International Journal of Aviation, Aeronautics, and Aerospace

Volume 7 | Issue 3 Article 7

2020

Automatic Gaze Classification for Aviators: Using Multi-task Convolutional Networks as a Proxy for Flight Instructor Observation

Justin Wilson
Southern Methodist University, wilsonj@mail.smu.edu
Sandro Scielzo
L3Harris, sandro.scielzo@l3harris.com
Sukumaran Nair
Southern Methodist University, nair@lyle.smu.edu
Eric C. Larson
Southern Methodist University, eclarson@smu.edu

A typical scenario for training military personnel to fly specialized aircraft is to have aviators fly mock missions in a simulator with an instructor nearby. When evaluating the performance of these aviators, flight instructors rely on observation and after-the-fact assessment. Scan patterns are vital aspects of context for a flight instructor, and are fundamental to basic flight. For example, a student might scan too rapidly, omit, or fixate - these are common errors when scanning the horizon and cross-checking instruments (United States Air Force [USAF], 2019). Anecdotally, flight instructors often cite that head and eye movements are pivotal for judging student intent and situational awareness. Virtual reality-based training environments with embedded eye-tracking offer the possibility to automate and provide more context to some aspects of instructor observations and potentially expedite the learning process. In this work, we evaluate how using eye-tracking (with machine learning) can objectively assess aviator scan patterns during training, which may reduce instructor overall workload. Therefore, two key **research questions** are:

1. Do flight instructors assess the quality of scan patterns of an aviator similarly?
2. If so, can machine learning techniques be used to automate the instructor evaluation of scan pattern quality for aviators in various phases of flight?

We hypothesize that both research questions can be answered in the affirmative. A *gaze* or *scan pattern* is a technique in which an aviator observes all requisite information inside and outside the aircraft in-order appropriately and safely fly that aircraft. The scan begins and ends in the same position, observing all applicable items - "systematically, thoroughly...complete, and continuous" (United States Navy [USN], 2019).

We propose a new method for gaze classification by transforming gaze or scan patterns into heatmaps and classifying them with deep convolutional neural networks (Krizhevsky, Sutskever, & Hinton, 2012). The patterns are classified into levels of "quality" that would typically require review from an instructor. Data are collected in a mixed-reality training environment using a physical flight simulator, a virtual reality environment, and a gaze-tracking sensor for monitoring eye movements within the virtual space. From these devices a heatmap is synthesized from the pattern created by the gaze of an aviator flying during a specified window of time. We detail the contributions of our work as follows:

- We designed and carried out a human subjects experiment for aviators in a variety of flight scenarios. We recruited three instructors to review the gaze patterns from these scenarios, and we analyzed inter-rater reliability; We conclude there is strong agreement between these expert raters on what gaze quality is for a given maneuver.
- We propose methods for gaze data augmentation specific to pilot gaze patterns that increase robustness of trained machine learning models.

- We investigate two competing, convolutional neural network architectures: a task agnostic model and a multi-task model. We evaluate the architectures with K-fold cross-validation, achieving greater than 93.0% average test accuracy compared to instructor observation.

The location of gaze within a reference frame, *i.e.* what the aviator is looking at as mapped to that reference frame, is deceptively complex when measured from an unconstrained head mounted system. The calculation is based upon: (1) the alignment of head and eye positions with respect to the subject's field of view (FOV), (2) the object observed, (3) the location on the object observed, (4) the reference frame into which the point-of-gaze is mapped, (5) the angular error and precision of that translation, and (6) the calibration error -therefore the mapped position within that reference frame can vary dramatically. Thus, the problem is not as simple as pinpointing the objects within the reference frame and defining a bounding box. Moreover, the regions of interest for a pilot gaze pattern will vary depending on the maneuver in question, and the number of regions within that reference frame can increase based on the complexity of the maneuver. A classifier that uses gaze must both scale with increased regions of interest and handle perturbations in gaze position as projected onto the reference frame. This is not just for fixed head position (among aviators), but in dynamic support of the yaw, pitch, and roll of head and eye movements of a given aviator - especially as it pertains to the movement of the subject's FOV about the heads-up display's (HUD) eyebox (Spitzer, Ferrell, & Ferrell, 2017). An obvious fixed-position example is that aviators have different abdomen heights and seat height preferences, which can lead to some aviators looking slightly downward or slightly upward towards the instrument panel and displays (USN, 2015). One machine learning model that can learn features that are robust to translations is a convolutional neural network (CNN) (LeCun, Bengio, et al., 1995). CNNs are capable of handling fixed and dynamic perturbations throughout a reference frame, and do not require individual labeling of regions of interest (Chollet, 2017a). Given these advantages of CNNs, we hypothesize these models are superior in classifying gaze.

LITERATURE REVIEW

Our work builds from a number of research communities. As such, we divide our discussion into five relatively disjoint areas: (1) sight picture, (2) situational awareness, (3) heatmaps in eye-tracking, (4) physiology and eye-tracking in aviation, and (5) other works in gaze classification.

Sight Picture

When training pilots to perform new maneuvers, instructors will often refer to the concept of a "sight picture" and assess pilots for their ability to recall and use different sight pictures. In the case of firearms, sight picture can be

referred to as the perspective, as viewed by the shooter, created by "the alignment of the sights of a firearm with the target" (Merriam-Webster, 2020). The concept of sight picture in aviation originated with fixed gunnery weapons (Miller & Gleason, 1947). When such a firearm is affixed to an aircraft the whole aircraft must be maneuvered in order to properly aim. Thus, an early concept of sight picture in aviation can be thought of as the perspective of the pilot—what he/she sees through a reflector plate of the cockpit—given the aircraft serves as the firearm. This requires perceptual abilities to evaluate aim and fly the aircraft (Miller & Gleason, 1947). Students would learn a series of sight pictures for a discrete number of angles of attack (AoA), which is further compounded by the realities of a moving target in air-to-air situations. Strong "perceptual memory" is required for both understanding what the correct sight picture is and correcting to it, given the current sight picture (Miller & Gleason, 1947). Applications and technology have evolved, but the concept of aligning an aircraft with the environment for the purpose of accomplishing a specific maneuver based on an understood mental picture of what is correct—from the perspective of the pilot in the seat of the cockpit—has become fundamental to basic flight (Federal Aviation Administration & Soucie, 2017; Kershner, 2001; USN, 2011, 2015, 2019). Thus, gaze patterns may have a strong relationship to the internal sight picture of an aviator, and we hypothesize that instructors can evaluate the quality of gaze patterns based on the expected patterns for a given flight maneuver.

Situational Awareness

In the context of aviation, situation or situational awareness (SA) is an established cognitive construct representing a state of knowledge about a dynamic environment, which is linked probabilistically with pilot performance (Endsley & Garland, 2000a). In the aviation community, the idea of SA is often associated with the pilot's ability to answer specific questions about their environment and maneuver. In aviation, maintaining accurate SA is crucial to ensure mission success, and the lack of SA is often associated with pilot error (Endsley & Robertson, 2000; Fuller, Johnston, & McDonald, 1995). As a result, the need to accurately measure SA is important to improve both training design and overall training outcomes (Endsley & Garland, 2000b). Much effort has been conducted towards the understanding of eye movements and gaze patterns, which "shed considerable light on [aviators'] real-time behavior" in aviation and aerospace as a whole (Valerie et al., 2005). For example, using nine gaze representations, (Newn, Velloso, Allison, Abdelrahman, & Vetere, 2017) showed that humans have a strong capacity to accurately infer intent. Our research could potentially provide a form of context or situation awareness to a broader evaluation system.

Flight instructors seek to comprehend the intent and awareness of their students - context that is currently derived from in-flight observation and post-flight examination. Because of this, situation-aware avionics - capable of assisting

either the flight instructor or the aviator in understanding the human-machine state (Calhoun, 2016) in real-time - is highly sought after. Key motivations include accelerating aircrew training, improving aircrew performance, and decreasing pilot workload. Our work expands upon these ideas by comparing gaze on an equal footing, using subject matter expert opinion (*i.e.*, flight instructors) to label gaze heatmaps.

Heatmaps in Eye-Tracking

Gaze can be defined as the direction of the visual axis within a reference frame. It is a summation of eye position relative to the head, and the head position relative to the same reference frame (Guitton & Volle, 1987). The visualization of gaze is a key research area in gaze tracking. A predominant gaze visualization is the heatmap. Heatmaps can provide clear depictions of aggregate gaze by combining gaze fixations while sacrificing the depiction of the order in which the fixations occurred (Duchowski, Price, Meyer, & Orero, 2012). Privitera (2006) found that different subjects are mostly consistent on what regions they observed, but are less consistent in the order they view them. Therefore, order is not necessarily as important as the locations observed (Duchowski et al., 2012). Spakov provides an in-depth examination on the methods for visualizing fixations, including heatmaps (Spakov, 2008; Spakov & Miniotas, 2007). Moreover, Stellmach, Nacke, and Dachselt (2010) examined gaze fixation in three-dimensional (3D) virtual environments surveying 3D scan paths, 3D heatmaps; and they provide a prototype toolkit for aiding future eye-tracking studies.

Given heatmaps are accumulated fixations, Stellmach *et al.* found that they are useful for indicating visual attention over a period of time because visualizing data in three dimensions enables a representation over a longer period of time. We chose to use two-dimensional heatmaps within three fixed reference frames around aviators because: (1) experienced aviators have highly developed perceptual memory and judgment (Miller & Gleason, 1947), (2) heatmaps provide a clear depiction of aggregate gaze (Duchowski et al., 2012), (3) subjects are consistent on regions they observed (Privitera, 2006), (4) gaze order is not necessarily as important (Duchowski et al., 2012), (5) heatmaps are useful for indicating visual attention over a period of time (Stellmach et al., 2010), and (6) humans have a strong capacity to infer intent through gaze representation (Newn et al., 2017).

In our experiments, we segmented flights into smaller phases, where similar gaze region abstractions can be expected. The three reference frames chosen provide full cockpit coverage. As an example, Figure 1 is a high-resolution heatmap aggregated over a full normal landing maneuver (the clouds are part of a background image only, and do not reflect what the pilot actually saw during the landing). The three reference frames are shown in the Figure and

overlaid with heatmaps denoting where the pilot scanned, from the perspective of each frame. From the heatmaps, it is apparent that the pilot looked directly at the HUD most of the time, while periodically scanning the horizon.

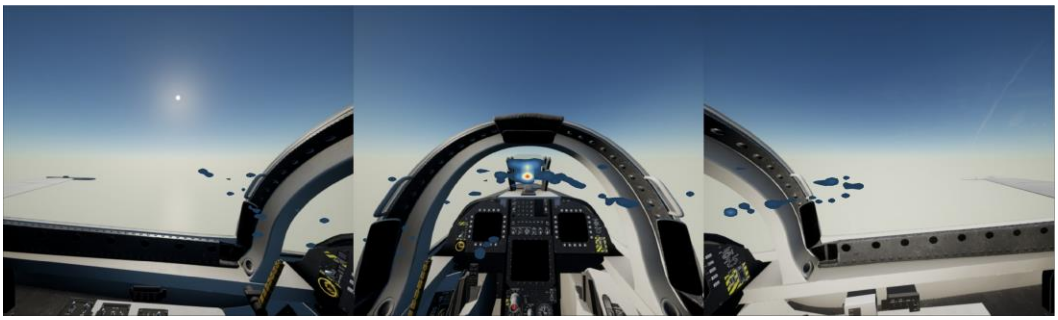


Figure 1. Aggregate high-resolution heatmap of a normal landing.

Physiology and Eye-Tracking in Aviation

Our work is not the first to look at eye tracking with aviation. However, previous works do not attempt to classify the quality of a gaze pattern. Instead, their focus is often on capturing the attention or cognitive workload of the pilot. Schnell, Keller, and Poolman (2008) sought to assess workload with a tool that works to unify flight data with physiological measures into a single framework in flight and in real-time. Weibel, Fouse, Emmenegger, Kimmich, and Hutchins (2012) looked at digital ethnography to understand visual attention of aircrew throughout varying phases of flight using a mobile eye-tracking system — they reported on techniques and methods to digest and visualize the dynamics of time-synchronized, multimodal, visual attention data. Specifically, they looked at visualizing tracking data, analyzing areas-of-interest, with infrared markers and the errors associated, visualizing the temporal dynamics, such as overlaying gaze on video frames, and gaze-to- object recognition. Weibel *et al.* sought to discern when an aviator's gaze fixed on an object of interest without IR markers. They did this with OpenCL by matching objects from one frame to all frames. Vrzakova and Bednarik (2012) sought to understand how mobile eye-tracking could work in a real cockpit. Recently, Lounis, Peysakhovich, and Causse (2018) looked to enhance aircrew-aircraft interaction. They monitored the attentional behavior of aircrew using a gaze tracker and developed a cockpit monitoring database that serves as an assessment tool. They expanded on their work by developing a flight eye-tracking assistant built on their database that uses thresholding of dwell times for areas of interest with audible alarms (Lounis, Peysakhovich, & Causse, 2020). While these previous works investigated physiology in the context of aviators, none explicitly address quality of gaze as observed by a subject matter expert.

Other Works in Gaze Classification

A number of works have investigated the classification of intent and attention from gaze data. While looking at a way to discern intent, Goldberg and

Schryver (1993) developed heuristics from multiple discriminant analysis to enable gaze-controlled UI zoom. Frutos-Pascual and Garcia-Zapirain (2015) look at attention performance with saccadic and fixation gaze data over 32 children. They achieved 88.0% accuracy with a random forest classifier. Abdelrahman et al. (2019) developed a way of classifying attention types through the use of thermal imaging and eye tracking. They developed several classifiers capable of classifying four types of attention (Sohlberg & Mateer, 1987) with an average AUC of 80.3%. Similar research to the work Weibel *et al.* conducted, Barz and Sonntag (2016) examined using gaze-to-object recognition with neural networks to classify objects and ensure the user draws attention to that object. They use a dispersion algorithm for fixation (Barz, Daiber, & Bulling, 2016) and thresholds for attention. Interestingly, a significant step towards gaze classification was conducted by Li, Mettler, and Andersh (2015). They investigated classifying gaze itself by breaking it into three fundamental patterns: (1) sac- cades (Carpenter, 1977; Guitton & Volle, 1987), (2) smooth pursuits (Robinson, 1965), and fixations (Robinson, 1964). They studied consumer helicopter drone pilots, while performing guidance and control tasks, as well as surgeons who conducted a peg transfer task (Sroka et al., 2010). They devised a scheme for converting gaze to a fixed reference frame for classification. They employed a spherical head centric coordinate frame, from a study of the receptive field of flies (Huston & Krapp, 2008), correlated with six-cameras. Using both empirical thresholding and hidden Markov models (HMM) to classify gaze data, they were able to accurately classify the three fundamental patterns with the use of gaze velocity and distance.

Our work is similar to these in that we classify gaze patterns based on discrete criteria. However, we use a more complex classification task — equating gaze patterns to human observation. Because the complexity of the classification task is increased, we also investigate multiple reference frames and a convolutional neural network architecture with increased predictive power. The advantage of this methodology is its ability to handle both perturbations in gaze position and scale with more complex gaze regions as the task and gaze patterns increase in complexity—all without the use of bounding boxes.

Figure 2. BBXR mixed-reality simulator (Hanson, 2018).

DATA COLLECTION

The data collection effort consisted of a repeated measures experiment (each maneuver was flown twice), including three flight maneuvers. The experiments were approved by a university IRB, application H18-105-LARE. We recruited 40 test subjects consisting of twenty-one pilots, nine operators, and ten novices. The pilot group involved individuals with military and commercial experience, all of whom had military flying experience in heavy, rotary, and/or fighter-type aircraft. Operators included naval flight officers (NFOs), combat systems officers (CSOs), remotely piloted aircraft (RPA) sensor operators, and avionics technicians; all with some flight or simulation experience. Novices consisted of those who had no aircraft experience at all. Gaze data and screen capture video were recorded for each maneuver. The three maneuvers flown during this experiment are listed as follows:

- *Cruise Maneuver:* the subject was instructed to fly straight and level maintaining 12,500 ft and 350 knots-indicated airspeed (KIAS) with tolerances of ±100 ft and ±15 KIAS for five minutes.
- *Normal Takeoff:* the subject was positioned on the centerline of the runway 13R at (simulated) Falon Naval Air Station (NAS) KNFL. The subject was asked to smoothly apply max power, rotate at 140 KIAS, and pitch between seven- and ten-degrees nosehigh – climbing 3,000 ft and leveling off. Tolerances included: ±1°, ±10 ft centerline, and ±2 deg runway heading.
- *Normal Landing*: the subject was positioned on final to runway 13R at Falon NAS. At decision height, 500 ft above ground level (AGL), the subject initiated a full- stop landing. The subject was instructed to verbalize his or her desired touchdown point upon nearing the runway. Tolerances included the nose-wheel within 10 feet of centerline.

Three other maneuvers were also flown. These maneuvers were collected during

the same experiment, but for other associated research. Using the concept of sight picture, they were useful in augmenting the original dataset, especially for less accurate gaze patterns (discussed further in a later section). The other three maneuvers flown are listed as follows:

- *Boundary Avoidance Tracking — Longitudinal Axis*: the subject was positioned be- hind a target aircraft that moved periodically, at random, in the vertical axis. As the target aircraft moved, the subject was asked to keep their aircraft's crosshair inside of a defined boundary about the target's longitudinal cross-section. With maneuver duration and subject piloting ability, the task difficulty was increased by reducing the boundary spacing.
- *Boundary Avoidance Tracking — Lateral Axis*: the subject joined on a target aircraft's right-wing. The target moved periodically and at random intervals in the vertical axis. As the target aircraft moved, the subject was asked to keep their aircraft's wing or canopy handle inside of a defined boundary about the target's lateral cross-section. With the progression of the maneuver, the task difficulty was increased by reducing the boundary spacing.
- *Air Intercept*: the subject begin flying straight and level. The subject obtained a radar lock on bandit aircraft. They offset his/her aircraft 30° left or right and descend 10° nose low to the bandit's altitude. At level-off, the pilot accelerated to $>$ 400 KIAS, and executed an intercept/escort profile. Subject closed for visual identification (VID) and verification of the aircraft markings (fin flash).

Boundary avoidance tracking (BAT) is a flight test technique used to understand the "pilot-in-the-loop handling qualities" (Gray, 2008). For both boundary avoidance tracking tasks, when the subject overshot a boundary, they were asked to rapidly recover and place the aircraft back into position and continue the maneuver. The BAT tasks were used to increase the required pilot workload to complete the maneuver. To this end, the simulator operator had the ability to manipulate the pitch control laws to increase/decrease the overall response of the aircraft. The operator altered the control laws in such a manner as to ensure the subject was stabilized before stepping to the next adjustment in a buildup manner. The subject flew each BAT maneuver for a minimum of five minutes.

Prior to data-collection each subject was given five minutes to familiarize themselves with the aircraft, during which any questions were answered. For Novices, they were shown how to use the stick, throttle, rudder pedals, and the heads-up-display (HUD) symbology was explained. During each maneuver the subject was asked to perform the maneuver as he or she normally would. Novices

were asked to perform the best they could, and no guidance was provided on proper gaze or scanning patterns. Given the air intercept task is a more complicated maneuver each subject was given the opportunity to practice the maneuver once; this demonstration data was also captured.

Equipment

The test simulator utilized was a prototype Blue Boxer Extended Reality Simulator (BBXR) (Hanson, 2018), provided by L3Harris Technologies. The Blue Boxer is a portable, training, mixed reality system that simulates the flight characteristics of aircraft. It utilizes virtual reality and high-precision hand tracking. Designed to be compact and portable, the system amalgamates physical and virtual mission equipment to simulate the flight environment. A key component of this system is the HTC VIVE Pro Eye Virtual Reality (VR) Headset for eye-tracking measurements. The HTC VIVE contains an integrated Tobii eye tracker, which is robust to head movements and can be worn with eye glasses. The Tobii eye tracker is similar in specification to the Tobii Glasses Pro line, with 90 Hz sampling rate, single point calibration, absolute pupil measurement, and slippage compensation if the headset moves unexpectedly. Gaze patterns over the course of each maneuver were collected. These are heatmap patterns that are traced from the visual eye path—containing eye fixations. The heatmap of gaze is calculated with a fixed origin relative to the virtual cockpit. As such, head position and orientation, relative to the cockpit, were collected.

Data Format

The BBXR provided gaze data at 90Hz, and this included the left, right, and combined x-position and y-position on a given field of view (FOV) which is projected onto a two- dimensional square reference frame also known as a screenspace. Three screenspaces were established with 60° FOV and with the same camera origin located just above the pilot's chair. This position did not change for the duration of the data collection effort.

The three screenspaces are located relative to the cockpit, as shown in Figure 1. One screenspace is center and in-front of the multi-function displays and heads-up display. The other two screenspaces are positioned parallel to each other. They are aft and perpendicular to the center screenspace. One is on the left-hand side of the cockpit, and the other is on the right-hand side, effectively surrounding the pilot within the cockpit. Gaze position is reported for each screenspace. If the subject is looking at something in a given FOV, the position for that screenspace is reported between 0 and 1 for both x and y coordinates.

Expert Review

Once data collection was completed, the screen capture video was separated by subject and maneuver. Heatmaps were generated, using a non-overlapping 30-second sliding window for the duration of each maneuver flown. Three subject matter experts labeled a subset of the gaze data using maneuver videos and

heatmaps to establish inter-rater reliability. The subject matter experts included two experienced instructor pilots (IP) and one experienced instructor combat systems officer (CSO). Table 1 lists the demographic information for each expert rater that reviewed the subject gaze patterns. Further details on the review process are discussed in a subsequent section.

Table 1
Rater Experience

Instructor Pilot I	Instructor Pilot II	Instructor CSO[a]
27 years flying 12 aircraft flown 5,025 total flight hours 2,000 instructor hours USN TOPGUN graduate and instructor Former USN TOPGUN commanding officer	40 years flying 16 aircraft flown 12,860 total flight hours 1,250 instructor hours USN TOPGUN graduate and instructor First Officer on B-757/767, B-737 and Airbus 320	13 years flying 31 aircraft flown 1,300 total flight hours 225 CSO instructor hours USAF Test Pilot School graduate

[a]Total hours encompass both CSO and pilot time

Demographics

Table 2 lists the subject demographic information for the 40 test subjects. This includes the sample mean, standard deviation, min and max values for subject age, flight hours, and number of aircraft flown. Further the percentages for type of flight experience are listed. Note that the flight experience is not exclusive to military or civilian, a subject can have both. Therefore, flight experience can add up to greater than 100%.

Table 2
Subject Demographics

	Age				**Hours**				**Aircraft Flown**				**Flight Experience**	
Subjects	$\bar{x}$	Std Dev	Min	Max	$\bar{x}$	Std Dev	Min	Max	$\bar{x}$	Std Dev	Min	Max	Civilian	Military
All Subjects	42.0	10.7	22	61	2816.0	2876.6	0.0	13000.0	4.2	3.7	0	12	25.0%	72.5%
Pilots	47.4	7.3	35	61	4631.8	2739.3	1700.0	13000.0	6.5	2.6	3	12	42.9%	100.0%
Operators [a]	37.9	20.0	23	43	1910.0	1550.0	0.0	4500.0	3.9	3.9	0	9	11.1%	88.9%
Novices	34.8	22.8	22	53	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0.0%	0.0%

[a]For operators, operator time and pilot time, if applicable, are reported combined.

GAZE QUALITY LABELING

"Gaze quality" in the context of this research is defined as a rating on a 3-class ordinal scale evaluating the subject's ability to scan his/her environment in-order to safely and correctly execute the assigned maneuver or task. It is based on instructor opinion among the three raters. To understand if instructors rated the gaze patterns similarly and to establish an appropriate scale, inter-rater reliability (IRR) was evaluated. The raters included two IPs and an instructor CSO. All three are seasoned instructors and evaluators with military experience. Given their experience, they have refined perceptual abilities, making them ideal for the labeling task.

Figure 3. Example of a concatenated video frame from a video that instructors reviewed for labeling gaze quality. Three heatmaps of different screenspaces line the top of the frame and the bottom consists of video from the pilot (left) and a zoomed heatmap of the HUD (right).

Review Interface Creation

In support of our labeling effort we created high-resolution (hi-res) heatmaps for the instructors to review along with the maneuver video. These hi-res heatmaps were shown in conjunction with maneuver video so that instructors could label the quality of segments of gaze data over a window of time incorporating additional context outside of the heatmap and resolution. The data collection videos and data were segmented based on subject and maneuver. Hi-res heatmaps

were generated over 30-second, consecutive, non-overlapping windows during a maneuver. If there were any leftover data points, a final overlapping 30 second heatmap was generated for the last 30-seconds of the maneuver. The start and stop times of these 30-seconds windows were saved so that our machine learning model could process the data in the same labeled time window. A 30-second window was chosen because the raters considered that interval to be a reasonable amount of time for an instructor to observe and discern the quality of a student's gaze pattern. The hi-res heatmaps were created by using a bivariate kernel density estimate (KDE) with a Gaussian kernel. A 500-level KDE overlay was generated on top of still images of the cockpit from each screenspace's FOV, as shown in Figure 3 (Top). The heatmap intensity was scaled over all three screenspaces. A zoomed version over the HUD was also created.

The interface used by the instructors for labeling consisted of the three screenspace heatmaps and a zoomed HUD heatmap concatenated to each video frame of the pilot's maneuver (Figure 3). This means that as the reviewer was observing the maneuver video, he/she was also viewing the hi-res heatmaps for the 30-second window of data being observed. The original maneuver video provided situational awareness on aircraft movement and position, but it also provided gaze convergence tracking represented by a green eye floating about the cockpit. This way, the instructors could review not only the heatmaps synthesized from gaze data, but they also observed what the pilot was staring at during their maneuvers. A reviewer always reviewed data for an entire maneuver and the windows of data were labeled in the order they appeared in the maneuver. Figure 3 provides an example of a concatenated video frame.

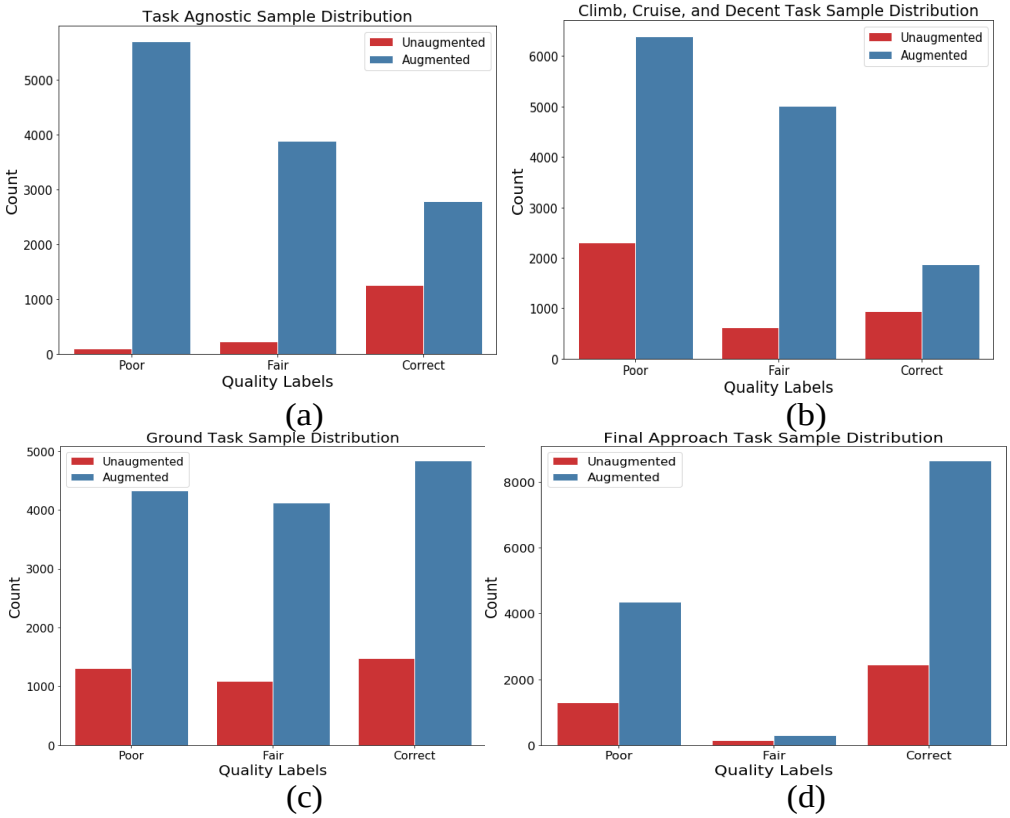


Figure 4. Sample counts for task agnostic and multi-task models.

Rating Process

For labeling, the three raters used a grading scheme of poor, fair, or correct. Within this scale the raters were further allowed to rate windowed data with scores that were "in- between" levels such as "poor-to-fair," yielding a five-class scale: (1) poor, (1.5) poor-to-fair, (2) fair, (2.5) fair-to-correct, and (3) correct. Pilot I reviewed 520 sets and Pilot II reviewed 517 sets of thirty-second windowed data, pulled from across all 40 test subjects. Of the total ratings, both pilot raters reviewed 109 gaze patterns that overlapped be- tween the two datasets to support the investigation of inter-rater reliability. The CSO was the primary annotator of the full dataset, and therefore provided ratings for all the data that the two pilot raters provided. The label count distributions are found on Figure 4.

Inter-Rater Reliability Results

When investigating inter-rater reliability, we employed Cohen's κ (Cohen, 1960, 1968), the correlation coefficient for the binary-rater case, Fleiss' κ (Fleiss, 1971) and Randolph's κ (Randolph, 2005) for the multi-rater case. We also investigated multiple levels of agreement by transforming the 5-class ratings into multiple 3-class variations. Specifically, the 5-class was transformed into 3-class

ratings by ceiling and flooring the class labels. A final 3-class version was calculated by taking the floor of the rating 1.5 and the ceiling of the rating 2.5. The results of each transformation and evaluation criteria are shown in Table 3.

Table 3
Inter-rater Reliability

Raters	Coef.	5−Class	Ceil [3−Class]	Floor [3−Class]	Floor/Ceil [3−Class]
Pilot I and II	*Cohen's κ*	.45	.60	.36	.71
	r	.66	.68	.50	.79
Pilot I and CSO	*Cohen's κ*	.30	.40	.24	.37
	r	.43	.46	.33	.44
Pilot II and CSO	*Cohen's κ*	.63	.66	.60	.71
	r	.78	.72	.68	.76
Pilot I, Pilot II, and CSO	*Fleiss'κ*	.30	.55	.39	.56
	Randolph's κ	.47	.78	.50	.78

For the binary-rater case, pilots I and II's Cohen's κ were strongest when the floor and ceiling were utilized, a κ of 0.71. Further, an r-value of 0.79 reveals a strong positive linear relationship. Pilot I and the CSO had less agreement than that of both instructor pilots. However, when the ceiling method was used a κ of .4 reflects moderate agreement. This was further observed by an r-value of 0.46 — indicating a positive linear relationship. Finally, Pilot II and the CSO exhibit strong agreement with a κ of 0.71 and an r-value of 0.76 when the floor and ceiling method was utilized. Overall, this signifies that Pilot I and the CSO have moderate agreement while both have strong agreement with Pilot II.

The multi-rater case yielded a strong inter-rater relationship. Again, the ceiling and floor/ceiling methods provided the highest reliability. The overall multi-way Randolph's κ was 0.78 and the multi-way Fleiss' κ was 0.56, indicating a strong inter-rater reliability was established. ***Therefore, we answer our first research question by affirming our hypothesis: Instructors can grade gaze quality into three levels with sufficient similarity.***

Of note, the rating label distribution is uneven and further addressed in a subsequent section. While both the ceiling and the floor methods provide high inter-rater reliability, the floor/ceiling method was chosen as it yielded more labels for "poor" class — the weakest class count across tasks from the unaugmented dataset.

(a) (b) (c) (d)

Figure 5. Four example center screenspaces (a-d) — While (d) would rate "poor" for all 3-tasks, (a-c) are each characterized as (a) "correct", (b) "fair", and (c) "poor" given the climb, cruise, and decent task. However, (b) is labeled "correct" for the ground task, as only airspeed and centerline are needed, and (c) is labeled "correct" for the final approach task, a pilot flying the AoA bracket.

Multi-Task Labeling

Throughout the labeling of the dataset, raters expressed concern that there can be more than one phase within a maneuver. For example, on takeoff the pilot transitions from ground roll to a climb. Each of these phases can have a different "correct" gaze pattern. There is a potential robustness issue for a machine learning model because a given gaze pattern might be judged "correct" in one phase, but would garner only a "fair" in another phase, see Figure 5. A more robust solution is to use a multi-task machine learning model that has the ability to interpret the same gaze pattern differently, depending upon the phase. From the maneuvers we grouped the phases into three generalized phases (tasks) according to the relative similarity of gaze pattern for the phases: (1) climb, cruise, and decent (CCD), (2) ground, and (3) final approach. Taking advantage of the idea that there are common gaze patterns among phases, the annotator was asked to label every window of gaze data according to all three generalized phases. That is, the annotator labeled quality as if the subject was flying each of the three generic phases described above, and providing a label for each - Figure 4.

DEEP LEARNING ARCHITECTURE

Now that it has been established that instructors can similarly rate the gaze pattern of aviators, we move our analysis to our second research question: Can machine learning be used to automate the classification of gaze quality? To investigate this, we choose to use a model that can make predictions directly from heatmaps of the input gaze patterns: a convolutional neural network. For our convolutional architecture we employ two key techniques: transfer learning and multi-task learning. Leveraging prior knowledge to hasten the learning of new tasks is known as transfer learning (Jonathan Baxter et al., 1995; Pan & Yang, 2010).

These methods preserve and take advantage of previously trained models from one task or domain and apply them to a second different task or target domain. More broadly, transfer learning allows for domains, tasks, and distributions between training and test to be different (Pan & Yang, 2010). Such methods can affect new training of accurate models for an entirely different task and/or source domain where labeled data may be limited (Pan & Yang, 2010). In our research, the use of pre-trained layers is helpful because we can leverage the ability of the model to find a number of low-level image features such as edges, shapes, and noise that generalize to our target task of classifying gaze patterns.

A related concept is the preservation of learned knowledge while training multiple tasks—multi-task learning. While multi-task learning can be considered a form of transfer learning, it traditionally differs in that the shared knowledge is learned at the same time, between tasks, and during the training process. A typical approach for multi-task learning is to uncover the shared latent features that can benefit each task (Pan & Yang, 2010). Ruder (2017) shows that multi-task learning models tend to prefer solutions that generalize.

In this work we take advantage of the learned weights from the Visual Geometry Group's 16-layer model (VGG16) from Simonyan and Zisserman (2015). VGG16 was trained on an ImageNet repository ILSVRC-2012 dataset (Russakovsky et al., 2015), a repository used for the 2012-2014 Large Scale Visual Recognition Challenge (ILSVRC). The ILSVRC-2012 dataset is built as a subset of ImageNet's (Deng et al., 2009) greater repository with a training dataset of 1000 categories and 1.2 million images. We seek to take advantage of the spatial representations learned at the shallower depths of the VGG16 model— where such representations are less complex, domain specific, and more applicable to the gaze domain. Though these representations are trained on a substantially different task, we expect that many can be re-purposed for gaze classification, while other representations will be ignored.

Two models were implemented - a task agnostic version which has no knowledge of the maneuver being performed, Figure 6, and a multi-task version with three tasks (such that a classification of quality is provided for each type of flight maneuver), Figure 7. Both models utilize the first seven weight layers of the VGG16 pre-trained model (pruned at the third max pooling layer), with all layer weights unchanged during training (that is, we do not optimize the weights copied from VGG16 in our network). The input remains a 224 x 244 3-channel tensor as with the VGG16 model. However, the three channels are no longer red, green, and blue. Instead, each channel consists of a screenspace - left, center, and right from the heatmaps synthesized by the test subject's gaze pattern. We note this is an abuse of the original training of the VGG16 model, but we found the results to still be satisfactory. We hypothesize that, while the replacement of red-green-blue channels with screenspaces does not make intuitive sense, the neural network was

still able to make sense of representations.

In an effort to provide reproducible results, we provide details of the training and modeling in the following paragraph. Many terms used here are used without definition, but are common in the machine learning community and are ubiquitous in the Tensorflow library. Our implementation utilized the Tensorflow version of VGG16, so all input tensors were standardized between -1 and 1. The pruned VGG16 model's output is routed through two separable convolution layers (Chollet, 2017b) with 256 filters each, a kernel size of 3x3, and same size output padding. That output is then passed through a max pooling layer, with 2x2 pooling window size, and flattened. The flattened output is passed to three dense layers with a 0.5 dropout rate (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014); each subsequent layer is a step down by a power of two of the previous dense layer. Specifically, 128, 64, and 32 nodes, form the bottleneck. Finally, for the task agnostic model, one 32-node dense layer is followed by a three-node dense layer with a softmax activation (Boltzmann, 1868; Goodfellow, Bengio, & Courville, 2016). For the three-task model, the bottleneck output is routed to three individual 32-node dense layers. Each are followed by a three-node dense layer with softmax activations. All hidden layers utilize a ReLU activation function (Hahnloser, Sarpeshkar, Mahowald, Douglas, & Seung, 2000; Jarrett, Kavukcuoglu, Ranzato, & LeCun, 2009; Nair & Hinton, 2010). We use a batch size of 64, but do not implement batch-normalization (Ioffe & Szegedy, 2015) given the suggestions of Simonyan and Zisserman (2015).

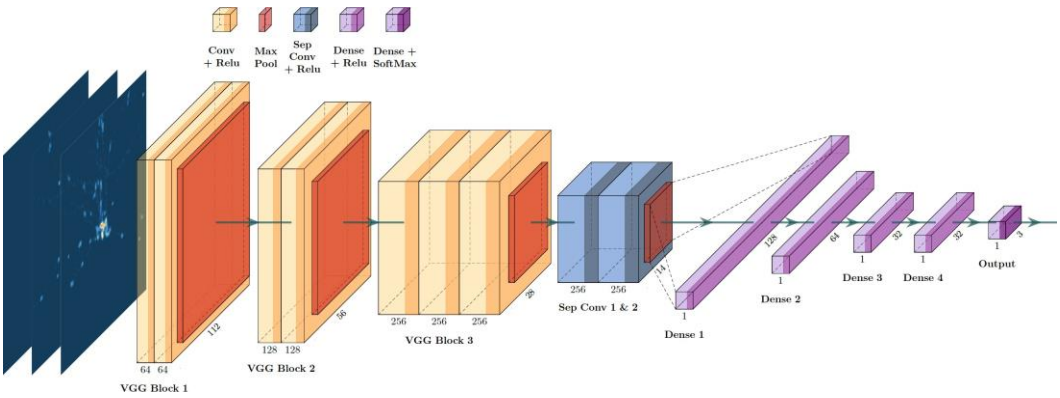


Figure 6. Task Agnostic Convolutional Neural Network.

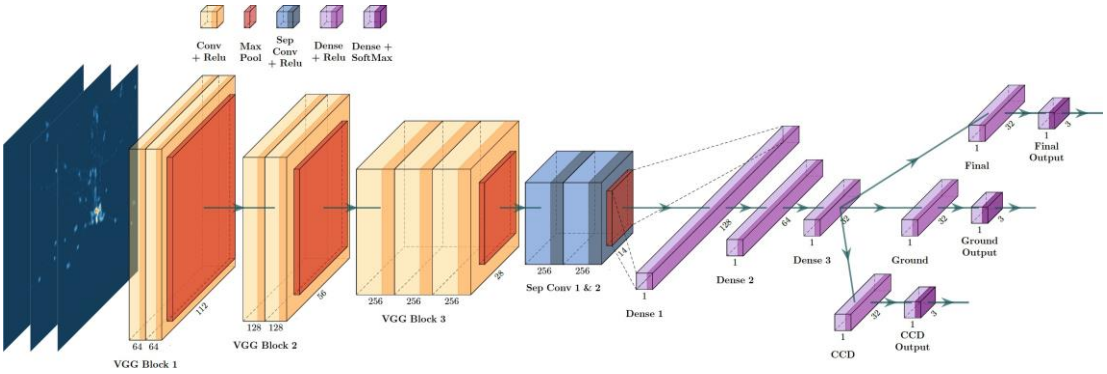


Figure 7. Multi-task Convolutional Neural Network.

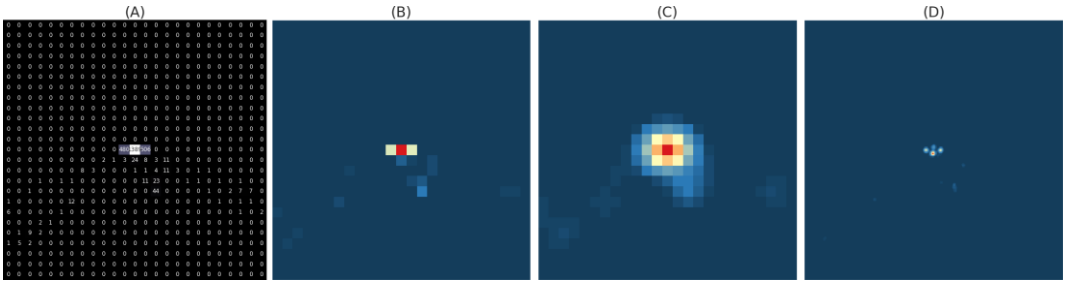


*Figure 8. Ce*nter Screenspace Scaled 25 X 25 (A) Count Matrix, (B) Heatmap, (C) Gaussian Blur, and (D) Actual 244 x 244.

Input Heatmap Generation

For input heatmap generation the gaze data provided by the BBXR was mapped to a two- dimensional tensor. The model input is of the shape 244 x 244 with three channels, the same shape as with VGG16. However, rather than using the RGB channels, each channel is a distinct screenspace generated from the gaze data over a thirty-second window, left, center, and right - in this order. While this method may introduce errors in the input distribution of VGG16, we did not observe any worrying behavior of the model; the VGG16 layers remained frozen and were not fine-tuned. Kernel density estimation, which was used to create visually appealing heatmaps for the instructors to review, was not used to create the heatmaps for the neural network because of the considerable amount of time needed to compute it. Instead, we employed a significantly faster approach for generating the heatmaps for the neural network. Specifically, to create the desired input tensor, we first mapped a consecutive thirty seconds of data to the desired resolution, 244 x 244. This was done by multiplying all values 0 and < 1 by 244 for both x and y values. Second, a count matrix was generated, where a 244 x 244 matrix of 0s is created and a 1 added to the position of each x and y pair. Third, the maximum value was calculated across all three screenspaces because the heat or

intensity was measured over all channels in the tensor, and all non-zero values are divided by this maximum value. Fourth, for a smoother input, a Gaussian blur with std of 1.0 was taken over the newly generated heatmap. Ultimately, this creates a lower resolution version of the Hi-def heatmap, for a given window of data, used for labeling—but this feature can be computed faster than realtime, making it appropriate for a number of applications. A scaled example of the center screenspace as it is transformed to a heatmap is shown in Figure 8. As noted, all three screenspaces are stacked to create a 244 x 244 x 3 tensor.

Training

For training we use the adaptive momentum stochastic optimization method (Kingma & Ba, 2017) with a learning rate range of 1e-6 to 1e-3, standard beta values, and gradient clipping range of 0.0 to 0.5. We chose the best hyper parameters based on the cross- validation results. Before training, we randomly separated 10% of the pilots for a final test set. The remaining 90% of the data was used for a 10-fold, across-subject cross-validation. That is, a given aviator cannot simultaneously be in a training fold and a validation fold. This resulted in validating folds consisting of about three aviators. We load balanced (*i.e.*, stratified) these test folds such that at least one of the three subjects was always a novice or operator and one was always a pilot. Otherwise the three subjects were chosen at random. Each fold was trained for fifty epochs. A final model was then trained with all the data, except the 10% portioned test dataset.

Data Augmentation

For deep learning applications, it is ideal to have as much data as possible for training. In the absence of numerous labeled data, an augmentation process can help to synthetically boost the number of training samples. This is commonly known as data augmentation — using existing labels and manipulating the input data to create new samples. Augmentation is only used for the training samples. That is, the testing samples remained unchanged.

For augmenting gaze data, we implemented several augmentation methods. These included: (1) removing or clipping a portion of the heatmap, (2) perturbing the heatmap within the reference frame along an axis, (3) mirroring the heatmap across the vertical or horizontal axes, and (4) having the rater label gaze patterns from unassociated maneuvers according to proper sight picture. This process resulted in approximately 9,400 training samples across all gaze quality classes in addition to the original 3,877 samples.

The clipping method takes advantage of the sight picture awareness and further modifies a window of data by creating a new heatmap. This involves generating a heatmap from a portion of the windowed data and relabeling the new pattern respective to the given phase. For this research we only modified heatmaps created from "correct" labeled data. For example, take the CCD generic phase. Aggregate the windowed data that are labeled "correct." These gaze patterns tend

to have a triangle-like shape over the HUD about the airspeed, altitude, and pitch ladder/flight path marker (FPM). We created a new heatmap by removing the left portion, any data with less than a ratio of 0.49 for the x-coordinate - gaze convergence over the airspeed indicator. We repeated this process again creating a second new heatmap by clipping the right side, any data greater than a ratio of 0.51 for the x-coordinate - gaze convergence over the altitude indicator. This method added twice the size of CCD class "correct" labels to the class "acceptable" labels. Post clipping, these heatmaps were verified or relabeled manually by the annotator and normalized by their largest value.

For this research we only perturbed the heatmaps vertically along the y-axis. Specifically, the heatmaps were adjusted up or down by ratios between -0.03 and 0.1. The heatmaps retained their original labels, and were further verified by the annotator. This method doubled the available training samples for our convolutional networks.

The mirroring method used flipped the heatmap about an axis, and then labeled appropriately given the sight picture, for each task. We only utilized mirroring about the y-axis. This was particularly useful for "poor" labeled gaze data that has heavy gaze fixation on one side or the other. Finally, gaze heatmaps from the other maneuvers, Section, were labeled. In conjunction with the concatenated videos we took advantage of the perceptual awareness of the raters, Section, by having the annotator mentally project the appropriate phase sight picture onto the maneuver being flown - both boundary avoidance and air intercept tasks. Augmented gaze pattern label distributions are found in Figure 4.

GAZE CLASSIFICATION RESULTS

As discussed, the training of each model included the 10-Fold, across-subject, cross-validation followed by a final fit over the entire dataset minus the 10% set aside for the test dataset. Results for both the 10% test set and the averaged cross-validation are presented. The training for both the task agnostic and multi-task models converged on or about the 20^{th} epoch of the optimization. The results of the task agnostic and multi-task models are discussed below, followed by a second inter-rater reliability analysis comparing the two models to the human gaze quality raters.

Task Agnostic Model

To evaluate the performance of each model, we choose to use the confusion matrix. A confusion matrix is a method for counting the number of observations where an instructor rated a gaze pattern quality one way and the model predicted the gaze pattern quality as either poor, fair, or correct. By counting the number of times the instructor and the model agree for each classification category and the number of times they disagree, we can formulate a

confusion matrix. Ideally, the matrix shows only values on the diagonal of the matrix, indicating that no confusions occurred. Figure 9(a) presents the task agnostic model's confusion matrix for both the average accuracy and average categorical true positive rates over all ten folds. Figure 9(b) characterizes the confusion matrix for test accuracy and true positive rates for each category.

The task agnostic model has an average fold accuracy of 89.9% with a test accuracy of 89.2%. The average true positive rates over all folds is 93.0% for poor, 89.0% for fair, and 84.0% for correct. The test dataset true positive rates are slightly different for each quality label with 90.0%, 68.0%, and 93.0% for poor, fair, and correct. The fair case is likely reflective of the potential task agnostic issues discussed previously. That is, an unseen heatmap from the test dataset is potentially classified incorrectly because it can have two different quality labels depending on the flight phase—and the agnostic model has no information regarding flight phase. Ultimately, this model accurately classifies gaze patterns that are poor 90.0%, fair 68.0%, correct 93.0% of the time. This may imply that a "correct" or "poor" gaze pattern generalizes across a number of flight phases, whereas fair patterns are more dependent on phase. Because of the low performance at grading "fair" gaze patterns, we conclude that the performance of the task-agnostic model is not reliable enough for use in automating the scoring of aviator gaze patterns. We therefore turn our focus to a more expressive model, the multi-task model, to understand if its performance is more consistent.

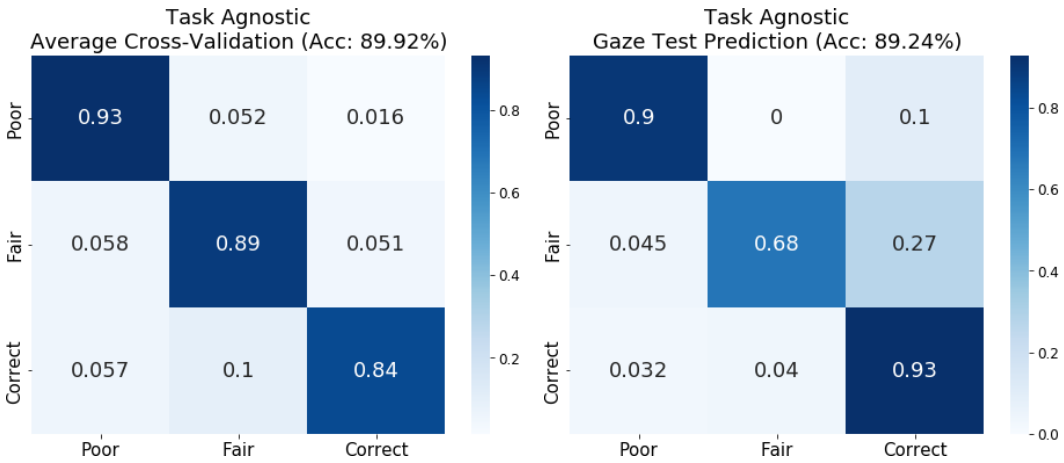


Figure 9. Task Agnostic model: (a) Combined confusion matrix over 10-folds (b) Confusion matrix over held out test dataset.

Multi-task Model

The confusion matrices for the multi-task model are more involved to interpret. Rather than presenting the confusion matrix overall, we must present one confusion matrix for each of the three flight maneuvers. Figure 10 (a-c, top row)

presents the multi-task model's confusion matrices as well as the average accuracy over all ten folds for each task. Figure 10 (d-f, bottom row) depicts the confusion matrix for test accuracy for each category. The "final approach" task was converted to a binary classification because too few examples were labeled as "fair" by the raters. Because there were not enough ground truth labels to train the model, we only report the binary classification "poor" versus "correct."

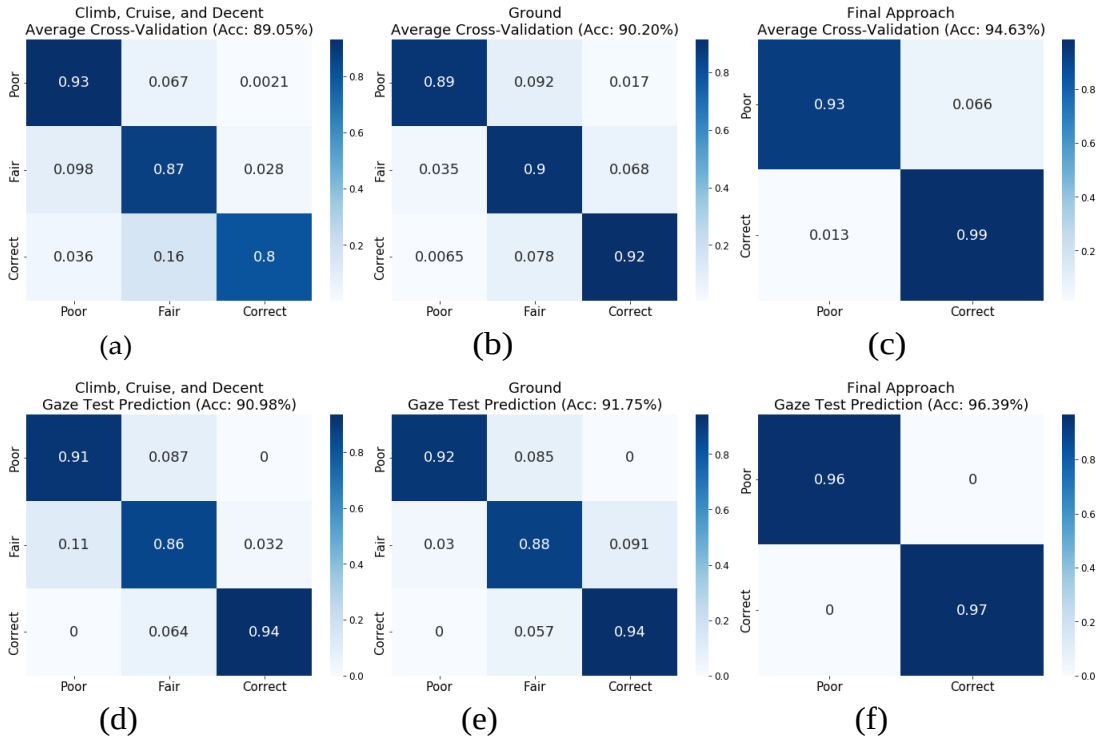


Figure 10. Multi-task model combined confusion matrices over: (a) 10-folds climb, cruise, and decent, (b) 10-folds ground, (c) 10-folds final approach (d) Climb, cruise, and decent test dataset, (e) Ground test dataset, and (f) Final approach test dataset.

The multi-task model has an average fold accuracy of 94.2% with a 93.0% average test accuracy across all tasks. Individually, the average fold accuracy for each task is 89.1%, 90.2%, and 94.63% with test accuracies of 91.0%, 91.8%, and 96.45%, for CCD, ground, and final approach, respectively. The test dataset true positive rates are comparatively stronger for each task than with our task agnostic model. CCD classifies 91.0%, 86.0%, and 94.0%, while the ground task classifies 92.0%, 88.0%, and 94.0%. Finally, the final approach task classifies 96.0% and 100.0% for "poor" and "fair" each. Overall, these results look promising and are substantially better than the task agnostic model. These results imply that the

model is robust to different tasks, but they do not indicate if the model is comparable to a human instructor. To elucidate if the model is significantly different than a trained instructor, we analyze inter-rater reliability, including the model as if it were a fourth instructor.

Human-Model Inter-Rater Reliability

We compared inter-rater reliability measures for the task agnostic model and multi-task model, separately. This analysis is identical to that carried out in Table 3 among raters, except we treat each model as if it were a human instructor. Table 4 presents the results of the IRR analysis between each rater and model, among all raters and each model, and among a subset of raters and each model. This analysis helps to answer the following question: *Does the model agree with human raters in a manner that is similar to how human raters agree with each other?*

Table 4
Inter-rater Reliability, Model as Additional Rater

Raters and Model	*r*	*Cohen's κ*	*Fleiss' κ*	*Randolph's κ*
Pilot I & Task Agnostic	.25	.18		
Pilot II & Task Agnostic	.42	.37		
CSO & Task Agnostic	.50	.43		
All Raters & Task Agnostic			.39	.61
Pilot I & Multi-task	.28	.23		
Pilot II & Multi-task	.73	.67		
CSO & Multi-task	.75	.70		
All Raters & Multi-task			.57	.77
Pilot I, II, & Multi-task			.53	.76
Pilot II, CSO, & Multi-task			.66	.85
Pilot I, CSO, & Multi-task			.36	.62

The **task agnostic model** has Cohen's κ of .18, .37, and .43 with Pilot I, Pilot II, and the CSO, respectively. These κ values signify some agreement with Pilot I and moderate agreement with Pilot II and the CSO. For the multi-rater variants, the Fleiss and Randolph κ's are .39 and .61, establishing moderate IRR agreement according to the scale defined by Landis and Koch (1977). This is consistent with our previous conclusion that the task agnostic model performance is not reliable enough for use in scoring aviator gaze patterns. For the **multi-task model** strong inter-rater reliability is achieved. The multi-task model and Pilot II have Cohen's κ of .67, substantial agreement. The relationship between the CSO and the model is a κ of .70, also a substantial agreement. This indicates strong agreement, especially given that CSO and Pilot II have a Cohen's κ of .71 (Table 3). Furthermore, Pilot II and the CSO each have an *r* value with the multi-

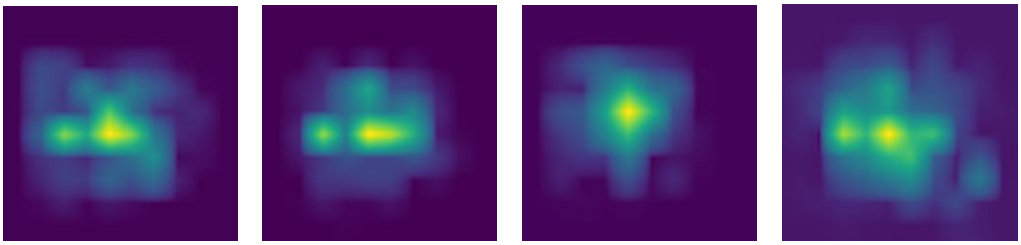
task model that signifies a strong linear relationship, with .73 and .75 respectively. The least agreement is between the multi-task model and pilot I, which has a κ of .23. While lower, this is similar to the CSO and Pilot I, which have a κ of .37 (Table 3) — the lowest agreement between human raters.

For the multi-rater case, the model and all three raters have a Fleiss' κ of .57—which is a slightly better Fleiss' κ than that of the all human raters, .56 (Table 3). Both rater combinations have the same Randolph's κ of .77, signifying substantial agreement. If Pilot I is removed from the multi-rater analysis, the highest multi-rater κ values are achieved, .66, Fleiss, and .85, Randolph - almost perfect agreement.

In summary, for inter-rater reliability among the raters and the models, it is apparent that both models are similar to the respective relationships among the full dataset annotator, the CSO, and the two pilot raters — that is the CSO and Pilot I having shown less agreement than the CSO and Pilot II. ***We therefore conclude that the performance of the multi-task model is sufficient to be used as an automated gaze scoring tool, which is directly affirming of our hypothesis for research question two.***

DISCUSSION

Overall, the performance of the multi-task model was comparable to an instructor for verifying gaze quality. Therefore, it should be possible to deploy this model for applications such as (1) augmenting instructor observations or (2) training pilots to better scan for different maneuvers automatically in a real-time environment. The processing time required for the model is primarily due to the time needed for collecting gaze points. If implemented as a pipeline and primed with the initial observation window that slides overtime, the model could support a frame rate of greater than 30 FPS. The actual time required for just the classification portion took, on average, less than 400 ms running on a 2018 Mac Book Pro. Therefore, the creation of an interface which, in real-time, displays the predicted gaze pattern quality for the pilot, is possible. This can assist pilots in adjusting their technique during practice sessions or in mission execution. One limitation of using a CNN to classify gaze is that the results are empirical and linked to the setup used for training. This means that the models investigated in this study are relevant only for the heads-up-display and instrument display in our simulator. While it is assumed that gaze pattern classification will generalize to other instrument display panels, each display configuration requires the training of a new CNN model, especially if it differs significantly from the panel used in this study.



(a) CCD task (b) Ground task (c) Final task (d) Negative example

Figure 11. Zoomed Grad-CAM heatmaps at the separable convolution activation layer of the multi-task model, (see Figure 7); all Grad-CAMs were computed respective of the 'correct' label except for the negative example, which is labeled 'poor.'

In the same manner, pattern quality could be used by an instructor in a dashboard for a class of student pilots. Such a dashboard might increase the number of simultaneous pilots an instructor could effectively observe in a training setting. One limitation of the current model is that it can only display the overall quality of the gaze pattern, but cannot display what corrective actions a pilot might take to increase the gaze pattern quality. In future work, it may be possible to use gradient class activation mapping (Grad-CAM), Figure 11, to trace back to what portions of the input heatmap cause the model to have degraded performance (Selvaraju et al., 2017). From this knowledge, it should be possible to interpret the Grad-CAM output to instructions such as "*Check your airspeed more often to improve gaze quality*", "*Keep your airspeed and wingman in your scan, you are fixated on the altitude indicator."* With this kind of approach, a multi-modal variant could be created that takes into account aircraft state, gaze pattern quality, Grad-CAM, and a student's site picture informing a student how to improve upon this site picture and scan with specific inputs. As an example, on final approach, the system could say something like, "*You are not keeping your airspeed in your scan, pitch down* 2° *for airspeed*." Further, there are several ways to solve a glide slope problem on approach. This kind of model could inform an optimal solution based on the student perspective - with repetition, potentially improving perceptual awareness (Miller & Gleason, 1947).

Given the performance of the multi-task model, it is preferable compared to the task agnostic model. However, this does introduce some complexities for deployment. For instance, in a real-time pilot feedback interface, the inference system would need to understand or be informed which flight phase the pilot was undertaking. Some of these maneuvers are easily categorized, such as using the "weight-on-wheels" signal of the aircraft to know when the aircraft is on the ground. Others, however, would require additional classification of the phase, sub-phase, or require the instructor/pilot to select what phase or maneuver is currently being undertaken. While potentially minor, it does add an extra layer to a

system that we prefer to be completely automatic. More complex gaze pattern classifiers can be built using fundamental gaze patterns, such as those shown in this study. Given continued development there is potential to further mitigate the need for manual task selection. The results of our research may also be applied outside of the virtual environment, in actual flight. For an apples-to-apples comparison, Figure 12 depicts the data from a complete final approach for both the mixed-reality simulator and a real-life flight conducted on a C-17 military cargo aircraft. This example is an instance of gaze pattern data having been collected on a real flight. What has not been shown is how the gaze pattern quality can be assessed automatically in real flight, there are several hurdles that will need to be overcome. The noise sources of sunlight, head/body movement from G-forces, and overall head movement in an actual flight could potentially reduce the accuracy of the model. In future work, deploying this model to real flight would need research into the noise sources and sensitivity of the model to this additional noise. Even so, we hypothesize that the model could work in real flights because the simulations are high quality, such that noise sources from focal length and maneuver specific noise are already captured well by the model. Future work will investigate this more systematically.

(a) (b)

Figure 12. Aggregate example heatmaps for final approach: (a) zoomed in heatmap of HUD, (b) Aggregate real-life flight on final approach in a C-17 aircraft (Martin, Calhoun, Schnell, & Thompson, 2019).

Finally, another limitation of our study is that our pilot population is comprised of individuals with military training. Moreover, the flight scenario employed used an instrument panel traditionally used in a fighter type aircraft. This may limit the generalizability of our findings to other styles of aircraft and training experience. Even so, modern commercial aircraft can be equipped with a HUD. The Boeing 787 includes a HUD as standard equipment (Nicholl, 2014). Furthermore, the positions of airspeed, altitude, and pitch ladder are typically standardized (Federal Aviation Administration, 2014). What is not standard is

HUD implementation (Nicholl, 2014). While there is no evidence to indicate the proposed methods would not work in a civilian aircraft, further investigation is warranted to ensure this conclusion.

CONCLUSION

In this research, we use convolutional neural networks to classify gaze or scan pattern quality for aviators in a multi-device, mixed reality aviation environment. We designed a human subjects experiment to inform the design and evaluation of these models with 40 subjects performing common flight maneuvers. We recruited three subject matter experts to rate the gaze patterns and analyzed their agreement, showing they have strong inter-rater reliability. Our multi-task convolutional neural network matched subject matter experts with greater than 93% average accuracy and strong multi-rater agreement, a κ of .77.

This result suggests that gaze patterns for various flight maneuvers can be automatically classified into three levels of quality with reliable accuracy and in near-real time. This automated gaze classification may be of use in establishing the context of an aviator while they are learning a particular flight maneuver. The scope of our conclusions is limited to gaze patterns in the scenarios from our experiments, but gaze classification for additional flight maneuvers or for other activities in other domains may also be applicable. We leave the investigation of gaze quality classification in additional flight maneuvers and other disciplines to future work.

REFERENCES

Abdelrahman, Y., Khan, A. A., Newn, J., Velloso, E., Safwat, S. A., Bailey, J., . . . Schmidt, A. (2019, September). Classifying attention types with thermal imaging and eye tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technology, 3*(3), 69:1–69:27. Retrieved from http://doi.acm.org/10.1145/3351227

Barz, M., Daiber, F., & Bulling, A. (2016). Prediction of Gaze estimation error for error-aware Gaze-based interfaces. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications* (pp. 275–278). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/ 10.1145/2857491.2857493

Barz, M., & Sonntag, D. (2016). Gaze-guided object classification using deep neural networks for attention-based computing. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (pp. 253–256). New York, NY, USA: ACM. doi:10.1145/2968219.2971389

Baxter, J., Caruana, R., Mitchell, T., Pratt, L. Y., Silver, D. L., & Thurn, S. (1995). *Post-NIPS*95 Workshop on Transfer in Inductive Systems.* Retrieved 2019-11- 10, from http://socrates.acadiau.ca/courses/comp/dsilver/ NIPS95\ _LTL/\\transfer.workshop.1995.html

Boltzmann, L. (1868). Studien über das Gleichgewicht der lebendigen Kraft zwischen bewegten materiellen Punkten. *Wiener Berichte, 5*, 517–560.

Calhoun, P. (2016). Darpa emerging technologies. *Strategic Studies Quarterly*, 91–113.

Carpenter, R. H. (1977). *Movements of the eyes*. London: Pion Limited.

Chollet, F. (2017a). *Deep learning with python, chapter 5* (1st ed.). Greenwich, CT: Manning.

Chollet, F. (2017b, July). Xception: DeeplLearning with Depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1800– 1807). (ISSN: 1063-6919) doi: 10.1109/CVPR.2017.195

Cohen, J. (1960, April). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20* (1), 37–46. Retrieved 2019-11-18, from http://journals.sagepub.com/ doi/10.1177/001316446002000104 doi: 10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70* (4), 213–220. doi: 10.1037/h0026256

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Duchowski, A. T., Price, M. M., Meyer, M., & Orero, P. (2012). Aggregate gaze visualization with real-time heatmaps. In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12* (p. 13). Santa Barbara, CA: ACM Press. Retrieved 2019-11- 21, from http://dl.acm.org/citation.cfm?doid=2168556.2168558 doi: 10.1145/2168556.2168558

Endsley, M. R., & Garland, D. J. (2000a, July). Pilot situation awareness training in general aviation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 44* (11), 357–360. Retrieved 2019-11-21, from https://doi.org/10.1177/154193120004401107 doi:10.1177/154193120004401107

Endsley, M. R., & Garland, D. J. (2000b). *Situation awareness analysis and measurement*. Boca Raton, FL: CRC Press.

Endsley, M. R., & Robertson, M. M. (2000). Training for situation awareness in individuals and teams. *Situation Awareness Analysis and Measurement*, 349–366.

Federal Aviation Administration. (2014). *Advisory circular "25-11B: Electronic flight displays*. Washington, DC: Author.

Federal Aviation Administration, & Soucie, D. (2017). *Airplane Flying Handbook (Federal Aviation Administration): FAA-H-8083-3B, Chapter 3 & 8*. Skyhorse.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76* (5), 378–382. doi: 10.1037/h0031619

Frutos-Pascual, M., & Garcia-Zapirain, B. (2015, May). Assessing visual attention using eye tracking sensors in intelligent cognitive therapies based on serious games. *Sensors, 15*(5), 11092–11117. doi:10.3390/s150511092

Fuller, I. R., Johnston, N., & McDonald, N. (1995). A taxonomy of situation awareness errors. In *Human Factors and Ergonomics Society Annual Meeting Proceedings, 32*. doi:10.1177/154193128803200221

Goldberg, J. H., & Schryver, J. C. (1993, December). *Eye-gaze determination of user intent at the computer interface* (Tech. Rep. No. CONF-9308228-1). Oak Ridge National Lab., TN (United States). Retrieved 2019-11-21, from https://www.osti.gov/biblio/10153103-T4ixQl/native/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: The MIT Press.

Gray, W. (2008). A generalized handling qualities flight test technique utilizing boundary avoidance tracking. In *2008 US Air Force T&E Days* (p. 1648).

Guitton, D., & Volle, M. (1987, September). Gaze control in humans: eye-head coordination during orienting movements to targets within and beyond the oculomotor range. *Journal of Neurophysiology, 58* (3), 427–459. Retrieved from https://www.physiology.org/. doi:10.1152/jn.1987.58.3.427

Hahnloser, R. H. R., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., & Seung, H. S. (2000, June). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature, 405* (6789), 947–951. Retrieved 2019-11-10, from https://www.nature.com/articles/35016072 doi:10.1038/35016072

Hanson, T. (2018, November). *L3 introduces first-ever high-fidelity, mixed reality deployable training simulator.* Retrieved 2019-11-05, from https://www.l3t.com/link/press/ l3-introduces-first-ever-high-fidelity-mixed-reality-deployable-training-simulator

Huston, S. J., & Krapp, H. G. (2008, July). Visuomotor transformation in the fly gaze stabilization system. *PLOS Biology*, *6*(7), e173. doi:10.1371/journal.pbio.0060173

Ioffe, S., & Szegedy, C. (2015, March). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167 [cs]*. Retrieved 2019-11-10, from http://arxiv.org/abs/1502.03167 (arXiv: 1502.03167)

Jarrett, K., Kavukcuoglu, K., Ranzato, M. A., & LeCun, Y. (2009, September). What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision* (pp. 2146–2153). Kyoto: IEEE. Retrieved 2019-11-10, from http://ieeexplore.ieee.org/ document/5459469/ doi: 10.1109/ICCV.2009.5459469

Kershner, W. K. (2001). *The student's pilot's flight manual: From first flight to private certificate* (9th ed.). Ames, IA: Iowa State Pr.

Kingma, D. P., & Ba, J. (2017, January). Adam: A method for stochastic optimization. *arXiv:1412.6980 [cs]*. Retrieved 2019-11-11, from http://arxiv.org/abs/1412.6980 (arXiv: 1412.6980 version: 8)

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 1097–1105). Curran As- sociates, Inc. Retrieved 2019-11-10, from http://papers.nips.cc/paper/ 4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. Retrieved 2019-11-19, from www.jstor.org/stable/2529310 doi: 10.2307/2529310

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks, 3361* (10), 1995.

Li, B., Mettler, B., & Andersh, J. (2015, October). Classification of human gaze in spatial guidance and control. In *2015 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 1073–1080). (ISSN: null)

doi:10.1109/SMC.2015.193

Lounis, C., Peysakhovich, V., & Causse, M. (2018). Intelligent cockpit: Eye tracking integration to enhance the pilot-aircraft interaction. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (pp. 74:1–74:3). New York, NY, USA: ACM. doi:10.1145/3204493.3207420

Lounis, C., Peysakhovich, V., & Causse, M. (2020). Flight eye tracking assistant (FETA): Proof of concept. In N. Stanton (Ed.), *Advances in Human Factors of Transportation* (Vol. 964, pp. 739–751). Cham: Springer International Publishing. doi:10.1007/978-3-030-20503-4_66

Martin, P., Calhoun, P., Schnell, T., & Thompson, C. (2019, September). Objective measures of pilot workload. *63RD Setp Symposium Proceedings*. Retrieved from https://secure.whoglue.net/setp_admin/papers/SETP%20Pilot%20Workload%20Study%20PA%20Released.pdf

Merriam-Webster. (2020, February). *Definition of sight picture* [Dictionary]. Retrieved 2020- 02-25, from https://www.merriam-webster.com/ dictionary/ sight+picture

Miller, N. E., & Gleason, J. G. (1947). *Psychological research on pilot training* (No. 8). Washington, DC: US Government Printing Office.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine.*

Learning. (pp. 807–814). USA: Omnipress. Retrieved 2019-11-10, from http://dl.acm.org/ citation.cfm?id=3104322.3104425 (event-place: Haifa, Israel)

Newn, J., Velloso, E., Allison, F., Abdelrahman, Y., & Vetere, F. (2017). Evaluating real-time gaze representations to infer intentions in competitive turn-based strategy games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '17* (pp. 541–552). Amsterdam, The Netherlands: ACM Press. Retrieved 2019-11-21, from http:// dl.acm.org/citation.cfm?doid=3116595.3116624 doi:10.1145/3116595.3116624

Nicholl, R. (2014). *Airline head-up display systems: Human factors considerations*. *A*vailable at SSRN 2384101.

Pan, S. J., & Yang, Q. (2010, October). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. doi:10.1109/TKDE.2009.191

Privitera, C. M. (2006, February). The scanpath theory: Its definition and later developments. In *Human Vision and Electronic Imaging XI* (Vol. 6057, p. 60570A). International Society for Optics and Photonics. Retrieved 2019-11-21, from https://www.spiedigitallibrary.org/ conference-proceedings-of-

spie/6057/60570A/The-scanpath-theory-its-definition-and-later-developments/10.1117/12.674146.short doi: 10.1117/12.674146

Randolph, J. J. (2005). *Free-marginal multirater kappa (multirater K[free]): An alternative to Fleiss' fixed-marginal multirater kappa*. Retrieved 2019-11-08, from https://eric.ed.gov/?id=ED490661

Robinson, D. A. (1964). The mechanics of human saccadic eye movement. *The Journal of Physiology*, *174*(2), 245–264. Retrieved 2019-11-21, from https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1964.sp007485doi: 10.1113/jphysiol.1964.sp007485

Robinson, D. A. (1965). The mechanics of human smooth pursuit eye movement. *The Journal of Physiology*, *180*(3), 569–591. Retrieved 2019-11-21, from https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1965.sp007718doi: 10.1113/ jphysiol.1965.sp007718

Ruder, S. (2017, June). An overview of multi-task learning in deep neural networks. *arXiv:1706.05098 [cs, stat]*. Retrieved 2019-02-01, from http://arxiv.org/abs/1706.05098 (arXiv: 1706.05098)

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252. doi:10.1007/s11263-015-0816-y

Schnell, T., Keller, M., & Poolman, P. (2008, October). Neurophysiological workload assessment in flight. In *2008 IEEE/AIAA 27th Digital Avionics Systems Conference* (pp. 4.B.2–1–4.B.2–14). (ISSN: 2155-7209) doi:10.1109/DASC.2008.4702827

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations.*

Sohlberg, M. M., & Mateer, C. A. (1987, April). Effectiveness of an attention-training program. *Journal of Clinical and Experimental Neuropsychology*, *9* (2), 117–130. Retrieved 2019-11-22, from https://doi.org/10.1080/01688638708405352

Spakov, O. (2008). *iComponent – device-independent platform for analyzing eye movement data and developing eye-based applications* (Doctoral dissertation). University of Tampere, Finland.

Spakov, O., & Miniotas, D. (2007). Visualization of eye gaze data using heat maps. *Elektronika ir elektrotechnika*, (2), 55–58. Retrieved 2019-11-21, from https://vb.vgtu.lt/object/elaba:6113807/

Spitzer, C., Ferrell, U., & Ferrell, T. (2017). *Digital avionics handbook, Chapter 17* (3rd ed.). Boca Raton, FL: CRC press.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014, June). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.

Sroka, G., Feldman, L. S., Vassiliou, M. C., Kaneva, P. A., Fayez, R., & Fried, G. M. (2010, January). Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room—a randomized controlled trial. *The American Journal of Surgery, 199*(1), 115–120. doi:10.1016/j.amjsurg.2009.07.035

Stellmach, S., Nacke, L., & Dachselt, R. (2010). Advanced gaze visualizations for three-dimensional virtual environments. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (pp. 109–112). New York, NY, USA: ACM. Retrieved 2019-11-21, from http://doi.acm.org/10.1145/1743666.1743693 (event-place: Austin, Texas)

United States Air Force. (2019, June). *Air Force manual 11-217 flight operations*. Retrieved from https://static.e-publishing.af.mil/production/1/af_a3/publication/afman11-217/afman11-217.pdf

United States Navy. (2011, May). *Flight training instruction primary formation T-6B.* Retrieved 2020-02-26, from https://www.cnatra.navy.mil/local/docs/pat-pubs/P-766.pdf

United States Navy. (2015, February). *Flight training instruction contact helicopter advanced phase TH-57C.* Retrieved 2020-02-26 from https://www.cnatra.navy.mil/ local/docs/pat-pubs/P-457.pdf

United States Navy. (2019, April). *Flight training instruction primary contact T–6B.* Retrieved 2020-02-26 from https://www.cnatra.navy.mil/local/docs/pat-pubs/P-764.pdf

Valerie, A., Huemer, Hayashi, M., Renema, F., Elkins, S., McCandless, J. W., & McCann, R. S. (2005). *Characterizing scan patterns in a spacecraft cockpit simulator: Expert vs. novice performance*. doi:10.1177/154193120504900119

Vrzakova, H., & Bednarik, R. (2012). Hard lessons learned: Mobile eye-tracking in cockpits. In *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction* (pp. 7:1–7:6). New York, NY: ACM. Retrieved 2019-11-19 from http://doi.acm.org/ 10.1145/2401836.2401843(event-place: Santa Monica, California)

Weibel, N., Fouse, A., Emmenegger, C., Kimmich, S., & Hutchins, E. (2012). Let's look at the cockpit: Exploring mobile eye-tracking for observational research on the flight deck. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 107–114). New York, NY: ACM.

Retrieved 2019-11-19, from http://doi.acm.org/10.1145/ 2168556.2168573 (event-place: Santa Barbara, California)

APPENDIX A: HEATMAP EXAMPLES

The following are examples of poor, fair, and correct labeled heatmaps for each model task (flight phase) over a 30 second window. Note that this is not exhaustive, for example Figure 13 (c) is also "correct" for the final approach task.

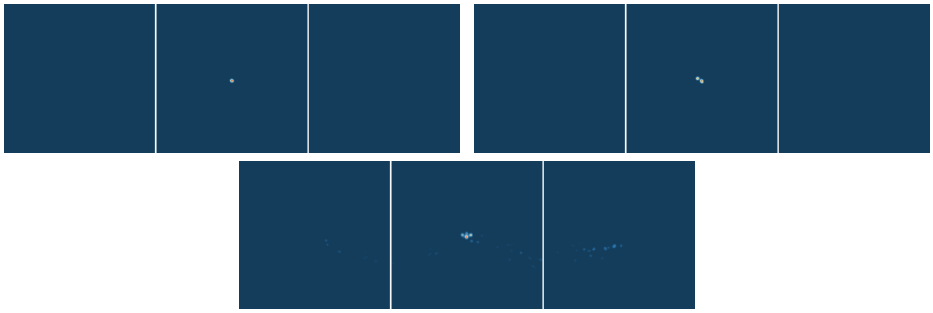


Figure 13. Climb, cruise, and decent task examples: poor (left), fair (right), and correct (bottom).

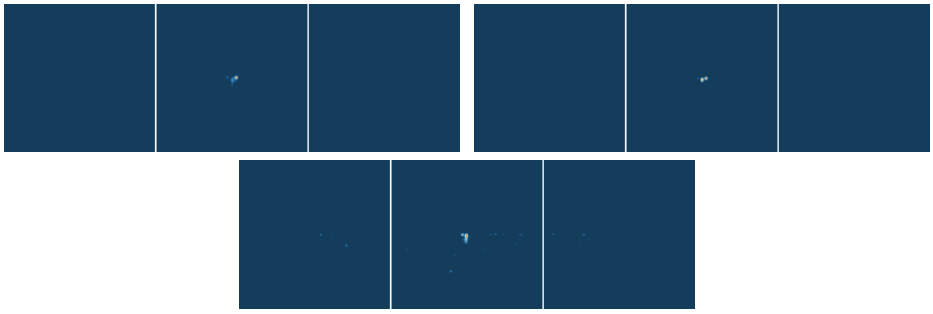


Figure 14. Ground task: poor (left), fair (right), and correct (bottom).

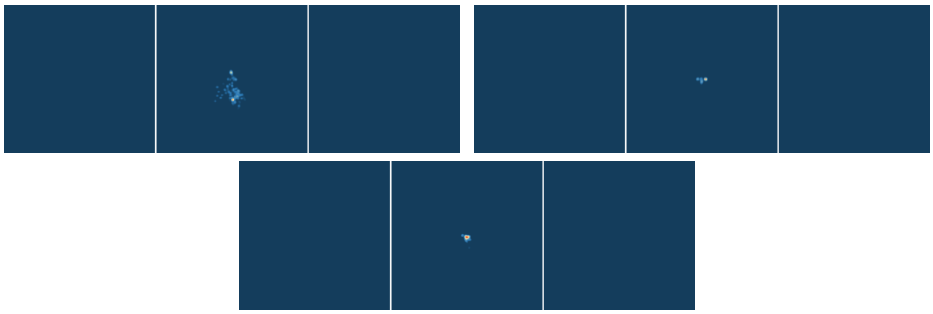


Figure 15. Final approach task: poor (left), fair (right), and correct (bottom).

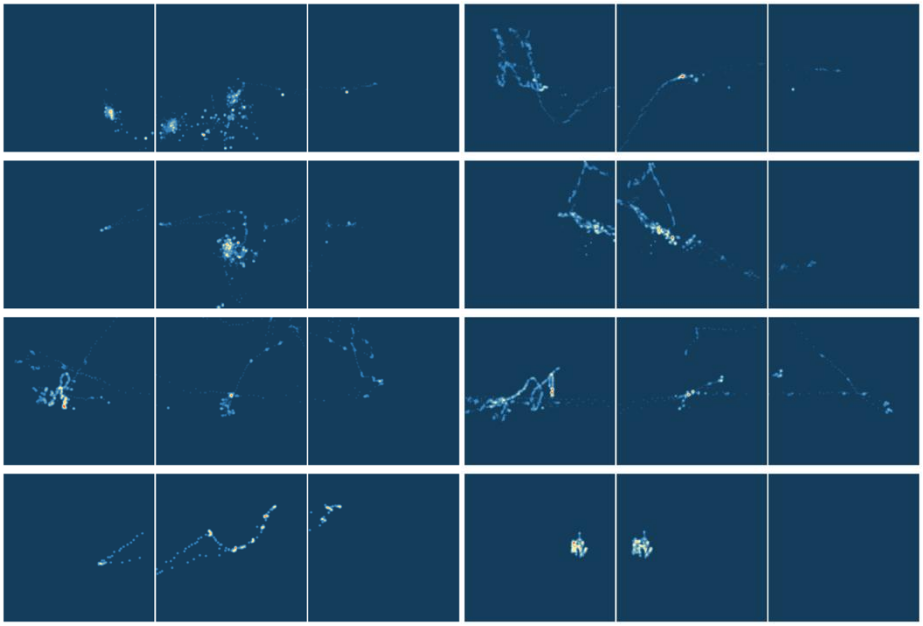


Figure 16. Negative heatmap examples: all heatmaps are rated 'poor' for CCD, ground, and final approach. All cover a 30-second window period.

For each example in Figure 16, these heatmaps are negative examples and are labeled poor for all three model tasks (flight phases) — CCD, ground, and final approach. All cover a 30-second window period.

APPENDIX B: HARDWARE AND SOFTWARE

For this research, we used the Python versions of Scikit-Learn, TensorFlow and Keras. Both video concatenation with high-resolution heatmaps and model training were conducted on the Southern Methodist University high-performance compute cluster (HPC). Because the pruned VGG model remained frozen and the dataset is small, training was also conducted on a desktop machine using an NVIDIA RTX 2080 ti, with some memory limitations. The need for the HPC is due to memory requirements and less so computational power. Multiple folds could not be stored in RAM. Further, for batch size, GPU memory was a factor.

At 870 TFLOPS, the **HPC** has 354 nodes, 11,276 AVX2 Intel CPU cores, 275,968 accelerator cores, 120 TB in total memory, 100 Gb/s node interconnect bandwidth, and 2.8 PB of scratch space. The accelerator nodes include 36 NVIDIA P100 GPUs with 16 GB CoWoS HBM2 memory and 24 NVIDIA V100 GPUs with 32 GB of CoWoS HBM2 memory. The **desktop machine** included a 12 core AMD Ryzen 9 3900X, 64 GB of DDR4 RAM, SSDs totaling 6 TB, and an NVIDIA RTX 2080 TI with 4352 cores and 11 GB GDDR6 memory.