Semi-automated Analysis of Collaborative Interaction: Are We There Yet?

THOMAS NEUMAYR, University of Applied Sciences Upper Austria, Austria and Johannes Kepler University Linz, Austria
MIRJAM AUGSTEIN, University of Applied Sciences Upper Austria, Austria
JOHANNES SCHÖNBÖCK, University of Applied Sciences Upper Austria, Austria
SEAN RINTEL, Microsoft Research, UK
HELMUT LEEB, University of Applied Sciences Upper Austria, Austria
THOMAS TEICHMEISTER, University of Applied Sciences Upper Austria, Austria

In recent years, research on collaborative interaction has relied on manual coding of rich audio/video recordings. The fine-grained analysis of such material is extremely time-consuming and labor-intensive. This is not only difficult to scale, but, as a result, might also limit the quality and completeness of coding due to fatigue, inherent human biases, (accidental or intentional), and inter-rater inconsistencies. In this paper, we explore how recent advances in machine learning may reduce manual effort and loss of information while retaining the value of human intelligence in the coding process. We present ACACIA (AI Chain for Augmented Collaborative Interaction Analysis), an AI video data analysis application which combines a range of advances in machine perception of video material for the analysis of collaborative interaction. We evaluate ACACIA's abilities, show how far we can already get, and which challenges remain. Our contribution lies in establishing a combined machine and human analysis pipeline that may be generalized to different collaborative settings and guide future research.

CCS Concepts: • **Human-centered computing** → **User studies**; *Computer supported cooperative work*; *Interaction devices*; **User studies**; *Computer supported cooperative work*; *Interaction devices*; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: collaboration analysis, empirical studies, observational data, data analysis, artificial intelligence

ACM Reference Format:
Thomas Neumayr, Mirjam Augstein, Johannes Schönböck, Sean Rintel, Helmut Leeb, and Thomas Teichmeister. 2022. Semi-automated Analysis of Collaborative Interaction: Are We There Yet?. *Proc. ACM Hum.-Comput. Interact.* 6, ISS, Article 571 (December 2022), 27 pages. https://doi.org/10.1145/3567724

Authors' addresses: Thomas Neumayr, thomas.neumayr@fh-hagenberg.at, University of Applied Sciences Upper Austria, Softwarepark 11, Hagenberg, Austria, 4232 and Johannes Kepler University Linz, Altenberger Strasse 69, Linz, Austria, 4040; Mirjam Augstein, mirjam.augstein@fh-hagenberg.at, University of Applied Sciences Upper Austria, Softwarepark 11, Hagenberg, Austria, 4232; Johannes Schönböck, johannes.schoenboeck@fh-hagenberg.at, University of Applied Sciences Upper Austria, Softwarepark 11, Hagenberg, Austria, 4232; Sean Rintel, serintel@microsoft.com, Microsoft Research, 21 Station Rd, Cambridge, UK, CB1 2FB; Helmut Leeb, helmut.leeb@fh-hagenberg.at, University of Applied Sciences Upper Austria, Softwarepark 11, Hagenberg, Austria, 4232; Thomas Teichmeister, thomas.teichmeister@fh-hagenberg.at, University of Applied Sciences Upper Austria, Softwarepark 11, Hagenberg, Austria, 4232.

2573-0142/2022/12-ART571
https://doi.org/10.1145/3567724

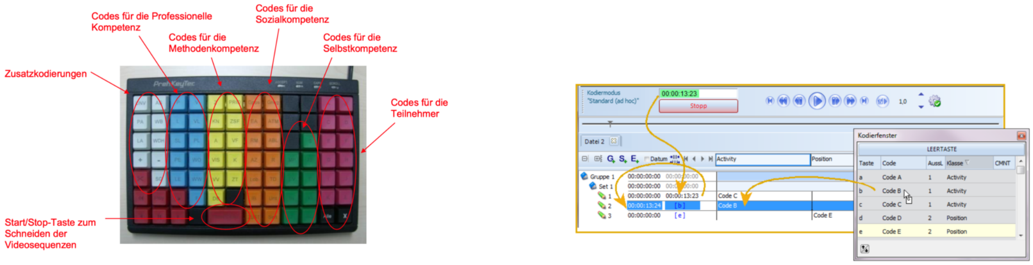
1 INTRODUCTION

Many of the advances in Computer-Supported Cooperative Work (CSCW) and Human-Computer Interaction (HCI) come from studying the intricacies of techno-social collaboration phenomena such as territorial functioning [41], collaborative coupling [16, 32, 42], awareness and collaborative work [43], and multi-device collaboration [6]. Such fine-grained analyses typically require extensive audio/video recordings of both people's activities in space and with devices, as well as recordings of what is happening on device screens. Even for small groups and short term use, the analysis and interpretation of such rich data has significant costs in terms of human time and effort. Researchers report that the observation and analysis of group interaction is "labour-intensive" [7], and "tedious and time consuming" [21]. The effort is considered necessary to obtain valid empirical findings, but its cost is so great that it creates a scaling problem [15] which effectively prohibits studies of large numbers of groups and/or longitudinal studies.

Systematic procedures for capturing group interaction behavior based on observational data collected in field studies have existed since at least the 1930s (see, e.g., [34, 45] as described in detail by Kauffeld & Meinecke [21]). In early studies, activities had to be captured in coding schemes immediately during observations, often by multiple parallel observers, which in turn led to first considerations around inter-rater reliability. Reliability has also been raised as a question because many coding schemes are quite idiosyncratic, based on dubious foundations, or leading to overly strong claims. Birdwhistell's kinegraphics was accused of all of the three problems, and despite his influence on more well-known researchers such as Hall, Kendon, Goffman, and Knapp, his coding schemes and theories have largely fallen out of favour [19].

Advances in audio and video recording capabilities in the 1960s and 1970s [22] made it possible to partially replace other data gathering approaches such as observations or interviews [26]. Accordingly, the cognitive load placed on researchers by simultaneous observation and coding was alleviated at least to the extent that the recorded material could be thoroughly analyzed at a later time. In some instances, this gave rise to new fields of research, such as Conversation Analysis (CA). CA researchers coalesced around Jefferson's [17] basic principles of audio transcription. In subsequent decades, CA researchers have added a range of elements for the system to cover a wide range of human multi-modal activity such as "grammar, lexicon, prosody, gesture, gaze, body postures, movements, manipulations of artifacts, etc." [28]. However, these additions have not propagated evenly across the research community, and may in some instances be either too detailed or not detailed enough, so individual researchers often create their own idiosyncratic transcription methods to capture the phenomena that they are exploring. The huge potential of such detailed coding approaches is in direct contrast to the immense and ever-increasing manual coding effort involved – and it takes considerable training just to learn [13]. According to Mondada [28], several days of *a posteriori* work are necessary to code a single one-hour video with all the detail above. Taken together, these issues of effort and reliability lead to an urgent need for methods, techniques, and technologies that help to reduce the amount of human effort involved in the process.

One approach in this direction is the introduction of tools such as the act4teams coding scheme [20] which is supplemented by customizable software and a specifically designed coding keyboard (see Figure 1). Another example is EagleView [7], a system for user studies of proxemics (the position of people and objects) that visualizes spatial interactions. Such approaches "enable more straightforward data analysis and efficient pattern tracking" [22] and can help reduce researchers' workload by taking over or simplifying the handling of routine tasks in the manual coding process and guiding researchers' coding decisions by standardized schemes. However, many of the fundamental challenges related to manual coding cannot be solved just by improving the manual coding


Codes für die Professionelle Kompetenz
Codes für die Sozialkompetenz
Codes für die Methodenkompetenz
Codes für die Selbstkompetenz
Zusatzkodierungen
Codes für die Teilnehmer
Start/Stop-Taste zum Schneiden der Videosequenzen
Kodiermodus "Standard (ad hoc)"
00:00:13:23
Stopp
1,0
Datei 2
Datum
Activity
Position
Gruppe 1
Set 1
1
2
3
00:00:00:00
00:00:13:23
00:00:13:24
[b]
[e]
Code C
Code B
Code E
Kodierfenster
LEERTASTE

Taste	Code	AussL	Klasse	CMNT
a	Code A	1	Activity	
b	Code B	1	Activity	
c	Code C	1	Activity	
d	Code D	2	Position	
e	Code E	2	Position	

Fig. 1. Example of specialized hardware [1] (left) and software [27] (right) to help with manual coding of audio-video data; annotations are in German.

process. These challenges include the amount of human effort, fatigue, inconsistent inter-rater reliability, and (intentional or accidental) human biases that might affect coding outcomes.

Our approach to these challenges is rooted in the idea of bringing machine perception into the analysis process of collaborative interaction (a "machine-in-the-loop" inversion of the "human-in-the-loop" concept). In this paper, we introduce a **semi-automated research pipeline** which uses multiple recent Artificial Intelligence (AI) technologies combined into a research toolset. We call this system ACACIA (AI Chain for Augmented Collaborative Interaction Analysis). In ACACIA, machine perception takes a first pass at time-consuming steps in the coding process that are easy for machines to solve but hard for people to do reliably over a large data set (e.g., identification of people, objects and regions in a video image including computation of proxemics or identification of interaction in certain regions or territories). This presents researchers with a range of searchable visualizations which may then be used to analyse collaborative activity starting from the same baseline of observable facts. As a proof-of-concept, we present an illustration and evaluation of a collaborative interaction scenario analyzable with the help of the ACACIA infrastructure interwoven with human endeavor. We hope that our exploration inspires other researchers to apply a similarly structured research process. Further, our insights into the use of different AI technologies applied to the analysis of collaborative interaction provide useful starting points for researchers to consider how to move the needle on new approaches to scaling empirical analysis.

In summary, the contributions of our paper can be outlined as follows. Foremost, we have (i) developed a framework that combines machine- and human-related efforts into a consistent research pipeline for collaborative contexts (which can be generalized and guide future related studies). Our approach further has the potential to (ii) free resources of tedious manual work (which might also reduce bias and inconsistencies), (iii) improve scaling issues in research endeavors (e.g., allowing for the analysis of user studies with more overall participants, longitudinal studies, or replication studies), and (iv) facilitate the analysis of larger volumes of data (e.g., a vast amount of video recordings).

2 RELATED WORK

Our overview of related literature is twofold. First, we describe exemplary manual coding approaches in the domains of HCI and CSCW, some of which will be later discussed for their potential to be (partially) automated with ACACIA. Second, we give an overview of related automation approaches that utilize AI in different domains.

2.1 Manual Coding Approaches in HCI and CSCW

A wealth of different coding approaches is described in *The Cambridge Handbook of Group Interaction Analysis* [5]. As noted above, the act4teams coding scheme aims to "measure the fine-grained problem-solving dynamics that occur in groups and teams" [20]. It focuses on verbal group communication and distinguishes four different categories of statements: problem-focused, procedural, socio-emotional, and action-oriented. With the addition of nearly 40 subcategories, the coding scheme has considerable descriptive power in analyzing collaborative settings such as "team meetings, group problem-solving conversations, and group creativity interactions". The authors estimate a time factor of about 1:8-15 for coding, which means that 1 minute of audio-video material to be coded results in 8 to 15 minutes of coding time. Additionally, the time required for training the coders is assumed to be approximately 200 hours. Further estimates of the effort of different coding schemes amount to factors up to 1:30 [38] or even 1:50 [39]. This leads to the aforementioned situation where one hour of audio-video material results in several days of coding—even by experienced coders.

Although not a coding scheme *per se*, Scott et al. [41] conducted extensive manual coding to shed light on territorial functioning in co-located collaboration at conventional table surfaces. In an initial study, they coded exactly where on a round table the participants interacted with their hands (see Figure 2) in conjunction with their other behaviour:

> "Sessions were videotaped and audiotaped, and field notes were recorded. We collected 29, 43, and 38 minutes of data from Groups 1-3, respectively.
> In order to analyze the participants' spatial interactions, their tabletop activity was transcribed from the video data. Transcripts included all tabletop actions, the initiator of each action, the location of each action, the location of each participant, and any conversation related to the tabletop actions. To facilitate our analysis, the tabletop workspace was divided into 16 directional zones [...], and 4 radial zones [...]"

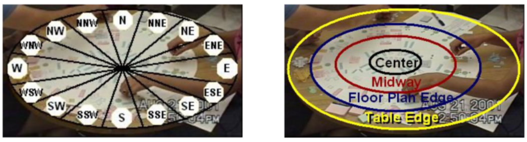


Fig. 2. Example of territorial codes manually assigned to interactions by Scott et al. [41].

The sheer amount of information that had to be manually extracted from the audiovisual recording gives an idea of how much effort must have been necessary to obtain the findings. Our vision is to enable future researchers to more easily conduct such endeavors—either with fully or partially automated support—reducing tedious coding activities and, consequently, stirring greater interest in investigating group interaction scenarios.

Another interesting example, this time involving a relatively high level of human interpretation, is the coding of collaborative coupling. First introduced by Tang et al. [42] this framework of analysis allows a fine-grained categorization of the closeness of collaboration among several persons into several so-called coupling styles. Tang et al.'s coupling styles were later extended by Isenberg et al. [16] with a revised list of eight coupling styles to fit their study setting. Their definitions of the coupling styles can be regarded as pointers towards how researchers interpreted and coded each collaborative interaction between participants as one of their eight coupling styles. For example, Isenberg et al.'s "Sharing of the same view" is defined to involve "[p]articipants either look at the same document reader or the same search result list together at the same time". One further extension of the concept of collaborative coupling is described in Neumayr et al. [32], who introduced the possibility of several coupling styles being in existence in parallel for multiple subgroups. We estimate that this task of deciding which coupling style is currently in place (which is also far from trivial for human coders) for which persons will also be the most challenging part of any automated approaches.

In essence, most coding activities (be it for act4teams, territoriality, or collaborative coupling) in this context consider collaborative settings with their entities and how relationships between those entities unfold over time. The entities can be either active (such as humans, intelligent user interfaces, or robots) or passive (such as objects like devices, tools, screens, artifacts, regions or territories). People as actors are mostly considered behaviorally by human coders through their implicit and explicit actions, which can be expressed through touching, looking at, pointing to, talking and gesturing, or moving around. These actions—while possible for isolated individuals (as active entities)—are of interest in collaborative settings mainly when conducted in relation to other (both active and passive) entities. The core of our **vision**, therefore, is to use AI to recognize all related **entities** and establish meaningful **relationships** between them.

2.2 Automation Approaches

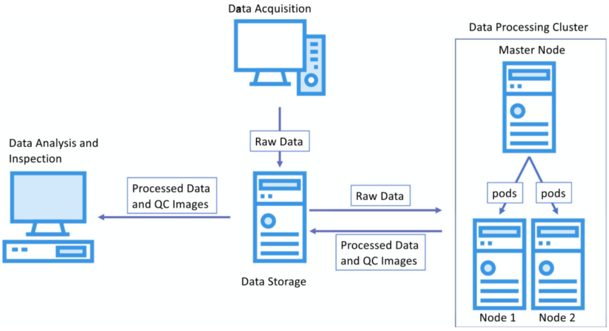


Fig. 3. Illustration of the overall automation approach of Gonzáles & Evans taken from [12]. Our general approach and system architecture is loosely inspired by their suggestions.

In the domain of biomedical research, Gonzáles & Evans [12] suggested a research pipeline for automated image processing of dark-field microscopy images. As in our proposed approach, they also expect "streamlining the path from experiment to conclusions". They suggest a number of technologies that could be used to achieve such a pipeline and present an overview of common cloud storage and computing providers. In particular, their contribution to present "a data processing pipeline based on AI to automatically process imaging data" (see Figure 3) is related to and inspired our approach. We also agree with them that the move "to full automation removes the burden of step-by-step manual data analysis and frees the scientist to focus on experiment design, data acquisition, and interpretation." In their "automated data processing pipeline for microscopy" they extract certain "regions of interest" and further information and this is then prepared as tabular data (to be later analyzed in a software such as R). Furthermore, they state that automated results need to be quality controlled by a human which is certainly even more important in a field that provides the basis of decision-making in medicine. A similar approach for the automated analysis of microscopy data is described by Kraus et al. [24] who present the application of a deep convolutional neural network to analyze such image data.

Hoey et al. [15] share our general motivation and idea of using AI to address the drawbacks concerning human coding effort of group interaction data. In their approach, they use machine learning to evaluate hundreds of randomly selected GitHub pull request comments and categorize them according to Bales' interaction process analysis [2] coding categories and several additional emotion words (such as "cautious", "happy", or "nervous"). In their preliminary results they achieved some poor F1-scores (defined as "evenly weighted precision and recall" [15]) which give an impression how reliably a category could be distinguished from all other categories. Yet, for some aggregated measures the F1-scores were much more promising, which encouraged us to combine several (partly experimental) services to achieve better results with aggregates.

Another prominent example of an automated approach using computer vision is the ASSESS MS project [30], which measures the motor skills of Multiple Sclerosis patients. The software and hardware system is designed for real-world clinical use and automatically interprets visual data (e.g., of diagnostic tasks such as the finger-to-nose test) with machine learning based on depth information provided by the Kinect sensor [8].

In research on verbal information, Ullmann [44] used several machine learning algorithms to automatically assess the quality of reflective writing in education according to several categories such as the depth of the text, personal belief, or perspective. He showed that most of the categories can be automatically assessed with "substantial or almost perfect reliability". Similarly, Portnoff et al. [35] used Natural Language Processing to extract the type, product, and price information of cybercriminal underground forum posts in an approach of automated analysis. They were successful in extracting and automatically classifying the information across as well as within several related forums.

Another approach for annotating data is Label Studio[1]. Here, video, audio, text, or images can be labeled both manually and in an automated way. For automation, it is possible to integrate own algorithms as plugins which then return annotated data. Our approach also gives researchers the opportunity to manually annotate data, but we support researchers with a toolset that can be used out of the box and is tailored specifically to the domain of analyzing collaborative interaction, as opposed to the more general orientation of Label Studio. Moreover, we also use data from Azure Kinect and synchronize several data streams with the help of \psi, which both Label Studio cannot easily process.

Further approaches that employ (semi-)automated analysis with the help of AI and aim at reducing effort or interobserver variability (or issues with inter-rater reliability) have been conducted and discussed in different disciplines such as medicine (only few of a wealth of examples are [18], [31], [25], [36], and [40]), astronomy [14], product configuration [3], management [22], or ecology [11]. While many of the approaches are aimed at automated analysis of still image data (mainly in medicine), some are based on digital text (mainly in CSCW) but to the best of our knowledge, there are no approaches that combine different sources (e.g., several images from a video recording taking into account the history, proximity, or gaze detection) for the analysis of collaborative interaction.

As we have shown in the related literature, manual approaches to coding are instrumental and prevalent but associated with a great deal of effort (potentially leading to various problems discussed above). While there are multiple approaches to automate part of these efforts and we can resort to their general strategies, none of them are usable in the domain of collaborative interaction out of the box. As a first step to close this gap, we designed, implemented, and evaluated ACACIA to answer the question how far we can get by introducing currently available AI services into the research process.

[1] https://labelstud.io, last access May 4th, 2022

3 DESIGN AND DEVELOPMENT OF ACACIA

We have two goals. First, we aim to reduce the tedious manual work usually associated with the analysis of audiovisual material obtained in the context of user studies on collaborative interaction. This should also help improve the scaling of research, facilitating the analysis of a larger amount of recordings. Second, we aim to prevent possible accidental or intentional biases or inter-rater inconsistencies in the analysis. In the following section, we present our conceptual architecture as well as a high-level overview of our prototype.

3.1 Conceptual Architecture

To analyze collaborative interaction our central premise is that it is essential to find out who (i.e., which active entity) was interacting with what (i.e., which passive entity) or whom (i.e., which other active entity) at which point in time, thus, establishing a relationship between entities. Our goal is to gather rich information about how people collaborate or interact with each other or with other tools and items (e.g., detecting the closeness of collaboration, use of territories, or regions of interest). Conceptually and behaviorally, such a relationship of "interacting with" can be understood in a variety of ways, often combining aspects of "looking at," "touching," "being close to," or "talking to" someone or something. A number of external APIs and AI services can be used to semi-automatically detect these aspects (cf. Figure 7 for an overview). For instance, to identify the passive and active actors, we can make use of automated face detection and identification mechanisms. To find out which objects (or regions) they interacted with, these objects first need to be recognized or manually annotated. To find out who was looking at someone, gaze tracking can help. Semi-automatic recognition of these entities and their relationships to each other enables later analysis of study data using established models and frameworks, as discussed in more detail in Section 5.

Most of the aforementioned questions can be answered on a per-frame basis (e.g., who is currently visible or looking at something). However, there are also questions which cannot be answered (with corresponding certainty and probability) based on a *single* frame but require the analysis of several frames depicting different points in time. One such question is: Where does a person move to (it is necessary to track the person in cases where the person is currently facing away from the camera or the face is currently covered)?

The other important aspect of the system is to track the relationship between objects and people. In a **pre-processing** step, **entities** (such as objects, regions, or people) are recognized and contextualized by the automated analysis. In this step, face detection and computer vision services are used to determine which persons, regions, or objects are visible in each frame and where they are located. The intention behind the **pre-processing step** is to allow for later advanced and configurable **post-processing** functions, such as creating **relationships between the detected entities**, without needing to repeatedly process all the data. Researchers can configure and start real-time post-processing to establish relationships between the entities, answering the question who focuses on what or whom.

3.2 ACACIA in a Nutshell

Our research prototype ACACIA (AI Chain for Augmented Collaborative Interaction Analysis) is a desktop application developed with Windows Presentation Foundation (WPF) to explore the current capabilities of readily available as well as experimental AI services for the analysis of collaborative interaction (cf. Figure 4). Our underlying aims were twofold. First, we strove to combine all the different services by integrating them in a plug & play-style manner as they all deliver JSON data, which is mainly a technical matter (and contribution). Second, we aimed at developing a prototoype well usable for interested researchers to aid the research process. The latter is a more general contribution to the field of HCI, a domain which is strongly characterized by the conduct of user studies. Please note that ACACIA is not intended as a consumer-style application concerning usability and user experience yet. In the following, we describe ACACIA, highlighting its capabilities from (i) a technical and (ii) an HCI perspective.

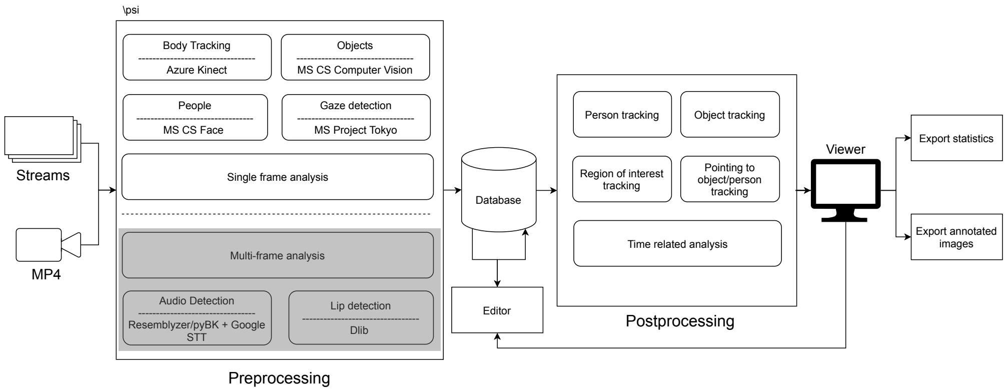


Fig. 4. High level depiction of ACACIA's research pipeline. (Pointing currently is in an experimental state and was not evaluated, yet).

Technical Perspective. ACACIA can handle two kinds of video formats: (i) the standard audio-visual container format mp4 and (ii) an Azure Kinect stream with additional information including depth information as input data. The depth information increases the accuracy of capturing a collaborative scenario (e.g., gaze estimation).The uniqueness of our approach stems from being able to combine multiple machine perception services (such as Microsoft's Project Tokyo[2], or Microsoft Azure Cognitive Services) with data sources (2D as well as 3D image information). We use a combination of built-in components and additional external services in order to detect the aspects we are interested in – in this case for analyzing collaborative interaction, but different components could be used for different phenomena.

The central component of our system is the Microsoft open-source Platform for Situated Intelligence (\psi), [4]. In general, \psi allows for performing different actions on data by connecting existing and self-developed components to process data. \psi is designed to write small components and connect them through a pipeline (or chain) which executes different actions (in parallel) in a predefined order. Furthermore, the framework provides various input sources such as real-time camera input, video files, or Azure Kinect sensors. We use \psi for recording videos and storing the data as streams. These streams contain all data (including images, depth images, etc.) which can later on be processed by our system. Since the data is stored, we can again read it and perform various analyses at a later point in time.

We also use \psi to perform different actions with different frame rates. For example, for tracking gestures one must use a higher frame rate since limbs are usually moving relatively fast. Instead, object detection does not necessarily have to be performed in such a high frame rate (many objects are comparatively stationary), thus, preventing unnecessary costs and processing time.

[2] https://www.microsoft.com/en-us/research/project/project-tokyo/, last access: February 28th, 2022

HCI Perspective. Based on our experiences with previous collaboration analysis and their according pain points (as outlined in Section 1), in an iterative design process we integrated solutions into ACACIA's frontend to tackle these requirements. Besides providing developers the opportunity to integrate additional services into the research pipeline in the backend (either in the pre or post-processing phases), ACACIA's frontend allows researchers to adjust and approve which information is automatically detected. Researchers are able to add, edit, and delete objects, regions, and people, using a dedicated editor. Editing the pre-processing phase to include combinations of automatically-detected and manually edited elements can be done without affecting the post-processing algorithm. For example, regions of interests can be added, or a supervisor of a study present during the study procedure can be removed (i.e., excluded from the analysis).

Finally, the results of the analysis can be viewed via ACACIA's GUI and exported as compressed data including different kinds of annotated images (see, e.g., Figure 5) and statistical information (see, e.g., Tables 1 and 2, which are discussed in more detail in Section 3.3). The annotations are drawn based on information received from the respective APIs. For example the position and size of rectangles around the faces originate from Face API's `FaceRectangle` class, the ones around objects from Computer Vision as properties x, y, w, and h (but all rectangles can be modified *a posteriori* in ACACIA's editor). The final analysis results are then stored in a local database for future usage. It is also possible to share the findings with remote researchers over a cloud database.

To wrap up, the main purpose of ACACIA is to record who is interacting with what or whom and tell us the respective location of these entities. This is usually the core of what researchers are interested in and manually searching for when analyzing collaborative interaction (as set out in Sections 1 and 2.1). Concerning customizability, our approach is to support researchers with a tool that is capable of answering these core questions out of the box without the need for training data or building own models. Our motivation is to allow researchers without a technical background to use ACACIA. To tackle cognitive load, visualizations can be selectively displayed (see Figure 5) and the export is categorized to prevent overwhelming numbers of columns in worksheets, accordingly. Apart from the possibility of focusing on the most interesting subset of the overall functionality, we did not include further customization from users' perspective (while developers can readily include other external services or data streams), as most questions of collaboration research manual coders would extract from images can be consequently answered, as we will further exemplify in Sections 5.3 and 5.4.

3.3 AI Services in ACACIA

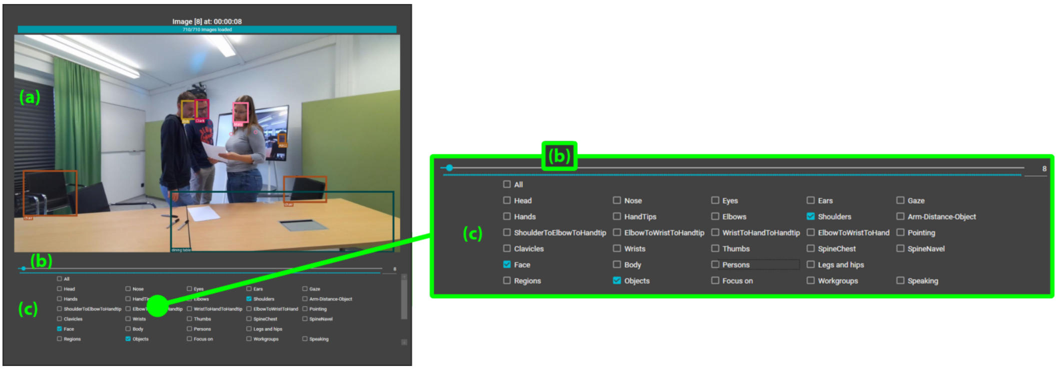
ACACIA uses several readily available as well as more experimental services which are registered as components via \psi. An overview of these external services is provided in Figure 7. We now turn

Table 1. Simplified example of the tool's export about identified entities.

Frame #	Persons	Objects	Regions	Location of all Entities
41	John, James, Jeannie	TV, Desk, Blackboard	Todo-List, Area on Desk	Locations [John=>102, 203, ...]
42	John, James, Jeannie	TV, Desk, Blackboard	Todo-List, Area on Desk	Locations [John=>112, 200, ...]
43	John, James	TV, Desk, Blackboard, Smartphone, Whiteboard	Todo-List, Area on Desk	Locations [John=>125, 205, ...]

Table 2. Simplified example of the tool's export about relations between entities.

Frame #	Person	Gazes at	Points to	Interacts With
41	John	James		James, Jeannie, TV
41	James	John	TV	John, Jeannie, TV
41	Jeannie	TV		John, James, TV



☐ All

☐ Head	☐ Nose	☐ Eyes	☐ Ears	☐ Gaze
☐ Hands	☐ HandTips	☐ Elbows	☑ Shoulders	☐ Arm-Distance-Object
☐ ShoulderToElbowToHandtip	☐ ElbowToWristToHandtip	☐ WristToHandToHandtip	☐ ElbowToWristToHand	☐ Pointing
☐ Clavicles	☐ Wrists	☐ Thumbs	☐ SpineChest	☐ SpineNavel
☑ Face	☐ Body	☐ Persons	☐ Legs and hips	
☐ Regions	☑ Objects	☐ Focus on	☐ Workgroups	☐ Speaking

Fig. 5. Left: main area of ACACIA's viewer. Right: call-out of frame control bar (top) and visualization options (bottom). (a) marks the visualization pane showing the selected visualizations on the current frame, (b) the frame control bar allows navigating the video frames, and (c) shows a list of visualization options to choose from. Please note that some options (e.g., Speaking or Workgroups) are experimental and have not been evaluated, yet.

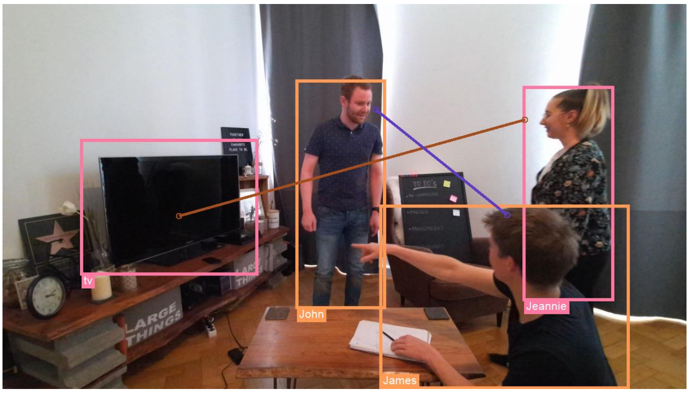


Fig. 6. Automatically generated example image (frame #41 from Table 1 and Table 2) with object detection, person identification, and gaze paths. With depth information, it is possible to intersect gaze paths with people, objects, or regions in the 3-dimensional room.

to discussing the AI services used to answer typical questions during the coding of collaborative scenarios.

3.3.1 *Entity Recognition.*

In this section, we describe the procedures conducted with the aim of detecting and contextualizing entities such as regions, objects, and people in video frames. Most of these activities happen during the pre-processing phase. **Questions: Who or what is there, and where are they?**

Person Detection and Identification (Recognition). **Question: Are there people, where are they, and who are they?** For the analysis of collaborative interaction, one of the most important pieces of information is the identification of the people participating in such a setting. Furthermore, we would like to detect if a person is interacting with any other people, regions, or objects in order to characterize the overall collaboration. Our approach mainly consists of three steps and employs

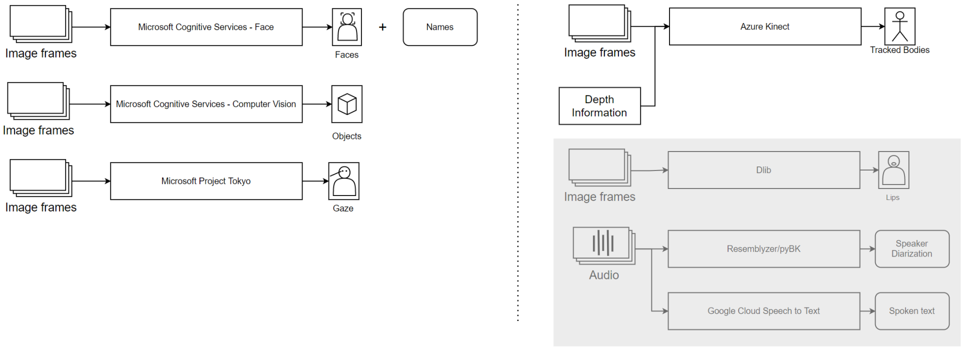


Fig. 7. Overview of external services used in our approach. Please note that Lip and Speech detection have not been systematically evaluated, yet, and should be seen as a suggestion for future work.

Microsoft Azure Cognitive Services[3], which is "a set of RESTful services capable of recognizing, understanding, and interpreting the content of pictures, speeches, live videos, written text, and much more, with a natural language description" [10].

First, all the participants must be registered in ACACIA via a name and a profile image, which is needed to detect people and to correctly assign a face to a person name (i.e., identification). In our tests, we usually took these profile images directly from the recorded RGB frames. From our experiences, the requirements on profile image picture quality in terms of resolution are rather low. In our system evaluation described in Section 4, we recorded RGB with a resolution of 1080p (i.e., 1920 x 1080 pixels) but took smaller segments of the images which depicted people's faces (mostly around 200 x 200 pixels). However, we achieved best results with front-facing imagery. Cognitive Services Face API's documentation provides some further guidelines and hints concerning the quality needed to detect[4] and recognize[5] faces. For example, for detecting faces the resolution must be between 36 x 36 and 4096 x 4096 pixels (depending also on the proportion between face resolution and overall image resolution). For recognition (i.e., identification), some limiting factors, such as "Extreme facial expressions" or "Obstructions that block one or both eyes" are also raised there.

After the profile images are registered in ACACIA, we perform face detection on individual frames of the video. After successfully detecting people, the final step is to identify them by using their profile image. Both tasks use Cognitive Services: we use the Face API for face detection and face recognition and receive the location and size of the identified faces from this service. The results are then visualized in ACACIA's Viewer, as can be seen in Figure 6 (left, rectangles around people with name printed below), as well as exported as annotated images and table data (see Table 1 in columns people and Location of all Entities).

Region and Object Detection. **Question: Which objects and regions are there and where are they?** In addition to the identification of the participating people, the recognition of objects and the consideration of regions are also central aspects in the analysis of collaboration. Object detection can be accomplished in either an automated manner with computer vision, or through

[3] https://azure.microsoft.com/en-us/services/cognitive-services/, last access February 28th, 2022
[4] https://docs.microsoft.com/en-us/azure/cognitive-services/face/concepts/face-detection, last access May 13th, 2022
[5] https://docs.microsoft.com/en-us/azure/cognitive-services/face/concepts/face-recognition, last access May 13th, 2022

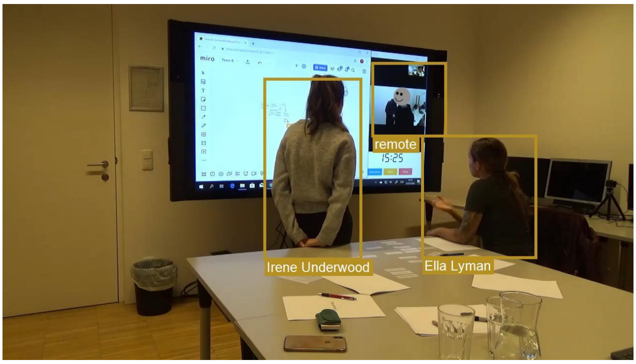


Fig. 8. Visualization of person detection and identification in ACACIA. Identified people receive a frame holding their name. Regions of interest also receive a frame and their name. The frame denoted with "remote" is such a region and marks the area where the camera image of the remotes can be seen. Color coding conveys where the individual people are currently focusing, that is, where their gaze is apparently directed (not visualized here). In this case, both co-located people focus on the remote participant, effectively gathering all participants' attention in the person space [9].

manually adding objects or regions to the database that was established in the pre-processing step (see Figure 4).

In the automated variant, each frame of a video is analyzed in ACACIA by the Computer Vision service of Azure Cognitive Services. ACACIA sends the images to the service, which returns a list with all detected objects with a corresponding name and position. After automatic detection, researchers can adapt detection of objects and introduce regions manually if necessary. Regions are objects or areas in frames which are of particular interest to the observers. Regions—although technically identical—are conceptually different from objects such as a table, a chair, or a laptop. In general, it is not possible to automatically detect regions because they are individual and specific to a collaboration scenario. For example, people may be sorting items on a segment of a table, while the rest of the table is reserved for other use. Another example can be seen in Figure 8, where the local participants are currently focusing on the particular area of a wall-sized display (which is also detected but not visualized in this example) showing the remote collaborators' camera image. Figure 6 (left) and Figure 9 further show examples of automatically detected objects (e.g., the TV screen).

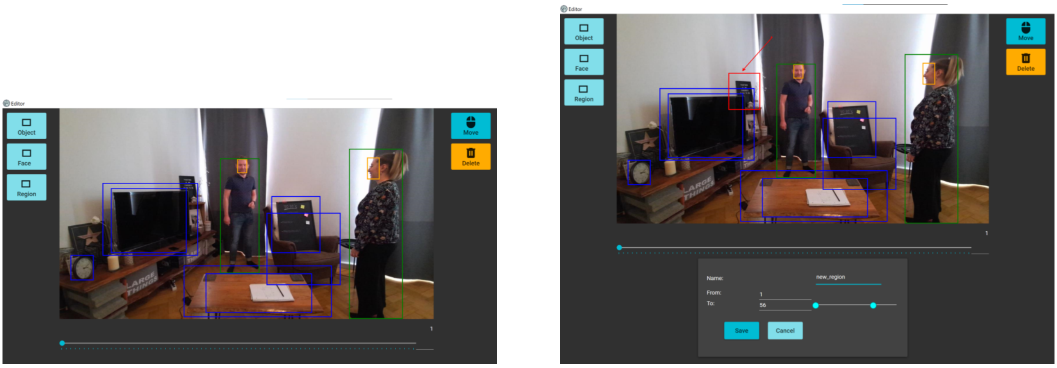
To facilitate flexibility and individualization, users can manually annotate images with regions in ACACIA's editor (see Figure 4 and Figure 9). Since regions are typically more stationary than objects, this should usually only take a few minutes of researchers' time. Additionally, it is possible to add other kinds of objects and adjust objects that have already been detected automatically. Objects and regions can be set active for a specific set of frames (between one and all frames of a video). All adjustments are stored in the local database and used for later post-processing, visualization (see, e.g., Figure 6 (left, rectangle around the TV)) as well as the statistics export function (see Table 1 in columns Objects, Regions, and Location of all Entities).

3.3.2 Establishing Relationships between Entities. As soon as all related objects, regions, and people have been detected and located, relationships between the entities can be sought. Most of the activities required for this establishment of relationships rely on our own algorithms and are conducted in the post-processing phase. **Questions: *Who* is looking at someone or something?**

Person Tracking. **Question: Is the person in this frame the same as the person in that frame?** Under most conditions, we are able to detect and recognize people reliably. However, since people can move around, potentially leading to the situation that a person's face is not visible for the camera at a certain point in time, or a second person is temporarily covering the first one, we decided to track people over time to alleviate tracking losses (and therefore establishing relationships between potentially unidentified people across different frames). Thus, the results collected for each frame are used to track people over time to prevent losing potentially important information. More concretely, our algorithm tries to bind identified faces to the according bodies (which are also detected in pre-processing). Should a face be suddenly not detected anymore (e.g., because they are currently facing away), ACACIA takes a look at nearby frames and searches for exactly this face and body combination. If such a combination is found, and the difference between the physical locations is below a configurable threshold per frame, we can fill in the detection gaps. The location difference is considered in order to disambiguate the face-to-body connections, because it is not plausible for humans to move faster than a certain threshold speed (we used a slow walking speed of around 1 m/s which should be appropriate in most lab study settings). Accordingly, when neither face nor body are detected, the algorithm fails to track people.

Gaze Detection. **Question: Who or what is a person currently looking at?** In addition to identifying persons and keeping track of their position, it is also fundamental to identify where a person is looking at. For the analysis of collaboration, we planned to detect if a person is currently looking at another person, region (e.g., the task space where the main collaborative work is happening, or the remote person space where we look at when talking to the remotes [9]), or object (e.g., a tablet, or monitor)—and if so—at which person, region, or object. This is essential since we can derive where the different people's focus is in each frame. To detect the gaze direction of people we use a model from Microsoft's Project Tokyo[6]'s framework (short: Tokyo), which is an endeavor to support blind or low-vision people by finding out "how [...] agent technologies [can] amplify existing skills and abilities to help people do more". Among many other results, the project led to the creation of an AI framework particularly capable of inferring pose and gaze from two-dimensional imagery [29]. We use Tokyo for such gaze and pose detection as an external service, embedded into

[6] https://www.microsoft.com/en-us/research/project/project-tokyo/, last access June 17th, 2021



Name: new_region
From: 1
To: 56

Fig. 9. ACACIA's editor giving researchers control over the results of the automated detection process. Objects, regions, and faces (for person identification) can be manually added, moved, and deleted for inclusion in the subsequent analyses. First, a frame is selected (left), then the changes can be done (right) as in this example, where a new region was added by drawing a rectangle (red rectangle with red arrow).

a \psi component. Tokyo not only gives us information about the gaze direction, it also gives us information about the poses of people in a video. This information helps us to perform simple body tracking in videos without depth information as a fallback solution. Moreover, the information of the pose is used to assign gaze to the correct person. Since it is already possible to detect and identify the faces of the people by the aforementioned algorithms, we can assign the results which we receive from Tokyo to the correct person. With the information of gaze, we can detect where a person is looking. Therefore, we applied different algorithms to estimate which object, region, or person the person might look at. Since the detection of where a person is looking at in a simple video file can be ambiguous without 3D information and may lead to inaccurate results, we extended our algorithms, insofar as in cases where the 3D information (recorded by Azure Kinect) is available, we combine this information with the gaze, to be more precise when detecting where the person is looking at in the room. The gaze vector is generally determined with Tokyo using the pose of a person. This vector is then extended in the 3D depth image and an attempt is made to check whether an intersection with the point cloud of the depth image is found in the direction of the gaze. If such an intersection is found, the gaze beam ends in the 3D space and the nearest entities (objects, persons, or regions) can be considered possible candidates as gaze targets. This leads to better results, especially if multiple target areas are arranged one after another or if people, regions, and objects are located side by side. However, using an Azure Kinect for recording a video including depth information–although recommended–is not a precondition to using ACACIA–we only use this information if it is available. However, when no 3D information is available, all gaze paths do not end inside the image, leading to potentially several gaze target sequential candidates and consequential ambiguity. Again, gaze paths are visualized in the Viewer and can be exported as annotated images (see, e.g., Figure 6, left, John's gaze is visualized with a purple line intersecting and stopping at James in the 3-dimensional room, Jeannie's gaze with a brown line stopping at the TV). Furthermore, the information is present in the exported spreadsheet data (see Table 2 in columns Person and Gazes at).

4 EVALUATION OF ACACIA

After earlier tests with our system in a fabricated setting (see, e.g., Figure 10) which yielded rather promising results, we wanted to evaluate ACACIA in a less controlled (i.e., a more realistic) study setting in a small-scale system evaluation involving users in a realistic collaboration scenario. Our goal was to assess the current performance of the system and its components and find out situations in which recognition works reliably and where further improvements are needed. Generally speaking, the aim was to answer our research question if **we are there yet** already, when it comes to (partly) automating analysis of collaborative interaction with off-the-shelf (or at least currently available) services. A second aim was to uncover current services' potential to be integrated into the research pipeline as a "machine-in-the-loop" inversion of the "human-in-the-loop" concept. In terms of experimental design, the capability of ACACIA to correctly detect and recognize entities and their relationships can be seen as our dependent variable. It is measured by comparing the system output concerning certain collaborative interaction situations to the impression of human coders. We regard mainly two independent variables as decisive: the quality of the recorded data (i.e., RGB imagery and depth information as well as perspectives, angles, etc.) and performance of the external services (which were used in combination).

4.1 Participants & Apparatus

For this purpose, we recruited five participants among our university staff, who were not among the authors (two professors and three research associates, 3f, 2m, mean age 33.6 years, SD=6.98, min=25, max=41). They gave informed consent, contributed voluntarily, and were not compensated with extrinsic incentives for their time. We gathered two participants in one room and connected them with the remaining three participants in another room by an audio-video link on two identical Microsoft Surface Hub 2S devices. To reflect a minimal possible study setup, each room was equipped with one Azure Kinect device. For reporting, the participants' names were changed to Alice and Holly (room 1) as well as Bob, Clark, and Mary (room 2). They were asked to solve several logic puzzle tasks[7] and could make use of an interactive whiteboard application.

4.2 Procedure

To allow for person identification, we registered all five participants in ACACIA's user profiles using pictures taken directly from the study material. For pre and post processing, ACACIA's computation time was around ten minutes for a one-minute video segment with one of our current PCs (Windows 10 64 bit, Intel Core i7-8850H CPU @ 2.60 GHz with 64 GB of RAM; main frame rate of 1 frame per second). Please note that bandwidth as well as throughput of external services affects the processing time. Currently, Face API allows 30,000 calls and Computer Vision 5,000 calls for free per month. Per single frame, ACACIA in our study performed 1 Computer Vision and 5-6 Face API calls (which is dependent on the number of people depicted in the frames).

Overall, we gathered around 60 minutes of collaborative interaction data (30 minutes per room) and discussed in a team of researchers which situations are most representative for illustrating ACACIA's current performance. Furthermore, after processing the data in ACACIA, two of the authors spent around 90 minutes to collaboratively judge the results in a first pass and manually added around 20 faces in total to the complete material with ACACIA's editor. This was necessary because of the extended periods of time some participants were completely hidden and our tracking algorithm was only able to fill in the gaps in such extreme cases with a little manual assistance. Then we evaluated the results and compared ACACIA's output to a human coder's impression. We followed an approach suggested by Saldaña [37] (combining advantages of "coding solo" and "team coding") with one researcher mainly responsible for the judgments who discussed ambiguous cases in several sessions with the remaining team of researchers in order to arrive at a shared understanding of what human coders would most probably have coded for each frame. The findings are presented below.

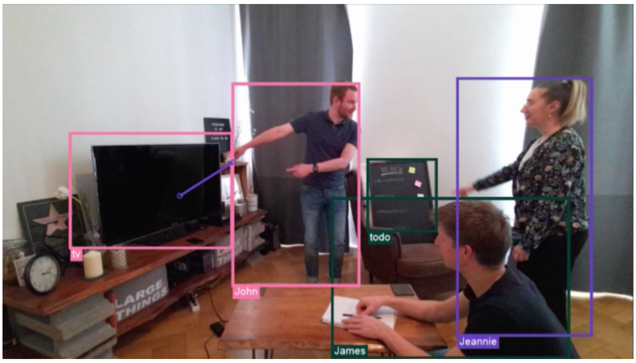


Fig. 10. Example of one scenario recorded with Azure Kinect (RGB and depth information). The pointing gesture of John's right hand at the TV is marked with a purple line. Please note that the line stops in the 3D space where the (object or region) "TV" is situated.

[7] https://logic.puzzlebaron.com, last access February 9th, 2022

4.3 Recognizing Entities

First we discuss to which extent we can already answer the following questions:

(1) Which people are in a room, and where are they?
(2) Which objects are in the room, and where are they?
(3) With manual annotation we can determine (according to the current research interest): Which regions of interest are in the room, and where are they?

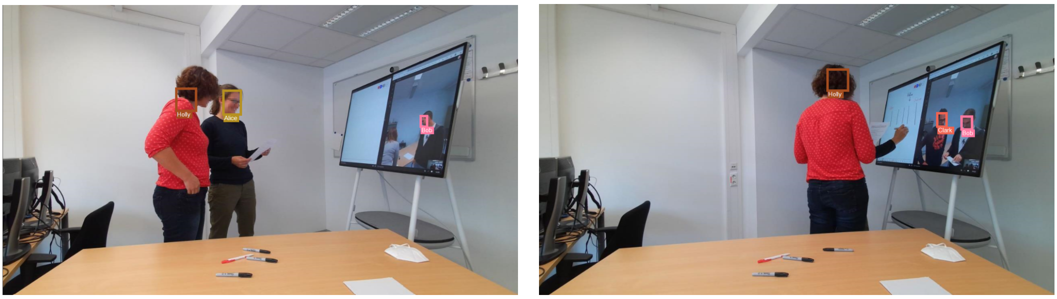


Fig. 11. Person identification: Two illustrative examples of our small-scale evaluation. Left: All co-located participants (Holly & Alice) are recognized and identified, one of two remotes is identified (Bob). Right: Only one co-located is identified (the other, Alice, is hidden behind Holly), both currently visible remotes (Clark & Bob) are identified.

Which people are in a room, and where are they? In our first example, we evaluated a 3.5 minute segment of one of our puzzle tasks in the room with the interacting dyad, starting from the task's beginning to mid-way of solving the puzzle. During the initial 40 seconds (1 fps) both co-located participants (Holly & Alice) were recognized correctly. Then, in frame #41, Alice is hidden behind Holly and she is not recognized anymore (see Figure 11, right). As this happened quite frequently in our setup, Alice is only recognized on 93 of the 211 frames (44.08%), while Holly in the foreground (i.e., without obstacles between the camera and her) is recognized in all 211 frames (100%). This leads to the rather disillusioning percentage of 44% for the question if all participants in the room were captured and identified correctly. However, we also saw several positive aspects.

Firstly, although Holly was mostly facing away from the camera, she was identified all the time thanks to our tracking algorithm discussed in Section 3.3.2. We could be so bold to say that as long as the camera had clear sight of the participants, the recognition and identification rate was near 100%. For (partly) hidden participants, our tracking algorithm tries to fill in the gaps but some manual fine tuning (i.e., occasionally adding missing entities in ACACIA's editor) might be necessary for persons who are hidden for an extended period of time.

Secondly, also the remotes were recognized to some extent which might be very interesting for future enhancements concerning targeted audio, etc. Therefore, we evaluated how often all participants—both co-located and remote—were correctly recognized and identified by ACACIA, as long as their head was visible (but not necessarily their face). The rationale behind this is that a human coder presented with a still image could in most situations identify remote depictions by seeing their head. Furthermore, they are likely unable to estimate where currently invisible persons are located with certainty or they can at least not be sure what they are currently doing (e.g., focusing on something)—unless they see (a larger part of) their head or face.

On each of the 211 frames, either 3, 4, or 5 persons (i.e., their faces or heads) are currently visible (see Table 3). We judged how many of them the system correctly identified. For instance, in

Table 3. Identification (implying recognition) rate of both in-room and remote persons in one representative example 3.5 minute video segment. Overall, 61.47% of the (co-located and remote) participants were successfully identified.

# of Identified Persons	# of Frames	Identification Percentage	Product
1 of 4	1	25.00%	0.25
1 of 5	17	20.00%	3.4
2 of 3	6	66.67%	4
2 of 4	27	50.00%	13.5
2 of 5	33	40.00%	13.2
3 of 3	3	100.00%	3
3 of 4	53	75.00%	39.75
3 of 5	42	60.00%	25.2
4 of 4	21	100.00%	21
4 of 5	8	80.00%	6.4
Sum:	**211**		**129.7**

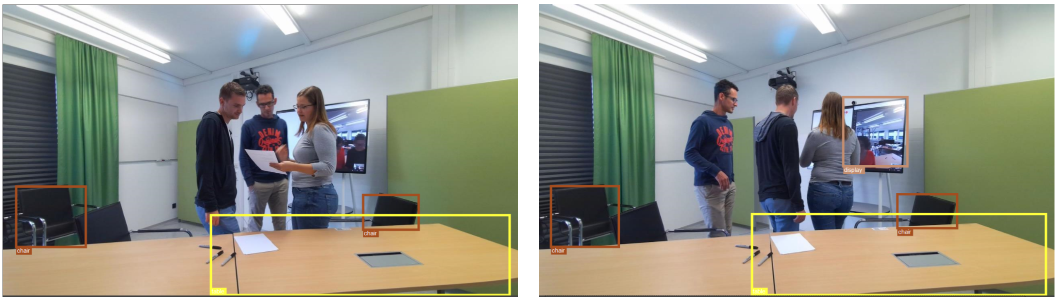


Fig. 12. Object recognition: On the left, currently three objects have been recognized, two chairs (brown) and one table (yellow). On the right, the portion of the Surface Hub's screen showing the remote participants was additionally recognized as a "display".

Figure 11 left, currently 3 out of 4 people are correctly identified (remote Mary is facing away and currently not recognized). On the right, currently 3 out of 3 are correctly identified (because we only see one ghost arm of Alice but not her face which is, however, instrumental for identification). Consequently, all frames received a percentage rating (e.g., 75% if 3 out of 4 were recognized, 60% if 3 out of 5 were recognized and so on). We added all of these percentages and found out that 61.47% of all participants, whether remote or co-located, were recognized and identified correctly as long as their faces were visible. So just by using one camera headed towards the screen depicting the remotes, we know for more than 60% who is there and where they (i.e., their virtual representations) are. As for the remotes our algorithm could only rarely resort to pinning the faces to their bodies (bodies were visible only occasionally), we deem this result as a good starting point for future continuous improvements.

Which objects are in the room, and where are they? The detection of objects in our evaluation showed two main issues. Firstly, the recognized objects were mostly too broad (e.g., a whole table area instead of sheets or pens on that table) and unspecific (e.g., ‘display‘) to be used as targets of interest out of the box. This makes human intervention necessary in most cases. Secondly, recognition was too volatile to be used reliably for establishing relationships between entities (e.g., the question which person focuses on which object). Still, we do not see these as major issues, because regions of interest can be manually added to the frames which is possible quickly (usually a matter of minutes)–solving both issues. Moreover, this allows a more focused approach towards the present research interest instead of relying on generic objects. To illustrate how well object recognition performs, we took a representative 2-minute sample (1 fps resulting in 120 frames) of our study in the room with the triad, because more potential objects were visible there (see Figure 12). However, only the two chairs to the left and to the right of the participants were recognized with 100% accuracy. The table in the front was either recognized as a "table" (52) or "dining table" (29) in 81 of the frames (67.5%). The area of the video call was recognized as either "television" (19) or "display" (18) in 37 of the frames (30.8%). Furthermore, there was not one coherent block of 81 or 37 frames in which these objects were recognized but there were many interruptions instead. Sporadically, other objects were recognized, for example the "chair" between the two outer chairs on 3 separate frames plus one frame as "seating" and the trousers of one participant on one frame (as "jeans").

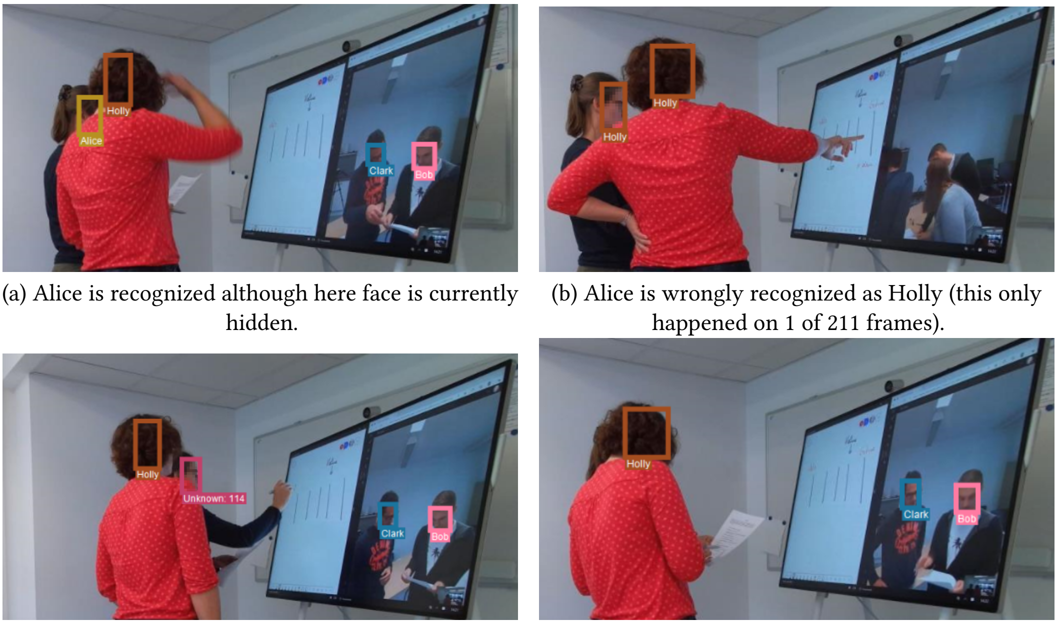
Which regions of interest are in the room, and where are they? This question is of course a bit misleading, because regions are added through manual annotation. In Section 4.4 we address how regions can be integrated into the research process. Taken together with persons' localization which is a byproduct of detection and identification, we can, therefore, say for each identified person where they currently were, relative to manually added regions (or zones) and other identified people. For example, this allows conclusions about proxemics (and could as a generalization help with the analysis of team sports, or military and police operations). We did not separately evaluate this because our findings about identification rate in combination with efforts of manual annotation can answer how well this works.

4.4 Establishing Relationships

As we expected, the questions requiring more reasoning are relatively more difficult to answer for our system:

(1) Is the person in this frame the same as the person in that frame?
(2) Who or what is a person currently looking at?

Regarding the **first question**, we already showed above that ACACIA can achieve up to 100% of correct person tracking as long as their face was once visible and then the person is not completely hidden again (as ACACIA tries to bind identified faces to their body). In our tests we mostly lost track of persons only when they were currently not visible (see Figure 13, (d)). The only exceptions are four (1.9%) misrecognitions of Alice although she or her face were somewhat visible (see Figure 13, (b) and (c)). Please note that in Figure 13, (b), Alice is wrongly recognized as Holly, although Holly is already recognized in this frame. Because one person cannot be in a frame multiple times, our system should instead report Alice as 'Unknown' (but a plausibility check was missing because in our previous tests such a case never occurred). Recognition worked reliably in cases where persons were currently not facing the camera and only the back of their head was visible (exemplified by Holly's recognition rate of 100%, also see Figure 13). In our evaluation, we only used one camera. Therefore, it can be assumed that in typical study situations, rooms do not need to be equipped with dozens of cameras from all angles, or participants need to stay within artificially defined boundaries, but rather they can move naturally and freely around the room. However, we suggest using two or three cameras per room as this would prevent people occluding other people which typically results in tracking losses. To further illustrate the problems which can arise when a person is (partly) hidden, please refer to Figure 13.



(a) Alice is recognized although here face is currently hidden.

(b) Alice is wrongly recognized as Holly (this only happened on 1 of 211 frames).

(c) Alice is wrongly recognized as an unknown person while she is partly visible (this only happened on 3 of 211 frames).

(d) Alice is not recognized because she is hidden (this happened frequently in our setup: in 114 of 211 frames).

Fig. 13. Different issues and their frequentness concerning person identification arising from a camera's blocked line of sight in one example 3.5 minute situation of our evaluation.

Summing up, we can say that as long as a camera has clear sight to a person which was identified before, we have a reliable identification, even when the person is presently facing away or partly hidden. Still, our current tracking algorithm may fail when a person is hidden too long (then it might be necessary to manually add a few faces in the editor).

To evaluate our **question two** about gaze, we again resorted to the example we used for person detection and identification (2 persons in one room, 211 frames). Again, we manually judged if each person actually looked at the area the system thought they looked at. As shown above, object recognition was too volatile to be used for the target areas of the gaze, so we invested a few minutes to add several regions of interest: a printed-out worksheet, the area of the Surface Hub showing the whiteboard application, and the area of the Surface Hub showing the remote participants (see Figure 14). Besides the regions of interest, the list of possible gaze targets included persons and blank spaces that were neither regions nor persons.

Above, we have shown that Alice was correctly identified on 93 frames, while Holly was identified on all 211 frames. Out of the 93 frames, Alice's calculated gaze was presumably on the correct region (or person) on 56 of the 93 frames she was recognized in (60.2%). Holly's gaze target was correct on 71 frames (33.7%). The number for Holly is so low, because she mostly faced away and so the estimate for her gaze was less precise. Although Alice was often not detected because she was hidden behind Holly, at least she was more often headed towards the camera. Furthermore, we analyzed if the system-suggested target area was simply wrong or if only the first priority (i.e., the one area considered as most probable) was incorrect, while the correct target area was farther

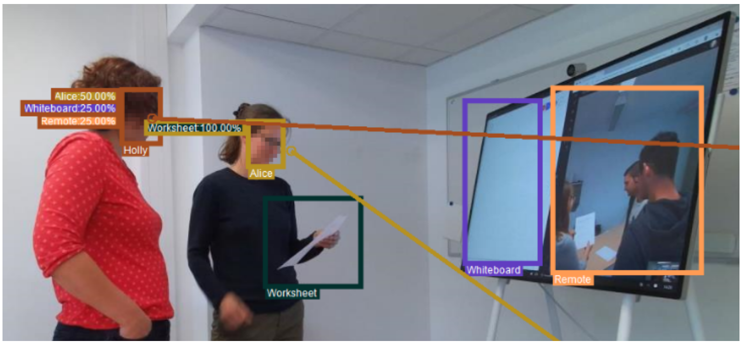


Fig. 14. The manually annotated regions of interest: a mobile worksheet, and two stationary areas on the Surface Hub. Further visualization of automatically detected persons, estimated gaze path, and list of possible gaze target areas. In this frame, Holly's gaze is estimated to target Alice, although she actually looks at the region "Remote" (which is at least in the list of candidates). Meanwhile, Alice's gaze is estimated sufficiently well to pin it to the worksheet.

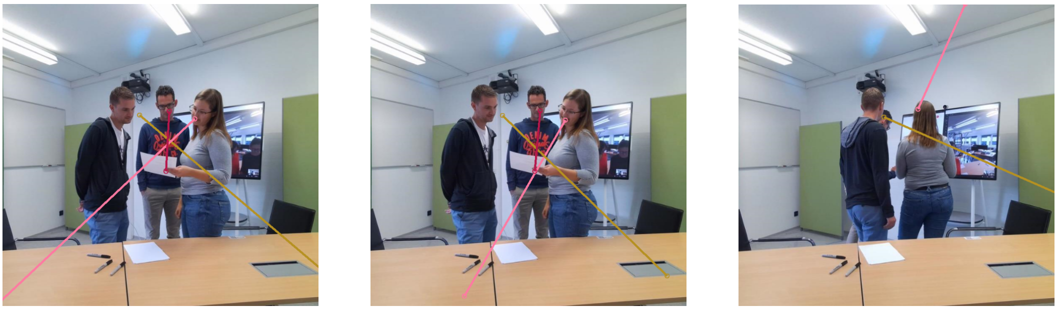


Fig. 15. Estimated gazes of the participants. On the left, only Clark's (person in the center) gaze can be located in the 3D space. In the center image, all three can be located (not necessarily correctly). On the right, we see a typical example when participants are hidden or facing away.

down the list of possible candidates (see Figure 14). The number of frames where the correct area is anywhere in the list of possible areas is 60 for Alice (64.5%) and 124 for Holly (58.8%).

Summary. The accuracy of gaze estimation is arguably not sufficiently reliable to be used right away but in our impression shows that there is real potential. During our evaluation, there was a serious issue that often we did not receive a calculated end point of the gaze beam in the 3D space (see Figure 15). However, we need such an end point to determine which of potentially several sequential target candidates is the most probable. Presently, our algorithm considers the first target a gaze beam crosses as the 'most probable' one in such cases without an end point. We identified a number of possible reasons why the percentage of non-3D-gaze in our evaluation was so high: either i) Azure Kinect was unable to detect the point cloud, ii) Tokyo did not recognize the eyes, iii) Tokyo miscalculated the eye direction (or z direction), or iv) ACACIA failed to map Tokyo's information to the Azure 3D space. Our analysis into this showed that reasons i) through iii) are the most likely ones because without changes to ACACIA, on one frame (e.g., the center image in Figure 15) we have gaze beams' end points but not so on another (e.g., the left image in Figure 15).

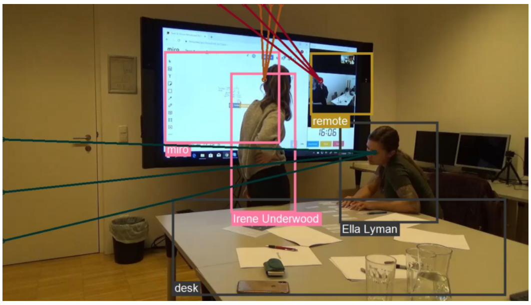


Fig. 16. Example image taken from an earlier user study, only 2D images without depth information were used as data source.

However, we would like to hold that all components are powerful tools but are certainly dependent on the material they receive.

5 DISCUSSION

In this section, we first discuss the performance of ACACIA and then go on to describe how we envision our automatically extracted data may assist with the analysis of study data using example models and frameworks from prior research (also see Section 2.1). There, we focus on the analysis of collaborative interaction but would like to hold that—from a generalization standpoint—any form of analysis that is interested in (identified) entities' location and/or relations to other entities could also profit from a system like ACACIA.

5.1 System Performance

Overall, both the recognition and identification of persons worked remarkably well as long as the camera had clear sight of the persons. Object identification proved to be too unspecific and volatile to be used for establishing connections between entities which is why we used the manual annotation of regions as a fallback solution—which, however, is usually possible within minutes. For such frames, where persons were correctly identified and not facing away from the camera, gaze estimation also worked to some extent (between approx. 33.7% if face orientation was suboptimal and 60.2% for persons more often—but still not permanently—facing the camera). We learned that an introduction of a second or maybe third camera is likely to improve the result by far. However, a synchronization mechanism still has to be developed to decide which camera's input should be considered based on estimates of the probability that an inference is correct or not (i.e., a robust confidence rating).

Although using Azure Kinect is optional in ACACIA, we strongly recommend it to improve gaze estimation. However, we discovered that the model from Project Tokyo can infer reasonable pose and gaze information from 2-dimensional video footage even with rather low resolution. For example, on video footage of an earlier study, the image of remote collaborators is visible on a display on an area of approximately 150 by 150 pixels in our video. Still, the Project Tokyo model was able to infer a gaze path, which even looked reasonably directed (see Figure 16, red gaze cone). Of course, this gaze path is not totally accurate because of offsets between the camera position and where the picture is presented but, nevertheless, led us to further considerations around future workspace awareness support mechanisms based on remotes' gaze (cf. [46]). Overall, our impression is that Project Tokyo was created with continuously updating pose and gaze information in mind, decreasing the requirements for precision in a single isolated frame because of this constant approximation. When applied to images taken from a video with a frame rate of one frame per second, the requirements for precision are relatively high, and smoothing algorithms often lack a higher number of interpolation points. For the future, we expect that ambiguities in gaze detection will decrease as external services improve their precision (several updates already lead to notable improvements during our development period) and increased computation power allows for more rapid processing of higher frame rate recordings. In summary, we think a combination of high-resolution 2D imagery, depth information, and a front-facing camera angle is best for reliable gaze estimation.

5.2 Privacy

ACACIA is capable of not only identifying persons but also inferring what they are currently doing and with whom they are interacting. Kranzberg's first law states that "Technology is neither good nor bad; nor is it neutral." [23]. In consequence, it will always be a combination of the two poles and media coverage about misuses of technology such as facial recognition—be it by authoritarian governments or encroaching enterprises—is widespread[8]. Concerning the future of such technology, we have to seriously pay attention to enable people to use such systems responsibly. Even if such technology is not abused, simply by the fact that images are uploaded to external services there are several dangers and risks involved. One suggestion we propose is to **never** use a person's real name—just as we did—when using profile images for the identification in ACACIA. Commercial providers of AI services could furthermore monitor the fair use of their services and a community-driven rating approach could further act as a safeguard. Finally, it is essential that we aim for voluntariness, properly inform participants about the procedures and ask if they consent. We do not see the future of our approach in the analysis of everyday collaborative work settings (which would imply an even more extensive number of critical privacy challenges) but exclusively for the analysis of scientific studies in the HCI environment.

5.3 Proxemics

Proxemics is an aspect of collaboration that is indicative of many other more concrete forms and elements of interaction, such as the ones we will discuss in the following Section 5.4. For instance, information on proxemics can help to identify presence/activity in a certain territory (see Section 5.4) or provide information on closeness of collaboration. For example, EagleView suggested by Brudy et al. [7] specializes on proxemics and provides exhaustive tools to support an entire research process in this regard (while our scope is more general). EagleView uses a Kinect v2 device to track people ("fixed objects" can be manually created) and we use an Azure Kinect but also use its RGB images to detect people, regions, and objects, and identify people. Analysis of proxemics in our system, conceptually similar to what has been suggested by Brudy et al. [7] but with a less complex setup works with relatively high reliability in ACACIA.

Current Status: High recognition and identification rate of persons when they are in clear sight of the camera (up to 100%). Regions can be annotated in a matter of minutes. Related cues can be derived from identification and localization of people and regions. Object detection was not as reliable as we hoped for. High precision of the detected entities' location in a frame.

Process suggestion in ACACIA: **(i)** Manually add regions of interest or missing objects in the editor if necessary, and **(ii)** run post-processing to obtain proximity information.

[8] https://www.theguardian.com/technology/facial-recognition, last access May 4th, 2022

5.4 Territoriality & Media Spaces

Broadly speaking, analyzing for territoriality and media spaces means that we are interested in who interacts in which territories or spaces. We envision that investigating this will be quite straightforward with ACACIA. Regions of interest can be manually created in our editor to represent the different personal, group, and storage territories on the one hand, or task, person, and reference space on the other hand present in a collaborative setting. As some areas (e.g., storage territory, person space) are more mobile than other ones (e.g., group territories) [41, p. 300], a slightly higher effort might be necessary to set up all regions of interest and keep track of their movement over the duration of the videos. We showed an example of how this is possible by creating a mobile "Worksheet" region during our evaluation (see Figure 14). Additionally, areas in proximity to participants are likely to be used as their personal territories [41, p. 301]. ACACIA may facilitate larger-scale user studies with a quantitative focus on **who interacts in what kind of territories or media spaces over time**.

Current status: Analyzing for territoriality or media spaces makes it necessary that relationships between entities are established. So far, we evaluated gaze as a possible way to estimate such relationships. Depending on camera angles of the depth camera, gaze can be estimated correctly in between one third and two thirds of the cases. This can on the higher end serve as a first rough estimate but needs further improvement.

Process suggestion in ACACIA: **(i)** Manually add regions as either territories or media spaces in the editor, **(ii)** run post-processing to obtain gaze, proximity, and touching (to be added) information, and **(iii)** evaluate quantitative distribution based on the statistics exports or qualitative insights in the Viewer.

6 CONCLUSION AND FUTURE WORK

In this paper, we suggested a process and approach to integrate semi-automated analysis of collaborative interaction into the research pipeline. We believe that semi-automated analysis steps have great potential to enhance and support HCI and CSCW research processes (with a specific focus on analysis of collaborative behavior and interaction) for several reasons (also see Section 1). First, systematic manual coding of collaborative interaction (which is, however, necessary if fine-grained insights in the nature of collaboration should be gained), is exceptionally time-consuming and labor-intensive for the researchers involved. Second, this might lead to consequential problems potentially critical for the quality of the coding result, such as loss of information and reliability due to human fatigue or inter-rater inconsistencies.

To answer our question, if we have already arrived at automating (parts of) the analysis process, judging from our systematic evaluation, we estimate that we are half-way there. Our evaluation showed that the biggest remaining issue is how to achieve optimal camera perspectives. Firstly, the tracking of hidden persons is frequently lost, albeit all attempts to alleviate this by pinning faces (which are identified) to bodies. Secondly, there is the dilemma that gaze estimation works best when persons are facing the camera which in turn limits a single camera's capabilities of judging the target area the gaze lands at. For example, a vertical display is best captured from behind a person operating it which is of course worst for gaze estimation. One solution is to use multiple cameras and we suggest future endeavors to create synchronization mechanisms for this which require robust confidence ratings for each instance.

These considerations also show that controlled lab studies are a more favorable first candidate for real-world application of such a system as opposed to field studies in classrooms or museums where larger crowds can be expected.

When we will arrive at this point in the near future, we do not, however, suggest to completely replace the human efforts involved in collaboration analysis processes by intelligent systems. On the one hand, this might seem to be a logical consequence of the above arguments, but on the other hand, it might also introduce further and potentially highly critical problems, e.g., due to biases built into these systems [33]. Further, many cases of collaboration might be rare or even unique and thus difficult or impossible to interpret correctly by data-driven artificial intelligence. Thus, we argue for interweaving human as well as machine efforts in the research process related to analysis of collaborative interaction.

The main contributions of our work are, on the one hand, the establishment of a semi-automated research pipeline, and on the other, the introduction and evaluation of our prototype system ACACIA, which provides a proof-of-concept for the underlying ideas. The specific results and lessons presented in this paper should, thus, be understood as exemplars which are highly dependent on the quality of the integrated hardware and software technology and services (e.g., the quality of semi-automatic analysis of video imagery depends on the quality of the camera which provided the recordings as well as the continuous visibility of entities to this camera). However, we believe that our findings can provide a good perspective and guideline to other researchers struggling with the known issues related to purely manual coding.

Future work is oriented alongside two different streams. The first stream will deal with synchronizing several capturing devices in the same room (to improve recognition rate by optimizing camera angles), as well as in different rooms (to facilitate automated capturing of remote and hybrid settings). Afterwards, in the second stream we plan to explore the system performance further in upcoming larger-scale user studies in different settings in the HCI domain: with varying participant numbers, different tasks and camera setups.

Besides gaze for establishing relationships between entities, we also investigated pointing which could be used in parallel for further disambiguation. Speech detection and speaker diarization are planned to further classify persons' activities or infer topics. All of this is, however, not mature enough to be systematically evaluated and needs improvements first.

With respect to the technological (and conceptual) achievements associated with ACACIA, there are also some general limitations in addition to the concrete (and partly technology-dependent) ones discussed in Sections 4 and 5.1. First, at the moment we cannot reliably detect and distinguish between fine-grained elements of a collaboration such as coupling styles (see Section 2.1). This requires either collecting and providing a huge amount of training data to feed to a classification algorithm or establishing a large set of partly interrelated rules that can be used as a basis for the distinction. Second, we found that calibration of the hardware setup is of immense importance. When recording earlier test material, the 3D space sensitive to interaction (which can later be analyzed in form of a point cloud) was not identical to the 2D image stream recorded by the same camera. Thus, some frames lack 3D information about part of the involved people (see Figure 17). We thus plan to provide a systematic setup and calibration guide for future studies. For instance, it might be helpful to add visible markers to the room indicating where in this setup detailed analysis will be possible.

In conclusion, we believe that the work presented in this paper represents an important step towards the facilitation and enhancement of the research process in the domain of collaborative interaction. There is additional potential to utilize our semi-automated analysis approach not only for the post-hoc analysis of collaboration, but also in the future to visualize certain aspects of a collaboration on the fly, to facilitate focused and individualized support of teams or sub-teams in potentially critical situations.

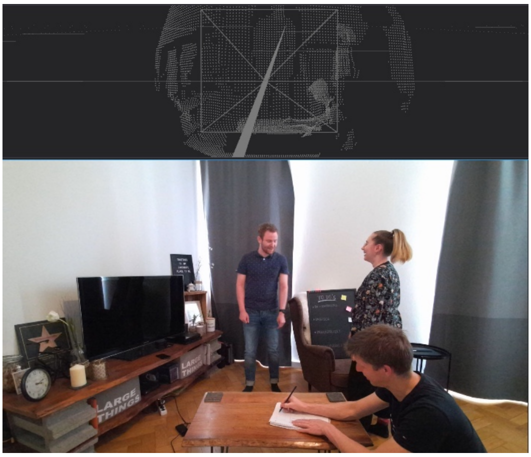


Fig. 17. In one of our earlier tests, we set the Azure Kinect to be more sensitive in the depth of the image which led to a narrower field of view (top) than the device's RGB camera (bottom) would record.

ACKNOWLEDGMENTS

Part of the work reported in this article has been conducted within the scope of the Hybrid Collaboration Spaces (HYCOS) project, funded by the Austrian Science Fund (FWF) [P 34928], and of the project Supporting Hybrid Collaboration for the Teams of Tomorrow, funded by the 2019 Microsoft Productivity Research Collaboration.

REFERENCES

[1] 4A-SIDE GmbH 2019. *act4teams® Handbuch Version 2.3*. 4A-SIDE GmbH, Spielmannstr. 19, 38106 Braunschweig, Germany.

[2] Robert F Bales. 1950. *Interaction process analysis; a method for the study of small groups.* Addison-Wesley, Boston, MA, USA.

[3] David Benavides, Alexander Felfernig, José A Galindo, and Florian Reinfrank. 2013. Automated analysis in feature modelling and product configuration. In *International Conference on Software Reuse*. Springer, Springer, Berlin, Heidelberg, Berlin, Germany, 160–175. https://doi.org/10.1007/978-3-642-38977-1_11

[4] Dan Bohus, Sean Andrist, Ashley Feniello, Nick Saw, Mihai Jalobeanu, Patrick Sweeney, Anne Loomis Thompson, and Eric Horvitz. 2021. Platform for Situated Intelligence. arXiv:2103.15975 [cs.AI]

[5] Elisabeth Brauner, Margarete Boos, and Michaela Kolbe (Eds.). 2018. *The Cambridge Handbook of Group Interaction Analysis*. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781316286302

[6] Frederik Brudy, Joshua Kevin Budiman, Steven Houben, and Nicolai Marquardt. 2018. Investigating the Role of an Overview Device in Multi-Device Collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 300. https://doi.org/10.1145/3173574.3173874

[7] Frederik Brudy, Suppachai Suwanwatcharachat, Wenyu Zhang, Steven Houben, and Nicolai Marquardt. 2018. EagleView: A Video Analysis Tool for Visualising and Querying Spatial Interactions of People and Devices. In *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces (ISS '18)*. Association for Computing Machinery, New York, NY, USA, 61–72. https://doi.org/10.1145/3279778.3279795

[8] Jessica Burggraaff, Jonas Dorn, Marcus D'Souza, Cecily Morrison, Christian P. Kamm, Peter Kontschieder, Prejaas Tewarie, Saskia Steinheimer, Abigail Sellen, Frank Dahlke, Ludwig Kappos, and Bernard Uitdehaag. 2020. Video-Based Pairwise Comparison: Enabling the Development of Automated Rating of Motor Dysfunction in Multiple Sclerosis. *Archives of Physical Medicine and Rehabilitation* 101, 2 (2020), 234–241. https://doi.org/10.1016/j.apmr.2019.07.016

[9] Bill Buxton. 2009. Mediaspace – Meaningspace – Meetingspace. In *Media Space 20 + Years of Mediated Life*, Steve Harrison (Ed.). Springer, London, 217–231. https://doi.org/10.1007/978-1-84882-483-6_13

[10] Alessandro Del Sole. 2018. *Introducing Microsoft Cognitive Services*. Apress, Berkeley, CA, 1–4. https://doi.org/10.1007/978-1-4842-3342-9_1

[11] Manuel Gonzalez-Rivero, Oscar Beijbom, Alberto Rodriguez-Ramirez, Dominic EP Bryant, Anjani Ganase, Yeray Gonzalez-Marrero, Ana Herrera-Reveles, Emma V Kennedy, Catherine JS Kim, Sebastian Lopez-Marcano, et al. 2020. Monitoring of coral reefs using artificial intelligence: A feasible and cost-effective approach. *Remote Sensing* 12, 3 (2020), 489. https://doi.org/10.3390/rs12030489

[12] Germán González and Conor L. Evans. 2019. Biomedical Image Processing with Containers and Deep Learning: An Automated Analysis Pipeline: Data architecture, artificial intelligence, automated processing, containerization, and clusters orchestration ease the transition from data acquisition to insights in medium-to-large datasets. *BioEssays* 41, 6 (2019), 1900004. https://doi.org/10.1002/bies.201900004 Publisher: Wiley Online Library.

[13] Jessica Harris, Maryanne Theobald, Susan Danby, Edward Reynolds, Sean Rintel, and Members of the Transcript Analysis Group (Tag). 2012. 'What's going on here?' The pedagogy of a data analysis session. In *Reshaping Doctoral Education*, Alison Lee and Susan Danby (Eds.). Routledge, London, UK, 109–121.

[14] Yashar D Hezaveh, Laurence Perreault Levasseur, and Philip J Marshall. 2017. Fast automated analysis of strong gravitational lenses with convolutional neural networks. *Nature* 548, 7669 (2017), 555–557. https://doi.org/10.1038/nature23463

[15] Jesse Hoey, Tobias Schröder, Jonathan Morgan, Kimberly B Rogers, Deepak Rishi, and Meiyappan Nagappan. 2018. Artificial intelligence and social simulation: Studying group dynamics on a massive scale. *Small Group Research* 49, 6 (2018), 647–683. https://doi.org/10.1177/1046496418802362

[16] Petra Isenberg, Danyel Fisher, Sharoda A. Paul, Meredith Ringel Morris, Kori Inkpen, and Mary Czerwinski. 2012. Co-located collaborative visual analytics around a tabletop display. *IEEE Transactions on visualization and Computer Graphics* 18, 5 (2012), 689–702. https://doi.org/10.1109/TVCG.2011.287

[17] Gail Jefferson et al. 2004. Glossary of transcript symbols with an introduction. *Pragmatics and beyond new series* 125 (2004), 13–34.

[18] Jelliffe Jeganathan, Ziyad Knio, Yannis Amador, Ting Hai, Arash Khamooshian, Robina Matyal, Kamal R Khabbaz, and Feroze Mahmood. 2017. Artificial intelligence in mitral valve analysis. *Annals of cardiac anaesthesia* 20, 2 (2017), 129.

[19] Stephen Jolly. 2000. Understanding body language: Birdwhistell's theory of kinesics. *Corporate Communications: An International Journal* 5 (Sept. 2000), 133–139. https://doi.org/10.1108/13563280010377518

[20] Simone Kauffeld, Nale Lehmann-Willenbrock, and Annika L. Meinecke. 2018. The Advanced Interaction Analysis for Teams (act4teams) Coding Scheme. In *The Cambridge Handbook of Group Interaction Analysis*, Elisabeth Brauner, Margarete Boos, and Michaela Kolbe (Eds.). Cambridge University Press, Cambridge, 422–431. https://doi.org/10.1017/9781316286302.022

[21] Simone Kauffeld and Annika L. Meinecke. 2018. History of group interaction research. In *The Cambridge handbook of group interaction analysis*. Cambridge University Press, New York, NY, US, 20–42. https://doi.org/10.1017/9781316286302.003

[22] Pawel Korzynski, Michael Haenlein, and Mika Rautiainen. 2021. Impression management techniques in crowdfunding: An analysis of Kickstarter videos using artificial intelligence. *European Management Journal* 39, 5 (2021), 675–684. https://doi.org/10.1016/j.emj.2021.01.001

[23] Melvin Kranzberg. 1986. Technology and History: "Kranzberg's Laws". *Technology and Culture* 27, 3 (1986), 544–560. https://doi.org/10.2307/3105385

[24] Oren Z Kraus, Ben T Grys, Jimmy Ba, Yolanda Chong, Brendan J Frey, Charles Boone, and Brenda J Andrews. 2017. Automated analysis of high-content microscopy data with deep learning. *Molecular systems biology* 13, 4 (2017), 924. https://doi.org/10.15252/msb.20177551

[25] Kenya Kusunose, Akihiro Haga, Takashi Abe, and Masataka Sata. 2019. Utilization of artificial intelligence in echocardiography. *Circulation Journal: Official Journal of the Japanese Circulation Society* 83, 8 (2019), 1623–1629. https://doi.org/10.1253/circj.CJ-19-0420

[26] Curtis Lebaron, Paula Jarzabkowski, Michael Pratt, and Greg Fetzer. 2017. An Introduction to Video Methods in Organizational Research. *Organizational Research Methods* 21 (10 2017). https://doi.org/10.1177/1094428117745649

[27] Mangold International GmbH 2014. *Interact 14 Benutzerhandbuch Version 14.1.4*. Mangold International GmbH, Graf-von-Deym Str. 5, 94424 Arnstorf, Germany.

[28] Lorenza Mondada. 2018. Multiple temporalities of language and body in interaction: Challenges for transcribing multimodality. *Research on Language and Social Interaction* 51, 1 (2018), 85–106.

[29] Cecily Morrison, Edward Cutrell, Martin Grayson, Anja Thieme, Alex Taylor, Geert Roumen, Camilla Longden, Sebastian Tschiatschek, Rita Faia Marques, and Abigail Sellen. 2021. Social Sensemaking with AI: Designing an Open-Ended AI Experience with a Blind Child. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 396, 14 pages. https://doi.org/10.1145/3411764.3445290

[30] Cecily Morrison, Kit Huckvale, Bob Corish, Jonas Dorn, Peter Kontschieder, Kenton O'Hara, ASSESS MS Team, Antonio Criminisi, and Abigail Sellen. 2016. Assessing Multiple Sclerosis with Kinect: Designing Computer Vision Systems for

Real-World Use. *Human-Computer Interaction* 31, 3-4 (January 2016), 191–226. https://doi.org/10.1080/07370024.2015.1093421

[31] Misgana Negassi, Rodrigo Suarez-Ibarrola, Simon Hein, Arkadiusz Miernik, and Alexander Reiterer. 2020. Application of artificial neural networks for automated analysis of cystoscopic images: a review of the current status and future prospects. *World journal of urology* 38 (2020), 2349–2358. Issue 10. https://doi.org/10.1007/s00345-019-03059-0

[32] Thomas Neumayr, Hans-Christian Jetter, Mirjam Augstein, Judith Friedl, and Thomas Luger. 2018. Domino: A Descriptive Framework for Hybrid Collaboration and Coupling Styles in Partially Distributed Teams. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 128. https://doi.org/10.1145/3274397

[33] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. 2020. Bias in Data-Driven Artificial Intelligence Systems – An Introductory Survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), 1–14. https://doi.org/10.1002/widm.1356

[34] M. B. Parten. 1932. Social participation among pre-school children. *The Journal of Abnormal and Social Psychology* 27, 3 (1932), 243–269. https://doi.org/10.1037/h0074524 Place: US Publisher: American Psychological Association.

[35] Rebecca S Portnoff, Sadia Afroz, Greg Durrett, Jonathan K Kummerfeld, Taylor Berg-Kirkpatrick, Damon McCoy, Kirill Levchenko, and Vern Paxson. 2017. Tools for automated analysis of cybercriminal markets. In *Proceedings of the 26th International Conference on World Wide Web*. ACM, New York, NY, USA, 657–666. https://doi.org/10.1145/3038912.3052600

[36] Ramachandran Rajalakshmi, Radhakrishnan Subashini, Ranjit Mohan Anjana, and Viswanathan Mohan. 2018. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. *Eye* 32, 6 (2018), 1138–1144. https://doi.org/10.1038/s41433-018-0064-9

[37] Johnny Saldaña. 2015. *The coding manual for qualitative researchers*. Sage, London, UK.

[38] Carsten C Schermuly and Wolfgang Scholl. 2012. The Discussion Coding System (DCS)—A new instrument for analyzing communication processes. *Communication Methods and Measures* 6, 1 (2012), 12–40. https://doi.org/10.1080/19312458.2011.651346

[39] Carsten C Schermuly and Franziska Schölmerich. 2017. Analyse von Gruppen in Organisationen. In *Handbuch Empirische Organisationsforschung*. Springer Fachmedien Wiesbaden, Wiesbaden, Germany, 491–512. https://doi.org/10.1007/978-3-658-08493-6_18

[40] Ursula Schmidt-Erfurth, Sebastian M Waldstein, Sophie Klimscha, Amir Sadeghipour, Xiaofeng Hu, Bianca S Gerendas, Aaron Osborne, and Hrvoje Bogunović. 2018. Prediction of individual disease conversion in early AMD using artificial intelligence. *Investigative ophthalmology & visual science* 59, 8 (2018), 3199–3208. https://doi.org/10.1167/iovs.18-24106

[41] Stacey D. Scott, M. Sheelagh T. Carpendale, and Kori M. Inkpen. 2004. Territoriality in collaborative tabletop workspaces. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. ACM, New York, NY, USA, 294–303. https://doi.org/10.1145/1031607.1031655

[42] Anthony Tang, Melanie Tory, Barry Po, Petra Neumann, and Sheelagh Carpendale. 2006. Collaborative coupling over tabletop displays. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, New York, NY, USA, 1181–1190. https://doi.org/10.1145/1124772.1124950

[43] John C Tang. 1991. Findings from observational studies of collaborative work. *International Journal of Man-machine studies* 34, 2 (1991), 143–160. https://doi.org/10.1016/0020-7373(91)90039-A

[44] Thomas Daniel Ullmann. 2019. Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education* 29, 2 (2019), 217–257. https://doi.org/10.1007/s40593-019-00174-2

[45] J. Wayne Wrightstone. 1934. An Instrument for Measuring Group Discussion and Planning. *The Journal of Educational Research* 27, 9 (May 1934), 641–650. https://doi.org/10.1080/00220671.1934.10880446 Publisher: Routledge _eprint: https://doi.org/10.1080/00220671.1934.10880446.

[46] Bin Xu, Jason Ellis, and Thomas Erickson. 2017. Attention from Afar: Simulating the Gazes of Remote Participants in Hybrid Meetings. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. ACM, New York, NY, USA, 101–113. https://doi.org/10.1145/3064663.3064720 event-place: Edinburgh, United Kingdom.