Performance changes in lung nodule detection following perceptual feedback of eye movements.

T. Donovan[a] *, D.J. Manning [a], T. Crawford [b]
[a]School of Medical Imaging Sciences, University of Cumbria, Lancaster, LA1 3JD, UK.
[b]Psychology Department, Lancaster University, LA1 4YF, UK

ABSTRACT

In a previously reported study we demonstrated that expert performance can decline following perceptual feedback of eye movements in the relatively simple radiological task of wrist fracture detection [1]. This study was carried out to determine if the same effect could be observed using a more complicated radiological task of identifying lung nodules on chest radiographs. Four groups (n=10 in each group) of observers with different levels of expertise were tested. The groups were naïve observers, level 1 radiography students, level 2 radiography students and experts. Feedback was presented to the observers in the form of their scan paths and fixations. Half the observers had feedback and half had no perceptual feedback. JAFROC analysis was used to measure observer performance. A repeated measures ANOVA was carried out. There was no significant effect between the pre and post "no feedback" condition. There was a significant difference between the pre and post "feedback" condition with a significant improvement following feedback (F(1,16)=6.6,p = 0.021). Overall the mean percentage improvement was small of 3.3%, with most of the improvement due to the level 1 group where the percentage increase in the figure of merit (FOM) was 8.4% and this was significant ($p<0.05$).

Eye tracking metrics indicate that the expert and naïve observers were less affected by feedback or a second look whereas there were mixed results between the level 1 and level 2 students possibly reflecting the different search strategies used. Perceptual feedback may be beneficial for those early in their training.

Keywords: perceptual feedback, expertise, eye-tracking, lung nodule detection

1. INTRODUCTION

It is recognised that visual dwell is a predictor of target recognition, and that most missed lesions are fixated. As a result of this feedback of eye fixations is often proposed as a mechanism for improving performance, and improvements have been demonstrated [2]. However in a previously reported study we demonstrated a decline in performance by experts and level 2 students in a relatively simple fracture detection task following perceptual feedback, with a small improvement for naïve observers and level 1 students [1]. So it is apparent that level of expertise and difficulty of radiological task are important considerations when considering the efficacy of perceptual feedback. This study was carried out to determine if the same effects could be observed in the more complicated radiological task of lung nodule perception.

The chest x-ray contains a great deal more information, and a wide range of contrasts, requiring good anatomical knowledge and a strong prototypical idea of normal appearances. Lung nodules are the most common of the localised radiographic abnormalities to be missed by radiologists [2], 30% of nodules are missed [3], of the missed nodules 60-70% are fixated [4]. Other studies testing subjects with different levels of expertise from novice to radiologist have found that it is the decision errors that that make the largest contribution to unreported pulmonary nodules irrespective of level of experience, and that inefficient, faulty or inadequate search make only a minor contribution to error rate [5]. The lung nodule task is essentially a visual search task and it is interesting that many of the reported error rates of 20 to 30% are similar to those in other search tasks such as industrial inspection tasks [6], airport security screening [7] or even some laboratory based visual search tasks [8]. As such the task is about the accuracy of decisions and the perceptual component of the radiology task rather than diagnostic outcomes. So although it is anticipated the naïve and novice observers will have a lower performance due to their lack of exposure to medical images, and possibly a different search strategy as they have no preconceptual guidance, they may still benefit from feedback.

*tim.donovan@cumbria.ac.uk; phone: 0044 1524384667

The experimental design consisted of two studies, one where the observers are given feedback and the other where observers don't receive feedback to help determine whether any performance changes are due to the perceptual feedback or simply due to the opportunity of having a second-look at the image.

2. EXPERIMENTAL DESIGN

The study design is a mixed factor 2 x 4 design with 'expertise' as the between group factor (4 levels of expertise) and feedback as a repeated measure factor (2 levels pre and post).Two studies were carried out, one with pre and post "feedback" and one with pre and post "no feedback". A repeated measures ANOVA was used to analyse the differences between the means of the groups.

2.1 Observer performance measurement

Observer performance was measured using JAFROC analysis software version 2.1.
JAFROC generates a figure of merit that allows quantification of search performance; it is defined as the probability that an observer will rate a lesion higher than the highest rated non lesion on a normal image.

2.2 Eye-tracking

Eye tracking metrics calculated were as follows:
a. Time to first fixation or time to first hit. This is the time from the beginning of the recording in milliseconds until the respective regions of interest (ROIs) were first fixated upon.
b. The average fixation duration in milliseconds is the average length of all fixations during all recordings on the respective ROIs.
c. The fixation count is the number of fixations in the respective ROIs. A fixation for the purposes of this experiment was defined as 100ms and 50 pixels (gaze deviation threshold).
d. The gaze time or dwell time is the total time of all fixations in the respective ROIs in milliseconds.
Data was analysed for each nodule, so although there were 5 observers in each group in each study, each group had 28 nodules.

3. METHODS

Ethics approval for the study was obtained.
Eye-tracking data was acquired with a Tobii x50 system (Tobii Eye Tracker and ClearView analysis software, Tobii Technology AB, 2005). This is a 50Hz scanner, i.e. it samples every 20ms.

3.1 Observer groups

All observers had normal or corrected to normal vision. Only one observer was unable to participate as it was not possible to obtain an acceptable calibration possibly due to stigmatism. All observers received £5 for participating. The groups were:

1. 10 experienced image reporters, 8 of these were radiologists and 2 reporting radiographers that routinely report chest radiographs. Age range 35 – 55 years (mean 47.2).
2. 10 level 2 undergraduate radiography students. These students would have had at least 28 weeks of clinical experience. Age range 20 – 54 years (mean 29.4).
3. 10 level 1 undergraduate radiography students. These students would have had at least 12 weeks of clinical experience. Age range 19- 48 years (mean 31.6).
4. 10 naïve observers, who were a mixture of students and staff from other disciplines within the university with no experience of medical images. Age range 22 – 40 years (mean 29.2).

Each participant in each expertise group was randomly allocated to either receive perceptual feedback or not receive perceptual feedback.

3.2 Test bank

Images were selected from a previously compiled test bank consisting of natural nodules which were histopathologically proven nodules and simulated nodules. A nodule for the purposes of this study is defined as a discrete opacity in the lung field or mediastinum measuring between 5- 30 mm in diameter. Nodules less than 5mm were not included as the perception of these occurs with less than random frequency [9].

30 chest films were selected, 15 films had lung nodules, of the 15, 9 images with one nodule, 3 images with 2 nodules, 1 image with three nodules, 2 images with 5 nodules. In total there were 28 nodules in the test bank, 4 of these were simulated.

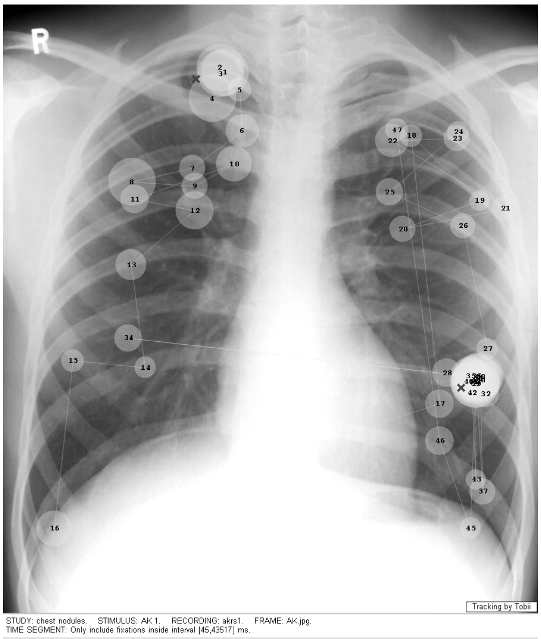
3.3 Procedure

The study was undertaken in a dedicated eye tracking laboratory in which ambient lighting could be controlled and optimum viewing conditions set up.

Subjects were told that the images may be normal or have one nodule or multiple nodules up to a maximum of five on any one image. Their task was to decide on the presence or absence of pulmonary nodules and disregard any other radiological findings. All subjects were shown two chest images, a normal one and one with multiple lung nodules prior to the start of the study. It was explicitly stated that feedback of areas that receive prolonged visual attention can be beneficial for performance, those just having a second look were told research indicates a second look can improve performance.

A 5 point calibration procedure was carried out and stored. The first image was presented, free search was allowed, subjects were instructed to click on each nodule and verbally indicate their rating using the four category rating scale. The subject terminated a search by pressing the space bar. The cases were then re-presented with feedback for 10 seconds, dependent on which group the subject was in. The feedback information consisted of the image overlaid with the scan path and fixations greater than 300ms presented as translucent circles (Figure 1). The longer the fixation the greater the diameter of the circle. The image was then re-presented with the observer again providing a rating and selecting each suspected nodule.

To obtain the data to determine the eye tracking metrics regions of interest (ROIs) needed to be defined. The ClearView ROI definition tool was used to define rectangular ROIs around each nodule, the size of which was defined to be 1.5 times the diameter of the lesion. Data from each of the 28 ROIs was then exported into a Microsoft Excel template and SPSS for further analysis.



STUDY: chest nodules. STIMULUS: AK 1. RECORDING: akrs1. FRAME: AK.jpg.
TIME SEGMENT: Only include fixations inside interval [45,43517] ms.

Figure 1. An example of the feedback presented to observers, the circles are fixations, and the crosses are where the observer has indicated there is a nodule.

4. RESULTS

4.1 JAFROC analysis

A repeated measures ANOVA was carried out. In the group that had “no feedback” no main effect was demonstrated (F(1,16) = 2.1, p= 0.164), therefore no post hoc tests were carried out. In the group that had “feedback” there was an effect (F (1,16) = 6.6, p = 0.021, partial Eta Squared = 0.3). So there was a significant difference between the pre and post “feedback” condition with a significant improvement following feedback (p = 0.021). Overall the mean percentage improvement was small of 3.3%, with most of the improvement due to the Level 1 group where the percentage improvement in the figure of merit (FOM) was 8.4% and this was significant (p<0.05).

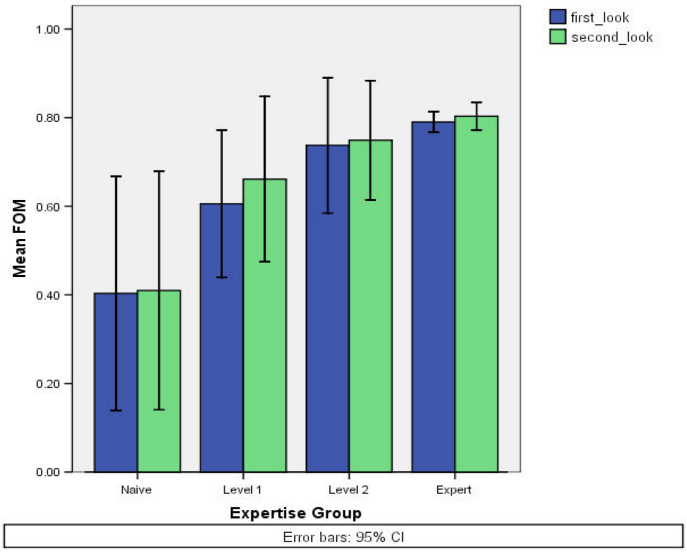


Figure 2. Mean figure of merit (FOM) pre and post feedback

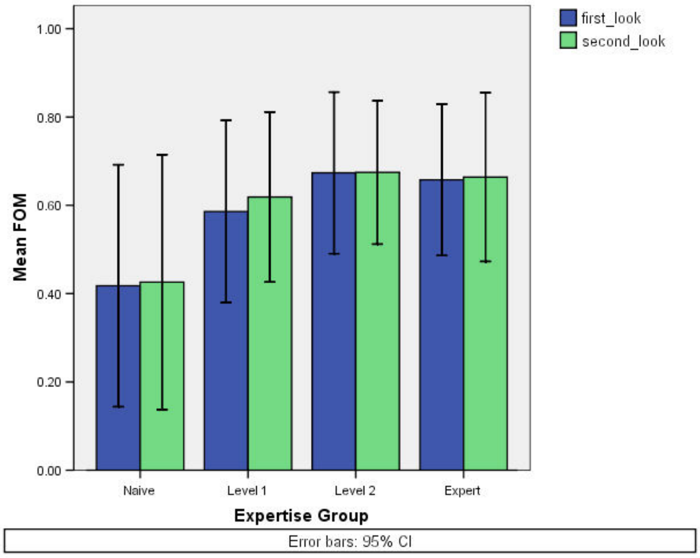


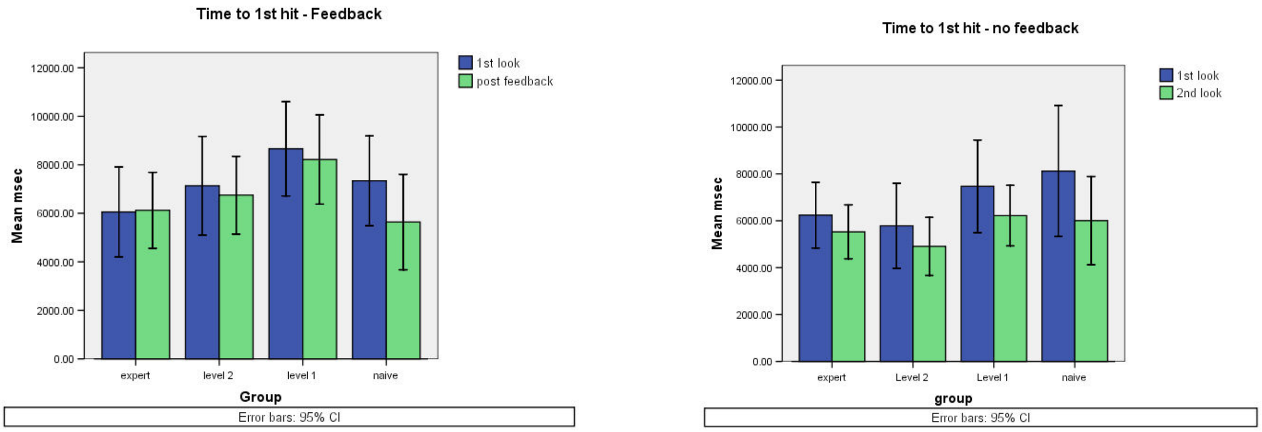
Figure 3. Mean figure of merit (FOM) pre and post a second look without feedback

4.2 Eye tracking results

For all eye tracking metrics a repeated measures ANOVA was carried out. As there were only two conditions the assumption of sphericity need not be assumed. Levene's Test of Equality of Error Variances showed that the assumption of homogeneity of variance was met in all cases.

4.3 Time to 1st hit

In the "no feedback" group there was a significant main effect (F(1,108) = 7.97, p=0.006), there were no significant interactions. In the "feedback" group there was no main effect.



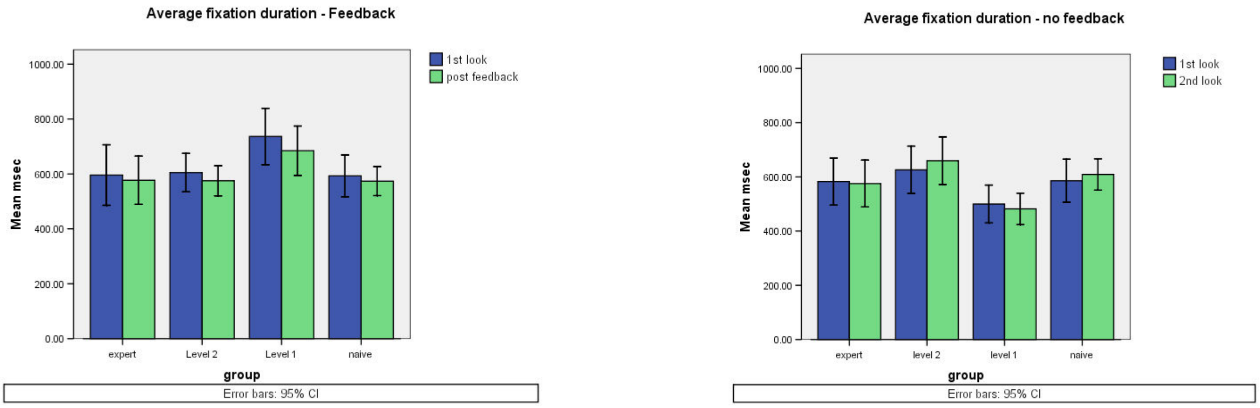
(a)

(b)

Figure 4. Time to 1st hit (msec) of the nodule in the ROI for (a) pre and post feedback, and (b) pre and post "no feedback"

4.4 Fixation duration

In both the "no feedback" group and "feedback" group there was no significant main effect. The mean fixation duration across all groups on the nodules was 600ms.



(a)

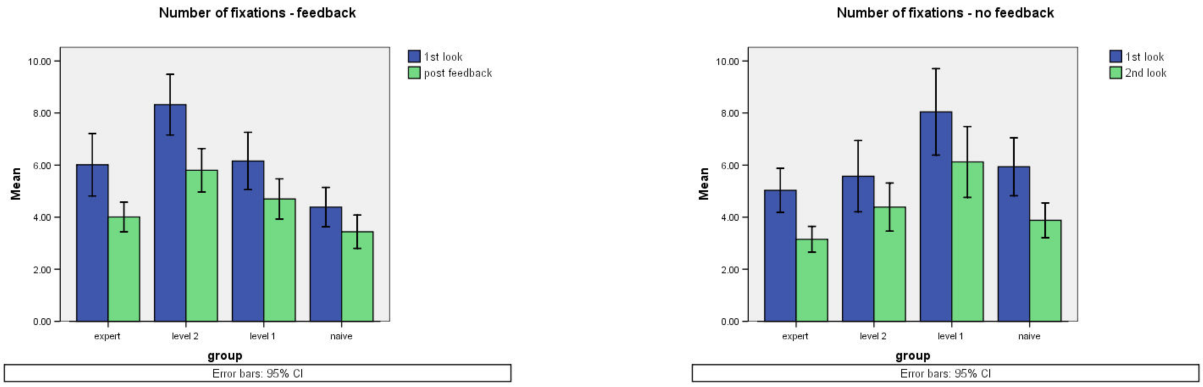
(b)

Figure 5. Average fixation duration (msec) in the ROI for (a) pre and post "feedback", and (b) pre and post "no feedback"

4.5 Number of fixations

In the “no feedback” group there was a significant main effect (F(1,108) = 60.36, p<0.001, partial Eta squared 0.34). The Tukey HSD post hoc test was carried out and the level 1 group was significantly different from the experts (p<0.001), level 2 (p=0.017), and naïve (p=0.015), with a greater number of fixations.

In the “feedback” group there was a significant main effect (F(1,108) =61.54, p<0.001, partial Eta squared 0.36). The Tukey HSD post hoc test revealed level 2 group was significantly different from the experts (p<0.001), level 1 (p=0.017), and naïve (p<0.001) with a greater number of fixations. The naïve group was also significantly different from level 1 (p=0.03) with a greater number of fixations.



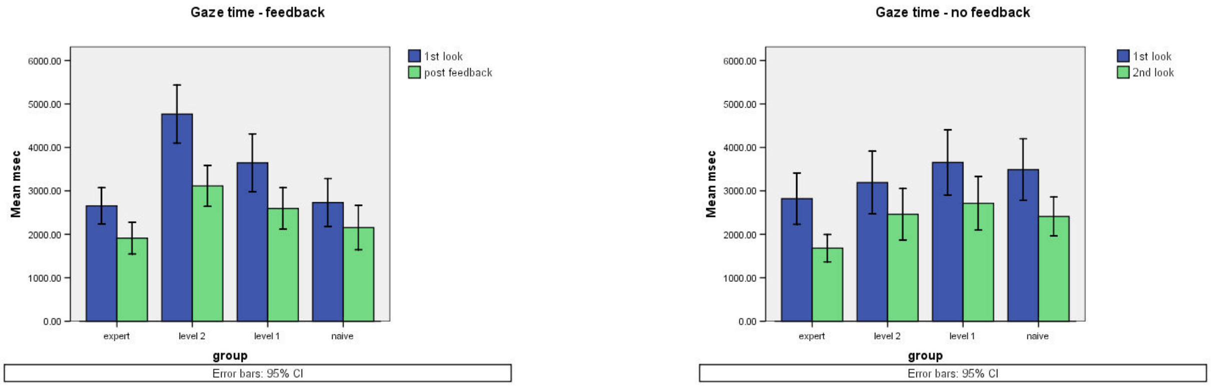
(a) (b)

Figure 6. The mean number of fixations in the ROI for (a) pre and post “feedback”, and (b) pre and post “no feedback”

4.6 Dwell time

In the “no feedback” group there was a main effect (F(1,108) 54.37, p<0.001, partial Eta 0.34), there were no significant interactions.

In the “feedback” group there was a main effect F (1,108) 74.47, p<0.001, partial Eta 0.41). The Tukey HSD post hoc test showed that level 2 was significantly different from experts (p<0.001) and the naïve group (p<0.001) with a greater dwell time.



(a) (b)

Figure 7. The dwell time (gaze time) in the ROI for (a) pre and post “feedback”, and (b) pre and post “no feedback”

5. DISCUSSION

5.1 JAFROC results

Performance measures such as the JAFROC results effectively summarise the general outcome of the visual search process in identifying nodules. The results in the "feedback" study are as would be predicted with performance improving with experience and interobserver variability reducing, although the pairwise comparisons show that only the naïve group has a significantly lower performance than the expert and level 2 group. In the "no feedback" study the expert group and level 2 were very similar, this is mainly due to the poor performance of two expert observers, one of whom had a low threshold for calling nodules (FOM 0.4577) and the other tended to only call one nodule on films with multiple nodules even though the eye tracking data indicated that all nodules were fixated (FOM 0.5786). The pairwise comparisons for the "no feedback" group reveal no significant differences between any of the groups. The setting for this study was an eye tracking laboratory and one should be cautious about the generalisability of the findings but ecological validity does seem to have been demonstrated with experts performing the best and naïve observers the worst, except for the two poorly performing experts who despite the experiment instructions, it appears that they applied their own criteria when viewing the test bank. The key point from the results however is the significant improvement in performance of the Level 1 group following feedback by 8.4%. This trend was also seen in the "no feedback" study, although not significant, with an improvement in the FOM of 5.3% by having an opportunity for a second look.

So why should this improvement occur in the Level 1 group. They were relatively inexperienced but will have had up to 12 weeks in clinical placement with exposure to many chest images and some anatomy and physiology lectures but with little training in pathology and pathological appearances.

It is probable that the perceptual mechanisms that allow efficient searching due to knowledge of normal appearances and knowledge of features that signal pathology [10] will not be developed in Level 1 students i.e they will not have developed a schema, they will be characterised by being slow and deliberate, possibly reflecting their use of explicit rules and reliance on knowledge in the form of rules or ideal cases [11], leading to an inability to modify a schema in response to additional and conflicting data [12]. To recognise an object is to decide that its perceptual representation is similar to a representation created during a previous experience and stored in memory, which is the way it is hypothesised that experts make decisions by depending on pattern recognition and exemplar based judgments [13]. In this study the decision time results (not shown in this paper) demonstrate the level 1 are no slower than any other group on the film with pathology but along with the Level 2 group are slower on the film with no nodules. It could be that the Level 1 find the feedback information useful precisely because they lack a schema and find any additional decision support constructive.

The lack of statistically significant differences between the groups is surprising, and may be due to lack of numbers in the study and large variability between subjects, but other studies have had similar results showing that novices can be surprisingly good at radiology tasks. A study by Manning [14] also using a lung nodule task found that the average performance of the novices was not significantly different from experienced radiographers about to undergo a course in image interpretation of the chest, they also showed a similar degree of inter-observer variance. Manning suggests that this could support the notion that a degree of innate skill exists; but it is likely that all radiographers by being exposed to chest radiographs on a daily basis will become perceptually adapted to their appearance, and improvements in perceptual performance occur quickly [15].

The results show that the performance improvement for the Level 1 group is greater following the provision of scan path information than "no feedback". Where scan path information has been presented to subjects in other studies it has usually been that of experts with the aim being to improve performance by offering subjects insight into the task or problem [16, 17,18]. In this study the scan path information is probably being used in a different way. In the same way as CAD highlights potential lesions, observers own fixations will identify potential perturbation in the image that could be nodules, yet the observer is unaware of their interest in the area until the feedback information makes it explicit. The results suggest that the level 1 students are more receptive to using this information and possibly adjusting their thresholds when their confidence in the decision is made, whereas those with greater experience or knowledge will be making their initial hypothesis quickly i.e. fast and frugal [19]. The naïve observers will be taking a common sense approach to the problem from their limited conceptual knowledge of a nodule using perceptual skills that we all have.

Eye tracking metrics

Whereas the FOM is a performance measure, eye tracking metrics are process measures and the utility of a process measure should rely on its explanation of performance [20]. Time to first hit is well recognised as a metric of expertise with experienced radiologists identifying pathology quickly and accurately [21, 22]. In this study the mean time for the experts pre "feedback" and pre "no feedback" was 6.05s and 6.24s respectively with a range of 0.6s – 18s and 1.1s – 17.4s and the pairwise comparisons demonstrate no significant differences between the expertise groups, so in this study, time to 1st hit is not apparent as a marker of expertise. Following "feedback" or "no feedback" it was the naïve group that had the biggest fall in time to first hit possibly suggesting they were using the post "feedback" and post "no feedback" as confirmation of their initial search rather than processing any additional information from the feedback.

Average fixation duration in the nodule ROIs was 600ms with no significant differences between groups, no time restrictions were imposed on subjects, if there had been then experts may have made brief eye fixations.

The number of fixations in the ROIs declines following "feedback" and "no feedback". This is to be expected with the adoption of an improved search strategy due to learning or familiarisation with the task. Alternatively observers could be looking elsewhere in the image. Once again differences between expertise groups are unclear. Experts do seem to make fewer fixations, as do the naïve group however they fail to detect many nodules as demonstrated by the data on nodules not fixated. There is a great deal of variability with the Level 1 and Level 2 groups. Previous research by Manning et al [14] has demonstrated distinct differences between radiologists and radiographers in that even though they may achieve a similar level in task performance radiologists have a tendency to fixate fewer times on the film and will exclude regions of the image dependent on the film, whereas radiographers are more regimented in the way they scan films. The dwell time data reflects the number of fixations with fewer return fixations.

The eye tracking data of the Level 1 and Level 2 groups displays a great deal of variability, this variability has been noted by Wooding et al [23] who report the variable fixation location of trainees is least similar to those of radiologists than those of any group even the controls, and is characterised by idiosyncratic eye movements. Most studies have looked at the transition from junior radiologist to experienced radiologist rather than radiography students, but the transition to experienced radiologist does involve a period of disorder. Kundel and Nodine [24] hypothesise that the loss of accuracy of intermediate film readers could be attributed to the conflict between perceptual and cognitive processes, and Lesgold [25] who took a cognitive approach using verbal protocol analysis to gain an insight into radiological expertise observed that intermediate individual demonstrate worse performance than novices and experts in a transition phase. It is likely the same thing is happening with Level 1 and Level 2 students as they create schema in order to process the information found in a chest radiograph.

6. CONCLUSION

The type of radiological task and level of expertise are important considerations if perceptual feedback is to be used as a method of improving performance. It appears that the Level 1 students, i.e. those with some experience gain the most from feedback, whereas it had no impact on the performance of experienced radiologists. This finding potentially has implications for CAD or decision support software that may not be appropriate for those who are already experts.

ACKNOWLEDGEMENTS

This work was supported in part by funding from the Centre for Health Research and Practice Development (CHRPD) at the University of Cumbria.

REFERENCES

[1] Donovan, T., Manning, D.J., Philips, P.W., Higham, S., Crawford, T., "The effect of feedback on performance in a fracture detection task," Proc. SPIE 5749, 79-85 (2005).

[2] Kundel, H. L., Nodine, C. F., & Krupinski, E. A., "Computer-displayed eye position as a visual aid to pulmonary nodule interpretation," Investigative radiology, 25(8), 890-896 (1990).

[3] Carmody, D. P., Nodine, C. F., & Kundel, H. L., "Finding lung nodules with and without comparative visual scanning," Perception & psychophysics, 29(6), 594-598 (1981).

[4] Nodine CF, Kundel HL. A visual dwell algorithm can aid search and recognition of missed lung nodules in chest radiographs. In Brogen D (Ed.), [Visual search] 1st edition Taylor S Francis, London 1990.

[5] Manning, D., Barker-Mill, S. C., Donovan, T., & Crawford, T., "Time-dependent observer errors in pulmonary nodule detection," The British journal of radiology, 79(940), 342-346 (2006).

[6] Gale, A.G., Human Response to Visual Stimuli, in Hendee, W. R., & Wells, P. N. T., [The perception of visual information (2nd ed.)] New York: Springer (1997).

[7] McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R., "Visual skills in airport-security screening," Psychological science : a journal of the American Psychological Society / APS, 15(5), 302-306 (2004).

[8] Wolfe, J. M., Horowitz, T. S., & Kenner, N. M., "Cognitive psychology: Rare items often missed in visual searches," Nature, 435(7041), 439-440 (2005).

[9] Brogden, B.G., Kelsey, C.A., Moseley, R.D., "Factors affecting the perception of pulmonary lesions," Radiologic Clinics of North America, 21, 4, 633-654 (1983).

[10] Myles-Worsley, M., Johnston, W. A., & Simons, M. A., "The influence of expertise on X-ray image processing," Journal of experimental psychology. Learning, memory, and cognition, 14(3), 553-557 (1988).

[11] Palmeri, T. J., Wong, A. C., & Gauthier, I., "Computational approaches to the development of perceptual expertise," Trends in cognitive sciences, 8(8), 378-386 (2004).

[12] Wood, B. P., "Visual expertise," Radiology, 211(1), 1-3 (1999).

[13] Posner, M. I., Petersen, S. E., Fox, P. T., & Raichle, M. E., "Localization of cognitive operations in the human brain," Science (New York, N.Y.), 240(4859), 1627-1631 (1988).

[14] Manning, D.J, Ethell, S.C., Donovan, T., "Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph", Br J Radiol, 77: 231-235 (2004).

[15] Sowden, P. T., Davies, I. R., & Roling, P., "Perceptual learning of the detection of features in X-ray images: A functional role for improvements in adults' visual sensitivity?" Journal of experimental psychology.Human perception and performance, 26(1), 379-390 (2000).

[16] Velichkovsky, B.M., "Communicating attention: gaze position transfer in cooperative problem solving," Pragmatics and Cognition, 3, 199-222 (1995).

[17] Nalangula, D., Greenstein, J.S., Gramopadhye, A. K., "Evaluation of the effect of feedforward training displays of search strategy on visual search performance," International Journal of Industrial Ergonomics, 36, 289-300 (2006).

[18] Gramopadhye, A. K., Drury, C. G., & Sharit, J., "Feedback strategies for visual search in airframe structural inspection," International Journal of Industrial Ergonomics, 19(5), 333-344 (1997).

[19] Gigerenzer, G., & Goldstein, D. G., "Reasoning the fast and frugal way: Models of bounded rationality," Psychological review, 103(4), 650-669 (1996).

[20] Duchowski, A. T., [Eye tracking methodology : Theory and practice], New York: Springer. (2003).

[21] Nodine, C. F., Kundel, H. L., Lauver, S. C., & Toto, L. C., "Nature of expertise in searching mammograms for breast masses," Academic Radiology*, 3*(12), 1000-1006 (1996).

[22] Krupinski, E. A., "Visual scanning patterns of radiologists searching mammograms," Academic Radiology, 3(2), 137-144 (1996).

[23] Wooding, D.S., Roberts, G.M., Phillips-Hughes, J., "The development of the eye movement response in the trainee radiologist," Proc. SPIE 3663, 136-145 (1988).

[24] Kundel, H. L., & Nodine, C. F., "A visual concept shapes image perception. Radiology", 146(2), 363-368 (1983).

[25] Lesgold, A., Rubinson, H., Feltovich, P., et al. "Expertise in a complex skill: diagnosing x-ray pictures," in Chi, M.T.H., Glazer, R., Farr. M.J. ed. [The Nature of Expertise] Hillsdale, NJ:LEA (1988).