Learning Image-Derived Eye Movement Patterns to Characterize Perceptual Expertise

Rui Li (rxl5604@rit.edu)
College of Computing and Information Science, 1 Lomb Memorial Drive
Rochester, NY 14623 USA

Jeff Pelz (pelz@cis.rit.edu)
College of Imaging Science, 1 Lomb Memorial Drive
Rochester, NY 14623 USA

Pengcheng Shi (spcast@rit.edu) Anne R. Haake (arhics@rit.edu)
College of Computing and Information Science, 1 Lomb Memorial Drive
Rochester, NY 14623 USA

Abstract

Experts have remarkable capability of locating, identifying and categorizing objects in their domain-specific images. Eliciting experts' visual strategies will benefit image understanding by transferring human domain knowledge into image-based computational procedures. In this paper, an experiment conducted to collect both eye movement and verbal description data from three groups of subjects with different medical training levels (eleven board-certified dermatologists, four dermatologists in training and thirteen novices) while they were examining and describing 42 photographic dermatological images. We present a hierarchical probabilistic framework to discover the stereotypical and idiosyncratic viewing behaviors exhibited within each group when they are diagnosing medical images. Furthermore, experts' annotations of thought units on the transcribed verbal descriptions are time-aligned with discovered eye movement patterns to interpret their semantic meanings. By mapping eye movement patterns to thought units, we uncover the manner in which these subjects alternated their behaviors over the course of inspection and how these experts parse the images.

Keywords: Eye movements; eye tracking; verbal description; multimodal data analysis; graphical model; user study; diagnostic reasoning

Introduction

Perceptual expertise is considered to be the crucial cognitive factor accounting for the advantage of highly trained experts (Hoffman & Fiore, 2007). Experts generate distinctively different perceptual representations when they view the same scene as novices (Palmeri, Wong, & Gauthier, 2004 ; Smuc, Mayr, & Windhager, 2010). Rather than passively "photocopying" the visual information directly from sensors into minds, visual perception actively interprets the information by altering perceptual representations of the images based on experience and goals. By analyzing the whole sequences of fixation and saccadic eye movements from groups with different expertise levels, significant differences in visual search strategies between groups show human expertise plays a great role in medical image examination. In (Manning, Ethell, Donovan, & Crawford, 2006), the nature of expert performance of four observer groups with different levels of expertise was investigated . They compared multiple eye movement measures and suggested these distinctive variations among the observations of the better performance from higher expertise level are due to the consequences of experience and training. In (Krupinski et al., 2006), an eye movement study was conducted on diagnostic pathology of light microscopy to identify distinctive viewing stereotypes for each level of experience . Their results suggest eye movement monitoring could serve as a basis for the creation of innovative pathology training routines.

In knowledge-rich domains, perceptual expertise is particularly valuable. Medical image understanding via manually marking and annotating become not only labor intensive for experts but also ineffective because of the variability and noise of experts' performance (Gordon, Lotenberg, Jeronimo, & Greenspan, 2009). For training and designing decision support systems, the basic perceptual strategies and principles of diagnostic-reasoning are also desired (Dempere-Marco, Hu, & Yang, 2011). To address this problem, it requires the ability of extracting and representing experts' perceptual expertise in a form that is ready to be applied. In this work, our contributions are: first, we discover and represent expertise-related eye movement patterns exhibited among multiple experts in an objective and unbiased way; second, to validate these patterns, we identify their semantic meanings by time-aligning them with standardized thought units annotated by additional experts. Third, we also characterize the eye movement patterns of three different expertise levels respectively which can be used to categorize users' expertise levels based on their visual inspection on medical images.

Human viewing behaviors are valuable yet effortless resources worth of exploiting. In specific domains experts perceptual expertise is considered to be more consistent and informative than their manual markings. Human vision is an active dynamic process in which the viewer seeks out specific information to support ongoing cognitive and behavioral activity (Henderson & Malcolm, 2009). Since visual acuity is limited to the foveal region and resolution fades dramatically in the periphery, we move our eyes to bring a portion of the visual field into high resolution at the center of gaze. Studies have shown that visual attention is influenced by two main sources of input: bottom-up visual attention driven by low-level saliency image features and top-down process in which cognitive processes, guided by the viewing task and scene context, influence visual attention (Torralba, Oliva, Castelhano, & Henderson, 2006 ; Loboda, Brusilovsky, & Brunstein, 2011). Growing evidence suggests that top-down information dominates the active image viewing process and the influence of low-level salience guidance is minimal (Castelhano, Mack, & Henderson, 2009). These theoretical outcomes provide us with the possibility to capture experts' cognitive strategies, perceptual expertise and expectations by investigating their stereotypical and idiosyncratic viewing behaviors, and decode their semantic meanings.

In our work we focus on medical images where domain knowledge and perceptual expertise are in demand. We elicit and model physicians' perceptual and conceptual expertise from their diagnostic reasoning process while inspecting medical images. Physicians examine medical images and verbally describe their thinking process as if teaching a trainee, and both their eye movements and verbal descriptions are recorded. In order to capture the stereotypical and idiosyncratic eye movement patterns exhibited among these physicians, we develop a hierarchical dynamic model. This model allows us to build a library of all the patterns exhibited by physicians' time-evolving eye movement series (scanpaths) and each eye movement pattern essentially corresponds to a particular statistical regularity of the temporal-spatial properties inferred from multiple eye movement series. Thus each physician's eye movement time series can be characterized by a particular combination of a subset of these patterns from this library. To investigate the relationships between visual and verbal conceptual processing by analyzing the verbal descriptions. additional experts annotate the transcribed verbal descriptions using standardized semantic labels (thought units) that describe the process of creating a differential diagnosis from their domain knowledge (Habif, Jr., Chapman, Dinulos, & Zug, 2005). After time-aligning these thought units with the eye movements patterns, we discovered significant correlations between them. This results indicate that the patterns we extracted from eye movement data possess distinct and specific semantic meanings in terms of human capabilities of image understanding.

Experiment

Subjects recruited for the eye tracking experiment belong to three groups based on their training level including 11 board-certified dermatologists (attending physicians), 4 dermatologists in training (residents) and 13 undergraduate lay people (novices). We also recruited physician assistant students who served as "trainees" in order to motivate dermatologists to verbalize their diagnosis reasoning using the Master-Apprentice scenario, which is known to be effective in eliciting detailed descriptions.

A SMI (Senso-Motoric Instruments) eye tracking apparatus was applied to display the stimuli at a resolution of 1680x1050 pixels for the collection of eye movement data and recording of verbal descriptions. The eye tracker was running at 50 Hz sampling rate and has accuracy of 0.5^o visual angle. The subjects viewed the medical images binocularly at a distance of about 60 cm. The experiment was conducted in an eye tracking laboratory with ambient light.

A set of 42 dermatological images, each representing a different diagnosis, was selected for the study. These images were presented to subjects on the monitor. Medical professionals were instructed to examine and describe each image to the students while working towards a diagnosis, as if teaching. The experiment lasted approximately 1 hour. The medical professionals were instructed not only to view the medical images and make a diagnosis, but also to describe what they see as well as their thought processes leading them to the diagnosis. The novice observers were instructed to examine the images and offer a detailed description as if describing to their doctors over the phone. Both eye movements and verbal descriptions were recorded for the viewing durations controlled by each subject. The experiment started with a 13-point calibration and the calibration was validated after every 10 images. Calibration is accepted if its variance is less then 0.5^o. The audio recordings of the verbal descriptions from the dermatologists were transcribed and annotated.

An annotation study was conducted on the transcripts to investigate the semantic interpretations of the estimated eye movement patterns. During annotation two highly trained dermatologists identified 9 thought units. A thought unit is a single word or group of words that receives a descriptive label based on its semantic role in the diagnostic process. The thought unit labels are patient demographics (DEM), body location (LOC), configuration (CON), distribution (DIS), primary morphology (PRI), secondary morphology (SEC), differential diagnosis (DIF), final diagnosis (Dx), and recommendations (REC). Words not belonging to a thought unit were designated as 'None'. These two physicians annotated transcribed verbal descriptions with these thought units. The annotation were then time-aligned with eye movement patterns. Using this method, each unit of eye movement data, which is composed of a fixation and its successive saccade, receives two labels: one is its pattern indicator inferred by the model and the other is its time-aligned thought unit annotated through the consensus of multiple experts. This result allows us to interpret the eye movement patterns by measuring the correspondence between them and the thought units.

Hierarchical Dynamical Model

A hierarchically-structured dynamical model was developed to capture both the common eye movement patterns shared among multiple expertise-specific groups of subjects and unique eye movement patterns exhibited by individuals. The hierarchical beta processes proposed by Thibaux et al.(Thibaux & Jordan, 2007) as a prior distribution of our model provides the flexibility of discovering more patterns as new eye movement data are observed. Since fixation and

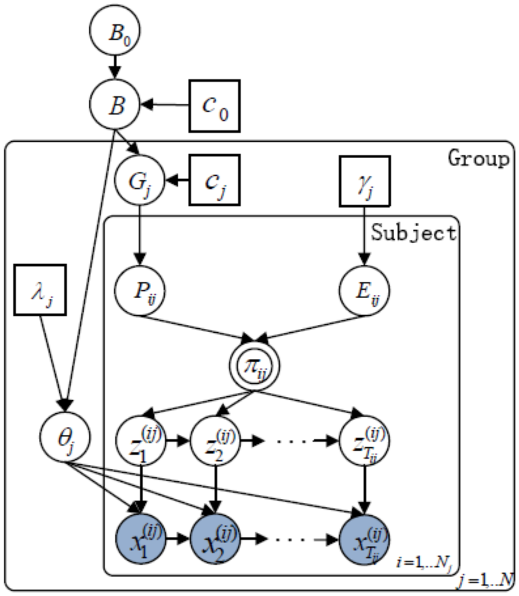


Figure 1: Hierarchical dynamical model. B_0 denotes a fixed continuous complete random measure as a global baseline on the space of Θ. B denotes a beta process to measure the eye movement patterns shared among N groups. G_j denotes a beta process to represent the eye movement patterns shared among N_j subjects of group j. The transition distribution π_{ij} of subject i in group j is deterministic. $z_{t_{ij}}^{(ij)}$ and $x_{t_{ij}}^{(ij)}$ denotes the hidden state variable and the observation variable of the hidden Markov model. θ_j denotes the emission distribution. The total number of eye movement patterns exhibited in the group j is denoted by K_j which is depend on B.

saccadic data are deployed in a sequential manner we use a hidden Markov model (HMM) as the likelihoods to characterize their temporal dynamic nature. Eye movements are inherently not smooth and highly correlated, the strong Markovian assumption of HMMs is inappropriate. We therefore employ autoregressive HMMs to relax the Markovian assumption by modeling eye movement data as a noisy linear combination of some finite set of past observations plus additive white noise. We utilize this hierarchical prior in the following specification based on our problem scenario.

Let B_0 denote a fixed continuous random base measure on a space Θ which represents a library of all the potential eye movements patterns. For multiple groups to share patterns, let B denote a discrete realization of a beta process given the prior $BP(c_0, B_0)$. Let G_j be a discrete random measure on Θ drawn from B following the beta process which represents a random measure on the eye movement patterns shared among multiple subjects within the group j. Let P_{ij} denote a Bernoulli measure given the beta process G_j. P_{ij} is a binary vector of Bernoulli random variables representing whether a particular eye movement pattern exhibited in the eye movement data of subject i within group j. This hierarchical construction can be formulated as follow:

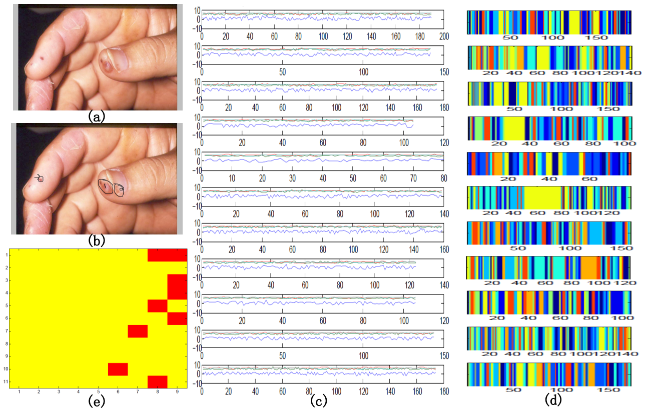


Figure 2: The model running on the eye movement data of 11 subjects viewing one case. (a) shows the original medical image. Images used with the permission Logical Images, Inc. (b) primary and secondary abnormalities were explicitly marked and numbered by an experienced dermatologist. (c) shows eleven time series, each observation of which is composed of 4 components: log values of fixation location (xy coordinate), fixation duration and saccade amplitude. (d) shows the HMM-derived eye movement pattern sequences for the corresponding 11 time series with 4 chains of 55000 sampling iterations. The color coding corresponds to the segments of each specific eye movement pattern. (e) shows the shared eye movement pattern matrix of which the row number indicates the subjects and the column number indicates the shared patterns. For example, yellow color at the first row represents the time series of subject 1 who exhibits pattern 1-7 but lacks pattern 8 and 9.

$$B|B_0 \sim BP(c_0, B_0) \tag{1}$$

$$G_j|B \sim BP(c_j, B) \qquad j = 1, ..., N \tag{2}$$

$$P_{ij}|G_j \sim BeP(G_j) \qquad i = 1, ..., N_j \tag{3}$$

where $G_j = \sum_k g_{jk}\delta_{\theta_{jk}}$. This term shows that G_j is associated with both a set of countable number of eye movement patterns $\{\theta_{jk}\}$ drawn from the eye movement pattern library Θ and their corresponding probability masses $\{g_{jk}\}$ given group j. The combination of these two variables characterizes how the common eye movement patterns shared among subjects within expertise-specific group j. Thus P_{ij} as a Bernoulli process realization from the random measure G_j is denoted as:

$$P_{ij} = \sum_k p_{ijk}\delta_{\theta_{jk}} \tag{4}$$

where p_{ijk} as a binary random variable denotes whether subject i within group j exhibits eye movement pattern k given probability mass g_{jk}. Based on the above formulation, for $k = 1...K_j$ patterns we readily define $\{(\theta_{jk}, g_{jk})\}$ as a set of common eye movement patterns shared among group j and

$\{(\theta_{jk}, p_{ijk})\}$ as subject i's personal subset of eye movement patterns given group j, as shown in Figure 1.

The transition distribution $\pi_{ij} = \{\pi_{z_t^{(ij)}}\}$ of the hidden Markov model at the bottom level governs the transitions between the i^{th} subject's personal subset of eye movement patterns θ_{jk} of group j. It is determined by the element-wise multiplication between the eye movement subset $\{p_{ijk}\}$ of subject i in group j and the gamma-distributed random variables $\{e_{ijk}\}$:

$$e_{ijk}|\gamma_j \sim Gamma(\gamma_j, 1) \tag{5}$$

$$\pi_{ij} \propto E_{ij} \bigotimes P_{ij} \tag{6}$$

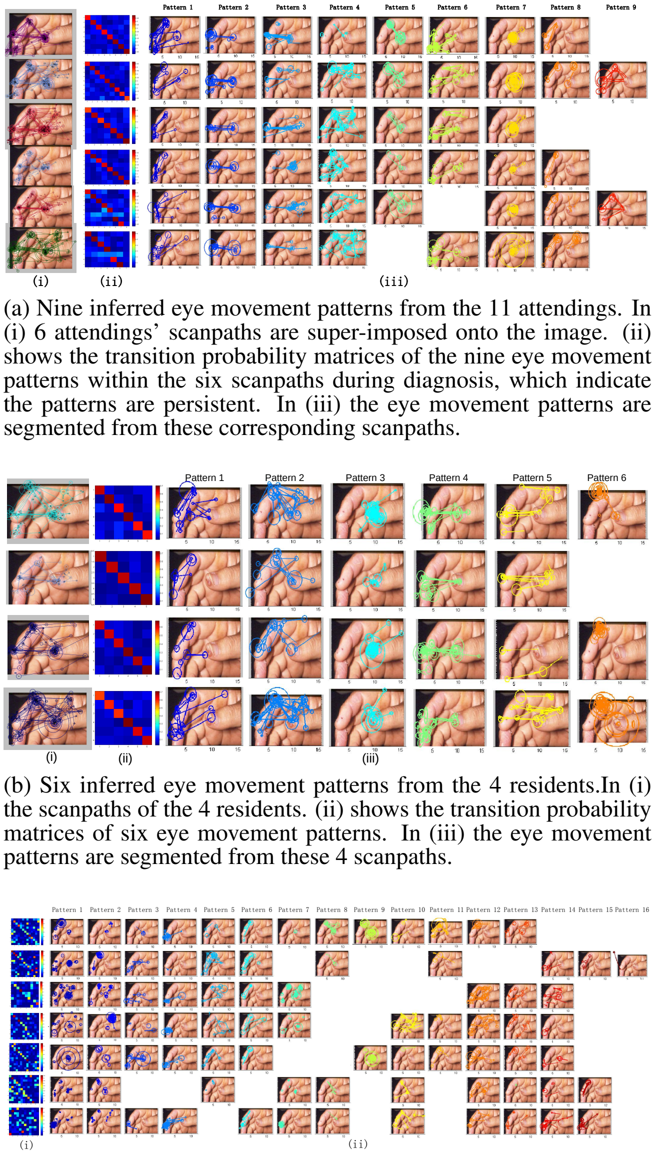
where $E_{ij} = [e_{ij1}, \ldots e_{ijK_j}]$. So the effective dimensionality of π_{ij} is determined by P_{ij}, which is inferred from observations.

We use Markov chain Monte Carlo sampler to do the posterior inference over this model. In one iteration of the sampler, each latent variable is visited and assigned a value by drawing from the distribution of that variable conditional on the assignments to all other latent variables as well as the observation. In particular, based on the sampling algorithm proposed in (Thibaux & Jordan, 2007), we developed a Gibbs sampling solution to the hierarchical beta processes part of the model.

Results and Discussion

In Figure 2, we illustrate one set of observed data and estimating processes from the framework of the 11 dermatologists diagnosing a case of a skin manifestation of endocarditis. In the medical image, there are multiple skin lesions spreading over the thumb nail and tip, the two parts of index finger and the middle finger as marked in (b) of Figure 2. A primary abnormality is on the thumb tip. The scanpaths in Figure 3a (i) indicate that dermatologists fixated on the primary abnormality heavily and switch their visual attention actively between and within the primary and secondary findings. The estimated patterns are color-coded as panels shown in (d) of Figure 2. These panels describe the time-evolving manner in which each individual alters eye movement patterns at the individual level.

Pattern occurrence and thought unit alignment resulted in assignment of each fixation to a specific pattern and to a thought unit (or None). Initial integration of eye movement patterns with thought units was accomplished by calculating correspondence in Figure 4. Analysis shows, for example, that primary morphology (PRI) is closely related to the combination of two specific patterns: Pattern 2 is characterized by fixations switching between the primary and the different secondary abnormalities; and Pattern 7 by long fixations only on the primary abnormality. These patterns suggest dermatologists were seeking meaningful ways to integrate these two findings for some principled reasons, although these informative findings are separable in the sense that they are operationally defined and measured independently of one another. Pattern 7 has strong relationship to location (LOC) which ap-



(a) Nine inferred eye movement patterns from the 11 attendings. In (i) 6 attendings' scanpaths are super-imposed onto the image. (ii) shows the transition probability matrices of the nine eye movement patterns within the six scanpaths during diagnosis, which indicate the patterns are persistent. In (iii) the eye movement patterns are segmented from these corresponding scanpaths.

(b) Six inferred eye movement patterns from the 4 residents.In (i) the scanpaths of the 4 residents. (ii) shows the transition probability matrices of six eye movement patterns. In (iii) the eye movement patterns are segmented from these 4 scanpaths.

(c) Sixteen inferred eye movement patterns from the the 13 novices.In (i) the transition probability matrices of the sixteen eye movement patterns, which suggest novices' visual behaviors are not persistent. In (ii) the patterns are segmented from these 7 scanpaths.

Figure 3: The inferred eye movement patterns of the three expertise-specific groups. Each observation unit of the eye movement sequences is composed of 4 components: fixation location (xy coordinate), fixation duration and saccade amplitude. We then apply our model on these sequential data to reveal the subtlety of the behavioral patterns varying over time. The inferred patterns were derived with 4 chains of 55000 sampling iterations. The color coding specifies the segments of each specific pattern.

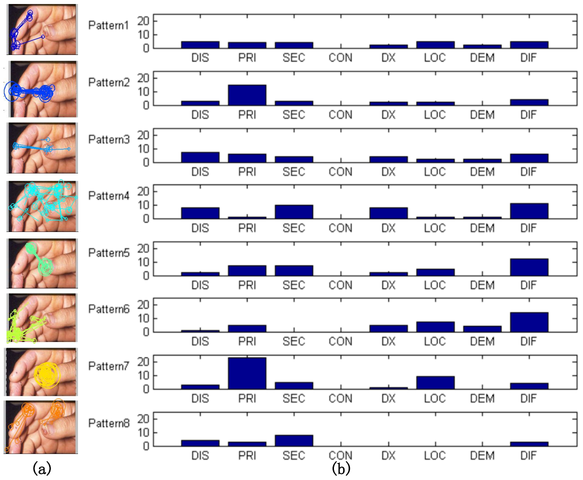


Figure 4: Correspondence between the 8 eye movement patterns of the 11 attendings (the rows) and their 8 thought units (the columns). (a) the representative patterns. (b) histograms show the corresponding relationship between discovered eye movement patterns and annotated thought units. For each pattern we plotted the counts of fixations which are labeled as the 9 thought units.

pears to correspond to the primary morphology location. Pattern 4 consists of scanpath segments which are characterized by shorter fixation durations and longer saccades. This scanning behavior strongly corresponds to thought units, including distribution (DIS), secondary morphology (SEC), diagnosis (DX) and differential diagnosis (DIF). Scanning pattern coupled with thought unit DX is possibly related to confirmation of secondary findings to support or rule out diagnostic hypotheses.

Some similar patterns also emerged in the resident group but is lacking in the novice group as shown in Figure 3b. This suggests that experts, equipped with domain knowledge organized in finer gradations of functional categories, can discriminate the significance of their findings in a particular context. In contrast, in Figure 3c the novices failed to do so, although they perceive the same abnormalities too. Compare Figure 3a (ii), Figure 3b (ii) and Figure 3c (i), the difference between the transition probability matrices of the three expertise-specific groups suggests professionals' eye movement patterns are more persistent than the novices'.

These results suggest that there exist structural regularities of experts' diagnostic-reasoning processes, and such perceptual and conceptual processing regularities can be captured and manifested through experts' eye movements. This is consistent with previous empirical studies (Patel, Arocha, & Kaufman, 2001). These discovered stereotypical eye movement patterns indicate that experts are able to rapidly invoke the appropriate specific knowledge and expertise, and initially detect a general pattern of disease. These capabilities lead them to a gross anatomic localization and narrow down the possible interpretations. On the other hand, novices have hard time to focus on the important structures and are more likely to maintain inappropriate interpretations.

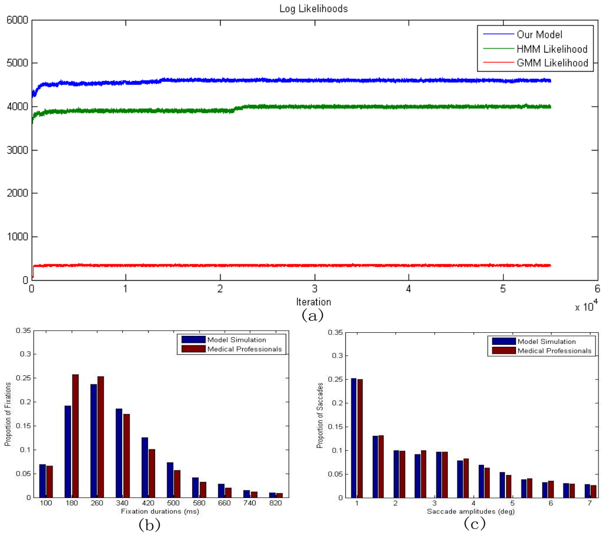
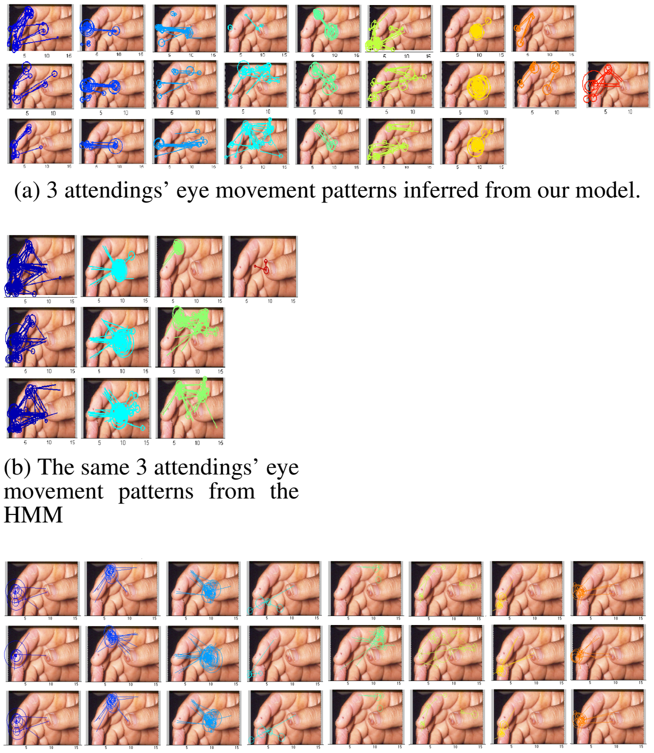


Figure 5: Quantitative performance evaluations. (a) The likelihood-value plots of a Gaussian mixture model, a hidden Markov model and our model after 55000 sampling iterations on our data-set. (b) The histogram of the fixation duration distributions of the 15 professionals (attendings and residents) and our model's simulations over 42 images. (c) The histograms of the saccade amplitude distributions of the 15 professionals and our model's simulations over 42 images.

To measure performance, we compared the log-likelihood values among our model, a hidden Markov model (HMM) and a Gaussian mixture model (GMM) as shown in Fig. 5 (a). To implement the HMM and GMM, we have to assume each eye movement sequence exhibits the same set of patterns. The log-likelihood values of our model and the GMM are 4000 vs. 300, which indicates our model fits the observation better. One possible cause is that the GMM makes a strong assumption that the eye movement data are independent which is hardly true. On the contrary, our model only assumes that the eye movement patterns are exchangeable in order. Additionally, our model and the HMM take sequential information of eye movements into account. We visualized the eye movement patterns from HMM and GMM in Fig. 6 (b)-(c) and make a comparison with our model's results in Fig. 6 (a). In Fig. 5 (b)-(c), our model generated 7356 fixation-saccade units to simulate the 15 professionals. It is worth noting that this result also validates the discovered eye movement patterns. Such simulation requires us to generate a set of realizations of eye movement patterns first from the hierarchical prior, simulate multiple possible sequences of these patterns, and then draw fixation-saccade samples from them.

Conclusions

Our approach identified and semantically interpreted both stereotypical and idiosyncratic expertise-specific eye move-



(a) 3 attendings' eye movement patterns inferred from our model.

(b) The same 3 attendings' eye movement patterns from the HMM

(c) The same 3 attendings' eye movement patterns from the GMM.

Figure 6: Illustrations of the attendings eye movement patterns from the three models.

ment patterns that only exist over time. In our future work, the discovered eye movement patterns will be related to image features by projecting the patterns from their temporal-spatial space into the image feature space. We will not only identify the most valuable image feature sets leading to a correct diagnosis but also uncover how a particular feature's importance changes over the course of the diagnostic reasoning process. These discoveries will provide training information to novices on how to look for relevant image features. Evaluation of a subject's expertise level is another future study. We can identify the expertise level given a subject's visual interaction with test images through calculating the model's posterior probability. Compared to simply calculating diagnosis error rates to evaluate expertise level, our approach can unveil which diagnostic reasoning steps lead to wrong diagnosis and the possible cognitive factors such as misconception, miscategorization and misperception, and form the basis of support systems.

Acknowledgements

We thank M.D. Cara Calvelli and Dr. Cecilia Ovesdotter Alm for many helpful discussions. This work was supported by research grants from the National Science Foundation (IIS-0941452) and the NIH/NLM (R21 LM010039-01).

Références

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing Task Influences Eye Movement Control during Active Scene Perception. *J. Vision*, *9*(3), 1-15.

Dempere-Marco, L., Hu, X., & Yang, G.-Z. (2011). A novel framework for the analysis of eye movements during visual search for knowledge gathering. *Cognitive Computation*, *3*, 206–222.

Gordon, S., Lotenberg, S., Jeronimo, J., & Greenspan, H. (2009). Evaluation of uterine cervis segmentations using ground truth from multiple experts. *J. Computerized Medical Imaging and Graphics*, *33*(3), 205-216.

Habif, T. P., Jr., J. L. C., Chapman, M. S., Dinulos, J. G., & Zug, K. A. (2005). Skin disease diagnosis and treatment. Elsevier Mosby.

Henderson, J. M., & Malcolm, G. L. (2009). Searching in the Dark Cognitive Relevance Drives Attention in Real-world Scenes. *Psychonomic Bulletin and Review*, *16*(5), 850-856.

Hoffman, R., & Fiore, M. S. (2007). Perceptual (re)learning : a leverage point for human-centered computing. *J. Intelligent Systems*, *22*(3), 79-83.

Krupinski, E., Tillack, A., Richter, L., Henderson, J., Bhattacharyya, A., Scott, K., et al. (2006). Eye-movement study and human performance using telepathology virtual slides. implications for medical education and differences with experience. *Journal of Human Pathology*, *37*(12), 1543–1556.

Loboda, T. D., Brusilovsky, P., & Brunstein, J. (2011). Inferring word relevance from eye-movements of readers. In *Proc. iui* (pp. 175–184). ACM Press.

Manning, D., Ethell, S., Donovan, T., & Crawford, T. (2006). How do radiologists do it? the influence of experience and training on searching for chest nodules. *Journal of Radiography*, *12*(2), 134–142.

Palmeri, T. J., Wong, A. C.-N., & Gauthier, I. (2004). Computational approaches to the development of perceptual expertise. *TRENDS in Cognitive Sciences*, *8*(8), 378–386.

Patel, V. L., Arocha, J. E., & Kaufman, D. R. (2001). A primer on aspects of cognition for medical informatics. *J Am Med Inform Assoc.*, *8*, 324–343.

Smuc, M., Mayr, E., & Windhager, F. (2010). The game lies in the eye of the beholder: The influence of expertise on watching soccer. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 1631–1636). Austin, TX : Lawrence Erlbaum Associates.

Thibaux, R., & Jordan, M. I. (2007). Hierarchical beta processes and the indian buffet process. *J. Machine Learning and Research*, *22*(3), 25-31.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, *113*(4), 766–786.