Crossmodal Pattern Discrimination in Humans and Robots: A Visuo-Tactile Case Study

Focko Higgen[1], Philipp Ruppel[2], Michael Görner[2], Matthias Kerzel[2], Sven Magg[2], Norman Hendrich[2]

***Abstract*— The quality of crossmodal perception hinges on two factors: The accuracy of the independent unimodal perception and the ability to enhance this accuracy by crossmodal integration. In elderly, the ability for crossmodal perception diminishes. To research to which degree the impediment of these two abilities in elderly contributes to this diminishment, we replicate a medical study on visuo-tactile crossmodal pattern discrimination utilizing state-of-the-art tactile sensing technology and artificial neural networks. We explore the perception of each modality in isolation as well as the crossmodal integration. We show, that the integration of complex high-level unimodal features outperforms the comparison of independent unimodal classifications. Our work creates a bridge between neurological research and embodied artificial neurocognitive systems.**

I. Introduction

In our daily life, it is crucial to continuously process simultaneous input from different sensory systems to adapt to changes in our surrounding [1]. With aging, performance decreases in several cognitive domains, one primary domain being the adequate processing of incoming stimuli [2]. Impairments in successfully integrating information from different sensory systems might be one of the reasons for the challenges of the elderly in daily life. As the percentage of older people in the population increases, understanding of mechanisms of age-related declines and development of adequate support approaches gain more and more importance. Interestingly, the design of high-performing artificial systems for crossmodal integration is likewise one of the most significant challenges in robotics. Adapting a neurological experiment, aimed to determine difficulties of elderly in crossmodal integration can help to establish common grounds in human and robotic research and the mutual exchange of theory. On the one side, it will allow for evaluation of performance of artifical systems compared to humans with different abilities and help to adapt more biologically plausible and performant artificial neural network (ANN) models. On the other side, network models might help to understand reasons for poor performance in elderly humans and can be a basis for the development of assistive devices. We employ embodied neurocognitive models to evaluate different hypotheses of the contribution of unimodal processing and crossmodal integration for a specific visuo-tactile crossmodal differentiation task.

[1] with the Medical Center Hamburg-Eppendorf (UKE), Germany.
[2] with the University of Hamburg, Germany.
f.higgen@uke.de; {ruppel, goerner, kerzel, magg, hendrich}@informatik.uni-hamburg.de

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

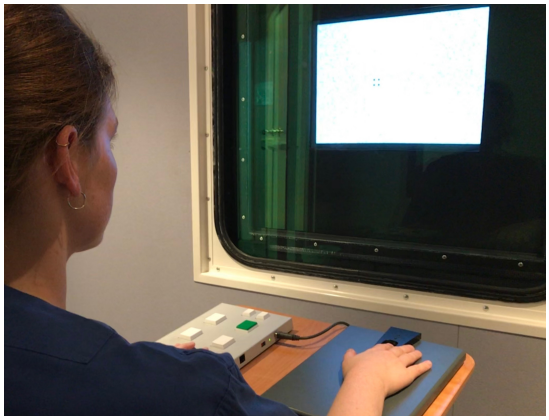

Fig. 1. The original human visuo-tactile discrimination experiment. The participant looks for a (four-dot) visual pattern on the monitor and feels a haptic pattern generated by a Braille actuator on the index finger of the right hand. The participant then decides whether both patterns are the same or different, and presses the green or red button accordingly. The task is made more difficult by blending the visual pattern in the background noise and by reducing the actuated pin height of the tactile display.

II. Visuo-tactile Discrimination in Humans

In our human-participant experiment, 20 young (aged 20-28) and 20 healthy elderly participants (aged 65-79) performed a visuo-tactile pattern discrimination task (adapted from [4]). In this task, participants had to compare Braille patterns presented tactilely to the right index fingertip with visual patterns presented on a computer screen (Fig. 1). A set of four clearly distinct patterns was used in the study because the untrained elderly participants had problems to distinguish more complex Braille patterns (Fig. 2).

During the experiment, stimulus intensity was adjusted individually based on an adaptive-staircase procedure with a target detection accuracy of approximately 80%, to ensure comparable detection performance across modalities and between elderly and young participants (Table I). Tactile stimulus intensity was adjusted by changing the height of

TABLE I

	Tactile threshold	Visual threshold	Performance in Matching
Elderly	1143.0	53.65	66,2%
Young	576.8	49	78,31%

Tactile threshold (Braille pin height in μm) and visual threshold (grey level) for 80% detection accuracy, and performance on the matching/discremination task, for young and elderly human participants.

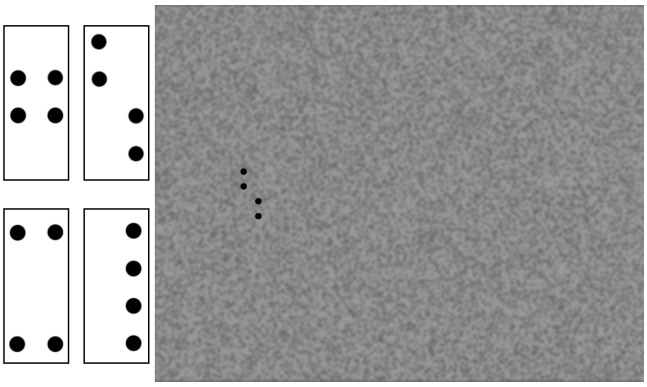

Fig. 2. (Left) The four visual braille patterns that were used for all experiments. (Right) One example input for pattern 3 with 100% intensity (i.e. full black).

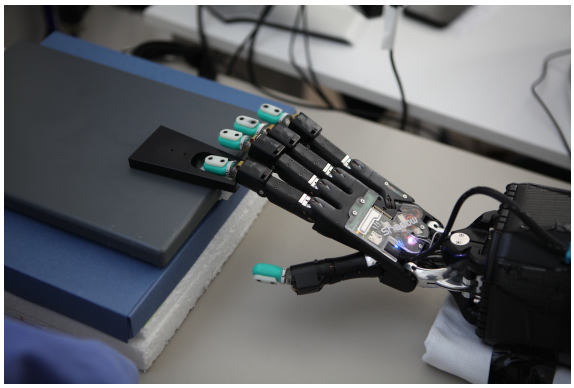

Fig. 3. Robot experiment setup using a Shadow C6 Dexterous Hand. The BioTac tactile fingertip of the first finger of the hand is placed on the Braille actuator.

the braille pattern, visual stimulus intensity by changing the patterns' contrast against a noisy background. Finally, participants performed the visuo-tactile discrimination task at the afore defined unimodal thresholds of 80% accuracy.

III. Robotic Adaptation

The setup described above was realized in a robotic experiment (Fig. 3). The Braille stimuli were applied to the fingertips of a Shadow C6 Dexterous Hand [5] equipped with BioTac tactile sensors [6], [7]. The sensor surface of the BioTac closely matches the size and shape of a human finger and it was possible to align and center the sensor onto the Braille actuator without modifying the setup.

To classify the haptic stimuli of the Braille actuator correctly, the sensor can detect multiple contacts through indirect measurement. The turquoise rubber shell is filled with a conductive liquid and held in place around an inner rigid "bone". When contacting an object, the rubber deforms, changing the overall pressure of the liquid (1 channel) and also the impedance between a set of electrodes patterned on the bone (19 channels). At the same time, liquid temperature changes due to the contact (2 channels). Raw data from the sensor combines the measured pressure, temperature and impedances, but is notoriously difficult to interpret [8], [9]. Because the temperature conditions during recording remained stable, we omitted the respective sensor readings and feed the 20 other channels into an ANN to learn the mapping from raw data to applied Braille stimuli.

As visual stimuli, we use the same visual stimuli employed in the human experiment. These stimuli are directly fed into the neural architecture without an intermediate sensor like a camera. As detailed below, the comparison with the human experiments relies on the exact gray values used in the stimuli; direct input of the images to the network avoids any level-shifts due to inconsistent camera exposure control.

As the detection and classification of the tactile and visual stimuli require offline learning, the adaptive staircase procedure could not be used. Instead, we recorded enough patterns of different complexity so that the required stimuli (corresponding to about 80% single-channel accuracy) could be presented to the trained ANNs after learning. In total, we recorded several hours of raw sensor data from the robot, labeled with the presented tactile and visual patterns.

IV. Computational Models

To evaluate the influence of the actual crossmodal integration of high-level unimodal features in contrast to just comparing unimodal classification we propose two neural architectures: The *V-architecture*, see Fig. 5, statically compares unimodal classification results. It consists of two separate networks that perform unimodal classification of the tactile and haptic input pattern, respectively. Eventually, both classification results are compared in the final layer.

The *Y-architecture*, see Fig. 6, instead integrates high-level feature representations of both modalities. It also has two separate columns for unimodal feature extraction on the visual and tactile data. However, instead of performing a unimodal pattern classification, the extracted features are concatenated and further integrated by a series of dense layers, the stem of the *Y-architecture*. This network performs a *late integration* of crossmodal information. (Early integration models and unified processing of both modalities will be adressed in future work.)

Empirical and automated optimization resulted in the

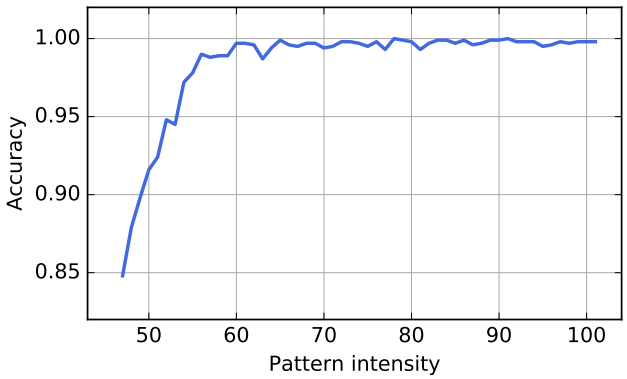

Fig. 4. Mean classification accuracy for visual channel over 1000 images with varying pattern intensity 47% - 100%

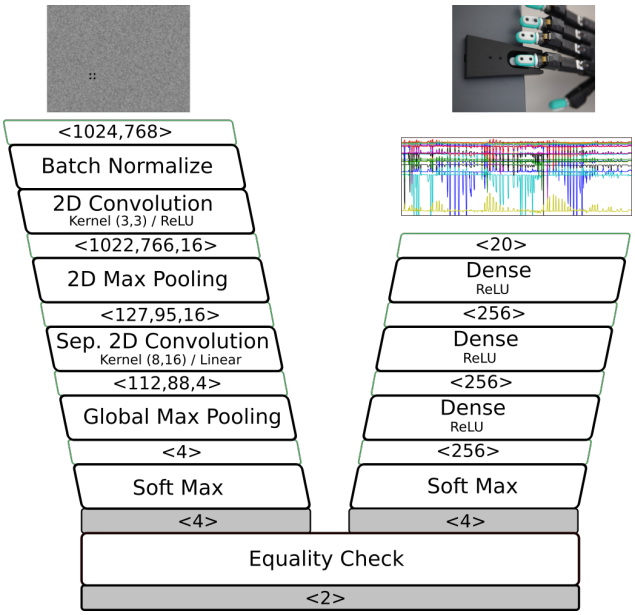

Fig. 5. Structure of the unimodal *V-architecture*. Visual stimuli (left column) and tactile data (right column) are processed separately and statically compared in the end.

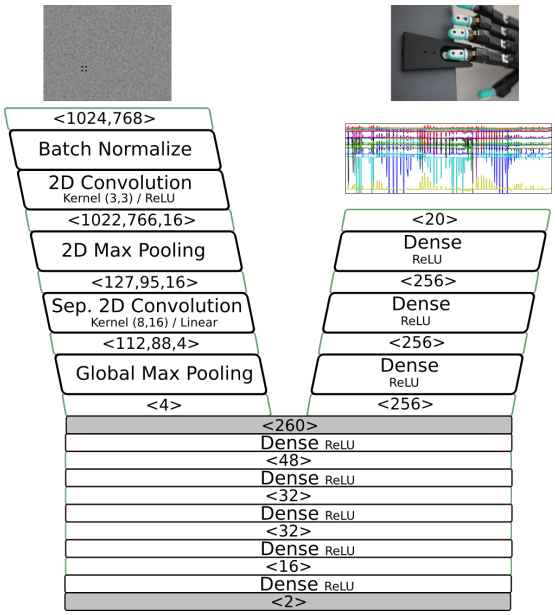

Fig. 6. Structure of the crossmodal integration network (*Y-architecture*). Both columns are first trained separately on visual and tactile data. Afterwards, a number of densely connected layers and a softmax output layer are added, and the network is trained again on the combined visual and tactile data.

following hyperparameters: For the visual columns, two convolution layers L_1 and L_2 after a batch-normalization step, are followed by a pooling layer each (max-pool after L_1, global max-pool after L_2). Kernel regularization was used for both CNN layers. For L_1, the squared sum of the weights was used to force the network to focus on differences of grayscale rather than absolute values. For L_2, L1 regularization with a factor of 0.01 was used to enforce a high level of sparsity. For the *V-architecture*, a final dense layer with soft-max activation with four patterns performs classification. In the *Y-architecture* the extracted high-level features are directly propagated. For the haptic modality, we use an MLP with three hidden layers (20 inputs from the BioTac sensor, three layers with 256 neurons each, followed by one softmax output layer with 4 neurons, corresponding to the 4 Braille patterns). Again, the last layer follows for the *V-architecture* only. Finally, the crossmodal integration in the *Y-architecture* is performed by a series of dense layers with a decreasing number of hidden units (48, 32, 32, 16) followed by a binary softmax layer for same or different patterns.

V. Unimodal and Crossmodal Training

Assuming random selection of the visual and tactile stimuli for the discrimination task (one out of four possible patterns each), always predicting a mismatch will already achieve an accuracy of 75%. This is the baseline that more complex models must surpass.

The training for both networks follows the same pattern: First, each unimodal column of the network is trained. In the case of the non-integrating *V-architecture* a static comparator follows. For the crossmodally integrating *Y-architecture*, a third training phase follows where the complete *Y-architecture* is trained. Both architectures are trained for a total of 100 epochs; in the case of the crossmodal integrating *Y-architecture*, 50 of these epochs were used for unimodal pretraining and 50 epochs for training the whole network. For all training phases, the Adam optimizer with a learning rate of 0.001 and a batch size of 32 was used.

The noisy visual input images are generated by placing one of four target Braille patterns (43x104px, see Fig. 2 left) randomly on one of 48 randomly generated background images (1024x768px) to generate 5000 images each for training and validation (Fig. 2 right). The background consists of a Perlis noise pattern with a gray range of between 40% and 60% (mean 53.7%). The stimulus intensity (i.e., gray level) of the pattern was selected to be between 47% and 100% (black).

The Braille patterns become increasingly difficult to see for humans as the gray levels of the patterns blend with the gray levels of the background. On the robot, it might be possible to achieve even higher classification accuracy using classical computer vision algorithms and prior knowledge about how the data was generated. This would, however, undermine the goal to create controllable unimodal classification performances. Similar to the visual modality, a sufficient number of haptic samples with different pin heights were collected. Depending on the pin height and the unimodal network, different classification results can be achieved.

VI. Results

To test the individual classification accuracy of both channels, and to compare them to the performance of the human participants in the original experiment, both models were fed inputs of varying difficulty (pattern intensity for the visual channel, pin height for haptic channel). The

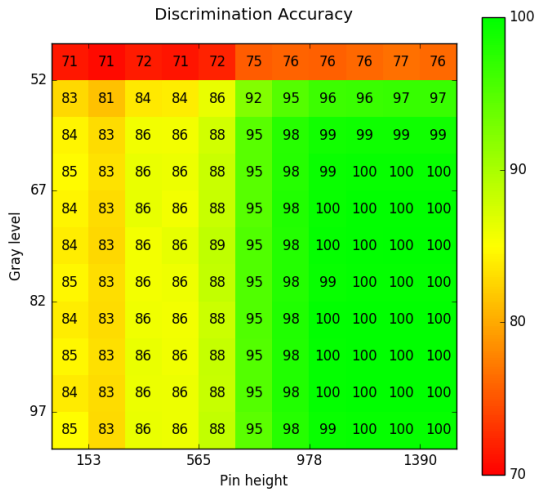

Fig. 7. Performance of the unimodal processing of the *V-architecture* on the discrimination task. Parameters are the grey level (% of black) of the visual pattern and the active pin height (μm) of the Braille actuator. Two separate networks are trained for the visual and tactile data, followed by a static comparison (compare Fig 5).

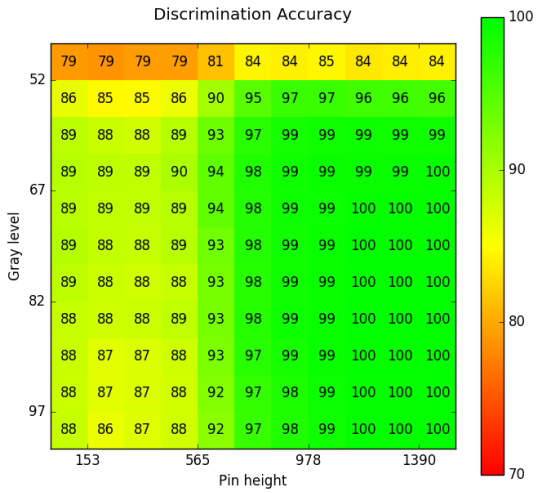

Fig. 8. Performance of the late-integration crossmodal network on the discrimination task. Parameters are the grey level (% of black) of the visual pattern and the active pin height (μm) of the Braille actuator. See the text for details (compare Fig 6).

results for the visual channel can be seen in Fig. 4. The classification accuracy was on average 98.43% and started dropping once the gray value of the pattern also appeared in the background image (values between 40% and 60%). The results of the unimodal processing *V-architecture* on the discrimination task are shown in Fig. 7. As expected, the performance of the network degrades when the channels are too noisy, but accuracy improves quickly as the signal quality (gray level, Braille actuator pin height) becomes better. The corresponding results for the *Y-architecture* that performs late crossmodal integration are shown in Fig. 8. As expected, this more complex network outperforms the unimodal network over the whole parameter range. More interestingly, it already surpasses the trivial (75% level) classifier when both the visual and tactile stimuli are very noisy.

Our results are also interesting in the context of the human experiment: Elderly showed significant higher thresholds for unimodal pattern detection compared to young (see Table 1). While young showed a stable performance in the crossmodal task at the unimodal thresholds, elderly showed a significant weaker performance, suggesting an impaired crossmodal integration in the aged brain which corresponds to independent unimodal classification (*V-architecture*).

VII. Discussion

In the unimodal visual condition, the performance of the *ANN* and young participants was comparable, whereas, in the unimodal tactile condition, young outperformed the *ANN*. In both conditions, the ANN performed distinctively better than elderly participants. Crossmodal performance, however, depends not only on unimodal pattern recognition but also on integration mechanisms. Interestingly, only young participants showed a stable performance in the visuo-tactile matching task at the unimodal thresholds. This result suggests a superior mechanism for crossmodal stimulus processing in the human brain. This finding is mirrored in our result, that the crossmodal integrating neural architecture outperforms the non-integration unimodal neural architecture even though the unimodal processing columns are identical.

It has been suggested that efficient stimulus processing in the human brain depends on recurrent neural networks and early sensory integration [3]. Adapting such approaches in future work might, on the one hand, improve the performance of artificial devices, but on the other hand, also give insights into the question which disturbances of the system lead to suboptimal functioning in the aged brain. Further research is needed, to answer the question of how young brains successfully integrate crossmodal information and which of these mechanisms can be adapted in artificial systems.

References

[1] Calvert, G. A. (2001). *Crossmodal processing in the human brain: insights from functional neuroimaging studies*, Cereb. Cortex, 11, 1110–1123.

[2] Freiherr, J., Lundstrom, J. N., Habel, U., & Reetz, K. (2013). *Multisensory integration mechanisms during aging*. Front Hum Neurosci, 7, 863.

[3] Ghazanfar, A.A. and Schroeder, C.E. (2006). *Is neocortex essentially multisensory?* Trends in Cognitive Sciences, 10(6), 278-285.

[4] Göschl, F., Engel, A. K., & Friese, U. (2014). Attention modulates visual-tactile interaction in spatial pattern matching. PLoS One, 9(9), e106896. doi: 10.1371/journal.pone.0106896

[5] Shadow Robot Dextrous Hand, online: `www.shadowrobot.com`

[6] Syntouch LLC, *BioTac Product Manual* (V20), SynTouch LLC, California, Mar 2015. `www.syntouchinc.com/wp-content/uploads/2017/01/BioTac\_Product\_Manual.pdf`

[7] N. Wettels, *Biomimetric Tactile Sensor for object identification and grasp control*, Ph.D. dissertation, Univ. of Southern California, 2011.

[8] C. H. Lin, T. Erickson, J. Fishel, N. Wettels, and G. Loeb, *Signal processing and fabrication of a biomimetic tactile sensor array with thermal, force and microvibration modalities*, in IEEE Int. Conference on Robotics and Biomimetics (ROBIO), 2009, pp. 129–134.

[9] C.H. Lin, J. Fishel, G Loeb. *Estimating Point of Contact, Force and Torque in a Biomimetic Tactile Sensor with Deformable Skin* SynTouch LLc, 2013.