Context in Human Emotion Perception for Automatic Affect Detection: A Survey of Audiovisual Databases

Bernd Dudzik*, Michel-Pierre Jansen†, Franziska Burger*, Frank Kaptein*,
Joost Broekens‡, Dirk K.J. Heylen†, Hayley Hung*, Mark A. Neerincx*, and Khiet P. Truong†
*INSY, Delft University of Technology, Delft (NL), 2628XE
Email: {B.J.W.Dudzik, F.V.Burger, F.C.A.Kaptein, H.Hung, M.A.Neerincx}@tudelft.nl
†HMI, University of Twente, Enschede (NL), 7500AE
Email: {m.jansen-1, d.k.j.heylen, k.p.truong}@utwente.nl
‡LIACS, Leiden University, Leiden (NL), 2333CA
Email: d.j.broekens@liacs.leidenuniv.nl

***Abstract*—An important aspect of human emotion perception is the use of contextual information to understand others' feelings even in situations where their behavior is not very expressive or has an emotionally ambiguous meaning. For technology to successfully detect affect, it must mimic this human ability when analyzing audiovisual input. Databases upon which machine learning algorithms are trained should capture the context of social interactions as well as the behavior expressed in them. However, there is a lack of consensus about what constitutes relevant context in such databases. In this article, we make two contributions towards overcoming this challenge: (a) we identify two principal sources of context for emotion perceptions based on psychological theory, and (b) we provide an overview of how each of these has been considered in published databases covering social interactions. Our results show that a similar set of contextual features are present across the reviewed databases. Between all the different databases researchers seem to have taken into account a set of contextual features reflecting the sources of context seen in psychological theory. However, within individual databases, these features are not yet systematically varied. This is problematic because it prevents them from being used *directly* as resources for the modeling of context-sensitive affect detection. Based on our findings, we suggest improvements for the future development of affective databases.**

***Index Terms*—Context, Automatic Affect Detection, Human Emotion Perception, Audiovisual Databases, Survey**

I. Introduction

One of the goals in Affective Computing research is automatic affect detection – providing computers with a human-like ability to perceive affective states in their users [1]. Affect detection is primarily approached by annotating and automatically analyzing behavioral signals captured as audio-visual data [1], [2]. In doing so, most research to date has focused on interpreting these behavioral signals in isolation, while largely ignoring their surrounding context. This forms a clear contrast to the effortless and continuous integration of diverse sources of information underlying emotion perception in humans, which takes into account additional knowledge about the observed person and the social situation for inferring experienced emotions [3]. This enables human observers to understand others' feelings even in situations where these display a behavior that is not very expressive or has an emotionally ambiguous meaning [4], [5]. Mirroring this human ability to use contextual information when inferring affective states could have similar benefits for technology [6], [7]. That is, it could improve estimates, even in situations where displayed behaviors might have multiple meanings. An important requirement for developing such context-sensitive affect detection is the existence of relevant datasets for training computational models. These datasets should systematically capture situations where human perceivers use contextual information to infer affective states. However, developing such datasets is a challenging endeavor, because it is often unclear what constitutes relevant context for affect detection [8], [9].

In this article, we make two contributions to overcome this challenge: (1) we identify two principal sources of context for emotion perceptions based on psychological theory, and (2) we provide an overview of what context related to each of these sources has been considered in published databases covering emotional social interactions. In this review, we focus on audio-visual databases. Firstly, because it has been the primary focus of automatic detection research in recent history [2]; and secondly, because the sensory information provided by auditory and visual modalities are important sources of information for emotion perception [10]. A structured overview of the contextual information captured within and across published databases benefits future research because: (1) it provides researchers with a common frame of reference for communicating about context that is relevant for automatic affect detection; and (2) it gives a preliminary overview of the current state of affairs, enabling the community to systematically and strategically develop future affective corpora.

This work has been partially supported by the 4TU research center Humans & Technology (H&T) project (Systems for Smart Social Spaces for Living Well: S4) and by the Netherlands Organization for Scientific Research (NWO) under the MINGLE project number 639.022.606

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

II. Context Information for Emotion Perceptions

A. Impact and Sources of Context Information

The Modified Brunswikian Lens Model (MBLM) [11] is a conceptual framework for the communication of affect that helps to understand the role of context in emotion perception. Specifically, it describes the process by which *behavioural signals* expressed by a *sender* are observed and interpreted by a *perceiver* when inferring affective states. It starts with the *encoding* of a sender's internal affective state into a behavioural signal possessing objectively measurable characteristics (e.g. pitch in verbal speech). By *transmission* through a medium (e.g. air), this signal reaches the perceptual system of perceivers. Here it is experienced as meaningful cues for *decoding* the sender's affective state.

Wieser and Brosch [12] provide a thorough overview of empirical research on what contextual information influences the decoding of behavioral signals in emotion perception. Guided by the MBLM, they distinguish between (1) additional perceived information about the sender and (2) perceived information about aspects in a scene that are separate from them. Furthermore, they highlight empirical findings demonstrating the impact of pre-existing knowledge on human affect recognition. Examples for this include the presence of racial stereotypes or learned affective associations [12]). Similarly, knowledge about cultural values and practices is widely recognized as influencing affective interpretations [3], [13], with perceivers' interpretations being more accurate if sender and perceiver share a cultural background [14]. Moreover, empirical evidence points to culture-specific differences in how perceivers use contextual information when interpreting the affective meaning of behavioral signals [5].

In summary, psychology highlights two primary sources of context that are used by perceivers when interpreting the emotional meaning of senders' behaviors:

- *Perceivable Encoding Context* – factors that are perceived alongside senders' behavioral cues, and that are experienced by perceivers as having plausibly affected the encoding of their emotion. Central aspects of this context relate directly to a *sender*, and the *situation* that he/she is embedded in [12]. Importantly, accessing contextual information of this kind involves reconstruction undertaken by perceivers based on sensory input. However, there may not be sufficient input available to reconstruct relevant influences on a senders' encoding, rendering these imperceivable in such circumstances.
- *Perceiver Knowledge and Experiences* – factors that potentially lead to the (re)construction or filtering of perceived information when decoding a behavioural signal. In particular, this comprises any relevant knowledge, skills, and personal experiences.

B. Identifying Context Information Captured in Databases

When creating an audiovisual database with annotations of perceived emotions, the perceivable encoding context is determined by the recorded audiovisual material that is presented to annotators. It constitutes the sensory input forming the foundation for the contextual information that these reconstruct about *senders* and the *situation* that these are in. This information is received alongside any *behavioral signals* emitted by senders (e.g. their facial expressions). Additionally, the individual background of each person that was selected as annotator determines what knowledge and experiences will influence their performance as perceivers. This difference is typically not explicitly captured in databases. However, creators of corpora may describe their annotators along certain dimensions they deem relevant (e.g., gender, age, or nationality). These *perceiver attributes* implicitly capture (some of) the individual differences in knowledge and experiences.

III. Method

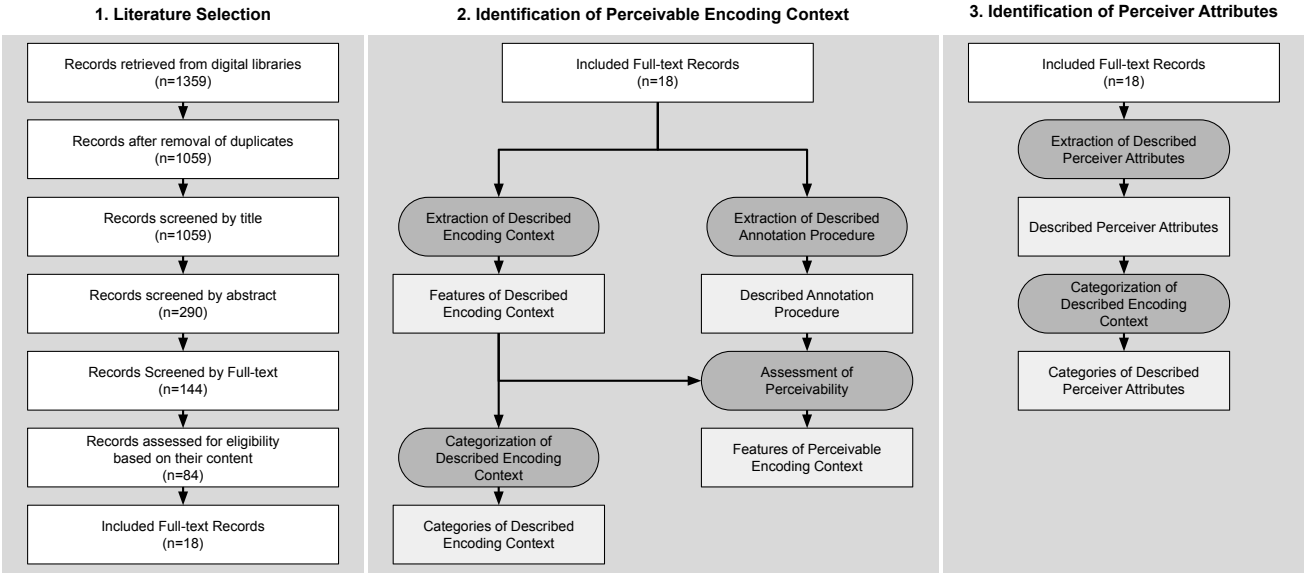
To provide an overview of the context for emotion perception that has been captured in existing corpora, we identify publications about relevant databases and extract any descriptions provided in them about (1) the specific social interaction(s) captured in a corpus' audiovisual records, and (2) the annotators providing data about the emotions they perceive. Figure 1 is an illustration of the methodology that we followed in this process.

A. Literature Selection

1) Search strategy: To identify relevant literature, we queried the following digital libraries: *Scopus*, *Web of Science*, *IEEE Xplore*, and *ACM Digital Libary*. The former two were selected based on their size and broad scope, while the latter two have a more specific focus on Affective Computing literature. The initial query covered the distinct core concepts: *affect*, *detection* and *audiovisual database*. In an iterative process, this query was modified based on the recall of papers from a list of publications previously identified as relevant.

2) Selection criteria and filtering: Using the constructed query, 1359 publications were retrieved. After removing duplicate records, 1059 unique entries remained for further review. This list was narrowed down through screening based on predefined criteria: a database paper is included when it (1) is retrievable, (2) introduces a database and does not only reference it, (3) includes perceived emotion or affect annotations. All papers were filtered sequentially through screening based on title, abstract, and the full text. This reduced the set first from 1059 to 290, then to 144, and finally to 84 publications. From within this list of records, we selected those 18 records describing databases that 1) contain and expose annotators to audiovisual recordings and 2) capture social interactions between at least two human beings. For reliability purposes, each of the filtering stages was performed independently by at least two authors. Interrater agreement was substantial for the title ($\kappa = .75$) and full-text filtering ($\kappa = .65$), and moderate for the abstract filtering ($\kappa = .42$). After independent filtering, all disagreements on specific publications were resolved through discussion between at least two authors.

Fig. 1. METHOD OVERVIEW



B. Identification of Perceivable Encoding Context

1) Extraction: We extracted from each full-text record any textual descriptions of the contents captured by the audiovisual material. Each feature so extracted was then categorized as either (1) relating to captured senders which encoded a behavioral signal (*SENDER features)*, or (2) the situation under which they did so (*SITUATION features*). Some databases comprise multiple subsets of audiovisual recordings with contents separately elaborated on by their creators (e.g. [15]). In such cases, we extracted the contextual features for each such subset. Together these features form a list of the *described encoding context* within a dataset.

2) Assessment of Perceivability: For each of the reviewed database publications, we identified passages that characterize the procedures followed to capture, segment and present the audiovisual material to annotators. Based on this information, one of the authors labeled each feature of the described encoding context as either perceivable, non-perceivable, or unclear. This process of categorization was repeated independently by a second author. We resolved cases in which both reviewers disagreed by the arbitration of a third author. For publications describing multiple different procedures for creation, segmentation, or presentation, we repeated this review process for each affected context feature. For example, if video material of the same social interaction had been segmented at three different lengths before annotators saw it, then we would have conducted three evaluations of the perceivability for all features of the described encoding context.

3) Categorization: Finally, we clustered all features of the described encoding context into distinct, non-overlapping categories that reflect a shared semantic meaning. This categorization process was based on a consensus decision among two of the authors, as were the labels assigned to each category created in this way.

The following categories were formed based on the extracted SENDER features:

- *Age* – descriptions of senders' biological age.
- *Cultural Embedding* – information about senders' nationality or their ethnic background.
- *Gender* – specifications of senders' gender.
- *Language Proficiency* – features describing the skill with which senders speak a particular language, e.g. whether they are native speakers or not.
- *Occupation* – descriptions of senders' profession or the educational program that they are currently enrolled in.
- *Other* – descriptions of senders for which no consensus among reviewers could be reached during categorization or that are highly corpus-specific.

The categories emerging for extracted SITUATION features were:

- *Cause of Emotion* – descriptions of the cause for the emotional behavior in the social interaction.
- *Conversation Content* – descriptions of the content that was discussed during the social interaction.
- *Conversation Partner* – information about senders' conversation partner present in the social interaction.
- *Conversation Language* – information about the language spoken during the social interaction.
- *Location* – descriptions/names of the locations in which the social interaction takes place.
- *Illumination* – descriptions of the lighting conditions during the social interaction.
- *Other* – descriptions of the social interaction for which no agreement could be reached for categorization or that are corpus-specific.

C. Identification of Perceiver Attributes

1) Extraction: We extracted *described perceiver attributes* from any relevant passages in each of the reviewed publi-

cations, i.e. sections in which database creators specify who provides the annotations for the emotional states of senders in the captured video data.

2) Categorization: We grouped the list of described perceiver attributes into non-overlapping categories based on consensus among the authors. The following list characterizes the attributes falling into each of them:

- *Age* – descriptions of perceivers' biological age.
- *Cultural Embedding* – features describing perceivers' nationality or their ethnic background.
- *Gender* – descriptions of perceivers' gender.
- *Language* – information about the languages understood by perceivers.
- *Occupation* – descriptions of the current profession or type of educational program that perceivers are currently participating in.
- *Other* – attributes for which no consensus among reviewers could be reached or that are highly corpus specific.

IV. Results

A. Perceivable Encoding Context

The categories of encoding context described in each of the reviewed database publications are listed in the *SENDER*- and *SITUATION*-sections of *Table I*. They show for each category (1) whether a feature belonging to it was described in a particular publication and (2) whether or not we assessed it as perceivable given the reported annotation procedure(s). Categories of sender-features that were generally reported and assessed as perceivable include *age* and *gender* as well as aspects of senders' *cultural embedding*. However, the latter is typically not varied within databases. For example, corpora contain recordings of social interactions between only Chinese [16] or only Philippine [17] nationals. While many datasets provide some information about senders' *occupation*, actual descriptions only span actors and students. Additionally, given the described annotation procedures, we judged it as unlikely that perceivers can infer this last information from the captured audiovisual material. Salient features with which situations in the databases are described include senders' *conversation partners* and/or the *conversation language* used in captured social interactions. This context is typically stable throughout the examples contained in databases, e.g. conversation language is Chinese [16] or Dutch [18] for all captured interactions.

An important category of perceivable situation features in the reviewed corpora is conversational information about the *Causes for Emotions* of senders. We deem this information mostly perceivable because the captured material is interactions between people that verbally communicate about these causes. For example, senders engage in discussions in which they had to reach conflicting goals [19]). Other categories of context information varied widely in their degree of perceivable across corpora. An example for this are conversation partners (e.g. [15], [17], [20]): less than half of the publications that explicitly mentioned the presence of a conversation partner in the captured social interactions also used an annotation procedure where perceivers are provided with audiovisual material about these people.

B. Perceiver Attributes

The categories of perceiver attributes mentioned in each of the reviewed databases can be found in the PERCEIVER-section of *Table I*. In comparison to the relative richness with which the reviewed publications describe encoding context, characterization of perceivers is very sparse: barely more than half of the corpora provide *any* information about the individuals serving as annotators. Descriptions that are provided cover mainly basic demographic information falling into the categories of age, gender, or cultural embedding.

V. Discussion

A. Similarity in Perceivable Encoding Context

Our findings show that there exist categories of perceivable encoding context that are frequently described across the reviewed databases. This indicates a form of common understanding among their creators as to what constitutes information that should be reported about the conditions under which emotional behavior is encoded and interpreted. Our review shows that, given the annotation procedures used in the reviewed corpora, it is likely that perceivers could infer features belonging to some of these categories from the audiovisual data that they were provided with. Consequently, these form plausible elements of the perceivable encoding context captured by these databases. However, while many databases contain such instances of contextualized human emotion perceptions, they do not yet form viable resources for the computational modeling of these perceptions. This is because the captured contextual features are not systematically varied within these corpora.

Nevertheless, the categories of perceivable encoding context extracted from the reviewed material provide a starting point for a systematic exploration and evaluation in affect detection technology. In particular, senders' age, gender and cultural embedding form good initial candidates: (1) they comprise information that was identified by psychological research as relevant for human emotion perception (e.g. see [33] for the influence of age or [14] for insights on the effects of senders' membership to specific ethnic groups), and (2) there exists dedicated computational research for extracting this information automatically from audiovisual data (e.g. gender [34] and ethnicity [35]). As such, they could be directly incorporated as high-level features within affect detection pipelines.

Apart from its direct use as information in affect detection systems, a systematic characterization of corpora in terms of the discovered categories for perceivable encoding context might also guide modeling efforts more generally. For example, machine learning approaches have emerged that focus on *transfer learning*. These algorithms attempt to learn predictive models by combining training input from different sources and exploit similarities between them [36]. Their application has the potential to result in models that are capable of affect detection across different contextual configurations, even for

TABLE I
CONTEXTUAL INFORMATION PRESENT IN THE REVIEWED PUBLICATIONS

	DATABASES																	
	BOZ	*BUS*	*CHO*	*CU*	*KAS*	*LEF1*	*LEF2*	*LI*	*LUB1*	*LUB2*	*MAH*	*MET*	*PER*	*RIN*	*SHU*	*SNE*	*WAN*	*WEI*
SENDER																		
Age	+	+	+	+	–	–	–	+	–	–	+	–	+	+	+	+	–	+
Cultural Embedding	–	–	–	–	–	?	–	+	+	+	+	–	–	–	–	+	–	–
Gender	+	+	+	–	–	+	+	+	+	+	+	+	+	+	+	+	+	+
Language Proficiency	–	–	–	–	+	–	–	–	+	–	–	–	+	+	–	–	–	–
Occupation	–	/	/	–	–	/	–	–	–	–	–	–	/	/	–	/	–	–
Other	–	–	–	–	+	–	+	–	/	–	/	–	–	–	+	–	+	–
	BOZ	*BUS*	*CHO*	*CU*	*KAS*	*LEF1*	*LEF2*	*LI*	*LUB1*	*LUB2*	*MAH*	*MET*	*PER*	*RIN*	*SHU*	*SNE*	*WAN*	*WEI*
SITUATION																		
Cause of Emotion	+	–	+	+	–	+	+	–	–	–	/	+	–	?	–	+	–	–
Conversation Content	–	/	?	–	–	/	–	–	–	–	–	–	–	–	–	–	+	–
Conversation Language	+	–	+	+	+	+	+	–	–	–	–	–	–	+	–	–	–	?
Conversation Partner	?	/	+	/	+	+	+	–	+	–	/	?	?	+	+	–	+	–
Location	–	–	–	/	+	–	–	–	–	–	–	–	+	+	–	–	–	–
Illumination	–	–	–	–	–	–	–	–	–	–	–	–	–	+	–	–	+	–
Other	–	–	–	–	–	–	–	+	–	+	–	/	+	+	–	–	–	–
	BOZ	*BUS*	*CHO*	*CU*	*KAS*	*LEF1*	*LEF2*	*LI*	*LUB1*	*LUB2*	*MAH*	*MET*	*PER*	*RIN*	*SHU*	*SNE*	*WAN*	*WEI*
PERCEIVER																		
Age	–	*	–	*	*	–	–	–	–	–	–	–	*	–	–	–	–	–
Cultural Embedding	–	*	–	*	–	–	–	–	–	*	–	–	–	–	–	–	–	–
Gender	–	*	–	–	*	–	–	–	–	–	–	–	*	–	–	*	*	–
Language	–	–	–	*	–	–	–	–	–	*	–	–	–	–	–	–	–	–
Occupation	*	–	*	–	–	–	*	–	–	–	–	–	–	–	–	–	–	–
Other	*	–	–	*	–	–	–	–	–	–	–	–	*	–	–	*	–	–

+ : Publication describes a feature of this category, and it is captured as perceiver context (i.e. it is both present and perceivable).
/ : Publication describes a feature of this category, but it is not captured as perceiver context (i.e. it is was not perceivable by annotators).
? : Publication describes a feature of this category, but it is unclear if it was captured as perceiver context (i.e. whether it was perceivable by annotators or not).
***** : Publication contains a perceiver attribute of this category.
– : Publication does not describe a feature/attribute of this category

BOZ: Bozkurt et al. 2017, [21]; BUS:, Busso et al. 2017, [15]; CHO: Chou et al. 2018, [16]; CU: Cu et al. 2012, [17]; KAS: Kasuriya et al. 2013, [22]; LEF1: Lefter et al. 2014, [18]; LEF2: Lefter et al. 2017, [19]; LI: Li et al. 2017, [23]; LUB1: Lubis et al. 2017, [24]; LUB2: Lubis et al. 2018, [25]; MAH: Mahmoud et al. 2011, [26]; MET: Metallinou et al 2016, [20]; PER: Perepelkina et al. 2018, [27]; RIN: Ringeval et al. 2013, [28]; SHU: Shukla et al. 2016, [29]; SNE: Sneddon et al. 2+012, [30]; WAN: Wang et al. 2016, [31]; WEI: Wei et al. 2014, [32];

cases where labeled training data is scarce. However, the knowledge transfer employed in these learning methods works best when there is a high degree of similarity between the different data sources used in training [37]. Access to a structured description of the perceivable context can aid practitioners in understanding how situations captured in affective corpora resemble or relate to each other, and thereby facilitate their targeted use in transfer learning.

B. Neglect of Perceivers' Knowledge and Experience

An implicit assumption underlying the use of corpora describing perceived emotions as training data for affect detection is that they form an acceptable approximation of the emotions that are experienced by senders. The primary motivation for this is that it is significantly easier to collect large corpora of the affective states that people perceive in audiovisual material (e.g. through crowdsourcing) than to collect the amount of training data about what people feel in a particular situation.

However, relevant knowledge and experiences vary between these perceivers, leading to differences in their ability to take senders' perspectives and to accurately relate to their affective state. For example, perceivers that are familiar with the culture in which a sender is embedded are better at recognizing their facial expressions [38]. This advantage in emotion perception seems particularly strong when perceivers stem from the same culture [14]. Similarly, some forms of knowledge and experience present in a perceiver – e.g. racial stereotypes [12] – may cause interpretations of particular senders' affective states to be systematically biased. Together, this makes the overall neglect of documented perceiver attributes in published databases problematic. Such contextual information could enable practitioners to detect and avoid undesired biases in the performance of emotion detection systems. The capacity to do so is increasingly recognized as important for the development of explainable and responsible AI technology.

Psychology has developed and validated numerous instruments that can be used to describe perceivers in terms of attributes that are relevant for modeling human emotion perception. For example, measures to assess individuals' general empathic capabilities (e.g. the MSCEIT [39] used by Perepelkina et al. [27] when developing their corpus). Additionally, personality measures can account for some of the differences in background across emotion perceivers [12]. Condensed versions exist for many of these psychometric measures (e.g. [40]), providing the possibility for reliable assessment even in time-constrained scenarios, such as when collecting data through online crowd-sourcing. Widespread adoption of such procedures and tools for describing annotators may be a step towards systematically capturing the impact of background in audiovisual corpora of emotion perceptions.

C. Limitations and Future Work

There are several limitations to the method that we used to identify and structure context information captured by the reviewed affective databases. For once, we conducted our review based on the publications in which authors described them, not the audiovisual material itself. Consequently, we are likely to have missed some information that was available as perceivable encoding context, but that is not explicitly reported. Similarly, we might have misjudged the perceivability of features in cases where annotation procedures were described with few details. Future research efforts could reduce this limitation by inspecting the audiovisual material of databases directly. Additionally, we limited our investigation in this particular review to databases that focus exclusively on emotions perceived in recorded interactions between humans. In future research, we plan to expand this survey to databases of other domains where human-like audiovisual affect detection is of interest, e.g. interactions between humans and robots, or media-induced emotional responses.

VI. Conclusion

In this paper, we provided an overview of how context has been considered in existing corpora for constructing affect detection systems. Drawing on literature from psychology, we identified two primary sources of contextual information that are accessed by perceivers when inferring the emotions of others during social interactions: perceivable elements of the encoding context, and their knowledge and experience.

Our findings highlight that researchers developing audiovisual databases show a degree of shared understanding of what information is relevant when describing affective interactions between humans. Moreover, a considerable amount of this information may also be accessible to the perceivers based on the captured audiovisual material. Unfortunately, while a structurally similar range of perceivable encoding context is considered across existing corpora, individual databases do not yet explicitly account for it through systematic variation. This prevents these corpora from forming directly usable repositories for exploiting this structure in computational modeling. Additionally, our findings reveal that information about the perceivers providing annotations in the reviewed corpora is scarce. However, such descriptions are crucial information for building robust and accurate affect detection, because they support accounting for perceivers' individual background in computational models.

Future development of affective corpora should consciously capture the context variables that we identified as present in existing databases. In particular, efforts should focus on those categories of perceivable encoding context that are (1) supported as relevant by empirical findings from psychology, and (2) that can be easily detected using existing automatic analysis of audiovisual sensor data. Examples include age, gender, and cultural embedding. This information should ideally be explicitly annotated in databases so that machine learning for affect detection systems can directly exploit it. This should go hand in hand with a careful selection and refined characterization of perceivers, allowing researchers to make informed decisions w.r.t. the data they use for computational modeling activities. Overall, such corpora would allow researchers to further explore which context variables indeed improve the accuracy and robustness of automatic affect detection.

REFERENCES

[1] R. Picard, *Affective Computing*. MIT Press, 2000.

[2] S. K. D'mello and J. Kory, "A Review and Meta-Analysis of Multimodal Affect Detection Systems," *ACM Computing Surveys*, vol. 47, no. 3, pp. 1–36, feb 2015.

[3] U. Hess and S. Hareli, "The influence of context on emotion recognition in humans," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, may 2015, pp. 1–6.

[4] A. Marpaung and A. Gonzalez, "Can an affect-sensitive system afford to be context independent?" in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10257 LNAI. Springer, Cham, 2017, pp. 454–467.

[5] H. Aviezer, N. Ensenberg, and R. R. Hassin, "The inherently contextualized nature of facial emotion perception," *Current Opinion in Psychology*, vol. 17, pp. 47–54, 2017.

[6] Z. Hammal and M. Kunz, "Pain monitoring: A dynamic and context-sensitive system," *Pattern Recognition*, vol. 45, no. 4, pp. 1265–1280, apr 2012.

[7] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion Recognition in Context," pp. 1667–1675.

[8] Z. Hammal and M. T. Suarez, "Towards Context Based Affective Computing," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, sep 2013, pp. 802–802.

[9] A. Vlachostergiou, G. Caridakis, and S. Kollias, "Investigating context awareness of Affective Computing systems: A critical approach," *Procedia Computer Science*, vol. 39, no. C, pp. 91–98, 2014.

[10] J. Zaki, N. Bolger, and K. Ochsner, "Unpacking the informational bases of empathic accuracy." *Emotion*, vol. 9, no. 4, pp. 478–487, aug 2009.

[11] K. Scherer, "Scherer, k.r.: Vocal communication of emotion: A review of research paradigms. speech communication 40, 227-256," *Speech Communication*, vol. 40, pp. 227–256, 04 2003.

[12] M. J. Wieser and T. Brosch, "Faces in context: A review and systematization of contextual influences on affective face processing," *Frontiers in Psychology*, vol. 3, no. NOV, p. 471, nov 2012.

[13] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in Emotion Perception," *Current Directions in Psychological Science*, vol. 20, no. 5, pp. 286–290, oct 2011.

[14] N. Ambady and M. Weisbuch, "On perceiving facial expressions: The role of culture and context," in *Oxford handbook of face perception*, A. J. Calder, G. Rhodes, M. H. Johnson, and J. V. Haxby, Eds. Oxford University Press Oxford, 2011, pp. 479–488.

[15] C. Busso, S. Parthasarathy, A. Burmania, M. Abdelwahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017.

[16] H. C. Chou, W. C. Lin, L. C. Chang, C. C. Li, H. P. Ma, and C. C. Lee, "NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus," in *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, vol. 2018-Janua. IEEE, oct 2018, pp. 292–298.

[17] J. Cu, K. Y. Solomon, M. T. Suarez, and M. Sta. Maria, "A multimodal emotion corpus for Filipino and its uses," *Journal on Multimodal User Interfaces*, vol. 7, no. 1-2, pp. 135–142, 2013.

[18] I. Lefter, G. J. Burghouts, and L. J. Rothkrantz, "An audio-visual dataset of human-human interactions in stressful situations," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 29–41, 2014.

[19] I. Lefter, C. M. Jonker, S. K. Tuente, W. Veling, and S. Bogaerts, "NAA: A multimodal database of negative affect and aggression," in *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, vol. 2018-Janua. IEEE, oct 2018, pp. 21–27.

[20] A. Metallinou, Z. Yang, C. chun Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations," *Language Resources and Evaluation*, vol. 50, no. 3, pp. 497–521, 2016.

[21] E. Bozkurt, H. Khaki, S. Keçeci, B. B. Türker, Y. Yemez, and E. Erzin, "The JESTKOD database: an affective multimodal database of dyadic interactions," *Language Resources and Evaluation*, vol. 51, no. 3, pp. 857–872, 2017.

[22] S. Kasuriya, T. Teeramunkong, and C. Wutiwiwatchai, "Developing a Thai emotional speech corpus," *2013 International Conference Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation, O-COCOSDA/CASLRE 2013*, pp. 1–5, 2013.

[23] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "CHEAVD: a Chinese natural emotional audiovisual database," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 6, pp. 913–924, 2017.

[24] N. Lubis, M. Heck, S. Sakti, K. Yoshino, and S. Nakamura, "Processing negative emotions through social communication: Multimodal database construction and analysis," in *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, vol. 2018-Janua. IEEE, oct 2017, pp. 79–85.

[25] ——, "Processing negative emotions through social communication: Multimodal database construction and analysis," *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, vol. 2018-Janua, pp. 79–85, 2018.

[26] M. Mahmoud, T. Baltrušaitis, P. Robinson, and L. D. Riek, "3D Corpus of spontaneous complex mental states," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6974 LNCS, no. PART 1, pp. 205–214, 2011.

[27] O. Perepelkina, E. Kazimirova, and M. Konstantinova, "RAMAS: Russian Multimodal Corpus of Dyadic Interaction for Affective Computing." Springer International Publishing, 2018, vol. 11096, pp. 501–510.

[28] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*, no. i, pp. 1–8, 2013.

[29] J. Shukla, M. Barreda-Ángeles, J. Oliver, and D. Puig, "MuDERI: Multimodal Database for Emotion Recognition Among Intellectually Disabled Individuals," in *Research on Education and Media*, ser. Lecture Notes in Computer Science, A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs, and H. He, Eds. Cham: Springer International Publishing, jun 2016, vol. 9979, no. 1, pp. 264–273.

[30] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The Belfast induced natural emotion database," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 32–41, 2012.

[31] K. Wang, Z. Zhu, S. Wang, X. Sun, and L. Li, "A database for emotional interactions of the elderly," *2016 IEEE/ACIS 15th International Conference on Computer and Information Science, ICIS 2016 - Proceedings*, no. April 2018, 2016.

[32] H. Wei, D. S. Monaghan, N. E. O'Connor, and P. Scanlon, "A New Multi-modal Dataset for Human Affect Analysis," pp. 42–51, 2014.

[33] M. Riediger, M. C. Voelkle, N. C. Ebner, and U. Lindenberger, "Beyond "Happy, angry, or sad?": Age-of-poser and age-of-rater effects on multi-dimensional emotion perception," *Cognition and Emotion*, vol. 25, no. 6, pp. 968–982, 2011.

[34] C. B. Ng, Y. H. Tay, and B. M. Goi, "Recognizing human gender in computer vision: A survey," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7458 LNAI, pp. 335–346, 2012.

[35] N. Srinivas, H. Atwal, D. C. Rose, G. Mahalingam, K. Ricanek, and D. S. Bolme, "Age, Gender, and Fine-Grained Ethnicity Prediction Using Convolutional Neural Networks for the East Asian Face Dataset," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, may 2017, pp. 953–960.

[36] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[37] Z. Wang, Z. Dai, B. Póczos, and J. Carbonell, "Characterizing and Avoiding Negative Transfer," nov 2018.

[38] H. A. Elfenbein and N. Ambady, "When Familiarity Breeds Accuracy: Cultural Exposure and Facial Emotion Recognition," *Journal of Personality and Social Psychology*, vol. 85, no. 2, pp. 276–290, 2003.

[39] J. D. Mayer, P. Salovey, D. R. Caruso, and G. Sitarenios, "Measuring Emotional Intelligence with the MSCEIT V2.0," *Emotion*, vol. 3, no. 1, pp. 97–105, 2003.

[40] S. D. Gosling, P. J. Rentfrow, and W. B. Swann, "A very brief measure of the Big-Five personality domains," *Journal of Research in Personality*, vol. 37, no. 6, pp. 504–528, 2003.