Predicting Where We Look from Spatiotemporal Gaps

Ryo Yonetani
Kyoto University
Yoshida Hon-machi, Sakyo,
Kyoto, Japan
yonetani@vision.kuee.kyoto-u.ac.jp

Hiroaki Kawashima
Kyoto University
Yoshida Hon-machi, Sakyo,
Kyoto, Japan
kawashima@i.kyoto-u.ac.jp

Takashi Matsuyama
Kyoto University
Yoshida Hon-machi, Sakyo,
Kyoto, Japan
tm@i.kyoto-u.ac.jp

ABSTRACT

When we are watching videos, there exist spatiotemporal gaps between where we look and what we focus on, which result from temporally delayed responses and anticipation in eye movements. We focus on the underlying structures of those gaps and propose a novel method to predict points of gaze from video data. In the proposed methods, we model the spatiotemporal patterns of salient regions that tend to be focused on and statistically learn which types of the patterns strongly appear around the points of gaze with respect to each type of eye movements. It allows us to exploit the structures of gaps affected by eye movements and salient motions for the gaze-point prediction. The effectiveness of the proposed method is confirmed with several public datasets.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

Keywords

Saliency map; eye movement; spatiotemporal gap

1. INTRODUCTION

This study presents a novel method of learning to predict where we look in videos, which take into account of reaction delays and anticipation in eye movements. The proposed method predicts the points of gaze based on the underlying structures of spatiotemporal gaps between the points of gaze and salient regions that tend to be focused on around the points. It allows us to significantly improve the performance of gaze point prediction for several public datasets.

Eye movement understanding has a great potential in many research fields such as human computer interaction, interface design, computer graphics and computer vision. Eyes sometimes act as a proxy for explicit user inputs in recommender systems [30]. Furthermore, researchers have long been engaged on a problem of analyzing implicit mental states from eyes, such as interests [6], attentive states [29], intentions [24], proficiency [4], etc. Another direction is to predict where humans look from image and video data. It provides a valuable help for content design [25], automatic image cropping [23], etc. In this context, modeling visual attention mechanisms, practically as a form of *saliency maps* [2], is now a long-standing topic in the fields of computer vision and visual psychology [1, 9, 10, 16].

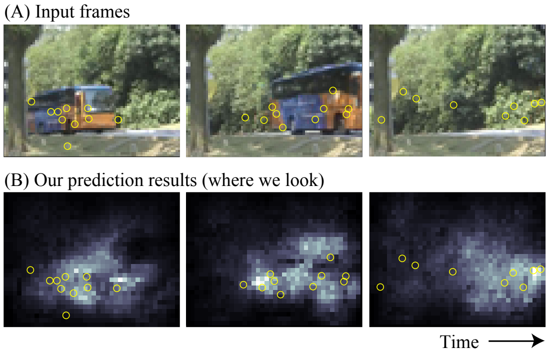

Figure 1: Results of our gaze-point prediction. The input frames in this figure are parts of the public dataset in [13]. Yellow points indicate ground truths of gaze points, where each point corresponds to one individual subject in [21]. (A) input frames that subjects watched. They provide spatiotemporal gaps between the points of gaze and the bus possibly being focused on. (B) Gaze-prediction maps. Luminance indicates the degree of gaze-point existence.

Apart from a great success in the studies on saliency maps for images, those for videos still have a particular and interesting issue when trying to predict gaze: a spatiotemporal gap between where humans look and what they focus on. That is, a point of gaze (where they look in a video frame) sometimes does not always correspond to that of covert attentional foci (what they focus on in the frame) when we are watching a video. In Fig. 1 (A), a bus went from the left side to the right, and parts of gaze points (depicted as yellow points) were following it. Even if the bus has gone out of the frame, some gaze points remained at the right of the frame. Here we can find a spatiotemporal gap as a form of reaction aftereffects; covert attentional foci were possibly on the bus in the previous frames while looking at irrelevant locations at the current frame. Riche *et al.* have tried

ICMI'13, December 09 - 13 2013, Sydney, NSW, Australia
Copyright 2013 ACM 978-1-4503-2129-7/13/12 ...$15.00.
http://dx.doi.org/10.1145/2522848.2522853 .

to overcome such a gap by broaden saliency maps based on morphological operations and smoothing [21].

On the other hand, the spatiotemporal gaps between points of gaze and those of covert attentional foci can form particular structures derived from various aspects of eye movements and salient regions being focused on, as will be reviewed in Sec. 2. We therefore focus on those structures and propose a novel method to predict where humans look in videos based on the modeled gap structures. While we follow a traditional framework of gaze point prediction based on a supervised learning framework such as [1, 10], the proposed method involves the following contributions:

1. We develop a novel model named *gap structure model (GSM)* to describe the underlying structures of spatiotemporal gaps. The GSM extracts salient regions around the points of gaze as the candidates of attentional foci and describes the gaps (relative positions to the points of gaze) and motion patterns of the regions. The fitting results of the GSM are utilized as a feature for the prediction.

2. We extend traditional learning-based methods like [1, 10] by taking into account of the types of eye movements that also affect the gaps. Specifically, we learn a model for gaze-point prediction with respect to each type of eye movements in a training phase and integrate outputs of the models into a single gaze-prediction map in a prediction phase (Fig. 1 (B)).

Particularly in the GSM, we introduce a codebook of simpler and localized patterns that efficiently describes gaps and motion patterns of salient regions. That is, gap structures are modeled by the mixture of localized patterns based on the codebook. Then, the training phase can be formulated as a problem of finding the types of localized patterns that strongly appear around the points of gaze.

Note that our method is not a family of saliency maps that find conspicuous regions from images and videos but a pure gaze prediction technique. It indicates “this point is likely to be looked at with a certain type of eye movements because there exist salient regions around the point with reasonable gaps and motion types”, which gives us reasoning to several points that conventional saliency maps have always regarded as a false negative. The applications of this study include proficiency estimation [4] and detection of developmental disorders [27], which need to know or to predict where humans actually look accurately rather than to extract conspicuous regions from images and videos.

Fig. 2 presents the overview of this study. Before introducing the proposed method, we first review several related studies and conduct some preliminary experiments to show that there surely exist gaps in Sec. 2. Then, Sec. 3 presents the GSM. It consists of the extraction of spatiotemporal patterns of salient regions and the modeling of the patterns based on the combination of localized patterns as described in Step 1 of the figure. With a codebook of the localized patterns, the gap structures can be described as a vector consisting of an activation for each of the localized patterns. In Step 2, we extract activation vectors in multiple scales so as to cope with various salient motions around the points of gaze. We then train a discriminative model from the activation vectors with respect to each type of eye movements. Finally, given a newly observed video, we evaluate the degree of gaze-point existence with all the models and integrate the outputs to obtain a single gaze-prediction map (Sec. 4).

2. PRELIMINARY STUDY

2.1 Related work

Empirically, there are many cases where we encounter a spatiotemporal gap when we are watching a video. For example, we sometimes fail to orient our eyes to salient objects captured in peripheral vision when the objects have already moved or gone out of the frame before the shift of gaze. Besides, when focusing on salient objects in fast motion, we can also observe the spatiotemporal gaps since it is hard to keep our eyes on the object regions.

Several studies have mentioned a gap between a point of gaze and that of covert attentional foci. As introduced in [7], the relationship or causality between attention and eye movements has been well studied from the early 20th century. Generally, eye movements are believed to require preceding shifts of visual attention in several cases: for example, the preview effect is a phenomenon in reading that humans fixate a word in a sentence while attending about-to-fixated word in their periphery [20]. In addition, disassociation of attention and saccadic eye movements, that is, the situation that humans move their eyes and their spatial attention to different locations, is discussed in [7].

In this study, we extract salient regions from saliency maps as a candidate for attentional foci. Since most saliency maps highlight possible locations where humans pay “location-based” (pixel-wise) attention, spatial gaps can be observed when they are paying “object-based” attention. Object-based attention, incremental grouping [22] for example, is a mechanism that spreads attentional resource within an object formed by Gestalt grouping. In such a situation, humans have the possibility to fixate locations which are not highlighted by the location-based saliency maps.

With regard to temporal gaps, we can predict a trajectory of object motions and attend the destination before the object arrives. Smooth pursuits can be indeed initiated before the beginning of the object motions [7]. In addition, experimental studies have revealed that there was a future field (or predictive remapping) mechanism in visual attention, which play an important role to predict where target would appear next [16]. As a study on a reaction delay, Rashbass has measured a saccadic response toward an object with sudden motions [19]. This study has revealed that humans required saccades with a reaction delay before smooth pursuits if they were trying to attend an object in fast motion.

Consequently, spatiotemporal gaps reflect many aspects in eye movements, and these can differ for the types of eye movements and those of salient motions. In other words, the spatiotemporal gaps can form a particular structure based on the types of eye movements and salient motions.

2.2 Statistical analysis of eye movements

This section is aimed at statistically analyzing how much spatiotemporal gaps exist in public datasets and how they are affected by the types of eye movements. As a measure of the matches between the points of gaze and salient regions, this study introduces one of traditional saliency map metrics, normalized scanpath saliency (NSS) [18]. Let $\boldsymbol{p} = (x, y, t)$ be a point in a spatiotemporal volume of videos where $(x, y) \in \mathbb{R}^2$ is a spatial location and $t \in \mathbb{N}$ is a frame

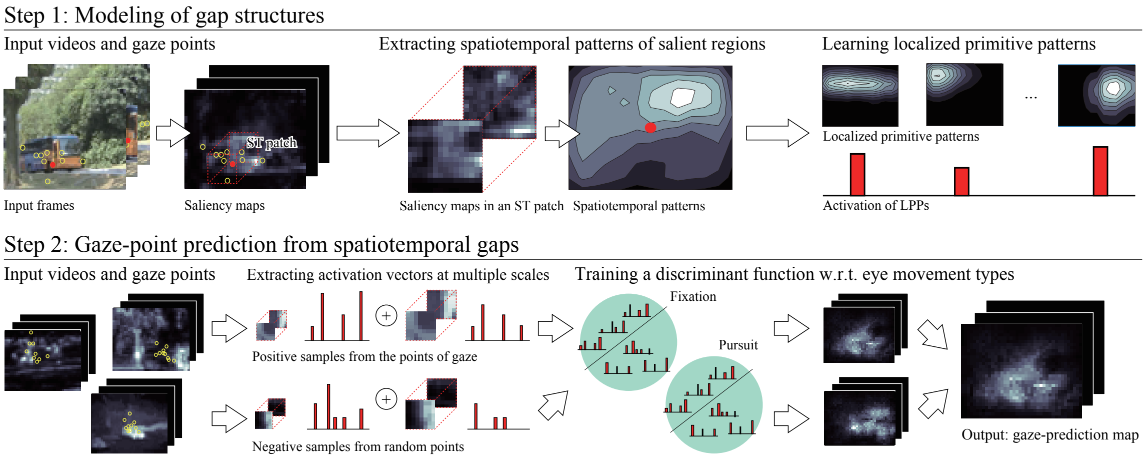
Step 1: Modeling of gap structures
Input videos and gaze points
Extracting spatiotemporal patterns of salient regions
Learning localized primitive patterns
ST patch
Localized primitive patterns
...
Input frames
Saliency maps
Saliency maps in an ST patch
Spatiotemporal patterns
Activation of LPPs
Step 2: Gaze-point prediction from spatiotemporal gaps
Input videos and gaze points
Extracting activation vectors at multiple scales
Training a discriminant function w.r.t. eye movement types
+
Fixation
Positive samples from the points of gaze
Pursuit
+
Negative samples from random points
Output: gaze-prediction map

Figure 2: Overview of the study. The input frames in this figure are parts of the public dataset in [13].

ID. A saliency map is denoted as $S : \mathbb{R}^2 \to \mathbb{R}$, where the degree of salience at point (x, y) is $S(x, y)$. The NSS evaluates the correlation between saliency maps (prediction results) and the points of gaze (ground truth labels), where a higher NSS indicates better results. Specifically, the degree of salience at point (x, y) in S is first evaluated as follows:

$$V(S, x, y) = \frac{S(x, y) - \mu(S)}{\sigma(S)}, \qquad (1)$$

where $\mu(S)$ and $\sigma(S)$ are the mean and the standard deviation of the degree of salience in S, respectively. Then, the NSS score is calculated by averaging evaluation score $V(S, x, y)$ over all the gaze points in frame t, $\mathcal{P}_t = \{\boldsymbol{p}_n = (x_n, y_n, t_n) \mid n = 1, \dots, N_t, t_n = t\}$ where N_t is the number of samples. When we deal with videos, that is, we have a sequence of saliency maps $S_1, \dots, S_T$ and corresponding gaze point sets $\mathcal{P}_1, \dots, \mathcal{P}_T$, we calculate NSS scores for all the pairs of S_t and $\mathcal{P}_t$ and average them.

In order to investigate a spatiotemporal gap statistically, we extend the NSS by calculating a mean and a standard deviation of salience in local patches of several different sizes with the center at (x, y). If NSSs at the points of gaze are high, the points of gaze possibly matches those of attentional foci. Otherwise, there are more salient regions around the gaze points; in other words, there exist gaps.

Since the degree of gaps can be affected by the types of eye movements, we divided a dataset into subsets based on the 4 types of eye movements: fixations (**FX**), slow pursuits (**SP**), fast pursuits (**FP**) and saccades (**SC**). The ascending order of eye motion speeds is **FX**<**SP**<**FP**<**SC**, and we annotated the type labels of eye movements based on the 3 thresholds of the eye-motion speeds, where the thresholds were derived by calculating 25, 50 and 75 percentile points of eye motion speed samples collected in a dataset.

In experiments, we adopted the following public datasets:

CRCNS-ORIG [8] (CRCNS) [1] contains 50 videos with a variety of genres including surveillance videos, game plays, TV news and commercials. Each video was watched by 4-6 subjects who were instructed to "follow the main actors and actions".

ASCMN database [21] (ASCMN) [2] contains 24 videos consisting of outdoor scenes, surveillance videos, videos of human crowds, etc. The videos in the database include parts of CRCNS [8], Vasconcelo's database [15][3] and a standard complex-background video surveillance database [13][4]. Parts of them contain objects with abnormal or sudden motions, which can possibly provide spatiotemporal gaps in eye movements. Each video has 10 subjects who were not instructed particularly during experiments.

We also adopted the following three models of saliency maps:

Itti's model [9] (IT) is one of traditional saliency maps, which now serves as a baseline in many studies. It calculates center-surround differences of various features such as color and motion at multiple scales and fuses them into a single map. We chose color, intensity, orientation and motion features so as to take dynamic changes in videos into account.

Cheng's model [3] (RC) is a family of salient region detection techniques that extract a region with statistical irregularity in a given image. The model first segments images into small superpixels [5] and evaluates their salience based on the rarity of color.

Torralba's model in Judd et al. [10] (TR) is a simple saliency model based on the rarity of responses from various subband pyramids, which is utilized in [10].

Table 1 demonstrates NSSs in a variety of combinatorial conditions. Note that we resized images into 80 × 60 pixels and used patches of 11 × 11 and 31 × 31 pixels to calculate NSSs. The results demonstrate that NSSs tend to decrease as an eye-motion speed increases. It indicates that there can be larger spatiotemporal gaps when eyes move faster.

In addition to the above finding, the NSSs tend to decrease as the sizes of patches get smaller. It indicates that gaze points capture globally salient regions and gaps should not

[1] `http://crcns.org/data-sets/eye/eye-1`

[2] `http://www.tcts.fpms.ac.be/attention/?article38/saliency-benchmark`

[3] `http://www.svcl.ucsd.edu/projects/background_subtraction/ucsdbgsub_dataset.htm`

[4] `http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html`

Table 1: NSS in various combinatorial conditions.

datasets	Saliency models (original NSS)	Types	31×31	11×11
CRCNS	IT (0.75)	FX	0.51	0.14
		LP	0.44	0.11
		FP	0.34	0.07
		SC	0.25	0.04
	RC (0.91)	FX	0.59	0.16
		LP	0.52	0.14
		FP	0.42	0.11
		SC	0.34	0.09
	TR (0.73)	FX	0.6	0.2
		LP	0.53	0.16
		FP	0.48	0.14
		SC	0.39	0.12
ASCMN	IT (0.62)	FX	0.38	0.09
		LP	0.31	0.05
		FP	0.21	0.01
		SC	0.17	0.01
	RC (0.59)	FX	0.34	0.09
		LP	0.26	0.05
		FP	0.21	0.04
		SC	0.18	0.03
	TR (0.39)	FX	0.29	0.08
		LP	0.24	0.05
		FP	0.15	0.02
		SC	0.14	0.02

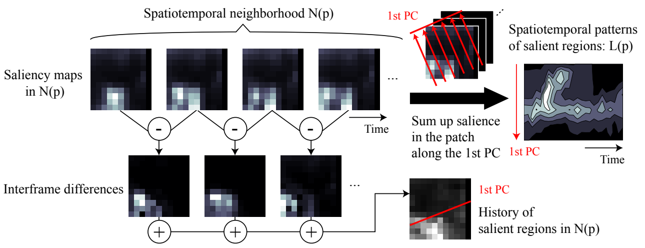
Spatiotemporal neighborhood N(p)
1st PC
Spatiotemporal patterns of salient regions: L(p)
Saliency maps in N(p)
...
-
Time
Sum up salience in the patch along the 1st PC
1st PC
Time
Interframe differences
...
1st PC
History of salient regions in N(p)
+

Figure 3: Extracting gap structures.

be so large. That is, we can look for salient regions being focused on in the neighborhood around the points of gaze.

3. GAP STRUCTURE MODEL

Now we present a model of gap structures observed between points of gaze and those of covert attentional foci, the *gap structure model (GSM)*. As discussed in Sec. 2, there are salient regions that are possibly focused on in the neighborhood around the points of gaze, and the degree of gaps can be affected by the types of eye movements and salient motions. Taking them into account, we first present a method to extract spatiotemporal patterns of salient regions in a neighborhood around a certain point (Sec. 3.1). Given a neighborhood around the points of gaze, these patterns indicate both gaps and motion patterns of salient regions, which we refer to as gap structures.

Then, the GSM describes the patterns based on a codebook consisting of simple and localized patterns, *localized primitive patterns (LPP)*. The codebook of LPPs allows us to describe complicated gap structures such as "there is a moving region until the point is looked at and at the same time a static region nearby the point for all the time" in a simple and efficient manner (Sec. 3.2).

3.1 Extracting spatiotemporal patterns of salient regions

Let us denote a spatiotemporal neighborhood around $\boldsymbol{p} = (x, y, t)$ as $\mathcal{N}(\boldsymbol{p}) = \{\boldsymbol{q} = (u, v, \tau) \mid \|x - u\| \leq \delta_x, \|y - v\| \leq \delta_y, \|t - \tau\| \leq \delta_t\}$, where $\delta_x, \delta_y, \delta_t$ define the size of the neighborhood. Then, spatiotemporal pattens of salient regions are observed in a sequence of saliency maps cropped by $\mathcal{N}(\boldsymbol{p})$. If points of gaze are given to $\boldsymbol{p}$, these patterns describe gap structures as the combination of motion patterns of salient regions and their relative positions to $\boldsymbol{p}$, i.e., gaps.

To visualize the gap structures and make the latter procedures easier, we describe the spatiotemporal patterns in $\mathcal{N}(\boldsymbol{p})$ by the patterns in a 2-d Euclidean space. Specifically, we look for an axis in a spatial domain, which describes variation of salient regions the best, and integrate the degree of salience along the axis. As depicted in Fig. 3, we first calculate inter-frame differences of saliency maps in $\mathcal{N}(\boldsymbol{p})$ and sum them up over time. The output of the above procedure provides the history of salient regions. Then, we approximate it by massive samples and calculate its 1st principal component as a direction of the maximum variation in the history. Finally, we sum up the degrees of salience along the direction of the 1st principal component for every frame to get 2-d representation of the spatiotemporal patterns, $L(\boldsymbol{p})$.

Fig. 4 depicts some examples of gap structures and corresponding situations. To visualize the locations and motion patterns of salient regions clearly, the spatiotemporal patterns of salient regions are represented by contour maps. These examples demonstrate that the points of gaze (red points) are close to salient regions but are not on the regions. In examples of smooth pursuits (**SP** and **FP**), regions are sometimes in motion, which indicate subjects tended to follow the target. Moreover, there are sometimes multiple salient regions in neighborhoods.

3.2 Learning localized primitive patterns

To model spatiotemporal pattern $L(\boldsymbol{p})$ with LPPs, we introduce an efficient description based on a codebook of LPPs obtained in a data-driven fashion. Let us denote a vectorized version of $L(\boldsymbol{p})$ as $\boldsymbol{l}(\boldsymbol{p}) \in \mathbb{R}_+^K$. Then, a codebook consisting of I LPPs is described as $\mathcal{M} = \{\boldsymbol{m}_i \in \mathbb{R}_+^K \mid i = 1, \ldots, I\}$, where m_i is a vectorized version of LPPs defined in the same spatiotemporal volume as $\boldsymbol{l}(\boldsymbol{p})$. Using this codebook, $\boldsymbol{l}(\boldsymbol{p})$ is transformed into $\boldsymbol{a}(\boldsymbol{p}) = (a_1, \ldots, a_I)^{\mathrm{T}} \in \mathbb{R}_+^I$, where a_i is an activation of LPP $\boldsymbol{m}_i$. Namely, high a_i around the points of gaze shows that there is a spatiotemporal gap between a point of gaze and salient regions, where the location and the motion pattern of the regions are described by $\boldsymbol{m}_i$.

To learn codebook $\mathcal{M}$, we adopt a non-negative matrix factorization (NMF) [11]. The NMF serves as an effective tool in various tasks such as face analysis [11], music transcription [26] and document clustering [28]. It decomposes a non-negative matrix into two non-negative factors, where one factor consists of structured bases and the other has sparse activation coefficients. Let us introduce N samples of spatial patterns $\mathcal{L} = (\boldsymbol{l}(\boldsymbol{p}_1), \ldots, \boldsymbol{l}(\boldsymbol{p}_N)) \in \mathbb{R}_+^{K \times N}$. Then, the NMF derives the two factors as $\mathcal{L} = MA$, where $M \in \mathbb{R}_+^{K \times I}$ represents a sequence of LPPs $M = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_I)$ (that is, the codebook $\mathcal{M}$) and $A \in \mathbb{R}_+^{I \times N}$ consists of activation coefficients, $A = (\boldsymbol{a}(\boldsymbol{p}_1), \ldots, \boldsymbol{a}(\boldsymbol{p}_N))$. Note that the NMF requires us to set I manually. We estimate I based on a cross validation scheme in experiments.

Fig. 5 illustrates an example of codebook $\mathcal{M}$. We obtain M and A by adopting the multiplicative update rules [12] implemented in [14].

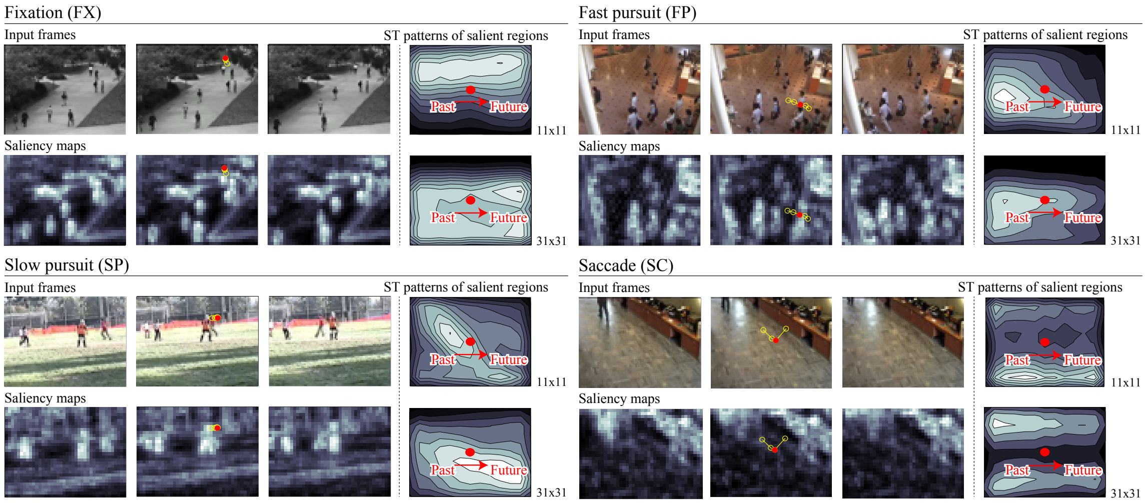

Figure 4: Examples of gap structures and corresponding situations. The input frames are parts of the public dataset in [8] (top-left), [15] (bottom-left) and [13] (top-right and bottom-right). Yellow lines in input frames and saliency maps describe the scanpath of an individual subject in [21]. Spatial size of local patches are 11×11-pixel and 31×31-pixel in 80×60 pixel-frames and temporal size δ_t is 0.4 sec. The red points in the gap structures correspond to the red gaze points in the input frames and saliency maps.

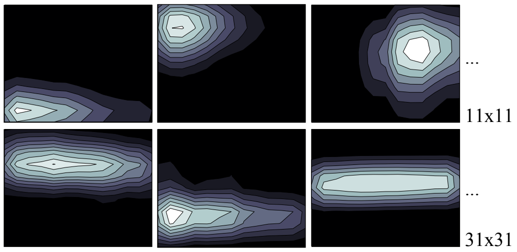

Figure 5: Examples of LPPs in a codebook.

4. PREDICTING WHERE WE LOOK FROM SPATIOTEMPORAL GAPS

Once we obtain a codebook of LPPs, we can predict the degree of gaze-point existence based on a supervised learning framework. As will be introduced in Sec. 4.1, we statistically learn which LPPs strongly appear around the points of gaze from a given sequence of saliency maps and corresponding eye movement data in a training phase. Then, we obtain a sequence of maps containing the degree of gaze point existence at each pixel (which we refer to as a gaze-prediction map) in a prediction phase.

As mentioned in the second contribution in Sec. 1, we extend the prediction framework by considering the relationships between gap structures and eye movement types. Specifically, we learn the models with respect to each type of eye movements, and integrate their outputs into a single gaze-prediction map (see Sec. 4.2).

4.1 Learning to predict where we look

As depicted in Fig. 4, the number of salient regions and their motion patterns can differ depending on the scales of neighborhood. Although they can all attract our attention, we cannot know which ones are actually focused on. We thus jointly consider multiple neighborhoods of H scales, $\mathcal{N}_h(\boldsymbol{p})$ $(h = 1, \ldots, H)$ to address this problem.

Specifically, the spatiotemporal patterns of salient regions in $\mathcal{N}_h(\boldsymbol{p})$ is first described as $L_h(\boldsymbol{p})$ and its vectorized version as $\boldsymbol{l}_h(\boldsymbol{p})$. We individually learn a codebook with respect to each of the scales, $\mathcal{M}_1, \ldots, \mathcal{M}_H$, after resizing each $\mathcal{N}_h(\boldsymbol{p})$ into the same patch size. Then, activation vectors for a scale h is described as $\boldsymbol{a}_h(\boldsymbol{p}) \in \mathbb{R}_+^{I_h}$ where I_h is the size of the codebook for scale h. Finally, we simply concatenate $\boldsymbol{a}_h(\boldsymbol{p})$ as $\boldsymbol{a}(\boldsymbol{p}) = (\boldsymbol{a}_1(\boldsymbol{p})^{\mathrm{T}}, \ldots \boldsymbol{a}_H(\boldsymbol{p})^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}_+^{I'}$ where $I' = \sum_h I_h$, to consider multiple scales jointly.

Given activation vector $\boldsymbol{a}(\boldsymbol{p})$, the degree of gaze-point existence at $\boldsymbol{p}$ can be evaluated as follows:

$$F(\boldsymbol{p}) = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{a}(\boldsymbol{p}), \tag{2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{I'}$ is a vector of model parameters. In a training phase, we fist give a binary label $\{1, -1\}$ to p, where 1 and -1 show positive (there exists a point of gaze) and negative (there does not), respectively. Then, we estimate $\boldsymbol{\beta}$ by learning a discriminant function (e.g., [1, 10]), $g(\boldsymbol{p}) = \mathrm{sgn}(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{p}) + \beta_0)$ where β_0 is a bias factor. Positive samples are collected from a set of points where subjects looked in a dataset while negative samples are from points with a lower probability of being looked at. In a prediction phase with newly observed videos, we evaluate $F(\boldsymbol{p})$ at each pixel $\boldsymbol{p}$ in the videos to obtain a sequence of gaze-prediction maps.

4.2 Introducing eye movement types

In this paper, we deal with the types of eye movements (**FX**, **SP**, **FP** and **SC** in Sec. 2.2 for example) under the assumption that each type of eye movements occurs with equal probability, independently and identically for spatial and temporal directions for simplicity. Let us introduce a set of eye movement types $E = \{e_1, \ldots, e_W\}$ to give label $e(\boldsymbol{p}) \in E$ to $\boldsymbol{p}$. We then train the discriminant function introduced in Sec. 4.1 individually from the positive samples with label $e_w \in E$ and negative samples to obtain parameter

$\boldsymbol{\beta}_{e_w}$. Based on the assumption presented above, we average model output $F_{e_w}(\boldsymbol{p}) = \boldsymbol{\beta}_{e_w}^{\mathrm{T}}\boldsymbol{a}(\boldsymbol{p})$ over e_w to evaluate the degree of gaze point existence at $\boldsymbol{p}$:

$$F_E(\boldsymbol{p}) = \frac{1}{W}\sum_{w=1}^{W} F_{e_w}(\boldsymbol{p}). \quad (3)$$

5. EXPERIMENTS

We verified the effectiveness of our gaze-point prediction with the public datasets CRCNS and ASCMN and models of saliency maps IT, RC and TR, which are introduced in Sec. 2.2. We used Eq. (2) (**GSM**) and Eq. (3) (**GSM+E**) as proposed methods. The obtained gaze-prediction maps were evaluated based on the NSS defined in Sec. 2.2.

5.1 Implementation

The proposed methods with the GSM have the following parameters to be trained:

- Scales of neighborhood $\mathcal{N}_1(\boldsymbol{p})\ldots\mathcal{N}_H(\boldsymbol{p})$ and the number of scales, H.
- LPP codebooks $\mathcal{M}_1,\ldots,\mathcal{M}_H$ and their sizes $I_1,\ldots,I_H$,
- $W-1$ thresholds of eye-motion speeds to give the types of eye movements, and the number of the types W.
- Model parameters $\boldsymbol{\beta}_{e_1},\ldots,\boldsymbol{\beta}_{e_W}$ for W types of eye movements.

H was empirically defined as $H = 2$ and the spatial sizes of $\mathcal{N}_1$ and $\mathcal{N}_2$, i.e., (δ_x, δ_y) were defined as $(5, 5)$ (11×11-pixel patch) and $(15, 15)$ (31×31-pixel patch) in 80×60 pixel-frames, respectively. The temporal sizes of $\mathcal{N}_1$ and $\mathcal{N}_2$, i.e., δ_t were both 0.4 sec. In addition, the number of eye movement types was set to $W = 4$. Those types correspond to FX, SP, FP and SC in Sec. 2.2. On the other hand, $\mathcal{M}, I, \boldsymbol{\beta}_{e_1},\ldots,\boldsymbol{\beta}_{e_W}$ and $W-1$ thresholds were estimated in a training dataset. $W-1$ thresholds were given as 25, 50, and 75 percentile eye-motion speeds in the dataset. When training a discriminant function, we adopted a Fisher's discriminant analysis so as to evaluate the effectiveness of our methods with a simple learning technique.

5.2 Evaluation schemes and baseline methods

In order to evaluate the generalization ability on videos, we conducted a leave-one-out scheme by splitting a dataset based on video IDs (and that is, we did not distinguish subjects). Specifically, we first divided the dataset consisting of C videos into $C-1$ training videos and 1 test video. From a training subset, we collected positive samples from a set of points where subjects looked. As for negatives, we randomly selected samples of the same size as positives from videos. Since the number of samples at each video frame is at most the number of subjects, the selection criterion can be regarded as the same as selecting from locations with a lower probability of being looked at. Then, we trained parameters so as to get the highest area-under-the-curve (AUC) score of a receiver operating characteristic curve with false-positive vs. true-positive rates. With a trained model, we evaluated the degree of gaze-point existence for all the pixels in the test subset to calculate NSS scores. We tested all the possible combinations of test and training videos and finally calculated an averaged NSS score over all the test subsets.

Table 2: NSS scores.

		ORIG	BS	BS+E	GSM	GSM+E
CRCNS	IT	0.75	0.86	0.85	1.14	**1.21**
	RC	0.91	1.00	1.02	1.15	**1.21**
	TR	0.73	0.86	0.89	1.10	**1.15**
ASCMN	IT	0.62	0.74	0.74	0.88	**0.90**
	RC	0.59	0.66	0.65	0.76	**0.77**
	TR	0.39	0.47	0.47	0.77	**0.82**

As a baseline method, we modified the method proposed in [21] for the sake of fairness. The original method basically utilized broadened saliency maps to fill spatiotemporal gaps. In our experiments, baseline method **BS** followed this idea and smoothed saliency maps, where the smoothing parameter was tuned so as to get the highest AUC score in a training subset. In addition, we regarded the degree of smoothed salience as a feature value for each pixel, and learn it in a discriminant function. BS can be also extended by training models with respect to eye movement types and average them over the types (**BS+E**). Consequently, we evaluated 4 methods, BS, BS+E, GSM and GSM+E under 2 datasets × 3 saliency models conditions.

5.3 Results and discussions

Table 2 shows NSS scores for all the conditions. Note that **ORIG** in the table shows NSS scores obtained from original saliency maps, which were utilized to find salient regions in the GSM. These results demonstrated the effectiveness of our proposed methods with the GSM. Although the NSS scores of ORIG, BS and BS+E had a variation with regard to saliency maps, the scores of our methods were very competitive. This fact indicates the independence of our methods to input saliency maps.

Fig. 6 depicts some qualitative results. In the 1st, 5th and 6th rows of the figure, targets were in motion and parts of gaze points were following them. These situations can provide a gap between salient regions in saliency maps (in the 2nd column) and gaze points. Our methods, GSM and GSM+E, can take into account of such a gap and show a higher degree of gaze-point existence at the points where subjects looked. On the other hand, when eye movements contain no gaps such as in the 4th row, original saliency maps and baseline methods provided higher scores.

Fig. 7 visualizes gap structures that well discriminated positive samples (points of gaze) from negatives (random points) by contour maps. We reconstructed the structures by giving coefficients of the trained discriminant function as activations of LPPs (higher values in coefficients contribute to the higher probability of gaze-point existence) and summing up the activated LPPs. In the smaller scale, salient regions tend to precede the points of gaze at any type of eye movements and saliency maps. Meanwhile for the larger scale, there are salient regions after the points of gaze for most of the cases. It indicates that subjects were somewhat predictive to salient regions, but could not accurately follow the regions without a temporal delay.

Comparing GSM+E with GSM, highlighted regions are more sparse in GSM+E as shown in Fig. 6. We can observe such outputs when one of prediction scores for different types of eye movements is particularly high. For this, Fig. 8 visualizes each of model outputs by the difference of color. In the 3rd row, we gave each pixel a 3-d value

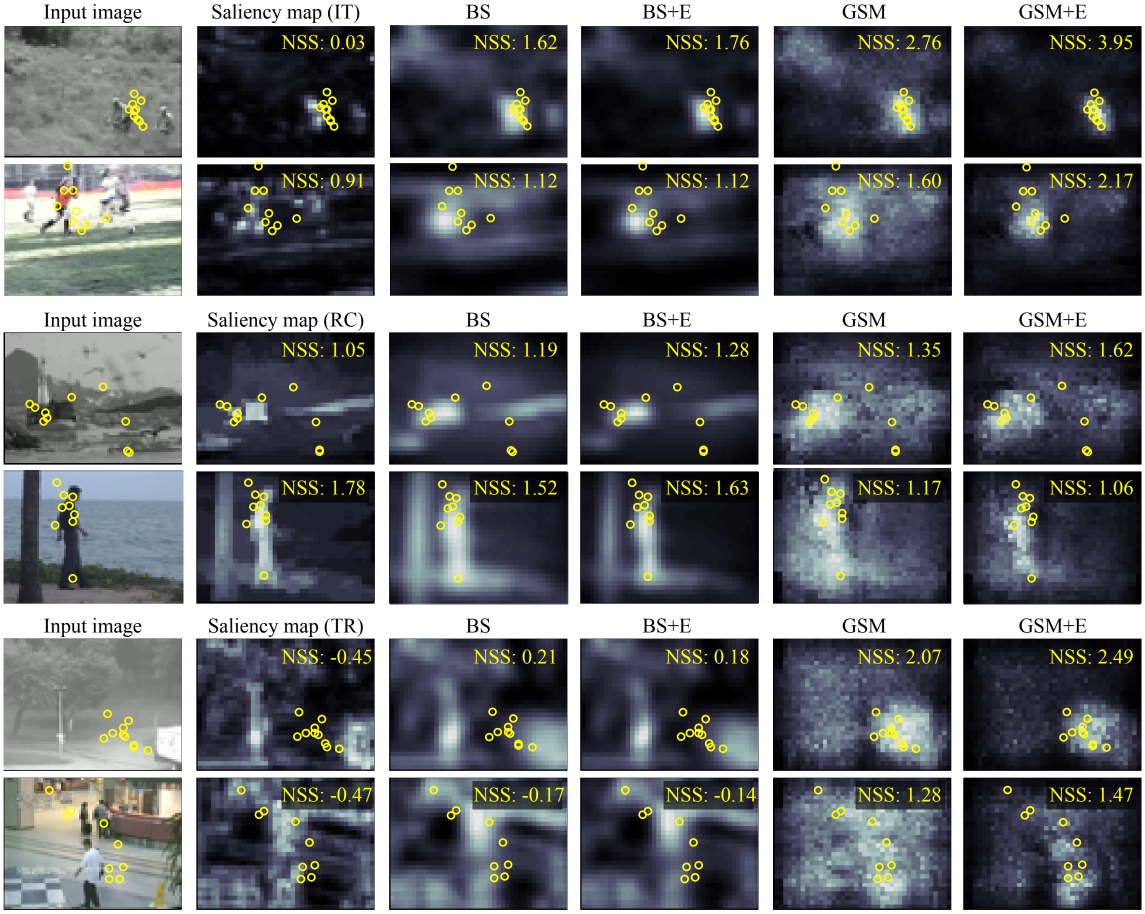

Figure 6: Qualitative results and corresponding NSS scores averaged over subjects in a frame. The input frames are parts of the public dataset in [8] (2nd row), [15] (1, 3, 5th rows) and [13] (4, 6th rows). Luminance indicates the degree of gaze-point existence. Yellow points indicate a set of gaze points, where each point corresponds to an individual subject in [21].

$(F_{e_1}(\boldsymbol{p}_n), 0.5(F_{e_2}(\boldsymbol{p}_n)+F_{e_3}(\boldsymbol{p}_n)), F_{e_4}(\boldsymbol{p}_n))$ in an RGB order where e_1, e_2, e_3, e_4 correspond to FX, SP, FP and SC, respectively. When there is a salient target in motion, model outputs of pursuits (as shown in green) become much higher than the others, and they make the final outputs sparse. In addition, there was a small probability of observing saccades when there was the target in motion, which tried to attend the target (points at the left side of the frame in the 3rd column), or to escape from the target (those at the bottom-right in the 4th column).

Finally, this study introduced a simple assumption for eye movement types, that is, each type can appear with equal probability, independently and identically for spatial and temporal directions. The prior probability on the types of eye movements can be biased, for example the saccadic eye movements can be less observed than other types. In addition, eye movement types at a certain spatiotemporal point can be statistically conditioned by those at its spatiotemporal neighborhood. In the experiments, the degree of improvements in NSSs from GSM to GSM+E is smaller than that from ORIG to GSM, and there is still room for further improvements by considering the aspect above. One promising approach is to introduce state-space models such as [17]. It assumes a Markov property for occurrences of eye movement types and gaze positions. By taking this into account, we can dynamically select models to be used based on the eye movement types which are likely to appear.

6. CONCLUSIONS

This study presented a method of gaze-point prediction based on the modeling of gap structures between the points of gaze and those of covert attentional foci. There are particular structures of gaps that depend on the types of salient motions around the points of gaze and those of eye movements. The proposed method involves those aspects by statistically learning the localized primitive patterns of salient regions with respect to each type of eye movements. We will extend our methods to consider dynamic changes of eye movement types in future work.

Acknowledgment. This work is in part supported by Grant-in-Aid for Scientific Research under the contract of 24·5573.

7. REFERENCES

[1] A. Borji. Boosting Bottom-up and Top-down Visual Features for Saliency Estimation. In *CVPR*, 2012.

[2] A. Borji and L. Itti. State-of-the-art in Visual Attention Modeling. *TPAMI*, 2012.

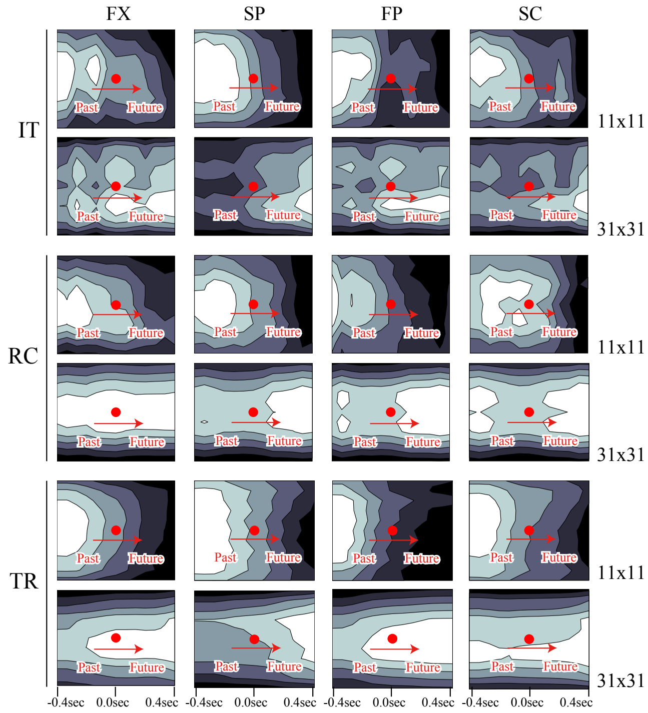

Figure 7: Examples of gap structures that well discriminated positive samples from negatives.

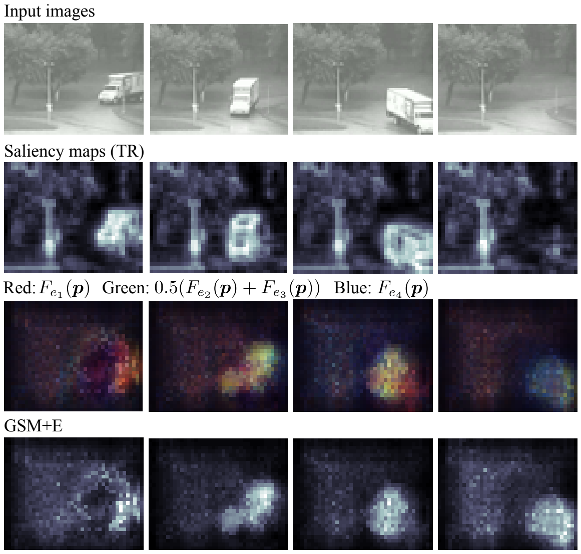
Input images
Saliency maps (TR)
Red: $F_{e_1}(\boldsymbol{p})$ Green: $0.5(F_{e_2}(\boldsymbol{p}) + F_{e_3}(\boldsymbol{p}))$ Blue: $F_{e_4}(\boldsymbol{p})$
GSM+E

Figure 8: Visualization of differences in the outputs of GSMs trained for each type of eye movements. The input frames are parts of the public dataset in [15]. In the 3rd row, red points show a high degree of gaze-point existence for fixations, green for pursuits and blue for saccades.

[3] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu. Global Contrast Based Salient Region Detection. In *CVPR*, 2011.

[4] S. Eivazi, R. Bednarik, M. Tukiainen, M. von und zu Fraunberg, V. Leinonen, and J. Jääskeläinen. Gaze Behaviour of Expert and Novice Microneurosurgeons Differs during Observations of Tumor Removal Recordings. In *ETRA*, 2012.

[5] P. Felzenszwalb and D. Huttenlocher. Efficient Graph-Based Image Segmentation. *IJCV*, 59(2):167–181, 2004.

[6] T. Hirayama, J. B. Dodane, H. Kawashima, and T. Matsuyama. Estimates of User Interest Using Timing Structures between Proactive Content-display Updates and Eye Movements. *IEICE Trans. on Information and Systems*, E-93D(6):1470–1478, 2010.

[7] J. Hoffman. Visual Attention and Eye Movements. In H. Pashler, editor, *Attention*, volume 31, chapter 3, pages 119–153. Psychology Press, 1998.

[8] L. Itti and R. Carmi. Eye-tracking Data from Human Volunteers Watching Complex Video Stimuli. CRCNS.org. http://dx.doi.org/10.6080/K0TD9V7F. 2009.

[9] L. Itti, C. Koch, and E. Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *TPAMI*, 20(11):1254–1259, 1998.

[10] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to Predict Where Humans Look. In *ICCV*, 2009.

[11] D. Lee and H. Seung. Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature*, 401(6755):788–791, 1999.

[12] D. Lee and H. Seung. Algorithms for Non-negative Matrix Factorization. In *NIPS*, 2001.

[13] L. Li, W. Huang, I. Y. Gu, and Q. Tian. Statistical Modeling of Complex Backgrounds for Foreground Object Detection. *TIP*, 13(11):1459–1472, 2004.

[14] Y. Li and A. Ngom. The Non-negative Matrix Factorization Toolbox for Biological Data Mining. *BMC Source Code for Biology and Medicine*, 8(1):10, 2013.

[15] V. Mahadevan and N. Vasconcelos. Spatiotemporal Saliency in Dynamic Scenes. *TPAMI*, 32(1):171–177, 2010.

[16] S. Mathot and J. Theeuwes. Evidence for the Predictive Remapping of Visual Attention. *Experimental Brain Research*, 200(1):117–122, 2010.

[17] D. Pang, A. Kimura, T. Takeuchi, J. Yamato, and K. Kashino. A Stochastic Model of Selective Visual Attention with a Dynamic Bayesian Network. In *ICME*, 2008.

[18] D. Parkhurst, K. Law, and E. Niebur. Modeling the Role of Salience in the Allocation of Overt Visual Attention. *Vision Research*, 42(1):107–123, 2002.

[19] C. Rashbass. The Relationship between Saccadic and Smooth Tracking Eye Movements. *The Journal of Physiology*, 159:326–338, 1961.

[20] K. Rayner. Parafoveal Identification during a Fixation in Reading. *Acta Psychologica*, 39(4):271 – 281, 1975.

[21] N. Riche, M. Mancas, and D. Culibrk. Dynamic Saliency Models and Human Attention: A Comparative Study on Videos. In *ACCV*, 2012.

[22] P. Roelfsema and R. Houtkamp. Incremental Grouping of Image Elements in Vision. *Attention, Perception & Psychophysics*, 73(8):2542–2572, 2011.

[23] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based Interaction for Semi-automatic Photo Cropping. In *CHI*, 2006.

[24] J. Simola, J. Salojärvi, and I. Kojo. Using Hidden Markov Model to Uncover Processing States from Eye Movements in Information Search Tasks. *Cognitive Systems Research*, 9(4):237–251, 2008.

[25] J. Simonin, S. Kieffer, and N. Carbonell. Effects of Display Layout on Gaze Activity During Visual Search. In *INTERACT*, volume 3585, 2005.

[26] P. Smaragdis and J. Brown. Non-negative Matrix Factorization for Polyphonic Music Transcription. In *WASPAA*, 2003.

[27] P.-H. Tseng, I. G. Cameron, G. Pari, J. N. Reynolds, D. P. Munoz, and L. Itti. High-throughput Classification of Clinical Populations from Natural Viewing Eye Movements. *Journal of Neurology*, 260(1):275–284, 2013.

[28] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *SIGIR*, 2003.

[29] R. Yonetani, H. Kawashima, , and T. Matsuyama. Multi-mode Saliency Dynamics Model for Analyzing Gaze and Attention. In *ETRA*, 2012.

[30] A. Yoshitaka, K. Wakiyama, and T. Hirashima. Recommendation of Visual Information by Gaze-based Implicit Preference Acquisition. In *MMM*, 2006.