Knowledge-augmented face perception: Prospects for the Bayesian brain-framework to align AI and human vision∗

Martin Maier a,b,*,1, Florian Blume a,c,*,1, Pia Bideau a,c, Olaf Hellwich a,c, Rasha Abdel Rahman a,b

a Cluster of Excellence Science of Intelligence, Technische Universität Berlin, Germany
b Department of Psychology, Humboldt-Universität zu Berlin, Germany
c Department of Computer Vision & Remote Sensing, Technische Universität Berlin, Germany

ARTICLE INFO

Keywords:
Bayesian brain
Facial expression perception
Top-down modulation
Computer vision
Deep learning

ABSTRACT

Human visual perception is efficient, flexible and context-sensitive. The Bayesian brain view explains this with probabilistic perceptual inference integrating prior experience and knowledge through top-down influences. Advances in machine learning, such as Artificial Neural Networks (ANNs), have enabled considerable progress in computer vision. Unlike humans, these networks do not yet adaptively draw on meaningful and task-relevant contextual cues and prior knowledge. We propose ideas to better align human and computer vision, applied to facial expression recognition. We review evidence of knowledge-augmented and context-sensitive face perception in humans and approaches trying to leverage such sources of information in computer vision. We discuss how both fields can establish an epistemic loop: Redesigning synthetic systems with inspiration from the Bayesian brain-framework could make networks more flexible and useful for human–machine interaction. In turn, employing ANNs as scientific tools will widen the scope of empirical research into human knowledge-augmented perception.

1. Introduction

Depending on knowledge-based expectations and the context in which we encounter an image, we can perceive it in many different ways. A powerful demonstration of this is the Kuleshov effect (Fig. 1a), named after Soviet filmmaker Lev Kuleshov, that formed the basis of a now ubiquitous film editing technique (Mobbs et al., 2006): the same camera shot of a man’s face can be interpreted as displaying sadness, appetite, or desire when juxtaposed with the image of a coffin, a bowl of soup, or an attractive woman, respectively. Human observers grasp a meaningful connection between juxtaposed scenes and use it to make sense of the man’s neutral expression. Now imagine a state-of-the-art neural network tasked with recognizing the man’s facial expression in each sequence. Given the same input, it would apply the same label each time. Even though, in isolation, this would be objectively accurate, it reveals a lack of context-sensitivity and would be misaligned with human perception in real world scenarios. If flexible integration of knowledge and contextual
information could be added to artificial neural networks, this could improve their performance across varying contexts, make them more useful scientific experiments addressing human intelligent perception, and benefit general human–machine interaction.

Since our work covers an interdisciplinary topic, we start out by clarifying some important conceptual terminology, and also refer readers to the glossary in Section 5. We use context information as a term for information contained in the situation in which visual input is encountered. This can be, for example, the background scene and situational or verbal cues preceding a visual stimulus. By context-sensitivity we mean that context information can alter perceptual outcomes, such that the interpretation of the same input can differ between different contexts. By knowledge-augmentation, we refer to the notion that human perception uses top-down processing (see Glossary; Section 5) to integrate additional semantic information not present in the current bottom-up visual input, derived, for example, from learned knowledge, beliefs, or linguistic structures. Readers with a background in machine learning might immediately think that the trained weights of a neural network represent (prior) knowledge, too, since they are based on meaningful statistical regularities in the training data. The difference to knowledge-augmentation as proposed here is that in neural networks, knowledge is only represented implicitly and bottom-up, initially not designed for flexible integration of additional sources of information during perceptual inference. The goal for knowledge-augmented computer vision will be to provide ANNs with the capabilities that top-down influences bring to human perception, in order to achieve more flexible and context-sensitive processing.

To highlight the gap between human and artificial visual perception, we primarily draw on the task of Facial Expression Recognition (FER) because of the well-documented effects of knowledge and context information on human perception of facial expressions (e.g., Wieser & Brosch, 2012; Aviezer, Ensenberg, & Hassin, 2017; Otten, Seth, & Pinto, 2017; Righart & Gelder, 2008; Suess, Rabovsky, & Abdel Rahman, 2015; de Gelder et al., 2006; Abdel Rahman, 2011) and the high scientific and societal relevance of its application in synthetic systems (Duffy, 2003; Hortensius, Hekele, & Cross, 2018). In our work, we use FER when referring to synthetic face processing to avoid any confusion with human face perception.

Like human visual perception in general, facial expression perception is characterized by top-down effects that enable augmenting factors like knowledge to influence perceptual processing in the brain from early to late stages (e.g., Wieser & Brosch, 2012; Wieser et al., 2014; Otten et al., 2017; Suess et al., 2015; Baum, Rabovsky, Rose, & Abdel Rahman, 2020; Lupyan, Thompson-Schill,
perceiver features include, for instance, knowledge stored in memory and other characteristics of the perceiver, like race bias. Suess inform perceptual priors. This provides an empirically fruitful way to explain growing evidence that human perception is influenced by top-down factors like knowledge and conceptual categories: whenever these factors influence task-relevant perceptual priors, they categorize them. Contextual cues within a face include direct vs. averted eye gaze (Adams, 2005; Adams & Kleck, 2003) and the dynamics of facial movements (e.g., Schwarz, Wieser, Gerdes, Mühlberger, & Pauli, 2013; Recio, Sommer, & Schacht, 2011). The example in Fig. 1b illustrates a within-sender cue, body posture, which has been demonstrated as a powerful modulator of facial expression perception. The Kuleshov effect illustrated in Fig. 1a demonstrates an influence of external features, i.e., the valence associated with the scene that the person is confronted with. Similar effects are achieved by verbal descriptions of socially relevant information about a person given immediately before the presentation of a face, for instance perception of a more negative expression (Hassin & Trope, 2000; Russell, 1997). Both the production of emotional facial expressions and the perceived meaning of facial expressions are inherently context-dependent (Hess & Bourgeois, 2010; Aviezer et al., 2017). Consider, for example, how in Fig. 1, the same face of a tennis player appears to display either frustration or relief, depending on whether the body posture suggests that the player won or lost a point (Aviezer, Trope, & Todorov, 2012).

Wieser and Brosch (2012) reviewed a wide range of context-effects on facial expression perception and provided a taxonomy to categorize them. Contextual cues within a face include direct vs. averted eye gaze (Adams & Kleck, 2005; Adams & Kleck, 2003) and the dynamics of facial movements (e.g., Schwarz, Wieser, Gerdes, Mühlberger, & Pauli, 2013; Recio, Sommer, & Schacht, 2011). The example in Fig. 1b illustrates a within-sender cue, body posture, which has been demonstrated as a powerful modulator of facial expression perception. The Kuleshov effect illustrated in Fig. 1a demonstrates an influence of external features, i.e., the valence associated with the scene that the person is confronted with. Similar effects are achieved by verbal descriptions of socially relevant information about a person given immediately before the presentation of a face, for instance perception of a more negative expression (Hassin & Trope, 2000; Russell, 1997). Both the production of emotional facial expressions and the perceived meaning of facial expressions are inherently context-dependent (Hess & Bourgeois, 2010; Aviezer et al., 2017). Consider, for example, how in Fig. 1, the same face of a tennis player appears to display either frustration or relief, depending on whether the body posture suggests that the player won or lost a point (Aviezer, Trope, & Todorov, 2012).

2. Knowledge-augmented (face) perception in humans

Facial expressions provide crucial cues in human communication, supporting inferences about other people’s emotional and cognitive mental states. In the traditional Basic Emotion view, reading from faces relies on a specific set of prototypical, universal facial expressions that can be decoded by human “receivers” (Ekman, 1971; Smith, Cottrell, Gosselin, & Schyns, 2005). However, there is now considerable evidence that not all relevant information can be read from the specific configurations of a face itself or its action units (e.g., Feldman Barrett et al., 2019; Wieser & Brosch, 2012). On the contrary, a lot may depend on what we read into a face (Hassin & Trope, 2000; Russell, 1997). Both the production of emotional facial expressions and the perceived meaning of facial expressions are inherently context-dependent (Hess & Bourgeois, 2010; Aviezer et al., 2017). Consider, for example, how in Fig. 1, the same face of a tennis player appears to display either frustration or relief, depending on whether the body posture suggests that the player won or lost a point (Aviezer, Trope, & Todorov, 2012).

2.1. Top-down effects and predictive processing

The Bayesian-brain framework has been fruitful in accommodating a variety of aspects of human visual perception (Friston, 2009; Clark, 2013; Lee & Mumford, 2003), including such seemingly distant perceptual phenomena as persistent perceptual illusions (e.g., Weiss, Simoncelli, & Adelson, 2002) and in-the-moment modulations of perception by knowledge, context, or language (Lupyan, Abdel Rahman, Boroditsky, & Clark, 2020; Lupyan & Clark, 2015; Clark, 2013). Accordingly, theoretical explanations of context-effects on facial expression perception increasingly focus on predictive processing (e.g., Otten et al., 2017; Trapp, Schweinberger, Hayward, & Kovács, 2018; Pereira, Barbosa, de Haan, & Ferreira-Santos, 2019; Feldman Barrett, 2016).

To reiterate the Bayesian-brain view of visual perception in a nutshell: Rather than a passive process providing us with an objective, veridical image of the physical world, perception is viewed as an active, constructive process with the purpose of providing a useful basis for adaptive behavior situated in our environment (Buzsáki, 2019; Lupyan & Clark, 2015; Seth & Tsakiris, 2018), even if that sometimes means to “usefully misrepresent” the world (Martin, Solms, & Sterzer, 2021). In this view, the content of perception is specified through predictive processing—as the brain’s “best guess” about the causes of its sensory input that minimizes the difference (prediction error) between actual sensory signals and the predictions based on constantly updated predictive models (Friston, 2009; Clark, 2013; Otten et al., 2017). Previous experiences and knowledge stored in memory, in addition to evolution and development, inform perceptual priors. This provides an empirically fruitful way to explain growing evidence that human perception is influenced by top-down factors like knowledge and conceptual categories: whenever these factors influence task-relevant perceptual priors, they
Influence perception as a whole. In this view, human perceptual processing is not only bottom-up and feedforward, but also strongly characterized by top-down feedback. Top-down feedback is implemented neurally by corticocortical backward connections that are well documented empirically and seem necessary to capture the dynamics of neural representations in human visual processing (Kietzmann et al., 2019; Mermillod et al., 2018; Bastos et al., 2012; Friston, 2005; Gilbert & Li, 2013). Predictions and prediction errors are exchanged between levels throughout the visual processing hierarchy. Prediction error signals are assumed to propagate up to higher-order levels until they are sufficiently minimized. Therefore, the processing level at which factors like prior knowledge and context optimally influence perception is—in the long run—determined both by the informativeness of the prior and that of the sensory input.

Insight into the time course of top-down modulated perception, leveraging the high temporal resolution of electrophysiological measures like EEG and MEG, aligns well with predictive processing. Specifically, event-related potentials (ERPs) revealed that semantic knowledge (e.g., associated with novel objects) and conceptual-linguistic categories influence early stages of visual processing (Abdel Rahman & Sommer, 2008; Maier & Abdel Rahman, 2018; Maier, Frömer, Rost, Sommer, & Abdel Rahman, 2015; Lupyán et al., 2020; Maier et al., 2014; Maier & Abdel Rahman, 2018; Maier & Abdel Rahman, 2019; Weller, Rabovsky, & Abdel Rahman, 2019). For objects and simple color contrast stimuli, effects have been observed as early as 100 ms after stimulus presentation, in the P1-component (Maier & Abdel Rahman, 2019; Maier & Abdel Rahman, 2018; Weller et al., 2019; Abdel Rahman & Sommer, 2008), which reflects processing of low-level visual features like contrast (Clark, Fan, & Hillyard, 1994; Haynes, Roth, Stadler, & Heinze, 2003). Top-down effects on face processing have been demonstrated as early as in the N170 component, associated with the structural encoding stage of visual (face) perception (see Section 2.2.1). These effects are observed even after carefully controlling for low-level stimulus confounds and pre-perceptual shifts of attention, and affect processing prior to perceptual awareness (Eiserbeck & Abdel Rahman, 2020; Eiserbeck, Enge, Rabovsky, & Rahman, 2021; Maier & Abdel Rahman, 2018, 2019; Weller et al., 2019). Notably, such top-down effects on early visual ERP components (P1 and N170) suggest cognitive penetrability of perception (see Glossary; Section 5), by showing that cognitive factors influence early vision (Macpherson, 2017; Pylyshyn, 1999).

Taken together, the Bayesian brain-framework provides a useful basis to understand and model top-down effects on the perception of facial expressions. This is in most cases adaptive, leading to efficient and flexible perception of facial expressions, dealing with often noisy, ambiguous, and inherently contextualized input. In some cases, though, knowledge- and context-based priors can lead to incorrect and potentially problematic inferences, for instance, when judgments of facial expressions and other persons are biased by prejudice or information from untrustworthy sources (Baum et al., 2020; Baum & Abdel Rahman, 2021).

### 2.2. Hallmarks of knowledge-augmented (face) perception

Previous reviews have focused on the types of context that can influence face perception (Wieser & Brosh, 2012), or predictive processing as a general framework to explain such effects (Otten et al., 2017). In this section we take into closer perspective five key hallmarks of knowledge-augmented face perception afforded by predictive processing. We hope that these concrete aspects will equally foster experimentation with human participants and provide promising starting points for implementing and probing the success of knowledge-augmented face perception in synthetic systems.

The first hallmark (H1) concerns insights into which different levels of information processing are targets of top-down influences on facial expression perception. The next three are potential mechanisms by which perception is augmented by top-down predictions: reducing uncertainty (H2), enhancing relevant features (H3), and guiding prediction (H4). The last concerns subjective appearance (H5): what are the characteristics of percepts that result from knowledge-augmented perception? By pinpointing how phenomenological appearance is altered, we may get access to the mental representations that act as priors in predictive processing and model them (Brinkman, Todorov, & Dotsch, 2017). Undoubtedly, all of these aspects mesh with each other—for instance, top-down guided predictions of upcoming input should facilitate enhancement of relevant visual features at the point in time at which they become relevant.

#### 2.2.1. Influences on different levels of visual processing (H1)

There is substantial evidence that knowledge and context have the potential to shape face perception at distinct stages in the visual processing hierarchy, including early visual processes associated with structural encoding of facial expressions (reviewed by Wieser & Brosh, 2012). More specifically, electrophysiological studies have identified modulations of relatively early (N170 component), intermediate (early posterior negativity, EPN) and late processing stages (late positive potential, LPP).

The N170 component has been extensively studied in relation to face perception (e.g., Bentin, Allison, Puce, Perez, & McCarthy, 1996; Eimer, Kiss, & Nicholas, 2010). It typically peaks at parieto-occipital electrode sites around 130 – 200 ms after stimulus onset and is assumed to reflect the structural encoding stage of visual processing. There are conflicting results concerning effects of context and (affective) knowledge on the N170. Faces presented in a fearful context elicited a larger N170 compared to faces presented in a neutral context (Righart & de Gelder, 2006). Affective information (e.g., descriptions of contemptible social behavior) have sometimes been found to elicit a larger N170 (Luo, Wang, Dzhelyova, Huang, & Mo, 2016) and sometimes not (Abdel Rahman, 2011; Baum & Abdel Rahman, 2021; Baum et al., 2020; Suess et al., 2015). Taken together, context information appears to influence structural encoding of facial expressions, though the exact conditions leading to top-down modulations at an early processing level remain to be better understood.

Several studies reported effects of affective person-related knowledge on the EPN component (Abdel Rahman, 2011; Suess et al., 2015; Luo et al., 2016; Xu, Li, Diao, Fan, & Yang, 2016; but see Baum et al., 2020). The EPN is typically observed as a relative difference between affectively charged and neutral stimuli at posterior electrode sites and is associated with early selective visual attention (Suess & Sommer, 2008; Maier, Frömer, Rost, Sommer, & Abdel Rahman, 2015; Lupyán et al., 2020; Maier et al., 2014; Maier & Abdel Rahman, 2018; Maier & Abdel Rahman, 2019; Weller, Rabovsky, & Abdel Rahman, 2019).
Consciousness and Cognition 101 (2022) 103301

et al., 2015; Schupp, Flasch, Stockburger, & Junghöfer, 2006). Crucially, the EPN has also been taken as one of the earliest electrophysiological markers of the processing of facial expressions (Schacht & Sommer, 2009; Schindler & Bublatzky, 2020), suggesting a close link between the actual perception of facial expressions and modulations induced by affective person-related knowledge on the processing of expressions (Abdel Rahman, 2011; Suess et al., 2015).

Current evidence is quite consistent regarding the LPP component, which peaks at centro-parietal electrodes between 400 and 600 ms and reflects sustained task-relevant evaluative processing of affectively charged stimuli (e.g., Schupp et al., 2006). Negative and, to a lesser extent, positive affective information, are typically associated with larger LPP amplitudes compared to neutral information (Abdel Rahman, 2011; Baum et al., 2020; Xu et al., 2016). This suggests that knowledge influences high-level stimulus evaluation, such as judging valence or task relevant attributes. Taken together, what we know or believe about other people’s previous social behavior can influence early structural encoding, selective visual attention, and high-level evaluation of faces and facial expressions.

The temporal dynamics of knowledge and context effects in human participants provide useful constraints on the types of neural networks required to model them (e.g., recurrent networks; Kietzmann et al., 2019), and guide experimentation with networks in which a combination of different layers can be targets of top-down feedback loops. We turn to a closer look on how this might work technically in the following computer vision part of our review (Section 3) and and revisit this topic in the discussion of our proposed analytic-synthetic loop (Section 4.2).

### 2.2.2. Reducing uncertainty (H2)

Perceptual input is often noisy, weak, or ambiguous. At the same time, a cognitive system can be in a state of higher or lower certainty, influencing the distribution of its predictions (Knill & Pouget, 2004). Additional information, e.g., provided by knowledge, can increase the informativeness of priors and reduce uncertainty. An explicit prediction that follows is that perception should be influenced more strongly by top-down processing when the input is little informative (e.g., weak or ambiguous) or task difficulty is high (Otten et al., 2017; Maier & Abdel Rahman, 2019). Indeed, there is evidence of top-down influences especially on perception of ambiguous facial expressions (reviewed by Aviezer, Ensenberg, & Hassin, 2017; Hassin, Aviezer, & Bentin, 2013), such as neutral expressions (Wieser et al., 2014; Suess et al., 2015; Abdel Rahman, 2011) or expressions that are easily confused when presented in isolation (such as anger and disgust or intense expressions of joy and frustration; Aviezer et al., 2008; Aviezer et al., 2012). Bublatzky, Kavcicglou, Guerra, Doll, and Junghöfer (2020) showed that inducing a context of fear made morphs between fearful and happy expressions more likely to be categorized as fearful. This was associated with effects on early processing stages especially for (more ambiguous) low-intensity expressions. There are a number of more general examples beyond facial expression recognition. For instance, visual and verbal cues help to recognize objects in degraded visual stimuli (e.g., Samaha, Boutonnet, Postle, & Lupyán, 2018). There is also evidence that top-down effects have a stronger influence on performance in difficult tasks, such as detecting targets during the attentional blink (Maier & Abdel Rahman, 2018), difficult as compared to easy visual search (Maier & Abdel Rahman, 2019; Constable & Becker, 2017; Witzel & Gegenfurtner, 2016), and mental imagery compared to object recognition (Abdel Rahman & Sommer, 2008; Suess & Abdel Rahman, 2015; Maier et al., 2015). Taken together, one hallmark of predictive, knowledge-augmented perceptual systems is that they can draw on additional sources of information to reduce perceptual uncertainty.

### 2.2.3. Enhancing relevant features (H3)

Some theoretical views have stressed the role of top-down feedback in enhancing processing of task-relevant, or diagnostic visual features (Bar, 2004; Lupyán, 2012). Factors like conceptual knowledge or linguistic representations may inform perceptual priors in a way that enhances processing of stimulus features relevant to the current perceptual goals. In object recognition, predictions about what kind of object one might be seeing help to narrow down which features to process in more detail. According to a hypothesis by Bar (2004), such coarse predictions are generated in higher-level cortical areas (orbitofrontal cortex) based on low spatial frequency information that is extracted quickly from visual scenes. These predictions would constrain potential candidate objects and feed back to object-selective areas (e.g., inferior temporal cortex) and areas associated with low-level processing of detailed features, relying more on high spatial frequencies (e.g., V2–V4). This idea is supported by evidence of effects of object-related knowledge on early visual processing observed in the P1 and N1 components of the ERP (Abdel Rahman & Sommer, 2008; Weller et al., 2019; Maier & Abdel Rahman, 2019). Another theoretical account focusing on enhancement of diagnostic features was proposed to explain effects of language on perception (Lupyán, 2012; Lupyán et al., 2020). Language may shape the way conceptual knowledge is stored, for instance by making representations more categorical. When linguistic representations are activated, visual features that are diagnostic of category membership are highlighted, while features unique to individual exemplars are discounted (Lupyán, 2012; Lupyán & Clark, 2015; Lupyán & Lewis, 2019). In line with this idea, words can activate information about visual object shape and guide perception (Noorman, Neville, & Simonova, 2018; Boutonnet & Lupyán, 2015). Visual search is more efficient when the target and distractors belong to different conceptual-linguistic categories (Maier & Abdel Rahman, 2019). Linguistic categories appear to enhance the low-level salience of color contrasts, leading to better detection performance in the attentional blink (Maier & Abdel Rahman, 2018). The latter effects go along with changes to early visual processing in the P1 component of the ERP, further suggesting that linguistic concepts influence the processing of low-level visual features.

While the features relevant for recognizing different facial expressions have been extensively studied, especially with a basic emotion approach in mind (e.g., Schyns, Bonnar, & Gosselin, 2002), we are not aware of studies into how knowledge dynamically modulates the relevance of facial features in facial expression perception. However, it is likely that mechanisms found in other domains like object perception would apply here as well. It could be that the target features for top-down enhancement of early face processing are found on the level of structural encoding of configurations of features (N170 component), rather than simple features like contrasts or orientations.
2.2.4. Guiding prediction (H4)

Perception as active inference also implies that we constantly predict—and, through action, influence—upcoming input (e.g., Clark, 2013; Friston, Daunizeau, Kilner, & Kiebel, 2010). For instance, we make predictive eye movements in dynamic natural scenes (Hayhoe, McKinney, Chajka, & Pelz, 2012; Renninger, Verghese, & Coughlan, 2007) and remap attention to where a saccade will land before we execute it (Rolfs, Jonikaitis, Deubel, & Cavanagh, 2011). Furthermore, expectations based on predictable timing of a stimulus onset are used to tune the excitability of relevant cortical areas at the optimal point in time (Solis-Vivanco, Jensen, & Bonnefond, 2018; Helfrich, Huang, Wilson, & Knight, 2017). In facial expression perception, there is evidence that people predict future states of dynamic facial expressions. In a study by Yoshikawa and Sato (2008), participants chose expressions continuing the momentum of a previously presented dynamic facial expression when tasked with selecting the last images perceived (see also Palumbo & Jellema, 2013). Dozolme, Prigent, Yang, and Amorim (2018) additionally observed a tendency to predict a return towards a more neutral expression after being shown very intense dynamic facial expressions. To our knowledge, the influence of knowledge and context on predicting the development of dynamic facial expressions over time, and how it is associated with enhancing diagnostic features, remains to be explored. But drawing on context to predict how facial expressions will develop next seems like a useful and efficient strategy in social interaction that may also benefit artificial agents.

2.2.5. Altering appearance (H5)

Several studies suggest that top-down factors like knowledge or linguistic categories can alter subjective appearance, i.e. the phenomenological experience of how we perceive something (Macpherson, 2017). To highlight a few examples outside of face perception: Dutch and German speakers perceive different prototypical colors of traffic lights that are named differently (orange vs. yellow) but are objectively the same in both countries (Mitterer, Horschig, Müßeler, & Majid, 2009). Greek speakers, who use two distinct linguistic categories for light and dark blue appear to perceive a stronger contrast between these colors than German or English speakers who have one linguistic category encompassing both colors, which increases their chances of detecting certain stimuli in the first place (Maier & Abdel Rahman, 2018; Thierry, Athanasopoulos, Wiggert, Dering, & Kuipers, 2009). In speech perception, knowing the word in which a phoneme occurs changes its perceived sound to match the predicted phoneme (Samuel, 2001). The debate whether cognition truly affects perceptual experience is still ongoing, however (e.g., Macpherson, 2017; Firestone & Scholl, 2016; Firestone & Scholl, 2015).

Concerning facial expression perception, Suess et al. (2015) showed that negatively valenced biographical information leads to perception of slightly negative expressions in objectively neutral faces (see also Abdel Rahman, 2011). When body cues affect emotional expression perception, participants indicate that they base their judgment mainly on what they see in the face, even though the posture objectively dominates judgments (Aviezer et al., 2012). When tasked to recreate perceived expressions seen in body-face combinations, participants tend to project the combined valence into the face alone. A number of studies have used the reverse correlation technique to visualize people’s prototypes of faces associated with certain characteristics and concepts, such as dominance and trustworthiness, ethnic prejudices or beliefs about trustworthy vs. criminal behavior (Dotsch & Todorov, 2012; Dotsch, Wigboldus, Langner, & van Knippenberg, 2008; Dotsch, Wigboldus, & van Knippenberg, 2013). Brooks and Freeman (2018) demonstrated a relationship between conceptual knowledge about emotions and visual prototypes. For instance, participants conceptualizing the emotion categories sad and angry as more similar also produced more similar classification images for sad and angry faces. Such visualizations of prototypical facial expressions based on conceptual knowledge could serve as a proxy for perceptual priors that would be useful in modeling top-down effects.

Taken together, knowledge and context information can influence appearance, including the appearance of facial expressions. This insight could prove very valuable in understanding and modeling knowledge-augmented face perception in terms of the Bayesian brain framework. Visualizing how mental representations of faces are altered by context could help to model perceptual priors in the implementation of context-effects in synthetic systems.

3. Knowledge-augmented (face) perception in synthetic systems

In the previous chapter, we reviewed latest research on the influence of context information in light of prior knowledge on human face perception. On the synthetic (i.e. machine learning) side, we see that there still exists a large discrepancy between cognitive processes in humans compared to AI, despite recent advances in this field (Sinz et al., 2019; Storrs & Kriegeskorte, 2019; Perconti & Plebe, 2020). This is not surprising since AI is still a relatively young—but quickly developing—research area. In agreement with many researchers in this field (Goyal & Bengio, 2020; Lake et al., 2017; Sinz et al., 2019), we believe that modeling human Bayesian-like perception in AI systems will help to bridge this gap. Here we provide an overview over various computer vision FER methods and review efforts towards knowledge-augmentation in machine learning in general, but also in FER in particular. This chapter is structured as follows: First, we provide a more technical definition of the task from a computer vision point of view. Next, we treat classic and deep learning approaches. In the last part of this chapter, we present approaches aiming at integrating context information and prior knowledge into machine learning in a dynamic and top-down way. For clarification on our terminology, we refer the reader to our Glossary in Section 5.

3.1. Facial Expression Recognition (FER)

In computer vision, the task of recognizing facial expressions in images and videos using synthetic systems is commonly called FER (Li & Deng, 2017; Huang et al., 2019), among other names like emotion recognition (Kosti, Alvarez, Recasens, & Lapedriza, 2017; Lee,
3.1.1. FER datasets

In machine learning, the term dataset denotes the input data for a computer algorithm. Given a dataset together with so-called ground truth annotations, this data can be used to train and to evaluate the respective algorithm. Ground truth annotations for facial expressions can be represented in various ways. Two prominent representations are the six basic Ekman expressions (Ekman, 1971) and valence-arousal ratings. Ekman’s basic emotions are anger, surprise, disgust, enjoyment, fear, and sadness. Valence-arousal ratings offer finer-grained nuances and turn FER into a regression task. To obtain such valence-arousal ratings, human subjects rate facial expressions on a continuous scale of how positive or negative and how aroused they appear. Valence-arousal ratings and discrete Ekman expression classes are commonly provided by FER datasets.

Table 1 lists a selection of datasets alongside an excerpt of relevant characteristics. More extensive dataset analyses were conducted by Li and Deng (2018) and Huang et al. (2019). Some of the datasets focus on the human face exclusively and deliberately omit other information by ensuring a neutral background (e.g., JAFFE; Lyons, Kamachi, & Gyoba, 2020). Others explicitly include the surrounding visual context (CAER; Lee et al., 2019 and EMOTIC; Kosti et al., 2017). EmotioNet (Benitez-Quiroz, Srinivasan, Feng, Wang, & Martinez, 2017) comes with additional Action Units (AUs) denoting facial muscle movements that form the respective expression. AUs are used in the so-called Facial Action Coding System (FACS) to obtain categorizations of facial expressions that depend less on subjective human ratings (Ekman & Friesen, 1978). Some datasets contain posed expressions taken from movies (AFEW; Dhall, Goecke, Lucey, & Gedeon, 2011 and CAER; Lee et al., 2019) or the internet (AffectNet; Mollahosseini, Hasani, & Mahoor, 2017 and AffWild; Zafeiriou et al., 2017). Artificially posed expressions might lead to very different face perception in humans compared with spontaneous ones. In real-world human–machine interaction, the latter is particularly relevant.

FER can be considered solved for the JAFFE dataset with an accuracy as high as 98.9% (Li & Deng, 2018; Huang et al., 2019), meaning almost all images got correctly classified. In comparison, currently best performing methods tested on FER2013 (Goodfellow et al., 2013) reach only up to 75.10% (Li & Deng, 2018). FER2013 is a dataset which consists primarily of everyday life scenes. The discrepancy between JAFFE and FER2013 can be explained by the different characteristics of each dataset. Natural images capturing everyday life scenes make it significantly more difficult to recognize facial expressions correctly. As, for example, the angle of the face relative to the camera changes or the surrounding introduces partial occlusions. Ambiguous expressions, which are often hard to estimate accurately, also occur more regularly. Thus, the capacity to reduce uncertainty through context is very likely to benefit FER.

To address this problem of ambiguous facial expressions, CAER (Lee et al., 2019) and EMOTIC (Kosti, Alvarez, Recasens, & Lapedriza, 2019) contain images that include the face together with its surrounding. The authors correctly note that it is often not possible to accurately predict a facial expression by evaluating the facial area only, without taking the visual context into account. Similarly, Sues et al. (2015) provide additional affective knowledge which accompanies images centered on a single human face (see Fig. 1c).

The presented datasets reveal implicit issues: Different image conditions make inter-dataset comparisons difficult. Images from JAFFE (Fig. 2f), CK+ (Fig. 2c) and and the Radboud database (Fig. 1c) are distinguishable from the others due to their clean appearance (focused on the face, no background, uniform clothing), which manifests in the accuracy achieved by synthetic systems.

Compared to computer vision approaches, human perception often favors defensive interpretations of the input, e.g., to avoid harm, in light of previous experience (knowledge). This means that human perception does not always strive for representing the immediate “ground-truth” (Martin et al., 2021; Seth & Tsakiris, 2018). Human annotators introduce subjectiveness in their annotations, which is only equalized (if ever) when enough data is collected. Subjectiveness applies in cases of ambiguous facial expressions in particular. To summarize, facial expressions are the results of complex internal mental states that are not directly accessible, hence a single label or even continuous valence-arousal ratings might not express the input appropriately. These aspects are discussed in current research (Hagendorff & Wezel, 2020) and should be explored further by future research in order to build more versatile AI systems.

The presented datasets clearly push ongoing research in FER by allowing comparisons between different approaches. However, the datasets’ purpose and limitations require a careful analysis and assessment before and during utilization.

3.1.2. Classic FER methods

In machine learning, classic methods refers to hand-crafted feature selectors which are designed by experts in the respective domain (e.g., computer vision). Typically, FER methods consist of two steps, (1) feature extraction and (2) classification. Viola and Jones (2001) use so-called Haar features to recognize facial expressions in static images in their acclaimed work. They employ filters that respond to face characteristics like eyes, mouth, or nose. The resulting features are then used to recognize the facial expression. Local Binary
Pattern (LBP) matching is a technique to capture the brightness relationships between pixels in an image in a multi-region histogram. It has successfully been employed for FER (Ahonen, Hadid, & Pietikainen, 2004; Luo, Wu, & Zhang, 2013). Facial expressions in videos can be detected by optical flow (Yacoob & Davis, 1996) and feature point tracking (Tie & Guan, 2013). The former estimates the displacement of pixels between two images, the latter detects key points and matches them between frames. Classification from intermediate image representations (i.e. features) can be achieved through various techniques. Three prominent and well-established methods in computer vision are k-Nearest Neighbors (kNN; Altman, 1992), Support Vector Machine (SVM; Cortes & Vapnik, 1995) and Adaptive Boosting (AdaBoost; Freund & Schapire, 1999). They require a training stage in which a few parameters are optimized on a training dataset with respect to some error function. During inference, the best matching cluster or class is selected as the final prediction. Such classic approaches laid an important foundation for AI and still offer advantages like few parameters and little data to train.

Table 1
Overview over a selection of FER datasets.

| Name       | Type     | # Examples | Labels                  | Condition          |
|------------|----------|------------|-------------------------|---------------------|
| AFEW       | Videos   | 1,809      | 6 Expressions           | Movies              |
| Affectnet  | Images   | ∼1 Million | 7 Expressions, Valence-Arousal | Web               |
| AFWild     | Videos   | 298, ∼1 M Frames | Valence-Arousal     | Web                |
| BU-4DFE    | Videos   | 60,600 Frames | 6 Expressions         | Lab                |
| CAER       | Videos   | 13,201     | 6 Expressions           | Movies              |
| CK+        | Videos   | 593 (10–60 Frames) | 7 Expressions       | Posed & Spontaneous |
| EMOTIC     | Images   | 23,571     | 26 Expressions, Valence-Arousal | Web              |
| EmotioNet  | Images   | ∼1 Million | 16 Expressions, Action Units | Web              |
| FER2013    | Images   | 35,867     | 6 Expressions           | Web                |
| JAFFE      | Images   | 219        | 6 Expressions           | Lab                |

Fig. 2. Example datasets commonly used in computer vision FER. In contrast, Fig. 1c shows data used in Suess et al. (2015) for experiments on human knowledge-augmented face perception. They are from the Radboud Faces Database (Langner et al., 2010) which contains stimuli used in many psychological studies (Schindler et al., 2019). Suess et al., 2015 provide additional affective person-related context information is also contained in this dataset. Superscripts: 1Dhall et al. (2011), 2Mollahosseini et al. (2017), 3Zafeiriou et al. (2017), 4Yin et al. (2008), 5Lee et al. (2019), 6Lucy et al. (2010), 7Kosti et al. (2019), 8Benitez-Quiroz et al. (2017), 9Goodfellow et al. (2013), 10Lyons et al. (2020).
3.1.3. Deep learning FER methods

Many theoretical foundations of ANNs were developed in the 80s and 90s (LeCun, Haffner, Bottou, & Bengio, 1999; Rumelhart, Hinton, & Williams, 1986). But only recently, deep learning has begun to dominate AI when training deep networks (i.e. consisting of multiple stacked layers) became feasible by exploiting the highly parallel structure of graphics cards (Krizhevsky, Sutskever, & Hinton, 2012). An important characteristic of a neural network is the architecture. A typical Convolutional Neural Network (CNN; specialization of ANNs; see Fig. 3) consists of convolutional layers (the main building blocks) and non-linearities, e.g., ReLU (Nair & Hinton, 2010).

**Artificial Neural Networks (A)**

Artificial neural networks (ANNs) essentially are high-dimensional function approximators. Each neuron realizes a non-linear function and consists of a weight vector \( w \) and a bias \( b \). These are the trainable parameters which are adjusted during training. The output of a neuron is computed as

\[
f(x) = g(wx + b)
\]

for some input \( x \), where \( g \) is the activation function that implements the non-linearity. Non-linearity of the neurons of a layer is essential since otherwise a neural network with multiple layers would collapse into a single linear function.

**Network Training (B)**

Training a neural network adjusts the weights of the neurons to produce some desired output. Updating weights requires a loss function which captures the error of the prediction of the network \( y' \) w.r.t. the desired ground truth \( y \). For example, it can penalize the quadratic distance between the two:

\[
L(y, y') = (y - y')^2
\]

In the beginning of the training process, the weights of the network are initialized randomly. A training sample is passed through the network and the error is computed. The gradient of the error \( \Delta L \) now acts as the weight update and is subtracted from the current weights - this can be seen as taking a step towards the lower loss. The gradient is passed through the network using the chain rule to compute the gradient for previous layers. This process is called (error-) backpropagation.

**Deep Learning (C)**

Deep learning is the term for training neural networks consisting of many layers. For many years, it was not possible to train such networks due to their computational complexity. But with an increase in computational power and the possibility to train neural networks on graphics cards, this task became feasible allowing multiple stacked layers. Such deep networks were able to improve computer vision performance by an order of magnitude.

**Convolutional Neural Networks (D)**

As can be seen in the figure in the upper right corner, a "classic" (fully-connected) artificial neural has dense connections, i.e. every neuron is connected to every previous neuron. This makes the network computationally inefficient and memory-intensive. One solution to this issue are convolutional neural networks (CNNs) which are loosely inspired by the human visual system. In CNNs, each layer has only a small set of filters (the weights) which convolve the input locally. This allows to perform the computations of a layer as a matrix multiplication which is highly parallelizable and can be executed efficiently in parallel on graphics cards. This yields unprecedented speeds in contrast to fully-connected networks.

**Recurrent Neural Networks (E)**

Recurrent neural networks (RNNs) denotes a neural network architecture that retains a state and thus allows recurrent processing with a simple memory model. RNNs excel at processing sequential input which develops over time, like speech, text or videos. They retain an internal state which acts as memory of past inputs.

![Fig. 3. Background on artificial neural networks. Own illustration, face image from Lyons et al. (2020).](image-url)
The previously discussed approaches did not address integrating context information and knowledge into FER algorithms in a top-down, dynamic way. But recently, a growing body of publications have diverged from developing ANNs with ever more parameters to increase accuracy on datasets and have turned to psychology and cognitive science to draw inspiration from human perception and cognition. Likewise, researchers from psychology and cognitive science have realized a potential of machine learning applications in their field. In the following, we highlight some of the recent developments from a computer vision perspective: First, we discuss recent opinions on how to advance AI using insights from psychology and cognitive science but also how AI could help understand human cognition better. We then examine concrete suggestions how to make ANNs context-sensitive in a top-down way and close this chapter discussing such networks for FER.

3.2. Towards knowledge-augmented FER

The previously discussed approaches did not address integrating context information and knowledge into FER algorithms in a top-down, dynamic way. But recently, a growing body of publications have diverged from developing ANNs with ever more parameters to increase accuracy on datasets and have turned to psychology and cognitive science to draw inspiration from human perception and cognition. Likewise, researchers from psychology and cognitive science have realized a potential of machine learning applications in their field. In the following, we highlight some of the recent developments from a computer vision perspective: First, we discuss recent opinions on how to advance AI using insights from psychology and cognitive science but also how AI could help understand human cognition better. We then examine concrete suggestions how to make ANNs context-sensitive in a top-down way and close this chapter discussing such networks for FER.

3.2.1. Intersections of cognitive sciences and AI

It is an ongoing discussion how AI can be improved with inspiration from human cognitive processes (e.g., Sinz et al., 2019; Goyal & Bengio, 2020; Lake et al., 2017; George, 2017; Sagel et al., 2020; Hassabis et al., 2017). Sinz et al. (2019) argue that neural networks do not take expressiveness power in the sense of the expressiveness of the functions they can model but do not incorporate the right inductive biases. Instead of making the next generation of neural networks deeper, the path researchers should take according to Sinz et al., 2019 is towards incorporating such biases. They propose a bias similar to the prior in Bayesian theory. The authors demand that under ambiguous input, the model should prefer certain interpretations over others depending on the context information available. This is in line with human perception dynamics as described in Hallmark H2 (reducing perceptual uncertainty; Section 2.2.2). The right set of (architectural) biases is also seen as the major requirement for the next generation of AI algorithms by various other researchers (e.g., Goyal & Bengio, 2020; George, 2017; Hassabis et al., 2017).

The reverse direction—that ANNs can help understand the human mind and brain better—has also been advocated (e.g., Ma & Peters, 2020; Chichy & Kaiser, 2019; Storrs & Kriegeskorte, 2019; Perconti & Plebe, 2020; Peterson, 2018; Lindsay, 2020). Sinz et al. (2019) propose that neural networks can be used to generate new stimuli for biological systems, as has been implemented in Walker et al. (2018). Generating stimuli for humans using neural networks could help understand biases and priors in processing of the human brain (see Discussion in Section 4 for further elaboration). A major drawback of deep neural networks is that the neurons cannot be directly compared to neurons in biological systems which are more complex by some orders of magnitude (Storrs & Kriegeskorte, 2019; Chichy & Kaiser, 2019). The backpropagation algorithm which is used to train artificial neural networks is biologically not entirely plausible since until now no neural equivalent to a back-propagated error has been discovered in humans (Storrs & Kriegeskorte, 2019). Still, neural networks have successfully been employed to predict brain activity (Perconti & Plebe, 2020; Lindsay, 2020) and were even used to generate optimal stimuli that yielded the largest response in neurons of macaques (Ponce et al., 2019).

3.2.2. How to approach knowledge-augmentation in AI

We now move from literature on how cognitive sciences and AI research can inform one another to work proposing how to achieve knowledge-augmented machine learning algorithms theoretically but also practically (e.g., Montoya, 2020; Cortese, De Martino, & Kawato, 2019; Kursuncu et al., 2020). Montoya (2020) suggests an abstract model that learns a representation of its environment that is continuously updated based on new observations. Cortese et al. (2019) present a processing model of regions in the human brain and their interplay. They emphasize that recurrence is a key aspect in the human brain which therefore should be an integral part of synthetic models.

Kursuncu et al. (2020) propose a theoretical model of a neural network that employs a so-called knowledge engine. The knowledge engine retains part of a knowledge graph and continuously updates the retained part based on the error of the overall system. This architecture is yet to be implemented and real experiments to be conducted, hence no practical results can vouch for its effectiveness.
Kursuncu et al., 2020’s theoretical model conforms to Montoya, 2020’s idea of a system that continuously builds a model of its environment, although updates happen only during training and not during inference in the former. Representing relations between humans in a knowledge graph to influence how a neural network perceives a face is worth investigating—especially with a focus on how to update the graph also during inference—and is in alignment with humans continuously updating their prior knowledge about the world.

Von Rueden et al., 2020 have published a comprehensive review and developed a taxonomy of incorporating knowledge into machine and deep learning. They use the term informed machine learning. Their taxonomy categorizes approaches along three aspects: The source of knowledge (e.g., natural sciences, commonsense knowledge, etc.), how knowledge is represented (e.g., by algebraic equations, knowledge graphs/semantic networks, etc.) and how it is integrated into the algorithm (e.g., through training data, hypothesis set, etc.). Von Rueden et al., 2020’s survey is drafted from a computer vision perspective primarily and lacks a psychology and cognitive science stance.

In their work on object classification, Zhang, Tseng, and Kreiman (2020) built a ANN architecture that consists of two branches—one processing the visual surroundings and one processing the actual facial expression—which feed their computed features into an LSTM. The authors conduct a comprehensive comparison with human performance in object classification, depending on the context. Their network exhibits performance patterns similar to humans. We argue that using LSTMs with context branches can be regarded as a very simple form of top-down implementation of context-sensitivity.

3.2.3. Efforts of integrating knowledge-augmentation in FER

Kosti et al. (2017) and Lee et al. (2019) (the authors of the EMOTIC and CAER datasets, see Section 3.1.1) recognize the fact that facial expressions substantially depend on the surrounding visual context and focusing only at the face is not sufficient to properly recognize expressions. They both present neural nets that consist of two branches: One branch is responsible for processing the face and a second branch uses the whole image as input to take the surroundings into account. A so-called fusion network then combines the estimated expression and the features obtained from the context information in a bottom-up way to produce the final prediction.

Surace, Patakchiola, Sönnmez, Spataro, and Cangelosi (2017) go in a similar direction. They draw on Bayesian Networks to recognize emotions in pictures showing groups of people more accurately. The BN takes a scene description as input, obtained from a so-called top-down module, and the averaged facial expressions of all faces in the image, estimated by a neural network. The BN then computes the posterior probability of the image belonging to one of the following categories: neutral, positive, negative. This boils down to reducing ambiguity in facial expressions (Hallmark H2, reducing perceptual uncertainty; Section 2.2.2).

Mermillod et al. (2018) employ so-called Simple Recurrent Networks (SRNs; Elman, 1990) to predict future facial expressions in videos based on the current and past frames (related to Hallmark H4, guiding prediction of future input; Section 2.2.4). The authors find that it can outperform a Multi-Layer Perceptron (MLP; simple dense neural network) especially in frames showing ambiguous expressions. The prior induced by past frames which is contained in the SRN reduces this ambiguity (related to Hallmark H2, reducing perceptual uncertainty; Section 2.2.2). Mermillod et al., 2018’s MLP and SRN architectures are rather shallow and simple, state-of-the-art neural networks will most probably outperform both. But the focus of their work is on the comparison between non- and recurrent networks, not on providing the overall best accuracy. Recurrency is also highlighted as a key consideration by some of the previously discussed publications (Cortese et al., 2019; Sinz et al., 2019; Lindsay, 2020).

As we have seen, there exist first approaches that draw on insights on human perception. But integrating context information and knowledge into ANNs in a top-down fashion has only just begun to become the center of attention. In the next chapter we want to outline some directions for future research.

4. Discussion

Our society is projected to integrate more and more synthetic agents that engage and communicate with humans in everyday situations, such as social robots used in care, stores, public transportation, or schools (Saygin, Chaminade, Ishiguro, Driver, & Frith, 2012; Bartneck, Suzuki, Kanda, & Nomura, 2007). To increase the chances of succeeding with this integration, synthetic agents should be able to correctly identify and appropriately engage with the intentions of human interaction partners (Kirtay et al., 2020). This requires them to adapt to human verbal and nonverbal communication (Wudarczyk et al., 2021), perception, and evaluation all of which are situated in and draw on semantic, affective, and embodied contexts. Current synthetic models, however, are not built for this type of flexibility. In this paper so far, we have elaborated on the qualitative differences between human facial expression perception and state-of-the-art computer vision, specifically concerning the ability to take into account knowledge and context information during information processing. We have reviewed evidence of knowledge-augmented face perception in humans, for instance to make sense of ambiguous expressions, or predict how a facial expression will develop across time. These effects are consistent with the Bayesian brain view of perception: knowledge and context inform priors that influence perceptual inference through top-down predictions. Computer vision, on the other hand, employs networks that—once trained—rely mostly on feedforward information processing. Current FER systems have largely been developed to optimize ground truth categorizations of more or less isolated facial expression images. However, even the ground truth in the used datasets is not a reliable reflection of real world situations, because the meaning of facial expressions is inherently context-dependent, both when producing and reading facial expressions. Therefore, chances are high that FER is systematically misaligned with human sociality and perception, which hampers useful and ethical applications of FER in human–machine interaction (see Section 4.3). In this final part, we aim to integrate both viewpoints of our review and discuss the prospects of starting to close this gap in an interdisciplinary endeavour. In our opinion, top-down effects on perception and the Bayesian brain framework lend promising inspirations to redesign FER. Ideally, combining the Bayesian brain framework with...
computer vision will eventually result in an epistemic loop (see Section 4.2), in which artificial networks inspired by human perception will serve as scientific tools to improve our understanding of facial expression perception and how it is shaped by prior knowledge.

4.1. Interpreting data in context

Machines’ capabilities can of course not measure up to human understanding, which incorporates knowledge shaped by evolution, rich embodied, emotional, social and enactive experience, linguistic and conceptual knowledge etc. However, deep neural networks excel at picking up even subtle statistical regularities in data they are trained on. This sometimes also leads to problems ranging from somewhat comical effects, like confounding context (e.g., green pastures) highly correlated with the presence of a certain object (e.g., sheep), with the object itself (i.e., labeling green pastures as sheep; Beery, van Horn, & Perona, 2018; Do neural nets dream of electric sheep, 2018), to serious issues like racial bias in face recognition (Drozdowski, Rathgeb, Dantcheva, Damer, & Busch, 2020). On the one hand, these examples demonstrate bugs of deep nets and a lack of conceptual understanding. On the other hand, there is potential for knowledge-augmentation: if represented explicitly as probabilistic prior information, the meaning embedded in such statistical regularities could inform perceptual predictions. Instead of directly labeling a green pasture as sheep, a knowledge-augmented network would be able to represent the fact that sheep are likely to be found on a green pasture as a prior, and bias top-down predictions accordingly (see Singh, Su, Jin, & Jiang, 2019, for a step in this direction).

We argue that such a network would have the potential to yield better generalization and become more robust against unwanted bias (e.g., induced by skewed training datasets) notably, by introducing (flexible) top-down bias. As reviewed in Section 3.2.2, there are approaches to formally represent semantic concepts, for instance in knowledge graphs. So in principle, networks are capable of learning meaningful connections, like the affective content of a scene (e.g., looking at a coffin, as illustrated in the Kuleshov effect, Fig. 1a) or a sentence describing a person’s previous behavior (as in Suess et al., 2015; Baum et al., 2020; Baum & Abdel Rahman, 2021). For instance, networks that have learned to associate certain contexts with matching facial or bodily expressions (Cowen et al., 2021) could be combined with networks processing facial expressions and inform their perceptual predictions.

4.2. An epistemic loop between cognitive science and synthetic FER

We have so far presented several arguments why computer vision should take inspiration from human predictive face perception. But crucially, progress on the synthetic side should feed back to cognitive science, bearing the potential for a better understanding of face perception in humans as well. Even though ANNs are limited in simulating the sheer complexity of brain dynamics, they add several key advantages: for instance, they process large amounts of data fast and cheaply, and allow to quickly test novel experimental manipulations for proof-of-principle demonstrations (Cichy & Kaiser, 2019). In psychophysical and neurocognitive studies, experimenters often choose a limited amount of faces (e.g., from well-controlled scientific image databases) displaying a limited range of facial expressions (e.g., a subsample of posed basic emotions) paired with hand selected range of contexts based on researchers’ experience and intuitions (e.g., negative, neutral and positive person-related information; Suess et al., 2015). In reverse-correlation studies, for example, participants need long sessions to respond to variations of the same base image several hundred times, such that studies yield classification images for only one or a few faces (Dotsch & Todorov, 2012). Neural networks could apply principles derived from such limited test cases to much larger face datasets. They could ultimately propose and even create novel, model-based stimuli to be validated with human participants. Such stimuli might be better optimized to experimental manipulations (Ponce et al., 2019) and achieve more nuance by combining a broader range of contexts and facial expressions than stimuli subjectively chosen by experimenters. This way, cognitive scientists might uncover novel contextual influences on face perception that would otherwise be missed.

4.2.1. Applying insight into the neural time course of top-down effects

Beyond the potential to discover new test cases, knowledge-augmented synthetic systems could be used to better understand the mechanisms by which context influences perception. For instance, it could turn out that some network layers exhibit more context-sensitivity than others. Since network layers are typically associated with the processing of specific features, model-based predictions could be made for neurocognitive experiments about which types of knowledge or context information influence which levels of visual processing in the brain (e.g., processing of low- vs. high-level features). By gaining access to a broader range of contexts than testable with human participants, artificial networks could then help to predict the impact of previously untested contexts.

One main lesson from the review of knowledge-augmented face perception in humans (specifically hallmark H1 discussed in Section 2.2.1) and and current synthetic approaches is that successful models will probably need recurrency. This is based on a few observations: First, the neural time course of top-down effects suggests entry points at several stages throughout the visual processing pipeline (early, intermediate, and late, see Section 2.2.1). Second, explicit modeling of brain dynamics during visual processing appears to require recurrent neural networks (Kietzmann et al., 2019). And third, recurrent networks are particularly successful in processing dynamic facial expressions, where predictions informed by previous states are beneficial (Mermillod et al., 2018). In the Bayesian brain view, human perception gains efficiency and flexibility by continuously weighing incoming sensory input with predictions based on perceptual priors that are informed by knowledge. Accordingly, recurrent neural networks could be designed to integrate information from other sources into priors and pass it on as top-down predictions at multiple stages or different iterations of processing. This could offer much more context-sensitivity akin to human perception.

The remaining hallmarks of knowledge-augmented face perception (2.2.2–2.2.5) provide concrete starting points for designing and experimenting with artificial systems. In the following, we make programmatic remarks on how they could be put to use. Ideally, the
same architectural principles will eventually afford several or all of the hallmarks of knowledge augmentation. For some of the aspects, additional psychological studies are still needed.

4.2.2. Reducing uncertainty in artificial networks

Regarding the use of top-down information to reduce uncertainty (Hallmark H2; Section 2.2.2), networks should be probabilistic, as assumed in the Bayesian brain framework. In some form, they should represent the likelihood of potential input stimuli or features in an image based on priors and the (un)certainty of their current predictions. This would allow, for instance, to draw more heavily on useful context information when the bottom-up input is less informative. In artificial agents, this could be extended with additional components of active inference, for instance by interacting with the environment or interaction partners to actively test predictions and reduce prediction error. Networks could in turn be trained to simulate effects observed in humans, like reading emotions into neutral facial expressions based on affective knowledge, or using contextual information to disambiguate subtle or easy-to-confound facial expressions like fear and surprise.

4.2.3. Potential of enhancing relevant features

As discussed earlier (Hallmark H3; Section 2.2.3), there is evidence that top-down factors like conceptual knowledge or linguistic categories may warp perceptual space by enhancing relevant features (Bar, 2004; Lupyan, 2012). Psychophysical studies have attempted to identify features most relevant in recognizing certain facial expressions (e.g., Schyns et al., 2002; Schyns, Petro, & Smith, 2007). To our knowledge, it has not been studied yet how the informativeness of specific features for classifying facial expressions changes with context and knowledge, but potential top-down influences should be studied in future experimental work. Data from human participants could be used to model top-down feedback in recurrent artificial neural networks to guide enhancement of relevant features, for instance through an attention mechanism, by biasing feature-selectivity in certain network layers. This would enable comparisons between human and machine diagnostic features and, when aligned properly, to gain new insights into human perception.

4.2.4. Leveraging top-down feedback to guide prediction

Concerning the use of top-down feedback to guide prediction (Hallmark H4; Section 2.2.4), can context-sensitivity, as discussed here, help artificial neural networks better predict what happens next in a visual scene, such as a developing dynamic facial expression? This will be especially relevant for real-world applications, for instance in social robots. As demonstrated by Mermillod et al. (2018), recurrent networks have advantages in processing dynamic expressions. Knowledge or context information could be integrated at each updating step as the network recursively predicts upcoming input (Kirtay et al., 2020). Taking inspiration from Zhang et al. (2020), a context-sensitive neural network could follow a two-stream design. Instead of processing objects, one branch could receive the facial expression as input and one the visual context. This architecture allows easy extension to other context information types while exhibiting top-down dynamics to some extent, due to its LSTM.

4.2.5. Visualized biases as perceptual priors

Finally, evidence from experiments on how semantic context shapes visual appearance (Hallmark H5; Section 2.2.5) could be used to implement top-down biases in synthetic systems. For instance, classification images (Fig. 1d) created by human participants can be proxies for knowledge-informed perceptual priors (Brinkman et al., 2017), which neural networks could use as training data. Further, generative neural networks could be trained to perform reverse correlation tasks and create classification images on their own, based on much larger datasets of semantic contexts and base images. These classification images could then be rated by human participants for validation. This way, networks could learn to predict what a facial expression would look like to human observers when seen in light of a given context. Such predictions could then be visualized to give human observers feedback about their potential biases.

4.3. Ethical considerations

Systems supposedly capable of automatically reading people’s emotions are already being marketed for diverse applications (like care, hiring, monitoring students surveillance; Feldman Barrett et al., 2019). However, as discussed extensively in this paper, current systems designed to read emotional states mostly ignore the fact that they are inherently and richly contextualized. Therefore, such systems suggest a false sense of accuracy or even objectivity. Another major criticism concerns bias introduced by skewed training data that intrinsically and inflexibly lead to classification errors, most notably including racial bias (Drozdowski et al., 2020) or gender bias (Domnich & Anbarjafari, 2021). Even though our focus is on basic science and a better understanding of facial expression perception, we believe that ethical considerations about potential applications should be kept in mind. Generally, a key cornerstone of ethical use of AI is that technologies should adapt to and support human behavior, and not vice versa. This is why FER systems should be better aligned to how humans perceive faces, and research should move away from the idea of trying to read affective states from out-of-context facial expressions. Augmenting FER with knowledge and making it adaptable to context information could make them less vulnerable to biases resulting from skewed training data. So far, such biases introduced by humans affect machine learning algorithms mainly implicitly. Making FER context-sensitive bears the risk of allowing new biases in. But FER systems capable of explicitly representing contextual effects aligned with human perception might ultimately also be helpful in alleviating human bias. Top-down effects can make people susceptible to influences from information based on prejudice or untrustworthy sources (Baum & Abdel Rahman, 2021; Baum et al., 2020). Knowledge-augmented synthetic systems could simulate human perception from different viewpoints, i.e. with different priors, and provide feedback about potential biases.
More similar underlying information processing could provide a common reference frame shared with humans and foster communication and more intuitive understanding (Friston & Frith, 2015; Kirtay et al., 2020). Systems able to make probabilistic predictions concerning visual input might also avoid false impressions of accuracy by signaling the (un)certainty of their perceptual inferences. They might also actively reduce uncertainty through additional data sampling or interaction (e.g., in a social robot).

4.4. Conclusions

Developing artificial models for FER aligned with human knowledge-augmented perception could add a crucial scientific tool to the interdisciplinary endeavour of better understanding perception of facial expressions. Inspired by the Bayesian brain-framework, artificial neural networks could be designed to put visual input into a meaningful context based on predictive information outside of the face itself, such as prior conceptual and affective knowledge. Potential progress in computer vision includes making FER more flexible, efficient, robust against skewed datasets, and better in real world situations that involve dynamically unfolding facial expressions. Progress in implementing predictive processing in synthetic systems can feed back to empirical research in cognitive sciences in an epistemic loop. Model-based experimentation on a far broader range of datasets and contexts than testable in the lab will help to assess the true scope of knowledge-augmented (face) perception.

5. Glossary

**Cognitive penetrability of perception** The (contested) notion that higher-level cognitive factors—like knowledge, beliefs, expectations, or language—influence perception. In the classical modular view, perception is informationally encapsulated, such that cognition only influences what happens before (e.g., where to direct attention) or after perception (e.g., how to interpret its output). According to different versions of the cognitive penetrability view, cognitive factors can influence (1) ongoing perceptual processing including early vision or (2) subjective perceptual experience (Macpherson, 2017). Recent electrophysiological studies into the time-course of potential top-down effects suggest influences on early perceptual processing (e.g., Maier & Abdel Rahman, 2019; Maier & Abdel Rahman, 2018).

**Context information, context-sensitivity and knowledge-augmentation** We use context information as a term for information contained in the situation in which visual input is encountered. Context-sensitivity means that context information can alter the perceptual outcome. Knowledge-augmentation means that human perception draws on top-down processing to integrate additional knowledge or context information that is not directly accessible in the bottom-up visual input. Knowledge includes, for instance, learned knowledge, beliefs, or linguistic structures. The goal for knowledge-augmented computer vision will be to provide artificial networks with the capabilities that top-down integration brings to human perception, in order to achieve more flexible and context-sensitive processing.

**Emotions vs. facial expressions** Emotions are complex psychological patterns in reaction to significant events that involve physiological, experiential and behavioral elements. Even though facial expressions often communicate emotions and intentions, there is no straightforward or universal relationship between the two. Configurations of facial movements do not uniquely map on the same emotions and do not generalize universally across different contexts and cultures (Feldman Barrett et al., 2019). Therefore, we describe facial expressions as inherently contextualized in this review.

**Top-down vs. bottom-up** The brain contains bidirectional connections implementing feedforward and feedback information processing. Anatomically speaking, bottom-up processing can be referred to as information flow from lower to higher cortical areas, with feedback loops only between neighboring, interconnected areas (Rauss & Pourtois, 2013). Accordingly, top-down modulation occurs when feedback from higher cortical areas that are further removed influences information processing at a lower level. In more cognitive terms, top-down modulation means that higher-order representations (e.g., the meaning of a word) influence lower-order processing (e.g. perception of an individual letter or sound). According to the Bayesian brain-framework, top-down information flow carries predictions, while prediction errors are propagated bottom-up (Clark, 2013; Friston, 2009).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdel Rahman, R. (2011). Facing good and evil: Early brain signatures of affective biographical knowledge in face recognition. Emotion, 11(6), 1397–1405. https://doi.org/10.1037/a0024717

Abdel Rahman, R., & Sommer, W. (2008). Seeing what we know and understand: How knowledge shapes perception. Psychonomic Bulletin & Review, 15(6), 1055–1063. https://doi.org/10.3758/PBR.15.6.1055

Adams, R. B., & Kleck, R. E. (2003). Perceived gaze direction and the processing of facial displays of emotion. Psychological Science, 14(6), 644–647. https://doi.org/10.1046/j.0956-7976.2003.psci.1479.x

Adams, R. B., & Kleck, R. E. (2005). Effects of Direct and Averted Gaze on the Perception of Facially Communicated Emotion. Emotion, 5(1), 3–11. https://doi.org/10.1037/1529-1055.5.1.3

Ahonen, T., Hadid, A., & Pietikäinen, M. (2004). Face Recognition with Local Binary Patterns. In T. Pajdla, & J. Matas (Eds.), Computer Vision - ECCV 2004 (pp. 469–481). Springer. https://doi.org/10.1007/978-3-540-24670-1_36.

Altman, N. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. https://doi.org/10.1080/00031305.1992.10475879.
Wudarczyk, O. A., Kirtay, M., Pischedda, D., Hafner, V. V., Haynes, J.-D., Kuhlen, A. K., & Abdel Rahman, R. (2021). Robots facilitate human language production. *Scientific Reports, 11*(1), 16737. https://doi.org/10.1038/s41598-021-95645-9. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals

Xu, M., Li, Z., Diao, L., Fan, L., & Yang, D. (2016). Contextual Valence and Sociality Jointly Influence the Early and Later Stages of Neutral Face Processing. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01258

Yacoob, Y., & Davis, L. (1996). Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(6), 636–642. https://doi.org/10.1109/34.506414

Yin, L., Chen, X., Sun, Y., Worm, T., & Reale, M. (2008). A High-Resolution 3D Dynamic Facial Expression Database. 6.

Yoshikawa, S., & Sato, W. (2008). Dynamic facial expressions of emotion induce representational momentum. *Cognitive, Affective, & Behavioral Neuroscience, 8*(1), 25–31. https://doi.org/10.3758/CABN.8.1.25

Zafeiriou, S., Kollias, D., Nicolaou, M.A., Papaioannou, A., Zhao, G., & Kotsia, I. (2017). Aff-Wild: Valence and Arousal ‘In-the-Wild’ Challenge. In 2017 *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1980–1987). https://doi.org/10.1109/CVPRW.2017.248.

Zhang, M., Tseng, C., & Kreiman, G. (2020). Putting Visual Object Recognition in Context. In 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12982–12991). https://doi.org/10.1109/CVPR42600.2020.01300.