Published in final edited form as:
*Med Image Anal*. 2012 August ; 16(6): 1280–1292. doi:10.1016/j.media.2012.04.007.

# Strategies for Improved Interpretation of Computer-Aided Detections for CT Colonography Utilizing Distributed Human Intelligence

**Matthew T. McKenna**[a], **Shijun Wang, Ph.D.**[a], **Tan B. Nguyen**[a], **Joseph E. Burns, M.D., Ph.D.**[a,b], **Nicholas Petrick, Ph.D.**[c], and **Ronald M. Summers, M.D., Ph.D.**[a]

[a]Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Building 10, Room 1C224, MSC 1182, Bethesda, MD 20892-1182

[b]Department of Radiological Sciences, University of California, Irvine Medical Center, 101 The vCity Drive South, Orange, CA 92868

[c]Center for Devices and Radiological Health, U.S. Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993-0002

## Abstract

Computer-aided detection (CAD) systems have been shown to improve the diagnostic performance of CT colonography (CTC) in the detection of premalignant colorectal polyps. Despite the improvement, the overall system is not optimal. CAD annotations on true lesions are incorrectly dismissed, and false positives are misinterpreted as true polyps. Here, we conduct an observer performance study utilizing distributed human intelligence in the form of anonymous knowledge workers (KWs) to investigate human performance in classifying polyp candidates under different presentation strategies. We evaluated 600 polyp candidates from 50 patients, each case having at least one polyp • 6 mm, from a large database of CTC studies. Each polyp candidate was labeled independently as a true or false polyp by 20 KWs and an expert radiologist. We asked each labeler to determine whether the candidate was a true polyp after looking at a single 3D-rendered image of the candidate and after watching a video fly-around of the candidate. We found that distributed human intelligence improved significantly when presented with the additional information in the video fly-around. We noted that performance degraded with increasing interpretation time and increasing difficulty, but distributed human intelligence performed better than our CAD classifier for “easy” and “moderate” polyp candidates. Further, we observed numerous parallels between the expert radiologist and the KWs. Both showed similar improvement in classification moving from single-image to video interpretation. Additionally, difficulty estimates obtained from the KWs using an expectation maximization algorithm correlated well with the difficulty rating assigned by the expert radiologist. Our results suggest that distributed human intelligence is a powerful tool that will aid in the development of CAD for CTC.

Corresponding Author and Reprint Requests: Ronald M. Summers, M.D., Ph.D., Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bldg. 10 Room 1C224D MSC 1182, BETHESDA MD 20892-1182, Phone: (301) 402-5486, FAX: (301) 451-5721, rms@nih.gov, Web: http://www.cc.nih.gov/drd/summers.html.

**Keywords**

Computed tomography colonography; observer performance study; crowdsourcing; distributed human intelligence; video analysis; labeling modeling

## 1 INTRODUCTION

Colorectal cancer is the second-leading cause of cancer death in Americans (Jemal et al., 2010). Colorectal cancer is a largely preventable disease as the removal of colorectal polyps, the precursor to malignancy, is known to be curative in most patients. Tests that are effective at detection of colorectal polyps include colonoscopy and CT colonography (Smith et al., 2010). Both colonoscopy and CT colonography (CTC) are tests that are performed and interpreted by trained physicians. In the past few years, computer-aided polyp detection (CAD) software has been developed and shown to improve the diagnostic performance of CT colonography when interpreted by radiologists (Li et al., 2009; Nappi and Yoshida, 2009; Summers et al., 2005; Suzuki et al., 2010; Wang et al., 2010; Zhu et al., 2010).

While CAD systems often improve the sensitivity of radiologists, this benefit is coupled with a drop in specificity (Dachman et al., 2010). Conversely, CAD markings on true polyps have been incorrectly dismissed as false positives by interpreting radiologists (Taylor et al., 2009). In clinical trials investigating CAD for CTC, wide ranges of sensitivities and high intra- and interobserver variability have been noted (Cotton et al., 2004; Johnson, 2008; Johnson et al., 2003; Petrick et al., 2008; Pickhardt et al., 2003; Rockey et al., 2005). It is of interest to understand the factors that lead to incorrect diagnosis at CTC by radiologists and by CAD. Methods to investigate training and interpretation techniques and to quantitatively evaluate various datasets also would be desirable. Such understanding could lead to higher, more consistent performance in CTC.

Distributed human intelligence, a form of web-based crowdsourcing, utilizes large numbers of lay people (referred to as “knowledge workers”) to complete various tasks requiring human intelligence. It is a relatively new phenomenon, but it has already been applied with great success in various areas of scientific research. An example is *FoldIt*, a software application structured in the form of an online game that lets users manipulate 3D protein models, trying to find the correct conformations (Cooper et al., 2010). The knowledge gained through this Internet game has led to improved algorithms to predict 3D protein structures from nucleotide sequences. Research into RNA folding and multiple sequence alignment in DNA also have been crowdsourced in the games *EteRNA* (http://eterna.cmu.edu/content/EteRNA) and *Phylo* (http://phylo.cs.mcgill.ca/). Crowdsourcing has expedited the annotation of datasets, which had been a major bottleneck in machine learning research. The *ESP Game* and *reCAPTCHA* system are used to collect annotations on image and text databases (von Ahn and Dabbish, 2004; von Ahn et al., 2008). In research to evaluate the validity of crowdsourced data, investigators have developed techniques to evaluate individual annotators and ensure data is comparable to that generated by experts (Ipeirotis et al., 2010; Raykar et al., 2010; Snow et al., 2008; Whitehill et al., 2009). Crowdsourcing can offer an efficient and reliable means to collect data, evaluate algorithm performance, and gain insight into human decision-making in a number of different areas of research.

We utilize distributed human intelligence to gain insight into the classification of polyp candidates from CTC data. We are interested in a hybrid approach to developing CAD systems – using both traditional, computational metrics as well as human vision. We are searching for a combination of detection characteristics and presentation format that will synergistically improve performance. We realize that large cohorts of experts are difficult and expensive to acquire; however, the utility of a system is measured by observer improvement. To address this problem, we propose and characterize a crowdsourced-human observer model.

In a previous study we found that anonymous knowledge workers (KWs) with minimal training could effectively differentiate polyps from non-polyps on CT colonography images with performance comparable to a CAD system for this classification task (Nguyen et al., 2012). Building on that study, we conducted an observer performance study with KWs to measure the effect of training and the CAD presentation style for the task of interpreting CTC CAD. Specifically, we evaluate KW performance in classification after observing a single image of a CAD mark and after watching a video fly-around of the mark. We also measure how a new rendering function and training data influence KW performance. We compare KWs with an expert radiologist in interpreting these CAD marks. We also comment on practical issues, such as compensation and KW filtering, in conducting these distributed intelligent observer studies.

In this paper, we show how distributed human intelligence can be used to conduct sophisticated observer studies to evaluate training and presentation style for medical imaging classification tasks. We evaluate the impact on reader performance when KWs are shown multiple perspectives of a polyp candidate compared with just showing a single view of the candidate along. We compare KW performance with that of an expert radiologist to evaluate the impact of reader training on performance. The purpose of this study is to demonstrate the feasibility of using distributed human intelligence to conduct observer performance studies for medical imaging applications areas. Specifically, we investigate how human performance in classifying CTC CAD polyp candidates is impacted by the presentation style used to display those candidates. The observations and experience gained from this online perception study may be useful in guiding CTC CAD algorithm development.

## 2 MATERIALS AND METHODS

The study was approved by our institution’s Office of Human Subjects Research, both for the retrospective use of the anonymized patient dataset and for the participation of the knowledge workers. The requirement for informed consent was waived.

### 2.1 Case Selection

We selected two independent sets of 50 patients each from a database of patients from three medical centers originally accrued during the study described by Pickhardt et al. (Pickhardt et al., 2003). The first of these sets served as training data to optimize parameters in our CAD system. The second set was used as test data to present to the KWs. We included all patients from our original study (Nguyen et al., 2012) in the test set. We supplemented this set with patients selected sequentially from each medical center. We required that each patient had at least one polyp •6 mm confirmed by histopathological evaluation following optical colonoscopy, and we rejected patients with poor preparation (i.e. poor insufflation) that prevented creation of videos as described in 2.7. Patient characteristics are shown in Table 1.

### 2.2 Bowel preparation and CT Scanning

Patients underwent a standard 24-hour colonic preparation (Pickhardt and Choi, 2003). Each patient was scanned in the supine and prone positions during a single breath hold using a 4-channel or 8-channel CT scanner (General Electric LightSpeed or LightSpeed Ultra, GE

Healthcare Technologies, Waukesha, WI) during a single imaging appointment. CT scanning parameters included 1.25- to 2.5-mm section collimation, 15 mm/s table speed, 1-mm reconstruction interval, 100 mAs, and 120 kVp.

### 2.3 Computer-aided polyp detection algorithm

CT images were analyzed using our computer-aided polyp detection software package described previously (Li et al., 2009). A support vector machine (SVM) committee classifier was trained on 5337 detections identified in the independent training set of 50 cases, for an average of 107 detections per patient. As each patient was scanned in both supine and prone positions (thus there were two CT scans per patient in the CAD data set), 53 of the 67 polyps in the training cases resulted in 91 of these detections. At the operating point corresponding to 10 false positives per patient, the classifier achieved a sensitivity of 0.86 with a specificity of 0.90 in the training set. When applied to the test set, the trained SVM classifier assigned a score to each detection, ranging from 0 to 1. Higher scores represent higher confidence that the detection is a true polyp. Characteristics of the polyps in our training data are shown in Table 1.

### 2.4 Experimental dataset selection

We applied our trained CAD system to our test set. The CAD system initially identified 4866 detections in the 100 testing set CT scans, for an average of 49 detections per scan (97 detections per patient). The system detected 65 of the 75 polyps confirmed by optical colonoscopy in this 50 patient data set. Of the 4866 detections, 112 were on the 65 true polyps distributed as follows. Forty-three of the polyps were detected once each on both the prone and supine scans, 11 were detected once on only the prone scan, 9 were detected once on only the supine scan, one polyp was detected twice on the prone scan and once on the supine, and one polyp was detected twice on the supine scan and once on the prone. To reduce this initial set of detections to a manageable dataset we could distribute to KWs, we used free-response operating characteristic analysis. We selected an operating point corresponding to 10 false positive detections per patient while maintaining a sensitivity of 0.741. We only used the 600 detections exceeding the SVM score threshold (•0.5714) corresponding to our operating point. Of these 600 detections, 88 represented 59 confirmed true polyps distributed as follows. Twenty-seven polyps were detected on both scans, 17 were detected on only the supine scan, 14 were detected on only the prone scan, and one polyp was detected twice on the supine scan and once on the prone. Characteristics of the polyps in our experimental set are shown in Table 1.

An expert radiologist categorized each detection based on its structure and labeled each detection as “easy”, “moderate”, or “difficult” based on perceived difficulty for a reader to correctly identify the detection as a true polyp or false positive. “Easy” detections were those whose categorizations as true or false positives were immediately obvious at a glance with limited training. “Difficult” detections were those whose categorizations were not immediately obvious and might require additional knowledge or information, or those that looked like an obvious polyp but were not polyps based on the reference standard. “Moderate” detections were those of intermediate difficulty.

### 2.5 Distributed human intelligence

In this study, we employed Amazon’s Mechanical Turk (MTurk) (https://www.mturk.com; Amazon.com, Inc.) web service to recruit initially untrained, anonymous workers to perform polyp classification on our dataset. MTurk is an Internet-based crowdsourcing platform that allows requesters to distribute small computer-based tasks to a large number of knowledge workers (KWs). KWs receive a small monetary reward from the requester for each human intelligence task (HIT) that they complete. Requesters also can reward high-performing KWs with additional monetary bonuses.

We generated and published one HIT on the MTurk platform for each CAD polyp candidate and asked 20 KWs to complete each HIT. By combining the results from multiple KWs who each worked on a different set of polyp candidates, we created a system of distributed human intelligence that reflected the KW's collective judgment. An expert radiologist also completed each HIT. Our aim in this study was to improve the performance of KWs from our original experiment. We tried to accomplish this in multiple ways by: (1) utilizing a new rendering scheme for the creation of images, (2) presenting the KWs with multiple perspectives on each detection in the form of a video, (3) implementing a new training module to better explain the task, (4) utilizing a qualification test to stratify and select KWs, and (5) offering a monetary reward for good performance.

### 2.6 Rendering Function

Volumetric ray-casting with perspective projection was used with a segmented, but uncleansed, CTC dataset to generate the images (Yao et al., 2004). A two dimensional opacity transfer function was created for each polyp candidate, which varied with local CT intensity values and gradient measures. This transfer function was subject to three design constraints: soft tissue (the colon wall and polyps) should be opaque, oral contrast in the colon lumen should be transparent, and high gradients at air-fluid interfaces should be semi-transparent.

We found that fluid artifacts and improper subtraction were a significant source of incorrect votes by the KWs in our initial experiments. Thus, we decided to create renderings using uncleansed datasets, to alert the KWs to the potential of subtraction artifacts. When the polyp candidate was located near contrast, the contrast was rendered transparent as it could affect proper interpretation of the detection. Otherwise, the contrast was rendered opaque. Such an approach seeks to illustrate improperly-segmented contrast and avoid the creation of polyp-like structures arising from poor segmentation (see Figure 4 for some examples of contrast artifacts). A color transfer function was set in conjunction with the opacity transfer function to illustrate the difference between tissue and contrast. Contrast was rendered white, matching its appearance in the 2D CT scan slices. Tissue was rendered a red color. A similar approach where tagged stool was color-coded and presented during interpretation was shown to increase efficiency of CTC reading while maintaining diagnostic accuracy (Park et al., 2008). The gradient measure was implemented to avoid volume averaging effects at air-fluid interfaces which result in a thin film with intensities comparable to tissue. The rendering pipeline was implemented with the Visualization Toolkit in Java using MATLAB (Version 7.10.0 (R2010a), MathWorks) (Schroeder et al., 2003). The rendering functions were inspired by those used in (Zhu et al., 2010).

### 2.7 Viewpoint Generation and Ranking

We used the colon segmentation result to generate multiple viewpoints for each polyp candidate. To generate the viewpoints, we first aligned a sampled hemisphere (81 points) with the measured surface normal of the polyp candidate. Using each point on the hemisphere to define a direction, a camera was iteratively moved in each direction starting at the polyp candidate centroid. The camera movement was stopped when it either hit tissue or exceeded a maximum distance from the centroid. A 2D illustration of this process is shown in Figure 1. Generating viewpoints in this way would ensure visibility of the centroid, and greater distances would allow us to display contextual information which could be of diagnostic importance.

We ranked each viewpoint to guide the creation of the video sequence. We iteratively populated a ranking list, initially using three criteria to judge the viewpoints: alignment with the principal components of the polyp candidate, distance from centroid, and alignment with the fluid normal. As the ranked list became populated we penalized subsequent viewpoints for sharing a similar alignment with a previously-selected viewpoint. In our observations, these criteria produced a range of informative viewpoints.

The three principal components of the polyp candidate were extracted from the 3D arrangement of the voxels marked as part of the detection using principal component analysis. Assuming viewpoints on the same side as the polyp candidate normal are more informative, the principal component vectors were aligned with the normal such that the scalar product of each vector with the normal was positive. We also assumed viewpoints situated at 45 degrees with respect to each principal component would be informative. Such viewpoints were chosen to illustrate the shape of the detection.

A distance score which increased with distance from the detection centroid was assigned to ensure the polyp candidate was not distorted in creating the image. Since we used perspective projection to generate the 2D images, a viewpoint very close to the polyp candidate could distort the candidate. Farther viewpoints also allow for capture of structures of possible diagnostic significance surrounding the detection.

Alignment with the fluid normal was considered to account for the presence of contrast in the images. Volume averaging effects at air-fluid interfaces resulted in a thin film with intensities comparable to tissue. While the gradient component of the transfer function was designed to render the film transparent, a viewpoint that runs parallel to such a surface would produce an occluded image of the polyp candidate. Rendering performance improves as the viewpoint is increasingly orthogonal to the fluid’s surface. Viewpoints that penetrate this air-fluid interface were assigned scores based on their alignment with the surface normal of the fluid.

To ensure different viewpoints were selected, a given viewpoint was penalized for having a similar alignment with previously selected viewpoints. Viewpoints that shared a similar viewing angle to a previously-selected viewpoint were given low scores.

### 2.8 Video Creation

Assuming that the attention span of the KWs would be limited, these ranked viewpoints had to be combined to efficiently illustrate the polyp candidate. We wanted to ensure that the higher-ranked viewpoints would be seen earlier in the video while still visiting as many predefined viewpoints as possible. Since the viewpoint selection process did not explicitly consider image properties, it was important to visit the lower-ranked viewpoints as these could be diagnostically significant. Colonic structures surrounding the polyp candidate also had to be avoided in generating the final camera path as passing through such structures would create occluded images. We utilized an undirected graph to solve this problem.

Each camera position was entered as a node on the graph, and pairwise linear connections were made between each camera position. The edges of the graph were set by the linear distance measure between each camera. These edges were weighted by a distance rank. For example, the closest viewpoint’s edge would be multiplied by 1, the next closest viewpoint’s edge would be multiplied by 2, etc. If no linear connection could be made, e.g. tissue was present between the two camera positions, no edge was constructed. After visiting a node, the value of each of its edges was increased to discourage revisiting the node. To define the sequence of camera positions, we used iterative runs of Dijkstra’s Algorithm to find the lowest-cost path connecting the two highest-ranked, unvisited camera positions (Dijkstra, 1959). By defining the graph in this way, the search was encouraged to proceed to closer neighbors instead of taking a direct route to the next highest-ranked position. A direct route would have been the lowest-cost path had we not modulated the distance by the distance rank. This approach increased the number of positions reached while still visiting the higher-ranked viewpoints and avoided colonic structures. The sequenced viewpoints were linearly interpolated to construct a final, smooth camera path around the polyp candidate.

When generating the video, the camera was focused on the polyp candidate centroid and was aligned with the surface normal of the polyp candidate. This ensured a smooth transition between frames in the video. The viewing angle of the camera was set to ensure the entire polyp candidate would be visible. The detection's voxels were first projected onto the plane running through the detection centroid and orthogonal to the viewing direction. The distance from the centroid to each projected point was measured, and the maximum distance, $R$, was recorded. The viewing angle set using the geometric relationship between the camera's distance from the detection centroid and $R$. The viewing angle was constrained between 70 and 120 degrees. These bounds were selected to allow surrounding structure to be seen while avoiding severe distortion of the polyp candidate in the images. The detection was marked by a green cube, and the video was generated at 8 frames per second. The final videos were each 1-minute, volume-rendered, intraluminal fly-arounds of the polyp candidate.

### 2.9 Single Image Creation

We had to generate a single image of each candidate to measure the difference between single image and video interpretation. The single image was created in the same way as the video using the top-ranked viewpoint from the video generation process.

### 2.10 Qualification Test and Training Module

Before working on our HITs, KWs were directed to a training site and asked to complete a qualification test. The training site contained information about colonic polyps and outlined an approach for completing the HITs. We asked the KWs first to examine the 2D CT scan sections to identify the primary material present in the detection and then to use the 3D reconstructions to evaluate the shape of the detection. The training page also included 6 examples of polyps and 6 examples of common false positives. Each example was illustrated with 2D CT scan sections and a 3D rendering, and the corresponding video was shown when the rendering was clicked. We provided a caption for each example to highlight important aspects of the images. The examples were selected to familiarize the KWs with the rendered images and to address common structures seen in CTC as noted from our original experiment. The training page was hosted separately from the MTurk interface and was available for reference at any time (http://mturk.dreamhosters.com/training.html NOTE TO EDITORS AND REVIEWERS: this link is currently available for review and can be made available in an online supplement if desired). This site contained more information than the training module used in the first experiment. The original training module used the same description of a polyp and showed examples of 5 true polyps and 6 false positives. Each example was illustrated with 2D CT scan sections and a volume rendering. Notably, the original training did not contain explicit instructions on how to interpret the images. Further, the examples in the original training were not captioned, only labeled with the ground truth information (true/false positive). Those examples also did not contain video.

The qualification test consisted of 5 questions. Each question showed a multi-perspective video and the axial, coronal, and sagittal CT scan sections of a detection from the training set. The KWs were asked to determine whether the images show a polyp. The detections in the test were similar to examples seen in the training, and 4 of the 5 questions were deemed “easy” by an expert radiologist. Two questions showed true polyps, and 3 showed common false positives. The KWs were required to correctly answer at least 4 of 5 to work on the HITs. The KWs were also expected to have an approval rating of >95% on the MTurk platform to participate in this study. A KW’s approval rating is defined as the ratio of assignments approved by MTurk requesters to the total number of assignments submitted by a KW, and it is part of each KW’s MTurk profile. These steps were taken to ensure the KWs were reliable and had a basic understanding of the task, and to eliminate the KWs who only voted “yes” in the first experiment.

To further encourage KWs to submit high-quality results, we offered a $5 USD bonus to be paid to the top 10 workers upon completion of the experiment. We defined “best workers” as those who had the highest average sensitivity and specificity across the single image and video interpretation ($Score = (Se_S+Sp_S+Se_V+Sp_V)/4$, where $Se_S$ and $Se_V$ are the fractions of true positive marks correctly classified using the single image and video respectively and $Sp_S$ and $Sp_V$ are the fractions of false positive marks correctly classified using the single image and video respectively). KWs had to complete more than 100 HITs to be considered for the reward. We informed the KWs of the bonus in the instructions to the qualification test: “As a special reward, the 10 best workers will each receive a $5 bonus for their work.” We did not reveal the specific ranking system to the KWs.

### 2.11 HIT Design

For each HIT, the KWs were asked to answer three questions. First, they were asked to examine three two-dimensional CT scan sections running through the detection from axial, sagittal, and coronal views. They had to identify whether the detection marked “Air”, “Tissue”, or “Fluid”. After answering, the single image of the polyp candidate was revealed. The KWs were asked to decide whether the image showed a polyp. Finally, the KWs were shown the video fly-around of the detection. KWs were required to watch at least 5 seconds of the video before answering whether the video showed a polyp. To account for the variable Internet bandwidth of KWs, we included a script in the HIT to start the video timer only when the video had finished buffering and had started to play. In this way, the KWs would have a smooth viewing experience, and we could remove some of the variability of KW bandwidth. We were not concerned about the bandwidth issue for the single image interpretations as those files were relatively small (around 65KB each for volume-rendered images and 8KB each for the CT scan sections). We recorded the decision times for each of these questions. KWs were blinded to the proportion of polyp candidates that were true positives and true negatives in the dataset. Each worker was given 20 minutes to complete the assignment and was paid $0.01 USD upon completion. This design to measure the difference between single image and video interpretation could be implemented on the MTurk platform and allowed for an efficient testing without bias of the difference (Obuchowski et al., 2010). A sample HIT can be seen in Figure 2.

### 2.12 Statistical analysis

The primary objective was to compare the KW’s area under the receiver operating characteristic (ROC) curve (AUC) using single images and using videos. The unit of analysis for constructing the ROC curves was the CAD polyp candidates, or detections. Sensitivity was defined as the fraction of true polyps correctly classified. Specificity was defined as the fraction of false positives correctly classified. Polyp classification for CAD was based on the SVM score for each detection. Polyp classification for KWs was based on the combined KW score calculated as the ratio of votes for true polyp to the total number of votes for each detection. Higher scores are expected to indicate higher confidence that the detection represents a true polyp. Empiric ROC curves were constructed and analyzed, and AUCs were calculated as Wilcoxon-Mann-Whitney statistics. AUCs were compared using a univariate z-score test statistic for correlation of areas from parametric ROC curves, for paired and unpaired data as appropriate, using ROCKIT 0.9.1 (University of Chicago) (Metz et al., 1998). Detections were assumed to be independent.

Several secondary analyses of the AUC were performed as described above. To quantify the effect of the new rendering scheme and training information, we compared the KW's AUC from our original experiment (Nguyen et al., 2012) with that from our current experiment. In our original experiment, KWs were trained using a different training module, and polyp candidates were presented with a set of five images – three two-dimensional CT scan sections from the axial, sagittal, and coronal views, and two three-dimensional reconstructions of the polyp candidate, one with a CAD mark and an identical image without the CAD mark. The three-dimensional reconstructions were generated using V3D-Colon (Viatronix, Stony Brook, NY). We only compared those detections that shared a volume overlap between the two experiments. AUCs for KWs and CAD also were compared for detections stratified by difficulty as determined by an expert radiologist. The AUC for KWs who scored a 5 on the qualification test was compared with the AUC for KWs who scored a 4 on the qualification test.

We also utilized the detectability index to describe our ROC curves. The detectability index, $d_a$, measures the normalized distance between the decision distributions for negative and positive cases. It can be calculated directly from the AUC using a standard formula, $d_a = \sqrt{2}z(A_w)$ (Burgess, 1995). In this formula, $A_w$ is the area under the ROC curve, and $z$ represents the z-score function calculated from the cumulative normal distribution. In practice, $d_a$ represents the perceptual signal-to-noise ratio as experienced by the KWs. While the AUC statistic fails to accurately capture changes in signal-to-noise ratio as it approaches unity, $d_a$ is designed to scale reasonably with the SNR. Larger $d_a$ values imply higher discriminatory ability.

Timing data were compared on a per-detection and per-worker basis. The interpretation time for each detection was defined as the average interpretation time of the KWs who evaluated the detection. This timing data was stratified by difficulty, and multi-variate ANOVA was used to compare the means. The interpretation time for KWs was normalized using the z-score statistic. Detections were collected over ranges in the z-score. Detections in each bin were treated independently, and sensitivity and specificity were calculated for each bin. Regression analysis was performed to measure how interpretation time correlated with sensitivity and specificity.

The performance of KWs over the course of completing HITs was investigated. We divided the HITs into two groups for each worker. The first group contained the first half of HITs each worker completed and the second group contained the second half of HITs each worker completed. We compared the ROC curves and timing data under each interpretation style for each half.

Several methods exist to learn the ground truth from the noisy, non-expert labeling of KWs (Dawid and Skene, 1979; Ipeirotis et al., 2010; Raykar et al., 2010; Whitehill et al., 2009). These methods use probabilistic approaches to estimate the ground truth when no absolute ground truth exists. These models often learn other parameters, such as worker quality, in estimating the ground truth. We investigated the perceived difficulty of each detection using the model proposed in (Whitehill et al., 2009). This model uses an expectation maximization algorithm to estimate ground truth labels from the binary labels assigned by KWs while accounting for worker quality and detection difficulty:

$$p(L_{ij}=Z_j \mid \alpha_i, \beta_j)=\frac{1}{1+e^{-\alpha_i \beta_j}} \qquad (1)$$

where $L_{ij}$ is the label assigned to image $j$ by KW $i$, $Z_j$ is the true label, $\alpha_i$ represents the expertise of KW $i$, $\beta_j$ represents the difficulty of image $j$, and $p$ is the probability that the label $L_{ij}$ matches the true label $Z_j$. $\alpha_i = (-\infty,\infty)$, where $\alpha = \infty$ means the KW always labels the image correctly, and $\alpha = -\infty$ means the KW is adversarial and always labels the image incorrectly. β is constrained to be positive, and 1/β represents image difficulty. $1/\beta = [0,\infty)$, where $1/\beta = 0$ means the image is so easy every KW will assign the correct label, and $1/\beta = \infty$ means the image is so difficult that even the best KW will have only a 50% chance of assigning a correct label. We defined the learned difficulty for each detection as the average of the difficulties calculated using the single-image and video results. We stratified the model predictions, $p$, by the difficulty assigned by an expert radiologist and compared the distributions using multi-variate ANOVA.

All data collection and analyses were performed with Amazon's MTurk web interface, Microsoft Office Excel, MATLAB, and ROCKIT. Numbers are reported as values ± standard error unless otherwise specified.

# 3 RESULTS

## 3.1 Experimental Characteristics

This experiment consisted of 600 HITs each completed by 20 KWs, for a total of 12,000 HIT results. These HITs were published on MTurk April 20, 2011, and the experiment concluded July 5, 2011. The distribution of detections by difficulties and detection categories can be seen in Figure 3. Some example detections are shown in Figure 4 and Figure 5.

## 3.2 Performance comparison

The detection-level AUCs and detectability indices, $d_a$, for KWs using single images and KWs using videos are shown in Table 2 with the corresponding ROC curves in Figure 6. The video AUC was significantly greater than the single image AUC, and we saw a 14% improvement in $d_a$ when moving from single image to video interpretation. The radiologist saw a large improvement in sensitivity with a small improvement in specificity when moving from the single image to the video.

## 3.3 Knowledge worker characteristics

Four hundred and fourteen KWs attempted the qualification test. Of those, 256 passed and 158 failed. One hundred and twenty-nine KWs scored a 5/5 on the test, 127 scored a 4/5, 76 scored a 3/5, 51 scored a 2/5, 17 scored a 1/5, and 14 KWs did not answer any questions correctly. Of the KWs who passed the qualification test, only 160 submitted HITs. Each knowledge worker completed, on average, 75 (±158) assignments. The average amount of time spent on each assignment was 51 (±71) seconds. Twenty-nine KWs completed more than 100 HITs, accounting for 86% of all completed assignments. The remaining 131 KWs completed only 14% of all completed assignments.

We show performance statistics for these groups in Table 3 with corresponding ROC curves in Figure 7. Sixty-seven KWs who scored a 4 on the qualification test ("four-workers") completed 5940 assignments, for an average of 89 (±175) assignments per worker. Ninety-three KWs who scored a 5 on the qualification test ("five-workers") completed 6060 assignments, for an average of 65 (±145) assignments per worker. Five-workers outperformed four-workers on both single-image (*p*=0.003) and video (*p*=0.065) interpretation. The $d_a$ measure suggest that the video interpretation aided four-workers more than the five-workers (21% improvement in four-workers versus 6% improvement in 5-workers) (Table 2). Interestingly, the top 10 KWs, ranked as described above, were comparable to a radiologist in completing this task (Figure 8).

### 3.4 Performance by detection difficulty

The detection-level AUC's for detections stratified by difficulty are shown in Table 2 and Table 4 with the corresponding ROC curves in Figure 9. CAD SVM performance is shown in Table 5. For easy detections, there is little difference between the AUC and $d_a$ for single image and video interpretation. There is an improvement in video performance over single-image performance for moderate detections. KWs outperform CAD on both easy and moderate detections ($p$<0.001 for each interpretation style). For difficult detections, the difference between video and single-image performance is larger and significant. However, no improvement over CAD was noted for these difficult detections ($p$ = 0.19 and 0.98 for single image and video interpretation respectively). The trend in $d_a$ reflects the AUC trend, showing a growing difference between presentation styles with increasing difficulty. Again, radiologist performance shows a decrease in both sensitivity and specificity with increased detection difficulty. This is consistent with the lower KW ROC curves observed with increasing detection difficulty.

### 3.5 Timing information

The average video interpretation times for each detection, stratified by difficulty, are shown in Figure 10. Easy, moderate, and difficult detections took 17.61 (±7.86), 19.38 (±7.31), and 20.36 (±7.50) seconds to interpret respectively. Increasing difficulty correlated with an increase in interpretation time. There was a significant difference ($p$<0.01) between the interpretation time for easy and difficult detections. We noted a similar pattern for single image interpretation times, but there was only a small difference in times among the difficulties. Easy, moderate, and difficult detections took 3.60 (±1.90), 3.65 (±1.72), 3.87 (±1.90) seconds to interpret respectively. None of these differences were significant ($p$>0.5). We noted that specificity tended to decrease as KW interpretation time increased as seen in Figure 11. The $R^2$ values for the single image and video specificity with respect to interpretation time were 0.95 and 0.93 respectively. We noted no such trend with sensitivity. Similar trends were noted in the radiologist's performance. The radiologist's average total work times (time from accepting HIT to submitting HIT) for easy, moderate, and difficult detections were 23.5 (±28.0), 33.3 (±22.8), and 47.6 (±23.0) seconds respectively. There was a significant difference between each of these groups ($p$<0.01).

### 3.6 Comparison of Trials

We found 183 detections, including 31 true positives and 152 false positives, in the current study that shared a volume overlap with a detection used in our original experiment. The corresponding ROC curves and radiologist's operating points are shown in Figure 12. The AUCs from our original experiment, current experiment using a single image, and current experiment using a video were 0.834 (±0.047), 0.909 (±0.036), and 0.947 (±0.029), respectively. We noted a significant increase in AUC in single image interpretation from our original experiment to the current experiment ($p$=0.0147). There was also a significant increase in AUC from single image to video interpretation in the current experiment ($p$=0.0290). The expert radiologist showed an increase in sensitivity and specificity between the original experiment and the current single image interpretation and between the current single image interpretation and the video interpretation. This trend is consistent with the changes in KWs ROC observed across the two studies.

### 3.7 KW Change over Time

The ROC curves for the first half of HITs compared to the second half is shown in Figure 13 with the corresponding AUC values in Table 6. We also show the average interpretation times for the halves in Table 6. We only considered the 29 workers who completed more than 100 HITs. We found significant differences between single image and video interpretation in each half ($p = 0.011$ and 0.001 respectively). However, we found no significant difference between performance on videos in the second half and single images in the first half ($p = 0.260$). We also noted a decrease in average interpretation time from the first half to the second half; however, these differences were not significant. The detectability index was consistent for video interpretation, but it fell 12% from the first half to the second half for single image interpretation.

### 3.8 Estimated Difficulty

The predictions of model (1) for each detection stratified by the difficulty assigned by an expert can be seen in Figure 14. We calculated two difficulty estimates ($\beta$) for each image, one from KW responses when using single images and the other from KW responses when using video. We took the average of those scores to represent image difficulty. We evaluated the model using those average difficulties for a worker with $\alpha_i = 1$. Thus, the model only varied with the difficulty estimates. The means for the model predictions ($p$ – the probability that the KW will assign the correct label) for easy, moderate, and difficult detections were 0.834 (±0.141), 0.771 (±0.133), and 0.702 (±0.098), respectively. We found a significant difference between each group ($p<0.01$). The model predicted that a KW would have a lower chance of assigning a correct label to a difficult detection than assigning a correct label to an easy detection. Given these distributions in Figure 14, it appears that the model was able to estimate reasonably well the difficulty from KW votes.

## 4 DISCUSSION

In this paper, we presented results from an observer study utilizing Internet-based distributed human intelligence to measure the effect of training and detection presentation in interpretation of CT colonography CAD. We found that there was a significant improvement in KW performance when a multi-perspective video was used over a single image in detection classification. KWs were able to outperform the classifier in our CAD algorithm for detections rated as easy or moderate. The new image generation technique and training also yielded a significant increase in performance over that seen in our original experiment.

Similar trends in interpretation were noted in both an expert radiologist and the KWs. Both experienced an overall increase in sensitivity with a small difference in specificity when interpreting the video and a decrease in performance with increasing difficulty. Both groups shared a similar opinion on the difficulty of detections, and both took longer to interpret more difficult detections. While there was a range of KW performance observed, the best KWs performed similarly to an expert radiologist. It has been noted in the radiology literature that longer interpretation times are associated with a decrease in specificity (Nodine et al., 2002; Saunders and Samei, 2006). We observed the same trend in our current experiment as specificity dropped as interpretation time increased. While we do expect that a radiologist would outperform KWs on a full reading of a CTC dataset, it is nevertheless interesting that the two groups achieved high performance and experienced similar trends in this focused task. Such findings open the possibility of using KWs to investigate other finite task based interpretation factors at a fraction of the cost and time of a traditional study involving radiologists. Additionally, KWs could be used as a platform to investigate various teaching paradigms for physicians in training to see how each paradigm translates to practice.

We used our first experiment as a benchmark to measure the difference in worker selection, training, and image generation. Analyzing the detections with volume overlap between the experiments, there was significant improvement in performance seen from our original experiment to our current experiment. In our first study, we noted that KW performance was consistent, noting no difference in KW AUC between two independent trials. We conclude then that the improvement observed in our current study can be ascribed to the changes implemented in our current experiment, notably the new training information, different image capture techniques, more stringent qualification requirements, and the possibility of a monetary reward. More detailed training and more stringent qualification requirements could further increase the performance and reliability of crowdsourcing experiments. However, time considerations must be made before changing requirements. Our current experiment took significantly longer than our first (1 week versus 11 weeks) most likely because of the requisite qualifications and compensation. The MTurk platform requires requesters to compete in order to attract KWs to their HITs, and higher wages may be the best way to attract a large KW base to quickly complete HITs.

We hypothesized that experience in interpreting CAD would serve as a form of training, and KW performance would improve with the number of detections interpreted. Judging by the decrease in interpretation time from the first half of detections to the second half, it seemed that the KWs did become more efficient at the task. We also noted that while single image and video interpretation in each half were significantly different, the performance using video in the second half was similar to the performance using single images in the first half. The detections were presented in a random order to KWs, so these changes probably reflect KW characteristics more than detection characteristics. These trends suggest that KWs may have changed their decision criteria as they completed more assignments. With an evolved definition of a polyp, the KWs may have only identified certain types of polyps in later HITs. While we tried to encourage consistently high performance by offering a $5 reward for the best KWs, this bonus and the $0.01 HIT compensation apparently were not sufficient to keep KWs operating at a high level as they completed increasing numbers of HITs. These results also could be explained by the “laboratory effect,” where performance in a laboratory setting is significantly worse than performance in the clinic (Gur et al., 2008). The KWs knew that their decisions had no impact on patient care, so they could have relaxed their standards in interpreting the detections in order to maximize their pay. A similar argument can be used to explain the decrease in time. KWs can make more money if they work faster and pay was not dependent on accuracy. In the clinic, both accuracy and throughput are relevant. Further experiments could investigate alternative strategies to maintain and improve performance in the course of completing HITs, such as randomly introducing training cases in the normal workflow and providing feedback on performance.

In evaluating KWs, the qualification test scores proved to be a more accurate indicator of performance than number of HITs interpreted. Five-workers significantly outperformed four-workers. Qualification tests may be a better way of evaluating readers than number of cases interpreted. This has potentially important implications as number of cases interpreted is frequently, and perhaps erroneously, used as a surrogate measure of expertise in many clinical publications.

Our current study demonstrates the sophistication that a crowdsourced experiment can achieve. The specific goal of this experiment was to investigate factors that may influence diagnostic accuracy in order to improve the design of CAD systems and improve CTC training. From this experiment, we can quickly gather data to evaluate different interpretation paradigms and identify areas for improvement. These experiments were both more practical and much less costly to conduct compared to observer studies utilizing radiologists. We hypothesize that data collected from crowdsourcing experiments will be useful in determining how best to refine and improve CAD algorithms as well as improving physician training by identifying ambiguous features that lead to incorrect interpretation. With developments in processing crowdsourced data, even more sophisticated measures can be made. We envision such data being incorporated in new CAD systems that fuse machine intelligence with human intelligence (Wang et al., 2011). Novel features could be extracted from the human perception of polyp candidates. Ensemble CAD systems could be developed in which classifiers separately handle detections grouped by perceived difficulty, as features that prove effective on easy detections may degrade performance on more difficult detections. Additionally, crowdsourcing could be used to evaluate datasets used by various CAD systems to allow for a more direct comparison of their performance.

While these results are interesting, it is important to realize the limitations in our experiment. The results were obtained from untrained observers removed from the clinic. It is encouraging that the changes we observed among our KW observers followed the trends of an expert radiologist who completed the experiment; however, we must establish a more substantial connection between trained radiologists and the KWs who complete our studies. We only used a single radiologist in this study to complete HITs and assign difficulty scores. Please note that our expert first encountered the images when completing the HITs. The radiologist completed the same HITs as the KWs. After completing each HIT, the ground truth of the detection was revealed, and we asked the radiologist to categorize the detection and assign a difficulty score at that time. Further, we only asked KWs to interpret detections identified by our CAD system. We tried to simply the task as much as possible, omitting some potentially important information to KW. For example, we chose not to describe the problem of residual stool – a type of false positive that causes much difficulty in practice. Despite these limitations, it is interesting that the KWs were largely able to complete this task, and the results point to the variety of data that can be collected by conducting a crowdsourced observer performance studies.

In summary, we have shown using distributed human intelligence that qualification tests, improved training and image rendering, the addition of video to static 2D and 3D images and the offer of a reward may lead to improved perception and classification of CTC CAD marks. KWs outperformed a trained CAD classifier for easy and moderate polyp candidates. KWs' performance compared favorably to that of an expert radiologist. Estimated CAD mark difficulty computed from KW scores using an expectation maximization technique correlated well with perceived difficulty assigned by an expert. Changes in individual KWs' performance over the course of the experiment and as a function of time-to-decision mimicked results published in the literature for the performance of expert observers. Our results indicate that CAD observer experiments conducted using distributed human intelligence may inform changes in data presentation and training that lead to improved performance of radiologists using CAD in the clinical setting, ultimately leading to increased performance of CT colonography.

## Acknowledgments

The intramural research programs of the National Institutes of Health, Clinical Center, supported this work. We thank Perry J. Pickhardt, M.D., J. Richard Choi, ScD, MD and William Schindler, DO, for supplying CTC data. The mention of commercial entities, or commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such entities or products by the Department of Health and Human Services.

## Abbreviations

| | |
|---|---|
| **CTC** | Computed tomography colonography |

| | |
|---|---|
| **KW** | knowledge worker |
| **MTurk** | Mechanical Turk |
| **HIT** | human intelligence task |

## REFERENCES

Burgess AE. Comparison of receiver operating characteristic and forced choice observer performance measurement methods. Med. Phys. 1995; 22:643–655. [PubMed: 7643805]

Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popovic Z, Players F. Predicting protein structures with a multiplayer online game. Nature. 2010; 466:756–760. [PubMed: 20686574]

Cotton PB, Durkalski VL, Benoit PC, Palesch YY, Mauldin PD, Hoffman B, Vining DJ, Small WC, Affronti J, Rex D, Kopecky KK, Ackerman S, Burdick JS, Brewington C, Turner MA, Zfass A, Wright AR, Iyer RB, Lynch P, Sivak MV, Butler H. Computed tomographic colonography (virtual colonoscopy) - A multicenter comparison with standard colonoscopy for detection of colorectal neoplasia. JAMA-J. Am. Med. Assoc. 2004; 291:1713–1719.

Dachman AH, Obuchowski NA, Hoffmeister JW, Hinshaw JL, Frew MI, Winter TC, Van Uitert RL, Periaswamy S, Summers RM, Hillman BJ. Effect of Computer-aided Detection for CT Colonography in a Multireader, Multicase Trial. Radiology. 2010; 256:827–835. [PubMed: 20663975]

Dawid AP, Skene AM. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics). 1979; 28:20–28.

Dijkstra EW. A note on two problems in connexion with graphs. Numerische Mathematik. 1959; 1:269–271.

Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, Perrin RL, Poller WR, Shah R, Sumkin JH, Wallace LP, Rockette HE. The "Laboratory" Effect: Comparing Radiologists' Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations1. Radiology. 2008; 249:47–53. [PubMed: 18682584]

Ipeirotis, PG.; Provost, F.; Wang, J. Proceedings of the ACM SIGKDD Workshop on Human Computation. Washington DC: ACM; 2010. Quality management on Amazon Mechanical Turk.

Jemal A, Siegel R, Xu J, Ward E. Cancer Statistics, 2010. CA: A Cancer Journal for Clinicians. 2010; 60:277–300. [PubMed: 20610543]

Johnson CD. Accuracy of CT Colonography for Detection of Large Adenomas and Cancers. N. Engl. J. Med. 2008; 359:1207–1217. [PubMed: 18799557]

Johnson CD, Toledano AY, Herman BA, Dachman AH, McFarland EG, Barish MA, Brink JA, Ernst RD, Fletcher JG, Halvorsen RA Jr, Hara AK, Hopper KD, Koehler RE, Lu DS, Macari M, Maccarty RL, Miller FH, Morrin M, Paulson EK, Yee J, Zalis M. Computerized tomographic colonography: performance evaluation in a retrospective multicenter setting. Gastroenterology. 2003; 125:688–695. [PubMed: 12949715]

Li J, Huang A, Yao J, Liu JM, Van Uitert RL, Petrick N, Summers RM. Optimizing computer-aided colonic polyp detection for CT colonography by evolving the Pareto front. Med. Phys. 2009; 36:201–212. [PubMed: 19235388]

Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Stat Med. 1998; 17:1033–1053. [PubMed: 9612889]

Nappi J, Yoshida H. Virtual tagging for laxative-free CT colonography: Pilot evaluation. Med. Phys. 2009; 36:1830–1838. [PubMed: 19544802]

Nguyen T, Wang S, Anugu V, Rose N, Burns J, McKenna M, Petrick N, Summers RM. Distributed Human Intelligence for Colonic Polyp Classification in Computer-aided Detection for CT Colonography. Radiology. 2012; 262:824–833. [PubMed: 22274839]

Nodine CF, Mello-Thoms C, Kundel HL, Weinstein SP. Time course of perception and decision making during mammographic interpretation. American Journal of Roentgenology. 2002; 179:917–923. [PubMed: 12239037]

Obuchowski NA, Meziane M, Dachman AH, Lieber ML, Mazzone PJ. What's the control in studies measuring the effect of computer-aided detection (CAD) on observer performance? Acad Radiol. 2010; 17:761–767. [PubMed: 20457419]

Park SH, Lee SS, Kim JK, Kim M-J, Kim HJ, Kim SY, Kim M-Y, Kim AY, Ha HK. Volume rendering with color coding of tagged stool during endoluminal fly-through CT colonography: Effect on reading efficiency. Radiology. 2008; 248:1018–1027. [PubMed: 18710990]

Petrick N, Haider M, Summers RM, Yeshwant SC, Brown L, Iuliano EM, Louie A, Choi JR, Pickhardt PJ. CT Colonography and Computer-aided Detection as a Second Reader: Observer Performance Study. Radiology. 2008; 246:148–156. [PubMed: 18096536]

Pickhardt PJ, Choi JH. Electronic cleansing and stool tagging in CT colonography: advantages and pitfalls with primary three-dimensional evaluation. AJR Am J Roentgenol. 2003; 181:799–805. [PubMed: 12933484]

Pickhardt PJ, Choi JR, Hwang I, Butler JA, Puckett ML, Hildebrandt HA, Wong RK, Nugent PA, Mysliwiec PA, Schindler WR. Computed tomographic virtual colonoscopy to screen for colorectal neoplasia in asymptomatic adults. N Engl J Med. 2003; 349:2191–2200. [PubMed: 14657426]

Raykar VC, Yu S, Zhao LH, Valadez GH, Florin C, Bogoni L, Moy L. Learning From Crowds. J. Mach. Learn. Res. 2010; 11:1297–1322.

Rockey DC, Paulson E, Niedzwiecki D, Davis W, Bosworth HB, Sanders L, Yee J, Henderson J, Hatten P, Burdick S, Sanyal A, Rubin DT, Sterling M, Akerkar G, Bhutani MS, Binmoeller K, Garvie J, Bini EJ, McQuaid K, Foster WL, Thompson WM, Dachman A, Halvorsen R. Analysis of air contrast barium enema, computed tomographic colonography, and colonoscopy: prospective comparison. Lancet. 2005; 365:305–311. [PubMed: 15664225]

Saunders RS, Samei E. Improving mammographic decision accuracy by incorporating observer ratings with interpretation time. The British journal of radiology. 2006; 79(Spec No 2):S117–S122. [PubMed: 17209116]

Schroeder, W.; Martin, K.; Lorensen, B. The Visualization Toolkit. 3 ed. Kitware, Inc.; 2003.

Smith RA, Cokkinides V, Brooks D, Saslow D, Brawley OW. Cancer Screening in the United States, 2010: A Review of Current American Cancer Society Guidelines and Issues in Cancer Screening. CA: A Cancer Journal for Clinicians. 2010; 60:99–119. [PubMed: 20228384]

Snow, R.; O'Connor, B.; Jurafsky, D.; Ng, AY. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii: Association for Computational Linguistics; 2008. Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks.

Summers RM, Yao J, Pickhardt PJ, Franaszek M, Bitter I, Brickman D, Krishna V, Choi JR. Computed tomographic virtual colonoscopy computer-aided polyp detection in a screening population. Gastroenterology. 2005; 129:1832–1844. [PubMed: 16344052]

Suzuki K, Zhang J, Xu JW. Massive-Training Artificial Neural Network Coupled With Laplacian-Eigenfunction-Based Dimensionality Reduction for Computer-Aided Detection of Polyps in CT Colonography. IEEE Trans. Med. Imaging. 2010; 29:1907–1917. [PubMed: 20570766]

Taylor SA, Robinson C, Boone D, Honeyfield L, Halligan S. Polyp Characteristics Correctly Annotated by Computer-aided Detection Software but Ignored by Reporting Radiologists during CT Colonography. Radiology. 2009; 253:715–723. [PubMed: 19789221]

von Ahn, L.; Dabbish, L. Proceedings of the SIGCHI conference on Human factors in Computing Systems. Vienna, Austria: ACM; 2004. Labeling Images with a Computer Game.

von Ahn L, Maurer B, McMillen C, Abraham D, Blum M. reCAPTCHA: Human-based character recognition via web security measures. Science. 2008; 321:1465. [PubMed: 18703711]

Wang, S.; Anugu, V.; Nguyen, T.; Rose, N.; Burns, J.; McKenna, M.; Petrick, N.; Summers, RM. Fusion of machine intelligence and human intelligence for colonic polyp detection in CT colonography; 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2011. p. 160-164.

Wang S, Yao J, Petrick N, Summers RM. Combining Statistical and Geometric Features for Colonic Polyp Detection in CTC Based on Multiple Kernel Learning. International Journal of Computational Intelligence and Applications. 2010; 9:1–15. [PubMed: 20953299]

Whitehill J, Ruvolo P, Wu T-f, Bergsma J, Movellan J. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. Advances in Neural Information Processing Systems. 2009; 22

Yao J, Miller M, Franaszek M, Summers RM. Colonic polyp segmentation in CT colonography based on fuzzy clustering and deformable models. IEEE Trans Med Imaging. 2004; 23:1344–1352. [PubMed: 15554123]

Zhu HB, Liang ZR, Pickhardt PJ, Barish MA, You JS, Fan Y, Lu HB, Posniak EJ, Richards RJ, Cohen HL. Increasing computer-aided detection specificity by projection features for CT colonography. Med. Phys. 2010; 37:1468–1481. [PubMed: 20443468]

## Highlights

- Changes to CTC CAD interpretation are assessed using distributed human intelligence
- Knowledge workers (KW) outperformed SVMs for certain categories of CAD
- KWs’ performance compared favorably to that of an expert radiologist
- Difficulty estimates from KW scores closely matched expert-assigned difficulty
- Changes in KW performance mimicked published results on expert observer performance

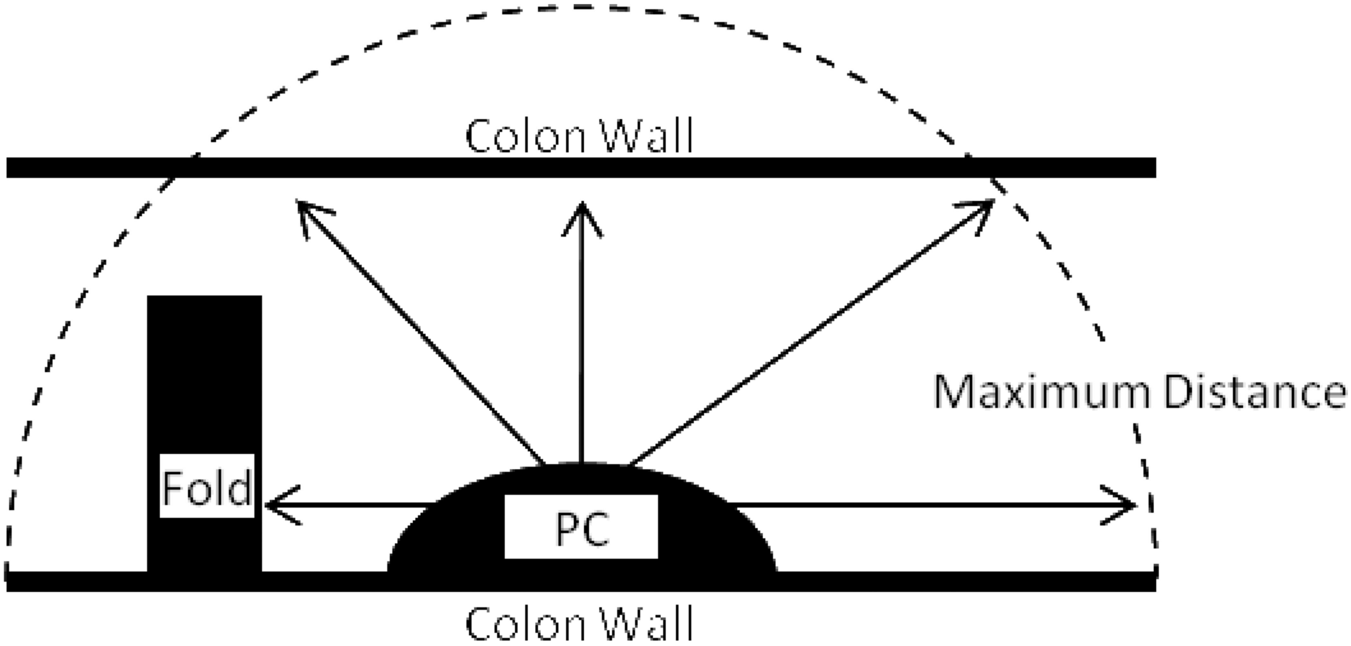
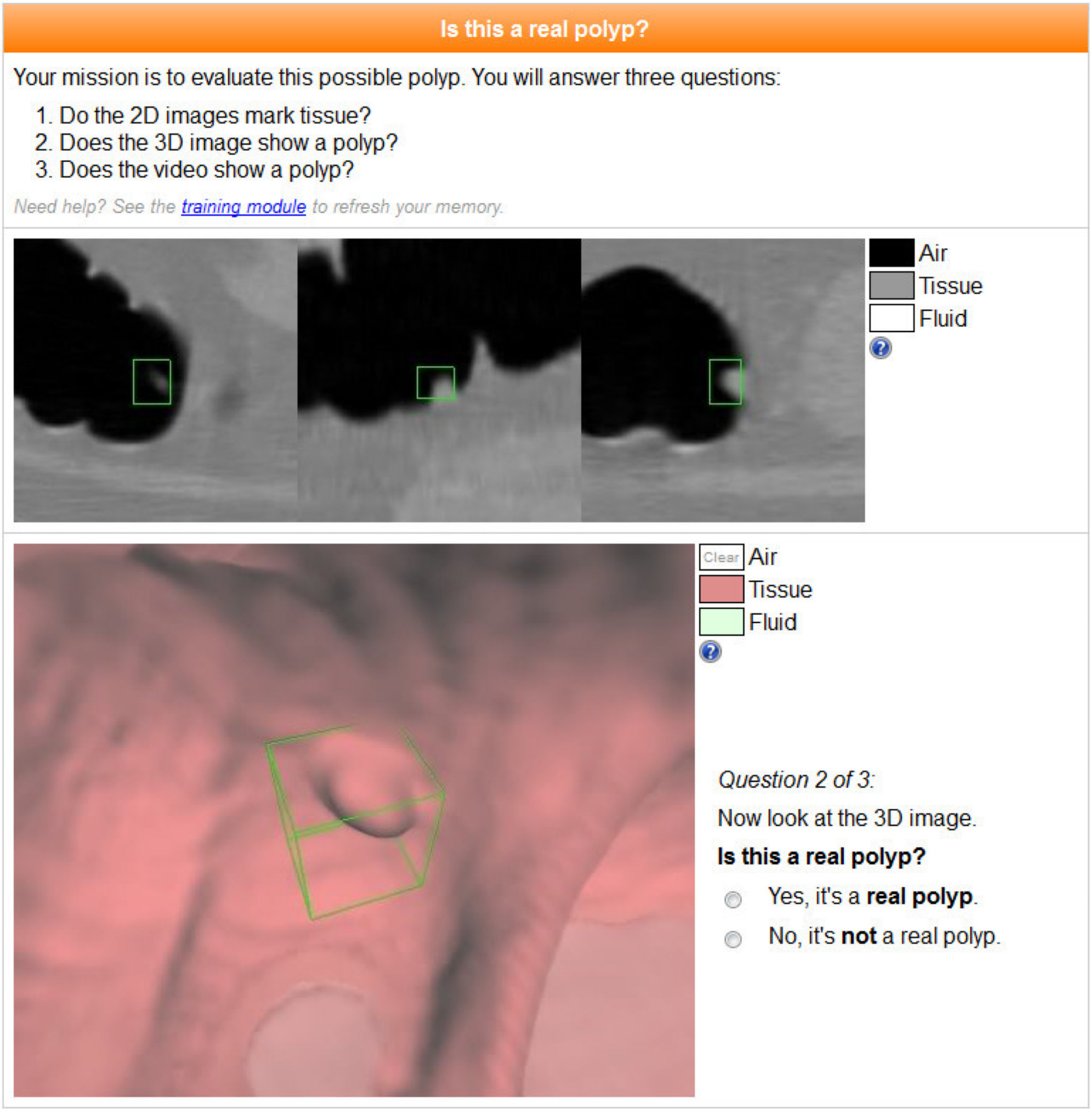


**Figure 1.**

Illustration of viewpoint generation process. Cameras (shown by arrows) are moved from the polyp candidate (PC) centroid in various directions until they hit tissue (e.g. colon wall or fold) or reach a maximum distance from the PC centroid.



**Is this a real polyp?**

Your mission is to evaluate this possible polyp. You will answer three questions:

1. Do the 2D images mark tissue?
2. Does the 3D image show a polyp?
3. Does the video show a polyp?

*Need help? See the training module to refresh your memory.*

■ Air
▨ Tissue
□ Fluid

Clear Air
🟥 Tissue
🟩 Fluid

*Question 2 of 3:*

Now look at the 3D image.

**Is this a real polyp?**

○ Yes, it's a **real polyp**.

○ No, it's **not** a real polyp.

**Figure 2.**

Human Intelligence Task Template. This is an example of the form used to collect answers from knowledge workers for each human intelligence task (HIT). This figure shows only the second of the three questions. The 3D rendering is hidden until the KW answers the first question. After answering the second question, the 3D rendering is replaced with its corresponding video fly-around. This HIT shows a 9 mm sessile polyp.

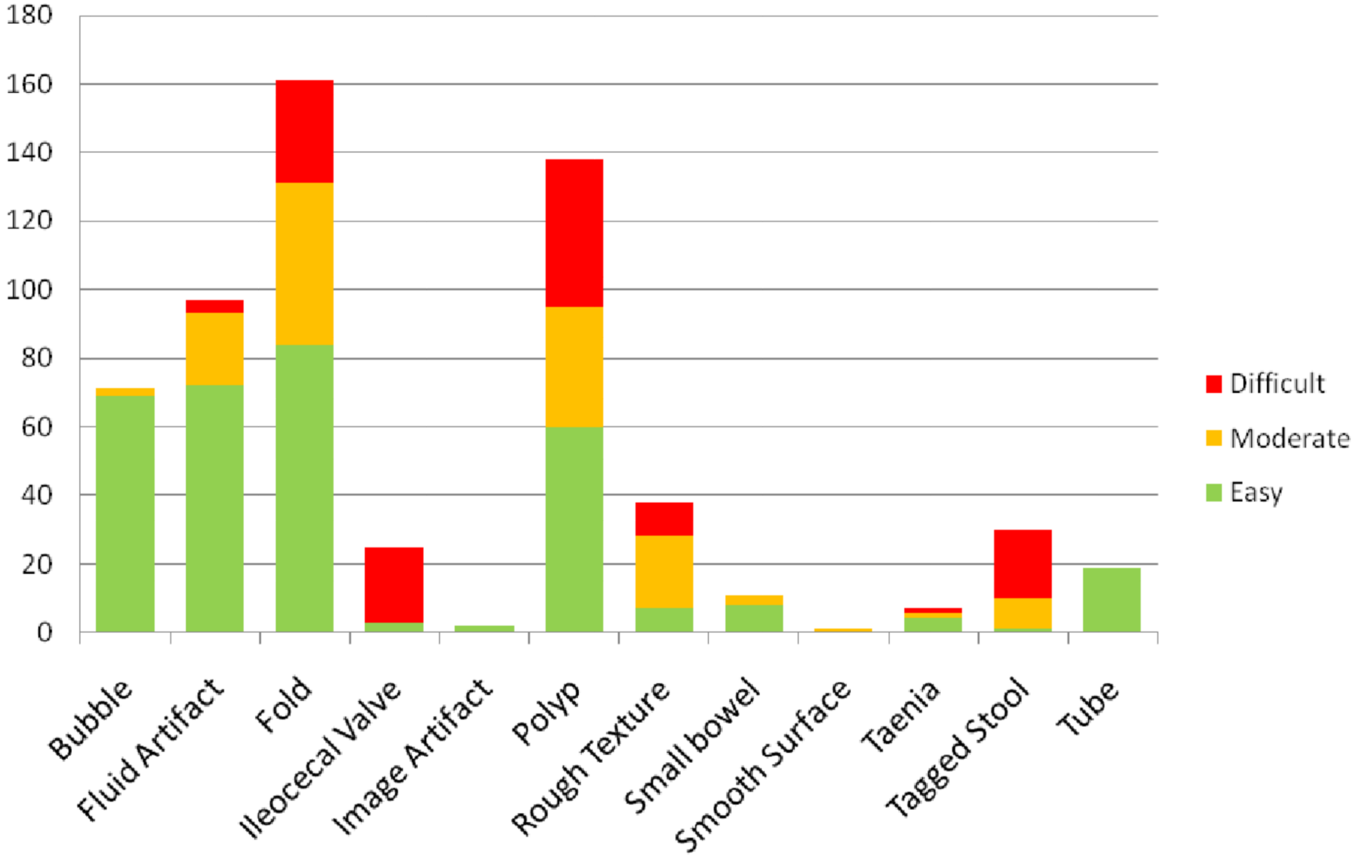


**Figure 3.**

Distribution of detection categories and difficulties for all CAD marks. Only the category labeled “polyp” represented true positive CAD marks; the remaining categories were false positive CAD marks.

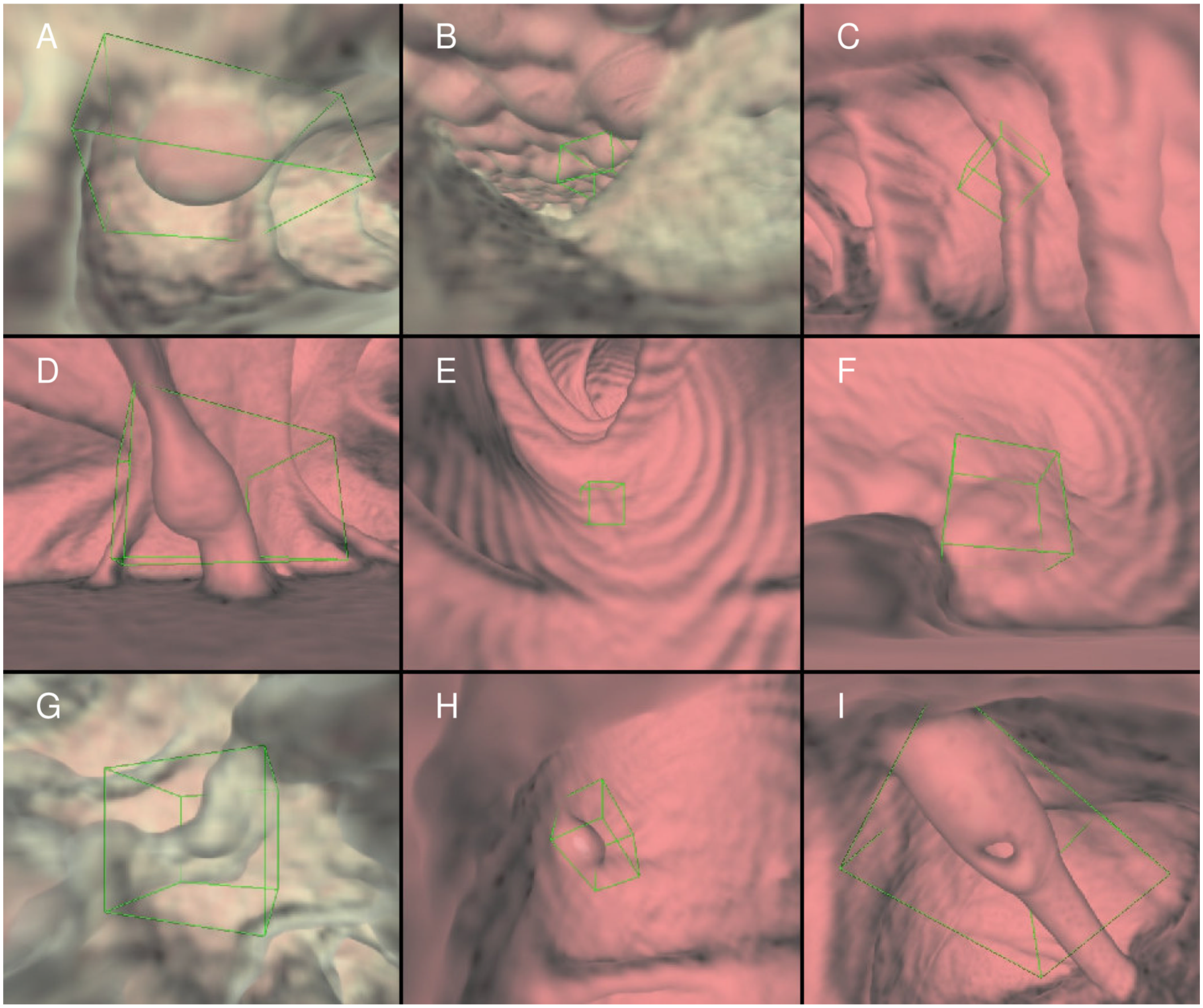


**Figure 4.**
Examples of common false positive CAD marks (difficulty of each mark listed in parentheses). (A) air bubble (easy), (B) fluid artifact (easy), (C) fold (easy), (D) ileocecal valve (difficult), (E) segment of the taenia coli (moderate), (F) rough surface (moderate), (F) small bowel (easy), (F) tagged stool (difficult), and (I) rectal tube used to insufflate the colon (easy).

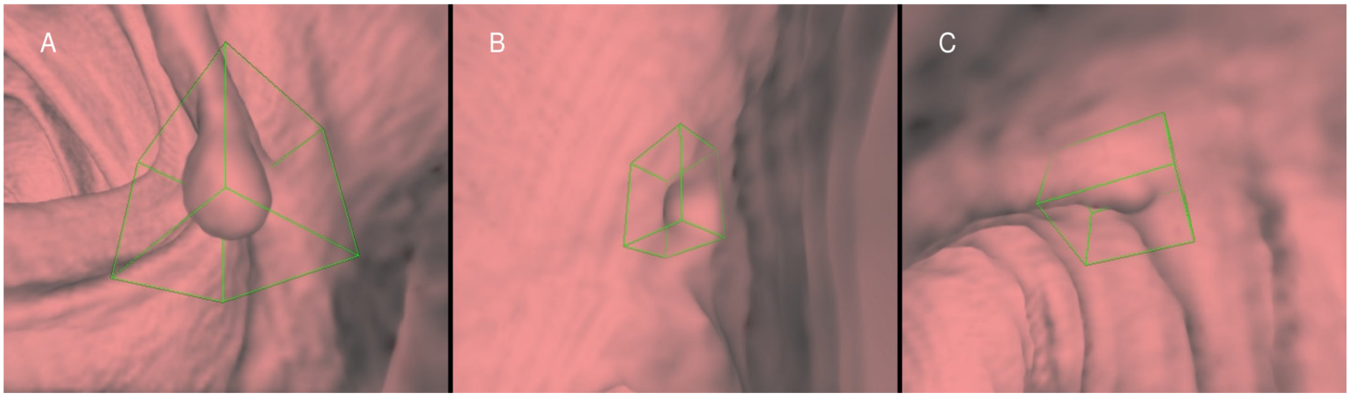


**Figure 5.**
Difficulty classification of CAD marks on true polyps. (A) 1.2 cm pedunculated polyp deemed easy to classify, (B) 7 mm sessile polyp deemed moderate, and (C) 8 mm sessile polyp deemed difficult.

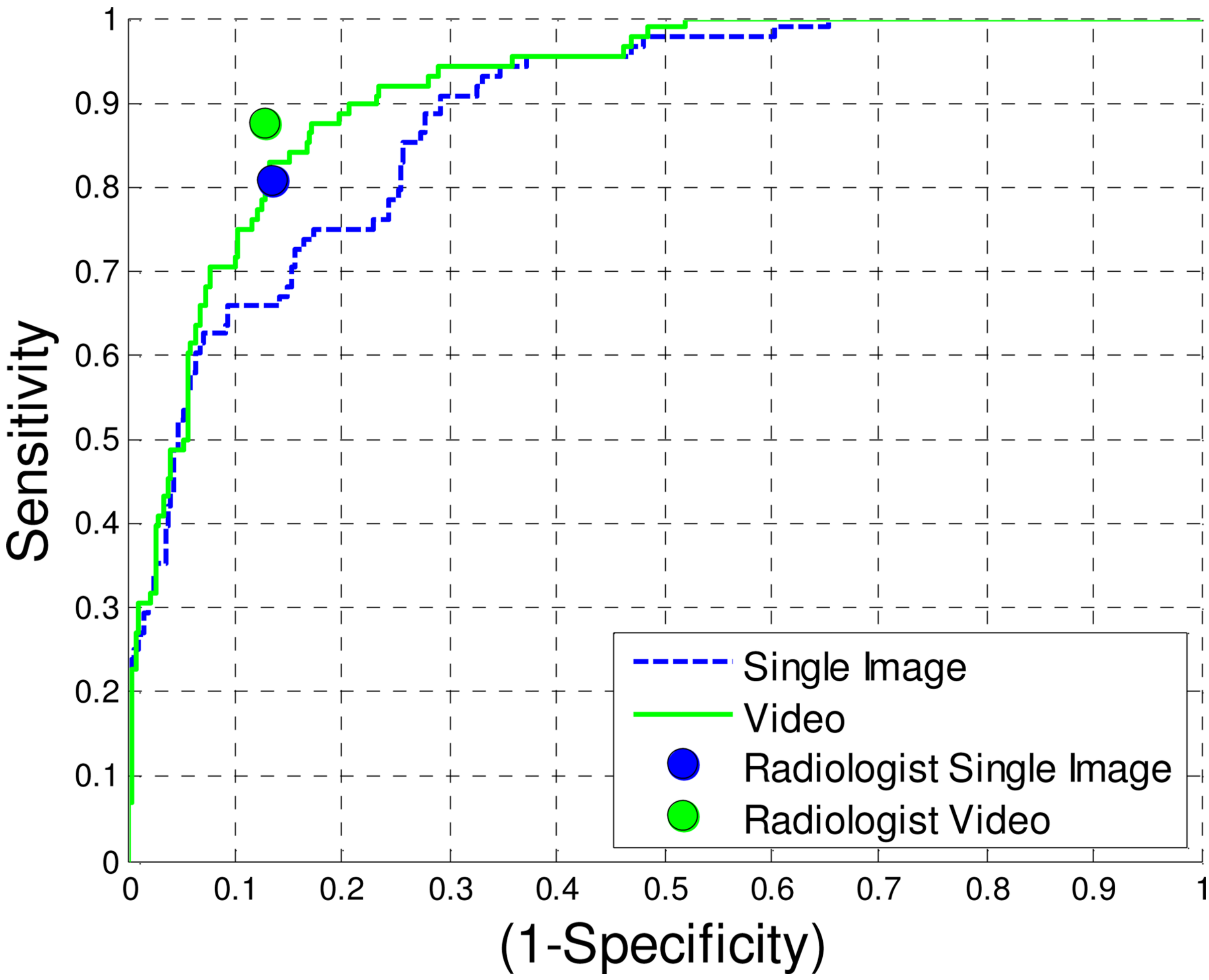


**Figure 6.**

Receiver operating characteristic curves for KWs and sensitivity and specificity of an expert radiologist in interpreting CAD marks presented using a single, static image and video. Both the KWs and expert radiologist improved with the video.

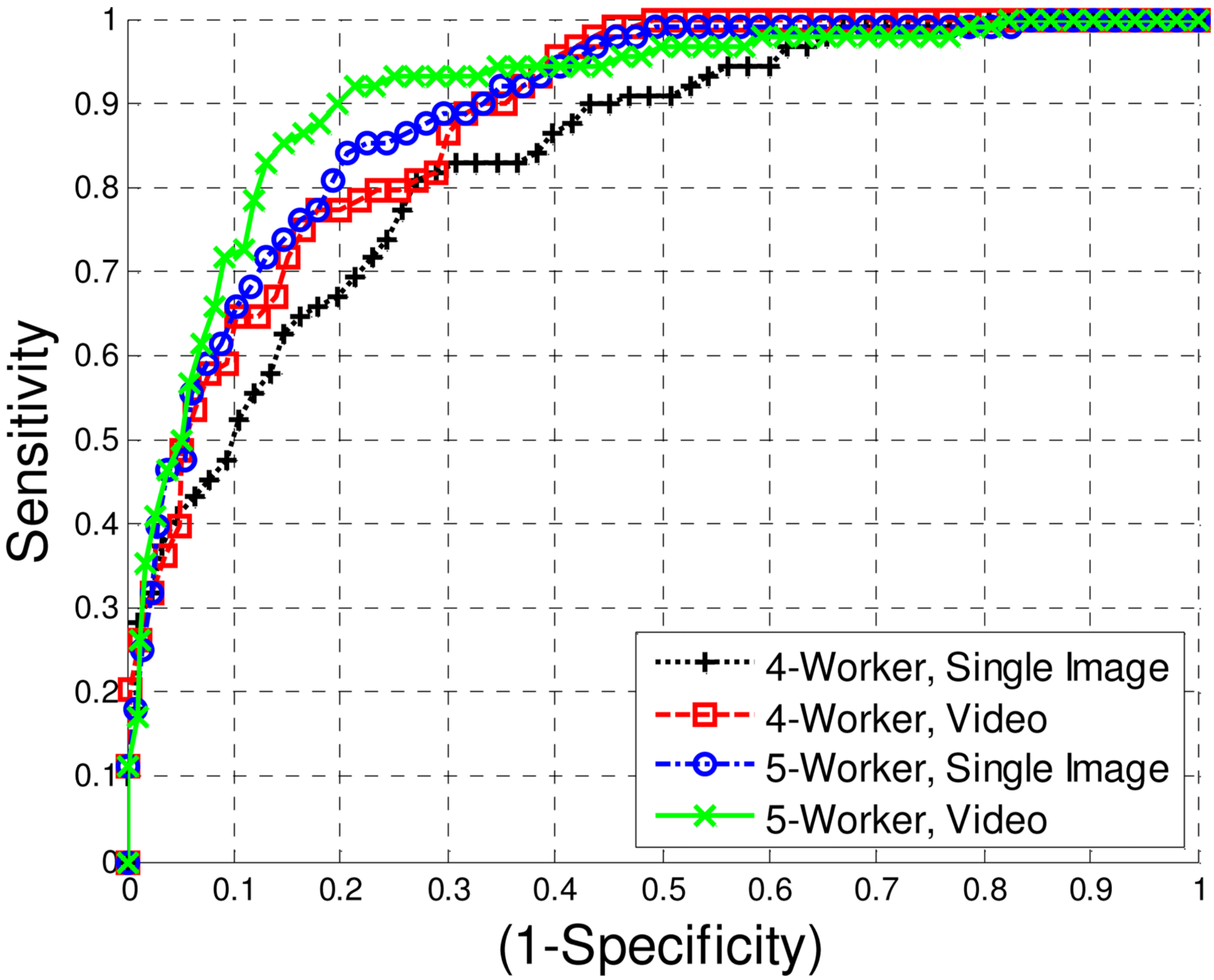


**Figure 7.**

Receiver operating characteristic curves for four-workers and five-workers. Five-workers are those knowledge workers who answered all five questions correctly on a five-question qualification test. Four-workers answered four of the five questions correctly. Five-workers outperformed four-workers on the human intelligence tasks using both single images and videos. AUC values are given in Table 2.

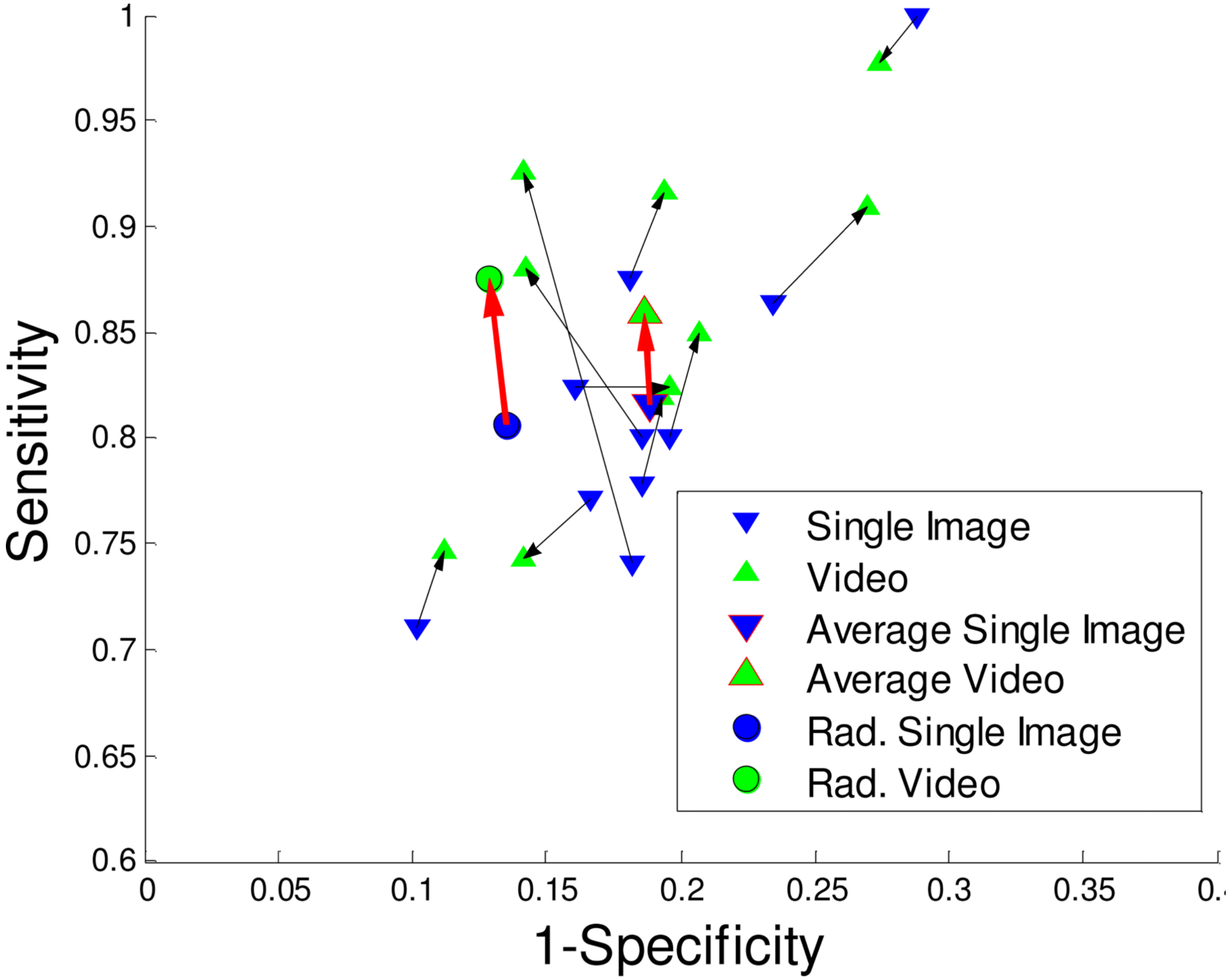


**Figure 8.**

Receiver operating characteristic operating points of the 10 best KWs and expert radiologist (Rad.). The 10 best workers, on average, saw an increase in sensitivity with a small increase in specificity. This improvement was similar to the improvement seen in an expert radiologist. The “Average” points in this figure represent the average from the 10 best KWs.

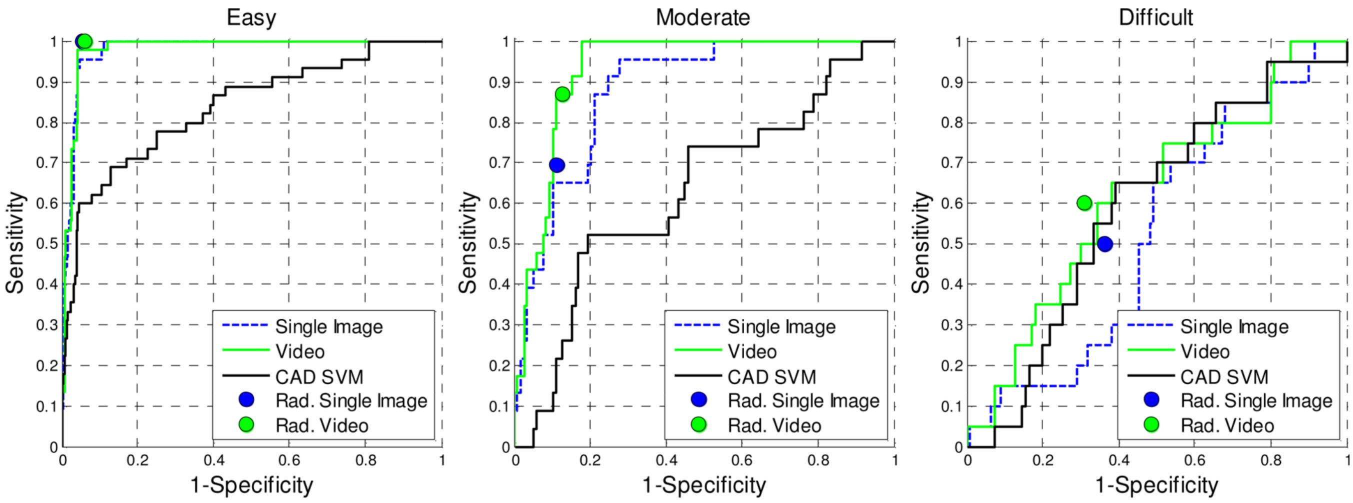


**Figure 9.**

KW performance stratified by level of difficulty of the CAD mark. KW and radiologist performance decreased with increasing difficulty. Note that the radiologist performance for easy detections is nearly identical for both single-image and video interpretation.

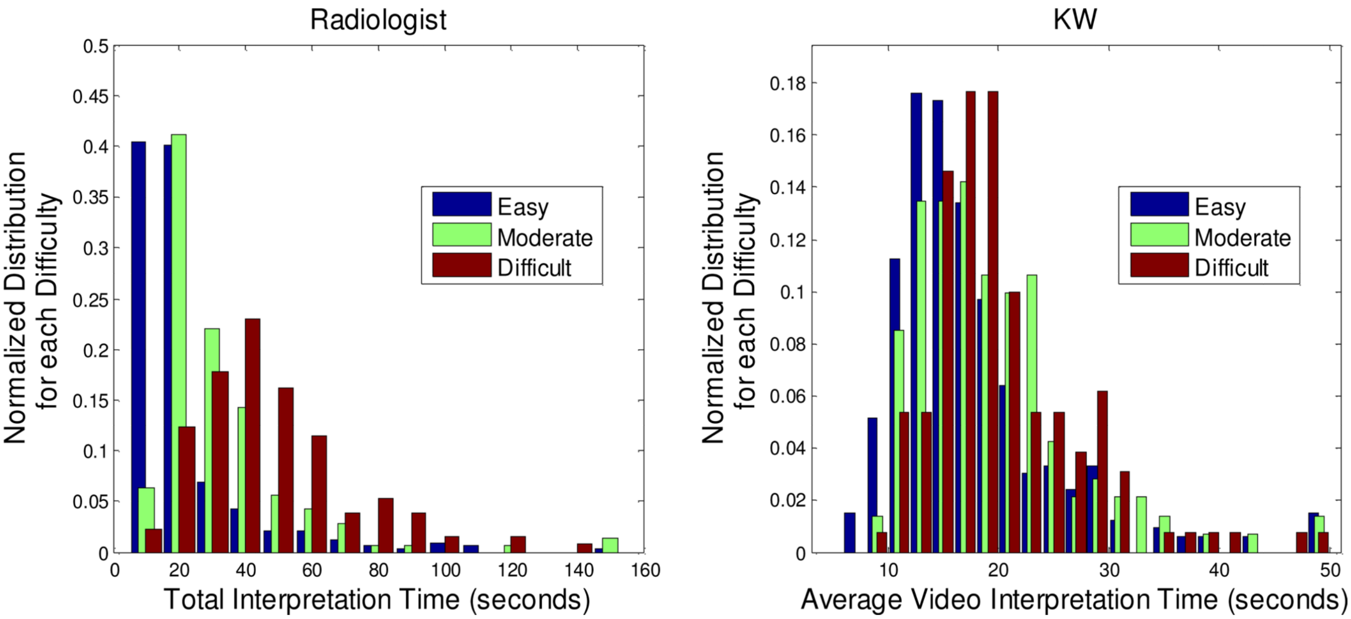


**Figure 10.**

Distribution of interpretation times by difficulty for an expert radiologist and KWs. Difficult detections took significantly longer to interpret than easy detections.

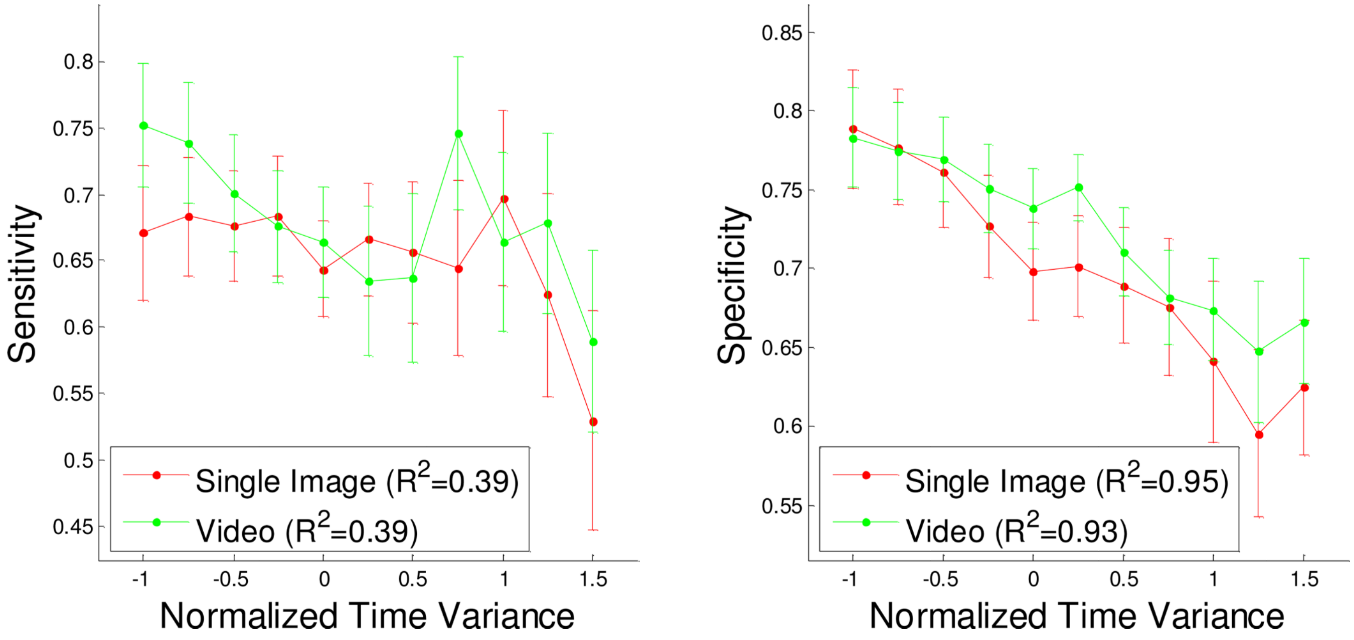


**Figure 11.**

Knowledge worker sensitivities and specificities as a function of normalized time variance. Normalized time variance for a particular worker is the interpretation time normalized by the variance of interpretation time for a particular worker. The error bars show the standard error of the estimate at each point. KW specificity decreases linearly as interpretation time increases.

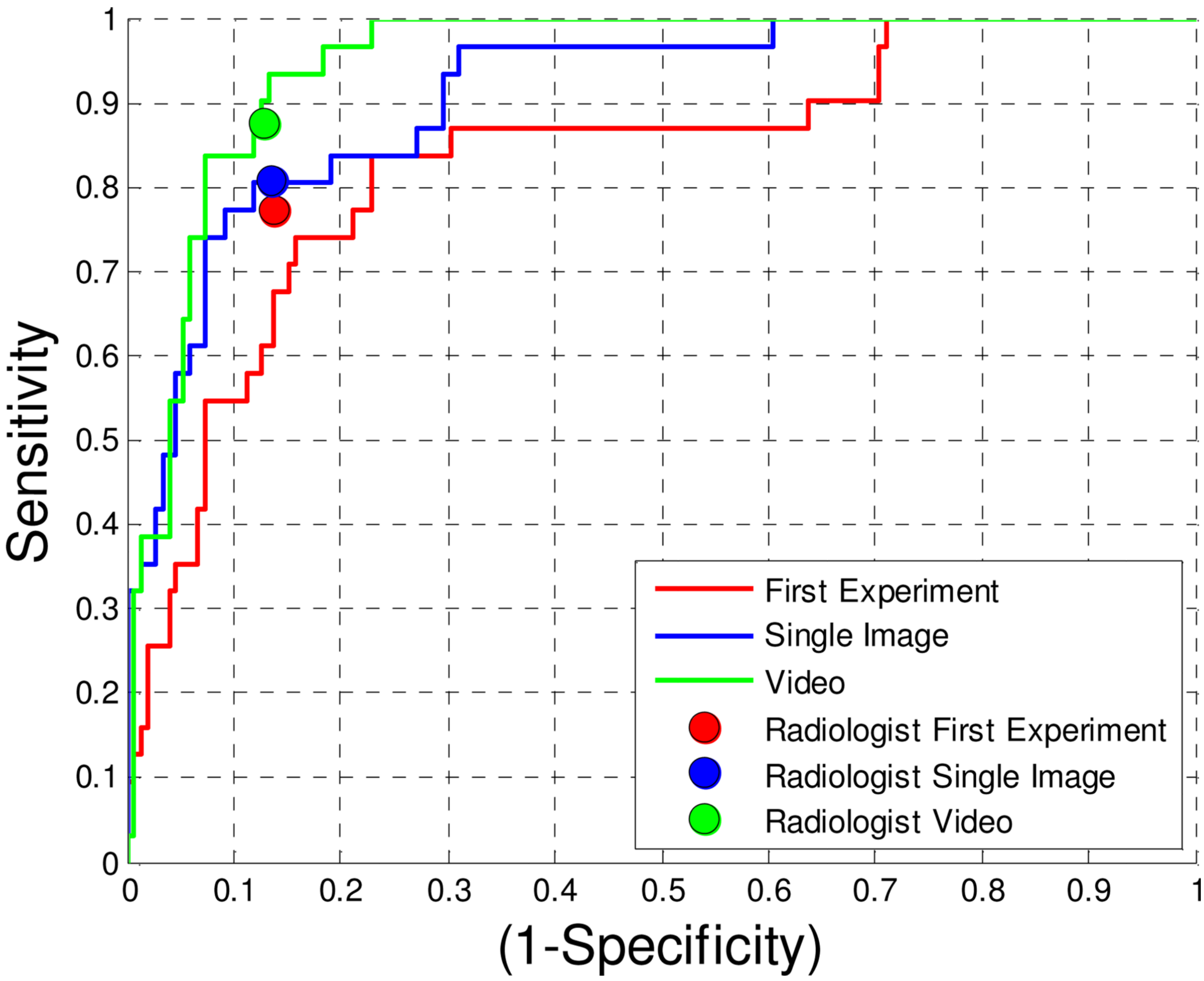


**Figure 12.**
Evaluation of training and image generation. We compare results from our current study with results from our original study (Nguyen et al., 2012) for the 183 CAD marks that shared a volume overlap between the studies. The changes in training, image generation, and qualification requirements resulted in an increase in AUC for KWs and increases in sensitivity and specificity for an expert.

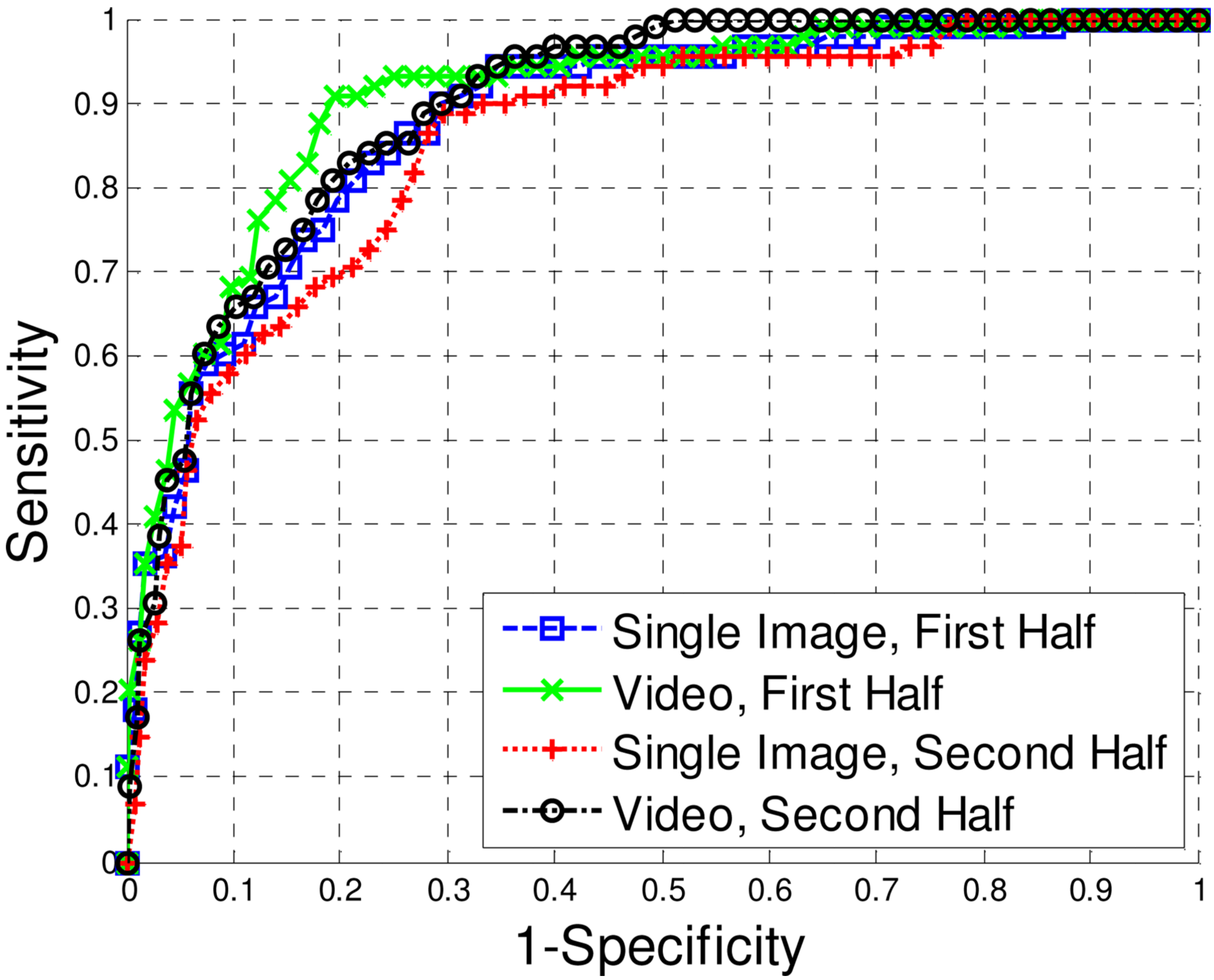


**Figure 13.**

ROC curves comparing the first half to the second half.

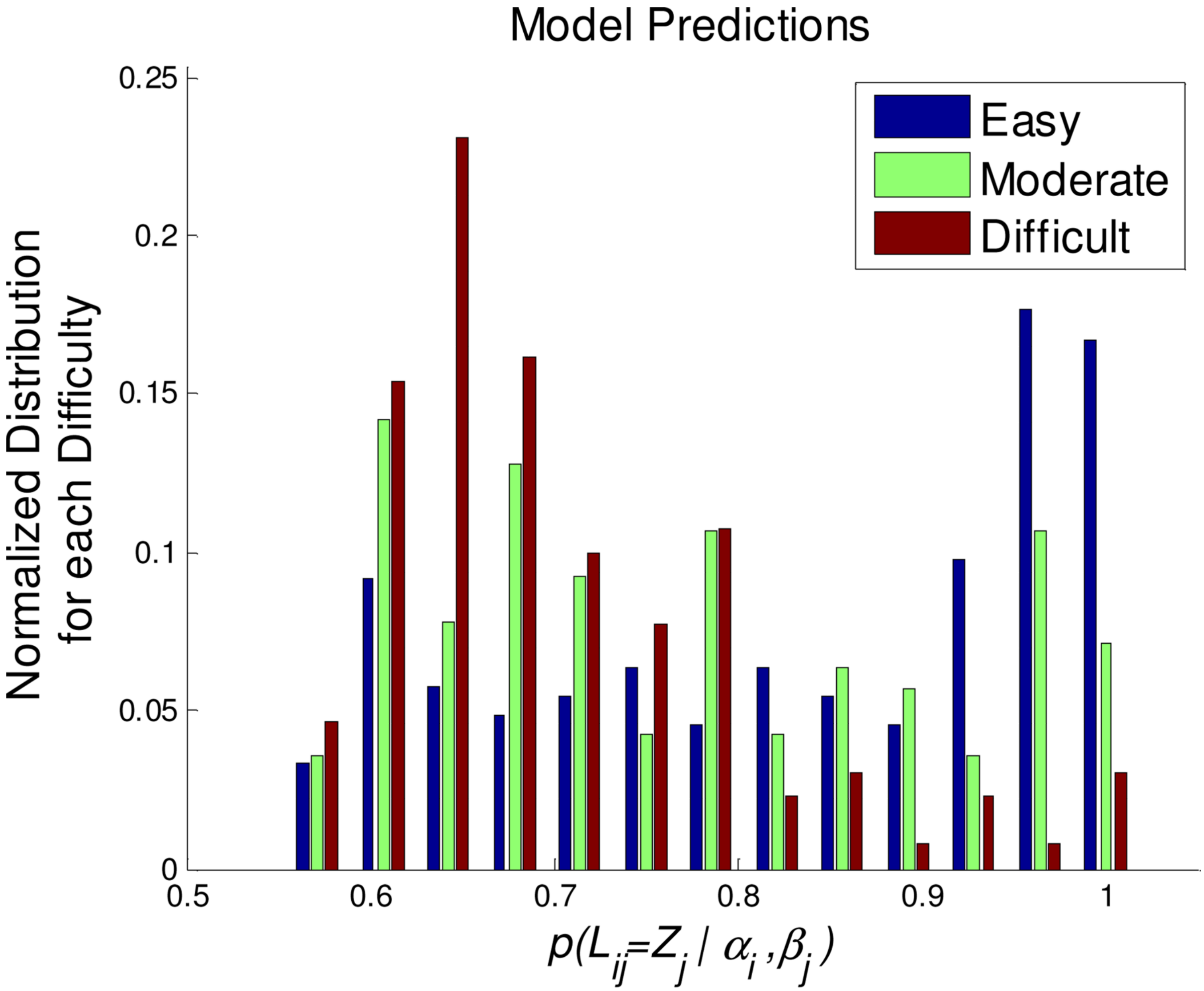


**Figure 14.**

Predictions ($p$) of model (1) stratified by expert-assigned difficulty. $p$ is the probability that worker $i$ will correctly assign a label $L_{ij}$ to image $j$ with true label of $Z_j$, given worker accuracy, $\alpha_i$, and image difficulty, $\beta_j$. For this figure, we fixed $\alpha_i = 1$ and used the $\beta_j$'s estimated from KW responses. The model was able to accurately estimate difficulty information. Detections labeled as "difficult" are estimated to have a lower probability of having a correct KW-assigned label. Easier detections are more likely to be marked correctly by a KW.

## Table 1

Patient and polyp characteristics. We only include those polyps found by our initial detection scheme.

| Characteristic | Experimental Set | Training Set |
|---|---|---|
| *Patient demographics* | | |
| Men/women | 33/17 (66/34%) | 29/21 (58/42%) |
| Mean age ± SD (range) | 60 ± 5.6 (51–73) | 59 ± 8.2 (47–77) |
| *Polyp size* | | |
| 6–9 mm | 43 (73%) | 38 (72%) |
| ≥10 mm | 16 (27%) | 15 (28%) |
| *Polyp histopathology* | | |
| Hyperplastic | 17 (29%) | 9 (17%) |
| Tubular adenomatous | 27 (54%) | 34 (64%) |
| Tubulovillous adenomatous | 7 (12%) | 7 (13%) |
| Other benign | 8 (14%) | 3 (6%) |
| *Polyp shape* | | |
| Sessile | 40 (68%) | 39 (74%) |
| Pedunculated | 12 (20%) | 10 (19%) |
| Flat | 6 (10%) | 3 (6%) |
| Other | 1 (2%) | 1 (2%) |
| *Polyp location* | | |
| Rectum | 10 (17%) | 8 (15%) |
| Sigmoid Colon | 15 (25%) | 14 (26%) |
| Descending Colon | 7 (12%) | 6 (11%) |
| Splenic Flexure | 2 (3%) | 1 (2%) |
| Transverse Colon | 6 (10%) | 7 (13%) |
| Hepatic Flexure | 2 (3%) | 1 (2%) |
| Ascending Colon | 12 (20%) | 12 (23%) |
| Cecum | 5 (8%) | 4 (8%) |

## Table 2

KW AUC and $d_a$ statistics. Table shows overall KW performance and performance stratified by both detection difficulty and worker qualification score.

| | | | AUC (± standard error) | | | $d_a$[†] (± standard deviation) | |
|---|---|---|---|---|---|---|---|
| | $n_{True}$ | $n_{False}$ | **Single Image** | **Video** | ***p*-value** | **Single Image** | **Video** |
| **All Detections** | 88 | 512 | .882 (±0.024) | .913 (±0.021) | 0.001 | 1.68 (±0.172) | 1.92 (±0.188) |
| **By Difficulty** | | | | | | | |
| **Easy** | 45 | 284 | .977 (±0.016) | .978 (±0.016) | 0.492 | 2.82 (±0.417) | 2.84 (±0.425) |
| **Moderate** | 23 | 118 | .877 (±0.048) | .927 (±0.038) | 0.094 | 1.64 (±0.334) | 2.06 (±0.388) |
| **Difficult** | 20 | 110 | .506 (±0.071) | .608 (±0.072) | 0.034 | 0.021 (±0.252) | 0.388 (±0.265) |
| **By Worker Qualification** | | | | | | | |
| **4-Workers** | 88 | 512 | .836 (±0.027) [*] | .881 (±0.024) [**] | 0.001 | 1.38 (±0.154) | 1.67 (±0.171) |
| **5-Workers** | 88 | 512 | .889 (±0.023) [*] | .903 (±0.022) [**] | 0.023 | 1.73 (±0.172) | 1.84 (±0.182) |

[†] $d_a$ is the detectability index, which represents the perceptual signal-to-noise ratio.

[*] *p*-value for comparison between 5-workers and 4-workers using a single image = 0.0031

[**] *p*-value for comparison between 5-workers and 4-workers using video = 0.0651

## Table 3

AUCs and $d_a$ of KWs stratified by difficulty and qualification score.

| Difficulty | Worker Group | AUC (± standard error) |  | *p*-value | $d_a$ [†] (± standard deviation) |  |
|---|---|---|---|---|---|---|
|  |  | Single Image | Video |  | Single Image | Video |
| **Easy** | 4-Worker | 0.954 (±0.020) | 0.969 (±0.018) | 0.059 | 2.38 (±0.294) | 2.65 (±0.368) |
|  | 5-Worker | 0.976 (±0.016) | 0.975 (±0.016) | 0.950 | 2.80 (±0.401) | 2.78 (±0.392) |
| **Moderate** | 4-Worker | 0.813 (±0.056) | 0.867 (±0.050) | 0.133 | 1.25 (±0.294) | 1.57 (±0.329) |
|  | 5-Worker | 0.883 (±0.047) | 0.929 (±0.038) | 0.042 | 1.68 (±0.338) | 2.08 (±0.397) |
| **Difficult** | 4-Worker | 0.488 (±0.070) [*] | 0.616 (±0.072) | 0.007 | −0.0443 (±0.248) [*] | 0.417 (±0.267) |
|  | 5-Worker | 0.501 (±0.070) | 0.582 (±0.072) | 0.039 | 0.00495 (±0.248) | 0.293 (±0.261) |

[†] $d_a$ is the detectability index, which represents the perceptual signal-to-noise ratio.

[*] AUC values below 0.5 and negative $d_a$ values are not realistic because they imply performance worse than guessing. The reported estimates likely signify KW performance no worse than random guessing (i.e., AUC=0.5 and $d_a$=0).

## Table 4

Radiologist performance statistics.

| | Sensitivity | | Specificity | |
|---|---|---|---|---|
| | **Single Image** | **Video** | **Single Image** | **Video** |
| **All Detections** | 0.8068 | 0.8750 | 0.8652 | 0.8711 |
| **Detections By Difficulty** | | | | |
| **Easy** | 1.0000 | 0.9437 | 1.0000 | 0.9401 |
| **Moderate** | 0.6957 | 0.8898 | 0.8696 | 0.8729 |
| **Difficult** | 0.5000 | 0.6364 | 0.6000 | 0.6909 |

## Table 5

CAD SVM performance statistics.

| | AUC (± standard error) | $d_a$ † (± standard deviation) |
|---|---|---|
| **Detections By Difficulty** | | |
| **Easy** | 0.835 (± .038) * | 1.38 (±0.217) |
| **Moderate** | 0.626 (± .067) | 0.454 (±0.250) |
| **Difficult** | 0.584 (± .072) | 0.300 (±0.261) |

† $d_a$ is the detectability index, which represents the perceptual signal-to-noise ratio.

* Significantly greater than moderate and difficult ($p$<0.05)

## Table 6

First half vs. second half statistics. The differences in time are not significant.

| | First Half | | Second Half | |
|---|---|---|---|---|
| | **Single Image** | **Video** | **Single Image** | **Video** |
| **AUC** | 0.876 (±0.025) | 0.898 (±0.023) | 0.848 (±0.027) | 0.894 (±0.023) |
| $d_a$ | 1.64 (±0.17) | 1.79 (±0.18) | 1.45 (±0.16) | 1.77 (±0.18) |
| **Time (s)** | 4.20 (±2.38) | 21.26 (±25.76) | 2.63 (±1.05) | 17.91 (±17.33) |