variant of accuracy in binary classification tasks, following [31], which recommend the use of several measures jointly. Moreover, we observed that for example error metrics, e.g. root mean square errors (RMSEs) and mean absolute errors (MAEs), yielded similar results and thus do not alter the conclusion of this paper.

We mainly report accuracies normalized by using test-retest reliability and chance level, as upper and lower bounds. A normalized accuracy of 1 means the performance equals the test-retest reliability, while a value of 0 means the performance is equal to the chance level, which in our case is 0.2.

5.4. Baseline models

We compare the proposed model with two types of baselines: k-nearest-neighbour (k-NN) models and a naïve bias model. The k-NN models are further divided into two models: the k-nearest-clip model and k-nearest-observer model. The k-nearest-clip is based on a common cognitive theory, exemplar theory [32], while k-nearest-observer is related to a well-known collective decision making strategy, WoC [10]. These provide good insight into how to view the results. Also note that these baselines require some rating information regarding observer and clip for parameter estimation, and thus are evaluated in the leave-one-out cross-validation scenario.

The k-nearest-clip model makes a prediction by using the majority voting of k-clips, regarding the target observer, that yielded ratings closest to those of the target clip. This is based on the exemplar theory, which assumes that individuals make categorical judgments by comparing new stimuli with instances already stored in a memory[3]. The k-nearest-observer model makes a prediction by using the majority voting of k-observers whose rating patterns were

3. Strictly speaking, according to the exemplar theory, the clips that were displayed after the clip being targeted should be excluded from the nearest clips. However, this paper uses such future clips by assuming that observers use prior knowledge established during their lives even for the first clip.

closest to that of the target observer. This model assumes that observers who gave similar ratings to some stimuli are also likely to provide similar ratings to other stimuli. The similarities between observers (clips) were measured by using the Euclidian distance between rating vectors with a length of $N_j - 1$, excluding clip j (a length of $N_i - 1$, excluding observer i). We chose the best k for both k-NNs.

When k is the maximum, k-nearest-clips is identical to our model that uses only $P(y|i)$ in Eq. 2 (we call it the $P(y|i)$ only model); both predict $\hat{y}_{ij}$ by using the majority voting of the ratings of observer i to the whole clips, except for clip j. Similarly, k-nearest-observers is identical to our $p(y|j)$ only model with the maximum k; both predict $\hat{y}_{ij}$ by using the majority voting of the all observers, except for observer i, to clip j. Accordingly, in summary we can say that our model exploits both $p(y|i)$ and $p(y|j)$ as they are (in addition to $p(y)$), by providing the same weight to all ratings, while the k-NNs only use one of their binary-weighted distributions based on their similarities.

The naïve bias model assumes that the rating scale is ordinal, and the observer rank, or her percentile on $P(y|j)$, is constant across clips js. For example, it assumes that an observer always gives the highest score among the entire observer group for all clips, while another observer gives the median score in the group. In the parameter estimation stage, each observer's rank was first calculated separately for each clip based on their rating scores; this yielded 97 ranks for each observer. The final observer's rank was determined as the average of 97 ranks. In the rating prediction stage, $P(y|j)$ was first calculated, and then the score at the estimated observer's rank was selected as the prediction.

6. Results

This section reports various validation results, including hypothesis testing, and prediction performance evaluation. All the results supports the validity of our proposed modelling.

6.1. Basic validation

6.1.1. Rating results. First, the proportion of the rating scores, namely $P(y)$, was (.17, .30, .20, .23, .10); it was well, but not perfectly, balanced, and around half the scores were "Similar" or "Slightly similar." Second, the mean labels output by the observers were reliable [33]; ICC(1,k)=.99, ICC(A,k)=.99, and ICC(C,k)=.99. The average correlation between a pair of observers was not high. The percent agreement, which corresponds to accuracy (our prediction measure), without distinguishing between ground truth and prediction, was .35, and Conger's kappa, a corrected agreement value, was .17. ICC(1,1)=.42, ICC(A,1)=.42, ICC(C,1)=.45. These match the results of previous studies, e.g. [9].

6.1.2. Test-retest reliability. Table 1 summarizes the results of a repeatability test. The values were obtained separately

TABLE 1. REPRODUCIBILITY TEST RESULTS WITH AWARENESS LEVEL

	Awareness level		
	Not at all	Weak	Clear
Proportion	.67	.28	.05
Reliability	.525	.417	.563

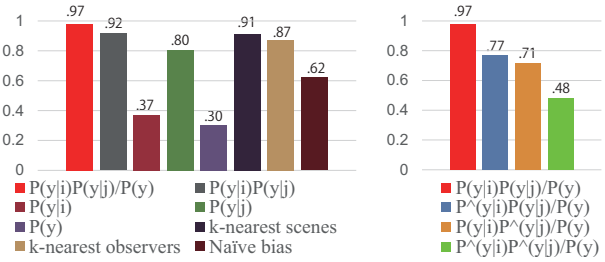

Figure 2. Normalized accuracy comparison: left) models in direct parameter estimation scenario, and right) models with indirect parameter estimation scenario. Vertical axes indicate normalized accuracy. The red and light green bars indicate our model in direct- and indirect-parameter-estimation scenarios, respectively.

for the awareness levels that they provided in the post questionnaire. The reliability was .525.

These results suggest the inherent difficulty of inferring affective ratings. The mean pairwise correlation (Pearson's r), among the observers who answered "Not aware at all" to all three retest clips and gave at least two ratings to the clips both in the test and retest sessions, was .49 ($N = 50$). This value is comparable to that reported in the literature, e.g. [34].

6.2. Evaluation of our modelling

6.2.1. Evaluation via hypothesis testing. The bootstrapping revealed that there was no interaction between observer class and clip class; the 95% confidence interval was $F(1, 9696) < 2.1$, $p > .15$, $\eta^2 < 0.001$. Even in the worst case, namely in the grouping that maximizes F-statistic, the interaction was not practically significant, although it was statistically significant[4]; $F(1, 9696) = 10.2$, $p = 0.0014$, $\eta^2 = 0.0011$. Thus, we can conclude that our independence assumption is valid. This is further supported by the performance evaluation in 6.2.2.

6.2.2. Evaluation of direct parameter estimation scenario. The left hand side of Fig. 2 shows the prediction performance of each model in a direct parameter estimation scenario. Strikingly, our model also achieved comparable results, namely a normalized accuracy of .98. Moreover, all three terms of our model in Eq. 2, contributed to the high performance, although the most powerful term, $p(y|j)$, achieved a normalized accuracy of .80. The good performance of k-nearest clips does not contradict the exemplar

4. The statistically significant but practically non-significant difference is not a surprising result given our large sample size, namely $N_i \times N_j \approx 10,000$ samples.

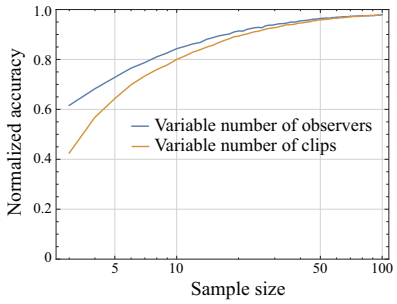

Figure 3. Performance curve for different training sample sizes. Horizontal axis indicates sample size, the number of observers (blue line) or the number of clips (orange line), and vertical axis shows normalized accuracy.

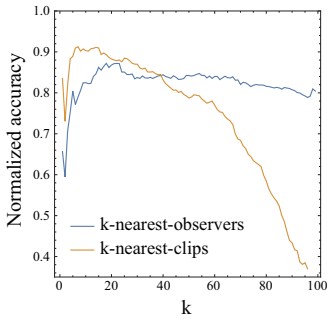

Figure 4. Accuracy curve of k-nearest neighbours for different ks

TABLE 2. MODEL COMPARISON ON BINARY CLASSIFICATION TASK

Model	SAGR [31]
Our model	.88
CUBAM (crowdsourcing aggregation) [17]	.81

theory. However, our model outperformed the baselines. The low performance of the naïve bias model does not suggest that the observers' ratings were simply biased toward the positive or negative side on the rating scale.

Figure 3 shows the impact of the training sample size, the number of observers N_i and the number of clips N_j, on our model[5]. The results basically follow our expectation. First, more is better, but 100 observers and 97 clips both look reasonably sufficient. Second, the curves almost converges at around the normalized accuracy of 1, the test-retest reliability. These results support both our models, and particularly the conditional independence assumption.

Figure 4 shows the impact of the size of k on k-NN performance. k-nearest-observers are largely insensitive to the k sizes over 20. On the other hand, k-nearest-clips clearly hit the peak at $k = 7$, and it outperformed the best k-nearest-observers. But, the performance suddenly deteriorated with larger ks.

Table 2 compares the prediction performance (SAGR) of our model with that of a well-known crowdsourcing aggregation method [17], which models each annotator's classification plane and has achieved good performance levels in various tasks. Our model outperformed it. Moreover, because the crowdsourcing aggregation method assumes binary classification, we converted the original prediction results of our model, obtained by Eq. 1, into binary labels with a simple thresholding technique; we considered the output to be '1' if the estimated score was +2 or +1, or '0' otherwise. For the candidate method, we also applied the same leave-one-out cross validation scenario, and we chose the best model dimensionality (two-dimensional model)[6].

6.2.3. Evaluation of indirect parameter estimation scenario. The right hand side of Fig. 2 shows the prediction performance of our model in the indirect parameter estimation scenario. The accuracies of the $\hat{P}(y|i)P(y|j)/P(y)$ model and the $P(y|i)\hat{P}(y|j)/P(y)$ model were comparable; they realized a normalized accuracy of around .75. When combined together, $\hat{P}(y|i)\hat{P}(y|j)/P(y)$ achieved the normalized accuracy of around .50. We consider this value promising, because this is a challenging task.

6.2.4. Individual differences. Figure 5 compares the observer-wise prediction accuracy of our model with those of the $P(y|i)$ only model and k-nearest clips. Our model outperformed or at least was comparable to these models for most observers. However, a few observers fitted better with either or both of these candidate models. Six observers reported approximately .1 higher (ranging between .10 and .13) accuracies for the $P(y|i)$ only model than for our model, while four observers reported .1 higher (ranging between .10 and .29) accuracies for k-nearest-clips; one observer was included for both. They suggest that these observers were more likely to determine their ratings to some clips without considering clip information ($P(y|i)$ only model) or with reference to similar clips (exemplar theory). Moreover, regarding the remaining models, the maximum accuracy difference from our model was smaller than .1. Although there is no clear reason for the threshold value of .1, we believe that this is meaningful in terms of grasping the rough characteristics of our model.

6.2.5. Confidence value. We found that there is a strong correlation between an individual's prediction accuracy with our model and the proportion of the WoC ratings that she gave; $r(98) = .73$, $p < .001$, as shown on the left hand side of Fig. 6. The proportion means how often the observer's rating matched the rating given by majority voting by the crowd or that given by the $P(y|j)$ only model. It means that the proportion of WoC ratings given can be used as an indicator of confidence value for the observer.

5. The curves were obtained with the procedures described below. First, we generated 2,000 bootstrap observer/clip sets randomly drawn from all the observers/clips excluding the target observer/clip to be inferred. Then, we calculated the performance by using each of these 2,000 sets. Finally, we averaged their performances.

6. This evaluation would be unfair for [17], because it does not take advantage of M levels of the ratings. However, we consider this comparison is sufficient to demonstrate the validity of the proposed modelling approach.

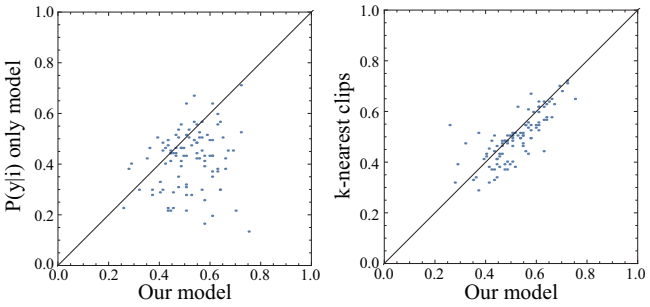

Figure 5. Unnormalized accuracies of our model (horizontal) versus left) $P(y|i)$ only model and right) k-nearest clips (vertical), respectively. Dots indicate observers.

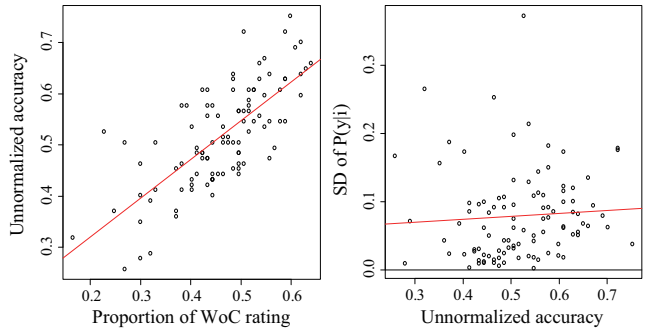

Figure 6. Left) Accuracy of our model strongly correlates with the proportion of given WoC ratings, and right) the prediction accuracy is independent of the standard deviation of $P(y|i)$. Dots indicate observers.

This is not a obvious result, because the proportion giving WoC ratings did not correlate with the standard deviation of $P(y|i)$, which is a reasonable indicator, taken from information theory, of how easily the observer's ratings can be predicted[7] ($r(98) = .47$, $p < .001$); $r(98) = .04$, $p = .73$, as shown in the right hand side of Fig. 6. Moreover, as an extreme case, if observers whose $P(y|i)$ distributions are identical to $P(y)$, our model assume that their ratings are always identical to those of WoC, because Eq. 2 is now $\hat{y}_{ij} = \arg\max_y P(y|j)$, our definition of the WoC rating. This would partly explain the reason for the high correlation between the prediction accuracy of our model and the proportion giving WoC ratings.

6.2.6. Validity of test-retest reliability. Our aforementioned claims using normalized accuracy rely on the test-retest reliability, which was calculated simply by using three clips. We thus further provide evidence of its validation.

The following results support the view that the test-retest reliability was not underestimated, which means the normalized accuracy of our model was not overestimated. Scene-wise unnormalized accuracy strongly correlated with the standard deviation of $P(y|j)$, $r(95) = .88$, $p < .001$, as

7. For example, a larger SD or a peaky distribution means limited rating scores were used more often than other scores, and thus it was easier to achieve a more accurate prediction.

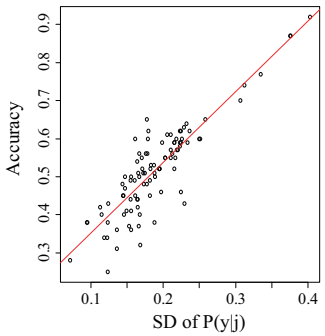

Figure 7. Standard deviation of $P(y|j)$ strongly correlates with the prediction accuracy of our model. Dots indicate clips.

shown in Fig. 7. The regressed line was accuracy = 0.17 + 1.85×sd. This estimates a mean accuracy of 0.47 for the three clips whose mean sd was 0.16, which is close to the obtained unnormalized accuracy for the three clips of 0.48. Accordingly, we conclude that we could be sufficiently confident about the obtained test-retest reliability.

7. Discussion

We have provided various pieces of evidence supporting our modelling approach. However, several issues still remain.

Firstly, the cognitive targets for which our modelling works well remain unclear. Our model is so general that a variety of discretized perception tasks are applicable, e.g. categorical and Likert-scale descriptions of perception. Some tasks might well match our modelling, but others might not. The applicability might depend on the difficulty of the task; for example, the relationship between the goodness of fit and the task difficulty might be linear or U-shaped. With the former we can assume a more difficult task, and the interaction between observer and clip becomes weaker or stronger. The latter assumes that the interaction is stronger only for moderately difficult tasks.

Secondly, this study predicted the rating score that a target observer will give to a target clip. This can be seen as a variant of choice behaviour, which differs from cognition or judgment itself [35]. We can expect various factors, for example, sensitivity to the intensity of emotional expressions [36], and the response style for simplifying tasks, to be mixed in our observer term. Separating these factors is outside of the scope of this paper, and will pose a challenging but interesting problem, where physiological signals would be required, like felt-emotion studies [3], [4], [5].

Thirdly, this study employed a discrete annotation procedure for both time and emotion space. To apply our work to continuous annotations, as with the recent trend in the affective community, e.g. [13], [14], our model must be extended, for example, by approximating the observer and clip distributions $P(y|i)$ and $P(y|j)$ with parametric functions, and introducing observer-specific delays, as in [37], [38], [39]. Moreover, it would be interesting to investigate

whether or not there is an interaction between observer and clip in continuous annotation scenarios.

8. Conclusion

This paper proposed a computational model that predicts the idiosyncratic perception of others' emotions by focusing on the observer effect and the video-clip effect. We modelled these terms separately based on Bayes' theorem on the assumption of their conditional independence. In a leave-one-out cross-validation scenario, our experiment with 100 observers and 97 clips demonstrated that 1) there is in fact no statistically and practically significant interaction between observer and clip, and 2) our Bayesian modelling achieves a 97 percent accuracy as a reference of test-retest reliability. Furthermore, in a more challenging, leave-one-subject-and-clip-out scenario for unknown observers and clips where no rating information was available, our model promisingly yielded an accuracy of around 50 percent. Using deep neural networks for both component models would be helpful in filling the gap between the performances of these two scenarios.

References

[1] D. Gatica-Perez, "Analyzing group interactions in conversations: A review," in *Proc. IEEE Int'l Conf. MFI*, 2006, pp. 41–46.

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. PAMI*, vol. 31, no. 1, pp. 39–58, 2009.

[3] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, 2012.

[4] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, 2012.

[5] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Comp. Intell. Magazine*, vol. 8, no. 2, pp. 20–33, 2013.

[6] D. A. Kenny and L. Albright, "Accuracy in interpersonal perception: A social relations analysis." *Psychol. Bull.*, vol. 102, no. 3, pp. 390–402, 1987.

[7] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources And Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[8] M. Nicolaou, H. Gunes, and M. Pantic, "Output-associative RVM regression for dimensional and continuous emotion prediction," in *Proc. IEEE FG*, 2011.

[9] S. Kumano, K. Otsuka, D. Mikami, M. Matsuda, and J. Yamato, "Analyzing interpersonal empathy via collective impressions," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 324–336, 2015.

[10] J. Surowiecki, *The Wisdom of Crowds*. New York: Anchor, 2005.

[11] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, no. 1-2, pp. 227–256, 2003.

[12] K. P. Truong, M. A. Neerincx, and D. A. V. Leeuwen, "Assessing agreement of observer- and self-annotations in spontaneous multimodal emotion data," in *In Proc. Interspeech*, 2008, pp. 318–321.

[13] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "FEELTRACE: An instrument for recording perceived emotion in real time," in *ISCA Workshop on Speech and Emotion*, 2000, pp. 19–24.

[14] G. N. Yannakakis and H. P. Martínez, "Grounding truth via ordinal annotation," in *Proc. ACII*, 2015, pp. 574–580.

[15] R. Cowie, "Describing the emotional states expressed in speech," in *ITRW on Speech and Emotion*, 2000, pp. 11–18.

[16] H. Meng, A. Kleinsmith, and N. Bianchi-Berthouze, "Multi-score learning for affect recognition: the case of body postures," in *Proc. ACII*, vol. 1, 2011, pp. 225–234.

[17] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds," in *Neural Information Processing Systems Conference (NIPS)*, 2010, pp. 2424–2432.

[18] A. Rui, O. Martinez, X. Binefa, and F. Sukno, "Fusion of valence and arousal annotations through dynamic subjective ordinal modelling," in *Proc. IEEE FG*, 2017 2017.

[19] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," *SIGKDD Explor. Newsl.*, vol. 17, no. 2, pp. 1–16, 2016.

[20] B. Zhang, G. Essl, and E. Mower Provost, "Automatic recognition of self-reported and perceived emotion: Does joint modeling help?" in *Proc. ACM ICMI*, 2016, pp. 217–224.

[21] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Appl. Statist.*, vol. 28, no. 1, pp. 20–28, 1979.

[22] T. L. Griffiths, C. Kemp, and J. B. Tenenbaum, "Bayesian models of cognition," in *Cambridge handbook of computational cognitive modeling*, R. Sun, Ed. Cambridge: Cambridge University Press, 2008, pp. 59–100.

[23] H. Zhang, "The optimality of naive Bayes," in *Proc. Int'l Florida Artif. Intell. Res. Soc. Conf.*, V. Barr and Z. Markov, Eds. AAAI Press, 2004.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[25] S. Kumano, K. Otsuka, M. Matsuda, R. Ishii, and J. Yamato, "Using a probabilistic topic model to link observers' perception tendency to personality," in *Proc. ACII*, 2013, pp. 588–593.

[26] S. Kumano, K. Otsuka, M. Matsuda, and J. Yamato, "Analyzing perceived empathy based on reaction time in behavioral mimicry," *IEICE Trans. on Information and System*, vol. E97-D, no. 8, pp. 2008–2020, 2014.

[27] S. D. Preston and F. B. de Waal, "Empathy: Its ultimate and proximate bases," *Behavioral and Brain Sciences*, vol. 25, no. 1, pp. 1–20, 2002.

[28] A. Smith, "Cognitive empathy and emotional empathy in human behavior and evolution," *The Psychological Record*, vol. 56, no. 1, pp. 3–21, 2006.

[29] H. Walter, "Social cognitive neuroscience of empathy: Concepts, circuits, and genes," *Emotion Review*, vol. 4, no. 1, pp. 9–17, 2012.

[30] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: A meta-analysis." *Psychological Bulletin*, vol. 128, no. 2, pp. 203–235, 2002.

[31] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vision Comput.*, vol. 31, no. 2, pp. 120–136, 2013.

[32] D. L. Medin and M. M. Schaffer, "Context theory of classification learning," *Psychological Review*, vol. 85, no. 3, pp. 207–238, 1978.

[33] K. L. Gwet, *Handbook of Inter-Rater Reliability (3rd Edition)*. Gaithersburg, MD: Advanced Analytics, LLC, 2012.

[34] T. Bänziger, D. Grandjean, and K. R. Scherer, "Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (MERT)," *Emotion*, vol. 9, pp. 691–704, 2009.

[35] R. Tourangeau, L. C. Rips, and K. Rasinski, *The psychology of survey response*. Cambridge: Cambridge University Press, 2000.

[36] P. Juslin and P. Laukka, "Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion," *Emotion*, vol. 1, no. 4, pp. 381–412, 2001.

[37] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *Proc. ICMI*, 2012, pp. 501–508.

[38] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. IEEE FG*, 2013, pp. 1–8.

[39] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Trans. Affect. Comput.*, vol. 6, no. 2, pp. 97–108, 2015.