2013 IEEE International Conference on Computer Vision

Semantically-based Human Scanpath Estimation with HMMs

Huiying Liu[1]*, Dong Xu[2], Qingming Huang[3], Wen Li[2], Min Xu[4], and Stephen Lin[5]

[1]Institute for Infocomm Research, Singapore, 138632

[2]Nanyang Technological University, Singapore, 639798

[3]University of Chinese Academy of Sciences, Beijing, China, 100190

[4]University of Technology, Sydney, Sydney, Australia, 2007

[5]Microsoft Research Asia, Beijing, China, 100080

Abstract

We present a method for estimating human scanpaths, which are sequences of gaze shifts that follow visual attention over an image. In this work, scanpaths are modeled based on three principal factors that influence human attention, namely low-level feature saliency, spatial position, and semantic content. Low-level feature saliency is formulated as transition probabilities between different image regions based on feature differences. The effect of spatial position on gaze shifts is modeled as a Levy flight with the shifts following a 2D Cauchy distribution. To account for semantic content, we propose to use a Hidden Markov Model (HMM) with a Bag-of-Visual-Words descriptor of image regions. An HMM is well-suited for this purpose in that 1) the hidden states, obtained by unsupervised learning, can represent latent semantic concepts, 2) the prior distribution of the hidden states describes visual attraction to the semantic concepts, and 3) the transition probabilities represent human gaze shift patterns. The proposed method is applied to task-driven viewing processes. Experiments and analysis performed on human eye gaze data verify the effectiveness of this method.

1. Introduction

An image contains a tremendous amount of visual information that would overload the brain if absorbed all at once. To protect from this, the brain employs a scheme of visual attention, in which parts of the visual information are selected and transferred in sequence for further processing. To promote efficiency, the human visual system naturally places information of higher attention value earlier in this pipeline.

Computational visual attention has become a significant research topic in computer vision because it can inform algorithms about important areas in an image. Toward this end, much work has focused on salient object detection [15][7] and gaze density estimation [8][29]. On the other hand, only a few works have considered the process of gaze shifting and recovered a temporal ordering of attention points over an image [10][28][27]. Estimation of such temporal orderings, which we refer to as scanpaths, has applications in image transmission, display, and compression. In this paper, we present a method to estimate gaze shifts and infer scanpaths such as shown in Fig. 1.

Figure 1. Illustration of gaze shifts. The left and right images show human scanpath segments and corresponding estimates from our algorithm, respectively, where the correspondences are indicated by matching colors.

Factors that influence gaze shift can be categorized into three types: low-level feature saliency, spatial position, and semantic content. Low-level feature saliency has been the most widely adopted and investigated cue for visual attention, and is often modeled based on feature contrast. In the estimation of gaze shifts, we utilize feature differences to calculate transition probabilities between different image regions, with more visually salient regions having greater attraction of gaze.

Spatial position has commonly been used to calculate transition probabilities in graph based and random walk based methods [28][8][7][32]. In [3], it was empirically shown that gaze shifting is a Levy flight process, which is a particular type of random walk with a step length that follows a heavy-tailed distribution. In our work, we incorporate Levy flight with a 2D Cauchy distribution to model the effect of spatial position on gaze shifts. Spatial position as well as low-level feature saliency are stimulus-driven rather

*Email: liuhy@i2r.a-star.edu.sg

1550-5499/13 $31.00 © 2013 IEEE
DOI 10.1109/ICCV.2013.401

than interpretation-driven factors, and as such they both lie in the domain of bottom-up attention.

The third factor, semantic content, provides a top-down attention component that has received little consideration due to its complexity. This component can be described as task-driven, since it is affected by the user's interpretation of the image. The task may be hidden (e.g., watching a street scene) or specific (e.g., searching for cars). Even without a specific task, our high-level interpretation of a scene affects our viewing behavior. In [5], it was experimentally shown that discrete objects attract more attention and predict visual fixation much better than early saliency cues. The experiments in [9] also demonstrate significant semantic guidance of eye movements for real-world scenes. In spite of the empirical support on the importance of semantic content in guiding attention, semantic content remains difficult to exploit because object segmentation and scene interpretation are still challenging problems. Recent methods have incorporated semantics to a degree, through the use of face or person detectors [11][32], or by learning task-related feature weights [13][19].

A practical, unsupervised approach for extracting semantic concepts is through latent semantic analysis [6]. Motivated by this approach, we attempt to infer latent semantic concepts and account for them in estimating gaze shifts. This is achieved using a Hidden Markov Model (HMM), a probabilistic model for time-series data that can be well adapted to this problem. In our work, latent semantic concepts which are difficult to discern are modeled by the hidden states, while the observations produced from these states are visible in the image and extracted as low-level descriptors. Gaze shift patterns are modeled by transition probabilities between the states, and the hidden states are obtained through an unsupervised training process convenient for application.

The main technical contribution of our human scanpath estimation method is the incorporation of semantic content through an HMM formulation in which latent semantics are represented by the hidden states, and gaze shift patterns are modeled in the transition matrix. This method is applied to task-driven attention, in which the HMM is learned from training images in the same general image class as the testing images. To evaluate the similarity of estimated scanpaths to ground truth, we employ a method based on gene sequence alignment. The results of our experiments on human gaze data provide strong support for this approach.

2. Related Work

There exist numerous works on visual attention that estimate saliency or the gaze distribution over an image. Relatively few techniques consider the dynamic process of gaze shifts and estimate scanpaths. In this section, we first review existing saliency calculation methods since saliency reflects gaze allocation and represents part of the basis for gaze shifts. We then review the methods for scanpath generation and other related works.

2.1. Saliency Calculation

The family of contrast based methods occupies a major position in the field of saliency calculation, and is motivated by the biological aspect of attention. The biological basis for the idea of contrast is rooted in the center-surround structure of the receptive field. This family includes Itti's saliency method [10] and its descendants, including those based on graphs [8] and proto-objects [27]. Itti's method calculates multi-scale contrast through a Gaussian pyramid, while other methods compute contrast in different ways, e.g., random walks on graphs [7] and color co-occurrence histograms [17].

The second important family of saliency methods is based on information pursuit and explores the psychological aspect of attention. The premise for this family of methods is that while viewing an image, the subject quickly gathers as much information from it as possible. The Attention based on Information Maximization (AIM) method [4] and the Super Gaussian Component (SGC) based method [26] measure the information of each image block based on the context within the given image. In contrast to these two methods, the Saliency Using Natural statistics method (SUN) [31] estimates information by obtaining the distributions of various features from natural image datasets and then using this prior knowledge to measure the information contained in each image block.

The third family consists of machine learning based methods, which learn from user data a model to measure saliency. Representatives of this family include methods based on conditional random fields [15][30], support vector machines [11], task-dependent influences [19][2], and probabilistic multi-task learning [13]. The significance of such methods is that they can learn from particular tasks to model the top-down aspect of attention [13][19][2].

2.2. Scanpath Generation

To generate scanpaths, Itti et al. fed a saliency map into a neural network and employed the Winner Take All (WTA) and Inhibition of Return strategies [10], while Walthera and Koch identified proto-objects in the image and ranked them according to saliency value [27]. These two methods generate scanpaths according to saliency but in fact what motivates gaze shifts is far more than that. Lee proposed that gaze shifting is due heavily to the radial decrease in resolution within the fovea, and that gaze shifting aims to maximize the information gain [12]. Renninger proposed that the purpose of gaze shifts is for information maximization [21], and further verified that gaze shifts aim to reduce local uncertainty [22]. Based on the above ideas, Wang simulated

human scanpaths by exploiting the properties of the human visual system, including the decrease of resolution in the fovea, the storing and fading of working memory, and information maximization on the residual image [28]. The primary difference of our work from the method in [28] is in the calculation of transition probabilities. Specifically, we employ an HMM to model shift patterns and to explore the impact of semantic content.

2.3. Other related works

There are a few works related to our HMM-based scanpath generation in that they use hidden states to model the invisible factors affecting gaze shifts. These include methods that model eye movements for camera control [23], and passive/active patterns [18] or brain states [1] to generate saliency maps. In all of these works, the hidden states are manually defined. By contrast, our method learns them from training data.

3. Our Method

In this paper, we denote a vector/matrix by a lowercase/uppercase letter in boldface. We also represent the transpose of a vector or a matrix by the superscript $'$. The ℓ_2 distance between two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ is defined as $\|\mathbf{a}-\mathbf{b}\| = \sqrt{(\mathbf{a}-\mathbf{b})'(\mathbf{a}-\mathbf{b})}$.

In our method, we segment the image into regions and model gaze shifts in terms of transition probabilities from one region to another. With the variables corresponding to the t-th gaze position denoted by the subscript t, the probability of a region to be chosen as the next gaze location g_{t+1} is determined by three components, namely low-level features $\mathbf{y}$, semantic content z, and image position $\mathbf{u}$:

$$\begin{aligned} & p(g_{t+1}|g_1, \cdots, g_t) \\ = \; & p(\mathbf{y}_{t+1}, z_{t+1}, \mathbf{u}_{t+1}|\mathbf{y}_1, z_1, \mathbf{u}_1, \cdots, \mathbf{y}_t, z_t, \mathbf{u}_t). \end{aligned} \tag{1}$$

We assume gaze shifts to be a Markov process, meaning that the next gaze location depends only on the current one. Thus Eq. (1) can be rewritten as

$$p(g_{t+1}|g_t) = p(\mathbf{y}_{t+1}, z_{t+1}, \mathbf{u}_{t+1}|\mathbf{y}_t, z_t, \mathbf{u}_t). \tag{2}$$

We also assume the three factors to be independent, so that Eq. (2) can be expressed as

$$p(g_{t+1}|g_t) = p(\mathbf{y}_{t+1}|\mathbf{y}_t)p(z_{t+1}|z_t)p(\mathbf{u}_{t+1}|\mathbf{u}_t). \tag{3}$$

The three parts of this formula are described in the following subsections.

3.1. Low-level feature saliency

The transition probabilities determined by low-level features are calculated through feature differences between image regions. Let $\mathbf{y}^{(r)}$ be the feature vector of region r, where $r = 1, 2, \cdots, R$, and R is the number of regions. We define the weight between region r and region s as the feature distance between them:

$$W_{r,s} = \|\mathbf{y}^{(r)} - \mathbf{y}^{(s)}\|. \tag{4}$$

The low-level features used in this paper are the YUV color values and Gabor features at five scales and eight orientations, since measures of intensity, color, orientation, and texture have been widely adopted and shown to be effective for estimating saliency [10][8].

The transition probability from region r to region s is calculated by normalizing the corresponding weight by the sum of outgoing weights from region r:

$$p(\mathbf{y}^{(s)}|\mathbf{y}^{(r)}) = \frac{W_{r,s}}{\sum_{s=1}^{R} W_{r,s}}. \tag{5}$$

A graph can be formed with the regions as nodes and the transition probabilities are the weights of the edges. Random walks on such graphs have been used to construct saliency maps [28][8].

3.2. Semantic content

We describe the influence of semantic content on gaze shifts using a hidden Markov model. The HMM is a statistical tool widely used for modeling time-series data, by characterizing an underlying process based on the visible output that it generates. The system being modeled is assumed to be a Markov process with unobserved (hidden) states. The states in an HMM are not directly visible but can be estimated from the visible output which is dependent on the state. This property of the HMM makes it a suitable choice for modeling semantic content in scanpath estimation, as the hidden states can represent latent semantic concepts while the output corresponds to descriptors for the visible image.

3.2.1 HMM-based prediction of gaze shifts

An HMM with M hidden states can be represented by three parameters, $\lambda = (\boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{\Phi})$. $\boldsymbol{\pi} \in \mathbb{R}^M$ is a vector that indicates the prior distribution of the hidden states. $\boldsymbol{\Theta} \in \mathbb{R}^{M \times M}$ is the transition matrix of the states, with entries $\theta_{i,j}$ representing the probability of transiting from state i to state j. $\boldsymbol{\Phi} \in \mathbb{R}^{M \times K}$ is the emission matrix which stores the probability of an observation given a certain state. K denotes the number of emissions, which in our case is the number of visual words. Observations of image regions are described by using the Bag-of-Visual-Words (BoVW) model with SIFT descriptors, as it has been widely used for recognizing various semantic categories [6] [16]. Let w_k be the k-th visual word. An image region is represented as a vector $\mathbf{x} = [x_1, x_2, \cdots, x_K]'$, with x_k denoting the frequency of occurrences of word w_k, normalized

to $\sum_{k=1}^{K} x_k = 1$. With the BoVW representation, $\phi_{i,k}$ (*i.e.*, the (i,k)-th entry of $\mathbf{\Phi}$) is the probability of word w_k given that the state for the current region is i. Let us denote $b_i(\mathbf{x}) = p(\mathbf{x}|z=i)$ as the probability of an observation $\mathbf{x}$ conditioned on state $z=i$. With the emission matrix, $b_i(\mathbf{x})$ can be expressed as

$$b_i(\mathbf{x}) = \prod_{k=1}^{K} (\phi_{i,k})^{x_k}. \tag{6}$$

Given a sequence of gazed image regions $\{g_1, g_2, \cdots, g_T\}$, we represent its BoVW representations as $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T\}$, where each $\mathbf{x}_t$ denotes the BoVW representation of the t-th region and T is the sequence length. Denoting $\alpha_{t,i} = p(z_t = i, \mathbf{x}_1, \ldots, \mathbf{x}_t)$ as the probability of state i at time t, we can estimate $\alpha_{t,i}$ by using the forward procedure [20] as follows:

$$\begin{aligned} \alpha_{1,i} &= b_i(\mathbf{x}_1)\pi_i, \\ \alpha_{t,i} &= b_i(\mathbf{x}_t)\sum_{j=1}^{M} \theta_{j,i}\alpha_{t-1,j}, \quad \forall t = 2, \ldots, T, \end{aligned} \tag{7}$$

where π_i (*i.e.*, the i-th entry of $\boldsymbol{\pi}$) is the prior probability of the i-th hidden state. Now, we obtain $p(z_{t+1}|z_t)$ in Eq. (3) by using the probability of the region to be gazed next through the HMM:

$$\begin{aligned} p\left(\mathbf{x}_{t+1}=\mathbf{x}|\mathbf{x}_1, \ldots, \mathbf{x}_t\right) &\propto p\left(\mathbf{x}_1, \ldots, \mathbf{x}_t, \mathbf{x}\right) \\ &= \sum_{i=1}^{M} b_i(\mathbf{x}) \sum_{j=1}^{M} \theta_{j,i}\alpha_{t,j}. \end{aligned} \tag{8}$$

3.2.2 Model learning

To learn the parameters of the HMM, $\lambda = (\boldsymbol{\pi}, \mathbf{\Theta}, \mathbf{\Phi})$, we use the Expectation-Maximization (EM) algorithm to obtain maximum likelihood estimates. We sketch the training process below and refer readers to [20] for more details.

Let us denote the training set as $\{\mathcal{X}_n|_{n=1}^{N}\}$, where $\mathcal{X}_n = \{\mathbf{x}_1^{(n)}, \ldots, \mathbf{x}_{T_n}^{(n)}\}$ is the n-th training sample (*i.e.*, a saccadic scanpath from one user on one image), and N is the total number of training samples. T_n is the length of the n-th sample. To distinguish the variables corresponding to different training samples, we will add the superscript (n) to those corresponding variables when necessary.

In the E-step, we calculate the probabilities $\alpha_{t,i}$ based on the current HMM model $\lambda = (\boldsymbol{\pi}, \mathbf{\Theta}, \mathbf{\Phi})$ by using the forward algorithm as in Eq. (7). Moreover, given state $z_t = i$, we can also estimate the probability of the partial sequence after time t (*i.e.*, $\beta_{t,i} = p(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T|z_t = i)$) by using the backward procedure as follows:

$$\begin{aligned} \beta_{T,i} &= 1, \\ \beta_{t-1,i} &= \sum_{j=1}^{M} \theta_{i,j} b_j(\mathbf{x}_t)\beta_{t,j}, \quad \forall t = 2, \ldots, T. \end{aligned} \tag{9}$$

Figure 2. Visualization of seven learned hidden states, arranged in columns. The regions with red boundaries correspond to the states. The numbers below the columns are the prior probabilities of the hidden states.

In the M-step, using the obtained $\alpha_{t,i}$ and $\beta_{t,i}$ values, we re-calculate the parameters $\boldsymbol{\pi}$, $\mathbf{\Theta}$ and $\mathbf{\Phi}$. Specifically, $\boldsymbol{\pi}$ is updated to

$$\pi_i = \frac{1}{N}\sum_{n=1}^{N} \alpha_{1,i}^{(n)}. \tag{10}$$

Recall that $\mathbf{\Theta}$ models the transition probability between any two states. We first define $\xi_{t,i,j} = p(z_t = i, z_{t+1} = j|\mathcal{X})$ as the probability of a sequence being in state i at time t and in state j at time $t+1$, which can be calculated as

$$\xi_{t,i,j} = \frac{\alpha_{t,i}\theta_{i,j}b_j(\mathbf{x}_{t+1})\beta_{t+1,j}}{\sum_{i=1}^{M}\sum_{j=1}^{M}\alpha_{t,i}\theta_{i,j}b_j(\mathbf{x}_{t+1})\beta_{t+1,j}}. \tag{11}$$

Then, $\mathbf{\Theta}$ is updated as

$$\theta_{i,j} = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n-1}\xi_{t,i,j}^{(n)}}{\sum_{n=1}^{N}\sum_{t=1}^{T_n-1}\sum_{j=1}^{M}\xi_{t,i,j}^{(n)}}. \tag{12}$$

Finally, we update $\mathbf{\Phi}$ based on the number of co-occurrences of word w_k and the state $z = i$. Based on Eq. (11) and (7), the probability of a state $z_t = i$ can be calculated as $\eta_{t,i} = \sum_{j=1}^{M}\xi_{t,i,j}$ for $t = 1, \ldots, T-1$ and $\eta_{T,i} = \alpha_{T,i}$. Therefore, by defining the co-occurrence as $c(z=i, w_k) = \sum_{n=1}^{N}\sum_{t=1}^{T_n}\eta_{t,i}^{(n)}x_{t,k}^{(n)}$ with $x_{t,k}^{(n)}$ being the k-th entry of $\mathbf{x}_t^{(n)}$, the emission matrix $\mathbf{\Phi}$ is updated by

$$\phi_{i,k} = \frac{c(z=i, w_k)}{\sum_{k=1}^{K} c(z=i, w_k)}. \tag{13}$$

This EM process is repeated to update the parameters until it converges to yield the final HMM.

3.2.3 Discussion

To visualize the hidden states, for each user scanpath we estimate the state of each gazed region (Eq. (7)). For each state, the regions with high probability are used for visualization, as illustrated in Fig. 2. Here, we use the NUSEF-portrait dataset (see Sec. 4.1) with seven hidden states. The

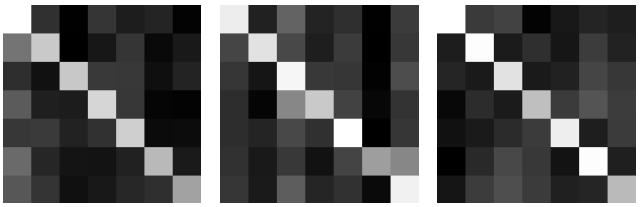


Figure 3. Visualization of three learned HMM transition matrices, with brighter shades indicating higher probability.

visualization shows that each hidden state has a consistent visual pattern (e.g., hair, hand, face, and eye) which indicates that they are able to represent semantic content to some degree.

The parameters of the HMM have practical meaning in the context of scanpath estimation. The prior distribution $\boldsymbol{\pi}$ is related to the attractiveness of each state (latent topics) with respect to the dataset. In Fig. 2, the prior probabilities of the states are given in the bottom row. We have found that certain objects always attract more attention, such as human faces and hands.

The transitions between states describe human gaze shift patterns. In [9], Hwang et al. found that human gaze tends to shift to similar concepts. Fig. 3 shows three visualized transition matrices learned in the experiment, on NUSEF-portrait (see Sec. 4.1) with 3-fold cross-validation. It can be seen that the transition matrices are consistent with the results of [9].

3.3. Spatial position

As mentioned previously, gaze shifting has been shown to be a Levy flight, which is a random walk with steps in an isotropically random direction and a step length subject to a heavy-tailed distribution [3]. Here, we use a 2D Cauchy distribution to model the gaze shift. Let $\mathbf{u}_t = (u_t, v_t)$ be the position of the t-th gaze position. The probability of transiting from $\mathbf{u}_t$ to position $\mathbf{u} = (u, v)$ is defined as

$$p(\mathbf{u}_{t+1}=\mathbf{u}|\mathbf{u}_t) = \frac{\gamma}{2\pi\left(\|\mathbf{u}-\mathbf{u}_t\|^2+\gamma^2\right)^{\frac{3}{2}}}, \quad (14)$$

where γ is the parameter of the Cauchy distribution. Under this model, the distribution of step lengths, $p(d)$ with $d = \|\mathbf{u}-\mathbf{u}_t\|$, is

$$p(d) = \frac{\gamma}{2\pi\left(d^2+\gamma^2\right)^{\frac{3}{2}}} \times 2\pi d = \frac{\gamma d}{\left(d^2+\gamma^2\right)^{\frac{3}{2}}}, \quad (15)$$

where $d \in (0, +\infty)$. Eq. (15) is a heavy-tailed distribution so the random walk defined by Eq. (14) is a Levy flight.

In several existing methods, a 2D Gaussian function is used to model the gaze shift [7][8][32]. However, as shown in Fig. 4 for human gaze data from the NUSEF dataset [25],

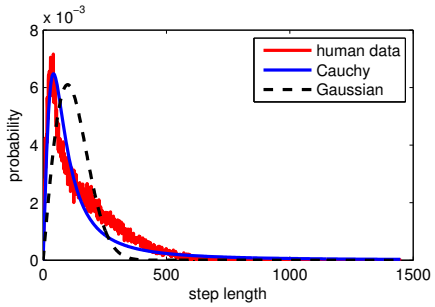


Figure 4. Step length distribution for human gaze shifts, with fitting results by using a Cauchy distribution and a Gaussian distribution.

a Gaussian function is less suitable than a Cauchy distribution for modeling gaze shift. $\gamma = 60$ is used in the figure and also in all the experiments.

3.4. Implementation details

We chose to use regions instead of pixels or image blocks as the basic perceptual unit for two reasons. First, visual grouping is a necessary step prior to semantic interpretation, where spatially neighboring pixels with similar features are perceived as a unit. Second, unlike image blocks, regions provide scale flexibility. To achieve scale invariance with block based methods, low-level feature saliency needs to be calculated at multiple scales [10][15]. For image segmentation, we employ the entropy rate superpixel segmentation method [14].

4. Experiments

In our experiments, we first evaluate performance with respect to different HMM settings, then examine the effect of each component in our method. Finally, we compare our method to other techniques. To determine the region that will be gazed next in a scanpath, we take the region with the highest probability computed from Eq. (3). Inhibition-of-return is used to prohibit the process from returning to previously gazed regions. The length of the scanpath is set to 20 for our method and the comparison techniques.

4.1. Test data and similarity metric

Dataset: We use two publicly available eye tracking datasets for evaluation, namely NUSEF [25] and JUDD [11]. Both record human gaze in a free viewing setting. The NUSEF dataset includes 758 images in total, among which we use the 476 images not protected by copyright. On average, the scanpaths of about 25 users are recorded for each image. The JUDD dataset consists of 1003 images with scanpaths of 15 subjects recorded by an eye tracking machine.

Since our method applies to task-driven viewing processes, either specific or hidden, the training and test images should ideally be from the same or similar categories.

For this we choose two subsets from the NUSEF dataset, namely NUSEF-portrait and NUSEF-face, which consist of 140 and 75 images respectively. The entire NUSEF and JUDD datasets are also used for evaluation.

Similarity metric: In [28], performance is evaluated by using the time delay embedding based distance. This method has the problem of multiple-to-one matching, i.e., multiple segments of a scanpath may be matched to the same segment of another one. Therefore, to evaluate the accuracy of an estimated scanpath, we compare it to measured ground truth using the Smith-Waterman local alignment algorithm [24], which yields a similarity measure widely used for DNA sequence comparison. Suppose we have two scanpaths, $\mathcal{X}_1$ and $\mathcal{X}_2$, of length T_1 and T_2. To identify similar scanpath segments between $\mathcal{X}_1$ and $\mathcal{X}_2$, we build a matrix $\mathbf{H} \in \mathbb{R}^{(T_1+1)\times(T_2+1)}$ with $h_{i,j}$ as the (i,j)-th entry:

$$\begin{aligned} h_{i,0} &= 0, 0 \leq i \leq T_1 \\ h_{0,j} &= 0, 0 \leq j \leq T_2, \end{aligned} \tag{16}$$

and the other elements are calculated as

$$h_{i,j} = max \begin{cases} 0 & \\ h_{i-1,j-1} + w(\mathbf{x}_i^{(1)}, \mathbf{x}_j^{(2)}) & (mis)match, \\ h_{i-1,j} + w(\mathbf{x}_i^{(1)}, -) & deletion, \\ h_{i,j-1} + w(-, \mathbf{x}_j^{(2)}) & insertion, \end{cases} \tag{17}$$

where

$$w(\mathbf{x}_i^{(1)}, \mathbf{x}_j^{(2)}) = \begin{cases} w(match), & \text{if } \mathbf{x}_i^{(1)} = \mathbf{x}_j^{(2)} \\ w(mismatch), & \text{if } \mathbf{x}_i^{(1)} \neq \mathbf{x}_j^{(2)}. \end{cases} \tag{18}$$

In this matrix, $h_{i,j}$ is the maximum similarity score between the first i entries of $\mathcal{X}_1$ and the first j entries of $\mathcal{X}_2$. To account for gaps in $\mathcal{X}_1$ and/or $\mathcal{X}_2$ during local alignment, "deletion" and "insertion" operations add a gap in $\mathcal{X}_1$ or $\mathcal{X}_2$, respectively, while adding a penalty of $w(\mathbf{x}, -)$ or $w(-, \mathbf{x})$ to the similarity score. In our method, we use the settings $w(match) = 1$ and $w(mismatch) = w(\mathbf{x}, -) = w(-, \mathbf{x}) = gap$, where gap can be set to $-\frac{1}{2}$, $-\frac{1}{3}$, or $-\frac{1}{4}$ to indicate a tolerance of gap length 1, 2, or 3 between the first and second matched elements. In calculating matches for Eq. (18), we allow a distance threshold of 50, meaning that if the distance between $\mathbf{x}_i^{(1)}$ and $\mathbf{x}_j^{(2)}$ is less than 50, they are considered to be matched. This match calculation is used to evaluate each method.

To find locally aligned pairs of segments from matrix $\mathbf{H}$, we start from its maximum element $h_{i,j}$ and trace backwards to position $(i-1, j)$, $(i, j-1)$, or $(i-1, j-1)$ depending on the direction of movement used to construct the matrix. This traceback is continued until reaching a matrix cell with zero value. The traced path describes a locally aligned segment pair between $\mathcal{X}_1$ and $\mathcal{X}_2$, which has a similarity score of $h_{i,j}$. We repeat this process to find other matched segments, starting from the next largest matrix element not included in a previously traced path. The sum of similarity scores from all the aligned segments is used as the similarity metric. For each image, we have multiple ground truth scanpaths from different users, so we compare the estimated scanpath with all of them and report the average similarity.

Three alignment examples are displayed in Fig. 1. In the red segment, our method estimates two points correctly. The green segment contains a redundant point, and the blue one has a mismatch. When $gap = -\frac{1}{2}$, the similarity of the red, green, and blue segments are 2, 1.67, 1.67, respectively.

4.2. HMM settings

We use the NUSEF-portrait dataset to examine the performance for different HMM settings and parameters. The two main HMM settings are the number of visual words (K) and the number of hidden states (M), while the number of training samples (N) also impacts performance. We set the number of regions to 300 in all the experiments.

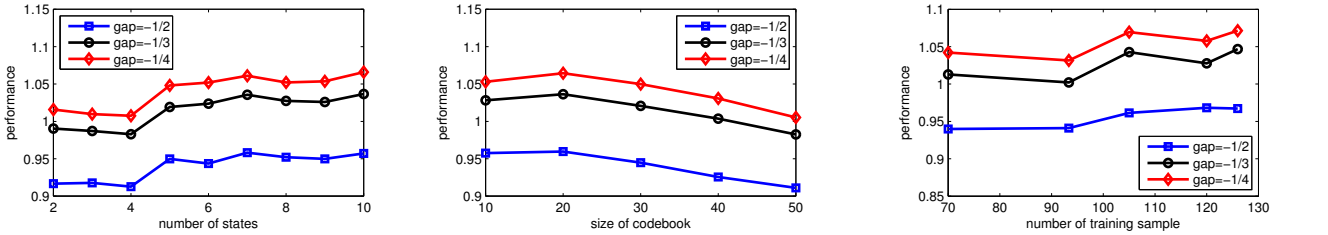
HMM setting Grid search is used to find the optimal HMM setting. We sample the number of states as $M = \{2, 3, \cdots, 10\}$ and the number of visual words as $K = \{10, 20, 30, 40, 50\}$. We use 3-fold cross-validation in this experiment. The results are displayed in Fig. 5. (a) shows the relationship between performance and number of states. The performance is averaged over all the sampled codebook sizes (number of visual words). From the figure, we can see that the performance increases quickly at first but more slowly after $M = 7$. (b) shows the performance with different codebook sizes, averaged over all the sampled numbers of states. Better performance is shown with smaller codebook sizes. Based on these tests, we choose $K = 10$ and $M = 7$ as the default HMM settings.

Number of training samples To examine the effect of the number of training samples, we performed the experiments using 2-fold, 3-fold, 4-fold, 7-fold, and 10-fold cross-validation. The results shown in (c) indicate that more training samples leads to improved performance.

4.3. Gaze factors

We compared the performance using each individual gaze factor (low-level feature saliency, semantic content with HMM, and spatial position with Levy flight) as well as the full gaze shift method with all three factors. Fig. 6 displays the results with $gap = -\frac{1}{2}$ on the four datasets, where *NUSEF* and *JUDD* denote the full datasets. Semantic content with HMM is shown to be effective and outperforms low-level feature saliency. The full combination of three factors is shown to perform significantly better than any of the factors individually. With $gap = -\frac{1}{3}$ and $gap = -\frac{1}{4}$, we have the same observation.

To determine the importance of the transition matrix,



(a) performance vs. number of states (b) performance vs. codebook size (c) performance vs. number of training samples

Figure 5. Effects of HMM settings.

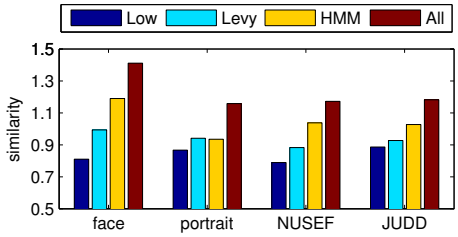


Figure 6. Effectiveness of gaze factors.

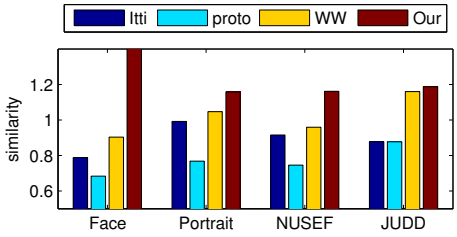


Figure 7. Comparison with other scanpath methods.

which is a key part in our approach, we conduct on the four datasets an experiment using the same HMM based approach except with a uniform transition matrix (i.e., $\theta_{i,j} = \frac{1}{M}, \forall i$ and j. With $gap = -\frac{1}{2}$, the results of the simplified HMM based approach are 0.87, 0.85, 0.79, and 0.86, which are lower than the 1.18, 0.97, 1.03, and 1.02 of our HMM based approach with the trained transition matrix (see the result from "HMM" in Fig. 6). This demonstrates the importance of modeling the gaze transitions.

More experiments are conducted to evaluate the performance for different numbers of regions, using the NUSEF dataset as an example. With $gap = -\frac{1}{2}$, the results are 1.07, 1.15, 1.19, 1.19, and 1.19 when using 100, 200, 300, 400, and 500 regions, which indicates that our approach is robust when the number of regions is within $[100, 500]$.

4.4. Comparison with other methods

To our knowledge, Itti's saliency based method (*Itti*) [10], Walthera's proto-object based method (*proto*) [27], and Wang's scanpath simulation method (*WW*) [28] are the only existing techniques for estimating scanpaths. Others output only saliency maps. We compared our method to the aforementioned techniques, using the code from [8] for *Itti*, and codes from the authors themselves for *proto* and *WW*. The results with $gap = -\frac{1}{2}$ are displayed in Fig. 7. We also conducted the t-test to examine the significance of the results. Based on the t-test, our proposed method surpasses Itti's method and the proto-object based method on all the datasets with a significance level of 0.05. Compared with Wang's method, our method is significantly better on the face, portrait, and NUSEF datasets, and is comparable to it on the JUDD dataset.

In Fig. 7, from the results for *proto*, we can see that sorting regions according to saliency does not provide good estimates of scanpaths. The improvements of our method over Wang's method on the JUDD dataset are less significant, due to its diverse content. Fig. 8 shows scanpath results for two images in the NUSEF-portrait set.

5. Conclusion

In this paper, we have proposed a human scanpath estimation method that employs an HMM to model the influence of semantic content, and uses Levy flight to account for spatial position. Experiments on challenging datasets show our method to outperform existing scanpath estimation techniques.

6. Acknowledgement

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, and in part by National Natural Science Foundation of China: 61025011. Dr. Dong Xu is partially supported by a Singapore MOE Tier 2 project.

References

[1] A. Borji, D. N. Sihite, and L. Itti. *An Object-Based Bayesian Framework for Top-Down Visual Attention*. AAAI, 2012. 3

[2] A. Borji, D. Sihite, and L. Itti. *Probabilistic learning of task-specific visual attention*. CVPR, 2012. 2

[3] D. Brockmann and T. Geisel. *Are human scanpaths Levy flights?* ICANN, 1999. 1, 5

[4] N. Bruce and J. Tsotsos. *Saliency based on information maximization*. NIPS, 2006. 2

[5] W. Einhäuser, M. Spain, and P. Perona. *Objects predict fixations better than early saliency*. Journal of Vision, 8(14):18, 1-26, 2008. 2

[6] L. Fei-Fei and P. Perona. *A Bayesian hierarchical model for learning natural scene categories*. CVPR, 2005. 2, 3

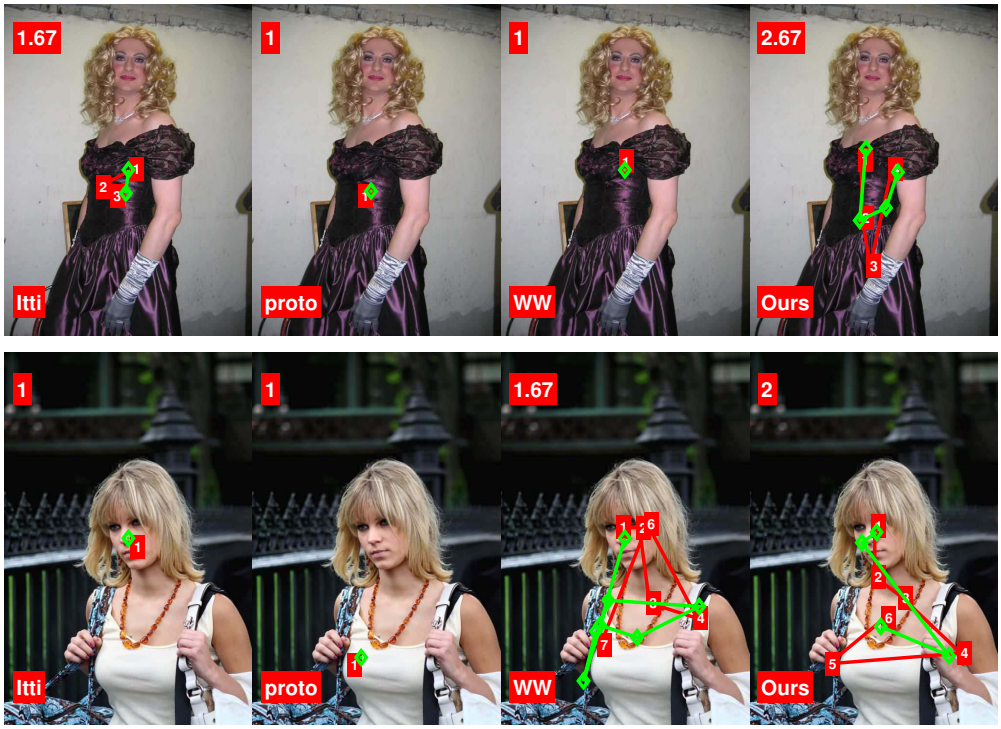


Figure 8. Scanpath comparisons. The best aligned segments are shown for each method. The red scanpaths with labeled gaze orderings are measured from users, while the green ones are estimated scanpaths. The corresponding methods are shown at the bottom-left corners, and the similarity scores for gap=-1/3 are given at the top-left corners.

[7] V. Gopalakrishnan, Y. Hu, and D. Rajan. *Random walks on graphs for salient object detection in images*. T-IP, 19(12): 3232-3242, 2010. 1, 2, 5

[8] J. Harel and C. Koch. *Graph-based visual saliency*. NIPS, 2006. 1, 2, 3, 5, 7

[9] A. D. Hwang, H.-C. Wang, and M. Pomplun. *Semantic guidance of eye movements in real-world scenes*. Vision Research, 51: 1192-1205, 2011. 2, 5

[10] L. Itti, C. Koch, and E. Niebur. *A model of saliency-based visual attention for rapid scene analysis*. T-PAMI, 20(11): 1254-1259, 1998. 1, 2, 3, 5, 7

[11] T. Judd, K. Ehinger, F. Durand, and A. Torralba. *Learning to predict where humans look*. ICCV, 2009. 2, 5

[12] T. Lee. *An information-theoretic framework for understanding saccadic eye movements*. NIPS, 2000. 2

[13] J. Li, Y. Tian, T. Huang, and W. Gao. *Probabilistic multi-task learning for visual saliency estimation in video*. IJCV, 90: 150-165, 2010. 2

[14] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa. *Entropy rate superpixel segmentation*. CVPR, 2011. 5

[15] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. *Learning to detect a salient object*. T-PAMI, 33(2): 353-367, 2011. 1, 2, 5

[16] D. G. Lowe. *Distinctive image features from scale-invariant keypoints*. IJCV, 60(2): 91-110, 2004. 3

[17] S. Lu and J. Lim. *Saliency modeling from image histograms*. ECCV, 2012. 2

[18] D. Pang, A. Kimura, T. Takeuchi, J. Yamato, and K. Kashino. *A stochastic model of selective visual attention with a dynamic Bayesian network*. ICME, 2008. 3

[19] R. J. Peters and L. Itti. *Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention*. CVPR, 2007. 2

[20] L. R. Rabiner. *A tutorial on hidden markov models and selected applications in speech recognition*. Proceedings of IEEE, 77(2), 1999. 4

[21] L. W. Renninger, J. Coughlan, P. Verghese, and J. Malik. *An information maximization model of eye movements*. NIPS, 2004. 2

[22] L. W. Renninger, P. Verghese, and J. Coughlan. *Where to look next? Eye movements reduce local uncertainty*. Journal of Vision, 7(3):6, 1-17, 2007. 2

[23] R. D. Rimey and C. M. Brown. *Controlling eye movements with hidden Markov models*. IJCV, 7(1):47-65, 1991. 3

[24] T. F. Smith and M. S. Waterman. *Identification of common molecular subsequences*. Journal of Molecular Biology, 147:195-197, 1981. 6

[25] R. Subramanian, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua. *An eye fixation database for saliency detection in images*. ECCV, 2010. 5

[26] X. Sun, H. Yao, and R. Ji. *What are we looking for: towards statistical modeling of saccadic eye movements and visual saliency*. CVPR, 2012. 2

[27] D. Walthera and C. Koch. *Modeling attention to salient proto-objects*. Neural Networks, 19: 1395-1407, 2006. 1, 2, 7

[28] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao. *Simulating human saccadic scanpaths on natural images*. CVPR, 2011. 1, 3, 6, 7

[29] W. Wang, Y. Wang, Q. Huang, and W. Gao. *Measuring visual saliency by site entropy rate*. CVPR, 2010. 1

[30] J. Yang and M. Yang. *Top-down visual saliency via joint crf and dictionary learning*. CVPR, 2012. 2

[31] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. *SUN: A Bayesian framework for salience using natural statistics*. Journal of Vision, 8(7):32, 1-20, 2008. 2

[32] Q. Zhao and C. Koch. *Learning a saliency map using fixated locations in natural scenes*. Journal of Vision, 11(3):9, 1-15, 2011. 1, 2, 5