Visual Cues for Disrespectful Conversation Analysis

Samiha Samrose, Wenyi Chu, Carolina He, Yuebai Gao,
Syeda Sarah Shahrin, Zhen Bai, Mohammed (Ehsan) Hoque
Department of Computer Science, University of Rochester, USA
{ssamrose,zhen.bai,mehoque}@cs.rochester.edu, {wchu6,clionhe,ygao41,sshahrin}@u.rochester.edu

***Abstract*—Toxic, abusive, or disrespectful behavior analysis is a non-trivial problem previously addressed mostly from the language perspective. In this paper, we present a novel video dataset containing disrespect and non-disrespect labels, and introduce such behavior analysis by using visual cues. The dataset is collected from YouTube news show videos of two-party conversations, in which a host and a guest interact through teleconferencing. Each video is labeled by three trained raters to identify disrespect expressed through face and gesture, voice, and language. By resolving confounding factors, we generate the corresponding pairwise samples of non-disrespect. To particularly show the influence of visual cues in disrespectful interactions, we present 222 labeled clips (duration=974.41(s), mean duration=4.39(s)). We extract and analyze the facial action units (AUs) prevalent in disrespectful behavior. Our result shows statistically significant differences after Bonferroni correction for Inner Brow raise (AU01), Lip Corner Depressor (AU15), and Chin Raiser (AU17). For prediction, we build two classifiers using logistic regression and linear Support Vector Machine with 62.61% and 61.48% accuracy, respectively. For an in-depth analysis of overall face and gesture features, we conduct a qualitative analysis using theme extraction. Our qualitative analysis provides further insights on leveraging synchronous and asynchronous features, along with combining text and audio data with visual cues to better detect disrespect.**

***Index Terms*—disrespect, visual cue, face and gesture.**

I. Introduction

With the continuous growth of online platforms facilitating easy exchange of opinions, the opinion polarization has deepened as well. Especially for topics involving politics, religion, etc., the discussion cannot remain civil if the opposing viewpoints are treated with disrespect [1]. Detecting disrespectful, toxic, abusive content has been explored mostly in the arena of natural language processing by using user-generated content from online discussion platforms [2]–[5]. Recently, publicly or privately sharing of video-content has gained momentum[1] within platforms like Skype, YouTube, Vine, Snapchat, etc. [6], [7]. Even though videos with hate, extremism, racism, etc. have been analyzed from language content and network formation perspectives [8], [9], demeaning or disrespectful behavior categorization within a multi-party conversation setting remains under-explored.

Disrespectful behavior analysis from naturalistic video-based conversations holds two main challenges. Firstly, the topics and the settings of the videos are highly sporadic, which make it difficult to construct a standard dataset. Secondly, video-based interaction analysis suffers from high dimensionality as it consists of multi-modal features involving audio-visual-language cues. The complexity also increases with speaker number.

In this paper, we present a novel dataset for studying disrespect in opinion exchange videos manually selected from YouTube[2]. To address the first challenge, we select news show videos in which the host and an invited guest discuss a political, controversial issue. We particularly consider news show videos, as (1) they have a professional setting for discussion, (2) the host and the guest both hold proper knowledge on

This work was supported by National Science Foundation Award IIS-1750380, and a Google Faculty Research Award.

[1]https://www.youtube.com/yt/about/press/
[2]General use policy followed: https://www.youtube.com/t/terms

Fig. 1. Two examples visual cues of disrespect from host-guest news show videos. Teleconferencing videos of such form maintains a standard template of split-screen. (Left) Finger pointing at the other speaker while talking. (Right) Wider eyes with raised eye-brows during speaking.

978-1-7281-3888-6/19/$31.00 ©2019 IEEE

the topic, and (3) the probability of having a disagreement is higher. For consistent measurement of behavioral patterns, we select those videos in which host-guest are teleconferencing by facing their cameras (shown in Fig 1). Such videos follow a standard pattern having a split-screen, where each part of the split-screen has a speaker facing frontwards. To remove any ambiguity regarding whom a disrespectful action is inflicted towards, we only consider dyadic discussion videos.

Because disrespect can be subjective [10]–[12], we formulate a definition with relevant constrains using related literature. Each selected video is watched and labeled by three raters trained on the definition constrains. The dataset includes the same length of *non-disrespect* samples as the *disrespect* samples from a video. As per our constrained definition the dataset holds inter-personal disrespect, thus it allows understanding the differences between disagreement and disrespectful features. Above all, being a naturalistic video-discussion dataset, this enables automatic analysis of verbal and non-verbal features for disrespectful content analysis. We have made the dataset publicly available: https://github.com/ROC-HCI/Disrespect. Our contributions are as follows:

- Constructing a *disrespect-nondisrespect* dataset of dyadic discussions from a naturalistic professional setting.
- Analyzing visual cues related to *disrespectful* behaviors by using facial landmark features.
- Applying a qualitative approach to face and gesture revealing the strength of synchronous-asynchronous features and multi-modal analysis for *disrespect* detection.

II. Related Work

Prior research has studied the importance of respect, or in other words the disruption caused by disrespect, in various settings. For a group discussion setting, commitment to the group or the discussion can vary with the level of respect or disrespect received [13]–[15]. Affective imbalance during conflict interactions negatively affects the discussion [16], whereas respect in the form of politeness and patience helps the discussion go better [2] and provides an overall better discussion experience [1].

During opinion exchange in difficult conversations, verbal and non-verbal features can play important roles in establishing or diminishing respect [1], [17], [18]. Derby et al. [18] show that without even realizing a person may start shouting out of excitement or anger. Therefore, even after knowing rules and norms of civility, during a conflict or disagreement people may inflict disrespect in the form of incidental personal threat or "face attack" [19]. However, intentional impoliteness is more related to disrespect than incidental attack [20]. Goffman [21] mentions face attack, threat, and malice as "calculated to convey complete disrespect... through symbolic meaning". Thus for any discussion, especially difficult ones, the effects of respect and disrespect potentially change the course of discussion.

Research has been done on detecting and regulating toxic language involving threat, aggression, harassment, racism, etc. using various NLP approaches [5], [22]. Wulczyn et al. [23] obtain Wikipedia comments with human ratings further facilitating ML model application to detect toxic textual content. Lorenzo-Dus et al. [9] identify impolite language from the comments of YouTube videos.

Even though impolite and toxic language have gained deserved attention, disrespect through audio-visual interactions has not gotten the same momentum. Research has emphasized affect moderation during a conversation [24], [25]. Group discussion with heated content has also been analyzed for remote- and co-located teams. However, the fact that (1) video-based interactions and thus content generation are increasing [6], and (2) verbal and non-verbal hostility in videos are also increasing is affecting overall content sharing [26]. As YouTube is a rich source of user-generated video content, there is an opportunity to observe and code disrespectful, uncivil, and impolite verbal and non-verbal components of interactions. As for coding, Coan et al. [27] introduce the Specific Affect Coding System (SPAFF) to learn to code behavior by observation. SPAFF mentions the facial action units, language patterns, etc. related to conflict and disagreement in a discussion. Combining the coding strategies for such disrespectful behaviors with toxic language components can provide a rich prediction of disrespect in audio-visual content.

III. Dataset

We construct a dataset having *disrespect* and *non-disrespect* labels from news-show YouTube videos.

A. Definition

To label *disrespect* in an interaction, first we define the problem based on constraints (shown in Table I).

Constraint 1. We do not identify any disrespect towards the topic of discussion. Rather, we identify a video-segment as disrespectful only if it involves disrespect inflicted towards the speakers' involved.

Constraint 2. We assume that the speakers involved have the same as well as the highest level of self-esteem. As from literature [28]–[30], a person with lower self-esteem might accept the insult inflicted upon them. Thus by applying this constraint we form a strong observer's perspective of high and equal self-respect for both speakers.

Constraint 3. We identify those disrespectful acts that are universal, not bounded by any specific norms or prior-notions. So we do not consider disrespect because of demographic and social identity, and past actions.

TABLE I
Labeling Instruction

Instruction	Explanation
Task	In a two-party conversation, identify the segments with disrespect towards each other.
Definition	Disrespect is an act that demeans someone's esteem.
Constraint	Judge each action assuming actors (people involved) have the same level of self-esteem. Do not consider cultural norms, demographic info (such as age, gender, race, etc.), social/professional rank, past action.

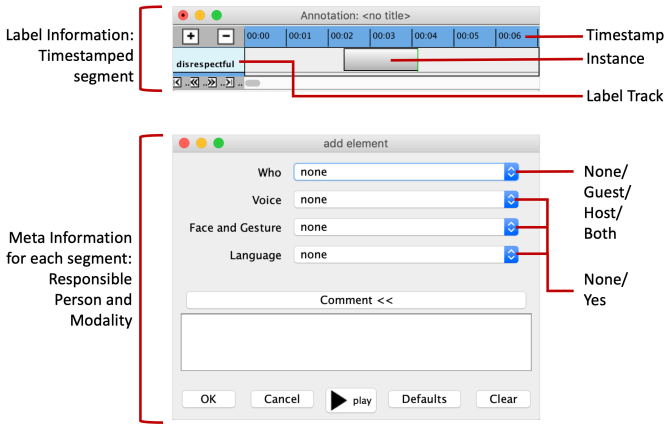


Fig. 2. Labeling Form.

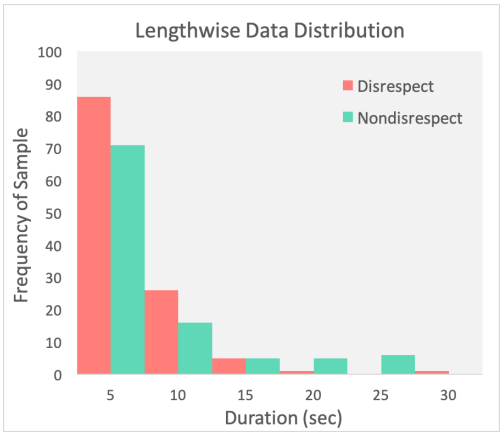


Fig. 3. Dataset duration frequency.

B. Data Extraction

We target discussions in which the setting is professional and prone to having disagreement. On this criterion, we select YouTube news show videos where a guest and a host talk about a particular political topic. Firstly, as with more than two people it might be hard for a system to understand towards whom the disrespect is directed, we only collect the videos with 1-1 guest-host conversations. Secondly, when host-guest are in a face-to-face discussion, they are facing each other, not the camera. This is problematic in capturing their full facial-bodily features. So we aim at those videos in which host-guest are teleconferencing facing the camera. These type of videos have a standard template with the screens split in half showing host and guest on each side of the split. The primary crawling is done from YouTube videos with search:{"heated+disagreement+debate+discussion+news"}. To keep consistent discussion topic and style, we select from there two popular US news media sources: CNN and Fox News. These two sources are included as they arguably hold different alignment of political views, and thus the target is to collect videos with potential *disrespectful* behaviors from both alignments. Then applying the two criteria mentioned beforehand we refine the desired dataset.

C. Segment Labeling

Next, we annotate the videos to extract the segments with *disrespect* label. Because of the constraints associated with our definition, instead of getting the label from crowd-sourcing platforms we include trained raters. Raters are trained by using related literature [31] and multiple pilot labeling iterations. Each video is labeled by three raters. We use the ANVIL tool[3] [32] for annotating the dataset. The labeling and meta information are shown in Fig 2.

In this stage, each rater individually watches a complete video. The videos are not masked for any particular modality which enables the raters to judge a video based on the complete context. The task of the rater is to identify and label segments containing *disrespect* generated from any modality, and also provide the label for that modality. To elaborate this step-by-step: For a video $v_x \in \{v_1, v_2, ..., v_n\}$, upon identifying a video-segment s_{x_y} as *disrespectful*, a rater labels it along with the corresponding start-timestamp i and end-timestamp j. Thus for each rater $r_z \in \{r_1, r_2, r_3\}$ of v_x, the labeled segments are extracted $s_x^z = \{s_{x_{ij}} \in v_x \mid D(s_{x_{ij}}) = 1, \forall_{i,j} 0 \leq i < j \leq len(v_x)\}$. For each segment, the rater also includes meta information that triggered $D(s_{x_{ij}}) = 1$ by specifying who is responsible for the *disrespectful* act (values: host/guest/both)[4], and which modalities are involved[5] (binary values)- face and gesture, voice, language. As this is time-series annotation with dynamic beginning-ending timestamps, discretizing measure for calculating inter-rater agreement is not applicable. An intersection approach is taken to find overlapping segments with two or more rater-agreements. Thus for a video v_x the final labeled segment is $s_x = \{s_x^p \cap s_x^q \mid p, q \in z \wedge p \neq q\}$.

D. Constructing Visual Cues Dataset

In this work, we present the visual cue analysis of the *disrespect* dataset. We discuss the extraction of *disrespect* and *non-disrespect* data specifically labeled based on visual cues:

1) Visual "Disrespect" Labeled Data

In a video, not every instance of s_x has {*meta-tag: face and gesture*} responsible for it. To get our desired visual-cues-datasegment where something in *face and gesture* is responsible for generating a *disrespectful* act, a refinement is applied on s_x such that, $s_x^- = \{s \in s_x \mid FG(s) = 1\}$; where $FG(s) = 1$ denotes those segments s where the *face and gesture* is labeled as responsible for the *disrespect* generation. For all videos, $VD^- = \bigcup_x s_x^-$. With this refinement, we end up with 119 sample clips, each of whose duration range from 0.1-25.4 seconds (detailed distribution shown using the 'red' bars in Fig 3).

2) Visual "Non-disrespect" Labeled Data

To conduct a comparison between visual cues related to *disre-*

[3]http://www.anvil-software.org/

[4]Even though 'None' value is provided, it is never used.

[5]Multiple modalities can be responsible in one segment.

spect and *non-disrespect*, we need to collect the latter samples by mitigating corresponding confounds [33]. To do so, for each video the total duration of the *non-disrespect* samples needs to be the same as that of the *disrespect* samples. Note that, for resolving confounding factors for text analysis, this pair generation is done with the aim of having the same number of positive-negative samples from a transcript. However, for video sample generation, in place of sample-count we reform the technique to have length or duration as the parameter for generating pairs. First, for each video we select candidate clip regions such that none of the raters marked these as *disrespectful*, $s_x^+ = \{s \in v_x \setminus s_x^-\}$. From these candidate regions of this video v_x, we extract n sample clips so that their total length is the same as the combined length of m samples of s_x^-, $VD_x^+ = \{s \in s_x^+ \mid \sum_{s \in VD_x^+} len(s) = \sum_{s \in s_x^-} len(s)\}$. Thus, $VD^+ = \bigcup_x VD_x^+$ gives 103 clips whose duration range is 0.01-22.9 seconds ('green' bars in Fig 3 shows detailed distribution).

Throughout the rest of the paper, by dataset we refer to this combined set of visual-cues samples with *disrespect* and *non-disrespect* labels.

IV. Analysis

A. Feature Extraction

We apply OpenFace [34] for automatic analysis of the Facial Action Coding System (FACS) [35]. Openface provides the values of 18 action units (AUs) (elaborated in Table II) for each video-frame recorded at 15 fps. Using this tool, the boolean values of the corresponding features are extracted for our dataset VD. Then the frequency of each feature is calculated for further analysis.

B. Statistical Analysis of Features

To compare the feature values and understand whether there is any difference among the two sample sets, a statistical analysis providing p-value is required. Our null hypothesis, H_0: There is no difference between *disrespect* and *non-disrespect* samples. As our data may or may not be normally distributed, we apply the Mann-Whitney U test [36]. Fig 4 shows the frequency comparison, and the statistically significant features at different significance levels. As we are considering 18 features and thus 18 hypotheses, we also apply the Bonferroni correction [37] on the p-values by multiplying them by 18.

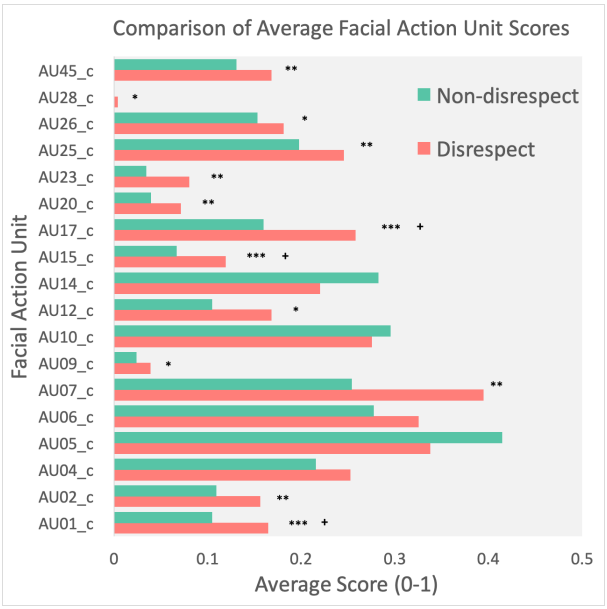


Fig. 4. MannWhitney U test results. * denotes statistically significant differences at the $p < 0.05(*), p < 0.01(**), p < 0.001(***)$ levels. + denotes statistically significant differences at the $p < 0.05$ level after Bonnferoni correction.

The final result in Fig 4 shows that the p-values of AU01 (Inner Brow Raiser), AU15 (Lip Corner Depressor), AU17 (Chin Raiser) stay statistically significant at $p < 0.05$. In each of the features, the frequency is higher for *disrespect* samples in comparison with that of the *non-disrespect* samples.

C. Classification

As from Section IV-B, we conclude that there are statistically significant differences between the AU features of the sample sets. As the difference is established, we strive to classify the samples into *disrespect* and *non-disrespect* classes. We apply and compare two machine learning models: (1) Logistic Regression Classifier [38], and (2) Linear Support Vector Machine (SVM) [39]. We use the Scikit-learn Python library [40] for the operations.

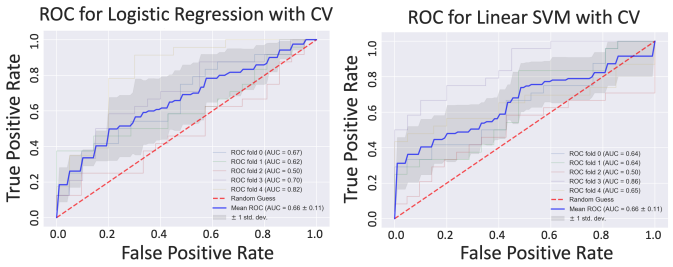
Both classifiers are applied on the extracted AU features to predict the two classes. For each feature, the average score per video clip is calculated to summarize and feed as input to the classifiers. The hyper-parameters are tuned by a randomized

TABLE II
Extracted Facial Action Units

AU#	Description	AU#	Description
AU01	Inner Brow Raiser	AU14	Dimpler
AU02	Outer Brow Raiser	AU15	Lip Corner Depressor
AU04	Brow Lowerer	AU17	Chin Raiser
AU05	Upper Lid Raiser	AU20	Lip Stretcher
AU06	Cheek Raiser	AU23	Lid Tightener
AU07	Lid Tightener	AU25	Lips Part
AU09	Nose Wrinkler	AU26	Jaw Drop
AU10	Upper Lid Raiser	AU28	Lip Suck
AU12	Lip Corner Puller	AU45	Blink

TABLE III
Classifier Analytics

	Logistic Reg.	Linear SVM
Accuracy	62.61%	61.48%
AUC	0.68	0.67
Precision	0.65	0.65
Recall	0.63	0.54
F1 Score	0.64	0.59



(a) ROC for Logistic Reg. (b) ROC for SVM

Fig. 5. Probabilistic ROC Curve.

5-fold cross-validation and 100 iterations. Fig 5 shows the ROC curve for one such iteration for both models. The logistic regression classifier gains an average accuracy of 62.61% with an average AUC of 0.68. As for the linear SVM, it holds an average of 61.48% accuracy with an average AUC of 0.67. Detailed performance of the classifier is reported in Table III.

Even though the classification accuracy is low, it shows that by using only the visual-cues, the *disrespectful* actions of an interaction can be identified.

D. Thematic Analysis

To better understand the overall visual cues associated with *disrespectful* interaction, we conduct a qualitative exploratory investigation by applying Thematic Analysis [41]. We randomly select 71 out of 119 clips with the *disrespect* visual-cue label for this exploratory investigation. The adopted steps for thematic analysis on our dataset is elaborated in Table IV.

TABLE IV
ADOPTED STEPS OF THEMATIC ANALYSIS

Step#	Instruction
Step: 1	Quickly watching all the clips to gain a first impression of all the clips.
Step: 2	For each clip: checking the metadata, watching the clip, writing down detail comment on that clip. Comments may include actions, concepts, patterns, ideas, something intuitive, something surprising, etc.
Step: 3	Searching for themes by organizing the comments/codes, thus generating potential themes and sub-themes.
Step: 4	Reviewing potential structure, and producing a thematic map showing relationships between themes and sub-themes.
Step: 5	Defining and naming themes.
Step: 6	Producing an analytic narrative with the merit and validity of the analysis.

Using theme extraction, we identify four main themes associated with the physical actions related to *disrespectful* interaction: (1) Eye (appeared 33%), (2) Hand (25%) (3) Mouth (23%), and (4) Head (19%). The themes and corresponding sub-themes are shown in Fig 7.

Note that, as these clips are very short it is unlikely to gain the context of disrespect from it. 44% of the time comments or codes express confusion and/or asks for context. Notably, previously during the annotation stage, the raters projected agreement regarding these segments while they watched full

Fig. 6. Thematic analysis of physical actions related to disrespect. (a-c) Hand-based, (d-f) Eye-based, (g-i) Mouth-based actions.

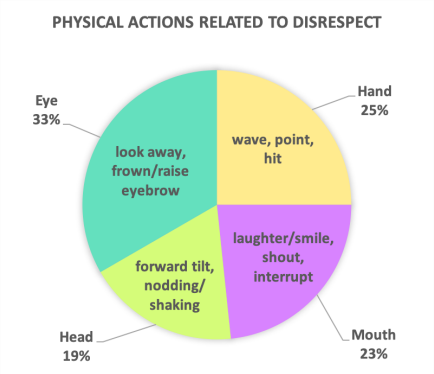


Fig. 7. Distribution of Themes and Sub-themes.

videos. Therefore, at this stage, for a short clip their confusion implies that, even for humans it might be difficult to identify a short instance of disrespect without proper context. This brings up another research direction to explore context-dependent understanding of disrespect.

The report of thematic analysis also sheds light on other physical actions, besides the face, which are responsible for disrespect. Fig 6 shows exemplary gestures brought up during sub-themes. For example, three sub-themes are constructed under hand movement: (a) pointing at the other person, (b) waving the hand in a dismissive manner, (c) hitting the table or nearby objects. For the case of eye movements: (d) looking away to avoid making eye-contact with the other speaker, (e) frowning, (f) raising eye-brows during a disagreement often can relate to disbelief towards the other speaker's opinions. The mouth movements can also express more information if correlated with other features: (g) shouting with wider mouth open gesture, (h) one-sided or single-person satirical smiling/laughing when a smile/laughter is absent for the case of the other speaker, (i) interrupting each other which can be observed by checking if both speakers' mouth-opens happen at the same time for a certain period of time. The last two examples also point out that by looking at synchronous and asynchronous physical features between the speakers can reveal more visual cues of disrespect.

V. Discussion

We find crucial insights to understand disrespect by using associated modalities. Our analysis shows that visual cues can differentiate between *disrespectful* and *non-disrespectful* behaviors. Therefore, it is evident that they contain information regarding the distinct classes. Secondly, using a single modality, it is hard for even humans to detect disrespect or impoliteness. If applied alone, any model that considers single modality would fail to take advantage of the diverse information available in videos. Thus, stronger models can be built on the dataset by considering all valuable modalities. Our future work involves multimodal analysis of this data, and studying the strengths and the weaknesses of each modality in comprehending *disrespect*.

Another interesting aspect is that, 44% of the instances humans asked for context to understand the label of a short clip. As discussed beforehand, disrespect is closely associated with intentional impoliteness rather than unintentional ones. Therefore, temporal pattern analysis also bears the capability to identify the candidate segments with potential *disrespectful* behavior. Thus, it might be possible to detect early signs of disrespect during a conversation. As a result, the framework can be incorporated in live online videos for automated moderation. In future, we intend to include temporal behavior patterns within the model to observe the performance changes.

At this stage, our dataset is small with only 222 clips. We plan to continue labeling and refining more videos to enrich the dataset. In our dataset, most of the clips are of short duration. However, we also include clips of varying lengths to diversify the dataset. As another refinement step, in future we plan to only include clips within a certain duration range by removing too short or too long outlying clips. Our thematic analysis shows the contribution of gesture (e.g., hand movement) in identifying visual cues of *disrespect*, the information which our current models do not hold. In future, we will explore building models with detailed visual cues of such information.

As disrespect may not be intentional in many cases, and people responsible for it may not realize conducting misbehaving. The detection of disrespect in video-based conversations thus also opens up the opportunity to provide necessary feedback. If early detection of disrespect is possible during an ongoing discussion, mediation at that moment can potentially mitigate further intensification of impoliteness.

VI. Conclusion

In this paper, we investigate the identification of the visual cues related to disrespectful behaviors in a two-person conversation setting. We construct a video discussion dataset of disrespect, refine it for face and gesture components, and make the dataset publicly available here https://github.com/ROC-HCI/Disrespect. Our analysis shows that during a disrespectful behavior, the facial actions unit features can be significantly different from that of a non-disrespectful discussion. We also predict the disrespectful behavior by only using facial features. Our qualitative analysis shows the potential improvement of the models by using other physical actions and context-awareness. While this is an exploratory study, we plan to investigate the pairwise correlation of the features for a speaker, the comparison of features for both speakers to detect synchronous and asynchronous behaviors, time-series data analysis to incorporate context-aware models, and multi-modal approach by combining text and audio feature with visual cues to detect disrespect.

Acknowledgment

We are thankful to Harry Reis, John Palowitch, Malte Jung, Jean Costa, Amir Zadeh, Mary Czerwinski, Daniel McDuff, Amanda Stent, and Reza Rawassizadeh for their invaluable feedback during various stages of this work. We thank the reviewers of the ACII conference in refining the paper.

REFERENCES

[1] J. Mansbridge, J. Bohman, S. Chambers, D. Estlund, A. Follesdal, A. Fung, C. Lafont, B. Manin, and J. l. Marti, "The place of self-interest and the role of power in deliberative democracy," *Journal of Political Philosophy*, vol. 18, no. 1, 2009.

[2] J. Zhang, J. P. Chang, C. Danescu-Niculescu-Mizil, L. Dixon, Y. Hua, N. Thain, and D. Taraborelli, "Conversations gone awry: Detecting early signs of conversational failure," *arXiv preprint arXiv:1805.05345*, 2018.

[3] B. van Aken, J. Risch, R. Krestel, and A. Löser, "Challenges for toxic comment classification: An in-depth error analysis," *arXiv preprint arXiv:1809.07572*, 2018.

[4] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Proceedings of the second workshop on language in social media*. Association for Computational Linguistics, 2012, pp. 19–26.

[5] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, "Deeper attention to abusive user content moderation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1125–1135.

[6] M. Duggan, "Photo and video sharing grow online," *Pew research internet project*, 2013.

[7] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 1–14.

[8] A. Sureka, P. Kumaraguru, A. Goyal, and S. Chhabra, "Mining youtube to discover extremist videos, users and hidden communities," in *Asia Information Retrieval Symposium*. Springer, 2010, pp. 13–24.

[9] N. Lorenzo-Dus, P. G.-C. Blitvich, and P. Bou-Franch, "On-line polylogues and impoliteness: The case of postings sent in response to the obama reggaeton youtube video," *Journal of Pragmatics*, vol. 43, no. 10, pp. 2578–2593, 2011.

[10] D. T. Miller, "Disrespect and the experience of injustice," *Annual review of psychology*, vol. 52, no. 1, pp. 527–553, 2001.

[11] J. Blanchard and N. Lurie, "Respect: Patient reports of disrespect in the healthcare setting and its impact on care," *Journal of Family Practice*, vol. 53, no. 9, pp. 721–731, 2004.

[12] H. J. Smith, T. R. Tyler, and Y. J. Huo, "Interpersonal treatment, social identity, and organizational behavior," *Social identity at work: Developing theory for organizational practice*, pp. 155–171, 2003.

[13] E. Sleebos, N. Ellemers, and D. de Gilder, "The carrot and the stick: Affective commitment and acceptance anxiety as motives for discretionary group efforts by respected and disrespected group members," *Personality and Social Psychology Bulletin*, vol. 32, no. 2, pp. 244–255, 2006.

[14] T. Tyler and S. Blader, *Cooperation in groups: Procedural justice, social identity, and behavioral engagement*. Routledge, 2013.

[15] D. De Cremer, "Respect and cooperation in social dilemmas: The importance of feeling included," *Personality and Social Psychology Bulletin*, vol. 28, no. 10, pp. 1335–1341, 2002.

[16] M. F. Jung, "Coupling interactions and performance: Predicting team performance from thin slices of conflict," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 23, no. 3, p. 18, 2016.

[17] R. C. Maia and T. A. Rezende, "Respect and disrespect in deliberation across the networked media environment: examining multiple paths of political talk," *Journal of Computer-Mediated Communication*, vol. 21, no. 2, pp. 121–139, 2016.

[18] E. Derby, D. Larsen, and K. Schwaber, *Agile retrospectives: Making good teams great*. Pragmatic Bookshelf, 2006.

[19] K. Tracy and S. J. Tracy, "Rudeness at 911: Reconceptualizing face and face attack," *Human Communication Research*, vol. 25, no. 2, pp. 225–251, 1998.

[20] J. Culpeper, D. Bousfield, and A. Wichmann, "Impoliteness revisited: with special reference to dynamic and prosodic aspects," *Journal of pragmatics*, vol. 35, no. 10-11, pp. 1545–1579, 2003.

[21] E. Goffman, *Interaction ritual: Essays in face-to-face behavior*. Routledge, 2017.

[22] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone can become a troll: Causes of trolling behavior in online discussions," in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. ACM, 2017, pp. 1217–1230.

[23] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 1391–1399.

[24] J. Costa, M. F. Jung, M. Czerwinski, F. Guimbretière, T. Le, and T. Choudhury, "Regulating feelings during interpersonal conflicts by changing voice self-perception," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 631.

[25] D. McDuff, A. Karlson, A. Kapoor, A. Roseway, and M. Czerwinski, "Affectaura: an intelligent system for emotional memory," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 849–858.

[26] P. J. Moor, A. Heuvelman, and R. Verleur, "Flaming on youtube," *Computers in human behavior*, vol. 26, no. 6, pp. 1536–1546, 2010.

[27] J. A. Coan and J. M. Gottman, "The specific affect coding system (spaff)," *Handbook of emotion elicitation and assessment*, pp. 267–285, 2007.

[28] L. L. Leape, M. F. Shore, J. L. Dienstag, R. J. Mayer, S. Edgman-Levitan, G. S. Meyer, and G. B. Healy, "Perspective: a culture of respect, part 1: the nature and causes of disrespectful behavior by physicians," *Academic medicine*, vol. 87, no. 7, pp. 845–852, 2012.

[29] D. Statman, "Humiliation, dignity and self-respect," *Philosophical Psychology*, vol. 13, no. 4, pp. 523–540, 2000.

[30] R. Wolf, "Respect and disrespect in international politics: the significance of status recognition," *International Theory*, vol. 3, no. 1, pp. 105–142, 2011.

[31] J. Coan and J. M. Gottman, *The Specific Affect Coding System (SPAFF)*, 04 2007, pp. 267–285.

[32] M. Kipp, "Anvil - a generic annotation tool for multimodal dialogue," in *INTERSPEECH*, 2001.

[33] D. B. Rubin, "The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials," *Statistics in Medicine*, vol. 26, no. 1, pp. 20–36, 2007.

[34] T. Baltruaitis, P. Robinson, and L. Morency, "Openface: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[35] P. Ekman and W. V. Friesen, "Facial action coding system (facs): Manual," 1978.

[36] J. S. Milton and J. C. Arnold, *Probability and Statistics in the Engineering and Computing Sciences*. McGraw-Hill Higher Education, 1986.

[37] A. Richardson, "Multiple comparisons using r by frank bretz, torsten hothorn, peter westfall," *International Statistical Review*, 2011.

[38] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[39] V. Vapnik and A. Chervonenkis, "A note on one class of perceptrons," *Automation and Remote Control*, vol. 25, 1964.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, 2011.

[41] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, no. 2, pp. 77–101, 2006.