A Comparison of Attention Estimation Techniques in a Public Display Scenario

Wolfgang Narzt(✉)

Johannes Kepler University, Linz, Austria
wolfgang.narzt@jku.at

Abstract. Human interaction with a public display presupposes a person's attention. An Interactive display, hence, aims at attracting attention by e.g. emitting a strong signal that makes the inattentive visitor turn towards it. The challenge in this regard is to reliably determine the attention of passers-by. In this article, we investigate different technical methods for estimating attention in a public display scenario by measuring physical expressive features, from which attention can be derived. In the course of an experimental setup we compare a Support Vector Machine, a neural network using a Multilayer Perceptron and a Finite State Machine and compare the results to a manual reference classification. We carve out strengths and weaknesses and identify the most feasible measuring method with regard to precision of recognition and practical application.

Keywords: Attention estimation · Support Vector Machines · Neural networks · Multilayer Perceptron · Finite state machines

1 Introduction

Public displays have evolved to an acknowledged enabling technology for shopping scenarios [9]. Retail traders recently aim at displaying interactive content and offer innovative (e.g., gesture-based [8, 10]) interaction in order to attract by-passers and to potentially increase sales. However, interaction with a public display presupposes a user stopping in front of the display and focusing to its content. Thus, enticing passers-by to a focused interaction (depending on how much attention they are already paying) has been recognized as a recent challenge in attention estimation research [3, 7]. An attention-aware display might want to present content in a way that indicates that a head-on looking visitor has been registered and is addressed individually.

Within advertising and marketing, such attentive user interfaces [1, 11, 16] are applied to raise attention of an already interested person, which is referred to as the AIDA principle (i.e. attract Attention, maintain Interest, create Desire, and lead customers to Action [12]). It either draws a user's attention or motivates interactions [19] and has led to a series of recent research projects (see e.g. [16, 20]), additionally stimulated by affordable and miniaturized sensing technology capable of measuring inadvertent cues from human subjects indicating their current state of attention.

Measuring attention is a means to the end of improving reaction of an interactive system to its users' attention. Attention-aware displays therefore need to continuously

F.F.-H. Nah and C.-H. Tan (Eds.): HCIBGO 2017, Part I, LNCS 10293, pp. 338–353, 2017.
DOI: 10.1007/978-3-319-58481-2_26

estimate the attention of persons they can perceive and need to have internal models for what types of signals may be appropriate to reach the goal of raising or lowering the attention a person is devoting to the display.

Most of suchlike attention-aware systems are closed and only locally react to estimated states of their viewers, whereas networked solutions aim at interacting with viewers across larger spaces with multiple displays and sensors. This, however, is a highly interesting scenario, especially for the e-commerce sector, where visitors are not only addressed differently depending on their state of attention, but also depending on the history of interaction.

Networked systems also deliver the background for our research: In the course of a federal initiative for building a smart city in a suburban area near Vienna ("Seestadt Aspern"), Siemens intends to install City Hubs at public places around the city, i.e. interactive networked displays providing tailored information or services for passers-by. A City Hub is supposed to automatically and context-sensitively draw a visitor's attention to its screen and to simultaneously address an already interacting person as well as a glimpsing by-passer in its peripheral field of view. Due to NDAs, we are not permitted to give any example of a City Hub application, nor can we provide further insights regarding location or number of installation sites. However, we may mention that City Hubs will be used for e-commerce services, as well, (amongst other application domains) and act as connected network nodes across a city area that "globally" capture contextualized information for individual human-machine interaction.

In this paper, we aim at evaluating different attention estimation methods in the context of public displays and assess our results in terms of effectiveness, accuracy, configuration costs and practical application. We therefore present an appropriate tailored attention model for public displays (derived from well-known and established models [3]) and experimentally (i.e. in a public scenario with arbitrary visitors in front of a display) oppose performance, accuracy and error rates to a manual reference classification. A critical assessment of our research and the examined measuring methods and considerations for further research, conclude our work.

2 State-of-the-Art

Human attention has primarily been investigated from a psychological point of view. It has been characterized by relations between attention allocation, attention capacity and task effort [6]. Attention, in very general terms, is the "*process by which organisms select a subset of available information upon which to focus for enhanced processing and integration*" [17]. The link to the computer vision domain arose in the 1980s and is based on a seminal psychological contribution by Treisman and Gelade [13], who proposed an attention model explaining the transition from the pre-attentive processing of pure features to the identification of objects characterized by a conjunction of features (Feature Integration Theory). It proved highly influential in many attention models developed in the context of computer vision applications.

Wickens and McCarley [18] derived a formula that yields an attention estimate for particular objects from quantifiable factors called Saliency, Effort, Expectancy and

Value (SEEV). This approach has proven popular as a simple base model that helps to structure the parameterization of attention factors without specifying a concrete attention formation process. Most of the proposed models either satisfy pure computer vision goals without describing biological and neurophysiological processes, or manage to straddle both the computer vision and biological requirements [14].

Research that is in the line of automatic detection of attention of human subjects is less interested in the internal attention processes, and more in external indicators of the results of those processes, in the extent to which they can be analyzed for the purpose of estimating a subject's attentional status. Such cues are called overt attention and comprise body, head and eye movement, whereas mental shifts of focus are called covert attention and can, if they occur without any overt components, generally not easily be measured without interrogating the subject [17].

Meanwhile, inadvertent overt cues can be measured using affordable sensing technology, which has recently stimulated research on attention-based interactive systems. Gollan et al. [5] e.g. employed a consumer-grade and accordingly low-cost Microsoft Kinect depth camera for measuring attention. A suchlike system is affordable and, most of all, unobtrusive, i.e. people need not be aware that they are being tracked. The authors aimed at clustering by-passers in an uncontrolled public display scenario into one of several distinct categories using Support Vector Machines and referring to Wickens' SEEV model in their reasoning. A "short time frame" (STF, spanning the past three frames only) comprised categories such as "Peripheral Visual Range" and "Concentrated Gaze", whereas the "long time frame" (LTF, beginning with the first in which a subject appeared, up to the current one), contained categories such as "Glimpser" and "Stopper" (see Fig. 1).

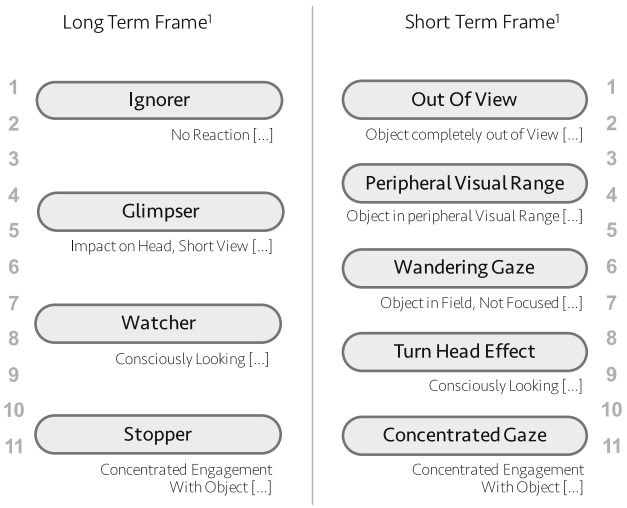

Fig. 1. Attention concept (Gollan et al.)

Validation of results was conducted against the results of manual tagging of the recorded material and showed an accuracy of 92.3% for LTF and 85.5% for STF. In

general, the research done by Gollan et al. seems most promising of the minimally obtrusive approaches (also see [3]).

More recently, the authors also approached the same experimental context (using the same data) from a different perspective: They showed that using the aforementioned SEEV-Model, the factor of Effort exerted by a subject to change his/her movement, position or head pose is correlated to various degrees with the probability of increasing his/her level of attention [4]. An important difference to the method before is that relative changes in attention could be estimated without supervised machine learning methods.

3 Approach

The attention model proposed by Gollan et al. has been selected as the basis for our investigations as it is promising threefold:

1. It is derived from the proven SEEV model and therefore claims to build upon mature attention estimation structures.
2. Its accuracy is compelling in terms of correct estimation of attention states even when using low-cost sensors or cameras.
3. It is not necessarily dependent on the use of machine learning methods in order to estimate attention states. Also, other approaches apply (see [4]).

Particularly, point 3 is of major interest, as machine learning approaches require preceding training phases and appear to be inflexible regarding e.g. changes of camera position or untrained events. Our intention is to investigate different estimation techniques (whereby the SVM approach has turned out to be the most common in this domain – with the drawback of lengthy initialization) pursuing two objectives:

1. Identification of the most accurate and performant method in terms of detecting the correct attention state (disregarding complexity, effort or costs)
2. Identification of the most practical method in terms of simplicity, operational readiness and flexibility regarding changes in the setup (e.g. camera shifts)

For these comparisons, we opposed a classical SVM, a second automatic learning technique with neural networks using a Multilayer Perceptron (MLP) and a straight approach performed by a Finite State Machine (FSM). First, though, we have to discuss a few adaptations we did on the Gollan model in order to customize it to the requirements of our aforementioned City Hub scenario: The original model proposes 11 distinct attention states both for STF and LTF (see Fig. 1), but aggregates them to 4 (for LTF), respectively 5 (for STF) labeled states for better human comprehensibility. The remaining subparts (secondary states) primarily divide their higher-ranking labeled states by means of time (for LTF, e.g. how long is a person glimpsing? Short-medium-long) or distance (for STF, how distant is the stimulus? Far-medium-close).

For our requirements, we on the one hand suggest to reduce the number of attention states (both labeled and secondary states) and on the other try to merge STF and LTF

classifications in a way that we always use current snap-shot data contextually considering the historic progress. This simplification is done in respect to practical applicability and is backed on the following arguments:

1. In the original model, it is difficult to distinguish between a "Glimpser" and a "Watcher" in the LTF showing a blurred borderline in the model definition when considering secondary states: How can we differentiate between a "long glimpse" and a "short conscious look"?
2. The "Concentrated Gaze" in the STF is characterized by a "concentrated engagement with an object" and distinguishes two secondary states with "medium stay" and "long stay" (again amongst other attributes). However, as the STF aggregates measurements from only 3 frames, we are not sure how to measure a medium or long stay in this context without a history of data.
3. Is an application using this attention model supposed to react on all attention states in each of the timeframes STF and LTF? As there is no justified answer to this question (it always depends on the application domain), it is up to the reader to subjectively rate on the quantity of states and on how he/she would react on a "long glimpse" in contrast to a "short conscious look".

The reduction process finally reveals three remaining (labeled) attention states "ignoring", "watching" and "ready to interact", which we use to estimate the degree of attention of passers-by (see Fig. 2).

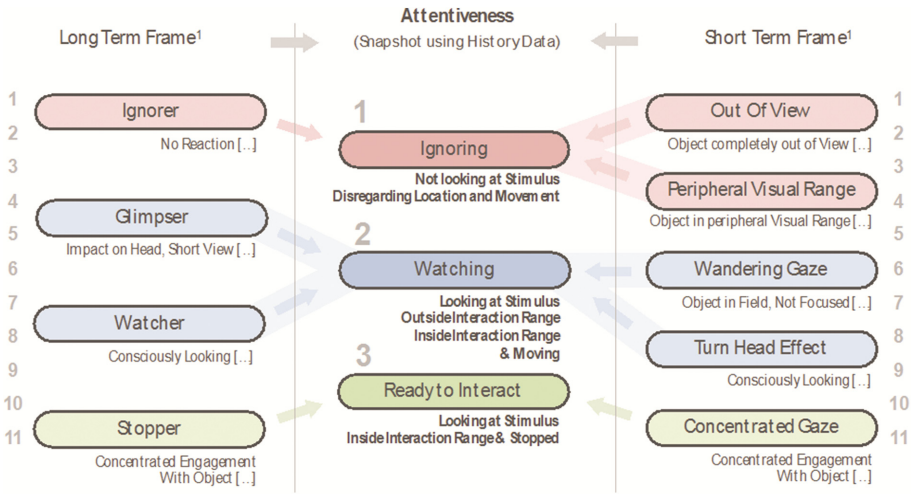

Fig. 2. Attention model deduction

These states are defined as follows:

1. The attention state ***A*** of a person ***p*** is "***ignoring***" when the person is not looking at a stimulus (display) regardless whether the person is moving or standing still at any location.

2. The attention state ***A*** of a person ***p*** is "***watching***" when the person is facing a stimulus, being too far away for interaction (outside an interaction zone). ***p*** remains "***watching***" inside the interaction zone while moving.
3. The attention state ***A*** of a person ***p*** is "***ready to interact***" when the person is facing a stimulus and has stopped inside the interaction zone (i.e., the person is close enough for interaction).

Following the example of the attention model of Gollan et al., our conceived labeled attention states are subdivided into secondary states offering a more focused distinction of what a person is actually doing. In our model, we introduce these secondary states using another layer of detail (see Fig. 3): Given that the three labeled attention states are settled on the highest layer representing the highest degree of abstraction (i.e. we put them on layer 3), the secondary attention states are on layer 2, including information on a person's focus, moving direction and distance, from which the layer 3 classifications can be inferred. Still, those states are abstract with features like "*moving away*" or "*moving perpendicular*", yet, they add information to the higher-ranking states on layer 3. The base layer 1 finally represents the technical basis and provides input data that can be measured by all kinds of sensors.

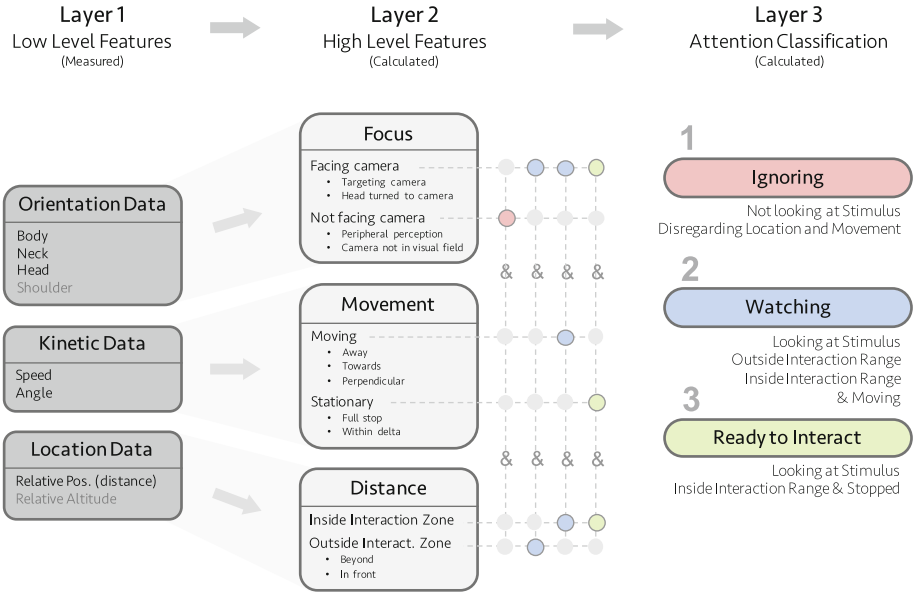

Fig. 3. 3-Layer attention model

Taking a closer look onto the model, we recognize that the attention states on layer 3 can be calculated by a logical combination of features from layer 2. While the attention state "*ignoring*" just utilizes the "*not facing stimulus*" feature of layer 2 (disregarding any movement or the location of a person), "*ready to interact*" combines focus, movement and distance in a way that a person must "*face the stimulus*" while "*stationary*"

and "*inside the interaction zone*". Thus, we conclude that the attention states on layer 3 can be calculated using a rule-based system (i.e. a Finite State Machine, FSM). These findings are the basis for our investigations on different attention estimation methods. As the FSM seems to consist of a simple rule-set (although the actual implementation is more complex than depicted in Fig. 3, using e.g. grading at the transition between states) it potentially represents a manageable alternative for estimating attention compared to established but complex machine learning methods.

4 Experiments

For comparison of different attention estimation methods, we created an experimental setup in the context of a museum scenario using a Microsoft Kinect v2 depth camera. The camera delivers position data, orientation and, as much as possible, gaze of the visitors, from which attention parameters can be estimated. As the museum scenario is uncontrolled we also wanted to have a guided setup enabling us to compare results disregarding environmental factors, thus, resulting in the following two setups:

1. Laboratory setup
   In the laboratory setup, we ensured optimal camera angles, good contrast and light conditions in order to minimize measurement errors. All participating persons were aware of the experiment, and we created a rough screenplay ensuring a balanced output of all three labeled attention states. We limited the distance between stimulus and person to a maximum of 4 meters in order to avoid measuring errors due to hardware restrictions as the depth sensor's (infrared) reliability decreases at larger distances. Moreover, only one person was in the camera's field of view avoiding occlusions and feature data loss.
2. Museum setup
   In this test setup, we collected data in a real-life scenario. In cooperation with the Ars Electronica Center, a museum for art, technology and society, we recorded visitors near an exhibit (a screen showing dynamic information in a loop). Visitors were aware about video surveillance in this area but had not been informed about the purpose of our experiment. And, as manual classification took place in a separate room, visitors were unaware that their attention was part of an experiment, so they acted naturally. Unfortunately, we were confronted with a series of issues that had to be considered when comparing results: Due to technical restrictions of the exhibit we were unable to place the sensor in an ideal position. Lighting conditions were suboptimal, and the distance between sensor and visitors was often >4 m. As a consequence, we were expecting measuring errors, particularly at larger distances.

Right after capturing the test persons on video, we classified the attention levels manually in order to create a ground truth for all further evaluation. To do so we developed a classification tool capable of synchronously playing the recorded video and feature data files generated by the Microsoft Kinect v2 with options to classify every single video frame in arbitrary replays. The person operating the tool was able to stop the video at any time and move through the video frame by frame. This process was

conducted by two persons independently in order to minimize subjective misinterpretations. In total, we classified about 23000 frames for all tests (with varying framerates between 15 and 30 fps – dynamically managed by the Kinect), i.e. we rely on video material of approx. 20 min net, where persons were classifiable (more than 2 h gross). We achieved accordance of 92.34% between the two classifiers. For differently rated frames we set the attention state by bilateral negotiations afterwards.

This ground truth was then used for further training and evaluation. As machine learning methods require training phases we provided 50% of the manually classified data as training data and performed 5-fold cross validation, i.e. we split the training set into 5 subparts, 4 of which contained training data, and used them to predict the attention levels in the remaining subpart. This procedure was repeated 5 times for every permutation for both SVM and MLP.

5 Results

We have evaluated three techniques for determining the attention. Due to its predominance in literature we trained a Support Vector Machine (SVM) on our datasets which proved to work reasonably well. However, the simplicity of our attention model and the low number of features used to estimate the three labeled attention states *ignoring*, *watching* and *ready to interact* (see Fig. 3) convinced us to implement a state machine for classification as this has several advantages over a machine learning approach:

- A state machine as a rule based system requires no training with a labeled data set, which in our scenario of public displays is a definite advantage.
- The state machine allows for easier adaption of sensor placement in relation to the stimulus via parameters.

While our implementation of a state machine doesn't achieve the same accuracy as an SVM it does work reasonably well to justify such an approach, especially considering the advantages it has over methods of machine learning in our scenario. In contrast, we compared the SVM with a neural network, in this case a Multilayer Perceptron (MLP), to see if results can be improved when pure accuracy is of utmost priority. Both SVM and MLP had problems with the quality of our recordings, as the dataset from the museum scenario turned out to be unbalanced regarding the three labeled attention states as predictably the majority of visitors captured by the sensor were ignoring the surveyed exhibit. This imbalance is expected in a public display setting; however, it does constitute a problem for training an SVM or MLP as this imbalance is likely to skew their respective predictions.

The dataset consists of 68678 frames recorded over two days. We then filtered this data, thus only frames containing valid values for head rotation remained. We also removed a small window of 10 frames directly around attention state changes to reduce the impact of sensor noise, which left us with 14177 frames. The skewed nature of the museum dataset has prompted the generation of a more uniformly distributed dataset; thus, this second dataset was recorded in our lab and contains 8591 usable frames with a roughly equal distribution in our defined attention states.

5.1 Support Vector Machine

A Support Vector Machine is a method of supervised machine learning introduced in [15]. The idea is to map input vectors to a high-dimensional feature space and then try to construct a decision surface to separate two given classes. Depending on the kernel used this surface can be linear, polynomial or with specialized kernels of arbitrary shape. As this method is only able to separate two classes there are tricks to generalize for multiclass problems. The predominant strategy is one-against-one where an SVM is trained for every pair of classes and the decision is made on which SVM gets the most votes. This is also the strategy employed by the implementation we use.

We trained several SVMs using 5-fold cross-validation with linear and Gaussian radial basis (RBF) kernels. Linear kernels proved to be unable to separate our classes satisfactorily but the RBF produced good results. The final parameters were selected by performing a grid search in a manageable parameter space. To satisfy the necessary condition for an SVM of samples being independent, they are randomized prior to training with values between 1 and 10 for costs C and 0.1 to 1 for *gamma*. The parameters for our final model were $C = 5$ and $gamma = 0.75$ as this parameter setting worked equally well for both our datasets.

As has already been stated, the dataset from our real-world scenario is heavily skewed towards one class. Therefore, we introduced class weights to the SVM algorithm to account for this imbalance. In the end, we achieved an overall accuracy of 97.83%. Table 1 shows the results of the SVM classifier detailed by class and distinguished by sensitivity (true positive rate, TPR) and specificity (true negative rate, TNR). The table reads as follows: We had 6424 matches for the attention state ignoring. 74 were estimated to be watching while their manual reference classification was ignoring, etc. The skew towards the ignoring state is immediately obvious.

Table 1. SVM results on the museum dataset.

		Predicted			Sensitivity (TPR)	Specificity (TNR)
		Ignoring	Watching	Ready to interact		
Actual	Ignoring	6424	74	6	98.77%	87.52%
	Watching	73	296	1	80.00%	98.90%
	Ready to interact	0	0	215	100.00%	99.90%
	Sum (frames)	7089			**Accuracy**	**97.83%**

Results of our second dataset from lab recordings are shown in Table 2. While overall classification accuracy for this dataset is slightly lower at 96.35% than for our museum dataset the accuracy per class is more balanced. Overall, the SVM classifier performs well both in real-world testing and on simulated data.

Table 2. SVM results on the lab dataset.

		Predicted			Sensitivity (TPR)	Specificity (TNR)
		Ignoring	Watching	Ready to interact		
Actual	Ignoring	2204	65	43	95.33%	98.08%
	Watching	41	1975	15	97.24%	97.43%
	Ready to interact	3	1	256	98.46%	98.66%
	Sum (frames)	4603			**Accuracy**	**96.35%**

5.2 Artificial Neural Network

In recent years, neural networks emerged as state-of-the-art for machine learning tasks due to the advances in GPU processing and now generally outperform SVMs. Therefore, we decided to compare results of our SVM classifier with an artificial neural network, specifically a Multilayer Perceptron (MLP). We used the implementation provided by the KNIME Analytics Platform. A Multilayer Perceptron is a feed-forward artificial neural network with a certain number of hidden layers. Each layer consists of perceptrons, which are simple linear classifiers. The input is fed into the network and passed from layer to layer where on each layer the perceptrons learn the appropriate threshold for the given classification problem. In contrast to a Support Vector Machine classifier tuning a neural network is more complex and generally requires more insight into the data at hand.

For our datasets, a shallow network with 3 layers and 50 neurons per hidden layer delivers good results but is not on par with the SVM classifier. Table 3 shows the results of the MLP classifier on the museum dataset. It reveals that the MLP also has problems with the imbalanced nature of the data but while the SVM classifier can take class weights into account this option does not exist for the MLP classifier we used. The MLP classifier achieves an overall accuracy of 97.97% on the museum dataset and therefore slightly beats the SVM classifier. However, this is mainly achieved by increasing accuracy of the over-represented ignoring samples.

Table 3. MLP results on the museum dataset.

		Predicted			Sensitivity (TPR)	Specificity (TNR)
		Ignoring	Watching	Ready to interact		
Actual	Ignoring	6465	54	3	99.13%	86.42%
	Watching	77	268	1	77.46%	99.07%
	Ready to interact	0	9	212	95.93%	99.94%
	Sum (frames)	7089			**Accuracy**	**97.97%**

On our lab dataset, the MLP classifier again achieves slightly higher overall accuracy at 97.41% when compared to the SVM classifier. However, when looking at the results in detail it can be observed that MLP still has slight problems with class imbalance (see Table 4).

Table 4. MLP results on the lab dataset.

		Predicted			Sensitivity (TPR)	Specificity (TNR)
		Ignoring	Watching	Ready to interact		
Actual	Ignoring	2204	26	2	98.75%	96.75%
	Watching	52	2054	4	97.35%	98.56%
	Ready to interact	25	10	226	86.59%	99.86%
	Sum (frames)	4603			**Accuracy**	**97.41%**

While we expect that it is possible to increase the accuracy of the MLP classifier for our task, this would require vastly more effort and more processing power. Therefore, we see little benefit to using a neural network instead of a SVM for our task.

5.3 Finite State Machine

The attention model as presented in Fig. 3 suggested the implementation of a state machine. This is a similar approach to Gollan and Ferscha [4]. While methods of machine learning generally have higher accuracy for such tasks this comes with certain drawbacks when compared to a rule-based system as discussed earlier. However, while the advantage of a state machine, namely easy configurability, is useful when setting up the system in different locations it also introduces a new problem: For each feature one or multiple thresholds have to be determined to optimize accuracy and specificity. Furthermore, implementations using methods of machine learning are able to handle noise better than simple linear thresholds. Therefore, we did expect a rule-based classifier to achieve the same accuracy as our SVM implementation, but, given the advantages, we would also be ready to accept a slightly worse accuracy.

To simplify parameterization of our state machine we use slightly different features than for the machine learning classifiers. Specifically, we calculate a single measure for gaze distance as the distance from the center of the focus object whereas the SVM and neural network classifiers are given the distances on the horizontal and vertical axis separately. Another difference is the use of a simple smoothing algorithm which suppresses state changes as a result of sensor noise. This is achieved by allowing a state change only after the majority of a certain number of consecutive frames (which is a parameter for our classifier) show a consistent changed state. This means that samples are no longer independent from each other as a previous sample has an effect on the prediction of the current frame, which is not the case for the other classifiers. However, in contrast to the other methods independent samples are no requirement for a state machine. Therefore, this doesn't pose a problem, but increases classification performance.

Table 5 shows the results of our state machine implementation on the museum dataset. As there is no need to hold back part of the dataset for verification, we used all 14177 frames. The offsets on the sensor plane to the center of the focus object is set to 25 cm to the left and 45 cm below the sensor. The smoothing algorithm was configured to consider the majority of the last 10 frames and the interaction range was set to 1.4 m as a simple threshold for gaze distance.

Table 5. FSM results on the museum dataset.

		Predicted			Sensitivity (TPR)	Specificity (TNR)
		Ignoring	Watching	Ready to interact		
Actual	Ignoring	11469	1238	292	88.23%	85.48%
	Watching	107	508	96	71.45%	90.72%
	Ready to interact	64	11	392	83.94%	97.17%
	Sum (frames)	14177			**Accuracy**	**87.25%**

Overall accuracy is 87.25% and therefore slightly short of our target of at least expected 90%. Threshold selection in this scenario is quite difficult and probably also inappropriate as thresholds tend to either under- or overestimate a subjects' attention state. Additionally, while we built in mechanisms to reduce the impact of sensor noise in the museum setting and also for the lab setting, our implementation is unable to compete with the SVM in this regard.

Again, in the more balanced and less noisy lab setting the results are quite a bit better and comparable to the results of the SVM implementation on the same dataset with an overall accuracy of 93.84% (see Table 6).

Table 6. FSM results on the lab dataset.

		Predicted			Sensitivity (TPR)	Specificity (TNR)
		Ignoring	Watching	Ready to interact		
Actual	Ignoring	4005	380	7	91.19%	96.67%
	Watching	122	3742	0	96.84%	91.92%
	Ready to interact	18	2	315	94.03%	99.92%
	Sum (frames)	8591			**Accuracy**	**93.84%**

6 Discussion

When looking at the results of our measurements by comparing prediction accuracy values for samples in different distance ranges it can be seen that the rule-based classifier FSM has difficulties predicting the class of samples in close proximity to the sensor. On

the museum dataset, for instance, samples in close proximity to the sensor were especially problematic for the state machine to predict correctly, as can be seen in Table 7.

Table 7. Performance of the FSM classifier on the museum dataset on samples between 0 and 2 m distance to the sensor.

		Predicted			Sensitivity (TPR)	Specificity (TNR)
		Ignoring	Watching	Ready to interact		
Actual	Ignoring	642	570	288	42.80%	91.12%
	Watching	26	424	96	77.66%	70.46%
	Ready to interact	64	11	392	83.94%	81.23%
	Sum (frames)	2513			**Accuracy**	**58.02%**

The machine learning based classifiers could handle these samples significantly better, indicating that a simple threshold is ill-suited to separate these samples. Results of the SVM classifier are shown in Table 8.

Table 8. Performance of the SVM classifier on the museum dataset on samples between 0 and 2 m distance to the sensor.

		Predicted			Sensitivity (TPR)	Specificity (TNR)
		Ignoring	Watching	Ready to interact		
Actual	Ignoring	716	54	6	92.27%	92.03%
	Watching	40	246	1	85.71%	94.55%
	Ready to interact	0	0	215	100.00%	99.34%
	Sum (frames)	1278			**Accuracy**	**92.10%**

Table 9. Performance of the MLP classifier on the museum dataset on samples between 0 and 2 m distance to the sensor.

		Predicted			Sensitivity (TPR)	Specificity (TNR)
		Ignoring	Watching	Ready to interact		
Actual	Ignoring	688	40	3	94.12%	91.79%
	Watching	40	225	1	84.59%	94.85%
	Ready to interact	0	9	212	95.93%	99.60%
	Sum (frames)	1218			**Accuracy**	**92.36%**

Performance of the MLP is similar to the SVM classifier, as can be seen in Table 9. This is true for all distance classes within the lab dataset.

On the lab dataset, this discrepancy in accuracy between 0 and 2 m vanishes and the FSM, SVM and MLP classifiers perform equally well (exact figures not shown here). This indicates that the FSM classifier in its current state has problems with the noisy environment of the museum setting. In the lab setting no clear advantage for either the FSM, SVM or MLP classifier emerges.

At medium distances (between 2 and 4 m) all three classifiers perform at equal accuracy levels, on the museum dataset with 92.32% accuracy for FSM, 97.56% for SVM and 98.38% for MLP, and on the lab dataset with 93.16% for FSM, 99.02% for SVM, and 99.06% for MLP. However, the samples in this distance class within the museum dataset are heavily skewed towards the *ignoring* class due to the experiment setup, distorting statements on the quality of our results. The samples within the lab dataset are more uniformly distributed across distance classes and therefore accuracy of the classifiers is evaluated on this dataset.

At far distances (>4 m) results are washed out. While we cannot provide accuracy values for the lab setup for this distance class (note that the experiment had been restricted to distances <4 m) figures for the museum setup range above 99% for all classifiers, which we believe is due to the unequal distribution of the state *ignoring* in this area (i.e. people were distantly passing the stimulus without recognizing it).

As a summary, our results show that the rule-based classifier, while delivering solid accuracy on our lab dataset, performs below expectations in the museum setting. As the features are the same for both our settings, we believe this to be a result of sub-optimal sensor placement. Additionally, the sensor we used had problems when groups gathered in front of the exhibit it was installed on. In our lab setting the sensor performed best with up to 4 subjects simultaneously, but deteriorated after that. This indicates that a better sensor would highly benefit our rule-based classifier. With several subjects in front of the exhibit the sensor was unable to track multiple targets reliably and would often lose track of subjects. This inhibited the smoothing algorithm of our rule based classifier and therefore introduced more noise to the results, which the SVM and MLP are obviously better at dealing with.

7 Conclusion

Estimating attention in public scenarios is growing more important as pervasive information systems become prevalent and previously manually controlled machinery becomes semi-automatic, deferring to human judgment only when necessary. Research initiatives in human-computer interaction such as *Raising Attention* [2] work towards establishing attention estimation as a tool in user experience design, in particular where information overload needs to be addressed.

Besides predominating machine-learning estimation techniques for determining human attention, we feel that the attempt of modeling a lightweight classifier immune to location changes of sensor or stimulus and without the necessity of training phases turned out to be a feasible alternative to ponderous machine learning approaches and merits

further research, especially comparing different sensors and techniques for extracting low level features (e.g. thresholds step functions). Particularly, in public display scenarios, training constitutes a significant effort, which could be avoided using FSMs.

At once, our "reduced" model of attention states is meant to provide a balance for practical application requirements: "*ignoring*", "*watching*" and "*ready to interact*" should already cover a wide range of scenarios, but an API designed around the model could also provide access to details of lower layers where necessary. The experimental results have shown that the three principal states are consistently separable.

References

1. Bulling, A.: Pervasive Attentive user interfaces. Computer **49**(1), 94–98 (2016)
2. Ferscha, A.: Raising attention. http://www.pervasive.jku.at/raisingattention/index.php/ra
3. Ferscha, A., Zia, K., Gollan, B.: Collective attention through public displays. In: International Conference on Self-Adaptive and Self-Organizing Systems, SASO, pp. 211–216. IEEE (2012). http://dx.doi.org/10.1109/SASO.2012.35
4. Gollan, B., Ferscha, A.: SEEV-effort - is it enough to model human attentional behavior in public display settings. In: Future Computing 2016, the Eighth International Conference on Future Computational Technologies and Applications (2016)
5. Gollan, B., Wally, B., Ferscha, A.: Automatic human attention estimation in an interactive system based on behavior analysis. In: Proceedings of the 15th Portuguese Conference on Artificial Intelligence (EPIA2011), pp. 978–989 (2011)
6. Kahneman, D.: Attention and effort. Am. J. Psychol. **88**(2), 339–340 (1973)
7. Kukka, H., Oja, H., Kostakos, V.: What makes you click: exploring visual signals to entice interaction on public displays. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2013), pp. 1699–1708. ACM (2013)
8. Mubin, O., Lashina, T., Loenen, E.: How not to become a buffoon in front of a shop window: a solution allowing natural head movement for interaction with a public display. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) INTERACT 2009. LNCS, vol. 5727, pp. 250–263. Springer, Heidelberg (2009). doi: 10.1007/978-3-642-03658-3_32
9. Muta, M., Masuko, S., Shinzato, K., Mujibiya, A.: Interactive study of WallSHOP: multiuser connectivity between public digital advertising and private devices for personalized shopping. In: Proceedings of the 4th International Symposium on Pervasive Displays (PerDis 2015), pp. 187–193. ACM, New York (2015)
10. Perry, M., Beckett, S., O'Hara, K., Subramanian, S.: WaveWindow: public, performative gestural interaction. In: ACM International Conference on Interactive Tabletops and Surfaces (ITS 2010), pp. 109–112. ACM, New York (2010)
11. Selker, T.: Visual attentive interfaces. BT Technol. J. **22**(4), 146–150 (2004)
12. Strong, E.K.: The Psychology of Selling and Advertising. McGraw-Hill, New York (1925)
13. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cogn. Psychol. **12**(1), 97–136 (1980)
14. Tsotsos, J., Rothenstein, A.: Computational models of visual attention. Scholarpedia **6**(1), 6201 (2011). http://dx.doi.org/10.4249/scholarpedia.6201
15. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer Science & Business Media, Heidelberg (1998)

16. Wang, M., Boring, S., Greenberg, S.: Proxemic peddler: a public advertising display that captures and preserves the attention of a passerby. In: Proceedings of the 2012 International Symposium on Pervasive Displays (PerDis 2012), 6 p. ACM, New York (2012). Article 3
17. Ward, L.: Attention. Scholarpedia **3**(10) (2008). http://dx.doi.org/10.4249/scholarpedia
18. Wickens, C., McCarley, J.: Applied Attention Theory. CRC Press, Boca Raton (2007)
19. Wolfe, J.M., Horowitz, T.S.: What attributes guide the deployment of visual attention and how do they do it. Nat. Rev. Neurosci. **5**(6), 495–501 (2005)
20. Zhang, Y., Bulling, A., Gellersen, H.: SideWays: a gaze interface for spontaneous interaction with situated displays. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2013), pp. 851–860. ACM, New York (2013). https://doi.org/10.1145/2470654.2470775