Towards Respectful Smart Glasses through Conversation Detection

Franziska Meirose[1], Sven Schultze[1], Sebastian Kuehlewind[1], Marion Koelle[1], Larbi Abdenebaoui[2], Susanne Boll[1]

University of Oldenburg, Oldenburg, Germany[1]
OFFIS – Institute for Information Technology, Oldenburg, Germany[2]

firstname.lastname@uni-oldenburg.de, firstname.lastname@offis.de

Abstract

Talking to each other is personal, maybe even intimate. Thus, privacy expectations are particularly high during interpersonal conversations, and image or audio recordings are problematic in these contexts. In consequence, smart glasses and other body-worn devices with "always-on" cameras are not well accepted during interpersonal conversations. Proposing a simple-to-implement computer vision procedure, we work towards a solution to this issue. Using imagery from a head-worn camera we detect face-to-face conversations in real-time, as well as distinguish between intimate, personal and social conversations based on intrinsic camera parameters. Starting from a fictive scenario, we illustrate how this knowledge can be used for interaction designs that increase both, the users' as well as their bystanders' privacy, e.g., by muting audio or disabling the camera. Finally, we suggest directions for future work.

1 Introduction

Body-worn camera devices recently hit the consumer market. While those devices allow for comfortable hands-free capture of videos and images from a first-person perspective, they also face significant acceptability issues. Their camera might be perceived to be "always on" or recording automatically, thus challenging both the user's and bystander's expectations about privacy and control of image collection and sharing. These issues have been investigated in the context of lifelogging, where Hoyle et al. found that the perceived privacy-sensitivity of an image depends on factors such as location, activity, and presence of persons or certain objects (Hoyle et al., 2014). In the context of smart glasses with integrated camera functionality, conversational scenarios are perceived as particularly sensitive (Koelle et al., 2015).

Prior work on detecting conversations, e.g., in the field of activity recognition, often relies on audio data (Choudhury et al., 2008), which causes another set of privacy issues, and complex architectures for training and classification.

Veröffentlicht durch die Gesellschaft für Informatik e. V. 2018 in
R. Dachselt, G. Weber (Hrsg.):
Mensch und Computer 2018 – Tagungsband, 02.–05. September 2018, Dresden.
 https://doi.org/10.18420/muc2018-mci-0336

In contrast, our approach performs a real-time analysis of RGB data. Thus it is light-weight and easy to implement for mobile platforms. We use facial features and the movement of the mouth to detect whether a conversation takes place. In addition, our apporach also considers the social context. Using the physical distance as a proxy, we distinguish between public, social, personal, and intimate conversations (c.f., Hall, 1966) as illustrated in Figure 1. Highlighting the user-centeredness of our approach, we start by introducing the chosen scenario and background, before going into detail on the prototype's implementation. We conclude with remarks on future directions.

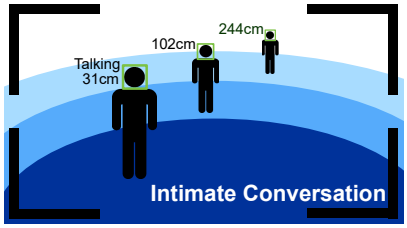

Figure 1: Basic principle; Facial features are analyzed to determine whether a person in the frame is Talking. Determining the person's distance to the camera allows to infer the type of conversation, e.g., intimate conversation.

2 Scenario and Background

In this work, we present an approach to conversation detection, and protection, that is based on proxemics, namely on how humans relate to their environment, and other persons in it. Simply put, if a conversation is detected, the proximity of the conversation partner is inferred from the camera's imagery, and defines whether the conversation is classified as intimate, personal, social or public. These classes are defined based on the human's comfort zones (Hall, 1966), as depicted in Figure 2, with the intimate zone being closest to the device wearer. In the following, we will illustrate how these zones relate to our prototypical implementation using a fictive, exemplary scenario.

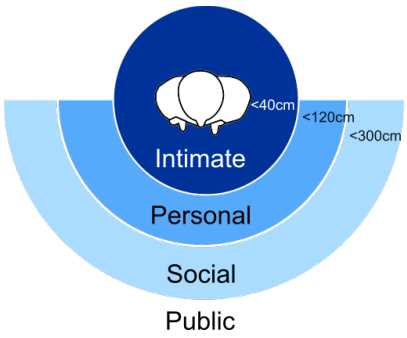

Figure 2: The distance between two conversation partners informs the type of conversation. According to Hall, distance zones can, with increasing distance to the user, be defined as intimate, personal, social and public (Hall, 1966).

On their Interrail trip through Italy, Lara and Paul are using smart glasses to automatically capture short videos and still images of memorable moments. At the moment, they are on a guided tour in the city center of Siena. When Lara is wearing the device while listening to the tour guide, the situation is classified as *public conversation* (distance > 300cm) and the experience may be captured as short clip. Ordering an ice cream or buying something from a market vendor would be classified as *social conversation* (between 300cm and 120cm distance), as a consequence, audio might be muted, and only still images would be captured. During lunch at a restaurant, a conversation with another member of her travel group or Paul would be classified as *private*, or *intimate*, depending on the proxemics. As a reaction, the system would stop the recording, and/or remind to Lara to take off the glasses to protect their privacy and to comply with social norms.

While this work mainly focuses on the detection of the different conversation types, it is not restricted to this exemplary travel scenario. A future system's interaction design could define suitable reactions depending on use case, location, or application type.

3 Prototype

In this section, we describe the proposed procedure for conversation detection. As a proof-of-concept, we implemented it in Python on a Raspberry Pi zero, using OpenCV's implementation of Viola-Jones' face detection (Viola and Jones, 2004). Using a live video feed we detect faces, and facial landmarks. Detected faces are tracked over multiple frames using a median flow algorithm. In the following, we detail on the conversation detection procedure building thereupon.

In a **preprocessing** step, each frame obtained from the live video stream is converted into grayscale and normalized, such that the darkest pixel in the frame is represented by 0 and the brightest by 255. Afterwards, camera calibration is applied based on the camera's lens parameters to minimize distortion (e.g., pincushion distortion).

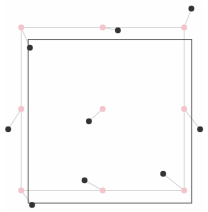

Figure 3: The x- and y-points of the shifted image as used by the median flow algorithm.

To perform **detection and tracking**, all faces present in the frame are localized using the Viola-Jones algorithm. In order to analyze a detected faces over multiple frames, which is necessary to detect whether the person is talking, we make use of the Median Flow algorithm (Kalal et al., 2010) to track faces unambiguously. In brief, this procedure allows to differentiate between multiple faces, by initializing a grid of points where the face is estimated. Secondly, all movements of the grid points are listed for the x- and y-dimension. For each dimension, the median of the distances is taken and the box around the face is moved accordingly. Then, a change in scale is determined by calculating the relative change in distances for each pair of frames. Subsequently, the median shift is selected, as visualized by Figure 3. As this approach only works well as long as the tracked face moves slowly, the Viola-Jones detection is re-applied every 30th frame, or when tracking is lost.

To **analyze** whether a conversation is taking place, the frame-wise detected face position is used to calculate the distance between the camera and the face. For our prototype, we assumed the camera to be at eye level. Thus, distance Z can be obtained as follows:

$$Z = \frac{F}{S_y} \cdot \frac{Y \cdot s_y}{y}$$

F	the camera's focal length [mm]
(S_x,S_y)	the camera's sensor size [mm]
Y	expected face height[1] [mm]
y	detected face height [mm]
(s_x,s_y)	frame resolution [px]
Z	estimated distance between detected face and camera [mm]

Based on this distance calculation, each face is categorized. In addition, facial landmarks are obtained for every face, which allows to determine the shape of the mouth. Subsequently, the distance between the upper and the lower lip ("lip distance") is calculated in pixels, and converted into millimeters (based on the distance between camera and face). We observe this value for an interval of ten frames, which allows to determine whether the person is talking.

[1] The implementation of our working prototype assumes 200mm as an adult's average face height. Depending on the use case this estimate might have to be adapted, e.g., to not exclude children.

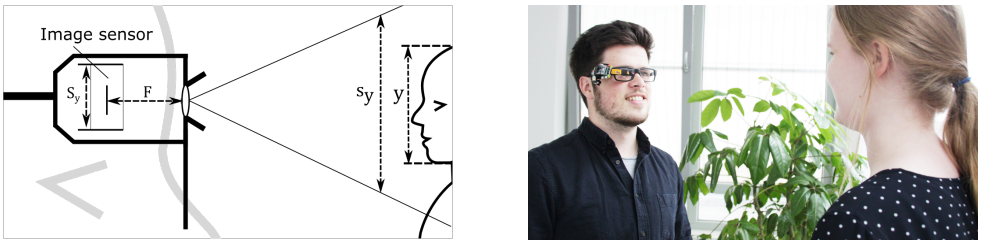

Figure 4: Left: simplified illustration of optical parameters used to calculate the distance between camera wearer and conversation partner. Right: working prototype consisting of a glasses frame with attached Vufine display, a Raspberry Pi Zero including camera and battery pack.

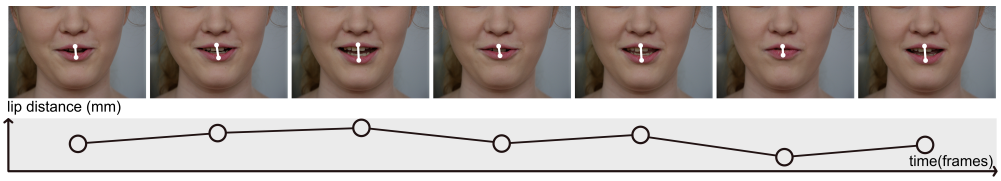

Figure 5: Measuring the variance in lip distance (bottom, from left to right) over multiple frames (top, indicated in white) allows to infer whether a person is talking.

Therefore, we first calculate the average difference of the lip distances from adjacent frames. We consider all measured lip distances of the past 10 frames, as illustrated in Figure 5. Then, we consider all differences in lip distances of adjacent frames. The average of these differences can be considered representative of how much the lips move. To counteract outliers, e.g., due to shadows or reflections, the average of differences is multiplied with the standard deviation of all measured lip distances in the current 10-frame interval. Finally, thresholding is applied to decide whether the person in the frame is talking. Joining these results with the determined distance, presence and type of conversation are concluded. Subsequently, the systems pre-defined reaction to a *public*, *social*, *personal*, or *intimate conversation* would be triggered.

4 Conclusion and Outlook

This paper is tied around privacy options for users and bystanders of head-worn devices with "always-on" cameras. We developed a working prototype and present a light-weight and easy-to-implement procedure that realizes conversation detection. Each detected conversation is classified into *public*, *social*, *personal*, and *intimate*. For future applications, we envision interaction designs that use knowledge about detected conversations and conversation type for contextual privacy protection, e.g., to blur imagery, mute audio or cover the camera lens. So far, we only have been collecting informal user feedback using our working prototype. However, a subsequent (long-term) user study will be necessary to provide insights on conversation detection and protection in-the-wild, and to substantiate the implications of this work.

References

Choudhury, T., Consolvo, S., Harrison, B., Hightower, J., LaMarca, A., LeGrand, L., … Wyatt, D. (2008). The mobile sensing platform: An embedded activity recognition system. *IEEE Pervasive Computing*, *7*(2). doi:10.1109/MPRV.2008.39

Hall, E. T. (1966). *The Hidden Dimension*. New York, NY, USA: Doubleday & Co.

Hoyle, R., Templeman, R., Armes, S., Anthony, D., Crandall, D., & Kapadia, A. (2014). Privacy Behaviors of Lifeloggers Using Wearable Cameras. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 571–582). UbiComp '14. Seattle, Washington: ACM. doi:10.1145/2632048.2632079

Kalal, Z., Mikolajczyk, K., & Matas, J. (2010). Forward-Backward Error: Automatic Detection of Tracking Failures. In *Proceedings of the 2010 20th International Conference on Pattern Recognition* (pp. 2756–2759). ICPR '10. Washington, DC, USA: IEEE Computer Society. doi:10.1109/ICPR.2010.675

Koelle, M., Kranz, M., & Möller, A. (2015). Don't Look at Me That Way!: Understanding User Attitudes Towards Data Glasses Usage. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 362–372). MobileHCI '15. Copenhagen, Denmark: ACM. doi:10.1145/2785830.2785842

Viola, P. & Jones, M. J. (2004). Robust Real-time Face Detection. *International Journal of Computer Vision*, *57*(2), 137–154.