Thomas Wittenberg*, Antonia Friedrich, Amelie Wittenberg, Stefan von Delius, Martin Raithel, Thomas Eixelberger, Sebastian Nowack

# Initial experiments of eye-tracking during AI-assisted polyp-detection in colonoscopy

**Abstract:** Currently, various AI-based systems for computer-assisted adenoma- and polyp-detection during colonoscopy have been brought to the market and are under clinical investigation. With these systems available to be used during routine screening colonoscopy and first results published about experiments and findings, it has become of interest how and to which extend such systems are used during the examination. Specifically, similarly to automotive navigation, it is of interest of how much visual focus is put onto the augmented image of the above-mentioned devices, signalling possible hypothesis of adenomas or polyps, and how much time-of-attention remains on the original colonoscopic video data. Thus, within a study, $N = 36$ participants using a prototype of a polyp-detection system have been observed with an eye-tracker-system, to capture and evaluate the *relative time of attention* with respect to the original and augmented video data and differentiate these values between various sub-groups based on experience, education and gender. T-tests were conducted to identify potential significant differences. Based on the obtained data, the augmented video data is used with a very high attention (up to 75%) depending on the regarded sub-group. Experienced as well as less-experienced users (with > 500 colonoscopies) both preferred looking at the original data. In contrast, gastroenterologists (in contrast to nurses, students, engineers) were more interested in the outcome of the novel AI-system. The female group preferred looking at the unobstructed data, while the male group was highly interested in the AI-based data.

**Keywords:** Artificial intelligence, Eye Tracking, Evaluation, Adenoma Detection

https://doi.org/10.1515/cdbme-2021-1031

## 1 Introduction

After almost thirty years of research and development in the field of machine learning and artificial intelligence (AI) with the goal design and develop devices to support gastroenterologists during colonoscopy [1], various commercially available AI-based systems for computer-assisted adenoma- and polyp-detection during screening colonoscopy have been brought to the market and are currently under investigation. Amongst the commercially available products, providing visual augmented hints about possible adenomas are e.g., *GI Genius* by *Medtronics* (USA) [2-4], the CAD EYE system from *FujiFilm* (Japan) [5, 6], or the *DISCOVERY* by *Pentax Medical* (Japan) [7,8]. With these systems available to be used during routine screening colonoscopy and first results published about experiments and findings [2-8], it has become of interest *how* and to *which extend* such systems are used during colonoscopy. Specifically, similarly to automotive navigation, it is of interest of *how much visual focus* is put onto the augmented image of the above-mentioned devices, signalling possible hypothesis of adenomas or polyps, and how much time-of-attention remains on the original (unaugmented) colonoscopic video data.
To this end within a study, $N = 36$ participants using a prototype of a polyp-detection system have been observed with an eye- / gaze tracker-system, to capture and evaluate (a) the *relative time of attention* $t_A$ with respect to the original and augmented video data, and (b) differentiate the values between various sub-groups based on experience, education and gender.

## 2 Related Work

Eye- and gaze tracking devices are already a common tool for market research as they can measure the point and duration of visual attention of the users. Using this technology, e.g., *Shi-*

***Corresponding author: Thomas Wittenberg:** Fraunhofer IIS, Am Wolfsmantel 33, Erlangen, Germany, E-Mail: thomas.wittenberg@iis.fraunhofer.de. **Antonia Stenzel, Sebastian Nowack**: formerly Fraunhofer IIS, Erlangen, Germany; **Thomas Eixelberger**: Fraunhofer IIS Erlangen, Germany; **Stefan von Delius**: Klinikum Rosenheim, Germany; **Martin Raithel**: Malteser Waldkrankenhaus, Erlangen, Germany, **Amelie Wittenberg**: Univ. Ulm

*nohara & Yamauchi* [9] objectively evaluated the skills during polyp-detection and snare-based polypectomy for eight novices and one experienced endoscopists by gaze tracking. The study showed that the experienced endoscopist detected the polyp faster than the novices, spent more time gazing at it and was less distracted by other events or structures like the snare-loop or searching the colon wall for polyps. *Meining et al* [10] used an eye-tracking device to compare image perception in gastrointestinal endoscopy with white-light endoscopy and narrowband imaging (NBI). 18 participants with different endoscopy experience were observed while assessing 23 image pairs measured the time spent on an image, the time until the first fixation of lesions, the total number of fixations per image and per lesion, and the number of fixations until finding the lesion. *Bernal et al* [11] applied an eye-tracking system to evaluate their approach of providing saliency maps in colonoscopy images with polyps, under the assumption that the observed gaze patterns of the physicians with respect to possible polyps in the scene are strongly related to the low-level image features (changes in edges, textures, colour) used to compute saliency maps. To this end they conducted a study with 22 participants. They also differentiated between groups related to experience.

# 3 Material and Methods

## 3.1 Video Data

From a wide selection of different colonoscopy videos (recorded with an Olympus EVIS EXERA III device) adequate image material was selected for evaluation. The following aspects were considered for the video material: They should depict scenes from real colonoscopy including various lesions, which are not too easy to recognize. The lesions should be recognizable by the AI-based detection system, while for evaluation purposes false-positives as well as false negatives were intentionally included manually. Thus, sub-sequences with no, one, or two polyps were compiled, whose entry and exit points were clearly recognizable.

As the complete evaluation was performed during a normal working day of the study group, the maximum processing time of the complete procedure (including questionnaires) was limited to 25 minutes and three videos with nine sub-sequences each and mean duration of approximately 3:40 minutes per video. All three videos are compilations of individual sequences from colonoscopy sequences. The content structure of these videos arose from the following considerations:

Videos #1 and #3 consist of the same sub-sequences and contain six polyps each, whereby the order of the individual sub—sequences were arranged differently in both videos. They are composed of 4,015 single frames and displayed at 25 fps. Video #2 was composed of 4,073 frames and depicted four polyps. The order of the sub-sequences in videos #1 and #3 was altered to allow a comparison for adenoma detection by the participants. Thus, video #1 was shown natively without the support of the AI-based detection software, while for video #3, the sequences appeared in an altered order in addition to the AI-based augmentation of detected polyps. The order of sub-sequences was altered that the participants would not notice that they had seen the content. In video #2, additionally to the AI-based detection, some false-positive and false-negative lesions were intentionally added. In video #2, two true positives were correctly augmented as well as two false-negatives (intentionally not marked by the AI-detection system) and two false positives (background tissue). The intentionally false-positives were used to determine the extent to which the users were influenced by the AI-software in the detection task.

The instructed task for all participants was to detect polyps under the aid of a novel AI-based polyp detection system.

## 3.2 KoloPol System

The KoloPol-system for automated real-time, low-delay adenoma and polyp detection during colonoscopy has been developed by the Fraunhofer IIS within a public-funded research project. The AI-based system is on one side based on a combination of methods from visual computing using a combination of low-level image features such as colour, texture, structures, edges, and their temporal developments, and on the other side on machine-learning approaches incorporating rule-based decisions related to human expert knowledge [12]. Using this system, the video data (described in the previous Section 3.1) has been analysed and prepared, thus yielding a known baseline for the experiments. Furthermore, as mentioned above, dedicated false-positives and false-negatives detection results have intentionally been added to the data, in order to check the attention of the study-participants.

## 3.3 Eye Tracking

The native and augmented videos were presented side-by-side to the participants of the study group, see *Figure 1.* In the left view an unaltered version of the colonoscopy video was presented while the right view provided the same video augmented with hypothesis for lesions. The gaze positions of the participants on the screen were recorded during the evaluation with a commercially available eye tracker system (*Tobii EyeX*) [13] and stored in an XML file. A software (Python) script was used

to assess whether the participants used the left or the right view of the polyp-detection software. See *Figure 2*, where the gaze positions over one second duration and one participant are overlayed onto a single colonoscopy frame.

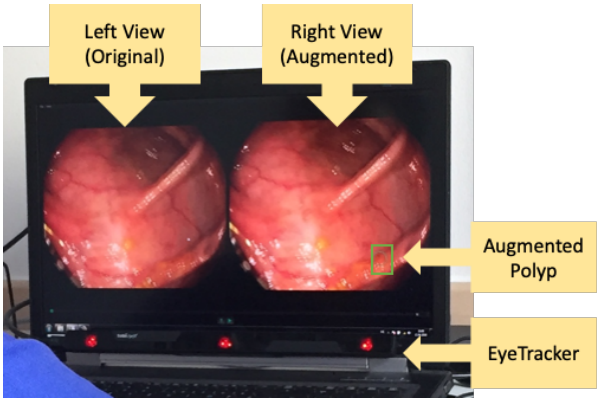


**Figure 1:** Technical Setup: in the left view, the original colonoscopy video was shown unchanged, in the right view the results from the AI-based detection system were displayed. The gaze was capture with an eye-tracking device mounted on the laptop.

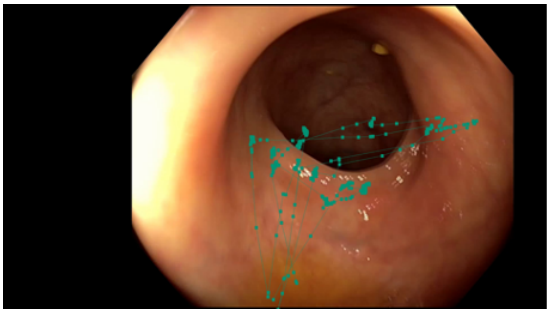

**Figure 2:** Example of gaze positions of one second overlayed on one colonoscopy frame

## 3.4 Study Group

The observed study group consisted of 39 participants, from which $N = 36$ completed the complete evaluation, while three subjects ended the evaluation prematurely. Furthermore, four datasets were excluded, as the eye tracker was not working correctly, so the final dataset consists of $N = 32$ participants. 53% ($N_{GE} = 17$) of these subjects were gastroenterologists, 31% ($N_{EN} = 10$) were endoscopy nurses, and the remaining 16% ($N_{other} = 5$) consisted of medical students and engineers. 56% ($N_{male} = 18$) of the participants were male, 44% ($N_{female} = 14$) female. The age of the subjects ranged from 20 to over 60 years. Based on their previous experience (with at least 500 colonoscopy procedures performed [14], $N_{experienced} = 21$ experienced and $N_{inexp} = 11$ less-experienced persons were identified.

## 3.5 Metrics

To evaluate the results the weighted average $\bar{x}^{G,P}$ within a certain target group $G \in$ {'experienced', 'less-experienced', 'gastroenterologists', 'other professionals'} and viewing panel $P \in$ {'left', 'right', 'both'} is calculated as

$$\bar{x}^{G,P} = \frac{\sum_{i=1}^{n} w_i x_i^{G,P}}{\sum_{i=1}^{n} w_i},$$

where $x_i^{G,P}$ denotes the $i$-th member in group $G$ viewing panel $P$. The scaling weight $w_i$ is computed as

$$w_i = \frac{N_i}{\sum_{i=1}^{n} N_i},$$

where $N_i$ denotes the number of frames in the regarded video. For the related standard deviation, we use the *weighted variation* as follows:

$$\sigma^{G,P} = \sqrt{\frac{\sum_{i=1}^{n} w_i \left(x_i^{G,P} - \bar{x}^{G,P}\right)^2}{\sum_{i=1}^{n} w_i}}$$

## 3.6 Statistical Analysis

Statistical analysis was conducted with IBM SPSS Statistics 26. Analysis was performed on a two-sided level of significance ($\alpha = .05$). The primary outcome (see Tables 1-3) was computed based on weighted mean values and standard deviation (see equations above). Assuming videos #2 and #3 of equal length, t-tests for independent samples were conducted on unweighted data. If there was no variance homogeneity, Welch-Test was used. Linear regression analysis was used exploratory to uncover potential dependencies in the data.

# 4 Results

## 4.1 Primary Results

The main results for the conducted eye-tracker study are presented in form of *percentual mean values* [μ] (see *Eq.* 1), weighted standard deviation [σ] (see *Eq.* 2), as well as the minimum and maximum [min; max] separated for the groups of "experienced" and "less-experienced" participants in *Table* 1 based on conducted colonoscopies, separated into "gastroenterologists" and the remaining group (endo-nurses, students, engineers) in *Table 2*, and split by gender in *Table 3*.

**Table 1:** Eye-tracking results, separated by experience

| | | Left View Original data [μ/σ] | Right view AI-detection [μ/σ]] | Both Views [μ/σ] |
|---|---|---|---|---|
| Experienced (*N*=21) | Video #2 | **56.60 %** | 42.36 % | 1.04 % |
| | | 38.68 % | 28.73 % | 0.89 % |
| | Video #3 | **53.25 %** | 45.61 % | 1.13 % |
| | | 39.40 % | 40.04 % | 1.06 % |
| less experiencedd (*N* = 11) | Video #2 | **51.41 %** | 47.23 % | 1.37 % |
| | | 28.01 % | 28.25 % | 0.67 % |
| | Video #3 | **62.17 %** | 36.47 % | 1.36 % |
| | | 25.96 % | 28.42 % | 0.64 % |

**Table 2:** Eye-tracking results, separated by education

| | | Left View Original data [μ/σ] | Right view AI-detection [μ/σ] | Both Views [μ/σ] |
|---|---|---|---|---|
| Physicians (*N*= 17) | Video #2 | 40.99 % | **58.16 %** | 0.85 % |
| | | 36.56 % | 36.91 % | 0,52 % |
| | Video #3 | 47.34 **%** | **51.34%** | 1.32 % |
| | | 34.26 % | 34.62 % | 0.87 % |
| Nurses /other (*N* = 15) | Video #2 | **70.07 %** | 28.45 % | 1.48 % |
| | | 27.09 % | 27.09 % | 0.99 % |
| | Video #3 | **71.05 %** | 27.72 % | 1.23 % |
| | | 33.25 % | 32.97 % | 1.23 % |

**Table 3:** Eye-tracking results, separated by gender

| | | Left View Original data [μ/σ] | Right view AI-detection [μ/σ] | Both Views [μ/σ] |
|---|---|---|---|---|
| Males (*N* = 18) | Video #2 | **35.60 %** | **63.18 %** | 1.32 % |
| | | 29.71 % | 30.21 % | 0,91 % |
| | Video #3 | **40,37 %** | **58.25 %** | 1,38 % |
| | | 32.90 % | 33.27 % | 1,03 % |
| Females (*N* = 14) | Video #2 | **78.43 %** | 20.51 % | 1,06 % |
| | | 26.70 % | 26,42 % | 0,75 % |
| | Video #3 | **75.43 %** | 23,57 % | 1.00% |
| | | 31.65 % | 31,52 % | 0,78 % |

The obtained extrema [min, max] are *not listed* in the tables as it seems that overall $n = 32$ regarded participants the achieved [min, max] values for looking at the left (original video) or right (AI-augmented videos) range from 0% (*never* looking at this side) to 100% (only looking at this side). Thus, it must be concluded, that at least one of the *experienced* participants as well as at least one *inexperienced* user preferred looking only on the left (original) side. Similar, at least one of the *experienced* participants as well as at least one *inexperienced* user preferred looking only on the right (augmented) side. These extrema also explain the relative high values of the weighted standard deviation.

## 4.2 Statistics

T-tests were performed based on differences in the *unweighted percentual mean* viewing time of the left video. There was no statistically significant difference between *experienced* and *less experienced* participants (Video #2: $t(26) = 0.28$ , $p = .781$; Video #3: $t(26) = 0.71$, $p = .486$).

A *statistically significant difference* was found between *physicians* and *nurses / others* with mean viewing time of the original data 30.36%-30.82% higher (Video #2: 95%-KI [6.58-54.13]; Video #3: 95%-KI [5.66-56.00]) for nurses / others, Video #2: $t(29) = 2.61$, $p \leq .01$; Video #3:$t(30) = 2.50$, $p < .05$).

There was a *statistically significant difference* between male and female participants with mean viewing time of the original data 33.90-42.64% higher (Video #2: 95%-KI[20.94-64.33]; Video #3: 95%-KI[9.19-58.69]) for female participants (Video #2: $t(30) = 4.01$, $p < .001$; Video #3: $t(30) = 2.79$, $p \leq .01$).

In exploratory linear regression analysis only gender (Video 2#: $\beta = 0.53$, $t(28) = 3.10$, $p < .05$; Video #3: $\beta = 0.4$, $t(28)=2.21$, $p < .05$) but *not profession* (Video 2#: $\beta = 0.20$, $t(28) = 1.20$, $p = .24$; Video #3: $\beta = 0.25$, $t(28) = 1.41$, $p = .17$) *nor experience* (Video 2#: $\beta = 0.84$, $t(28) = 0.55$, $p = .59$; Video #3: $\beta = 0.22$, $t(28) = 1.34$, $p = .19$) were significant predictors for viewing the original data. This suggests that differences in viewing time (left vs. right view) between physicians and nurses / others are potentially explained by gender effects.

# 5 Discussion

Eye- and gaze-tracking is an important tool to assess the visual attention of an endoscopist during colonoscopy with respect to various goals, such as education, training as well as focus on augmented data.

Based on the obtained data it seems that the augmented video data provided by such devices is used with a very high atten-

tion (up to 63%), depending on the sub-group (experienced vs. less-experienced, professional vs. non-professional, gender).
From the data shown in *Table 1*, it seems that the *experience* (related to at least 500 colonoscopies performed) is not a parameter to be taken in account, as independent of their experience, the participants mostly looked at the original data (51% – 62%) and *experienced* and *less experienced* participants did not significally differ in viewing time. – In contrast (see *Table 2*), it seems that the group of *professionals* (gastroenterologists) were quite interested (up to 58%) in the outcome of the novel AI-system (probably based on their curiosity of the novel AI-system as well the a-priory knowledge about possible false-positives and false negatives in the setting), while the *non-professional* group (endo nurses, students engineers) preferred looking at the original data (up to 75%). This is supported by significant differences in viewing time of the original data between professionals and non-professionals. However, exploratory linear regression analysis suggests that this effect might be explained by gender differences. – Finally, based on the *gender-split* (see *Table 3*), it can be observed, that the female group preferred looking at the unobstructed data (up to 78%) on the left panel, while the male group was highly interested (up to 63%) in the AI-based data presented in the right panel. The significant differences between different genders are also evident in the statistical analysis
Nevertheless, as the described experiments were made at a point of time (2017) where such AI-based detection-devices were still new to the community and not yet commercially available, it can be assumed that the captured attention with respect to the augmented data was higher that it would be after six months of using such devices, when the users have gained trusting experiences.
We believe that such experiments of attention measurements within endoscopy (and beyond) are of eager interest to the community in order to learn and classify where the focus of attention and concentrations is, and how endoscopist are possibly attracted to or distracted by to visual augmentation of video data and how to smoothly integrate such eye and gaze tracking systems into the clinical routine without to yielding an information overflow.

**Author Statement:** Research funding: This work has partially been supported by the German Ministry of Education and Research (BMBF) under project ‘KoloPol’ (03V0669). The Authors state no conflict of interest. Informed consent has been obtained from all individuals included in this study.

## References

[1] Wittenberg, Raithel. Artificial Intelligence-Based Polyp Detection in Colonoscopy: Where Have We Been, Where Do We Stand, Where Are We Headed? Visc Med. 2020 Dec;36(6):428-438.
[2] Hassan et al: New artificial intelligence system: first validation study versus experienced endoscopists for colorectal polyp detection Gut 2020;69: 799-800.
[3] Repici, Badalamenti et al.: Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroent. 159(2) 2020 pp 512-520.e7.
[4] Allescher, Mangold, Weingart: Künstliche Intelligenz in der Endoskopie – neue Wege zur Polypendetektion und Charakterisierung. *Gastroenterologe* 16, 3–16 (2021).
[5] Kubesch, Stratmann, Wittenberg, …(2021): Real-World Daten zur Anwendung von kommerziellen Artificial Intelligence (AI) Systemen zur Polypendetektion. Endoscopy Campus (EC) Magazin, 1.2021. S. 72-73.
[6] Weigt et al: Mit Hilfe eines validierten Polypendetektions- und Charakterisierungssystems können unerfahrene Untersucher Expertennieveau erreichen. Z Gastroenterol 2020; 58(08): e181
[7] Kiesslich: Highlights der Digestive Disease Week: Gastrointestinale Endoskopie. *Gastro-News* 7**,** 51–57 (2020). https://doi.org/10.1007/s15036-020-1378-6
[8] Antonelli, Gkolfakis, Tziatzios et al: Artificial intelligence-aided colonoscopy: Recent developments and future perspectives. *World J Gastroenterol*. 2020;26(47):7436-7443. doi:10.3748/wjg.v26.i47.7436
[9] Shinohara & Yamauchi: Eye Tracking As a Tool for Evaluating Colonoscopic Polypectomy Skill: A Feasibility Study. SAGES Meeting 2013 https://www.sages.org/meetings/annual-meeting/abstracts-archive/eye-tracking-as-a-tool-for-evaluating-colonoscopic-polypectomy-skill-a-feasibility-study/
[10] Meining, Atasoy, Chung, Navab, Yang: ‚Eye-tracking‘ for assessment of image perception in gastrointestinal endoscopy with narrowband imaging compared with white-light endoscopy. Endoscopy 2010; 42(8): 652-655
[11] Bernal, Sánchez, Fernández-Esparrach et al.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput Med Imaging Graph. 2015 43:99-111.
[12] Klare, Sander, Prinzen et al.: Automated polyp detection in the colorectum: a prospective study (with videos): Gastrointestinal Endoscopy 2019, 89(3): 576-582.e
[13] https://help.tobii.com/
[14] Spier, Benson Pfau et al. Colonoscopy training in gastroenterology fellowships: determining competence. Gastrointest Endosc. 2010 Feb;71(2):319-24.