How do we perceive the intensity of facial expressions? The PIFE dataset for analysis of perceived intensity

Marina Tiuleneva
DiPSCO
University of Trento
Rovereto, Italy
mari.tyuleneva@gmail.com

Emanuele Castano
DiPSCO
University of Trento
Rovereto, Italy
ICST, CNR
emanuele.castano@unitn.it

Radoslaw Niewiadomski
DIBRIS
University of Genoa
Genoa, Italy
radoslaw.niewiadomski@unige.it

***Abstract*—Decoding the intensity of facial expressions is of primary importance for humans. Modeling this computationally, however, is not an easy task. Here, we propose a new dataset composed of circa 400 videos and 1,000 images automatically extracted from several movies, and rated by humans on intensity. Each stimulus presents facial expressions of one person only, but overall, the stimuli represent a large variety of expressions in individuals of different age, gender, and ethnicity, in fictional yet natural movie settings. Each video was rated by 5 people in terms of perceived intensity and variability using a 7-point Likert scale; each image was rated by 5 people only for intensity. In total, 90 people participated in the ratings, and the average inter-rater ICC agreement is 0.63 for videos and 0.66 for images. For each video and image we also extracted intensity values on 15 action units using the OpenFace software.**

We report results for both human and computer-assisted intensity ratings, and propose a baseline regression model capable of estimating the perceived intensity in images and videos with a mean squared error of 0.74. We conclude our paper by discussing potential applications of a general computational model of perceived intensity.

***Index Terms*—facial expressions, dataset, intensity, perceptive study, regression**

I. Introduction

The intensity of human facial expression is an important cue that humans use to understand others and thus predict their behavior. Several attempts to computationally model intensity have been proposed [11], [17], [37], [40], [43]. Most of these approaches focus on identifying different phases within one facial expression (such as the apex or onset) and analyzing the temporal profile of intensity changes. These computational approaches are often fine-grained, focusing on single moments (e.g., single image frames) of an expression, the activation of specific action units (AUs), or basic expressions. While they are certainly appropriate when working with the precisely measured activation of certain facial muscles (e.g., through motion capture technology, facial landmarks, or EMG), they might not be optimal for modeling *the perception* of dynamic human facial expressions. In daily life, humans evaluate facial expressions in a holistic manner (e.g., "expression X is very intense") or compare two expressions (e.g., "expression X is stronger than Y"). We thus propose a more holistic approach to modeling perceived facial intensity, defined as the global (or overall) intensity attributed by (naive) observers to a facial expression. A facial expression, for the purpose of this work, is defined as any visible movement or positioning of the facial muscles that conveys information, including emotions, social attitudes, intentions, cognitive processes, communication feedback, cultural signals, and personality traits.

Our approach consists of providing one global descriptor to the whole expression without considering expression phases. With this approach, we address both theoretical and practical questions: First, how is *perceived* intensity related to the physical (i.e., measurable) activity of single facial features (e.g., AUs) or their combinations (RQ1)? Second, is it possible to build a general computational model to estimate the perceived intensity of any expression (RQ2)?

The lack of suitable datasets presents the first challenge in addressing these questions. In existing datasets, intensity is typically considered locally and within a specific context [1], [31]. As a result, facial expression variability is rather low. Additionally, the existing works largely ignore the fact that facial intensity can be considered in contexts other than emotion displays, such as expressions of dominance [6] or politeness [5]. Consequently, in this paper, we propose a novel dataset dedicated to intensity of facial expressions: the Perceived Intensity of Facial Expressions (PIFE) dataset. THE PIFE consists 409 short (ranging from 3 to 5 seconds) video segments and 998 images, extracted from several movies. Each video and image was rated by humans for intensity of the facial expression on a scale of 1 to 7. Additionally, for all segments, we extracted information about the action units activation using freely available software.

In relation to our research questions, previous studies have

This work supported by the PRIN 2022 project COCOA (D53D23008850001), funded by the Next Generation EU (NGEU) Programme, National Recovery and Resilience Plan (PNRR) and by the Italian Ministry of University and Research (MIUR).

analyzed whether perceived intensity varies linearly with the physical activation of facial features [16], [29]. However, these studies are limited to expressions of emotions, and their results are contrasting (see Section II-A). Additionally, if such a linear relation existed, then designing computational models for facial intensity recognition should be relatively straightforward. However, existing studies on this topic, especially recent works using complex deep learning architectures, suggest that the relationship between perceived intensity and measurable physical activation of facial features is likely to be more complex.

The present work contributes to the literature on automated assessment of facial expression, using a large variety of facial expressions collected from a broad spectrum of individuals, which are not manipulated (e.g., by morphing methods).

II. Related Works

A. *Perceived intensity and Physical Features of Facial Expressions*

The relationship between the physical (or measurable) features and perceived intensity of facial expressions was assessed by Hess et al. [16] for expressions of anger, disgust, sadness, and happiness using a set of pre-selected manipulated images. The results revealed that perceived intensity varied linearly with the intensity of the expression. Importantly, in that study, the stimuli were artificially modified using a morphing technique, which does not ensure that the stimuli correspond to real expressions. Indeed, Becker et al. [4] highlighted issues with stimuli generated using such techniques. They found that real video recordings were perceived to have greater emotional intensity than corresponding dynamic morphs.

In another study, 250 spontaneous laughter video segments (and corresponding facial MoCap data) recorded in natural settings were rated in terms of intensity using a 5-point Likert scale [25], [29]. The results indicated that perceived intensity was strongly correlated only with the measured (using MoCap) intensity of AU25/AU26, but not with other analyzed action units (i.e., AU6, AU12, AU4).

B. *Relevant datasets*

A detailed overview of the datasets for studying intensity was recently proposed by Mehta et al. [22]. One of the most frequently used datasets is the Cohn-Kanade (CK+) dataset [19], which contains both images and videos of seven basic emotions and action units (AUs). It has often been used to develop intensity models, e.g., [9], [17], [34], [43]. The BU-3DFE database [41] contains videos of basic emotions performed by 101 participants. Each expression starts and stops on a neutral face, allowing for easy comparison between several frames of the same expression. The expressions are annotated with four intensity levels. The BP4D [42] contains spontaneous displays of eight affective states elicited in lab conditions and includes FACS annotations with intensity levels (for certain AUs only). It was used for the Facial Expression Recognition and Analysis Challenge (FERA 2015) [39] and in many research papers (e.g., [40]). The MMI dataset [31] consists of images and videos of 19 subjects performing posed and spontaneous expressions of basic emotions and specific AUs. The faces in the videos start neutral, change to expression apex, and then return to neutral. The analysis of AU temporal activation patterns (onset, apex, offset) is also provided. MMI was used in several works, e.g., [2], [34]. Similarly, the MUG dataset consists of image sequences of posed and induced basic facial expressions, used in studies such as [34]. In [21], 27 young adults were video-recorded by a stereo camera while they watched online video clips intended to elicit spontaneous emotional expressions. Each video frame was manually coded for the presence, absence, and intensity of facial action units using ordinal scores ranging from 0 (not present) to 5 (maximum intensity). This dataset was used, among others, in [40]. Differently, Dhall et al. [11] focus on group happiness intensity. They propose the HAPPEI database, composed of images posted online and labeled with group-level mood intensity from ‘neutral’ to ‘thrilled’. Additionally, 8,500 single faces from this set were annotated in terms of happiness intensity, occlusions, and poses. A study was conducted to identify which factors influence how people perceive the group’s happiness (e.g., the number of people, the distance between them, occlusions).

C. *Modeling Intensity of Facial Expressions*

One of the early approaches for intensity estimation was proposed by Delannoy and McDonald [10], who considered it a multi-class classification problem and employed one-against-all Support Vector Machines with three degrees of intensity: low, medium, and high. Chang et al. [9] introduced a framework that combines low-level image features with ordinal regression to estimate expression intensity of single images. Similar to Delannoy and McDonald’s paper [10], three levels of intensity are also considered in [9], but this approach takes into account the relative order between the images.

Zhao et al. [43] also applied ordinal regression to estimate frame-level expression intensity by exploiting intensity labels given by human raters to selected frames. Features are extracted from facial landmark points, local binary patterns, and Gabor wavelet coefficients. In addition to the standard databases containing expressions of basic emotions, the model was also tested on the UNBC-McMaster shoulder pain dataset [20].

Kamarol et al. [17] present a framework for facial expression recognition and intensity estimation using a combination of kNN, weighted voting, and Hidden Markov Models. The algorithm classifies a sequence of features (corresponding to a video) into an expression class and quantifies the intensity of the sequence as the expression changes from neutral to apex. It was tested on a standard dataset of basic emotions [19]. Recently, more complex deep learning architectures for AU intensity estimation have been proposed. For example, Walecki et al. [40] explored Convolutional Neural Networks.

Beyond machine learning methods, Uddin and Canavan [37], [38] propose a set of mathematical formulas for quantifying spatial and temporal expressiveness based on action

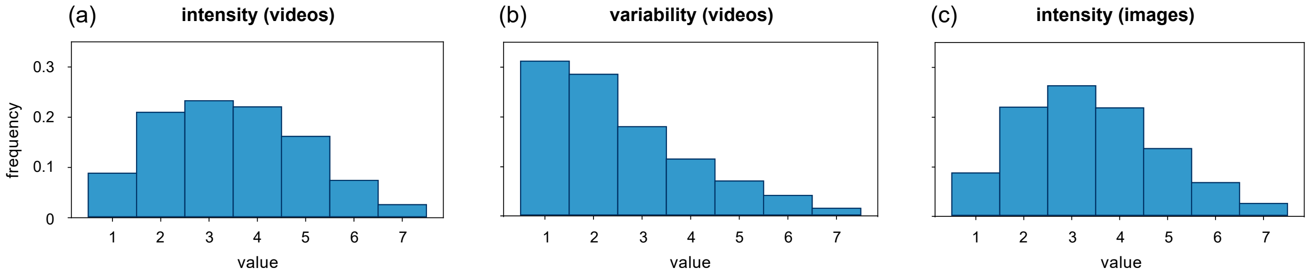

Fig. 1. The labels frequency for: a) videos intensity, b) videos variability, c) images intensity. The frequency is normalized to 1 on Y axis.

units data, landmarks, head pose, and gaze. For instance, in [37] they computed a bounded score at the video frame level, giving more importance to action units with high intensities and large temporal changes.

III. PIFE Dataset

The PIFE comprises of video segments and images from movies. Movies are often used to build datasets for affective computing, both as stimuli to elicit affective reactions [23], [24], as well as sources of nonverbal behavior data, e.g., for emotion classification [12]. They are an easily accessible resource providing a wide variety of expressions related to emotions and other internal states, expressed by a large number of individuals. While there may be differences between spontaneous and fake (or "acted") expressions [14], for our present purposes this is not a major concern. Another reason for using movies is that the montage techniques employed in movie production lead to scene segmentation, where actors' meaningful facial activity is shown in a single shot. Especially in fiction movies, there are not many casual or unrelated expressions.

Using movies allows us to capture a variety of facial expressions that are representative of real life. This contrasts with typical data collections in laboratory settings (e.g., [1], [19], [31]), which are limited to a small number of emotions. The work most similar to ours is by Dhall et al. [12], which introduces the AFEW (video) and SFEW (image) datasets of affective displays derived from over 50 movies using a semi-automatic recommender system and analyzing movie subtitles. The data was annotated in terms of (basic) emotions, and characteristics of the actors such as gender, age. To the best of the authors' knowledge, AFEW and SFEW were not evaluated in terms of perceived intensity. We believe that the characteristics discussed above make our dataset suitable for studying the perception of intensity of facial expressions and for developing computational models to estimate facial intensity.

Twelve movies with an average resolution of 736 × 363 and a mean duration of 116 minutes were used to extract the stimuli presented to human raters (minimum height and width: 640 × 308, maximum height and width: 858 × 464, minimum duration: 94 minutes, maximum duration: 152 minutes). The movies were processed using OpenFace 2.0 [3], a widely used tool in the research community (recent examples include [13], [15], [18], [27]) to extract information about facial activity and to divide movies into video segments.

A. *Choice of video segments*

The movie segments, where a human face is detected by OpenFace, were automatically extracted by comparing the distances between detected faces in consecutive frames following to the procedure described in [36]. Only the frames that were successfully (i.e., the value of success given by OpenFace is 1) analyzed by OpenFace were taken into consideration. The initial pool contained segments with multiple faces, where certain segments A_i and A_j may overlap, if face F_a is tracked in A_i and face F_b is tracked in A_j simultaneously. Next, the following selection criteria were applied to the initial pool:

- Duration: Segments lasting a minimum of 3 and a maximum of 5 seconds were retained. This was done to ensure a minimum of information for effective rating by humans, while at the same time limit the duration so that a segment does not contain multiple expressions and multiple sources of variability.
- Face display: Segments featuring more than one face were excluded to simplify the ranking task and mitigate ambiguity.
- Dimension: To ensure clear visibility for human raters, segments where a human face covers less than 20% of the frame were ruled out.

The script written in Python was used to select automatically the segments that meet the above mentioned criteria resulting in 434 segments. From this pool, we removed manually segments containing unrealistic deformations created with special effect techniques and masked faces, as well as the segments that were erroneously extracted. The latter occurs when two different faces appear in exactly the same position in two consecutive frames due to movie montage. The final dataset consists of 409 video segments.

Each video was edited to start and finish with a white frame. The audio was removed from all the videos. It should also be noted that some of the segments contain the actors speaking. This is an intentional decision, as facial activity during speech (e.g., lip movement) can influence the perception of intensity (e.g., strong visible articulation). Thus, in our view, these movements contribute to their perceived intensity.

B. Choice of images

Nearly a thousand single frames were extracted from the same set of movies, ensuring that the same number of images was taken from each movie. Each image depicts just one person's face. The images were selected randomly, but our intention was to balance the number of potentially high- and low-intensity expressions in the set. For this purpose, we employed very simple heuristics to estimate the facial intensity. Specifically, we computed, for each image, the average intensity of action units (AUs) using the AU intensity values extracted with OpenFace. Subsequently, we divided the images into four intensity quartiles and randomly selected the same number of images from each quartile. It's important to note that while this approach may not accurately simulate human perception, it provides a simple way to ensure that stimuli exhibit a certain variety in terms of facial intensity. Similarly to what we did for the videos, for images we also excluded manually some frames with highly deformed faces. After removing these images, we repeated the selection procedure to balance the number of stimuli per movie. A total of 998 images were retained.

IV. Rating study

409 videos and 998 images were rated in the online study organized on the Qualtrics platform. Each stimulus was followed by three questions for the video part and two questions for the image part. First, the participants rated the perceived intensity and variability of the facial expressions using two Likert scales from "very low" to "very high."

The second variable was added because we anticipate that this information may be useful for developing computational models in the future (see Figures 2 and 3). Finally, participants were also asked to specify the gender of the portrayed person. Their answers to the third question were not used for any data analysis; it was included to motivate participants to watch the stimuli attentively.

Videos were divided into 8 and images into 10 nearly equally sized sets. Each person was asked to rate only one out of 18 sets. Thus, each participant was assigned either videos or images but not both. This division was primarily made to ensure the task could be completed within a reasonable amount of time, typically around 20 minutes in practice. The stimuli within each set were presented in random order, and the sets were randomly assigned to the raters. Raters were allowed to replay a video segment multiple times before answering the questions. Once the rater submitted their answers, they were not allowed to change them. The stimuli were presented one per page in a standard web browser.

After expressing interest, participants were forwarded to the Qualtrics page. Before starting the task, they participated in a short online training phase. During that phase, they were given the following definition of intensity:

> *We refer to the strength or clarity with which signals are conveyed through facial movements. "Intense" in this context refers to the strength, vividness, or prominence of the expressions displayed on individuals' faces.*

The variability was explained as follows:

> *We refer to the diversity or range of facial expressions observed in the given video. In this context "variable" implies the degree of differences or variations in facial expressions displayed by individuals in the video.*

They were also asked to test the interface, rate the a sample of stimuli and answer a sample of attention questions.

To detect careless responding we included four attention questions per questionnaire. In the video survey, participants were presented with a white number against a black backdrop and were required to adjust both sliders to match this number. In the image survey, the attention questions featured images with more than one face, and participants had to select using a slider a number corresponding to the number of seen faces.

No identifying information was collected to ensure anonymity and participants were instructed that they could interrupt the task whenever they wanted without consequences.

After rating all the stimuli in the pool, participants were asked to provide basic demographic data and indicate how many movies and actors they recognized. To answer these two questions, they had to choose one out of five labels ranging from *I do not recognize any* to *All of them*.

V. Results

In the video rating study, 44 human annotators participated; 39 volunteers and 15 recruited via Prolific (an online platform for running surveys) and paid £5. Four participants were excluded due to failing more than one attention question. In the image rating study, 52 raters participated (25 were recruited via Prolific and paid £5) - two participants were excluded due to failing more than one attention question.

Figure 1 shows the distribution of video intensity and variability ratings, and of images intensity ratings. Both intensity ratings are fairly normally distributed, while variability ratings (of videos) shows a skewed distribution, suggesting that the video set contains just a few videos that display high variability in facial expressions. This is partially due to the fact that the upper limit for duration was defined as 5 seconds - thus limiting variability.

To check the inter-rater agreement, we used the inter-class correlation *ICC2k* implemented in Python [32]. We applied it separately to each stimuli group (8 video groups and 10 image groups), as all stimuli within one group were rated by the same 5 raters. For the images, the ICC values range from 0.373 to 0.866, with an average of 0.664. For videos, the ICC range for intensity is from 0.464 to 0.757, with an average of 0.63, and the ICC range for variability is from 0.555 to 0.813, with an average of 0.7. We conclude that the inter-rater agreement for videos' variability is good, while the agreements for video and image intensity are slightly lower but still acceptable.

Regarding the post-rating questionnaire, Table I presents the raters' answers. It can be seen that most of the participants

TABLE I
THE RESPONSES TO THE POST-RATING QUESTIONS CONCERNING FAMILIARITY WITH MOVIES AND ACTORS.

Label	Videos		Images	
	movies	actors	movies	actors
I did not recognise any	0.049	0	0.034	0.034
A few	0.463	0.244	0.448	0.310
Several	0.293	0.195	0.328	0.362
Many	0.170	0.512	0.172	0.293
All of them	0.024	0.049	0.017	0

recognize some movies and many (but not all) actors. At the current stage of this research, we do not utilize this information further in this paper.

The dataset consists of video clips with 167 female and 241 male faces, with a slightly skewed distribution favoring males. The inter-rater reliability, measured by Fleiss' Kappa, averages at 0.9505, indicating high agreement, with 12 videos where only one participant answered 'I cannot tell'. While this question was used to encourage the raters to watch the stimuli and we did not aim to make any conclusions the "correctness" of their answers, we note that high agreement likely indicates that the raters did indeed view the segments, suggesting that the rating procedure was successful.

VI. Data analysis

To address RQ1, the action units' intensities were extracted from all the video segments and images using OpenFace. For the videos, the extracted AU intensities were subsequently filtered using the Savitzky–Golay filter (see Figures 2-3).

A. Temporal Profiles of AU-Computed vs. Perceived Intensity

The first stage is to compare time series for certain segments that are similar in terms of perceived intensity. Figure 2a presents the maximum and average intensity detected by OpenFace for all action units in four segments rated by humans as highly intense and highly variable, while Figure 2b presents the same plots for four segments rated as highly intense but with low variability. In Figure 2b, it can be seen that the activation of action units is more or less constant throughout the entire segment. However, the plots in Figure 2a show variation in facial activity within segments. This observation is consistent with the reported variability values.

Next, we analyze intensity scoring separately for the three subsets corresponding to eyebrow activity (AU1, AU2, and AU4), middle face activity (AU5, AU6, AU7, and AU9), and mouth area activity (see Figures 3a and 3b). It can be seen that the mouth area often displays much more variability than the other two, which is probably caused by the fact that the person in the video is speaking.

B. Regression

To demonstrate the feasibility of modeling perceived intensity using the PIFE dataset (addressing Research Question 2), we conducted a series of experiments utilizing standard regression techniques, including Support Vector Regression (SVR), Random Forests (RF), Ridge Regression (RR), and Multilayer Perceptron (MLP). At this stage of the research, our objective is not to find the optimal solution, but rather to propose a baseline model.

We trained a set of regressors for images and videos separately. For images, we used the 15 AU intensity values extracted by OpenFace and the average ratings of perceived intensity. For the videos, the same information was extracted from each frame of a segment, and then the intensity values were aggregated by computing the 1) average and 2) maximum. We applied the StandardScaler [35] to the feature vector to match it with the 1-7 interval used in the ranking study. We compared the results with simple methods that compute the average and maximum intensity of all action units for each frame, and then aggregate them using the maximum and average, giving four additional methods in total (for images, we just compute average and maximum without applying aggregation). Optuna [30], an open-source hyperparameter optimization framework, was used to find the optimal parameters. For MLP, the best parameters were chosen by considering the following: $activation \in [tanh, relu], solver \in [sgd, adam], \alpha \in [1e-5, 1e-1], learning\_rate\_init \in [1e-5, 1e-1], n\_units\_l \in [10, 500], n\_layers \in [1, 10]$. For SVR, the best parameters were chosen by considering the following: $C \in [1e-3, 1e3], \epsilon \in [1e-4, 1e-1], \gamma \in [scale, auto], kernel \in [linear, poly, rbf, sigmoid]$.

Five-fold cross-validation was applied to each model, and average results in terms of mean square error (MSE) and mean absolute error (MAE) are reported in Table II (intensity scores are in the range 1-7). The best results were obtained with SVR using a linear kernel and $C = 0.009, \epsilon = 0.74$ (videos) and $C = 0.003, \epsilon = 0.004$ (images). However, in general, differences between different regression methods are small.

C. Discussion

The results show that simple averaging or taking the maximum of AUs is not sufficient to estimate the perceived intensity (see Table II, rows 11-14). This can also be seen in Figure 4b, where the correlations between the average and maximum of AUs and perceived intensity in the case of images are low (0.16 and 0.24, $p < 0.01$). All other approaches that result in more complex models using AUs data, rather than simple arithmetic operations, show much better results.

In relation to RQ1, these results indicate that the perceived intensity of facial expression is a complex phenomenon to model and cannot be accurately expressed using the maximum or average of measurable local facial activity. At the same time, the differences between various models are very small (see Table II, rows 1-10). In particular, standard linear regression (see Table II, rows 5 and 10) obtains similar results in terms of MSE and MAE compared to more advanced methods. Additionally, Figure 4a shows that the best model produces relatively small errors for low and middle-range intensities but larger errors for extreme values of perceived intensity.

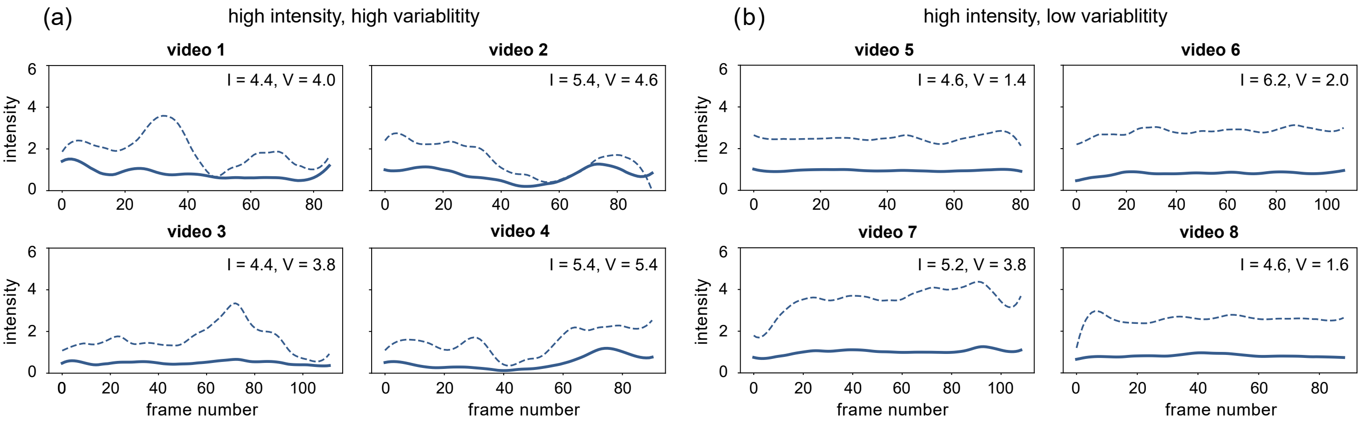

Fig. 2. a) the maximum and average of intensity for all action units in four segments rated as highly intense and highly variable; b) the maximum and average of intensity for all action units in four segments of high intensity and low variability.

TABLE II
REGRESSION RESULTS FOR VIDEOS AND IMAGES. "AGG" DENOTES THE AGGREGATION METHOD USED FOR VIDEOS, AS THEY CONSIST OF SEQUENCES OF FRAMES.

Method	Videos			Images	
	Agg	MSE	MAE	MSE	MAE
avg	avg	9.224	2.899	9.017	2.833
max	max	1.637	1.050	3.034	1.462
max	avg	6.234	2.323	-	-
avg	max	3.540	1.645	-	-
RR	max	0.774	0.709	0.823	0.753
RF	max	0.788	0.714	0.785	0.728
MLP	max	0.779	0.709	0.854	0.743
Linear	max	0.786	0.710	0.849	0.747
SVR	max	0.766	0.702	**0.784**	0.733
RR	avg	0.756	0.703	-	-
RF	avg	0.786	0.719	-	-
MLP	avg	0.756	0.698	-	-
Linear	avg	0.770	0.702	-	-
SVR	avg	**0.744**	**0.694**	-	-

On the other hand, there are many more middle-range cases in the current dataset, while segments of extremely low or extremely high intensity are few (see Figure 1). Consequently, the question of whether the perceived intensity is a linear combination of measured AUs intensities remains open. Our belief is that more advanced models are needed that take into account the temporal evolution of the single AUs intensity to improve the results. Applying such models requires more data.

Considering the simplicity of the regression methods used, the results presented in Table II can be considered only as a baseline computational model (RQ2). The MSE error obtained for the best model is comparable to the error reported in previous works, e.g. [43], which used a similar 7-point scale for intensity but a different dataset.

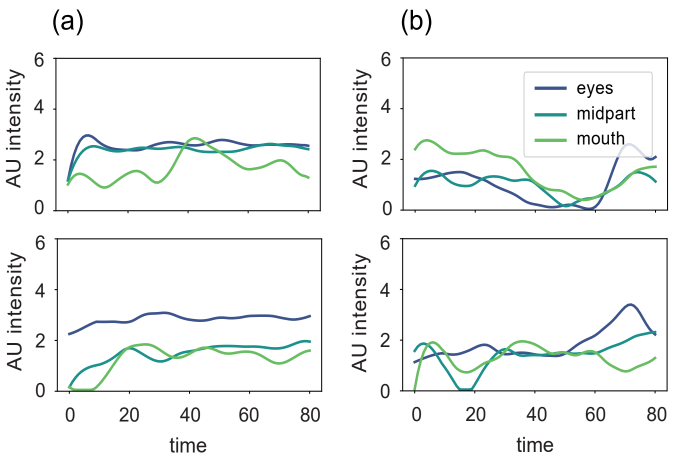

Fig. 3. Facial expressions and corresponding AUs: a) highly intense but low variability expressions; b) highly intense and highly variable.

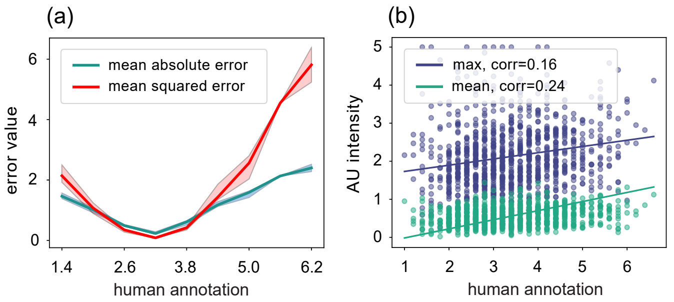

Fig. 4. Results: a) mean square error and mean absolute error for the best regression model, b) correlations between average and maximum of AUs for images and average of perceived intensity in human rankings.

VII. Applications

A computational model of perceived facial intensity may have several applications across a wide spectrum of topics, ranging from Human-Computer Interaction and entertainment (e.g., video games and Virtual Reality systems with affective feedback) to medicine (e.g., pain estimation). Here, we will focus on two different examples in more detail.

First, we believe that such models of perceived intensity can contribute to more effective Human-Robot Interaction. Existing models are usually restricted to a small number of emotional labels (e.g., Ekmanian) or action units, and they may not perform accurately when applied to the wide range of facial displays occurring in real-life interaction. Interactions with social robots may involve many unexpected events that elicit a variety of emotional (and other internal) states in humans [26], [33], which go well beyond basic emotions. Thus, robots need to perceive the intensity of all expressions, regardless of whether these expressions are basic emotions or not. Moreover, to be able to react appropriately, social robots (and other artificial companions) would need to analyze the intensity at a global level rather than at the frame level, resulting in one "global intensity score." In these applications, modeling the global perception of intensity is crucial.

Second interesting application of such a model can be the study of expressiveness across various movie genres. Recently, several works have exploited computational approaches to analyze and compare various literary genres [7], [8]. However, the lack of appropriate computational tools to process human nonverbal behavior has prevented similar comparative analyses for visual content, such as movies. In this context, it is particularly important to have tools that model the *perception* of facial expressions intensity.

VIII. Conclusion and Future Work

The main contribution of this work is the creation and validation of a new dataset called the Perceived Intensity of Facial Expressions Dataset. The PIFE comprises of videos and images depicting a large variety of facial expressions of several individuals in various natural settings and contexts (such as indoor and outdoor, artificial and natural lighting, partial shadows and occlusions, different face dimensions and face angles, and co-occurring activities such as speaking). All stimuli in the PIFE were evaluated in terms of intensity (videos and images) and variability (videos only) by 90 raters, with each stimulus rated five times. The dataset is designed for studying and modeling facial expression intensity.

Compared with existing datasets (see Section II-B), the following characteristics stand out:

- A large variety of expressions that go beyond basic facial displays of emotions, including non-affective expressions in various contexts;
- Holistic rankings of the perceived intensity by non-expert observers (i.e., no FACS experts);
- Rankings of the variability of facial activity in segments.

The data of facial landmarks, AU intensity values, and ratings of intensity and variability are available for research purposes [1]. Moreover, we provide baseline models for perceived intensity using the AU intensities extracted by OpenFace. While the obtained results are promising, there is room for improvement, and we offer some suggestions below.

Some limitations of this study should be mentioned. First, the quality of the models depends on the quality of the extracted data. It would be beneficial to repeat the experiments reported in Section VI using data extracted with tools other than OpenFace. Moreover, the final model should also be tested on other datasets. Second, we did not pose any restrictions on the participants other than age. It is not clear, however, whether some cultural and social factors may influence the perception of intensity. Future studies should address these issues. In this line, we added two questions in the questionnaire to check whether familiarity might have an impact on reports; however, the corresponding data (see Table I) were collected but not analyzed yet.

Future work will focus on designing better general computational models of facial intensity. Firstly, temporal profiles of AU intensities should be addressed in the future. Utilizing more advanced machine learning methods that take into account temporal information may require extending the dataset. To address this shortcoming, we are going to extract additional segments from another set of 12 movies. In our approach, segmentation is performed automatically, while collecting more ratings can be achieved using crowdsourcing, which proved to be quite successful in this study with only a small number of rejected raters. Secondly, the additional information on facial expressions collected in this study, that is, variability and its relation to perceived intensity, is worth deeper exploration in the future. Future research should also focus on the perceived intensity in multimodal expressions (e.g., facial expressions, body movements, and voice). Preliminary work in this direction was proposed recently by Niewiadomski et al. [28], but it is restricted only to the expressions of laughter. Last but not least, while the inter-rater agreements obtained in this work are generally satisfactory, stronger discrepancies were observed for some stimuli. It would be interesting to study in the future whether stimuli for which discrepancy is particularly high share some common characteristics.

Ethical Impact Statement

This work aims to answer a theoretical question regarding the perception of intensity and modeling the intensity perception. The stimuli used in this study come from available materials (movies), which are copyrighted and present publicly known persons (actors) performing intentionally some activities with the knowledge of being recorded and that these recordings (movies) will be available to a large audience. While facial information is revealed from images and videos, we neither use identity-specific information nor base our claims on a specific religion, race, or gender of the displayer. The video segments show actors of various ethnicities, age groups, and genders.

The raters performing the rating task are anonymous, and it is impossible to identify individual raters or judge them based on their given answers. The raters were informed before starting the task that they could withdraw at any moment without

[1] https://github.com/estiei/PIFE-Perceived-Intensity-of-Facial-Expression-Dataset

any consequences. Some of the raters received compensation corresponding to the time dedicated to the task.

According to the authors, at the current stage, this research does not bring any negative impact or consequences. However, it can contribute to the development of models that may potentially be misused in the future, similarly to any other model for analyzing human facial expressions. The proposed solution is non-obtrusive and reuses existing videos and online tools to collect the ratings. Several potential benefits are expected from the applications of this study, which are discussed in the paper.

References

[1] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, pages 1–4, 2010.

[2] Maren Awiszus, Stella Grasshof, Felix Kuhnke, and Jörn Ostermann. Unsupervised features for facial expression intensity estimation over time. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1199–11998, 2018.

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[4] Casey Becker, Russell Conduit, Philippe Chouinard, and Robin Laycock. A dynamic disadvantage? social perceptions of dynamic morphed emotions differ from videos and photos. *Journal of Nonverbal Behavior*, 01 2024.

[5] Paul Brunet, Roderick Cowie, Hastings Donnan, and Roddy Cowie. Politeness and social signals. *Cognitive processing*, 13 Suppl 2:447–53, 10 2011.

[6] Dana R Carney. The nonverbal expression of power, status, and dominance. *Current Opinion in Psychology*, 33:256–264, 2020. Power, Status and Hierarchy.

[7] Emanuele Castano. Less is more: How the language of fiction fosters emotion recognition. *Emotion Review*, 0(0):17540739241232350, 0.

[8] Emanuele Castano, Jessica Zanella, Fatemeh Saedi, Lisa Zunshine, and Luca Ducceschi. On the complexity of literary and popular fiction. *Empirical Studies of the Arts*, 42(1):281–300, 2024.

[9] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Intensity rank estimation of facial expressions based on a single image. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 3157–3162, 2013.

[10] Jane Reilly Delannoy and John McDonald. Automatic estimation of the dynamics of facial expression using a three-level model of intensity. In *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–6, 2008.

[11] Abhinav Dhall, Roland Goecke, and Tom Gedeon. Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing*, 6(1):13–26, 2015.

[12] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41, 2012.

[13] Euodia Dodd, Siyang Song, and Hatice Gunes. A framework for automatic personality recognition in dyadic interactions. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8, 2023.

[14] P. Ekman and W.V. Friesen. Felt, false, and miserable smiles. *Journal Nonverbal Behavior*, 6:238–252, 1982.

[15] Léo Hemamou, Arthur Guillon, Jean-Claude Martin, and Chloé Clavel. Don't judge me by my face: An indirect adversarial approach to remove sensitive information from multimodal neural representation in asynchronous job video interviews. In *2021 9th International Conference on Affective Computing and Intelligent Interaction*, pages 1–8. IEEE, 2021.

[16] Ursula Hess, Sylvie Blairy, and Robert E Kleck. The intensity of emotional facial expressions and decoding accuracy. *Journal of nonverbal behavior*, 21:241–257, 1997.

[17] Siti Khairuni Amalina Kamarol, Mohamed Hisham Jaward, Heikki Kälviäinen, Jussi Parkkinen, and Rajendran Parthiban. Joint facial expression recognition and intensity estimation based on weighted votes of image sequences. *Pattern Recognition Letters*, 92:25–32, 2017.

[18] Su Lei and Jonathan Gratch. Sources of facial expression synchrony. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, 2023.

[19] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010.

[20] Patrick Lucey, Jeffrey F. Cohn, Kenneth M. Prkachin, Patricia E. Solomon, and Iain Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition*, pages 57–64, 2011.

[21] S. Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Trans. on Affective Computing*, 4(2):151–160, 2013.

[22] Dhwani Mehta, Mohammad Faridul Haque Siddiqui, and Ahmad Javaid. Recognition of emotion intensities using machine learning algorithms: A comparative study. *Sensors*, 19, 04 2019.

[23] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 12(2):479–493, 2021.

[24] Michal Muszynski, Leimin Tian, Catherine Lai, Johanna D. Moore, Theodoros Kostoulas, Patrizia Lombardo, Thierry Pun, and Guillaume Chanel. Recognizing induced emotions of movie audiences from multimodal information. *IEEE Transactions on Affective Computing*, 12(1):36–52, 2021.

[25] R. Niewiadomski, J. Urbain, C. Pelachaud, and T. Dutoit. Finding out the audio and visual features that influence the perception of laughter intensity and differ in inhalation and exhalation phases. In *Proceedings of 4th International Workshop on Corpora for Research on Emotion, Sentiment & Social Signals, LREC 2012*, pages 25–31, Turkey, 2012.

[26] Radoslaw Niewiadomski, Merijn Bruijnes, Gijs Huisman, Conor Patrick Gallagher, and Maurizio Mancini. Social robots as eating companions. *Frontiers in Computer Science*, 4, 2022.

[27] Radoslaw Niewiadomski, Gabriele De Lucia, Gabriele Grazzi, and Maurizio Mancini. Towards commensal activities recognition. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, page 549–557, New York, NY, USA, 2022. Association for Computing Machinery.

[28] Radoslaw Niewiadomski, Yu Ding, Maurizio Mancini, Catherine Pelachaud, Gualtiero Volpe, and Antonio Camurri. Perception of intensity incongruence in synthesized multimodal expressions of laughter. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 684–690, 2015.

[29] Radoslaw Niewiadomski and Catherine Pelachaud. Towards multimodal expression of laughter. In Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn Walker, editors, *Intelligent Virtual Agents*, volume 7502 of *Lecture Notes in Computer Science*, pages 231–244. Springer Berlin Heidelberg, 2012.

[30] Optuna. https://optuna.org. Accessed: 2024-06-22.

[31] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*, pages 5 pp.–, 2005.

[32] Pinguin. https://pingouin-stats.org/build/html/generated/pingouin.intraclass_corr.html. Accessed: 2024-06-22.

[33] Maria Elena Lechuga Redondo, Radoslaw Niewiadomski, Francesco Rea, Sara Incao, Giulio Sandini, and Alessandra Sciutti. Comfortability analysis under a human–robot interaction perspective. *International Journal of Social Robotics*, 16:77–103, 2024.

[34] Motaz Sabri and Takio Kurita. Facial expression intensity estimation using siamese and triplet networks. *Neurocomputing*, 313:143–154, 2018.

[35] Scikit. https://scikit-learn.org. Accessed: 2024-06-22.

[36] Marina Tiuleneva, Emanuele Castano, and Radoslaw Niewiadomski. Towards the dataset for analysis and recognition of facial expressions intensity. In *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*, AVI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[37] M. Uddin and S. Canavan. Quantified facial temporal-expressiveness dynamics for affect analysis. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3955–3962, Los Alamitos, CA, USA, jan 2021. IEEE Computer Society.

[38] Md Taufeeq Uddin and Shaun J. Canavan. Quantified facial expressiveness for affective behavior analytics. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 131–140, 2022.

[39] Michel F. Valstar, Timur Almaev, Jeffrey M. Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F. Cohn. Fera 2015 - second facial expression recognition and analysis challenge. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, volume 06, pages 1–8, 2015.

[40] Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, Björn Schuller, and Maja Pantic. Deep structured learning for facial action unit intensity estimation. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5709–5718, 2017.

[41] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A high-resolution 3d dynamic facial expression database. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6, 2008.

[42] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. Best of Automatic Face and Gesture Recognition 2013.

[43] Rui Zhao, Quan Gan, Shangfei Wang, and Qiang Ji. Facial expression intensity estimation using ordinal information. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3466–3474, 2016.