Using machine learning to determine attention towards public displays from skeletal data

Jonas Lacher
jonas.lacher@unibw.de
University of the Bundeswehr Munich
Munich, Germany

Laura Bieschke
laura.bieschke@unibw.de
University of the Bundeswehr Munich
Munich, Germany

Florian Michalowski
florian.michalowski@unibw.de
University of the Bundeswehr Munich
Munich, Germany

Johannes Münch
johannes.muench@unibw.de
University of the Bundeswehr Munich
Munich, Germany

ABSTRACT

We develop a classifier model trained to analyze anonymized skeletal data of passers-by at interactive public displays to determine whether an interaction has occured. The test setup and data collection methods are described. The skeletal data is preprocessed to highlight more relevant bodyparts. The performance of the finished model will be evaluated statistically and compared to approaches using human observers from other research.

KEYWORDS

machine learning, attention, skeletal data, public display, evaluation

1 INTRODUCTION

Public displays are becoming an increasingly important part of the public world. It is hard to imagine subway stations, universities and other institutions without them. For scientific research, screens with interactive elements are of particular interest. However, this is often done in an inefficient way. When studying other papers, we noticed that they often want to investigate the effect of a small change by means of field tests with small samples, by having a lot of data directly recorded by human observers. Interpreting human behavior is easier for a human observer than for a machine. One idea in the field of public display research is the evaluation of an attention value. This value should indicate how much attention a person pays to the public display. Our intention is to determine this attention value with the help of a machine learning algorithm working on skeletal data of people in front of the public display. The attention value should be between 0 and 1. One means very attentive (interacts) and zero means no interest (public display is ignored). Interaction data from the public display will be used as ground truth value. Should it be possible to train a machine with a model for an attention value that can do this similarly well, this will greatly reduce the cost of such studies, and at the same time expand the scope of the data studied.

Veröffentlicht durch die Gesellschaft für Informatik e.V. in P. Fröhlich & V. Cobus (Hrsg.): Mensch und Computer 2023 – Workshopband, 03.-06. September 2023, Rapperswil (SG)

https://doi.org/10.18420/muc2023-mci-ws13-293

Consider, for example, the work of Parra et al. [8], who have studied in more detail the behavior of individuals in relation to public display installations in public spaces in order to categorize behavior or responses. In these studies, researchers have analyzed the behavior of individuals. In turn, there already exist papers [1, 2] that work with an attention value that they determine with self-developed algorithms. However, it is important to note that the attention values that have been created for these papers are each tailored to their specific problems. In this paper, a general attention value is determined using anonymized skeletal data.

The goal of our work is to determine an attention value using machine learning. For our investigation we use a public display installation in the public area of the Universität der Bundeswehr München [5]. The installation itself consists of an interactive screen that is used as a community mirror. The public display has an extensive log that, among other things, records all user interactions via the touch interface with accurate timestamps. The installation also includes a depth camera that records anonymized skeletal data, also with timestamps. In our investigative setup, we chose to use the public display log data as confirmation of attention. This approach allows us to dispense with a field experiment by an active observer. Using the interaction data, we classify the training dataset, which consists of the skeletal data. With this, we train a machine learning model that assigns an attention value between 0 and 1 to the skeletal data.

The main difference between this work and others is that we present a method that can facilitate further research in the context of public displays.

2 STATE OF THE TECHNOLOGY

Several papers reflecting the current state of research are summarized below.

Müller et al. [7] present a new taxonomy that is intended to upgrade and specify the way public displays are described. Besides the creation of specific descriptions, it also lists problems that can occur during research in the field of public displays. Getting passing people interested in a public display without getting on people's nerves is described as one of the biggest challenges facing research and development of public displays. This is the basis for our consideration to determine an attention value. The underlying intention of the research area is to sensitize people for a public display. If we want to examine the behavior more closely, we have to look at the

interest that is shown towards the installation. A tool that could facilitate this research is our attention value.

Fischer and Hornecker [4] had a similar focus. Their published paper deals with public displays at house façades. This paper elaborates that one encounters some challenges in working with public displays (on house facades). One major problem is the evaluation of the installations under study. There are no clear guidelines according to which an evaluation could be done. A solution for the elaborated problem has not been presented in the paper. However, new terminology is provided. It should help to describe the installations and the environment of the public displays. The paper supports the thesis that the environment around the public display is of great importance to the installation itself when investigating, designing and researching it and should not be neglected when considering it. In the generation of our attention value, we have chosen to illuminate the time around the touchscreen interaction. We consider 20 seconds before and after the interaction. We became aware of the importance of the space around the display through the paper just mentioned [4] and Memarovic et al. [6], which is briefly summarized in the following paragraph. With the time window of 40 seconds around the interaction time, we include different spatial positioning of the recorded persons in our investigations. The new terminology presented in the paper incorporates this assumption.

Memarovic et al. [6] is a paper that also publishes new terminologies in the form of a new model. The new model is used to categorize social needs in order to better match public displays to the defined needs. This is expected to increase the attractiveness of public displays. In addition to the model, a field experiment was also described that is conceptually based on the paper's new social needs model. The public display is called "FunSquare." The interaction of people with the "FunSquare" was studied by the researchers. With the help of the observation results, the conceptualizations of the subdivisions of the surrounding space of public displays have been presented. The established model and the new terminologies should support the analysis and development of public displays.

As already mentioned, Memarovic et al. [6] drew our attention to the space around public displays. Moreover in this paper we could clearly see that field studies are time consuming and therefore can only run over a relatively short period of time. Such field studies can now be facilitated or completely replaced by the described attention value.

Dostal et al. [2] is a paper that expresses interest in an attention value similar to our research area. In this paper, the attention value is determined using two values: the distance of the recorded person to the public display, which in this research was a wall screen, and the eye movement of the recorded person. The focus of this work was to create a toolkit that helps designers without much programming experience to generate attention values and incorporate them into the development. The newly presented toolkit is named SpiderEyes. The paper included a structure description, as well as a functional description of the entire toolkit.

With this paper, we had access to a concept that also aims to generate an attention value. We compared the approaches used with our own possibilities and concluded that it should be possible to determine an attention value using the data available to us.

The paper that is probably most similar to our work is Alt et al. [1]. As the title suggests, features were compared that can be used to assess the attention of passers-by towards public displays. The features used were the recognizability of a face, walking speed, walking distance, shoulder orientation, head orientation and gaze direction, as well as different combinations of these. The whole thing was examined in a laboratory test. The evaluation was also based on schemes to suggest whether passers-by can be identified or recognized. The paper concludes that the highest precision with identified passers-by is achieved by a combination of position data. For unknown pedestrians, facial data works best, but is still barely more precise than the baseline. Instead of a laboratory test as in Alt et al. [1] we aim in our work at a data evaluation by a machine learning model. The evaluation of the machine learning model is our attention value.

The last paper that we consider important for our articles is Stanley [9]. This paper is about using the Microsoft Kinect to determine an attention value. Similar to our approach, a black box procedure is performed using collected data. The collected data is categorized and through the categorization, an attention value of the recorded individuals is determined. Unlike us, the setup described by Stanley [9] is used to record students in class and during exams to support learning research. The tools described by Stanley [9] can be easily modified, as our work shows. So that we use the skeletal data of the Kinect camera to determine an attention value for the public display at the Universität der Bundeswehr München.

3 APPROACH

3.1 Test Environment and Data

Our work is based on the public display, which is installed in a public space at the Universität der Bundeswehr München [5]. The public display, which acts as a CommunityMirror, is composed of a touchscreen and a depth camera. The display is able to record interactions with the display as interaction logs. The depth camera records the movements of people in close proximity to the display and stores the recordings as skeletal data, so that all recordings are anonymized. The exact structure of the CommunityMirror with the description of the skeletal data can be found in Fietkau [3] and Koch et al. [5]. One frame is dedicated to one person. If more than one person is in front of the camera at the same time, up to three frames are created, one for each person detected. No more than three people can be tracked at the same time. The perceptual range of the camera is focused on the display and its near environment. The interaction logs of the display and the skeletal data, which are recorded in frames, are important for our considerations.

The interaction logs include time stamps indicating when the screen was touched. The skeletal data consists of points in a 3D space that reflect the recorded area of the camera (see Figure 1). The individual points describing a skeleton are part of a frame. Timestamps of the display and frames of the camera are matched. The intention is to assign the matching timestamps to the frames. This assignment can be represented by a newly created label for the skeletal data. The label should only contain a zero or one. Zero should be set at the beginning for all skeletal data. If a time stamp can be assigned to a skeletal dataset, i.e. the recorded frame was created at an interaction time, the label is set to one. In addition, all frames recorded 20 seconds before and after the assigned interaction time are also set to one. By labeling the skeletal data in this way,

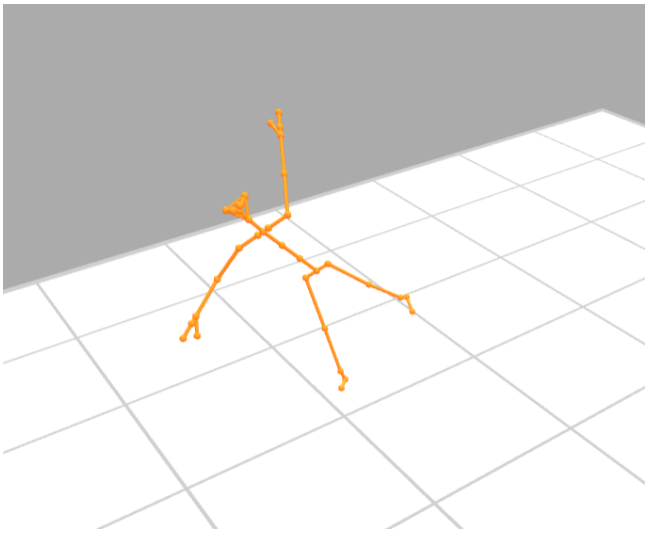

Figure 1: Visualized skeletal data

we create a dataset that divides skeletal data into two groups. One is the data that is important for describing the interaction with the display. Zero are the skeletal data, which are not important for the interaction with the display. The chosen time period of a total of 40 seconds around the interaction time has been chosen at the discretion of our research group and can be adjusted as desired. In our considerations, we found that a time window of about 20 seconds is necessary to understand the content of the screen in order to then interact with it.

This is now based on labeled skeletal data that falls into category one (important for interaction with the display), or zero (not important for interaction with the display). This data structure is to be used by a machine learning model. The model is to use the data to learn to determine an attention value for a skeletal dataset between 0 and 1. As mentioned before, the attention value 0 describes no interest in interacting with the public display, whereas 1 represents a high interest in interacting with the public display.

3.2 Model and Training

When thinking about the concept of our machine learning model and the preparation of the data, we had to deal with some problems.

In the proposed structure of our model, we decided on a simple categorization of the dataset. It is important to note that we made simplifications to facilitate the analysis of the data. The effect of this simplification is that our approach omits those cases where people are looking at the public display attentively but not interacting with it. By the way of the intended labeling, these cases are labeled with a zero. It would be laudable if a system could recognize such a situation. However, we call attention without interaction exceptions. We assume this case to be negligible. Moreover, it should be emphasized that the intention of a public display is interaction. Thus, only people who interact with the display are of interest. People who become attentive but do not build up enough interest to interact with the display are of lesser importance in the study of behavior near public displays. The attention value we define is intended to evaluate whether the skeletal data considered describes an interaction or not. The smaller the attention value, the less likely it is that the person described by the skeletal data shows interest in the public display.

One might consider whether our attention value corresponds to an interaction value. However, this is not the case. A simple interaction value could be determined from the relative distance from the skeletal data to the public display. The relative proximity would determine the interaction value. As a result, the interaction value increases as you get closer to the display and decreases as you move away.

The difference to our approach is that we consider the temporal space before and after the interaction as relevant, thus also skeletal data located further away from the display. The model should be trained with all skeletal data. Which means that data with label one, as well as with label zero function as training data.

Of course, it is not one hundred percent certain what the machine learning model will do in categorizing the skeletal data, since our concept is somewhat like a black box process. A dataset is given to the black box and the black box determines commonalities independently. Proximity to the display will almost certainly have an important factor in determining the attention value. But in our approach, it is not the only reference point for categorization. Thus, our attention value is not equal to an interaction value.

Due to our chosen machine learning model, a stochastic evaluation of the attention values to be generated is not possible. We repeat: In order to train our model, we label the available skeletal data. Our label is a binary categorization into interaction relevant (one) and interaction irrelevant skeletal data (zero). The labeling of the data is to be done by matching the recorded frames from the camera and the interaction logs from the display, as well as a fixed time window of 20 seconds before and after the interaction log. It is envisioned that the model will now learn to score the skeletal data between zero and one, using the labeled data set. An attention score between zero and one, to be computed by our model, cannot be evaluated using e.g. a confusion matrix [10]. To work with a confusion matrix, the output of the model must be rounded up or down. This would mean producing an output of zero or one. However, the intended attention value is to be a value between zero and one, which eliminates the need for an analysis by the confusion matrix. We are well aware that an analysis of the attention value would be possible through a field study.

3.3 Validation/Future Work

The reasonableness of our attention value could be proven by case examples as follows. The examples are defined in advance with the desired attention value that should be output by the model. Then the case studies are given to the model for evaluation and the calculated attention value is compared with the expected attention value. If the calculated value is within the error tolerance range from the expected value, we assume that the model works. Otherwise, we have shown that the model is not able to realize our intention: to determine a meaningful attention value by skeletal data.

The following is an experiment. We divide the description of the recorded behavior of a person into three phases. The approach phase: this is the phase of movement towards the public display.

The interaction phase with interaction occurring or no interaction occurring: this is the phase when the recorded person has the smallest relative distance to the display. And the distance phase: this is the case when a person moves away from the public display. To evaluate the attention value, one can define any number of case studies and the corresponding expected attention value. The more case studies are defined and checked, the more the reliability of the the attention value is shown.

An exemplary procedure for formulating a experiment would be: We first describe a scenario assumed to be realistic. This is described and understood as ground truth. Then the described scenario is divided into the three defined movement phases and a justified expected value is assigned to each of these movement phases, which must later be compared with the determined attention value. When comparing the expected value and the determined value, only a predefined error rate may occur, so that our attention value is meaningful for this experiment. Otherwise the examination shows that the attention value is inaccurate.

Our experiment includes the following scenario: a person walks through the image relatively far at the edge of the camera perspective. The person has a constant speed and does not stop at any point. The approach phase is any movement that decreases the distance to the public display. It starts with a distance to the public display that is relatively large. The person decreases the distance to the display by constantly moving to the closest possible point to the public display. The short moment where the person is closest to the display represents the interaction phase, which in this experiment consists of only one moment. After that, the person moves away from the public display and is thus in the distance phase. In our experiment, we choose at least one upper bound. Where if the limit is exceeded, the attention value is to be understood as imprecise.

An attention value between 0 and 0.3 is to be expected for the approach phase, since the person moves towards the public display, but does not slow down or even stop, or comes really close to the public display. Thus, the determined attention value should not exceed the value 0.5. In the interaction phase, which in this case describes only one point in time, the attention value should not exceed 0.5, anything less is acceptable. In the distance phase the attention value should decrease. The next point to the public display is exceeded and the person moves away. Thus, the value should decrease from the maximum back to zero.

We consider this form of validation as a possibility, since we assume that there is a finite set of behaviors in front of a public display, so these behaviors can also be covered by finitely many case studies.

4 CONCLUSION

Our work has shown that the determination of an attention value by a trained machine learning model is possible. Categorizing the data by a trained model is a legitimate approach, as it saves time and human resources. The main challenge faced with this approach is the analysis of the data obtained by the model. As our experiment descriptions have already shown, validating the results is more complex than with other evaluation procedures performed using trained models. The main difference is that our data cannot be classified into fixed categories that can be represented by integers.

In summary, the determination of an attention value based on recorded skeletal data by a trained model is possible. However, new methods of evaluation are necessary to determine the accuracy of the determined attention value.

REFERENCES

[1] Florian Alt, Andreas Bulling, Lukas Mecke, and Daniel Buschek. 2016. Attention, Please! Comparing Features for Measuring Audience Attention Towards Pervasive Displays. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) *(DIS '16)*. Association for Computing Machinery, New York, NY, USA, 823–828. https://doi.org/10.1145/2901790.2901897

[2] Jakub Dostal, Uta Hinrichs, Per Ola Kristensson, and Aaron Quigley. 2014. SpiderEyes: Designing Attention- and Proximity-Aware Collaborative Interfaces for Wall-Sized Displays. In *Proceedings of the 19th International Conference on Intelligent User Interfaces* (Haifa, Israel) *(IUI '14)*. Association for Computing Machinery, New York, NY, USA, 143–152. https://doi.org/10.1145/2557500.2557541

[3] Julian Fietkau. 2023. A New Software Toolset for Recording and Viewing Body Tracking Data. In *Mensch und Computer 2023 – Workshopband*, Peter Fröhlich and Vanessa Cobus (Eds.). Gesellschaft für Informatik e.V., Bonn, Germany, 4 pages. https://doi.org/10.18420/muc2023-mci-ws13-334

[4] Patrick Tobias Fischer and Eva Hornecker. 2012. Urban HCI: Spatial Aspects in the Design of Shared Encounters for Media Facades. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 307–316. https://doi.org/10.1145/2207676.2207719

[5] Michael Koch, Julian Fietkau, and Laura Stojko. 2023. Setting up a long-term evaluation environment for interactive semi-public information displays. In *Mensch und Computer 2023 – Workshopband*, Peter Fröhlich and Vanessa Cobus (Eds.). Gesellschaft für Informatik e.V., Bonn, Germany, 5 pages. https://doi.org/10.18420/muc2023-mci-ws13-356

[6] Nemanja Memarovic, Marc Langheinrich, Florian Alt, Ivan Elhart, Simo Hosio, and Elisa Rubegni. 2012. Using Public Displays to Stimulate Passive Engagement, Active Engagement, and Discovery in Public Spaces. In *Proceedings of the 4th Media Architecture Biennale Conference: Participation* (Aarhus, Denmark) *(MAB '12)*. Association for Computing Machinery, New York, NY, USA, 55–64. https://doi.org/10.1145/2421076.2421086

[7] Jörg Müller, Florian Alt, Daniel Michelis, and Albrecht Schmidt. 2010. Requirements and Design Space for Interactive Public Displays. In *Proceedings of the 18th ACM International Conference on Multimedia* (Firenze, Italy) *(MM '10)*. Association for Computing Machinery, New York, NY, USA, 1285–1294. https://doi.org/10.1145/1873951.1874203

[8] Gonzalo Parra, Joris Klerkx, and Erik Duval. 2014. Understanding Engagement with Interactive Public Displays: An Awareness Campaign in the Wild. In *Proceedings of The International Symposium on Pervasive Displays* (Copenhagen, Denmark) *(PerDis '14)*. Association for Computing Machinery, New York, NY, USA, 180–185. https://doi.org/10.1145/2611009.2611020

[9] Darren Stanley. 2013. *Measuring attention using Microsoft Kinect.* Master thesis. Rochester Institute of Technology. https://scholarworks.rit.edu/theses/4768/

[10] Dennis T. 2019. Confusion Matrix Visualization. Medium. https://medium.com/@dtuk81/confusion-matrix-visualization-fc31e3f30fea