Robot Perception of Human Groups in the Real World: State of the Art

Angelique Taylor and Laurel D. Riek
Department of Computer Science and Engineering,
University of California San Diego
email: amt062@eng.ucsd.edu, lriek@ucsd.edu

Abstract

As robots enter human spaces and begin to work proximately with people, it is important that they understand human social interaction. They must be able to perceive human social signals and understand how to adapt to groups. The goal of our work is to design robot perception algorithms that allow robots to understand human group dynamics via social cues, and understand how to behave collaboratively in groups. In this paper, we discuss the current state-of-the-art of two fields that have contributed methods to achieve this goal, social signal processing and computer vision. We describe recent advances in these fields, as well as some of the challenges faced when adapting them to mobile robots.

1 Introduction

Robots are transitioning into unstructured environments where they will work proximately with people (Riek 2013). As this transition happens, humans will have expectations of how these robots will behave, appear, and interact. Social, cultural, situational norms, and context play a significant role in both how people formulate these expectations, as well as how they might behave around robots. Thus, it is important robots are able to sense and understand the contextual world around them and human social signals in order to respond appropriately to it (Nigam and Riek 2015; O'Connor and Riek 2015).

Researchers in human-robot interaction (HRI) have explored how this contextual awareness might be accomplished by enabling robots to recognize and respond to (synthesize) human social signals. For example, some have designed models for robots to appropriately approach a human to initiate a conversation (Satake et al. 2013), or build proxemic-sensitive gesture and speech patterns (Mead and Matarić 2015). Others have explored detecting and synthesizing head motion, gaze patterns, and synchronous mechanisms as a means for building rapport, enhancing likability, or sustaining engagement during interaction (Riek, Paul, and Robinson 2010; Rich et al. 2010; Khoramshahi et al. 2016).

These challenges have inspired researchers to transition from dyadic interaction to group interaction between humans and robots. Group interaction has been studied for over a decade in fields such as social psychology, linguistics, sociology, computer vision, and robotics (Kendon 1990; Ricci et al. 2015). It provides rich information about interdependence between group members, group cohesion, and how people communicate both verbally and nonverbally. More specifically to robotics, group dynamics can provide information about how social signals resonate throughout groups, how a robot can cooperate with a group, and also may allow a robot to sense the affective states of groups. Therefore, robots need sensing abilities that allow them to perceive group social dynamics to facilitate effective face-to-face communication.

There are several challenges in detecting and interacting with groups from a mobile robot. First, there is the problem of sensing. Where are the sensors located? (On the robot? On a person? In the room?) Can a robot see all members of the group, even as they (and it) are in motion? Then, even if these sensing challenges are surmounted, a robot still needs to process these social signals correctly, which is non-trivial (Riek 2013). Finally, robots need to be able to process all this information in near real-time, often with limited computational resources.

There has been some recent work done in this domain within the robotics community. For example, Iqbal et al. (2016) designed an algorithm which can anticipate high-level group behavior, calculate the dynamics of the group, and adapt a robot's behavior to humans in real time. They experimentally validated this method across a range of multi-party interaction scenarios, and found it to be successful. Recently, they have adapted the model to leverage tempo as a mechanism for robots to adapt to humans Iqbal et al. (2016).

One gap in this prior work is that it was conducted using external sensors (four Microsoft Kinect sensors), and within a experimental interaction paradigm. We are interested in integrating robots into uncontrolled, naturalistic, unpredictable environments where the robot freely interacts with groups of people. To accomplish this, we plan to leverage multidisciplinary ideas from the social signal processing and computer vision fields to sense and respond to groups.

In this paper, we discuss the current state-of-the-art in social signal processing and computer vision fields for detecting groups, and explore how mobile robots might leverage these mechanisms.

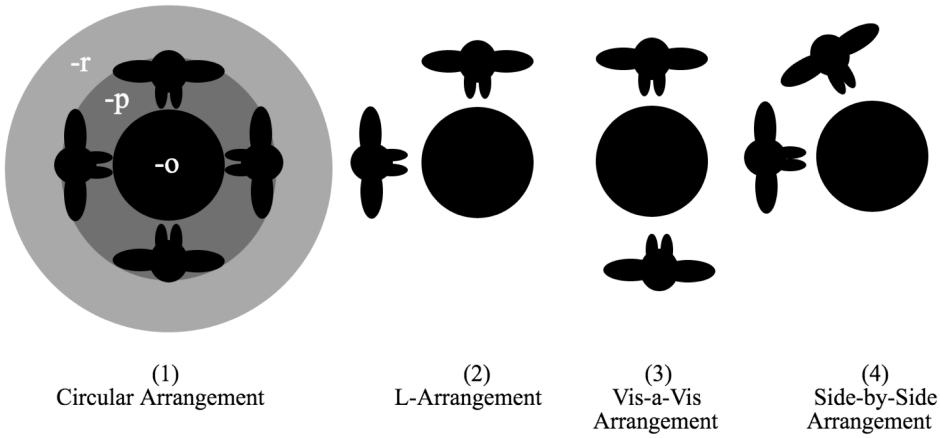

Figure 1: Human Spatial F-Formation Arrangements: (1) Circular Arrangement: Defines three social spaces. o-space is the convex hull in the middle of the group members, p-space is the area surrounding the o-space (where the humans are positioned), and r-space is the area beyond p-space; (2) L-Arrangement is formed between two humans standing perpendicular to one another in an L-shaped position in p-space; (3) Vis-a-Vis Arrangement is formed when two group members stand directly facing one another; and (4) Side-by-Side Arrangement is formed when two group members are facing the same direction (not toward one another) (Kendon 1976).

2 Background and Scope

Social signal processing (SSP) "aims at providing computers with the ability to sense and understand human social signals" Vinciarelli et al. (2009). These social signals include nonverbal behavioral cues such as facial expression, body posture, gesture, and proxemics. Social signals are important for robots to understand, because leveraging the affective states of a group can inform the robot's behavior. Computer vision techniques are also important for robots as they address perception techniques for human motion.

Some computer vision approaches can help address robot perception problems, such as pedestrian tracking and F-Formation detection. The problem of pedestrian tracking is to provide semantic information about pedestrians in a scene. F-Formation is a formal system that identifies groups in social environments (Kendon 1976). An F-Formation system arises when two or more humans sustain a spatial and orientation relationship, which divides the spatial relationship between groups into a p-space and o-space. As shown in Figure 1, o-space is the area in the center of the group, while p-space is the area where humans stand in a group surrounding o-space, and r-space is the area beyond p-space.

Several researchers in the computer vision community have been leading a transition from studying individuals to studying groups, and several have been leveraging the F-Formation system. This allows them to explore characteristics of groups such as dimension (small groups or crowds), durability (ephemeral, ad-hoc, or stable), and organization (Setti et al. 2015).

Researchers in robotics can also leverage these characteristics to learn about social interaction. The F-Formation concept can be used to systematically analyze social group behavior from a mobile robot. Robots can leverage dimension, durability, and organization to appropriately adapt its behavior to human group motion. In addition, this concept can help robots perceive where people are relative to itself, how they are moving, and how the robot can use this information to join groups.

2.1 Egocentric vs. Exocentric Perception

In order for robots to be able to detect groups and interact fluently during face-to-face interaction with group members, it is important that algorithms are designed using egocentric (or robot-centric) data; otherwise, these same algorithms may fail.

Egocentric vision has been used for many vision problems including: activity recognition, navigation, video summarization, action-object detection, and 3D saliency detection (Betancourt et al. 2015; Soo Park and Shi 2015). For the purposes of designing algorithms for robots in social spaces, egocentric vision is also an important problem as exocentric perspectives would be approached differently than egocentric perspectives. For example, egocentric vision incorporates features of the robot that are in the field-of-view of the camera; otherwise, the robot itself is not seen in the data. Also, egocentric vision is typically at human height; whereas, exocentric vision can have any spatial orientation and include the robot in the data as well.

2.2 Sensing Technologies

The sensing technologies used to perceive people in the robotics, computer vision, and social signal processing fields vary depending on the problem, but typically are differentiated as intrusive and non-intrusive sensors. Intrusive sensors record physiological signals or positional information, such as heart rate, galvanic skin response, or location. Some examples of intrusive sensors used in the literature include: accelerometers used to measure acceleration or speed (e.g. gesture), wearable sensors that contain accelerometers used

to identify emotions, cellular phones used to measure location or proxemics, and inertial measurement units (IMU) used to measure body specific force, angular rate, and sometimes magnetic fields (e.g. detect user activity) (Palaghias et al. 2016).

Non-intrusive sensors are typically sensors placed in the environment, usually far away from people, which collect data. Examples include: RGB-D cameras (e.g. the Microsoft Kinect can track skeleton motion), Panoramic cameras, sonar sensors, or thermal imaging sensors.

Because we are interested in methods for mobile robots, we limit our discussion to methods that use data from mobile sensors as this is most suitable for mobile robotic applications.

3 Methodologies

The problem of better understanding how to integrate robots into human groups have been approached by leveraging work from the social signal processing and computer vision fields. We discuss state-of-the-art methods used in each of these respective fields to promote the analysis of groups from a robot-centric (egocentric) perception perspective.

3.1 Social Signal Processing

Social signals are the expression of one's attitude toward a social situation. This emcompasses nonverbal behavioral cues such as body posture, gesture, facial expression, conversational analysis (e.g., turn taking), and proxemics (Moosaei, Hayes, and Riek 2015; Vinciarelli, Pantic, and Bourlard 2009). In this paper, we focus on analysis of gestures, body posture, and proxemics, as these are behavioral cues most readily detectable from a mobile robot, and can be very fruitful in understanding social groups.

The problem of gesture recognition has been approached from an intrusive and non-intrusive sensing perspective. Researchers looked at gestural analysis from a physiological and positional point-of-view using technologies such as accelerometers, inertial sensors, and textile capacitive sensor arrays (Palaghias et al. 2016; Singh et al. 2015). However, intrusive sensors can induce a Hawthorne effect (Wickstrom and Bendix 2000), and these sensors may alter participants' behavior. As a result, a nonintrusive sensing paradigm is often necessary to collect naturalistic gestural data, and many researchers have employed the Microsoft Kinect in their work.

However, the Kinect has raised challenges in gesture recognition. One problem is the noisy skeleton trajectory data, which increases the difficulty of accurately detecting gestures. Another problem is modeling the temporal and spatial dynamics of gestures using RGB, depth, and infrared for analysis. Gestures are analyzed frame-by-frame; therefore, a temporal model is needed to capture the time varying characteristics of gestures. Gestures vary spatially as well so this must also be captured in the model for it to be robust to these spatial-temporal variations (Pitsikalis et al. 2015).

Therefore, multimodal fusion techniques were developed to combine information from several modalities at once to increase accuracy of state-of-the-art models (Pitsikalis et al. 2015). This combined with machine learning techniques allow researchers to extract features and employ effective classifiers.

Typically, preprocessing techniques are employed on the data such as noise removal and signal smoothing (Escalera, Athitsos, and Guyon 2016). Then, features are extracted using spatio-temporal and salient characteristics of the data e.g. Gaussian temporal smoothing, Histogram of Oriented Gradients, and SIFT (Song, Demirdjian, and Davis 2011; Dalal and Triggs 2005; Lowe 1999). The most popular classifiers include Support Vector Machines, Random Forests, Conditional Random Fields, Dynamic Time Warping, Hidden Markov Models, and Deep Learning (Escalera, Athitsos, and Guyon 2016; Wu et al. 2016). Evaluation metrics of these methods typically include confusion matrices, spotting, Jaccard indices, and F1-scores. A comprehensive list of datasets can be found in recent surveys by Escalera et al. (2016) and D'Orazio et al. (2016).

Although gesture recognition research has made great headway, there remains a lack of work done to understand gestures from multiple humans simultaneously and in naturalistic scenes. So far, many gesture recognition systems are limited to lab settings where real world challenges are mitigated. Therefore, there is still a gap in the literature that addresses these issues on a mobile robot, which is necessary to integrate robots into human social environments and adeptly interact with people.

Another important behavioral cue during social interaction is body posture. Body posture has been used to identify affective states of humans. Many researchers studied body posture visually, and provide a stimuli to invoke predefined affective states from participants (Karg et al. 2013; Mota and Picard 2003).

Researchers leveraged intrusive sensors to identify peoples' affective states. Some studies used actors, and some studies used stimuli to induce authentic affective states. Typically, pressure sensors are installed in study participant's chair during data acquisition and the researchers use these pressure readings to identify when a participant leaned forward, leaned backward, sitting upright, slumping back, etc. (Mota and Picard 2003; Karg et al. 2013). Then, they perform feature extraction by modeling the pressure readings using a probability distribution and train a classifier to recognize the affective states of the participant (Karg et al. 2013). Common classifiers include: Naive Bayes, Nearest Neighbor, Support Vector Machine, Recurrent Neural Networks, Hidden Markov Models, and Decision Trees (Escalera, Athitsos, and Guyon 2016).

Others have employed non-intrusive sensors to study body posture, such as the Kinect, and used methods to infer states such as engagement, frustration, and focused attention (Grafsgaard et al. 2012; Liu et al. 2015; Lee et al. 2015). The state-of-the-art in body posture recognition suggests that identifying body postures is non-trivial; however, the problem is not as straightforward when identifying body posture from a group of humans situated in an circular orientation. In order to analyze each group members' posture while they are situated in a circular orientation, multiple sensors would be required to capture skeleton data of the

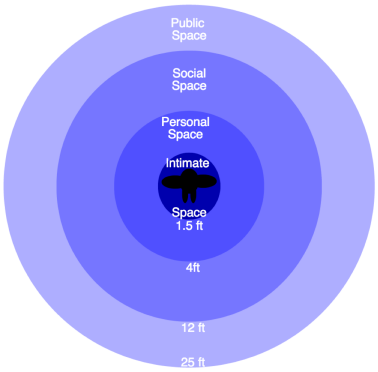

Figure 2: Proxemic Zones: Hall defines intimate space as between 6-18 inches around a person, personal space as 1.5-4 feet, social space as 4-12 feet, and public space as 12-25 feet around a person (Hall 1966)

group (depending on the group size). In this case, participants would be aware that they were under observation. Therefore, an open question is how can group body posture be analyzed from a mobile robot in a naturalistic setting without inducing a Hawthorne effect?

Proxemics is another behavioral cue that defines humans' relative subjective view of intimate, personal, social, and public space. Hall, the founder of the field, defines intimate space as 6-18 inches, personal space as 1.5-4 feet, social space as 4-12 feet and public space as 12-25 feet, as shown in Figure 2 (Hall 1966). Proxemics has been studied in HRI, though to date has mainly focused on understanding proxemic factors that affect human perception of the robot (Mead and Matarić 2015; Oosterhout, Visser, and others 2008). These factors include: speed, appearance, and direction of approach (Rios-Martinez, Spalanzani, and Laugier 2015).

For example, Butler and Agah (2001) found that humans become uncomfortable when a robot approaches them with fast speed, namely 1m/s. Their study suggests that an acceptable speed is 0.254m/s and 0.381m/s. They also studied how anthropomorphism affects proxemics. They found that humans that preferred the humanoid robot's appearance were more acceptable to closer proximity than other humans. Mumm and Mutlu's (2011) key findings suggests that when a robot gazes at a human, this same human is more likely to decrease their proximity with the robot and increase their proximity with the robot in the latter case. In addition, Oosterhout and Visser (2008) studied how proxemic preferences change with age and gender.

Kruse et al. (2012) incorporated proxemics into a robot navigation system to avoid discomfort of humans. Dautenhahn et al. (2006) found that people felt more threatened when a robot approached them directly while carrying an object. Also, Mead and Mataric (2015) studied how well robots could predict social signals such as speech and gesture from different proxemic locations of study participants.

Proxemics is very important for integrating robots into human spaces. The aforementioned key findings suggest that robots must use social signals as an indicator of whether it should approach humans. In addition, the findings suggest that appearance, speed, and direction of approach impacts humans perception of robots so these factors must be taken under consideration for face-to-face interaction between humans and robots.

The work discussed in this section provided an overview of the state-of-the-art of gesture recognition, body posture recognition, and proxemics. These problems encompass the key challenges for detecting groups in social spaces for a mobile robot from an SSP perspective. Next, we will discuss problems in the computer vision field that can help robots perceive human motion.

3.2 Computer Vision

The field of computer vision has studied many problems that are applicable for mobile robots to interact with groups, such as pedestrian and F-Formation detection. Common challenges between pedestrian and F-Formation detection include data occlusion, clutter, illumination changes, low contrast, and pose variation.

Pedestrian tracking is a challenging problem. Many methods have been proposed to address pedestrian detection over the past decade. However, some methods may perform well on one dataset, but perform poorly on others (Dollar et al. 2012; Dalal and Triggs 2005). This challenge has resulted in tracking accuracy being highly dependent on how models are trained. For example, accuracy is dependent on quality of training data and feature selection. Also, some datasets have stationary cameras, and some datasets have mobile cameras, which also impacts transferability of one method to another on different datasets.

Methods employed on data from mobile cameras have an additional challenge such as predicting the camera's motion as well as predicting motion of pedestrians. Pedestrian tracking algorithms have to account for unstable camera motion and high frequency changes of pedestrians moving in and out of the scene. Methods which rely on background subtraction are unsuitable for this task, so different approaches are needed. For example, Choi et al. (2013) designed an algorithm that simultaneously tracks camera motion and pedestrians. Other researchers have used a multimodal approach by leveraging different sensing technologies such as combining vision with either thermal/infrared camera, stereo, or laser sensors to track pedestrians (Walia and Kapoor 2016).

Because the computer vision community has been working on this problem for the past few decades, they have progressed from single-target to multi-target pedestrian tracking, and then leveraged social groupings to improve multi-target tracking accuracy (Choi et al. (2013); Leal et al. (2011)).

Multi-target tracking typically involves an approach similar to SSP via machine learning and can be addressed as an optimization or estimation problem. The first step is preprocessing, which most methods use background subtraction or Histogram of Oriented Gradients (Dalal and Triggs 2005) to isolate the target and pass the data to a tracking model. The tracking problem is formulated as an optimization problem. The optimization problem aims to find the global optimum

of all observations so that the model can generalize these observations well; hence, the model can detect many variations of human sizes, shapes, and color variations. Then, the optimization problem becomes an estimation problem, which can be approached using probabilistic inference or deterministic optimisation. The estimated location is the location of the pedestrian in an image and helps account for non-linear camera motion (Walia and Kapoor 2016).

Using probabilistic inference, a sampling method such as Particle filters or a Reversible Jump Markov Chain Monte Carlo particle filter can be used to generate a distribution to build a model for estimation (Breitenstein et al. 2011; Choi, Pantofaru, and Savarese 2013). Probabilistic inference is typically used to predict and update pedestrian locations.

In contrast, deterministic optimization techniques find the global optimal solution for predicted locations. Some deterministic optimization techniques include min-cost max-flow network flow, bipartite graph matching, max weight independent sets, or dynamic programming (Choi and Savarese 2010; Qin and Shelton 2012; Brendel, Amer, and Todorovic 2011; Andriyenko and Schindler 2010). Then, the models must be initialized, trained, and the parameters must be tuned to achieve the best accuracy (Walia and Kapoor 2016).

More recently, researchers used social factors such as group membership to improve pedestrian detection accuracy (Leal-Taixé, Pons-Moll, and Rosenhahn 2011). The approach to this problem is very similar to probabilistic inference and deterministic optimization; however, researchers employ additional methods on top of tracking schemes to track groups as well as individuals. For instance, Yigit and Temizel (2015) leveraged characteristics of groups such as joining, merging, and splitting of group members to build a model that uses a particle filter to track in-group (members in a group) and out-group (members not in a group) pedestrians. Leal et al. (2011) combined the minimum-cost network flow problem with the social force model to detect and track groups and individual pedestrians (Yamaguchi et al. 2011). In addition, Qin and Shelton (2012) used a clustering approach by applying Lagrangian optimization to design a two-stage iterative algorithm that employed Hungarian K-means clustering to track and detect groups.

Although pedestrian tracking methods contribute to detecting groups in social spaces, this problem becomes more challenging when deployed on a mobile robot (Luber and Arras 2013). For instance, some of the aforementioned methods for detecting individuals and groups are computationally expensive, which are not suitable given most platforms have limited on-board computational resources. Another issue is that many of the aforementioned techniques employed overhead cameras, which are rarely practical for real robots in the real world.

Detecting and tracking F-Formations using these types of methodologies could help mitigate some of these issues. However, this problem is non-trivial. It is challenging to approximate head pose, body posture, position, and spatial orientation of people in groups. To combat these issues, many methods have been proposed, which leverage machine learning and computer vision techniques. Both machine learning and computer vision provide strategies for approaching the F-Formation detection problem as a statistics and classification problem.

For example, Cristani et al. (2011) detected F-Formations by modeling the o-space as a random Gaussian distribution and used a Hough voting approach. Setti et al. (2013) built on this work to detect F-Formations of different scales using maximum weighted Boltzmann entropy.

Tran et al. (2013) designed a graph based clustering algorithm that incorporates how much humans are interacting to detect F-Formations. In addition, Vascon et al. (vascon2014game) extracted features from body-worn accelerometers and employed a Hidden Markov Model to estimate conversational groups. Ricci et al. (2015) performed multi-target tracking, extracted head and body pose features from the tracking observations, and then jointly estimated head pose, body pose, and F-Formations.

Setti et al. (2015) published a comprehensive comparison of all these F-Formation systems using publicly available F-Formation datasets. Also, they compared their most recent algorithm for detecting F-Formations using a graph-cut framework for clustering groups.

F-Formation systems have also been explored in the robotics community by Vazquez et al. (2015). They were the first to use lower body orientations to predict F-Formations as many robots are shorter than humans. While this was a great contribution to integrating robots into social spaces, they evaluated their method on a 2D, overhead video data set, so the method may not easily scale to being deployed on a physical mobile robot, which does not have access to exocentric cameras (Zen et al. 2010, cf). Therefore, an open question remains of how to deploy an F-Formation detection system on a mobile robot in naturalistic settings.

After reviewing the aforementioned computer vision problems, an open question is how do these techniques adapt to working on a robot, as opposed to a standalone machine? Algorithms in robotics must have near real-time performance, the computing devices must be light-weight (low memory), and the algorithms must consider unpredictability in human spaces. On the other hand, computer vision researchers do not usually have to concern themselves with these factors, and usually sacrifice computational efficiency to achieve high accuracy.

In fact, many computer vision algorithms are GPU based to speed up the algorithms, but roboticists do not have this luxury as their platforms typically have hardware resource constraints. Therefore, robots need computationally inexpensive detection and tracking methods with comparable accuracy as pedestrian tracking methods used in computer vision.

4 Discussion

We have reviewed the current state-of-the-art in human group perception, and discussed how this work might be adaptable to mobile robots. Low cost, non-intrusive sensors such as the Kinect have greatly advanced the fields of social signal processing and computer vision; however, a few major challenges remain. For example, it remains extremely difficult to analyze gesture and postural data from multiple people simultaneously in real-world, dynamic settings.

Other sensing modalities with larger fields-of-view might help overcome this issue, such as panoramic monocular or depth cameras (Viswanathan, Pires, and Huber 2016).

Proxemics may provide another path forward to understanding group behavior; prior work in HRI suggests it may be a beneficial mechanism. Perhaps global feature processing can be employed by robots to develop rough approximations of human group proxemics (Nigam and Riek 2015). Another approach is to explore satisficing approaches to group proxemic estimation (Riek 2013).

Pedestrian tracking has been studied for a long time and researchers have advanced the tracking capabilities of their algorithms. Some of the methods discussed in this paper are computationally expensive, which are not suitable given that most robotic platforms have limited on-board computational resources. This suggests that algorithms for pedestrian detection must be computationally inexpensive without sacrificing accuracy. We have been doing work recently in this space by designing new methods to suggest regions of interest and reduce the search space for existing pedestrian detectors, and have been able to significantly reduce the computation time of leading algorithms (Chan and Riek 2016).

The computer vision field has made great strides in perception problems; however, some open questions remain. For example, might these techniques discussed in this paper work in the absence of overhead cameras? Much of the prior work employed fixed, overhead cameras to sense groups, which will be impractical for robots in real-world settings. What is the path forward to sensing groups from an egocentric perspective? Can egocentric methods teach robots how to approach or join groups in social spaces?

Some future directions to progress toward a mobile robot that can socially interact with humans, is to explore how a robot can adapt to social behavior of humans in a group. Although there has been work done to adapt a robot's behavior to a group of humans (Iqbal, Rack, and Riek 2016; Iqbal, Moosaei, and Riek 2016), this work was applied in a laboratory setting; whereas, we are interested in robots that can interact with groups "in the wild." This approach to group analysis will lay the groundwork for robots that can handle real world challenges.

We encourage roboticists to address the challenges discussed and this work as they contribute to enabling robots to perceive, learn, and adapt to humans in naturalistic settings.

5 Acknowledgements

Some research reported in this article is based upon work supported by the National Science Foundation under Grant No. IIS-1527759.

References

Andriyenko, A., and Schindler, K. 2010. Globally optimal multi-target tracking on a hexagonal lattice. In *Computer Vision*. Springer.

Betancourt, A.; Morerio, P.; Regazzoni, C. S.; and Rauterberg, M. 2015. The evolution of first person vision methods: A survey. *Circuits and Systems for Video Technology* 25(5).

Breitenstein, M. D.; Reichlin, F.; Leibe, B.; Koller-Meier, E.; and Van Gool, L. 2011. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE transactions on pattern analysis and machine intelligence* 33(9).

Brendel, W.; Amer, M.; and Todorovic, S. 2011. Multiobject tracking as maximum weight independent set. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Butler, J. T., and Agah, A. 2001. Psychological effects of behavior patterns of a mobile personal robot. *Autonomous Robots* 10(2).

Chan, D., and Riek, L. D. 2016. A RGB-D region proposal generator for faster robot perception - in review.

Choi, W., and Savarese, S. 2010. Multiple target tracking in world coordinate with single, minimally calibrated camera. In *European Conference on Computer Vision*. Springer.

Choi, W.; Pantofaru, C.; and Savarese, S. 2013. A general framework for tracking multiple people from a moving camera. *Pattern Analysis and Machine Intelligence* 35(7).

Cristani, M.; Bazzani, L.; Paggetti, G.; Fossati, A.; Tosato, D.; Del Bue, A.; Menegaz, G.; and Murino, V. 2011. Social interaction discovery by statistical analysis of f-formations. In *BMVC*, volume 2.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *IEEE Computer Vision and Pattern Recognition (CVPR'05)*, volume 1. IEEE.

Dautenhahn, K.; Walters, M.; Woods, S.; Koay, K. L.; Nehaniv, C. L.; Sisbot, A.; Alami, R.; and Siméon, T. 2006. How may i serve you?: a robot companion approaching a seated person in a helping context. In *ACM SIGCHI/SIGART Human-robot interaction*. ACM.

Dollar, P.; Wojek, C.; Schiele, B.; and Perona, P. 2012. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence* 34(4).

D'Orazio, T.; Marani, R.; Renò, V.; and Cicirelli, G. 2016. Recent trends in gesture recognition: how depth data has improved classical approaches. *Image and Vision Computing*.

Escalera, S.; Athitsos, V.; and Guyon, I. 2016. Challenges in multimodal gesture recognition. *Journal of Machine Learning Research* 17(72).

Grafsgaard, J. F.; Boyer, K. E.; Wiebe, E. N.; and Lester, J. C. 2012. Analyzing posture and affect in task-oriented tutoring. In *FLAIRS Conference*.

Hall, E. T. 1966. The hidden dimension. *Garden City, N.Y., Doubleday* 14.

Iqbal, T.; Moosaei, M.; and Riek, L. D. 2016. Tempo adaptation and anticipation methods for human-robot teams. *Robotics, Science and Systems (RSS), Planning for Human-Robot Interaction*.

Iqbal, T.; Rack, S.; and Riek, L. D. 2016. Movement coordination in human-robot teams: A dynamical systems approach. *IEEE Transactions on Robotics. In press.*

Karg, M.; Samadani, A.-A.; Gorbet, R.; Kuhnlenz, K.; Hoey, J.; and Kulić, D. 2013. Body movements for affective expression: A survey of automatic recognition and generation. *IEEE Transactions on Affective Computing* 4(4).

Kendon, A. 1976. The f-formation system: The spatial organization of social encounters. *Man-Environment Systems* 6.

Kendon, A. 1990. *Conducting Interaction: Patterns of Behavior in Focused Encounters*, volume 7. The Press Syndicate of the University of Cambridge.

Khoramshahi, M.; Shukla, A.; Raffard, S.; Bardy, B. G.; and Billard, A. 2016. Role of gaze cues in interpersonal motor coordination: towards higher affiliation in human-robot interaction. *Plos One* 11(6).

Kruse, T.; Basili, P.; Glasauer, S.; and Kirsch, A. 2012. Legible robot navigation in the proximity of moving humans. In *Advanced Robotics and its Social Impacts (ARSO)*. IEEE.

Leal-Taixé, L.; Pons-Moll, G.; and Rosenhahn, B. 2011. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *Computer Vision Workshops (ICCV)*. IEEE.

Lee, D.; han Yun, W.; kyu Park, C.; Yoon, H.; Kim, J.; and Park, C. 2015. Measuring the engagement level of children for multiple intelligence test using kinect. In *Machine Vision (ICMV 2014)*. International Society for Optics and Photonics.

Liu, Z.; Zhou, L.; Leung, H.; and Shum, H. 2015. Kinect posture reconstruction based on a local mixture of gaussian process models. *IEEE transactions on visualization and computer graphics*.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999*, volume 2. Ieee.

Luber, M., and Arras, K. O. 2013. Multi-hypothesis social grouping and tracking for mobile robots. In *Robotics: Science and Systems*.

Mead, R., and Matarić, M. J. 2015. Proxemics and performance: Subjective human evaluations of autonomous sociable robot distance and social signal understanding. In *Intelligent Robots and Systems (IROS)*. IEEE.

Moosaei, M.; Hayes, C. J.; and Riek, L. D. 2015. Performing facial expression synthesis on robot faces: A real-time software system. *AISB Symposium on New Frontiers in Human-Robot Interaction*.

Mota, S., and Picard, R. W. 2003. Automated posture analysis for detecting learner's interest level. In *Computer Vision and Pattern Recognition*, volume 5. IEEE.

Mumm, J., and Mutlu, B. 2011. Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Human-robot interaction*. ACM.

Nigam, A., and Riek, L. D. 2015. Social context perception for mobile robots. In *Intelligent Robots and Systems (IROS)*. IEEE.

O'Connor, M. F., and Riek, L. D. 2015. Detecting social context: A method for social event classification using naturalistic multimodal data. In *Automatic Face and Gesture Recognition (FG)*, volume 3. IEEE.

Oosterhout, T. v.; Visser, A.; et al. 2008. A visual method for robot proxemics measurements. *Technical Report*.

Palaghias, N.; Hoseinitabatabaei, S. A.; Nati, M.; Gluhak, A.; and Moessner, K. 2016. A survey on mobile social signal processing. *ACM Computing Surveys (CSUR)* 48(4).

Pitsikalis, V.; Katsamanis, A.; Theodorakis, S.; and Maragos, P. 2015. Multimodal gesture recognition via multiple hypotheses rescoring. *The Journal of Machine Learning Research* 16(1).

Qin, Z., and Shelton, C. R. 2012. Improving multi-target tracking via social grouping. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Ricci, E.; Varadarajan, J.; Subramanian, R.; Rota Bulo, S.; Ahuja, N.; and Lanz, O. 2015. Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos. In *Computer Vision*. IEEE.

Rich, C.; Ponsler, B.; Holroyd, A.; and Sidner, C. L. 2010. Recognizing engagement in human-robot interaction. In *ACM/IEEE Human-Robot Interaction (HRI)*. IEEE.

Riek, L. D.; Paul, P. C.; and Robinson, P. 2010. When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. *Journal on Multimodal User Interfaces* 3(1-2).

Riek, L. D. 2013. The social co-robotics problem space: Six key challenges. *In Proceedings of Robotics: Science, and Systems (RSS), Robotics Challenges and Visions*.

Rios-Martinez, J.; Spalanzani, A.; and Laugier, C. 2015. From proxemics theory to socially-aware navigation: a survey. *International Journal of Social Robotics* 7(2).

Satake, S.; Kanda, T.; Glas, D. F.; Imai, M.; Ishiguro, H.; and Hagita, N. 2013. A robot that approaches pedestrians. *IEEE Transactions on Robotics* 29(2).

Setti, F.; Lanz, O.; Ferrario, R.; Murino, V.; and Cristani, M. 2013. Multi-scale f-formation discovery for group detection. In *Image Processing*. IEEE.

Setti, F.; Russell, C.; Bassetti, C.; and Cristani, M. 2015. F-formation detection: Individuating free-standing conversational groups in images. *PloS one* 10(5).

Singh, G.; Nelson, A.; Robucci, R.; Patel, C.; and Banerjee, N. 2015. Inviz: Low-power personalized gesture recognition using wearable textile capacitive sensor arrays. In *Pervasive Computing and Communications (PerCom)*. IEEE.

Song, Y.; Demirdjian, D.; and Davis, R. 2011. Multi-signal gesture recognition using temporal smoothing hidden conditional random fields. In *Automatic Face & Gesture Recognition*. IEEE.

Soo Park, H., and Shi, J. 2015. Social saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4777–4785.

Tran, K. N.; Bedagkar-Gala, A.; Kakadiaris, I. A.; and Shah, S. K. 2013. Social cues in group formation and local interactions for collective activity analysis. In *VISAPP (1)*.

Vázquez, M.; Steinfeld, A.; and Hudson, S. E. 2015. Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation. In *Intelligent Robots and Systems (IROS)*. IEEE.

Vinciarelli, A.; Pantic, M.; and Bourlard, H. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27(12).

Viswanathan, A.; Pires, B. R.; and Huber, D. 2016. Vision-based robot localization across seasons and in remote locations. In *Robotics and Automation (ICRA)*. IEEE.

Walia, G. S., and Kapoor, R. 2016. Recent advances on multicue object tracking: a survey. *Artificial Intelligence Review* 46(1).

Wickstrom, G., and Bendix, T. 2000. The" hawthorne effect"what did the original hawthorne studies actually show? *Scandinavian journal of work, environment & health*.

Wu, D.; Pigou, L.; Kindermans, P.-J.; Nam, L.; Shao, L.; Dambre, J.; and Odobez, J.-M. 2016. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Yamaguchi, K.; Berg, A. C.; Ortiz, L. E.; and Berg, T. L. 2011. Who are you with and where are you going? In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Yigit, A., and Temizel, A. 2015. Particle filter based conjoint individual-group tracker (cigt). In *Advanced Video and Signal Based Surveillance (AVSS)*. IEEE.

Zen, G.; Lepri, B.; Ricci, E.; and Lanz, O. 2010. Space speaks: towards socially and personality aware visual surveillance. In *ACM Multimodal pervasive video analysis*. ACM.