Interpreting uncertainty in future human behaviour with sequential Neural Processes

Responsible Supervisor : Hayley Hung (h.hung@tudelft.nl)
Co-supervisor : Chirag Raman (c.a.raman@tudelft.nl)

Neural Processes (NPs) [1] are a family of latent variable models that aim at combining the best of neural networks (NNs) and Gaussian processes (GPs). Like GPs, NPs define distributions over functions, are capable of rapid adaptation to new observations, and can estimate the uncertainty in their predictions. Like NNs, NPs are computationally efficient during training and evaluation but also learn to adapt their priors to data. Recent extensions
like Sequential Neural Processes [2] aim at modeling stochasticity in temporal transitions.

They are therefore an interesting family of models for the task of predicting future social conversational behaviour. These interactions are dynamic in nature where the behaviour of each person influences that of other interaction partners. Moments of low uncertainty in future behaviour in this context can indicate important events; for instance, a social robot could use anticipation of these events to form an engagement policy.

This project specifically deals with exploring the space of Sequential Neural Models to answer the question: how can we perform causal inference using modeled uncertainty to identify observed behavioural patterns that cause future events of high certainty?

[1] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S.M. Ali Eslami, and Yee Whye Teh. “Neural Processes”. arXiv prepint arXiv:1807.01622

[2] Gautam Singh, Jaesik Yoon, Youngsung Son, and Sungjin Ahn. “Sequential Neural Processes”. arXiv preprint arXiv:1906.10264

Body Orientation Estimation using wearable sensors and cameras

Responsible supervisor : Hayley Hung (h.hung@tudelft.nl)

Co-supervisor : Stephanie Tan (S.Tan-1@tudelft.nl)

Accurate estimation of orientations of human body parts orientation is of interest to the computer vision, wearable sensing, and robotics community. A popular method is based on inertial and magnetic sensing, which is achieved by Inertial Measurement

Units (IMUs). IMUs contain a tri-axial accelerometer, a tri-axial gyroscope, and a tri-axial magnetometer. The aim of this project is to transfer knowledge from the signal processing community [1] to the computer vision domain and the application domain of human social signals. We hope to use the IMU signals as well as the rotation vectors to understand how humans orient themselves in a crowded social setting. One possibility is to develop a new sensor fusion algorithm by jointly using IMU signals with RGB video data to more robustly estimate orientations. The IMU signals and recovered orientations may be important for downstream tasks such as conversation group membership estimation and visual focus of attention estimation [2]. A real-world dataset capturing interactions among conference attendees, similar to [3], will be used for these tasks.

[1] Kok, Manon, Jeroen D. Hol, and Thomas B. Schön. "Using inertial sensors for position and orientation estimation." arXiv preprint arXiv:1704.06053 (2017).

[2] Ricci, Elisa, et al. "Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[3] Alameda-Pineda, Xavier, et al. "Salsa: A novel dataset for multimodal group behavior analysis." IEEE transactions on pattern analysis and machine intelligence 38.8 (2015): 1707-1720.

Investigating domain adaptation strategies for human social action recognition

Responsible Supervisor: Hayley Hung (h.hung@tudelft.nl)

When considering the analysis of social action such as speaking, laughter, or gesturing, these behaviours can be carried out in both standing conversational scenarios as well as seated ones. Prior work has shown that such behaviours can be highly personal [1,2] in a free standing conversational setting. The aim of this project is to develop techniques to adapt to a different domain (seated conversations) when data from the same person in both contexts is available when no labels in the new domain are available. One possibility to evaluate the data in such a setting is to exploit second or third order tasks to check the effectiveness of the domain shift. In this case, the MatchNMingle Dataset [3] will be used where speed dates are available with labels of attraction are available [4] or labels for when social interactions occur.

[1] E. Gedik and H. Hung, "Personalised models for speech detection from body movements using transductive parameter transfer", Personal and Ubiquitous Computing, 2017

[2] J. Vargas Quiros and H. Hung, “CNNs and Fisher Vectors for No-Audio Multimodal Speech Detection”, MediaEval 2019

[3] L. Cabrera-Quiros*, A. Demetriou* , E. Gedik, L. v. d. Meij and H. Hung. "The MatchNMingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates", in IEEE Transactions on Affective Computing, 2018.

[4] Oyku Kapcak, Vargas, J., & Hung, H. (2019). Estimating Romantic , Social , and Sexual Attraction by Quantifying Bodily Coordination using Wearable Sensors. *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*.

F-formation evolution in large social gatherings

Responsible Supervisor: Hayley Hung (h.hung@tudelft.nl)

Co-supervisor : Stephanie Tan (S.Tan-1@tudelft.nl)

During large scale human social events such as networking events or parties, conversing groups, known as F-formations are known to form, grow, split and merge in a bottom up self-organised system. Not much us known about how or why such groups form. Using the publicly available Idiap poster dataset which consists of over 50 people interacting during a poster and coffee break for over 2 hours, this project will study these links. We know already from prior literature that F-formation detection images can be formulated as a maximal clique in an edge-weighted graph [1].

The aim of this project is to study the extent to which extensions of graph clustering approaches can be used to model the evolution of conversation partners over time. A promising approach to doing this is by exploiting notions of clique overlap and hierarchical clustering. In addition, methods for the construction of the affinity matrix (measuring the closeness between people in the scene) can also be considered based on the video data or other sensor data that captures the coordination dynamics between conversing partners. Promising directions include considering that some conversation partners might potentially belong to multiple conversing groups [3]. An alternative hypothesis is that multiple simultaneous conversations can exist in a single F-formation [4] and that modelling the possibility that members of an F-formation might belong to multiple conversations within the same F-formation.

[1] H. Hung and B. Krose, "Detecting F-formations as Dominant Sets", in International Conference on Multimodal Interaction, Alicante, Spain, November 2011.

[2] Pavan and Pelillo, "Dominant sets and hierarchical clustering," Proceedings Ninth IEEE International Conference on Computer Vision, Nice, France, 2003, pp. 362-369 vol.1.

[3]Torsello, A., Bulo, S. R., & Pelillo, M. (2008, December). Beyond partitions: Allowing overlapping groups in pairwise clustering. In 2008 19th International Conference on Pattern Recognition (pp. 1-4). IEEE.

[4] Raman, C., & Hung, H. (2019). Towards automatic estimation of conversation floors within F-formations. arXiv preprint arXiv:1907.10384.

Multimodal No-audio Speaker Detection by exploring the relationship between speech and gesture

Responsible Supervisor: Hayley Hung (h.hung@tudelft.nl)

Co-supervisor : Jose Vargas Quiros (j.d.vargasquiros@tudelft.nl)

In this project, the task of estimating the speaking status of people in a crowded networking event will be investigated. Due to the privacy sensitive aspect of recording audio, the aim of this project is to use other modalities such as video and wearable acceleration to predict when people are speaking or not. Whilst speaker detection is traditionally performed as an audio based task, we know from social science that people's vocal behaviour during speaking is accompanied by gestures [1]. We exploit these gestures in order to learn the relationship between speech and body movements.

In prior works [2, 3, 4], there has been no audio data available. In this project, an investigation will be made into the effect that more accurate ground truth, as extracted from audio, can impact the learning of a more accurate speaking detection model. This will involve the analysis of what aspects of speech and possibly dialogue acts are better predicted and how to improve the overall performance of no-audio speaking detection systems. The rhythm of gestures during conversation is known to correlate with speech prosody and so existing gesture detection methods could be exploited [5]. The project will leverage other collected data available in the SPC Lab for carrying out this investigation. This project can also involve the investigation of the social roles of people in a conversation.

This task has potential applications in smart office environments where the analysis of social behaviour is vital for monitoring team function and performance.

[1] McNeill, D.: Language and gesture, vol. 2. Cambridge University Press (2000)

[2] L. Cabrera-Quiros*, A. Demetriou* , E. Gedik, L. v. d. Meij and H. Hung. "The MatchNMingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates", in IEEE Transactions on Affective Computing, 2018.

[3] A Rosatelli, E Gedik, H Hung, “Detecting F-formations & Roles in Crowded Social Scenes with Wearables: Combining Proxemics & Dynamics using LSTMs”, arXiv preprint arXiv:1911.07279, 2019

[4] J. Vargas Quiros and H. Hung, “CNNs and Fisher Vectors for No-Audio Multimodal Speech Detection”, MediaEval, 2019

[5] L. Cabrera-Quiros, D. M. J. Tax and H. Hung, "Gestures in-the-wild: detecting conversational hand gestures in crowded scenes using a multimodal fusion of bags of video trajectories and body worn acceleration," in IEEE Transactions on Multimedia, 2020.

Deep methods for the analysis of behavioral entrainment from body movement

Responsible Supervisor: Hayley Hung (h.hung@tudelft.nl)

Co-supervisor : Jose Vargas Quiros (j.d.vargasquiros@tudelft.nl)

In dyadic and group interaction, our behavior is necessarily entrained with that of the other participants in the conversation. This phenomenon occurs over different modalities and extends to body movement [1]. We tend to imitate, synchronize and adapt to our conversational partners. Different correlation and mutual information based measures have been proposed to measure similarity, synchrony and convergence in body movement behavior. Previous work [2] showed that such simple measures, obtained from tri-axial chest-worn accelerometers can be used to predict the outcome of speed dates. This work also suggested the importance of methods that are robust against small differences in the timing of imitated behavior, a property of convolutional networks with max-pooling. Such networks have also been used successfully for the detection of speaking status from similar accelerometer data [3]. In this work, we will explore the use of deep convolutional methods for obtaining measures of movement similarity that can be used to reach conclusions about leading and following behavior in social interactions and to predict related labels like attraction. The project would also seek to provide insights about the mechanisms of mimicry and adaptation of body movement in groups of three or more conversing people.

[1] Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., & Cohen, D. (2012). Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, *3*(3), 349–365. https://doi.org/10.1109/T-AFFC.2012.12

[2] Oyku Kapcak, Vargas, J., & Hung, H. (2019). Estimating Romantic , Social , and Sexual Attraction by Quantifying Bodily Coordination using Wearable Sensors. *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*.

[3] J. Vargas Quiros and H. Hung, “CNNs and Fisher Vectors for No-Audio Multimodal Speech Detection”, MediaEval 2019

Romantic attraction estimation from speed dates by exploiting multimodal behaviour data

Responsible Supervisor: Hayley Hung (h.hung@tudelft.nl)

This project builds on previous investigations about exploiting synchrony and convergence patterns between conversing partners using a body worn accelerometer [1, 2] to estimate various forms of attraction as well as future behaviour. Past works have been exclusively single modality and have shown success. The aim of this project is to exploit investigate multimodal measures of synchrony and convergence to see if the estimation of attraction can be improved in cases of social, romantic and sexual attraction. The aim of the project is to gain insights into what coordinated behaviours are most closely linked to these different forms of attraction. This could include looking at the relationship between mimicry of atomic behaviours such as co-laughter, gesturing, or head orientation shifts.

[1] A. Veenstra and H. Hung, "Do They Like Me? Using Video Cues to Predict Desires during Speed-dates", IEEE International Conference on Computer Vision Workshops,p. 838-845, Barcelona, Spain, 2011.

[2] Oyku Kapcak, Vargas, J., & Hung, H. (2019). Estimating Romantic , Social , and Sexual Attraction by Quantifying Bodily Coordination using Wearable Sensors. 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW).

Multimodal Estimation of Cohesion in Teams

Responsible Supervisor: Hayley Hung (h.hung@tudelft.nl)

Team function is a vital in organisational settings. Yet our understanding of why they function well is still limited especially because the dynamics of team processes are barely studied. Team cohesion, which is described as a force that binds people together with an intrinsic motivation to continue working together [1, 2]. In keeping with the current consensus in the teams literature, this project studies task cohesion, the extent to which a team is united in striving for goal accomplishment, and social cohesion, an interpersonal bond among team members.

There have been a few prior attempts to analyse task and social cohesion [3, 4]. The aim of this project is to investigate multimodal approaches of synchrony and convergence which have been shown to be important indicators of social cohesion. On the other hand, task cohesion is generally less easy to automatically detect. This could be in part related to past model assumptions about task cohesion being similar in nature to social cohesion. This project could therefore investigate how to better estimate task cohesion by considering behaviours that are specific to task directed behaviour.

Finally, outside of the team meeting environment there is also the possibility to investigate cohesion in teams in a more pervasive long term setting [5].

[1] Casey-Campbell, M., & Martens, M. L. (2009). Sticking it all together: A critical assessment of the group cohesion - performance literature. International Journal of Management Reviews, 11, 223-246.

[2] Salas, E., Grossman, R., Hughes, A. M., & Coultas, C. W. (2015). Measuring team cohesion: Observations from the science. Human Factors, 57(3), 365-374.

[3] Nanninga, M. C., Zhang, Y., Lehmann-Willenbrock, N., Szlávik, Z., & Hung, H. (2017, November). Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry. In Proceedings of the 19th ACM International Conference on Multimodal Interaction (pp. 206-215). ACM

[4] Hung, H., & Gatica-Perez, D. (2010). Estimating cohesion in small groups using audio-visual nonverbal behavior. IEEE Transactions on Multimedia, 12(6), 563-575.

[5] Zhang, Y., Olenick, J., Chang, C. H., Kozlowski, S. W., & Hung, H. (2018). TeamSense: assessing personal affect and group cohesion in small teams through dyadic interaction and behavior analysis with wearable sensors. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2(3), 150

Group based speaking status detection from wearable acceleration

Responsible Supervisor: Hayley Hung (h.hung@tudelft.nl)

Co-supervisor : Jose Vargas Quiros (j.d.vargasquiros@tudelft.nl)

Speech and gesture have been shown to be closely intertwined [1]. Previous work has shown that body-worn accelerometers can be used for the detection of speaking status [2]. In this project we seek to enhance speaking status detection by jointly predicting speaking status for groups, instead of individually. This involves the exploration of pooling strategies for sequence and convolutional models for this task [3]. The challenge is to have a model capable of learning the dynamics of conversation in groups, where the occurrence of multiple simultaneous speakers is commonplace. The project could also provide insights into how these occurrences could signal salient moments in the conversation.

[1] Pouw, W. (2019). Quantifying Gesture-Speech Synchrony, (May). https://doi.org/10.17619/UNIPB/1-815

[2] Gedik, E., & Hung, H. (2017). Personalised models for speech detection from body movements using transductive parameter transfer. *Personal and Ubiquitous Computing*, *21*(4), 723–737. https://doi.org/10.1007/s00779-017-1006-4

[3] Goel, K., Fei-Fei, L., Savarese, S., Alahi, A., Robicquet, A., & Ramanathan, V. (2016). Social LSTM: Human Trajectory Prediction in Crowded Spaces, 961–971. https://doi.org/10.1109/cvpr.2016.110