A Deep Neural Model for Emotion-driven Multimodal Attention

German I. Parisi[1], **Pablo Barros**[1], **Haiyan Wu**[2], **Guochun Yang**[2], **Zhenghan Li**[2], **Xun Liu**[2] **and Stefan Wermter**[1]
[1] Knowledge Technology Institute, Department of Informatics, University of Hamburg, Germany
[2] CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing, China

Abstract

The ability of the brain to integrate multimodal information is crucial for providing a coherent perceptual experience, with perception being modulated by the interplay of different cortical and subcortical regions in the brain. Recent research has shown that affective stimuli play an important role in attentional mechanisms, with behavioral studies supporting that the focus of attention in a region of the visual field is increased when affective stimuli are present. This work proposes a deep neural model that learns to locate and recognize emotional expressions modulated by an attentional mechanism. Our model consists of two hierarchies of convolutional neural networks: one to learn the location of emotional expressions in a cluttered scene, and the other to recognize the type of emotion. We present experimental results processing facial and body motion cues, showing that our model for emotion-driven attention improves the accuracy of emotion expression recognition.

Introduction

Audiovisual spatial attention allows animals and humans to process relevant environmental stimuli while suppressing irrelevant information. Several brain areas and neural mechanisms have been identified to be involved in the processing of spatial attention during perception (Driver 2001). For instance, it has been found that the superior colliculus (SC), a subcortical part of the brain, plays a crucial role in spatial attention in terms of target selection and producing motor responses such as head-eye movements (Krauzlis, Lovejoy, and Zénon 2013). The integration of audiovisual stimuli in the SC has been extensively investigated from a neurophysiological perspective (Ursino, Cuppini, and Magosso 2014), with different computational approaches modeling the integration of multiple perceptual cues for triggering spatial attention in line with neurobehavioral evidence, e.g. with the use of a self-organizing neural architecture (Bauer, Magg, and Wermter 2015).

The SC is connected to higher cortical areas such as the visual and the auditory cortex, which are both able to process information events that unfold over larger time scales such as body actions and speech. Neurons selective to actions in terms of time-varying patterns have been found in a wide number of brain structures, such as the superior temporal sulcus (STS), parietal, premotor and motor cortex (Giese and Rizzolatti 2015). It has been argued that the STS in the mammalian brain may be the basis of an action-encoding network with neurons driven by audiovisual stimuli (Barraclough et al. 2005). Thus, the STS area is hypothesized to be an associative learning device for linking unimodal representations and accounting for the mapping of naturally occurring, highly correlated features such as body pose and motion, the characteristic sound of an action and linguistic stimuli.

Top-down connectivity from cortical areas is used by the SC to modulate attentional shifts. For instance, converging findings suggest that selective attention is modulated by the affective significance of sensory inputs (Vuilleumier 2005). In particular, it has been argued that emotional salience has a direct influence on attention and that neural processes responsible for emotional attention may supplement and even compete with other top-down mechanisms of perception. Behavioral studies have shown that people pay more attention to emotional rather than neutral stimuli, and that these effects often are reflexive and involuntary, e.g. visual targets expressing an emotion such as happy or angry are found faster among distractors than targets without such emotional values (Williams, Mathews, and MacLeod 1996; Vuilleumier and Schwartz 2001). Together, these findings indicate that emotional salience has a strong role in capturing attention, and that this emotional bias is also subject to a set of different non-affective regulatory effects.

Different computational models have been proposed for the detection and recognition of emotional expressions (Arkin et al. 2003). Different cues may carry emotional information such as face expressions, sound (voice pitch and intensity), and body movements (Gu, Mai, and Luo 2013). The combination of these cues increases recognition accuracy (Castellano, Kessous, and Caridakis 2008), suggesting that models for the robust processing of emotional states should integrate multiple modalities for the meaningful processing of a set of available perceptual cues.

The goal of our research is to introduce a cortico-collicular architecture aimed to model multimodal attention and that accounts for the interplay between the SC and cor-

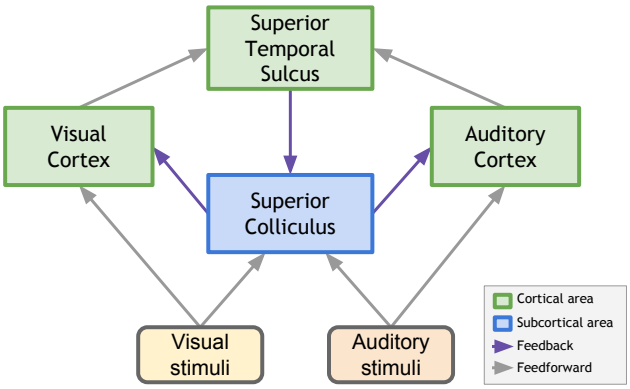

Figure 1: Diagram of the proposed cortico-collicular architecture for audiovisual inputs.

tical processing (Fig. 1). In particular, we investigate the use of convolutional neural networks (CNN) for stimulus localization (Speck et al. 2016) and CNN architectures for obtaining mid-level feature representations and spatial information from raw audiovisual stimuli. This representation serves as input for a cortical visual-auditory integration model to learn inherent spatio-temporal structure, e.g. recognition of emotions from face, body features, and auditory cues. The output from cortical areas is used as feedback for the SC model, thereby modulating attentional shifts as an interplay between bottom-up and top-down processing mechanisms.

We present experiments in the visual domain, where a deep neural architecture is used to learn the location of emotional stimuli (Barros et al. 2016). Emotion-driven attention is the input to the cortical architecture responsible for emotion recognition. Our experimental results show our attention model improves the performance of the recognition of emotional expressions.

Methods

Our learning architecture comprises the interaction of two modules: the attention model for locating emotionally salient areas in an image, and the recognition model for classifying emotions from face and body motion. Both models use extended CNN architectures with the aim to classify emotion classes from raw images. Although the input is composed of a single image sequence containing both face and body motion, the model will autonomously learn separate cue-specific filters. The recognition model classified emotional expressions modulated by emotion-driven attention. A diagram of our neural architecture is illustrated in Fig. 2.

For the recognition model, we extended a Cross-channel Convolution Neural Networks (CCCNN) (Barros, Weber, and Wermter 2015). This approach introduces the use of multi-channel and cross-channel learning for emotion expression recognition, and presented competitive performance and good generalization capabilities. The proposed model is able to learn simple and complex features and to model the dependencies of these features in a sequence. Using a multichannel implementation, it is possible to learn different features for each stimulus. We use the concept of cross-channel learning to deal with differences across modalities. This allows us to have regions of the network specified to learn features from face expressions and body movements but the final representation integrates both specific features.

The attention model uses the filtering capability of the convolution layers to learn the location of emotional expressions conveyed by two visual cues: face expressions and body movements. To modulate our recognition model, we train an attentional model that comprises a CNN architecture to distinguish between neutral and happy expressions conveyed by facial features and body movement. In contrast to traditional CNN learning models using discrete target labels for modulating the learning process, we use probability distributions that allow the model to estimate the location of interest, i.e. the region in the image that triggers selective attention. The attention model has two output layers, each one responsible for describing positions on the 2D visual field. An example of an emotionally salient region of an image and network output is shown in Fig. 3. Furthermore, using a probability distribution allows faster convergence with respect to having precise image coordinates.

To reduce the necessity of many layers, we use shunting inhibitory fields (Fregnac et al. 2003) in the last layers of our attention model. The shunting neurons act as an over specification tool, which makes the filters in a layer learn complex patterns. When applied to low-level features such as edges and contours, the shunting neurons tend to destroy the generalization properties of deep neural networks. However, when applied to high-level features with a very abstract specification, the shunting neurons tend to filter noise and learn only the most relevant features. Interestingly, after using teaching signals with only one emotional expression in the image, experiments have shown that the model is able to produce congruent probability distributions for more than one expression present in the scene.

To integrate the CCCNN and the CNN models, we connect the filters in the second convolutional layer of the recognition model with the attention model. This means that our attention model feeds specific facial and movement features to the second layer of the CCCNN. The second layer was chosen because the face channel already extracts features which are similar to the ones coming from the attention model, thus very related to the final facial features. The body motion channel still needs one more convolution layer to learn specific movement features.

Experiments

We evaluated our system with a bi-modal face and body benchmark dataset (FABO) (Gunes and Piccardi 2009), showing that the combination of facial features and body movements significantly improve the detection of emotion-relevant areas in the image. This corpus contains recordings of the upper torso of 23 subjects while performing 11 different emotion expressions. We extracted the face expression and body movement of the FABO corpus and located it in a random position in our meeting scene background. The

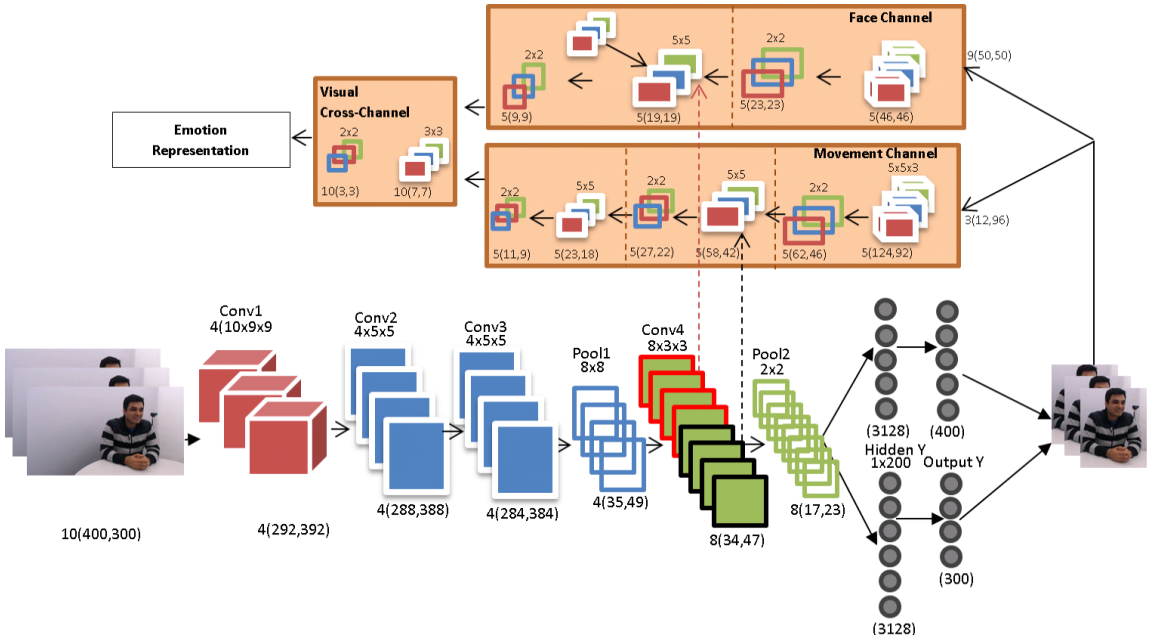

Figure 2: Our attention modulation model, which uses the specific features learned from the attention architecture as input for the specific channels of deeper layers of the CC-CNN. In this figure, the red dotted line represents the specific facial feature maps and the black dotted line represent the movement feature maps.

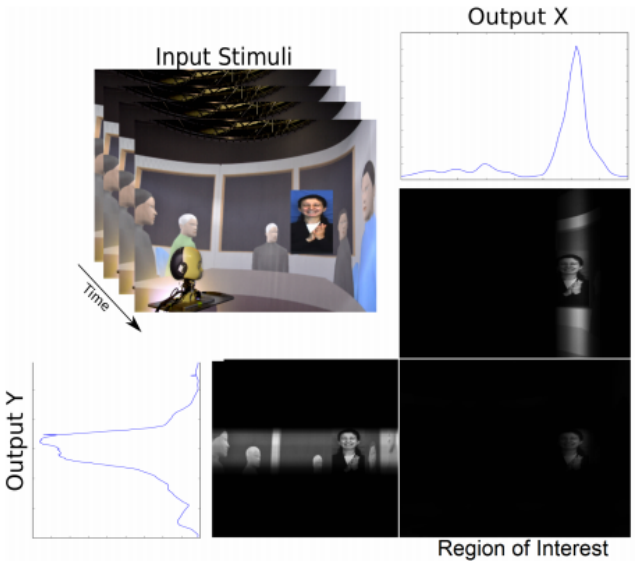

Figure 3: Example of the output of the attention model for an image with a happy expression.

Table 1: Reported accuracy and standard deviation (%) for the visual channels of the CCCNN trained with the FABO corpus with and without emotion-driven attention.

Class	No attention	With attention
Anger	95.9 (1.3)	95.7 (1.1)
Anxiety	91.2 (1.6)	95.4 (0.7)
Uncertainty	86.4 (1.1)	92.1 (1.3)
Boredom	92.3 (1.7)	90.3 (0.8)
Disgust	93.2 (1.8)	93.3 (1.6)
Fear	94.7 (0.6)	94.5 (0.7)
Happiness	98.8 (0.2)	98.0 (0.3)
Negative Surprise	99.6 (0.1)	99.7 (0.2)
Positive Surprise	89.6 (1.1)	94.8 (0.2)
Puzzlement	88.7 (1.2)	93.2 (0.8)
Sadness	99.8 (0.1)	99.5 (0.3)
Average	93.65 (1.0)	95.13 (0.7)

expressions were the same size, while their position in the images were randomly selected.

The classification accuracy of our architecture using the 11 classes of the FABO dataset is shown in Table 1, using 70% of the data for training and 30% for testing. We performed the experiments 30 times and computed the average accuracy. Table 1 also shows the results obtained after training the recognition model with the FABO corpus modulated from the attention model. We can see that the average recognition rate increased from 93.65% to 95.13% with the

use of the attention modulation. Expressions such as Boredom, Fear and Happiness led to slightly smaller accuracy. Our interpretation is that these expressions probably rely on presented hand-over-face or very slight movements, which were determined by the recognition model but ruled out by the attention model.

Conclusion

We presented a deep neural model that learns to locate and recognize emotional expressions modulated by emotion-driven attention as supported by biological and behavioral studies (Barros et al. 2016). Our approach is based on convolutional neural networks using the filtering capability of the convolution layers to learn the location of emotional expressions conveyed by facial and body motion cues. This resulted in a model that uses unlabeled expressions to locate regions of interest around faces and body motion, especially when they convey affective information.

The obtained results motivate the extension of our current architecture for the integration of auditory information. We expect that the addition of auditory cues will increase the precision of the model and approximate our computational approach to neurobiologically motivated neural mechanisms for multimodal integration and attention. To integrate auditory information, a new decision layer would have to be included to identify the region of interest based on the auditory and visual information. This decision layer would integrate both spatial cues, one coming from our visual attention model, and the other coming from a model for sound source localization (Bauer, Magg, and Wermter 2015). An auditory convolution layer could be used to identify emotional aspects of the auditory signal, such as arousal and valence, and use this information in the integration level (Barros and Wermter 2016).

By modeling the underlying neural mechanisms of multimodal attention in terms of cortico-collicular interaction, we aim at reproducing behavioral responses measured by psychological studies on attentional shifts from audiovisual stimuli. Future studies may examine and improve our model comparing results with human behavior in emotional expression recognition. In addition, the study on neural imaging data is a possible direction to test the model.

Acknowledgments

This research was partially supported by the German Research Foundation (DFG) and the National Science Foundation of China (NSFC) in project Crossmodal Learning, TRR-169.

References

[Arkin et al. 2003] Arkin, R. C.; Fujita, M.; Takagi, T.; and Hasegawa, R. 2003. An ethological and emotional basis for human–robot interaction. *Robotics and Autonomous Systems* 42(3):191–201.

[Barraclough et al. 2005] Barraclough, N. E.; Xiao, D.; Baker, C. I.; Oram, M. W.; and Perrett, D. I. 2005. Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *Journal of Cognitive Neuroscience* 17(3):377–391.

[Barros and Wermter 2016] Barros, P., and Wermter, S. 2016. Developing crossmodal expression recognition based on a deep neural model. *Adaptive Behavior* 24(5):373–396.

[Barros et al. 2016] Barros, P.; Parisi, G. I.; Weber, C.; and Wermter, S. 2016. Emotion-modulated attention improves expression recognition: A deep learning model. *Neurocomputing, Forthcoming*.

[Barros, Weber, and Wermter 2015] Barros, P.; Weber, C.; and Wermter, S. 2015. Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction. In *Humanoid Robots (Humanoids), 2015 IEEE-RAS 15th International Conference on*, 582–587. IEEE.

[Bauer, Magg, and Wermter 2015] Bauer, J.; Magg, S.; and Wermter, S. 2015. Attention modeled as information in learning multisensory integration. *Neural Networks* 65:44–52.

[Castellano, Kessous, and Caridakis 2008] Castellano, G.; Kessous, L.; and Caridakis, G. 2008. Emotion recognition through multiple modalities: Face, body gesture, speech. In Peter, C., and Beale, R., eds., *Affect and Emotion in Human-Computer Interaction*, volume 4868 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 92–103.

[Driver 2001] Driver, J. 2001. A selective review of selective attention research from the past century. *British Journal of Psychology* 92(1):53–78.

[Fregnac et al. 2003] Fregnac, Y.; Monier, C.; Chavane, F.; Baudot, P.; and Graham, L. 2003. Shunting inhibition, a silent step in visual cortical computation. *Journal of Physiology* 441–451.

[Giese and Rizzolatti 2015] Giese, M. A., and Rizzolatti, G. 2015. Neural and computational mechanisms of action processing: Interaction between visual and motor representations. *Neuron* 88(1):167–180.

[Gu, Mai, and Luo 2013] Gu, Y.; Mai, X.; and Luo, Y.-j. 2013. Do bodily expressions compete with facial expressions? time course of integration of emotional signals from the face and the body. *PLoS ONE* 8(7):736–762.

[Gunes and Piccardi 2009] Gunes, H., and Piccardi, M. 2009. Automatic temporal segment detection and affect recognition from face and body display. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 39(1):64–84.

[Krauzlis, Lovejoy, and Zénon 2013] Krauzlis, R. J.; Lovejoy, L. P.; and Zénon, A. 2013. Superior colliculus and visual spatial attention. *Annual review of neuroscience* 36(1):165–182.

[Speck et al. 2016] Speck, D.; Barros, P.; Weber, C.; and Stefan, W. 2016. Ball localization for robocup soccer using convolutional neural networks. In *RoboCup Symposium*.

[Ursino, Cuppini, and Magosso 2014] Ursino, M.; Cuppini, C.; and Magosso, E. 2014. Neurocomputational approaches to modelling multisensory integration in the brain: A review. *Neural Networks* 60:141 – 165.

[Vuilleumier and Schwartz 2001] Vuilleumier, P., and Schwartz, S. 2001. Emotional facial expressions capture attention. *Neurology* 56(2):153–158.

[Vuilleumier 2005] Vuilleumier, P. 2005. How brains beware: neural mechanisms of emotional attention. *Trends in cognitive sciences* 9(12):585–594.

[Williams, Mathews, and MacLeod 1996] Williams, J. M. G.; Mathews, A.; and MacLeod, C. 1996. The emotional stroop task and psychopathology. *Psychological bulletin* 120(1):3.