Modeling Cognitive Status through Automatic Scoring of a Digital Version of the Clock Drawing Test

Alexander Prange
German Research Center for Artificial Intelligence (DFKI)
Germany
alexander.prange@dfki.de

Daniel Sonntag
German Research Center for Artificial Intelligence (DFKI)
Germany
daniel.sonntag@dfki.de

ABSTRACT

The Clock Drawing Test is used as a cognitive assessment tool in geriatrics to detect signs of dementia or to model the progress of stroke recovery. The result is scored manually by a trained professional. We implement the Mendez scoring scheme and create a hierarchy of error categories that model the test characteristics of the clock drawing test, based on a set of impaired clock examples provided by a geriatrics clinic. Using a digital pen we recorded 120 clock samples for evaluating the automatic scoring system, with a total of 2400 error samples distributed over the 20 error classes of the Mendez scoring scheme. Error classes are scored automatically using a handwriting and gesture recognition framework. Results show that we provide a clinically relevant cognitive model for each subject. In addition, we heavily reduce the time spent on manual scoring. We compare manual scoring results with results produced by our automated system.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI)*; • **Applied computing** → *Health informatics.*

KEYWORDS

Clock Drawing Test, Cogntive Assessment, Digital Pen, Handwriting Features

ACM Reference Format:
Alexander Prange and Daniel Sonntag. 2019. Modeling Cognitive Status through Automatic Scoring of a Digital Version of the Clock Drawing Test. In *27th Conference on User Modeling, Adaptation and Personalization (UMAP '19), June 9–12, 2019, Larnaca, Cyprus.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3320435.3320452

1 INTRODUCTION

For more than 50 years, the Clock Drawing Test (CDT) is used as an assessment tool for cognitive impairment. It is a simple paper and pencil test in which the participant is asked to draw a clock face and indicate a certain time. The task is primarily designed to test the visuospatial ability and is often used in geriatrics to screen for signs of dementia, such as Alzheimer's disease, or other neurologic conditions, including Parkinson's disease, traumatic brain injuries, and stroke recovery. Usually a trained professional observes the clock drawing task and scores the final sketch based on a scoring scheme, which takes up to a few minutes.

An automatic scoring system has several benefits: first, it significantly reduces the time caregivers have to invest into administering the CDT; and second, it is likely to produce more objective scores and potentially enables a more detailed analysis [24]. Not all scoring schemes are equally well suited for automation, since most of them have been designed to be quick and easy to be interpreted by human testers. We selected the 20 point Clock Drawing Interpretation Scale (CDIS) by Mendez et al. [11], which is well suited for automation because it contains clear test parameters that can be modelled mathematically and computationally. In addition, the manual scoring procedure of CDIS is very time-consuming and would highly benefit from automatic computation.

The CDIS contains items such as "All numbers 1-12 are present", which are to be rated 0 if not fulfilled and 1 if satisified. All 20 individual scores are then added up and the final score indicates the severity of cognitive impairment. For example, a score of less than 18 is likely to indicate Alzheimer's. Based on CDIS and real clock drawings (provided by a geriatrics daycare clinic of a large hospital), we develop a categorization of individual scores subjects are likely to make during the CDT. We create a direct mapping from the error classes to the scoring parameters of the selected scoring scheme (1 or 0 scores). A digital pen is used to record clock sketches, which are then analyzed by using a set of handwriting and multi-stroke features, as well as classifiers based on heuristics and the semantic parameters of the clock drawing task. We compute the score per item and the sum of all items, which equals the final CDIS score.

The paper is structured as follows. Section 2 describes related work. Based on the original Mendez scoring scheme and real clock drawings we create a structure of error classes and compute a value for each of these classes, which we present in section 3. Because of the mapping to the scoring scheme, we are therefore able to provide a detailed automatic assessment of the drawn clock according to CDIS (section 3). The resulting technical architecture is presented in section 4. In order to evaluate how well our system predicts scores towards its usage in clinical routine, we design and conduct a lab experiment, in which we let cognitively healthy subjects simulate error classes. Participants are presented with a random selection of two error examples and are then asked to sketch a clock that satisfies both error classes using a digital pen. In addition we let two human expert raters with clinical background score each clock based on the original scoring scheme (section 5). We present the results of our evaluation (section 6) and provide a further discussion (section 7) and a conclusion (section 8).

UMAP '19, June 9–12, 2019, Larnaca, Cyprus

ACM ISBN 978-1-4503-6021-0/19/06…$15.00
https://doi.org/10.1145/3320435.3320452

2 BACKGROUND AND RELATED WORK

Digitizing and comparing paper-and-pencil assessments has been introduced recently [25]. Clinically relevant examples of cognitive assessments include the Rey-Osterrieth Complex Figure (ROCF) test [1], which can be used for various purposes, such as diagnosing the periphery [2]. During the ROCF, participants are asked to copy a complex geometrical figure while looking at the template. Then the template is taken away and the figure has to be recalled and sketched again. After 30 minutes, participants are asked to sketch what they still recall. Another very popular assessment tool is the Mini-Mental State Examination (MMSE) introduced in 1975 by Folstein et al. [6]. The MMSE is a 30-point questionnaire, which is administered by a trained professional, who leads the subject through the questionnaire, while taking notes. Afterwards, the trained professional manually evaluates the results, based on his notes and a predefined scoring scheme. Two tasks of the questionnaire require the subject to perform handwriting or sketching. Administration takes on average 5 to 10 minutes. Due to its standardization, validity, short administration period, and ease of use, the MMSE is widely applied as a screening tool for dementia [7]. The CDT is often administered together with the MMSE, because both assess different cognitive abilities [14]. Recent work by Niemann et al. [18] shows how the MMSE and its sketched parts can be automatically recognized using speech and handwriting gesture recognition. Compared to other cognitive assessment tools, the CDT is well accepted by patients and practitioners, because of its easy and quick application. It is being used worldwide for decades as a neuropsychological screening test and by now there exists a plentitude of variations and scoring schemes. Different cognitive abilities, including attention, comprehension, memory, visuospatial skills, motor and executive functions, can be measured using the CDT [11]. Factors such as age, education and language are known to influence the performance [16]. Although the majority of practitioners are familiar with the CDT and can distinguish a heavily impaired clock from a correct one without specific scoring criteria, it is the detailed analysis of sketch characteristics as represented by scoring schemes that allow us to identify disturbances with diagnostic sensitivity and specificity. The scoring schemes differ greatly in the complexity of the scoring procedure, which in turn has an effect on their application in the clinical domain. Some contain many and very detailed parameters that have to be evaluated manually and are therefore time consuming when scored by the human rater. Others contain rather unspecific descriptions and examples of how a clock has to look like in order to be considered an impaired clock.

In order to analyze the recorded handwritten strokes of CDT, we make use of so called handwriting features, which are mathematical representations of several geometrical aspects of the sketched input. Traditionally, stroke level features are most often used for statistical gesture recognition. One of the most prominent set of features was presented by Dean Rubine in 1991 [21]. It contains a total of 13 features that have been designed to reflect the visual appearance of strokes in order to be used in a gesture recognizer. More recent work by Don J.M. Willems and Ralph Niels [30] defines a total of 89 features using formal mathematical descriptions and algorithms. Adrien Delaye and Eric Anquetil introduced the HBF49 Feature Set [3], which contains 49 features and was specifically designed for different sets of symbols and as reference for evaluating symbol recognition systems. We also include14 features described by Sonntag et al. [26] to distinguish between written text and other types of gestures in handwriting recognition. In total, we include over 100 handwriting features implemented by Prange et al. [17] to predict the error classes of the CDT.

3 CATEGORIZATION OF ERROR CLASSES

Following Ehreke et al. [5] and Patocskai et al. [15] we distinguish between two types of scoring schemes, namely quantitative and qualitative evaluations. They considered quantitative analyses as those represented by numerical scales, whereas qualitative approaches classify the drawing of the clock based on descriptions of typical errors by considering the whole clock in their analysis and using a subjective approach [5, 15]. Considering the automation of scoring schemes we determined that not all of them are equally well suited to be computed automatically. Especially qualitative approaches, which are susceptible to subjective ratings, have proven to be difficult to be modelled computationally, whereas the numerical scales used in quantitative approaches are much more suitable. In order to find an appropriate scoring, we examine several of the most popular schemes that are used in daily practice and clinical research. Our selection of potential scoring schemes is based on expert interviews, a recent study conducted by Spenciere et al. [27] and a study presented by Tuokko et al. [28]. We consider a total of eight scoring schemes, the qualitative approaches by Libon et al. [8], the mixed approach by Rouleau et al. [19] and the quantitative scorings by Shulman et al. (both versions 1986 and 1993) [22, 23], Royall et al. [20], Manos and Wu [9], Watson et al. [29] and Mendez et al. [11].

Considering our selection of scoring schemes and real-world examples of impaired clock sketches, taken from the geriatrics daycare clinic of a large hospital, we categorize common mistakes in a hierarchy of error classes summarized in table 1. The main differentiation is done between errors concerning the numbers of the clock-face and the presence of hands indicating a specific time. Error classes are not necessarily mutually exclusive, e.g., "All numbers missing." will always result in "Digits are partially missing." being true as well. Vice versa if "All numbers missing." is true other categories such as digit misplacement "Digits are placed counterclockwise." cannot be satisfied at the same time. We do not consider this to be a problem, because several scoring schemes are designed this way and real-world impaired sketches rarely trigger only a single condition.

After careful consideration, we decided to focus on the Mendez et al. [11] CDIS scoring scheme, a 20 point quantitative scale with detailed conditions compared to other scorings. For example, the Shulman [22] scoring consists of six categories and the clock sketch has to be interpreted as belonging to one, category 1 is a "perfect" clock and category 6 is described as "No reasonable representation of a clock". The categories in between are described vaguely (e.g., category 2 - "Minor visuospatial errors") including examples, such as "Mildly impaired spacing of times". Such categories leave room for subjective interpretation and thus potentially lead to human bias based on the experience of the tester. Having no clear cut seperations between categories can result in one tester rating a clock

Table 1: Categorization of typical error classes based on existing CDT scoring schemes and real-world examples.

Digits	Missing	All missing
		Only 4 main (12, 3, 6, 9) present
		Partially missing
	Misplacement	Wrong angle in regard to center
		Counter-clockwise direction
		Varying distance to circle
		Varying spacing between
		Not inside circle
		Upside down
	Duplicates	Same number
		Synonyms (e.g. 14 and 2)
	Other	Roman numerals
		Slurred/Unreadable
		Not in range (1-12)
		Substitution for words
		Markings instead
Hands	Missing	Both are missing
		One is missing
	Misplacement	Outside circle
		Not connected to center
		Longer indicates wrong minutes
		Shorter indicates wrong hour
		No attempt to indicate time
	Length	Both hands same length
	Other	More than 2 hands
Other	No attempt to indicate a time	
	Correction gestures (e.g., crossing out)	
	Markings and helping lines dominate the clock face	
	Text written (e.g., "11:10")	
	Circle is re-traced	

as midly impaired, while another one considers the same sketch as belonging the the moderately impaired category. In comparison, the Mendez score items are much more detailed and most of them are easy to interpret, e.g., "There are no repeated or duplicated number symbols.". A complete overview of the CDIS scoring can be found in table 2. Another reason to select the Mendez scoring scheme is that, because of its size, it includes the majority of conditions that are also present in other scoring schemes. These properties make the CDIS a very good candidate for automatically detecting and interpreting its error classes based on multi-stroke characteristics.

4 TECHNICAL ARCHITECTURE

We record clock sketches using a digital pen and paper imprinted with a nearly invisible microdot pattern. The NeoSmartpen N2[1] is a ballpoint pen with an intergrated infrared camera near the tip, which recognizes the microdot pattern on the paper and records the exact position, timestamp and pressure of the pen. It allows us to analyze the digital ink including temporal aspects and time stamps (which is not possible with a pure image recognition approach). The technical architecture is shown in figure 1. First, we

[1]https://www.neosmartpen.com/

Table 2: The original *Clock Drawing Interpretation Scale (CDIS)* by Mendez et al. [11] consists of 20 items, where each item scores 0 or 1. The sum is used as the total score.

Item	Description
1	There is an attempt to indicate a time in any way.
2	All marks or items can be classified as either part of a closure figure, a hand, or a symbol for clock numbers.
3	There is a totally closed figure without gaps.
4	A "2" is present and is pointed out in some way for the time.
5	Most symbols are distributed as a circle without major gaps.
6	Three or more clock quadrants have one or more appropriate numbers: 12-3, 3-6, 6-9, 9-12 per respective clockwise quadrant.
7	Most symbols are ordered in a clockwise or rightward direction.
8	All symbols are totally wihtin a closure figure.
9	An "11" is present and is pointed out in some way for the time.
10	All numbers 1-12 are indicated.
11	There are no repeated or duplicated number symbols.
12	There are no substitutions for Arabic or Roman numerals.
13	The numbers do not go beyond the number 12.
14	All symbols lie about equally adjacent to a closure figure edge.
15	Seven or more of the same symbol type are ordered sequentially.
16	All hands radiate from the direction of a closure figure center.
17	One hand is visibly longer than another hand.
18	There are exactly two distinct and separable hands.
19	All hands are totally wihtin a closure figure.
20	There is an attempt to indicate a time with one or more hands.

annotate each stroke with a corresponding gesture type (such as circle, hand, number etc.). Second, we compute the values of the 20 error classes according to Mendez et al. [11]. In order to compute the CDT error classes, we create 20 rule-based classifiers corresponding to the conditions of the CDIS scoring. We employ a set of stroke-level *syntactic features*, which cover geometrical properties of the sketched strokes, such as length, curvature or compactness (see categorization of error classes). The second type of features are called *semantic features* and describe task-dependent properties. For example, the distance to the clock center, or the location of a number inside a clock quadrant are used as semantic features for the CDIS classes. We then summarize individual scores and receive the final CDIS score, which indicates the overall level of cognitive impairment indicated by the drawing.

The system differentiates between 7 gesture types a stroke can belong to: part of the circle, digit, hand, center, helping line, text and unknown/uncategorized. In an iterative streaming process, we look

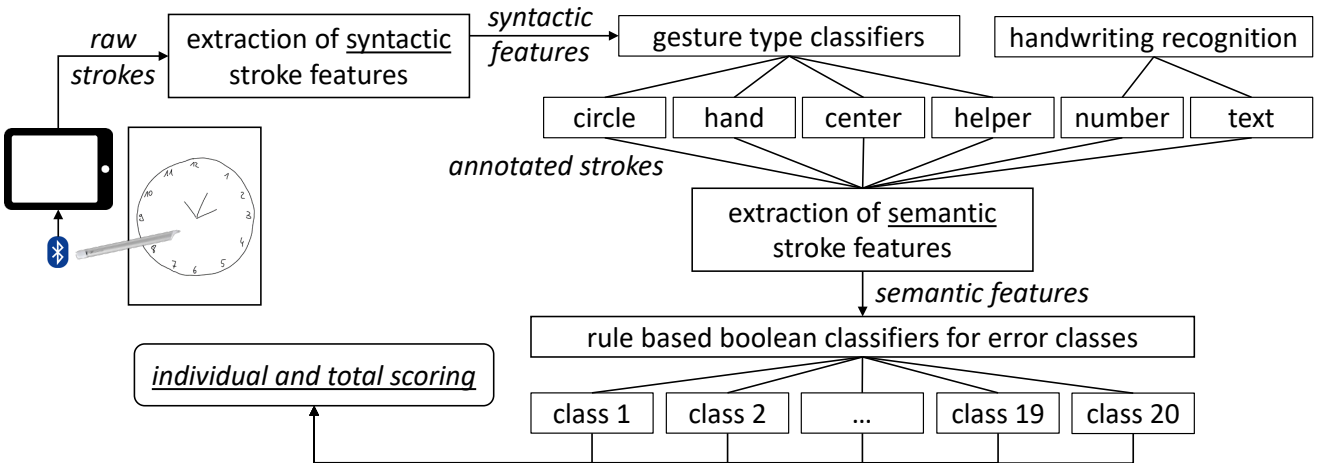


Figure 1: Architecture of our classification system for the automatic prediction of error classes as described in section 4.

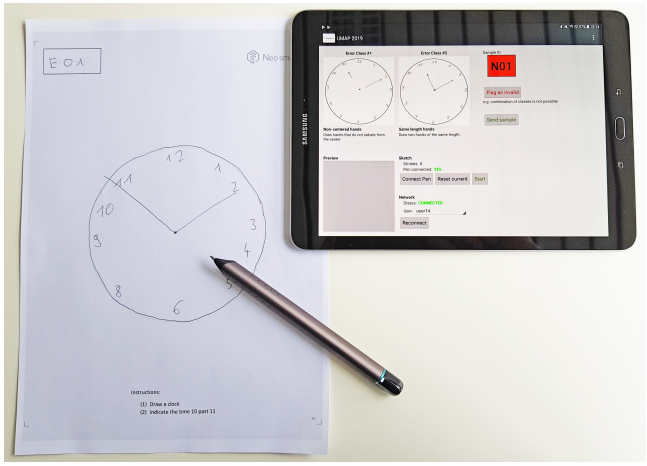


Figure 2: Picture of the experiment setup with touch-screen tablet, digital pen and paper.

at stroke patterns that match to one or more of the gesture classes based on the stroke level features. For example, a straight line has the same initial and final directional vector as a circle, whereas parts of the circle show a distinct curvature. Other gestures are classified based on their spatial location, e.g., markings on the outer circle are classified as being helper markings. For the detection of digits and text we use a commercial handwriting recognition engine by MyScript[2]. Example samples before and after annotation are depicted in figure 4. An intuitive example is error class 10 ("All numbers 1-12 are indicated."), where we iterate through all strokes classified as numbers and check if all are present (figure 4d). Other error classes require the use of additional stroke features, such as error class 18 ("There are exactly two distinct and separable hands."). Some error classes, e.g., class 14 ("All symbols lie about equally adjacent to a closure figure edge.") leave room for interpretation.

[2]https://www.myscript.com/

These error classes are implemented using a mixture of thresholds and mean values.

5 EVALUATION

In order to measure the performance of our automated system, we design and conduct an experiment in which we ask participants to draw impaired clocks on purpose, based on our previously investigated error classes that are common in the CDT. A total of 12 subjects (3 female, 8 male) participated in our study, ranging from 19 to 60 years of age. Subjects were recruited from university, including co-workers and students from varying domains and fields of work, and none where diagnosed with cognitive impairments. We consider the following questions:

- Is the experiment design reasonable: Are human raters able to correctly predict the error classes that participants were asked to draw?
- Are the parameters of the scoring scheme well-defined?
- Do human raters disagree?
- How accurately does the system predict error classes, how is the performance compared to human raters?
- For which classes do medical experts and the system produce similar scores, for which not and why?

5.1 Apparatus & Procedure

Our experiment setup consists of a touch-screen tablet and a digital pen and paper as shown in figure 2. The tablet is used to present instructions and for recording the data which is streamed by the pen via bluetooth. Participants sit in a distraction-free room at a table with the pen and paper in front of them. After being instructed about the task, subjects were asked to draw several clocks and answer a short demographics questionnaire. Each clock was drawn on a seperate sheet of paper and marked by adding a pseudo-anonymous identifier. The sheets were then scanned and distributed to two expert raters from the Charité in Berlin, with profound background knowledge about scoring schemes and the CDIS scoring

scheme [11] in particular. The results of the human expert scorings are then compared to the predictions made by the automated system.

5.2 Task Design

Participants are asked to draw clocks based on the original instructions of the Mendez [11] version of the CDT. They are instructed to draw a clock-face first and then indicate the time "10 past 11". Based on the error classes we create a set of instructions (see table 3) and visual examples for each of the 20 items (see figure **??**). For each clock the subject is given two distinct and randomly selected error classes and is instructed to combine both of them into the sketch to produce a single artificially impaired clock sketch. The descriptions and example clocks for the two errors are displayed on the touch-screen tablet. For example, if instructed to combine error classes 8 ("Outside the circle - Write any symbols outside the circle.") and 13 ("Beyond 12 - Include numbers beyond 12."), the resulting clock should look like this: instead of number 1-12 the clock has, e.g., numbers 13-24 and the numbers are written outside the circle, instead of on the inside. Another example can be seen in figure 4c, which combines the error classes 14 ("Varying distance - Vary the distance of symbols to the circle.") and 17 ("Same length hands - Draw two hands of the same length."). Subjects are able to skip a sample if they have the impression that they are unable to combine both errors into one sketch. We decided to ask subjects for a combination of two classes, because they have a higher probability to respresent real-world examples and after a pilot study we determined that combining more than two error classes would be too challenging for the participants. Each subject was asked to draw a total of 10 clocks, covering all 20 error classes per participant.

6 RESULTS

We analyze a total of 120 clock sketches while covering each error class with at least 12 positive samples. Considering the task design subjects flagged 5 combinations of error classes as invalid: (1 & 4), (1 & 17), (1 & 18), (1 & 20) and (17 & 20). For each of the 120 clock sketches we have 2 manual scorings performed by human expert raters as the ground truth. Ideally we would expect a distribution of samples where for each error class we have the same amount of positive and negative samples. But since error classes are not mutually exclusive, we get a slightly distorted distribution (figure 5). Based on the experiment design, at least 10% (12/120) positive examples per class would be anticipated, but as can be seen most classes have a clear imbalance in favour of negative samples, two have even less than 10% positive samples.

We now compare the ground truth classes with the classes the system computes. We compare the true positives (the drawn class was correctly recognized by the rater) and the false negatives (the rater did not recognize the class that was asked to be drawn) for human raters and the system. Based on that, we show the sensitivity and miss rate values in table 4. Human raters achieve on average a miss rate of 17% compared to the system with 28%. Notably, class 6 ("Three or more clock quadrants have one or more appropriate numbers") has a miss rate of 91% for human raters. This class and class 15 ("Seven or more of the same symbol type are ordered sequentially.") reach a 100% miss rate for the automated system. This means that for these classes neither the human raters nor the system was predicting the same error classes that were asked to be drawn by participants during the experiment.

Next we calculate the interrater reliability by comparing the individual scores across all 20 classes for each of the samples. Per sample we check for each class if a rater disagrees on the scoring with one of the others (human 1, human 2 and system). As we have established previously, the error classes that participants were asked to draw do not necessarily match the scores produced by the raters. This is why we look closer at cases where both human raters agree on the scoring, but where the system produces a different value. The results of these comparisons are displayed in table 5 and visualized in figure 6. Having calculated the overall percentage agreement we find that human raters agree in 93.1% and all 3 raters agree in 92.4% of cases. Considering only the instances where both humans agree, the system scores the same value per class on 82.6%. Human raters disagree in one-third of cases for classes 2 ("All marks or items can be classified as either part of a closure figure, a hand, or a symbol for clock numbers.") and 5 ("Most symbols are distributed as a circle without major gaps."). We calculate the interrater reliability using Cohen's kappa [10]. Comparing human raters the kappa value is $\kappa = 0.81$, while for human 1 and the system the value is $\kappa = 0.50$, and for human 2 and the system $\kappa = 0.49$. From the individual scores per class we can calculate the sum of scores which gives us the final CDIS value.

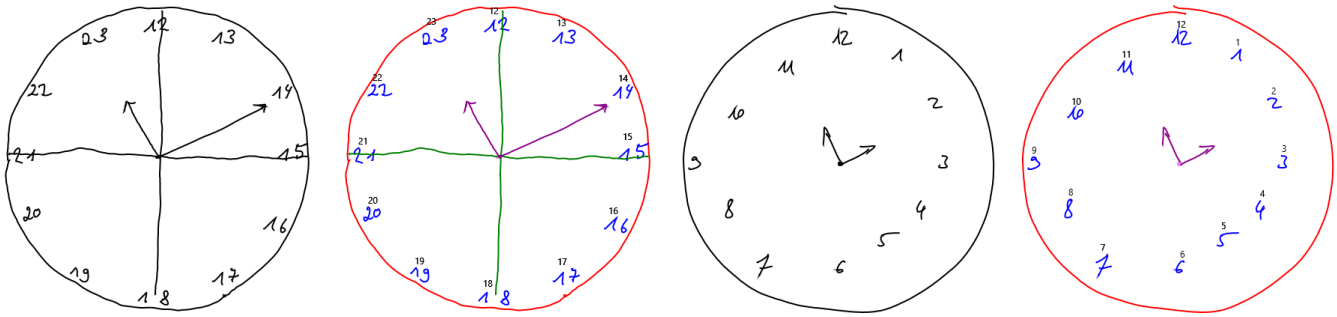
7 DISCUSSION

As explained above, based on the experiment design one would expect a distribution of at least 10% negative samples for each scoring item (see figure 5). Looking at the samples and scorings of classes 1 and 14, we determined that the instructions for drawing these errors leave enough room for interpretation so that the sketch is biased by the subjective interpretation of "indicate a time in any way" and "about equally distributed". Noticeably, the disagreement rates in table 5 show that raters agreed in the majority of cases for classes 1 and 14, resulting in a discrepancy in how drawing participants interpreted the instructions and how raters interpreted the scoring. What we see here is a prime example of the problem of qualitative scoring approaches, namely the divergence between interpretation of vaguely described conditions. We conclude from this that these instructions need to be clearer and more concise, leaving less room for subjective interpretation.

By looking at table 4 we see that classes 1, 6 and 20 have a high miss rate, meaning that raters where not able to detect this error in a sample even though the participant who sketched the sample was asked to add that error. For class 1 and 20 we have identified the above discussed subjective interpretation as the root cause, in which the phrase "an attempt" can be interpreted openly by raters. However, for class 6 the drawing instructions stated to "Create at least two quadrants with incorrect or missing numbers inside." in contrast to the scoring instruction "Three or more clock quadrants have one or more appropriate numbers: 12-3, 3-6, 6-9, 9-12 per respective clockwise quadrant.". Having compared the samples with the scores we conclude that our instruction should have been more strict in requiring the exact amount of missing numbers. Analyzing the other end of the scale we can see that

Table 3: Drawing instructions per error class as given to participants. The error classes correspond to the items of the Mendez scoring scheme (e.g., error class 10 corresponds to the Mendez scoring item "All numbers 1-12 are indicated.").

Error Class	Error Title	Error Description
1	No indication of time	Do not indicate the time in any way, no hands, no circling of numbers etc.
2	Superfluous symbols	Add additional markings/symbols which are not part of the clock.
3	Unclosed figure	Draw an incomplete circle, e.g. with one or more gaps.
4	Missing minute	Do not add a 2 or do not indicate it as part of the time.
5	Varying spacing	Vary the spacing between symbols.
6	Incorrect quadrant	Create at least two quadrants with incorrect or missing numbers inside.
7	Counter-clockwise	Order symbols counter-clockwise.
8	Outside the circle	Write any symbols outside the circle.
9	Missing hour	Do not add a 11 or do not indicate it as part of the time.
10	Missing numbers	Leave out at least one number.
11	Duplicates	Add at least one number symbol twice.
12	Substitution	Use substitutions for Arabic or Roman numerals.
13	Beyond 12	Include numbers beyond 12.
14	Varying distance	Vary the distance of symbols to the circle.
15	Non-sequential	Order at least 6 symbols non-sequentially.
16	Non-centered hands	Draw hands that do not radiate from the center.
17	Same length hands	Draw two hands of the same length.
18	Multiple or not separable hands	Draw more than two or not separable hands.
19	Hands outside	Draw at least one hand completly or partially outside the circle.
20	No time	Do not indicate the time using hands. You may indicate the time otherwise.



(a) B01 original. (b) B01 analyzed. (c) F09 original. (d) F09 analyzed.

Figure 4: Selection of produced clocks as drawn by the participants (plain) and after being analyzed by our automated system (analyzed). Red indicates the stroke was interpreted as part of the circle, magenta indicates hands, while numbers are blue and superflous symbols are marked in green. In sample B01 the participant was asked to draw error classes 2 (superflous symbols) and 13 (numbers beyond 12), while sample F09 contains error classes 14 (varying distance) and 17 (same length hands).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
positive	92.5%	71.3%	88.8%	65.0%	74.2%	81.7%	80.8%	88.8%	58.8%	62.5%	89.2%	90.0%	89.2%	91.7%	80.0%	66.3%	61.7%	69.2%	73.8%	82.9%
negative	7.5%	28.8%	11.3%	35.0%	25.8%	18.3%	19.2%	11.3%	41.3%	37.5%	10.8%	10.0%	10.8%	8.3%	20.0%	33.8%	38.3%	30.8%	26.3%	17.1%

Figure 5: Distribution of samples per error class, based on the average scoring of human raters.

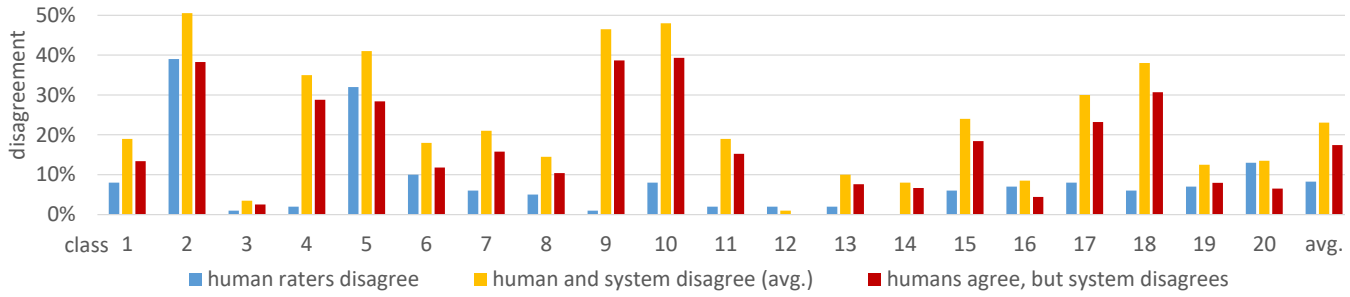


Figure 6: Visualization of the comparison of disagreement between raters, based on table 5.

Table 4: Comparison of samples between error classes that were asked to be drawn during the experiment and error classes that were recognized later by human raters and the automated system. Based on true positives and false negatives we calculate the sensitivity/recall and the miss rate.

class	human sensitivity	system sensitivity	human miss rate	system miss rate
1	0.58	1.00	0.42	0.00
2	0.92	0.92	0.08	0.08
3	1.00	0.75	0.00	0.25
4	0.75	0.42	0.25	0.58
5	0.75	0.17	0.25	0.83
6	0.09	0.00	0.91	1.00
7	0.88	1.00	0.12	0.00
8	0.92	1.00	0.08	0.00
9	1.00	0.75	0.00	0.25
10	0.96	1.00	0.04	0.00
11	0.96	1.00	0.04	0.00
12	0.92	0.92	0.08	0.08
13	1.00	1.00	0.00	0.00
14	0.75	0.67	0.25	0.33
15	0.96	0.00	0.04	1.00
16	0.92	0.83	0.08	0.17
17	0.88	0.42	0.12	0.58
18	0.96	0.92	0.04	0.08
19	0.92	0.83	0.08	0.17
20	0.55	0.83	0.45	0.17
avg.	0.83	0.72	0.17	0.28

Table 5: Comparison per sample and class on how strongly raters disagree in their scoring. We consider all 12 participants, each with 10 clock samples and 20 CDIS scoring items (n=2400). Human raters are indicated as human 1 and human 2. Note that the last column only considers instances where both human raters agree.

class	human 1 and 2	human 1 and system	human 2 and system	both humans and system
1	8%	23%	15%	13.39%
2	39%	39%	62%	38.27%
3	1%	4%	3%	2.52%
4	2%	35%	35%	28.81%
5	32%	44%	38%	28.41%
6	10%	18%	18%	11.82%
7	6%	22%	20%	15.79%
8	5%	14%	15%	10.43%
9	1%	46%	47%	38.66%
10	8%	46%	50%	39.29%
11	2%	19%	19%	15.25%
12	2%	1%	1%	0%
13	2%	9%	11%	7.63%
14	0%	8%	8%	6.67%
15	6%	23%	25%	18.42%
16	7%	7%	10%	4.42%
17	8%	32%	28%	23.21%
18	6%	39%	37%	30.70%
19	7%	15%	10%	7.96%
20	13%	20%	7%	6.54%

classes 3 ("There is a totally closed figure without gaps.") and 9 ("An "11" is present and is pointed out in some way for the time.") show a miss rate of 1% for the human raters, meaning that for these classes the raters scored almost exactly what participants intended to draw. For class 14 ("All symbols lie about equally adjacent to a closure figure edge.") human raters even agree in all of the cases. We conclude that if the scoring and drawing instructions are phrased carefully and leave little room for interpretation our experiment design can be considered reasonable.

We consider for which classes human raters and the system produce similar scores, for which not and why. As we can see from table 5 and figure 6, there are several classes were both humans and the system score very similarly, such as classes 3, 12, 13, 14 and 16. The discrepancies in classes 9 and 10 can be explained with the error rate of the number recognizer, which often does not recognize the 11 correctly. Class 2 is difficult because of the subjectiveness of interpreting which symbols can be counted as belonging to a clock an which not, e.g., are helping lines and markings part of a clock? If we look closer at classes 2 and 5, we can see that the automated system has to cope with disagreement between human expert annotations (39% for class 2 and 32% for class 5). Since we consider only cases where both humans agree (high probability of the error class actually being present) the automated system cannot score better than the disagreement rate between human raters. All in all our automated system reaches over 82% accuracy in predicting the same error classes as human raters.

8 CONCLUSION

Based on real world examples and existing scoring schemes we have categorized common types of errors that can happen during the CDT and have presented an experiment that evaluates the performance of our automated system in comparison to human raters. Such systems have the potential to significantly reduce the time of experts spent on manual scoring and reduce the influence of human bias on the scoring results, making the assessment of cognitive impairment more objective.

One of the limitations is that our current classifiers are manually crafted. We are currently working on reducing this effort by creating interactive machine learning models that can predict these error classes automatically. As not all existing scoring schemes are equally suitable to be automated, we are also investigating if and how the results of qualitative classification approaches, as presented here, can be mapped to quantitative approaches, which are less susceptible to subjective interpretation. Our system is currently in the process of being deployed for evaluation in the geriatrics daycare clinic of a large hospital. Ink features cannot only be used for gesture or sketch recognition, but also for characterisation of handwriting behaviour. Drotar et al. [4] have shown that the analysis of in-air movement can be used as a marker for Parkinson's disease. As cogntive tests often consist of more than one modality, e.g., speech and writing, we are also investigating how ink features can be used in multimodal scenarios [12, 13], where they may enhance the prediction of cognitive and emotional states [31].

Acknowledgements: This research is part of the Intera-KT project, which is supported by the Federal Ministry of Education and Research (BMBF) under grant number 16SV7768. Thanks go out to the anonymous reviewers, and Anika Steinert and Antje Latendorf from the Charité in Berlin.

REFERENCES

[1] R O Canham, S L Smith, and A M Tyrrell. 2000. *Automated scoring of a neuropsychological test: The Rey Osterrieth Complex Figure.* IEEE COMPUTER SOC, A406–A413.

[2] Daniel R. Coates, Johan Wagemans, and Bilge Sayim. 2017. Diagnosing the Periphery: Using the Rey-Osterrieth Complex Figure Drawing Test to Characterize Peripheral Visual Function. In *i-Perception.*

[3] Adrien Delaye and Eric Anquetil. 2013. HBF49 feature set: A first unified baseline for online symbol recognition. *Pattern Recognition* 46, 1 (Jan. 2013), 117–130. https://doi.org/10.1016/j.patcog.2012.07.015

[4] Peter Drotar, Jiri Mekyska, Irena Rektorova, Lucia Masarova, Zdenek Smekal, and Marcos Faundez-Zanuy. 2014. Analysis of in-air movement in handwriting: A novel marker for Parkinson's disease. *Computer Methods and Programs in Biomedicine* 117, 3 (2014), 405 – 411. https://doi.org/10.1016/j.cmpb.2014.08.007

[5] L. Ehreke, T. Luck, M. Luppa, H. H. Konig, A. Villringer, and S. G. Riedel-Heller. 2011. Clock drawing test - screening utility for mild cognitive impairment according to different scoring systems: results of the Leipzig Longitudinal Study of the Aged (LEILA 75+). *Int Psychogeriatr* 23, 10 (Dec 2011), 1592–1601.

[6] M. F. Folstein, S. E. Folstein, and P. R. McHugh. 1975. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* 12, 3 (Nov 1975), 189–198.

[7] Lindy E. Harrell, Daniel Marson, Anjan Chatterjee, and Jo Ann Parrish. 2000. The Severe Mini-Mental State Examination: A New Neuropsychologic Instrument for the Bedside Assessment of Severely Impaired Patients With Alzheimer Disease. *Alzheimer Disease & Associated Disorders* 14, 3 (2000).

[8] D. J. Libon, R. A. Swenson, E. J. Barnoski, and L. P. Sands. 1993. Clock drawing as an assessment tool for dementia. *Arch Clin Neuropsychol* 8, 5 (Oct 1993), 405–415.

[9] Peter J. Manos and Rae Wu. 1994. The Ten Point Clock Test: A Quick Screen and Grading Method for Cognitive Impairment in Medical and Surgical Patients. *The International Journal of Psychiatry in Medicine* 24, 3 (1994), 229–244. https://doi.org/10.2190/5A0F-936P-VG8N-0F5R arXiv:https://doi.org/10.2190/5A0F-936P-VG8N-0F5R PMID: 7890481.

[10] M. L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 22, 3 (2012), 276–282.

[11] Mario F. Mendez, Thomas Ala, and Kara L. Underwood. 1992. Development of Scoring Criteria for the Clock Drawing Task in Alzheimer's Disease. *Journal of the American Geriatrics Society* 40, 11 (1992), 1095–1099. https://doi.org/10.1111/j.1532-5415.1992.tb01796.x

[12] Mira Niemann, Alexander Prange, and Daniel Sonntag. 2018. Towards a Multimodal Multisensory Cognitive Assessment Framework. In *31st IEEE International Symposium on Computer-Based Medical Systems, CBMS 2018, Karlstad, Sweden, June 18-21, 2018.* 24–29. https://doi.org/10.1109/CBMS.2018.00012

[13] Sharon Oviatt, Björn Schuller, Philip R. Cohen, Daniel Sonntag, Gerasimos Potamianos, and Antonio Krüger (Eds.). 2017. *The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations - Volume 1.* Vol. Volume 1. Association for Computing Machinery and Morgan Claypool, New York, NY, USA.

[14] D. Palsetia, G. P. Rao, S. C. Tiwari, P. Lodha, and A. De Sousa. 2018. The Clock Drawing Test versus Mini-mental Status Examination as a Screening Tool for Dementia: A Clinical Comparison. *Indian J Psychol Med* 40, 1 (2018), 1–10.

[15] A. T. Patocskai, M. Pakaski, G. Vincze, M. Fullajtar, I. Szimjanovszki, G. Drotos, K. Boda, Z. Janka, and J. Kalman. 2014. Is there any difference between the findings of Clock Drawing Tests if the clocks show different times? *J. Alzheimers Dis.* 39, 4 (2014), 749–757.

[16] E. Pinto and R. Peters. 2009. Literature review of the Clock Drawing Test as a tool for cognitive screening. *Dement Geriatr Cogn Disord* 27, 3 (2009), 201–213.

[17] Alexander Prange, Michael Barz, and Daniel Sonntag. 2018. A categorisation and implementation of digital pen features for behaviour characterisation. *CoRR* abs/1810.03970 (2018). arXiv:1709.01796 https://arxiv.org/abs/1810.03970

[18] Alexander Prange, Mira Niemann, Antje Latendorf, Anika Steinert, and Daniel Sonntag. 2019. Multimodal speech-based dialogue for the Mini-Mental State Examination. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* ACM, New York, NY, USA, forthcoming.

[19] Isabelle Rouleau, David P Salmon, Nelson Butters, Colleen Kennedy, and Katheryn McGuire. 1992. Quantitative and qualitative analyses of clock drawings in Alzheimer's and Huntington's disease. *Brain and Cognition* 18, 1 (1992), 70 – 87. https://doi.org/10.1016/0278-2626(92)90112-Y

[20] D. R. Royall, J. A. Cordes, and M. Polk. 1998. CLOX: an executive clock drawing task. *J. Neurol. Neurosurg. Psychiatry* 64, 5 (May 1998), 588–594.

[21] Dean Rubine. 1991. Specifying Gestures by Example. In *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '91).* ACM, New York, NY, USA, 329–337. https://doi.org/10.1145/122718.122753

[22] Kenneth I. Shulman, Dolores Pushkar Gold, Carole A. Cohen, and Carla A. Zucchero. 1993. Clock-drawing and dementia in the community: A longitudinal study. *International Journal of Geriatric Psychiatry* 8, 6 (1993), 487–496. https://doi.org/10.1002/gps.930080606

[23] Kenneth I. Shulman, Ralph Shedletsky, and Ivan L. Silver. 1986. The challenge of time: Clock-drawing and cognitive function in the elderly. *International Journal of Geriatric Psychiatry* 1, 2 (1986), 135–140. https://doi.org/10.1002/gps.930010209 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/gps.930010209

[24] Daniel Sonntag. 2017. Interakt - A Multimodal Multisensory Interactive Cognitive Assessment Tool. *CoRR* abs/1709.01796 (2017). arXiv:1709.01796 http://arxiv.org/abs/1709.01796

[25] Daniel Sonntag. 2018. Interactive Cognitive Assessment Tools: A Case Study on Digital Pens for the Clinical Assessment of Dementia. *CoRR* abs/1810.04943 (2018). arXiv:1810.04943 http://arxiv.org/abs/1810.04943

[26] Daniel Sonntag, Markus Weber, Alexander Cavallaro, and Matthias Hammon. 2014. Integrating Digital Pens in Breast Imaging for Instant Knowledge Acquisition. *AI Magazine* 35, 1 (2014), 26–37. http://www.aaai.org/ojs/index.php/aimagazine/article/view/2501

[27] B. Spenciere, H. Alves, and H. Charchat-Fichman. 2017. Scoring systems for the Clock Drawing Test: A historical review. *Dement Neuropsychol* 11, 1 (2017), 6–14.

[28] H. Tuokko, T. Hadjistavropoulos, S. Rae, and N. O'Rourke. 2000. A comparison of alternative approaches to the scoring of clock drawing. *Arch Clin Neuropsychol* 15, 2 (Feb 2000), 137–148.

[29] Yasmira I. Watson, Cynthia L. Arfken, and Stanley J. Birge. 1993. Clock Completion: An Objective Screening Test for Dementia. *Journal of the American Geriatrics Society* 41, 11 (1993), 1235–1240. https://doi.org/10.1111/j.1532-5415.1993.tb07308.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1532-5415.1993.tb07308.x

[30] D.J.M. Willems and R. Niels. 2008. *Definitions for Features used in Online Pen Gesture Recognition.* Technical Report. NICI, Radboud University Nijmegen. http://unipen.nici.ru.nl/NicIcon/

[31] Kun Yu, Julien Epps, and F. Q. Chen. 2011. Cognitive Load Measurement with Pen Orientation and Pressure. In *MMCogEms: Infering Cognitive and Emotional States from Multimodal Measures.*