Computer Vision and Image Understanding 151 (2016) 138–152

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Modeling eye movement patterns to characterize perceptual skill in image-based diagnostic reasoning processes

Rui Li*, Pengcheng Shi, Jeff Pelz, Cecilia O. Alm, Anne R. Haake

Golisano College of Computing and Information Science, Rochester Institute of Technology, 1 Lomb Memorial Drive Rochester, NY 14623, USA

ARTICLE INFO

Article history:
Received 24 February 2015
Accepted 1 March 2016

Keywords:
Visual attention
Multi-modal data
Diagnostic reasoning
Probabilistic modeling
Nonparametric Bayesian method
Markov chain Monte Carlo
Combinatorial stochastic processes

ABSTRACT

Experts have a remarkable capability of locating, perceptually organizing, identifying, and categorizing objects in images specific to their domains of expertise. In this article, we present a hierarchical probabilistic framework to discover the stereotypical and idiosyncratic viewing behaviors exhibited with expertise-specific groups. Through these patterned eye movement behaviors we are able to elicit the domain-specific knowledge and perceptual skills from the subjects whose eye movements are recorded during diagnostic reasoning processes on medical images. Analyzing experts' eye movement patterns provides us insight into cognitive strategies exploited to solve complex perceptual reasoning tasks. An experiment was conducted to collect both eye movement and verbal narrative data from three groups of subjects with different levels or no medical training (eleven board-certified dermatologists, four dermatologists in training and thirteen undergraduates) while they were examining and describing 50 photographic dermatological images. We use a hidden Markov model to describe each subject's eye movement sequence combined with hierarchical stochastic processes to capture and differentiate the discovered eye movement patterns shared by multiple subjects within and among the three groups. Independent experts' annotations of diagnostic conceptual units of thought in the transcribed verbal narratives are time-aligned with discovered eye movement patterns to help interpret the patterns' meanings. By mapping eye movement patterns to thought units, we uncover the relationships between visual and linguistic elements of their reasoning and perceptual processes, and show the manner in which these subjects varied their behaviors while parsing the images. We also show that inferred eye movement patterns characterize groups of similar temporal and spatial properties, and specify a subset of distinctive eye movement patterns which are commonly exhibited across multiple images. Based on the combinations of the occurrences of these eye movement patterns, we are able to categorize the images from the perspective of experts' viewing strategies in a novel way. In each category, images share similar lesion distributions and configurations. Our results show that modeling with multi-modal data, representative of physicians' diagnostic viewing behaviors and thought processes, is feasible and informative to gain insights into physicians' cognitive strategies, as well as medical image understanding.

1. Introduction

Solely behavioral variables from task manipulations, such as response time or accuracy, are insufficient to determine whether a particular cognitive process is engaged or whether a particular cognitive architecture theory is correct. Since visual attention, as a selective dynamic cognitive process, is dominated by knowledge, interest, and expectations of the scene [7,23], it is possible to acquire insight into some aspects of subjects' interests or cognitive strategies by analyzing their eye movement sequences while they are pursuing certain tasks in domains of expertise where perceptual skills are paramount. One key step to manifest perceptual skill and uncover underlying cognitive processes is to discover expertise-specific perceptual viewing behaviors and differentiate the stereotypical and idiosyncratic behavioral patterns that characterize a group of subjects at the same training level. Addressing this problem requires segmenting an eye movement sequence into a set of time intervals that have a useful interpretation, as well as summarizing the commonality of eye movement patterns shared within and between expertise-specific groups. Furthermore, these meaningful patterns enable us to uncover time-evolving properties of subjects' perceptual reasoning processes and to understand images at a domain-knowledge level.

* Corresponding author.
E-mail addresses: lr8032@gmail.com, rxlics@rit.edu (R. Li), pengcheng.shi@rit.edu (P. Shi), jeff.pelz@rit.edu (J. Pelz), coagla@rit.edu (C.O. Alm), anne.haake@rit.edu (A.R. Haake).

http://dx.doi.org/10.1016/j.cviu.2016.03.001
1077-3142/

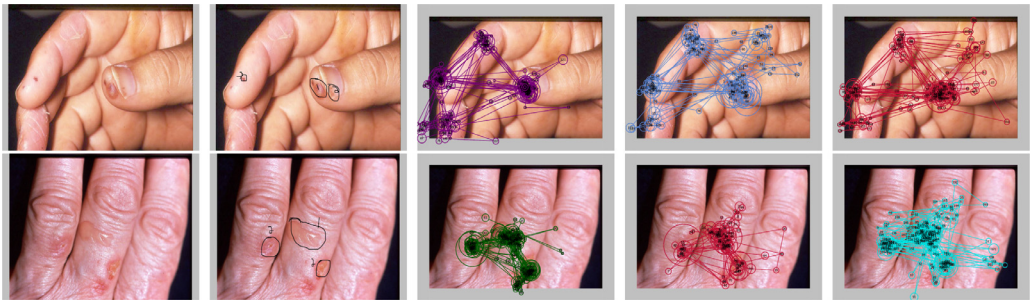


Fig. 1. Two example dermatological images examined by the subjects. The images from left to right are the original images, the primary and secondary abnormalities marked and numbered by an experienced dermatologist and three subjects' complete eye movement sequences acquired during the inspection process super-imposed onto the image, respectively. To visualize eye movement sequences, each circle center represents a fixation location and the radius is proportional to the duration time on that particular fixation. A line connecting two fixations represents a saccade.

Perceptual skill is considered to be the crucial cognitive factor accounting for the advantage of highly trained experts [24]. Experts generate distinctively different perceptual representations when they view the same scene as novices [38,45]. 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 or no expertise, significant differences in visual search strategies between groups show that expertise plays a key role in medical image examination. In such knowledge-rich domains, perceptual expertise is particularly valuable but poses challenges to its extraction, representation and application. Analyzing medical image understanding via traditional knowledge acquisition methods such as experts' marking on images (as shown in Fig. 1), verbal reports, and annotations is not only labor intensive but also ineffective because of the variability and noise of experts' performance [21]. In contrast, experts' perceptual skill is a valuable yet effortless resource worth exploiting, particularly for training and designing decision support systems where knowledge regarding the basic diagnostic strategies and principles of diagnostic-reasoning are desired [9]. We propose that this subconscious knowledge can be acquired by extracting and representing experts' perceptual skill in a form that is ready to be applied.

In this article we describe human-centered experimental approaches, which actively engage humans in the experimental process, to observe and record their perceptual and conceptual processing while inspecting medical images such as in Fig. 1. We subsequently profile the shared time-evolving eye movement patterns among physicians through our novel computational model, and also time-align eye movement patterns with semantic labels annotated by independent experts based on other dermatologists' verbal descriptions. We are then able to integrate these multimodal data towards understanding diagnostic reasoning processes and the dermatological images as well.

1.1. Visual attention

Attention is a critical contribution to perception in that focus of attention determines the portion of the sensory input from the external environment that will be readily available to perceptual processes. Complex visual information available in real-world scenes or stimuli exceeds the processing capability of the human visual system. Consequently, human vision is an active dynamic process in which the viewer seeks out specific information to support ongoing cognitive and behavioral activity [23]. Since high 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.

A series of fixations and saccades are used to describe such eye movements. Fixations occur when the gaze is held at a particular location, whereas saccades are rapid eye movements used to reposition the fovea to a new location. Both the number of fixations and their durations are commonly assumed to indicate the depth of information processing associated with the visual fields. Saccade amplitudes, which are rarely considered in the analysis of eye movement data, may also influence some conclusions drawn from the visual processing [7,39,51].

Studies have shown that visual attention is influenced by two main sources of input: bottom-up visual attention driven by low-level saliency features which are image properties that are distinctively different from their surroundings [27], and top-down cognitive processes, guided by the viewing task and scene context, influence visual attention [8,32,48]. Growing evidence suggests that top-down information dominates the active viewing process and the influence of low-level salience guidance is minimal [7]. It is acknowledged that covert visual attention can be dissociated from overt visual attention manifested by eye movements [22]. Nonetheless, studies have shown that overt and covert visual attention are tightly coupled in complex information processing tasks, such as reading and scene perception [40]. In particular, saccades which direct gaze to a new location usually follow a shift of covert attention to this location, leading to speculation that covert attention serves to plan saccades [26]. These theoretical findings provide us with the support to pursue the underlying engaged cognitive processing based on observed eye movements.

The concept of the saliency map [27] is based on the Feature Integration Theory [49]. A saliency map characterizes the bottom-up distinctiveness of a particular location relative to that of other locations in the scene through its conspicuousness. One derived computational model concerned with understanding people's visual attention deployments on natural images was developed [26]. The researchers built a computational model to evaluate the saliency level of an image based only on extracted low-level visual features such as intensity, color, and orientation. According to the computed saliency map, they attempted to predict people's visual attention allocation. The model has been tested over various image sets, and its performance is generally robust. Particularly in regards to man-made images, its performance is consistent with observations in humans. More recent research has moved beyond using only low-level visual features to compute the salient image areas, and has begun to investigate multiple cognitive factors that influence visual attention. The main additional factors include one's expectations about where to find information and one's current information need, as well [25]. To further formulate these cognitive factors, image saliency was redefined in terms of the combination of both

top-down and bottom-up cognitive influence and computed to predict users' viewing behaviors from the perspective of probability theory [37,54], and users were found to adapt their visual search in order to optimize the expected information gain [50]. The above series of modeling studies attempt to test visual attention theories by modeling the engaged cognitive factors. However, in specific domains requiring expertise, without guidance of perceptual skill and domain knowledge, scenes cannot be interpreted effectively purely based on factors such as visual features or goals. This motivates us to investigate how to formalize perceptual skill and reason about semantic meanings of image contents from observed experts' viewing behaviors and task performance.

1.2. Perceptual skill

Perceptual skill has been studied across various domains where it is profoundly exploited such as watching soccer games [45], playing chess [5], analyzing geo-spatial images [29], airport security screening [35], and examining photographical materials in clinical diagnosis [28,33]. Empirical perceptual studies of medical image-based diagnosis suggest that subjects vary their eye movement behaviors while they proceed in diagnosis on medical images. Furthermore, 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 that human expertise plays a great role in medical image examination. One related study investigated the nature of expert performance of four observer groups with different levels of expertise [33]. They compared multiple eye movement measures and suggested that these distinctive variations and better performance of the higher expertise group are due to the consequences of experience and training. Eye movement studies were conducted on diagnostic pathology by light microscopy to identify typical viewing behaviors for three expertise levels: pathologists, residents, and medical students [28]. Their results suggest that eye movement monitoring could serve as a basis for the creation of innovative pathology training routines.

Although capturing perceptual skills is challenging, comprehension of the cognitive basis could benefit a wide range of research areas in medical informatics such as medical image retrieval, proactive human-computer interaction, and domain training. We approach this challenge by working closely with medical specialists (dermatologists) using human-centered experimental approaches to observe and record their overt perceptual and conceptual processing while inspecting medical images towards diagnosis. The inherent dynamic property and complexity of experts' diagnostic reasoning motivates our investigation into the temporal dynamics of this perceptual-conceptual-interleaving process.

Previous studies fill the gap between physicians' interpretation and the statistics of pixel values by experts' manual annotation on segmented images and mapping into a domain knowledge ontology so as to perform medical image analysis at a semantic level [2,53]. However, there is great inter-variability between experts and intra-variability with which a single expert's performance changes from time to time also hinders this approach [21]. Moreover experts' perception, as tacit knowledge, functions below the level of consciousness. The eye tracking technique allows researchers to study experts' subconscious image viewing behaviors by objectively measuring eye movements and is a promising way to address these challenges. Recently, more and more studies have tried to incorporate human perceptual skills into image understanding approaches, treating eye movements as a static process by directly mapping eye movement data into the image feature space or by weighting image segments. However, the fact that meaningful perceptual patterns sometimes exist only over time and that the observed eye movement data are noisy and inconsistent undermine the reliability and robustness of these methods. In particular, inferring latent patterns underlying these observable human behaviors is a critical intermediate step in terms of advancing image understanding. One of the important contributions of our work is that we computationally discover and capture the spatial-temporal patterns in eye movement data.

1.3. Metrics of visual behaviors

Differentiation between stereotypical and idiosyncratic visual behaviors is considered a key aspect to investigate perceptual expertise using eye tracking data in domain-specific tasks [9]. This requires similarity measures to compare and evaluate visual behavior patterns between different observers. To capture medical specialists' (dermatologists') stereotypical and idiosyncratic visual behaviors from their eye movements, there has been significant progress in developing metrics for comparing and evaluating large amount of fixation and saccadic eye movement data represented as scanpaths [12,13,20,44,52]. To compare two eye movement sequences, normally the distances between their fixations are calculated. These methods can be broadly categorized into two classes.

One class of these algorithms is based on predefined Areas Of Interest (AOIs) [52]. A temporal sequence of AOIs is defined based on either dividing a scene into equally spaced bins or segmenting semantically meaningful regions in the scene. Then string-edit algorithms can be used to compare different sequences. These algorithms calculate the distance between two strings as the minimum number of edits required to transform one into the other. However, there are some issues: human intervention is still needed with respect to defining AOIs or specifying the size of the square regions and their locations; fixation durations are not taken into account; and string editing comparison among multiple scanpaths fail to measure meaningful variations between scanpaths.

The other analysis methods are based on clustering algorithms [18,44]. Clusters of fixation points are first grouped via parametric or non-parametric clustering algorithms based on their relative locations. After these clusters are labeled, pairwise comparisons can be conducted through various string editing methods. However, the clusters are not always meaningful, and fixation durations or saccade information is still not taken into account.

To compensate the above limitations, the Earth Mover's Distance (EMD) metric was proposed to measure the similarity of different visual behavior sequences in a pairwise way [10]. The similarity between eye movement sequences are viewed as a transportation problem by defining one sequence as a set of piles of earth and another sequence as a collection of holes and by setting the cost for a pile-hole pair to equal the ground distance between fixation in the two sequences. EMD thus compare eye movement sequences in a pair-wise manner, and the thresholds have to be heuristically specified. In contrast to EMD, our approach summarizes the similarity of multiple sequences simultaneously as well as of multiple related but distinct groups. Moreover, unlike EMD method using a number to characterize the similarity of whole sequences, we are able to cluster the similar sequence segments into consistent patterns based on their statistical properties.

1.4. Dynamic modeling approaches

Some studies adopted HMMs to profile subjects' perceptual processing based on their eye movements [1,34,41,42]. The disadvantage of these approaches is that they either have to heuristically predefine the number of hidden states or use standard parametric model selection methods to identify a "best" single number, the strengths and weaknesses of which in this problem setting is unknown. Two alternatives to HMMs are AOI-based or

clustering-based methods mentioned above. Although comprehensive eye movement features are taken into account in recent studies [11], current pair-wise comparison algorithms among multiple scanpaths are sensitive to data noise and minor variations between scanpaths. Furthermore, meaningful patterns may only exist over the time of whole processes, rather than comparing them piece by piece. This suggests a Markovian framework in which the model transitions among eye movement patterns that are associated with perceptual expertise and domain knowledge.

Recently there has been significant interest in augmenting dynamic systems' capabilities of modeling time series by combining stochastic processes. The hierarchical Dirichlet process (HDP) based HMMs allow the number of hidden states to be learned from observations by treating transition distributions as realizations of the HDP over countably infinite state spaces [3,14,46]. The infinite factorial HMM models a single time-series with emissions dependent on a feature with potentially infinite dimensionality which evolves with independent Markov processes [17]. Beta process (BP) based HMMs model multiple time series and capture an infinite number of potential dynamical modes which are shared among the series using the Indian buffet process (IBP) by integrating over the latent BP [15]. However, these approaches lack the capability of modeling multiple related but distinct groups of time series. This modeling requirement in our problem scenario motivates us to develop a novel hierarchically-structured dynamic model which is capable of profiling stereotypical and idiosyncratic patterns from multiple expertise-specific groups of eye movement sequences.

2. Experiment

We designed and conducted an eye tracking experiment to collect multi-modal data to investigate the conceptual and perceptual processing involved in subjects' medical image inspection [30]. These data were used to evaluate the new model.

2.1. Subjects

Subjects recruited for the eye tracking experiment belong to three groups based on their dermatology training level including eleven board-certified dermatologists (attending physicians), four dermatologists in training (residents) and thirteen undergraduate students who were lay people (novices). We also recruited physician assistant students who served as "trainees" in order to motivate dermatologists to verbalize their diagnostic reasoning using a modified Master-Apprentice scenario [4], which is known to be effective for eliciting tacit knowledge.

2.2. Apparatus

An 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 with reported accuracy of 0.5° 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.

2.3. Materials and procedure

A set of 50 dermatological images, each representing a different diagnosis, was selected for the study. These images were collected from the database of Logical Images Inc. and our collaborating author Cara Calvilli MD. These images were presented to subjects on a monitor. Viewing time limit on each image is 90 s. The subjects were instructed not only to view the medical images and make a diagnosis, but also to describe what they saw as well as their thought processes leading them to the diagnosis to the student sitting beside them as if they were in a training process. Both eye movements and verbal narratives 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. The audio recordings of the verbal narratives from the dermatologists were transcribed and annotated, as described below in Section 3.2.

3. Multi-modal data analysis

The fixations and the saccades were detected and identified using the position-variance method [43]. The dispersion threshold was set to 1.5 degree of visual angel which corresponds to about 44 pixels, given the subjects' distance and the monitor resolution. The minimum duration threshold was set to 80 milliseconds. Our statistical analysis of the eye movement events uncovered the attentive behavior difference between the three groups.

3.1. Statistical properties of eye movement data

Analysis of both fixation duration and saccade amplitude were conducted as a function of ordinal fixation number for the three expertise-specific groups to determine whether the two eye movement events, which are used as eye movement observation features, change over the time course of diagnosis and whether the differences as a function of expertise levels might be revealed at ordinal time points as shown in Fig. 2.

The first 20 fixations show a significant monotonically increasing trend for all three groups based on an ANOVA analysis ($F(19, 200) = 1.4,\ p < 0.01$; $F(19, 60) = 2.92,\ p < 0.01$ and $F(19, 240) = 1.98,\ p < 0.01$ respectively) and both attendings and residents have significantly longer averaged fixation durations than lay people ($F(2, 273) = 12.5,\ p < 0.001$), which is presented through the histogram of fixation duration distribution as shown in Fig. 2. Similar analysis on saccade amplitudes of the three expertise-specific groups shows that the first 20 saccade amplitudes of both dermatologists and residents follows a significant monotonically decreasing trend ($F(19, 200) = 1.24, p < 0.01$; and $F(19, 60) = 1.19,\ p < 0.01$). There was no effect for the lay people' average saccade amplitudes. The shorter fixation durations and longer saccade amplitudes at the initial stage suggests that both attending and residents started examining images with a quick image scan. After that, fixation duration became longer and saccade amplitudes decreased, suggesting a more thorough examination. In contrast, lay peoples' fixation durations increased at the initial stage but there is no statistically significant change for their saccades.

In sum, the above descriptive statistical analysis indicates that the expertise-level difference can be manifested via perceptual viewing behaviors, which is consistent with previous studies [28,33]. We then apply our model on these time series data to reveal the subtlety of the behavior patterns varying over time.

3.2. Annotation analysis on transcribed verbal descriptions

An annotation study was conducted on the experts' transcripts to investigate the verbalized cognitive processes of dermatologists on their paths toward a diagnosis [36]. After transcribing the experts' narration of the images, independent experts identified conceptual units of thought (corresponding to particular steps or information in the diagnostic process) in the transcripts. These *thought units* were subsequently time-aligned with the recorded speech and eye movement patterns in the speech analysis tool Praat [6]. Two highly trained dermatologists annotated transcribed

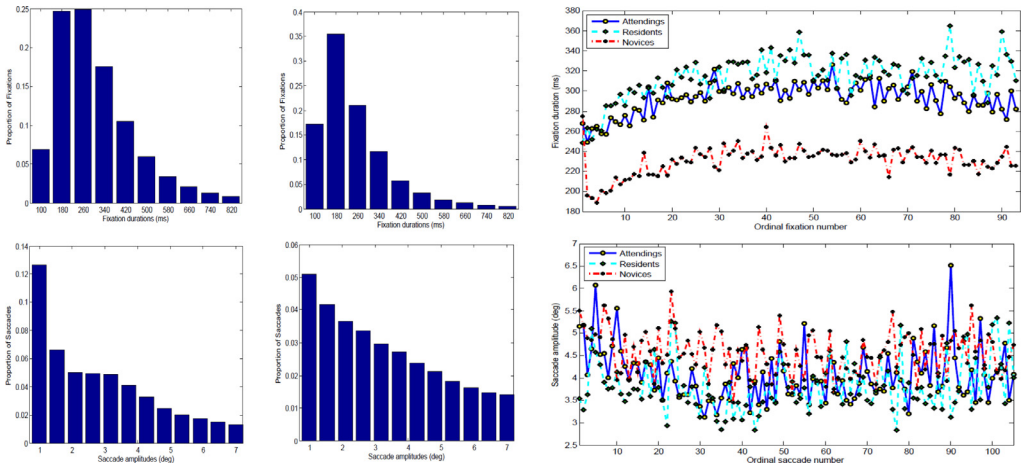


Fig. 2. Analysis of eye movement data. On the top left are two histograms of the fixation duration (msec) distribution for 15 physicians (attendings and residents), and 13 lay persons, respectively. On the top right are the average fixation durations by ordinal fixation number over the course of diagnosis of all 50 images for the three expertise-specific groups: attendings (blue), residents (cyan) and lay persons (red). On the lower left are the histograms of the saccade amplitude (degree) distribution for experts and lay people. The lower right graph shows the average saccade amplitude by ordinal fixation number for three expertise-specific groups over the course of diagnosis with the same color coding. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

verbal descriptions with these 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. Nine basic thought units, provided by a dermatologist, were used for annotation. The provided 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'.

The conceptual thought unit annotations were then linked to the model's inferred perceptual eye movement patterns, as discussed below in Section 4.

4. The computational modeling approach

The modeling approach of the expertise-specific groups' eye movements for the dermatological images is diagrammed in Fig. 3.

In Fig. 3(a), the hierarchy represents the heterogeneous structure produced by individuals with different expertise levels examining medical image viewing strategies. A group of subjects with the same expertise level share a set of behavior patterns based on their knowledge. In accordance with these common behavior patterns, each group member's time-evolving behaviors also display their individualized temporal patterns in terms of unique subsets of behaviors and/or their unique sequential combinations. At the lowest level, each behavior is measured based on observed eye movements. Fig. 3(b) shows the graphical representation of the hierarchical dynamic model corresponding to (a)'s structure. B_0 is the global base measure on the space of all possible behaviors Θ. The common behavior pattern of the group defined as $\{(\theta_k, E_k)\}$ is characterized by the shared behaviors among p group members and the probabilities that it possesses each particular behavior is encoded by B_0. A group member p performs individualized behavior pattern defined as$\{(\theta_k, S_{pk})\}$ which is a Bernoulli process realization of the group common pattern $\{(\theta_k, E_k)\}$. The transition matrix π_p follows a Dirichlet distribution specified by the non-zero entries of S_p.

4.1. Dynamical likelihoods

Autoregressive-HMMs has been proposed to be a simpler but often effective way to describe dynamical systems [16]. Let $y_t^{(ij)}$ denote the eye movement data of the ith subject at time step t in the jth group. We associate each time-step's observation with one fixation and its successive saccade as one observation unit. Let $x_t^{(ij)}$ denote the corresponding latent dynamic mode. We have

$$x_t^{(ij)} \sim \pi_{x_{t-1}^{(ij)}} \tag{1}$$

$$y_t^{(ij)} = A_{x_t^{(ij)}} \tilde{y}_t^{(ij)} + e_t(x_t^{(ij)}) \tag{2}$$

where $e_t^{(ij)}(k) \sim N(0, \Sigma_k)$ which is an additive white noise, $A_k = [A_{1,k}, ..., A_{r,k}]$ as the set of lag matrices, and $\tilde{y}_t^{(ij)} = [y_{t-1}^{(ij)}, ..., y_{t-r}^{(ij)}]$. In our case, we specify $r = 1$.We thus define $\theta_k = (A_k, \Sigma_k)$ as one eye movement pattern.

4.2. Hierarchical prior

The hierarchical beta-Bernoulli processes proposed by Thibaux et al. [47] is a suitable tool to describe the situation where multiple groups of subjects are defined by countably infinite shared features following the Levy measure. We utilize this process in the following specification based on our problem scenario.

Let B_0 denote a fixed continuous random base measure on a measurable space $\Theta = \{\theta_k\}$ which represents a library of all the potential eye movements patterns. To characterize patterns shared among multiple groups, 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 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 sequence of subject i within group j. This hierarchical construction can be formulated as follow:

$$B|B_0 \sim BP(c_0, B_0) \tag{3}$$

$$G_j|B \sim BP(c_j, B) \qquad j = 1, ..., N \tag{4}$$

$$P_{ij}|G_j \sim BeP(G_j) \qquad i = 1, ..., N_j \tag{5}$$

where $B = \sum_k b_k \delta_{\theta_k}$ with $\{\theta_k\}$ drawn from the library Θ and coupled with their weights b_k. b_k is beta-distributed given b_0 and c_0. Furthermore, $G_j = \sum_k g_{jk} \delta_{\theta_{jk}}$ shows that G_j is associated with

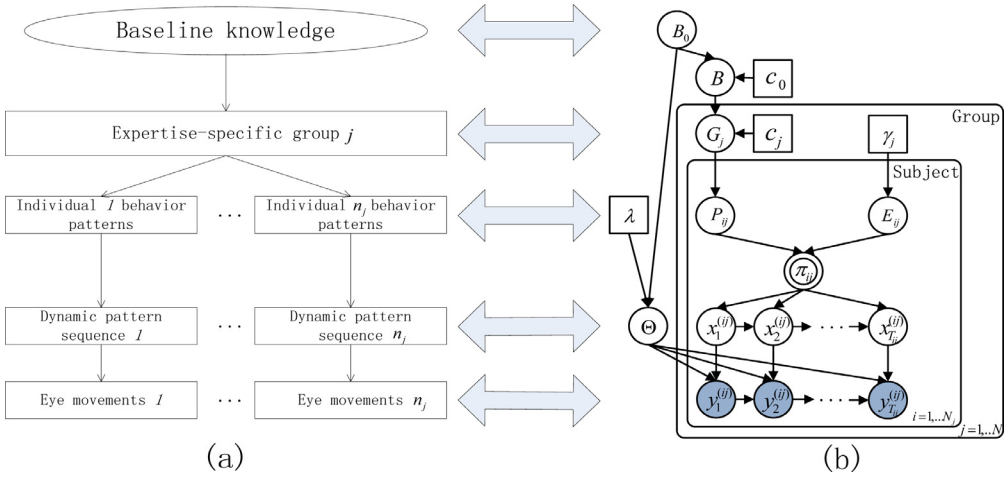


Fig. 3. The hierarchical dynamic model represented by graphical model schemes. The detailed description is in Section 4.

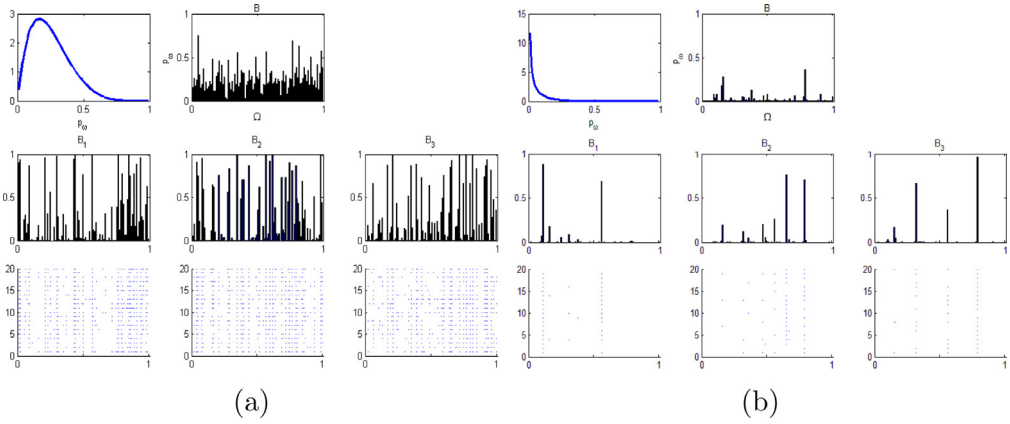


Fig. 4. Realizations from a hierarchal beta process with n=3 and $n_1 = n_2 = n_3 = 20$. We vary the concentration parameters: c_1, c_2 and c_3, and the base measure parameters: a_0 and b_0. In (a), the parameters are $c_1 = c_2 = c_3 = 1$ and $a_0 = 2$, $b_0 = 6$. In (b), the parameters are $c_1 = c_2 = c_3 = 1$ and $a_0 = 2$, $b_0 = 0.6$.

both $\{\theta_{jk}\}$ which is a subset of countable number of eye movement patterns drawn from $\{\theta_k\}$ and their corresponding probability masses $\{g_{jk}\}$ given group j. $\{g_{jk}\}$ is also beta-distributed given b_k and c_j. The combination of these two variables characterizes how the 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:

$$P_{ij} = \sum_k p_{ijk}\delta_{\theta_{jk}} \tag{6}$$

where p_{ijk} as a binary variable denotes whether subject i within group j exhibits eye movement pattern k given probability mass g_{jk}.

Based on the above formulation, for $k = 1, \ldots, K_j$ patterns $\{(\theta_{jk}, g_{jk})\}$ characterize how a set of common eye movement patterns likely shared among group j and $\{(\theta_{jk}, p_{ijk})\}$ represent subject i's personal subset of eye movement patterns given group j. According to the above equations, we illustrate two sets of hierarchical beta process each of which contains three groups of beta-Bernoulli processes from a common beta process with specified parameters in Fig. 4. This illustration highlights the effect of the concentration parameters for c_{1-3} and mass parameters for (a_0, b_0).

The transition distribution $\pi_{ij} = \{\pi_{x_t^{(ij)}}\}$ of the auto-HMMs at the bottom level governs the transitions between the ith 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 variables $\{e_{ijk}\}$:

$$e_{ijk}|\gamma_j \sim Gamma(\gamma_j, 1) \tag{7}$$

$$\pi_{ij} \propto E_{ij} \bigotimes P_{ij} \tag{8}$$

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

4.3. Posterior inference with Gibbs sampler

We use the Gibbs sampler to do the posterior inference. 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 [47], we developed a Gibbs sampling solution to the hierarchical beta processes part of the model.

We adopt normal-inverse-Wishart distribution to provide an appropriate conjugate matrix prior to pattern space Θ. The conjugate prior on the set of design matrix A and the noise covariance Σ is the matrix normal-inverse-Wishart prior. This distribution places a conditionally matrix normal prior on A given Σ:

$$p(A|\Sigma, M, K) = \frac{|K|^{\frac{d}{2}}}{|2\pi\Sigma|} \exp\left\{-\frac{1}{2} tr((A-M)^T \Sigma^{-1}(A-M)K)\right\} \tag{9}$$

and an inverse-Wishart prior on Σ

$$\Sigma \sim \mathcal{W}(\nu, \Delta) \tag{10}$$

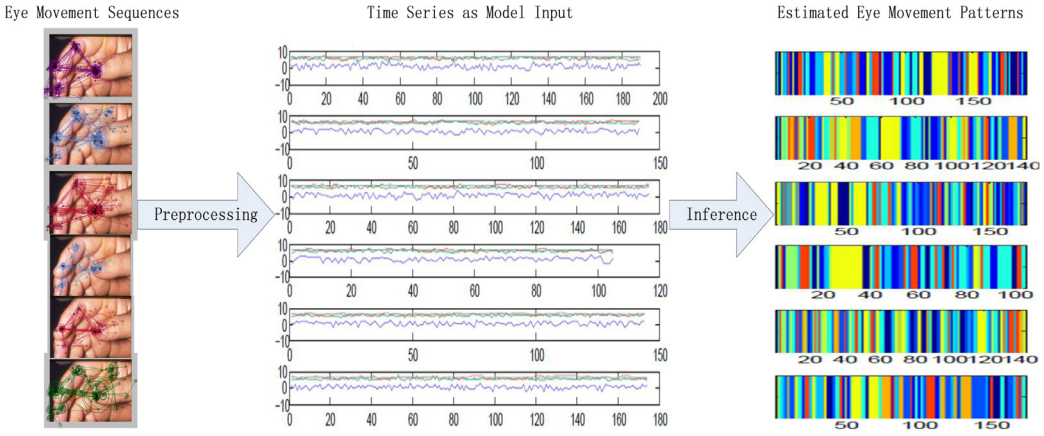


Fig. 5. Inference of the eye movement patterns. From left to right, six subjects' eye movement sequences are illustrated on a dermatological image. The sequences are represented as time series which is composed of 4 components: log values of fixation location (x-y coordinates), fixation duration and saccade amplitude. The model-derived eye movement pattern sequences for the corresponding time series are estimated with 4 chains of 55,000 sampling iterations. The color coding corresponds to the segments of each specific eye movement pattern.

Consider a set of observations $D = \{X, Y\}$, the posterior distribution of $\{A, \Sigma\}$ can be decomposed as the product of posterior A as $\mathcal{MN}(A; S_{yx}S_{xx}^{-1}, \Sigma, S_{xx})$ with $S_{xx} = XX^T + K$, $S_{yx} = YX^T + MK$, and $S_{yy} = YY^T + MKM^T$ and the marginal posterior of Σ as $\mathcal{W}(\nu + N, \Delta + S_{y|x})$ where $S_{y|x} = S_{yy} - S_{yx}S_{xx}^{-1}S_{yx}^T$.

When sampling the pattern indicator matrix P_j of group j, we need to address two situations separately. For a pattern which has non-zero probability because of either its priori or having already been instantiated by at least one subject, we compute its posterior as follows.

Let $\{\omega\}$ denote the atoms (eye movement patterns) that have been observed at least once. We define the variables to perform inference: $b_0 = B_0(\{\omega\})$, $b = B(\{\omega\}) = \sum_k b_k\delta_\omega$, $g_j = G_j(\{\omega\}) = \sum_k g_{jk}\delta_\omega$, and $p_{ij} = P_{ij}(\{\omega\}) = \sum_k p_{ijk}\delta_\omega$. According to Eqs. (3)–5, these variables from their respective processes have the following distributions:

$$B(\omega) \sim Beta(c_0B_0(\omega), c_0(1 - B_0(\omega))) \tag{11}$$

$$G_j(\omega) \sim Beta(c_jB(\omega), c_j(1 - B(\omega))) \tag{12}$$

$$P_{ij}(\omega) \sim Ber(G_j) \tag{13}$$

We marginalize out G using conjugacy. Let $m_j = \sum_{i=1}^{n_j} p_{ij}$, and use $\Gamma(x+1) = x\Gamma(x)$, the posterior distribution of b given P_j:

$$p(b|b_0, P) \propto p(b|b_0)\frac{\Gamma(m_j + c_jb)\Gamma(n_j - m_j + c_j(1-b))}{\Gamma(c_jb)\Gamma(c_j(1-b))} \tag{14}$$

This posterior is log-concave, which we can use adaptive rejection sampling method [19] to approximate in our Gibbs sampler. We can sample g_j from its conditional posterior distribution by conjugacy:

$$p(g_j|b, P) \propto Beta(c_jb + m_j, c_j(1-b) + n_j - m_j) \tag{15}$$

Given the ith subject's eye movements data sequence $y_{1:T_{ij}}^{(ij)}$ in the group j, transition variable E_{ij} and within-group-j shared pattern set $\theta_{1:K_j}$, the current sampling pattern indicator p_{ijk} of pattern k exhibited by subjects i in group j follows this posterior distribution:

$$\begin{aligned} &p(p_{ijk}|P^{(-ijk)}, y_{1:T_{ij}^{(ij)}}, \theta_{1:K_j}^{(-ijk)}, E_{ij}, B_0) \propto \\ &p(p_{ijk}|P^{(-ijk)}, B_0)p(y_{1:T_{ij}^{(ij)}}|P_{ij}, E_{ij}, \theta_{1:K_j}^{(-ijk)}) \end{aligned} \tag{16}$$

where $P^{(-ijk)}$ denotes the set of all P_{ij} except p_{ijk}. In particular, for the instantiated patterns

$$p(p_{ijk}|P^{(-ijk)}, B_0) = \int p(p_{ijk}|G_j)\int p(G_j|B, P)p(B|B_0, P)dBdG_j \tag{17}$$

Both $p(G_j|B, P)$ and $p(B|B_0, P)$ can be sampled as in Eq. (15) and Eq. (14), respectively.

For the yet-instantiated patterns of group j, since they can be directly sampled from the conjugate prior distribution of Θ, we only need to infer the distribution of their number the prior distribution of which is Poisson-distributed $K \sim Poi(\frac{c_0\lambda}{c_0+k-1})$. Given that all other patterns from all other groups are zero:

$$\begin{aligned} &p(k_{ij}|P_{ij}, y_{1:T_{ij}^{(ij)}}, \theta_{1:K_j}^{(-ijk)}, E_{ij}, \lambda) \\ &\quad \propto p(p_{ijk}|P^{(-ijk)}, \lambda)p(y_{1:T_{ij}^{(ij)}}|P_{ij}, E_{ij}, \theta_{1:K_j}^{(-ijk)}) \end{aligned} \tag{18}$$

Given transition distributions $\pi^{(i)}$, shared patterns $\{\theta_k\}$, and observations $y_{1:T_{ij}}$, within a massage passing algorithm, we compute the backward messages:

$$m_{t+1,t}(x_{t_{ij}}) \propto p(y_{t+1:T_{ij}}^{(ij)}|x_{t_{ij}}, \pi_{ij}, \{\theta_k\}) \tag{19}$$

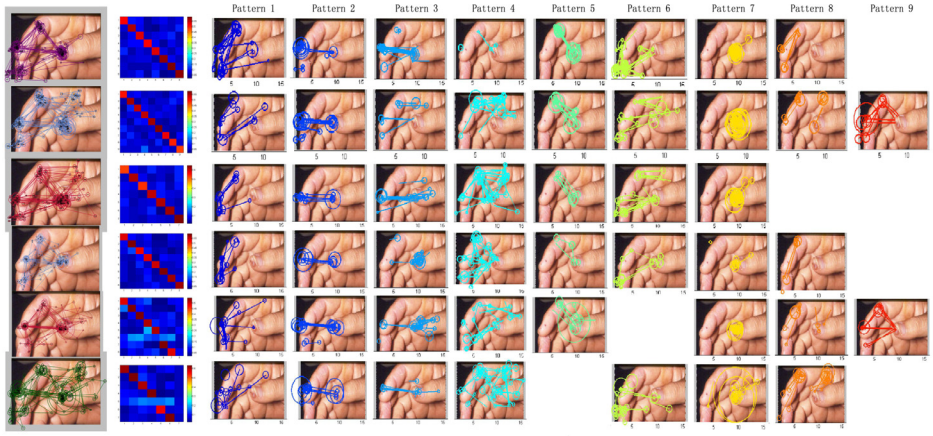
to update the hidden state sequences $x_{1:T_{ij}}^{(ij)}$ by sampling from:

$$\begin{aligned} &p(x_{t_{ij}}|x_{t_{ij}-1}, y_{1:T_{ij}}^{(ij)}, \pi_{ij}, \{\theta_k\}) \\ &\quad \propto \pi_{x_{t_{ij}-1}}(x_{t_{ij}})N(y_{t_{ij}}^{(ij)}; A_{x_{t_{ij}}}\tilde{y}_{t_{ij}}^{(ij)}, \Sigma_{x_{t_{ij}}})m_{t+1,t}(x_{t_{ij}}) \end{aligned} \tag{20}$$

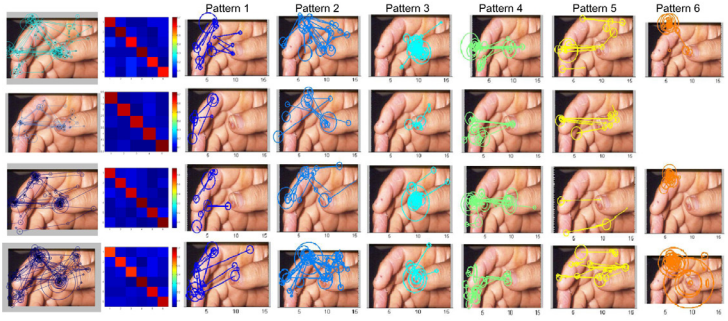
5. Interpretations of the eye movement patterns

In Fig. 5, the eye movement sequences are represented as four dimensional time series of fixation x-y coordinates, fixation duration, and saccade amplitude. The time series are the input of our model. For a given time series, the eye movement patterns are estimated by our model as a set of time series segments. the segments with the same color correspond to a specific eye movement pattern shared across the time series.

Furthermore, Figs. 6 and 7 visualize the discovered eye movement patterns from three expertise-specific groups on two dermatological images. The color-coding is consistent with in Fig. 5. We defined a fixation and its subsequent saccade as one fixation-saccade unit. Each eye movement pattern is visualized by a subset



(a) Nine inferred eye movement patterns from the attending group. The first column is the eye movement sequences. The second is the transition matrices indicating the patterns' persistency. The right are the visualized patterns. The color coding for each pattern is the same as Figure 5



(b) Six inferred eye movement patterns from the residents.

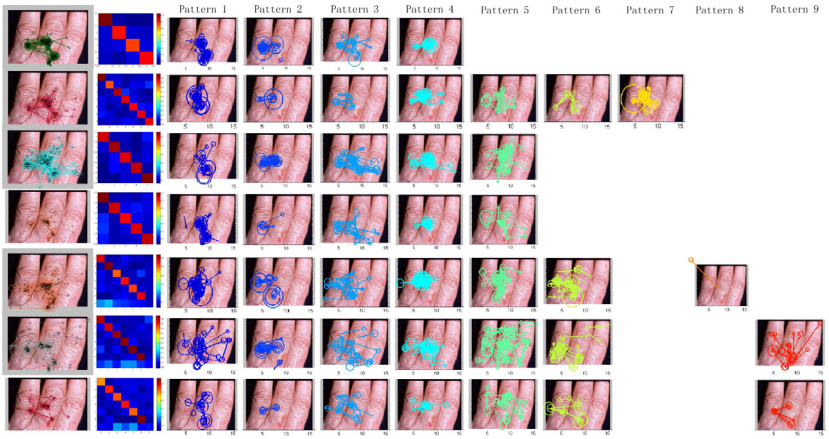
(c) Sixteen inferred eye movement patterns from the novices.

Fig. 6. The eye movement patterns of the three expertise-specific groups signify the different perceptual behaviors between the experts and the novices.

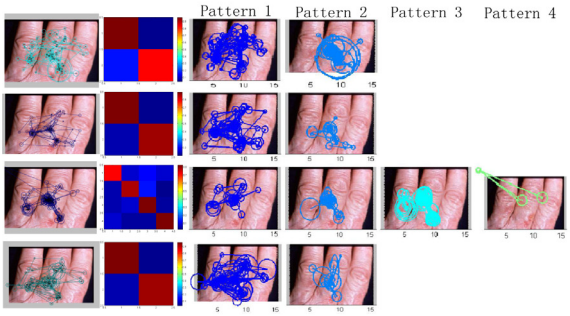
of fixation-saccade units which have similar fixation durations, fixation spatial coordinates, and saccade amplitudes. These two images are among the most difficult cases to make a correct diagnosis, and some of the patterns exhibited on them are critical to inform some properties of the images.

Taking the first illustrated image in Fig. 6 for example, there are multiple skin lesions spreading over the thumb nail and tip, the two parts of index finger and the middle finger. A primary abnormality is on the thumb tip. The eye movement sequences indicate that attending dermatologists fixated on the primary abnormality heavily and switched their visual attention actively between and within the primary and secondary findings. The same patterns are also exhibited in the resident dermatologist group. The reason for lacking other patterns is probably because the number of participants at this expertise-specific group is limited in the dataset (only four participants). In contrast, the novice group exhibits significantly different eye movement patterns compared to the other groups. According to the novices' patterns, we can see shorter saccades so as to leave long fixation durations at the center of the image as seen in Pattern 1 and 9 of Fig. 6(c) and do not exhibit the eye movement switching between primary and secondary abnormalities as dermatologists' Pattern 2, 3, and 5 in Fig. 6(a).

In a transition probability matrix, a row corresponds to a current eye movement pattern, and its cell values represent a discrete probability distribution over subsequent patterns given the current



(a) Nine inferred eye movement patterns from the attending group. We illustrate the visualized eye movement sequences, the transition matrices, and the color-coded patterns.



(b) Four inferred eye movement patterns from the residents.

(c) Ten inferred eye movement patterns from the novices.

Fig. 7. The inferred eye movement patterns of the three expertise-specific groups.

pattern. For example, in Fig. 6(a) the matrices have high values in diagonal, which means that each pattern is persistent (If the current pattern is A, then the subsequent pattern is most likely to be A). The more random transition matrices in Fig. 6(c) indicate that novice group's patterns are not persistent, which suggests that novices' focus of attention is unstable when viewing the image. We reason that these relatively unstable viewing behavior reflect that fact that the novice cannot perceive the important diagnostic relationships among the multiple abnormalities and fail to prioritize them. All the pattern differences between expertise-specific groups holds for the other images studied here.

Some shared patterns emerged in the attending and the resident groups but are lacking in the novice group as shown in Fig. 6(c). This suggests that experts, equipped with domain knowledge organized in finer gradations of functional categories [24], can discriminate the significance of their findings in a particular context. In contrast, in Fig. 6(c) the novices failed to do so, although their eye movement patterns indicate that they notice the same abnormalities too. When comparing the transition probability matrices between the expertise-specific groups in the second column of Fig. 6(a–c) and Fig. 7(a–c), it becomes clear that professionals' eye movement patterns are more persistent than the novices'.

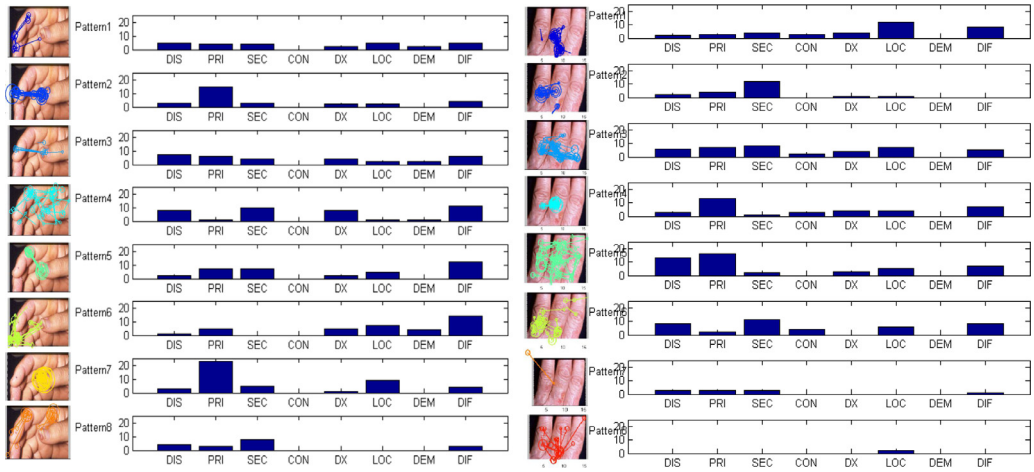


Fig. 8. Analysis of the correspondence between eye movement patterns and thought units for the two example images. Histograms show the relationship between discovered eye movement patterns and annotated thought units. For each pattern we plotted the counts of fixations which are labeled as the corresponding thought units. The pattern numbering is consistent with previous figures.

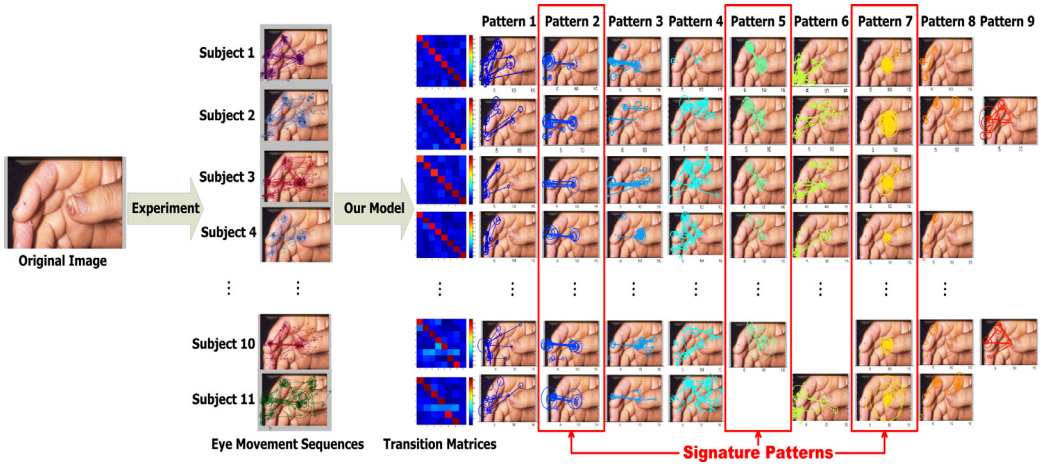


Fig. 9. Six out of eleven eye movement sequences super-imposed onto one dermatological image are illustrated here. Our model decomposes the eleven eye movement sequences on this image into nine eye movement patterns (color-coded) with the transition probability matrices. In this way, each eye movement sequence is represented by a certain number out of nine patterns and their corresponding transition matrix. On the right is the shared eye movement pattern matrix of which each row corresponds to a subject's eye movement sequence and each column indicates one shared eye movement pattern among multiple subjects. In this case, three patterns are recognized as *signature patterns* based on their self-transition probabilities, temporal-spatial properties, and diagnostic semantics.

To further analyze the meanings of the discovered eye movement patterns, we mapped thought units (see Section 3.2) to patterns discovered in the eye movement data in order to see whether they correspond consistently during the diagnostic process. Pattern occurrence and thought unit alignment resulted in assignment of each fixation in a complete eye movement sequence to a specific pattern and to a thought unit such as PRI or LOC (or None). Although thought units are often spread out across eye movement patterns, some trends can be discerned. Initial integration of eye movement patterns with thought units was accomplished by calculating the counts of their time-aligned correspondence in Fig. 8. Analysis on the left column diagram of Fig. 8 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. It is worth to point out that identification of the primary morphology is an early key diagnostic step which helps the physician to place the lesion in the correct category. Pattern 7 has relationship to location (LOC) which appears to correspond to the primary morphology location. Pattern 4 consists of eye movement sequence segments which are characterized by shorter fixation durations and longer saccades. This scanning behavior corresponds to the thought units, including distribution (DIS), secondary morphology (SEC), diagnosis (DX) and differential diagnosis (DIF). For example, the scanning pattern coupled with thought unit DX is possibly related to confirmation of secondary findings to support or rule out diagnostic hypotheses.

6. Image analysis based on signature patterns

Our model converges, after 5000 sampling iterations, to generate 387 eye movement patterns based on eleven subjects diagnosing fifty dermatological images. These results allow us to analyze and describe the dermatological images based on a novel perspective of experts' perceptual strategies.

6.1. Eye movement pattern estimation

Fig. 9 illustrates the eleven dermatologists diagnosing a case of a skin manifestation of endocarditis by showing one set of observed eye movement sequences and the model's discovered eye movement patterns shared by the dermatologists which correspond to descriptively meaningful perceptual units. 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. A primary abnormality is on the thumb tip. The eye movement sequences in Fig. 9 indicates that dermatologists examine the image in a highly patterned manner by fixating on the primary abnormality heavily and switching their visual attention actively between and within the primary and secondary abnormalities. Our model decomposes each eye movement sequence into several subsets of its segments. Each subset is characterized by one estimated latent state and a Gaussian emission distribution which summarizes the similar temporal-spatial properties shared among multiple sequences. The way that the patterns are shared among the subjects is also indicated by their matrix in Fig. 9. For example the first subject's eye movements evolve over time with the first eight out of nine patterns, and the eleventh subject has seven patterns except pattern 5 and pattern 9. In other words, most but not all patterns are shared by all physicians, as one would expect when modeling human behaviors where there almost certainly exist some variation and some individual differences. Again, our model is able to capture both the shared (stereotypical) behaviors and the individualized (idiosyncratic) ones. Transition probability matrices indicated these patterns are persistent with high self-transition probabilities which measure the likelihood of a given pattern transiting into itself in our dynamic model.

6.2. Signature pattern recognition

When expanding the analysis to multiple images, we discover several basic yet distinctive types of patterns shared across multiple images which we term as *signature patterns* with respect to the patterns' fixation duration and saccade amplitude.

We define a type of signature patterns by three criteria. First, its self-transition probability, which is indicated by the transition matrix, is no less than the median 0.65, so the signature patterns are stably retained by experts. Second, it manifests clear diagnostic regions, for example pattern 7 in Fig. 9 corresponds to a long fixation duration on the primary abnormality. Third, the temporal-spatial properties of signature pattern exemplars within each type are similar but distinctive from other types, which is elaborated in Fig. 10. The other discovered patterns are not identified as signature patterns because they lack one or more of the three criteria. In the illustrated case in Fig. 9, there are three patterns recognized as the signature patterns. Pattern 2 and Pattern 5 are characterized by fixations switching back and forth between the primary and the different secondary abnormalities with long saccade amplitudes and relatively short fixation durations. These patterns suggest that subjects compare and associate the two types of abnormalities. Pattern 7 is characterized by a series of long-duration fixations only on the primary abnormality with extremely short saccades. This pattern suggest that subjects fixate on the primary abnormality to make a diagnosis.

Based on the eye movement patterns generated from our model over fifty images, we are able to specify three types of signature patterns. The first type is named *Concentrating Pattern* which is characterized by a series of long-duration fixations and short-amplitude saccades usually fixating on primary abnormalities. The second is the *Switching Pattern* characterized by a series of relatively short-duration fixations and long-amplitude saccades usually switching back and forth between two abnormalities. And the third is *Clutter Pattern* characterized by a series of shorter fixations and relatively long saccades usually scanning within localized abnormal regions. To quantify the temporal-spatial properties of the three types of signature patterns, we illustrate some of their exemplars in Fig. 10.

The estimation of the signature patterns based on their exemplar features can be solved using different classification techniques. Since Gaussian mixture is one intuitively appropriate tool to describe the distributions of these signature patterns according to Fig. 10, we first adopt quadratic discrimination analysis (QDA) by assuming a simple parametric model for the densities of the temporal-spatial properties of the eye movement units. A training set includes 217 eye movement units of 12 exemplar patterns from 10 images, which are shown in Fig. 10. We test the validity of the classifier through comparing the image categorization performance based on QDA with K nearest neighbors (K-NN) and experts' performance.

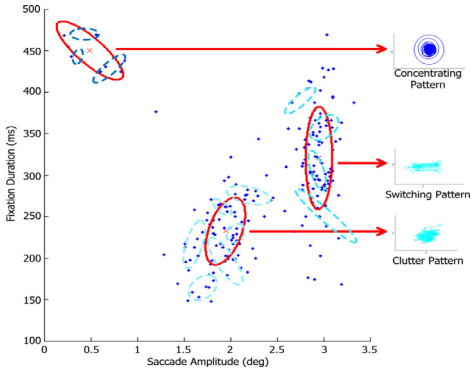


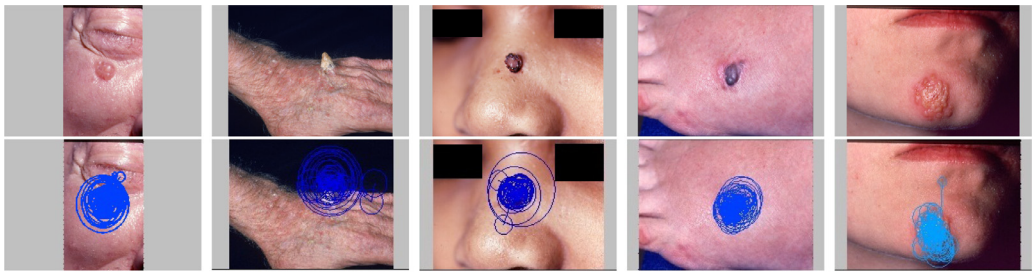
Fig. 10. Distinctive temporal-spatial properties of 217 fixation-saccade pairs from 12 exemplars forms the three types of signature patterns. Each blue dot represents one eye movement unit from a signature pattern exemplar. The exemplars are indicated by dash-line Gaussian emission distributions estimated from our model. Both eye movement units and their corresponding exemplars are projected from a four-dimension space (including x-y coordinate, fixation duration and saccade amplitude) onto this space. The signature patterns are characterized by a three-component Gaussian mixture. The one on the upper left represents *Concentrating Pattern*, the one on the right captures *Switching Pattern*, and the one on the lower middle represents *Clutter Pattern*. For each type, we project the units back into x-y coordinate space centered on the origin and visualize them on the right side of the main diagram. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

6.3. Perceptual category specification

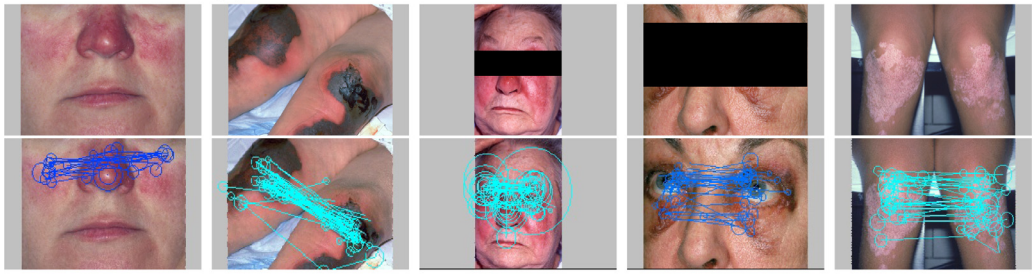
Based on our consulting dermatologist's suggestion, we propose four broad perceptual categories in terms of lesion distribution and configuration [31]. We further determine the associations between the combinations of the exhibitions of these three types of signature patterns and the four specified categories:

- If the set of eye movement patterns exhibited on an image only includes *Concentrating Patterns*, the image is categorized as *Localized* which means that the image contains a solitary lesion as primary abnormality.
- If the set of eye movement patterns exhibited on an image solely includes *Switching Patterns*, the image is categorized as *Symmetrical* which means that the lesions in the image are symmetrically distributed.
- If the set of eye movement patterns exhibited on an image includes both *Concentrating Patterns* and *Switching Patterns*, the image is categorized as *Multiple Morphologies* which means that the lesions in the image belong to different dermatological morphologies and usually one lesion is identified as primary abnormalities and the other are secondary ones.
- If the set of eye movement patterns exhibited on an image includes *Clutter Patterns*, the image is categorized as *High-Density Lesions* which means that the image contains multiple lesions distributed in either scattered or clustered manner.

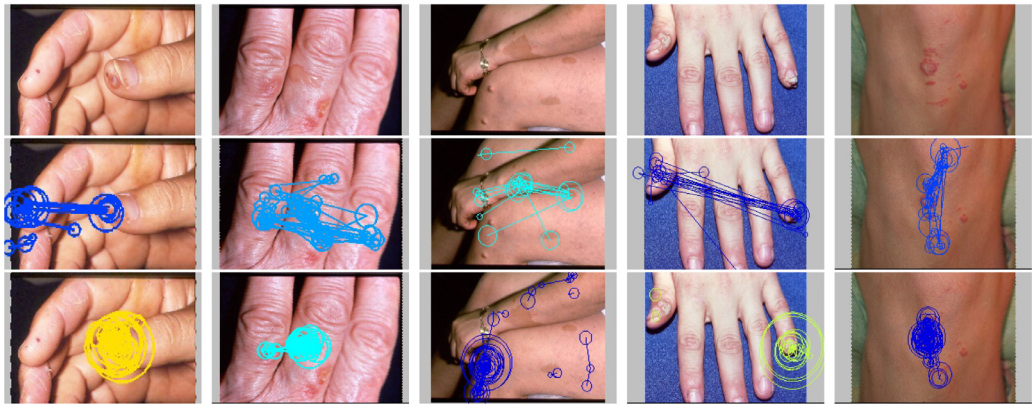
According to the signature patterns recognized on the images, we can catalogue the images into the four categories as shown in Fig. 11(a)–(d).



(a) *Localized* lesions with *Concentrating Pattern*.



(b) *Symmetrical* distribution with the *Switching Pattern*.



(c) *Multiple Morphologies* with the *Switching* and the *Concentrating Patterns*.

(d) *High-density Lesions* with the *Clutter Pattern*.

Fig. 11. For each category five example images are illustrated with the signature patterns recognized.

The difference between *Multiple Morphologies* images and *Symmetrical* images is that the eye movement patterns exhibited on the latter do not contain *Concentrating Pattern*. This is because the symmetrical visual-spatial structures imply that lesions are equivalent important without single primary one for the subjects to concentrate their focus on as shown in Fig. 11. Since the specifications of signature patterns are determined heuristically, we may be able to improve the categorization performance by identifying additional meaningful and distinctive eye movement patterns, and these extra patterns may also lead to image categorization at a finer detailed level.

7. Results and discussion

To measure the performance of our image categorization approach, we conduct an experiment following the same procedure by recruiting another ten dermatologists and using a different set of forty dermatological images as stimuli. These images are also randomly selected from a dermatological image database. Our three consulting dermatologists achieve consensus to categorize the forty images into the four perceptual categories. We use 232 estimated eye movement patterns on these images and the ones from the previous experiment as a testing set. In Fig. 12, we

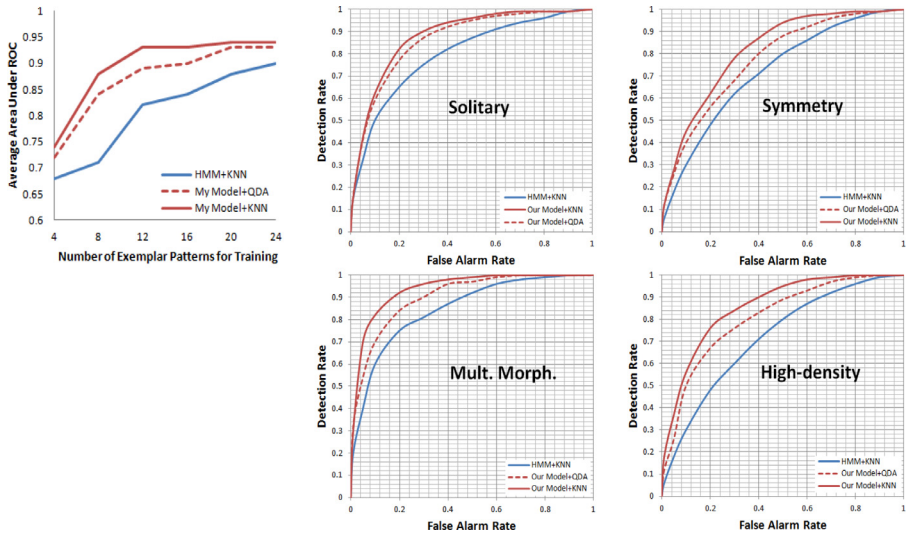


Fig. 12. ROC curves summarizing categorization performance for the four perceptual categories. Left: Area under average ROC curves for different numbers of exemplar patterns. Right: We compare our model using two different classification techniques with canonical Hidden Markov Models.

examine categorization performance given training sets containing between 4 and 24 exemplars. We assume each eye movement sequence exhibits the same set of patterns in order to implement the canonical HMMs. We see that our model lead to significant improvements in categorization performance, particularly when few training exemplars are available. The highest accuracy is achieved on detection of the "Multiple Morphologies" category. This may be caused by the requirement of detections of the two different signature patterns to determine the varied distributions and significance of the lesions. The difference between "Multiple Morphologies" images and "Symmetry" images is that the eye movement patterns exhibited on the latter do not contain "Concentrating Pattern". This is because the symmetrical visual-spatial structures imply that lesions are equivalent important without single primary one for the subjects to concentrate their focus on as shown in Fig. 11b. Since the specifications of signature patterns are heuristic, we may be able to improve the categorization performance by identifying extra meaningful and distinctive eye movement patterns, and these extra patterns may also lead to image categorization at a finer detailed level.

We obtain certain aspects of experts' domain-specific knowledge by summarizing their perceptual skills from their eye movements while diagnosing images. The domain-specific knowledge unveils the meaning and significance of the visual cues as well as the relations among functionally integral visual cues without segmentation or processing of individual objects or regions. This will benefit the traditional pixel-based statistical methods for image understanding by evaluating perceptual meanings and relations of the image features which spatially correspond to the eye movement patterns. This combination of expert knowledge and image features allows us to generalize our approach to images on which there is no experts' eye movements recorded.

The dermatological images are taken and collected by dermatologists for diagnosis and training purpose. Since the photographers tend to center primary abnormalities and preserve related contextual information such as patients' demographic information, body parts, lesion size and so on, these high-resolution images have complex backgrounds and large variations in luminance and camera angles. These factors cause some false alarms. In particular, photographic scales of some lesions in the images tend to influence our model's performance. For instance, the solitary lesions are at large scale in some images, leading to cluttered eye movement patterns rather than concentrating ones as shown in the fifth image of Fig. 11(d). Since both the number of fixations and their durations are indicative of the depth of information processing associated with the particular image regions, the exhibition of *Concentrating Pattern* usually corresponds to a localized primary abnormality as shown in Fig. 11(a) and (c). The saccade amplitudes of *Switching Pattern* and *Clutter Pattern* inform both the image visual-spatial structures (symmetry) as in Fig. 11(b) and distributions of multiple abnormalities (primary abnormality versus secondary abnormality) as in Fig. 11(c).

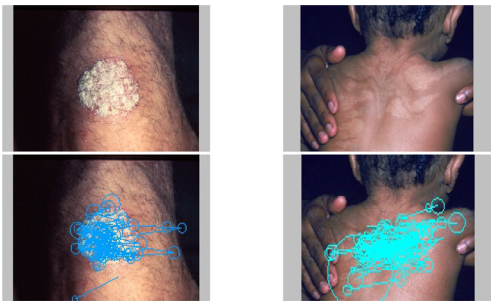


Fig. 13. Two false positive cases. The left panel shows an image labeled as *Localized* lesion with *Clutter Pattern* recognized on it. The right panel shows a case labeled as *symmetric* lesion with *Clutter Pattern* recognized on it. Image used with permission from Logical Images, Inc.

Since the dermatological images are collected for future diagnosis, and training purposes, the dermatologists took them in a particular way. They tend to center primary abnormalities and preserve as much related contextual information as possible, such as patients' demographic information, body parts, lesion size and so on. Nonetheless, these high-resolution images have complex backgrounds, and large appearance variations for luminance and camera angles. These factors cause some false alarms, as shown in Fig. 13. In particular, photographic scales of some lesions in the images tend to influence our model's performance. For instance, the localized lesions are at large scale in some images, leading to cluttered eye movement patterns rather than concentrating ones as shown in Fig. 13(a). In another case shown in Fig. 13(b) there is an angle between the camera and the patient's back, so the symmetric shape lesions are skewed in the image. When dermatologists are examining this image, they tend to focus on the half of the lesion that are closer. This leads to a *Clutter Pattern* instead of

a *Symmetric Pattern*. Since both the number of fixations and their durations are indicative of the depth of information processing associated with the particular image regions, the exhibition of *Concentrating Pattern* usually corresponds to a localized primary abnormality as shown in Fig. 11(a) and (c). The saccade amplitudes of *Switching Pattern* and *Clutter Pattern* inform both the image visual-spatial structures (symmetry) as in Fig. 11(b) and distributions of multiple abnormalities (primary abnormality versus secondary abnormality) as in Fig. 11(c).

Note that the different viewing times of dermatologists yield length-varying eye movement sequences. Since each sequence is modeled with one HMM separately, the emission distributions of which group multiple fixation-saccadic units into one pattern exhibited repeatedly. Thus longer sequence means that its corresponding longer HMM draws more pattern samples from the prior distribution, so besides containing more repeated common patterns, it likely has some unique patterns.

8. Conclusions

We proposed a hierarchical dynamic model by combining hierarchical beta processes as the prior and autoregressive-HMMs as the data model to discover dynamical patterns from three expertise-specific groups of eye movement sequences. Our approach identified expertise-specific eye movement patterns that exist over time. Center bias effect is also discovered in free viewing of scenes when there is no task (e.g. searching a particular object) assigned to subjects. However, when the dermatologists are examining and diagnosing a dermatological image, they tend to move their eyes actively and strategically. Dermatology images and experts are an appropriate test-bed, but we can also apply our approach to other problem domains. We analyzed 50 images and delivered an extensive discussion on three illustrated cases. As our future work, we will use the discovered meaningful patterns to parse corresponding image features, which correspond to deep perceptual skills (as opposed to detailed surface features only), and that, accordingly, have potential to fill the semantic gap described at the paper's beginning.

We successfully discover certain aspects of experts' domain-specific knowledge by summarizing their perceptual skills from their eye movements while diagnosing images. The domain-specific knowledge unveils the meaning and significance of the visual cues as well as the relations among functionally integral visual cues without segmentation or processing of individual objects or regions. This will benefit the traditional pixel-based statistical methods for image understanding by evaluating perceptual meanings and relations of the image features which spatially correspond to the eye movement patterns. This combination of expert knowledge and image features will help us to generalize our approach to images for which there is no experts' eye movements recorded.

Acknowledgments

Our work was supported by the National Science Foundation under grant IIS-0941452 CDI, and the National Institutes of Health under grant 1 R21 LM010039-01A1.

References

[1] M.G. Armentano, A.A. Amandi, Recognition of user intentions for interface agents with variable order markov models, in: Proceedings of UMAP, 2009, pp. 173–184.
[2] L. Ballerini, X. Li, R.B. Fisher, J. Rees, A query-by-example content-based image retrieval system of non-melanoma skin lesions, in: MCRCDS workshop of MICCAI, Springer Press, New York, 2009, pp. 312–319.
[3] M.J. Beal, Z. Ghahramani, C.E. Rasmussen, The infinite hidden markov model, in: Proceedings of NIPS, 2002, pp. 577–584.
[4] H. Beyer, K. Holtzblatt, Contextual design: defining customer-centered systems, Morgan Kaufmann.
[5] M. Bilalic, R. Langner, M. Erb, W. Grodd, Mechanisms and neural basis of object and pattern recognition: a study with chess experts, J. Exp. Psychol. Gen. 139 (2010) 728–742.
[6] P. Boersma, D. Weenink, Praat: doing phonetics by computer (Version 5.1.05), 2007.
[7] M.S. Castelhano, M.L. Mack, J.M. Henderson, Viewing task influences eye movement control during active scene perception, J. Vis. 9 (3) (2009) 1–15.
[8] M. DeAngelus, J.B. Pelz, Top-down control of eye movements: yarbus revisited, Vis. Cogn. 17 (2009) 790–811.
[9] L. Dempere-Marco, X. Hu, G.-Z. Yang, A novel framework for the analysis of eye movements during visual search for knowledge gathering, Cogn. Comput. 3 (2011) 206–222.
[10] L. Dempere-Marco, X.-P. Hu, S.M. Ellis, D.M. Hansell, G.-Z. Yang, Analysis of visual search patterns with emd metric in normalized anatomical space, IEEE Trans. Med. Imaging 25 (8) (2006) 1011–1021.
[11] R. Dewhurst, M. Nyström, H. Jarodzka, T. Foulsham, R. Johansson, K. Holmqvist, It depends on how you look at it: scanpath comparison in multiple dimensions with multimatch, a vector-based approach, Behav. Res. Methods (2012).
[12] A.T. Duchowski, J. Driver, S. Jolaoso, B.N. Ramey, A. Robbins, Scanpath comparison revisited, in: Proceedings of ETRA, 2010, pp. 219–226.
[13] M. Feusner, B. Lukoff, Testing for statistically significant differences between groups of scan patterns, in: Proceedings of ETRA, 2008, pp. 43–46.
[14] E.B. Fox, E.B. Sudderth, M.I. Jordan, A.S. Willsky, An hdp-hmm for systems with state persistence, in: Proceedings of ICML, 2008, pp. 312–319.
[15] E.B. Fox, E.B. Sudderth, M.I. Jordan, A.S. Willsky, Sharing features among dynamical systems with beta processes, in: Proceedings of NIPS, 2009, pp. 549–557.
[16] E.B. Fox, E.B. Sudderth, M.I. Jordan, A.S. Willsky, Bayesian nonparametric methods for learning markov switching processes, IEEE Signal Process. Mag. 27 (2010) 43–54.
[17] J.V. Gael, Y.W. Teh, Z. Ghahramani, The infinite factorial hidden markov model, in: Proceedings of NIPS, 2009, pp. 1697–1704.
[18] F. Galgani, Y. Sun, P.L. Lanzi, J. Leigh, Automatic analysis of eye tracking data for medical diagnosis, in: Proceedings of CIDM, 2009, pp. 195–202.
[19] W.R. Gilks, N.G. Best, K.K.C. Tan, Adaptive rejection metropolis sampling within gibbs sampling, J. Roy. Stat. Soc. 44 (1995) 455–472.
[20] J.H. Goldberg, J.I. Helfman, Scanpath clustering and aggregation, in: Proceedings of ETRA, 2011, pp. 227–234.
[21] S. Gordon, S. Lotenberg, J. Jeronimo, H. Greenspan, Evaluation of uterine cervis segmentations using ground truth from multiple experts, J. Comput. Med. Imaging Graph. 33 (3) (2009) 205–216.
[22] von Helmholtz, Handbuch der physiologischen optik, Leipzig: Voss.
[23] J.M. Henderson, G.L. Malcolm, Searching in the dark cognitive relevance drives attention in real-world scenes, Psychon. Bull. Rev. 16 (5) (2009) 850–856, doi:10.3758/PBR.16.5.850.
[24] R. Hoffman, M.S. Fiore, Perceptual (re)learning : a leverage point for human–centered computing, J. Intell. Syst. 22 (3) (2007) 79–83.
[25] L. Itti, C. Koch, Computational modelling of visual attention, Nature Rev. Neurosci. 2 (4) (2001) 194–203.
[26] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 20 (11) (1998) 1254–1259.
[27] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Human Neurobiol. 4 (1985).
[28] E. Krupinski, A. Tillack, L. Richter, J. Henderson, A. Bhatacharyya, K. Scott, A. Graham, M. Descour, J. Davis, R. Weinstein, Eye-movement study and human performance using telepathology virtual slides. implications for medical education and differences with experience, J. Human Pathol. 37 (12) (2006) 1543–1556.
[29] E. Levin, A. Zarnowski, C.A. Cohen, R. Liimakka, Human centric approach to inhomogeneous geospatial data fusion and actualization, in: Proceedings of ASPRS, 2010, pp. 1–5.
[30] R. Li, J. Pelz, P. Shi, A.R. Haake, Learning image-derived eye movement patterns to characterize perceptual expertise, in: Proceedings of CogSci, 2012, pp. 190–195.
[31] R. Li, P. Shi, A.R. Haake, Image understanding from experts' eyes by modeling perceptual skills of diagnostic reasoning processes, in: Proceedings of CVPR, 2013, pp. 2187–2194.
[32] T.D. Loboda, P. Brusilovsky, J. Brunstein, Inferring word relevance from eye–movements of readers, in: Proceedings of IUI, ACM Press, 2011, pp. 175–184.
[33] D. Manning, S. Ethell, T. Donovan, T. Crawford, How do radiologists do it? The influence of experience and training on searching for chest nodules, J. Radiogr. 12 (2) (2006) 134–142.
[34] S. Mathe, C. Sminchisescu, Action from still image dataset and inverse optimal control to learn task specific visual scanpaths, in: Proceedings of NIPS, 2013, pp. 1923–1931.
[35] J.S. McCarley, A.F. Kramer, C.D. Wickens, E.D. Vidoni, W.R. Boot, Visual skills in airport security screening, Psychol. Sci. 15 (2004) 302–306.
[36] W. McCoy, C.O. Alm, C. Calvelli, R. Li, J.B. Pelz, P. Shi, A. Haake, Annotation schemes to encode domain knowledge in medical narratives, in: Language Annotation Workshop VI of ACL, 2012, pp. 95–103.
[37] A. Oliva, A. Torralba, M.S. Catelhano, J.M. Henderson, Top-down control of visual attention in object detection, in: Proceedings of ICIP, ACM, 2003, pp. 253–256.
[38] T.J. Palmeri, A.C.-N. Wong, I. Gauthier, Computational approaches to the development of perceptual expertise, TRENDS Cogn. Sci. 8 (8) (2004) 378–386.

[39] S. Pannasch, B.M. Velichkovsky, Distractor effect and saccade amplitudes: further evidence on different modes of processing in free exploration of visual images, Vis. Cogn. 17 (6) (2009) 1109–1131.
[40] K. Rayner, Eye movements in reading and information processing: 20 years of research, Psychol. Bull. 124 (3) (1998) 372–422.
[41] R.D. Rimey, C.M. Brown, Controlling eye movements with hidden markov models, Int. J. Comput. Vis. 7 (1991) 47–65.
[42] D.D. Salvucci, Inferring intent in eye-based interfaces: Tracing eye movements with process models, in: Proceedings of CHI, 1999, pp. 254–261.
[43] D.D. Salvucci, J.H. Goldberg, Identifying fixations and saccades in eye tracking protocols, in: Proceedings of the Eye Tracking Research and Applications Symposium, 2000, pp. 71–78.
[44] A. Santella, D. DeCarlo, Robust clustering of eye movement recordings for quantification of visual interest, in: Proceedings of ETRA, 2004, pp. 27–34.
[45] M. Smuc, E. Mayr, F. Windhager, The game lies in the eye of the beholder: The influence of expertise on watching soccer, in: Proceedings of CogSci, Lawrence Erlbaum Associates, Austin, TX, 2010, pp. 1631–1636.
[46] Y.W. Teh, M.I. Jordan, M.J. Beal, D.M. Blei, Hierarchical dirichlet processes, J. Am. Stat. Assoc. 101 (476) (2006) 1566–1581.
[47] R. Thibaux, M.I. Jordan, Hierarchical beta processes and the indian buffet process, J. Mach. Learn. Res. 22 (3) (2007) 25–31.
[48] A. Torralba, A. Oliva, M.S. Castelhano, J.M. Henderson, Contextual guidance of eye movements and attention in real-world scenes: the role of global features on object search, Psychol. Rev. 113 (4) (2006) 766–786.
[49] A.M. Treisman, G. Gelade, A feature-integration theory of attention, Cogn. Psychol. 12 (2) (1980) 97–136.
[50] Y.-C. Tseng, A. Howes, The adaptation of visual search strategy to expected information gain, in: Proceedings CHI, ACM Press, 2008, pp. 1075–1084.
[51] P.J.A. Unema, S. Pannasch, M. Joos, B.M. Velichkovsky, Time course of information processing during scene perception: the relationship between saccade amplitude and fixation duration, Vis. Cogn. 12 (3) (2005) 473–494.
[52] J.M. West, A.R. Haake, E.P. Rozanski, K.S. Karn, Eyepatterns: Software for identifying patterns and similarities across fixation sequences, in: Proceedings of ETRA, 2006, pp. 149–154.
[53] J.W. Woods, C.A. Sneiderman, K. Hameed, M.J. Ackerman, C. Hatton, Using umls metathesaurus concepts to describe medical images dermatology vocabulary, J. Comput. Biol. Med. 36 (2006) 89–100.
[54] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, G.W. Cottrell, Sun: a bayesian framework for saliency using natural statistics, J. Vis. 8 (32) (2008) 1–20.