IPSJ Transactions on Computer Vision and Applications Vol.5 75–79 (July 2013)

[DOI: 10.2197/ipsjtcva.5.75]

Express Paper

Learning Spatiotemporal Gaps between Where We Look and What We Focus on

Ryo Yonetani[1,a)] Hiroaki Kawashima[1,b)] Takashi Matsuyama[1,c)]

Received: March 11, 2013, Accepted: April 24, 2013, Released: July 29, 2013

Abstract: When we are watching videos, there are spatiotemporal gaps between where we look (points of gaze) and what we focus on (points of attentional focus), which result from temporally delayed responses or anticipation in eye movements. We focus on the underlying structure of those gaps and propose a novel learning-based model to predict where humans look in videos. The proposed model selects a relevant point of focus in the spatiotemporal neighborhood around a point of gaze, and jointly learns its salience and spatiotemporal gap with the point of gaze. It tells us "this point is likely to be looked at because there is a point of focus around the point with a reasonable spatiotemporal gap." Experimental results with a public dataset demonstrate the effectiveness of the model to predict the points of gaze by learning a particular structure of gaps with respect to the types of eye movements and those of salient motions in videos.

Keywords: saliency map, eye movement, spatiotemporal structure

1. Introduction

Visual attention is a built-in mechanism of the human visual system, which enables us to quickly select meaningful regions in our sights. Modeling this attention mechanism, practically as a form of *saliency maps*, is now a long-standing topic in the fields of computer vision and vision psychology [1], [6], [8], [9], [11]. This study is aimed at predicting where humans look *in videos* with help from saliency maps, which has many applications such as proficiency assessment [3] and automatic image cropping [13].

"Saliency" is originally the degree of differences between a point and its spatiotemporal neighborhoods, which measures how much a target attracts visual attention in a bottom-up manner [2]. Although the saliency is an effective cue to predict a point of attentional focus (PoF; the point on a screen where humans direct their attention) and furthermore, that of gaze (PoG; the point where they actually direct their eyes), these two points sometimes have a spatiotemporal gap. For example, when humans try to attend salient objects in motion, eye movements sometimes contain temporally delayed responses to the objects (see **Fig. 1**). It causes a mismatch between computed saliency maps and PoGs, and thus has the potential to make some evaluation methods irrelevant, which measure the degree of salience at PoGs (e.g., the normalized scanpath saliency and area under the curve adopted in Refs. [1], [9], [11]). Riche et al. have tried to evaluate the saliency maps under such a situation by smoothing both distributions of PoGs and saliency maps [12].

On the other hand, the gaps between PoGs and PoFs can form a particular structure because they reflect temporal aspects of eye movements such as delayed response. Therefore, we focus on the structures of gaps and propose a novel learning-based model, the *gap structure model* (*GSM*), to predict where humans look. The main contribution of this study is that we learn the underlying structures of gaps in a data-driven fashion, and exploit them to predict PoGs (the 3rd column of Fig. 1). Specifically, the GSM tells us "this point is likely to be looked at because there is a PoF around the point with a reasonable gap." Namely, it gives us reasoning to several PoGs which have always been regarded as a false negative of saliency maps, and it directly benefits the systems that work based on gaze inputs, such as Refs. [3], [13].

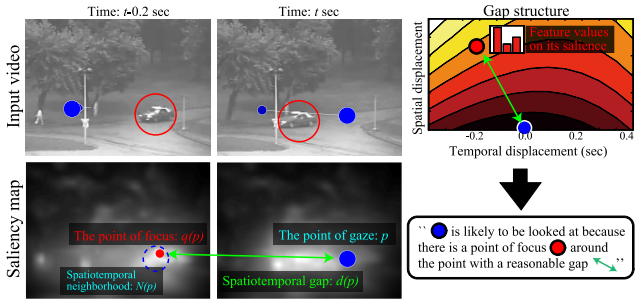

Fig. 1 Examples of eye movements with delayed response. At time $t - 0.2$ sec, a subject finds the target highlighted by a red circle. The subject then tries to shift his gaze (blue points) to the target (t sec), but cannot find it because the target has already moved. The proposed model learns spatiotemporal gaps between the point of gaze and that of focus (a red point) selected in the neighborhood of gaze point (blue dotted circle) to predict where humans look.

2. The Gap Structure Model

Let $\boldsymbol{p} \in \mathbb{R}^3_+$ be a point in a spatiotemporal volume of videos. The goal of this study is to estimate the degree of gaze-point existences $F_{\mathrm{GSM}}(\boldsymbol{p}) \in \mathbb{R}$ for the given point $\boldsymbol{p}$.

[1] Graduate School of Informatics, Kyoto University, Kyoto 606–8501, Japan
[a)] yonetani@vision.kuee.kyoto-u.ac.jp
[b)] kawashima@i.kyoto-u.ac.jp
[c)] tm@i.kyoto-u.ac.jp

The main idea of this study is to incorporate the structures of spatiotemporal gaps between the PoGs and PoFs into the model. We thus first scan the spatiotemporal neighborhood around $\boldsymbol{p}$: $N(\boldsymbol{p}) \subset \mathbb{R}^3_+$ to select a PoF $\boldsymbol{q}(\boldsymbol{p}) \in N(\boldsymbol{p})$ (practical implementation of $N(\boldsymbol{p})$ will be given in Section 3).

To select $\boldsymbol{q}(\boldsymbol{p})$, we utilize traditional saliency computation techniques such as signal-processing-based saliency maps. We select a relevant PoF in the neighborhood of PoGs, which takes the maximum value of salience in the neighborhood. Then we calculate a gap $\boldsymbol{d}(\boldsymbol{p}) = \boldsymbol{p} - \boldsymbol{q}(\boldsymbol{p})$ and evaluate it by the learned structure of gaps (see also Fig. 1). Since the saliency at $\boldsymbol{q}(\boldsymbol{p})$ also affects $F_{\mathrm{GSM}}(\boldsymbol{p})$, we jointly learn saliency with the gap structures.

In Section 2.1, we first propose a model of spatiotemporal gap structures. We then introduce learning-based models to compute saliency, and propose a technique to jointly learn the saliency and the gap structures (Section 2.2). Section 2.3 presents how to select PoFs.

2.1 Spatiotemporal Gap Structures

Psychological studies have revealed that eye movements require preceding shifts of visual attention in some cases [7]. Thus, there are sometimes spatial gaps between PoGs and locations of actual targets that a human tended to focus on. We can also observe the spatial gaps when focusing on salient objects in fast motion, since it is hard to keep our eyes on the object regions.

At the same time, there are temporal gaps as well when watching general videos. For example, we sometimes fail to orient our eyes to salient objects captured in peripheral vision, when the objects have already moved or got out of the frame. Alternatively, we can anticipate where salient objects will move next, and can attend destinations before their arrival. Consequently, the spatiotemporal gaps reflect many aspects in eye movements, and they are expected to form a particular structure. In addition, they can differ for the types of eye movements and those of salient motions.

As a first step toward the modeling of gap structures, this paper presents a basic model that only involves the following two properties: (1) There are possible locations that a PoF exists in the spatiotemporal neighborhood around the PoGs $N(\boldsymbol{p})$, and thus the degree of gap occurrences has local extrema in $N(\boldsymbol{p})$, and (2) The variations of gap occurrences along spatial and temporal directions are correlated, e.g., the PoFs are possibly spatially distant from PoGs as they are temporally distant when targets of focus are in motion. By taking them into account, we evaluate a gap $\boldsymbol{d}(\boldsymbol{p}) = \boldsymbol{p} - \boldsymbol{q}(\boldsymbol{p})$ by a quadratic function $G(\boldsymbol{d}(\boldsymbol{p}))$ of the following form:

$$G(\boldsymbol{d}(\boldsymbol{p})) = \gamma_1\|\boldsymbol{d}_s\| + \gamma_2 d_t + \gamma_3\|\boldsymbol{d}_s\|d_t + \gamma_4\|\boldsymbol{d}_s\|^2 + \gamma_5 d_t^2, \quad (1)$$

where $\boldsymbol{d}_s \in \mathbb{R}^2$ and $d_t \in \mathbb{R}$ are spatial and temporal components of $\boldsymbol{d}(\boldsymbol{p})$, respectively (i.e., $\boldsymbol{d}(\boldsymbol{p}) = (\boldsymbol{d}_s, d_t)$). Note that we do not consider spatial orientations for simplicity.

For modeling structures, similar approaches have been proposed for object detection [5] and face recognition [14]. They learn the relative positions of object (or facial) parts by introducing a quadratic function $G(\boldsymbol{d}(\boldsymbol{p})) = \gamma_1\|\boldsymbol{d}(\boldsymbol{p})\| + \gamma_2\|\boldsymbol{d}(\boldsymbol{p})\|^2$. Our model of gap structures is different from them in terms of describing not only spatial but temporal relationships.

2.2 Learning Gap Structures with Saliency

The degree of salience at the PoFs $\boldsymbol{q}(\boldsymbol{p})$ affects the existence of PoGs around $\boldsymbol{q}(\boldsymbol{p})$. For example, the PoFs with strong salience tend to attract eyes much more. Therefore, it is essential to consider such saliency information when learning gap structures.

In this paper, we combine learning-based saliency maps (LBSM) [1], [9], [11] with the gap structure model in Eq. (1) and learn them jointly. Let us denote a sample at point $\boldsymbol{x} \in \mathbb{R}^3_+$ as a K-dimensional feature vector $\boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^K$. Besides, each $\boldsymbol{x}$ is given a binary label $l(\boldsymbol{x}) \in \{1, -1\}$, where 1 and -1 show positive (salient and it attracts attention) and negative (non-salient), respectively. Then, models of LBSMs with binary labels [1], [9] can be described as follows:

$$F_{\mathrm{LBSM}}(\boldsymbol{x}) = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}), \quad (2)$$

where $\boldsymbol{\beta} \in \mathbb{R}^K$ is a vector of model parameters. Basically, positive samples with label $l(\boldsymbol{x}) = 1$ are collected from a set of points where subjects looked in a dataset, while negative samples with label $l(\boldsymbol{x}) = -1$ are from points with a lower probability of being looked at. With pairs of $\boldsymbol{\phi}(\boldsymbol{x})$ and $l(\boldsymbol{x})$, the model is trained as a classifier such as a support vector machine (SVM). Then, the score of $F_{\mathrm{LBSM}}(\boldsymbol{x})$ describes the degree of salience at the point $\boldsymbol{x}$, or how much $\boldsymbol{x}$ attracts our visual attention.

Based on the LBSM, we compute the salience at the PoFs: $F_{\mathrm{LBSM}}(\boldsymbol{q}(\boldsymbol{p}))$ (that is, we collect positive samples from $\boldsymbol{q}(\boldsymbol{p})$ instead of $\boldsymbol{p}$). The degree of gaze-point existences $F_{\mathrm{GSM}}(\boldsymbol{p})$ is finally derived as follows:

$$F_{\mathrm{GSM}}(\boldsymbol{p}) = F_{\mathrm{LBSM}}(\boldsymbol{q}(\boldsymbol{p})) + G(\boldsymbol{d}(\boldsymbol{p})), \quad (3)$$

where model parameters comprise $\boldsymbol{\beta}$ and $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)$.

2.3 Selecting the Points of Focus

Selecting the points that humans focus on around actual PoGs is a non-trivial problem, because this is a general issue in a visual attention study itself. To estimate the PoFs, we introduce a traditional technique to compute saliency maps via signal processing (e.g., Refs. [6], [8]), or their combination with top-down features (i.e., object detection scores using Refs. [5], [15] for instance). Specifically, we compute an evaluation score $S(\boldsymbol{x})$ via saliency computation preliminarily (i.e., $S(\boldsymbol{x})$ means the degree of salience at $\boldsymbol{x}$), and select the point $\boldsymbol{q}(\boldsymbol{p}) \in N(\boldsymbol{p})$ that takes the maximum value of $S(\boldsymbol{x})$ as the PoF: $\boldsymbol{q}(\boldsymbol{p}) = \arg\max_{\boldsymbol{x}}\{S(\boldsymbol{x}) \mid \boldsymbol{x} \in N(\boldsymbol{p})\}$. The specific definitions of $S(\boldsymbol{x})$ which we adopted in experiments are shown in Section 3.

Note that signal-processing-based saliency maps are generally task-invariant and available without eye movement data, whereas the LBSM adopted in the GSM (Eq. (2)) is capable of modeling task-specific (data-specific) aspects of visual attention with help from actual eye movement data. Therefore, the GSM can be regarded as the model that selects the task-invariant salient point as a PoF, and learns task-specific saliency at the selected point with a gap structure to predict surrounding PoGs.

3. Model Implementation

Equation (3) provides a general form of the GSM, and allows an arbitrary sample description $\boldsymbol{\phi}(\boldsymbol{x})$, an evaluation score to select a PoF $S(\boldsymbol{x})$ and a neighborhood $N(\boldsymbol{p})$. This section introduces the implementation example adopted in our experiments.

Sample Description

With regard to sample description $\boldsymbol{\phi}(\boldsymbol{x})$, we took the same approach as several state-of-the-art models [1], [9]: introducing low-level (biologically plausible) and high-level (top down) features. In order to verify the effectiveness of the GSM itself, we employed simpler features compared to existing literature. For low-level features, we extracted 4 types of center-surround difference maps: intensity, color, orientation and motion channels of graph-based visual saliency (GBVS) [6]. High-level features included detection scores of 3 object classes: people, faces and cars. We used deformable part models [5] for detecting people and cars and the Viola-Jones detector [15] for faces. Finally, $\boldsymbol{\phi}(\boldsymbol{x})$ was described by 7-dimensional vectors.

Selecting the points of focus

Since $\phi(\boldsymbol{x})$ comprises saliency-related features, we utilized parts of $\phi(\boldsymbol{x})$ to compute evaluation score $S(\boldsymbol{x})$. Specifically, we implemented the following three: (1) spatial saliency: the sum of intensity, color and orientation channels of the GBVS, (2) motion saliency: a motion channel of GBVS, and (3) top-down objectness: the sum of detection scores of people, faces and cars computed via Refs. [5], [15].

Regarding neighborhood $N(\boldsymbol{p})$, we considered the spatiotemporal correlations in gap occurrences introduced in Section 2.1, and defined $N(\boldsymbol{p})$ as $N(\boldsymbol{p};\alpha,\varepsilon,\tau) = \{(\boldsymbol{x}_s, x_t) \mid \|\boldsymbol{x}_s - \boldsymbol{p}_s\| - \alpha\|x_t - p_t\| < \varepsilon, \|x_t - p_t\| < \tau\}$, where $\boldsymbol{p}_s \in \mathbb{R}^2_+$ and $p_t \in \mathbb{R}_+$ are spatial and temporal components of $\boldsymbol{p}$, respectively. α is a scale factor, ε and τ are thresholds of spatial displacements at $\|x_t - p_t\| = 0$ and temporal ones, respectively.

4. Experiments

Since gaps between the PoGs and PoFs reflect various aspects in eye movements, their structures can differ for the types of the eye movements. In the experiments, we therefore evaluated our model with several detection tasks: fixation, pursuit and saccade detection from videos. That is, given a sample at $\boldsymbol{p}$, we verified if the model $F_{\mathrm{GSM}}(\boldsymbol{p})$ gives higher scores where subjects looked at $\boldsymbol{p}$ with an eye movement of interest (indicated by $l(\boldsymbol{p}) = 1$) and low otherwise.

4.1 Experimental Setups

Dataset

We used the ASCMN database [12] [*1], which comprised the following five types of general videos:

- **ABN** 5 videos with objects in abnormal motions.
- **SUR** 4 surveillance videos with no special motion event.
- **CRO** 5 videos of dense crowds.
- **MOV** 4 videos taken with a moving camera.
- **NOI** 6 videos with several sudden motions.

[*1] http://www.tcts.fpms.ac.be/attention/?categorie13/databases

Each video contains eye movement data of 10 subjects. As noted in Ref. [12], videos of each type differ with respect to the types of salient motions they contain. At the same time, all the videos have no scene changes, and types of salient motions are similar among videos in each type.

We thus conducted experiments for each type of videos independently, since the types of salient motions can affect spatiotemporal gaps. Consequently, we evaluated our model under 3 (the number of eye movement types) × 5 (the number of video types) = 15 situations in the experiments.

In addition to those situations, we also conducted evaluation with a) eye movements of each type with videos of all the types (**V-ALL**), b) eye movements of all the types with videos of each type (**E-ALL**) and c) eye movements of all the types with videos of all the types.

Annotating the types of eye movements

Fixation, pursuit and saccade were semi-automatically annotated. These three types were classified based on the speed of gaze shifts. We manually defined the thresholds of the minimum speed and length of gaze shifts to annotate saccades. For annotating pursuits, we sampled several example patterns of pursuit from the dataset and detected others which took high correlation scores with the examples.

Collecting samples for evaluation

Since subjects were not asked to follow any specific task in the ASCMN database, the spatiotemporal gaps between the PoGs and PoFs do not necessarily appear. That is, there seem to be no spatiotemporal gaps when subjects easily attended salient regions in videos. To clarify the effectiveness of the proposed model, we only used PoG samples with lower salience, which are likely to have a more salient point in their neighborhood. Specifically, we gave $\mathbf{1}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x})$ ($\mathbf{1}$ is an all-ones vector) as the degree of salience and computed a 50-percentile point as a threshold to extract the samples.

For positive samples, we randomly selected the same number of fixations, pursuit and saccade samples. For negatives, we randomly selected samples of the same size as positives from videos. Since each video frame contains at most 10 PoGs (the number of subjects), the selection criterion can be regarded as the same as selecting from locations with the lower probability of being looked at.

Evaluation scheme and comparative models

We conducted a 5-fold cross validation by splitting data into 5 subsets. A linear SVM [4] was served to jointly learn the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, where the penalty parameter C in Ref. [4] was tuned via 4-fold cross validation in the training subsets. A receiver operating characteristic curve with false-positive vs. true-positive rates and the area under the curve (AUC) was calculated from output scores.

The proposed model can adopt arbitrary evaluation scores $S(x)$ to select PoFs. In experiments, we introduced several variants of scores presented in Section 3: spatial saliency (GSM-spat), motion saliency (GSM-mot), and top-down objectness (GSM-td). Parameters in the neighborhood $N(\boldsymbol{p};\alpha,\varepsilon,\tau)$ were given as $\tau = 6$ frames (0.4 sec), $\varepsilon = 16$ pixels in a frame of 640x480 pixels. α was given as the second per frame, $\alpha = 1/15$. We also adopted the

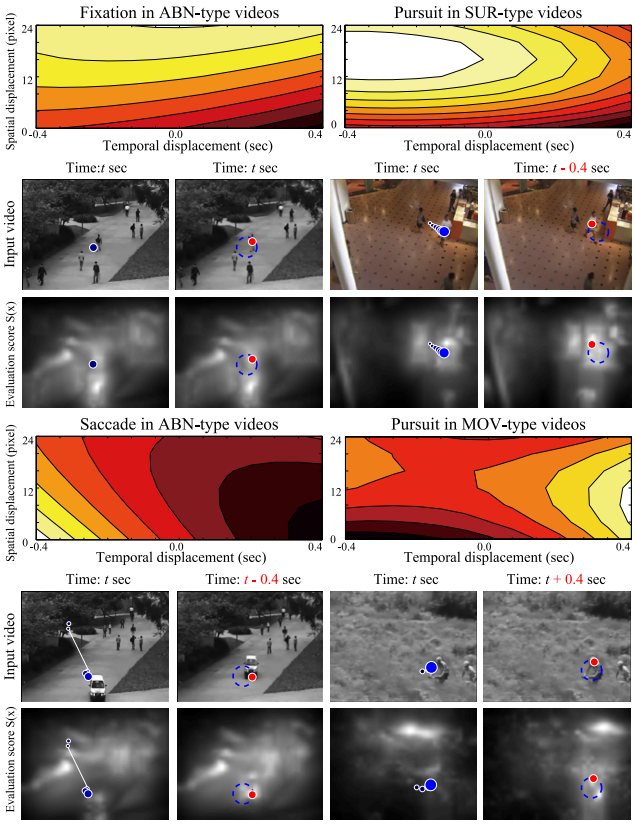

Fig. 2 Selected results of learned gap structures and corresponding situations. The horizontal and the vertical axes in the contour figures respectively show temporal displacement d_t and spatial displacement $\|\boldsymbol{d}_s\|$, where brighter regions have higher probabilities of focus existences (that is, higher degree of gap occurrences). Blue and red points describe the points of gaze and those of focus respectively, where the focused points were selected in the neighborhoods of the gaze points (blue dotted circles).

original LBSM in Eq. (2) with positive samples collected from $\boldsymbol{p}$ as a baseline.

4.2 Results and Discussions

Figure 2 visualizes selected results of learned gap structures $G(\boldsymbol{d}(\boldsymbol{p}))$ and corresponding situations in data. They indicate that there is a particular gap structure for each type of videos and eye movements. For fixation (top-left), the PoFs tend to be distant from PoGs independent of temporal displacements. Such gap structures seem to be appropriate for fixations when objects move all the time, such as ABN-type videos that contain many people walking on a street. The top-right and bottom-right results reflect typical behaviors of pursuits: pursuing targets to follow them (top-right) and precede them (bottom-right). Indeed, comparing SUR-type and MOV-type videos, MOV-types involve camera motions to capture objects at the center of the frame, and motions of the objects are slow enough to precede. Regarding ABN-type videos, saccades with anticipation were hardly observed (bottom-left), since there were many people in a frame and subjects frequently changed targets with delays.

Table 1 describes AUC scores for each experiment. The increases of scores indicate that several PoGs were correctly detected, which were regarded as false negatives of LBSM. For ABN-type videos and fixation in MOV-type videos, the motions of objects tend to be faster than that of eye movements. In addition, NOI-type videos contain objects with sudden motions. These cases can produce a temporal delay in eye reactions and bring larger improvements by the proposed methods. In fixation and saccade detection for CRO-type videos, the LBSM got the highest scores. These videos contain human crowds and it is originally difficult to extract relevant saliency information as discussed in Ref. [12]. Then, the proposed methods are likely to fail the estimation of PoFs and their gap with PoGs.

Table 1 AUC scores for each type of videos and eye movements.

Fixation	ABN	SUR	CRO	MOV	NOI	V-ALL
LBSM	0.652	0.632	**0.652**	0.640	0.673	0.613
GSM-spat	**0.712**	**0.643**	0.622	0.645	**0.689**	**0.637**
GSM-mot	0.684	0.633	0.641	0.652	0.665	0.621
GSM-td	0.644	0.640	0.613	**0.725**	0.686	0.611
Pursuit	ABN	SUR	CRO	MOV	NOI	V-ALL
LBSM	0.704	0.643	0.644	0.694	0.674	0.651
GSM-spat	0.744	0.673	0.636	0.694	**0.701**	**0.664**
GSM-mot	0.735	0.669	**0.655**	0.704	0.696	0.653
GSM-td	**0.752**	**0.676**	0.633	**0.707**	0.689	0.657
Saccade	ABN	SUR	CRO	MOV	NOI	V-ALL
LBSM	0.577	0.516	**0.640**	0.699	0.632	0.591
GSM-spat	**0.646**	0.541	0.610	0.651	**0.652**	**0.604**
GSM-mot	0.619	0.530	0.615	0.703	0.614	0.582
GSM-td	**0.646**	**0.549**	0.638	**0.704**	0.628	0.585
E-ALL	ABN	SUR	CRO	MOV	NOI	V-ALL
LBSM	0.616	0.595	0.614	0.632	0.645	0.612
GSM-spat	0.634	**0.623**	0.609	0.594	**0.666**	**0.629**
GSM-mot	**0.659**	0.596	**0.627**	0.636	0.642	0.614
GSM-td	0.654	0.640	0.598	**0.639**	0.646	0.614

In V-ALL and E-ALL that merged video and eye movement types respectively, we found smaller improvements than in the other 15 situations which divided data based on pairs of video and eye movement types. Furthermore, effective evaluation criteria to select the PoFs differed with regard to the types of videos and those of eye movements. Currently the GSM picks up the only PoF with one evaluation criterion for each of the PoGs, and we need to incorporate an automatic model selection scheme into its framework so as to cope with videos and eye movements with a variety of types. One promising technique is to introduce state-space models conditioned by eye movement types as well as video types such as in Ref. [10]. In addition, selecting multiple PoF candidates for each of the PoGs can improve the reliableness for the detected PoFs.

5. Conclusions

We presented a novel learning-based model to predict where humans look in videos, which exploit a spatiotemporal gap between the PoGs and PoFs for the prediction. Note that our experiments conducted a pixel-wise evaluation of models, which collected samples randomly from videos. Thus they are simpler than traditional schemes that build whole saliency maps and evaluate them. Future work will build saliency maps using our gap structure model, and evaluate them with state-of-the-art saliency maps and their modification by smoothing techniques [12].

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

References

[1] Borji, A.: Boosting Bottom-up and Top-down Visual Features for Saliency Estimation, *Proc. IEEE Conference on Computer Vision and*

Pattern Recognition (*CVPR*) (2012).
[2] Borji, A. and Itti, L.: State-of-the-art in Visual Attention Modeling, *IEEE Trans. Pattern Analysis and Machine Intelligence* (*PAMI*) (2012).
[3] Eivazi, S., Bednarik, R., Tukiainen, M., von und zu Fraunberg, M., Leinonen, V. and Jääskeläinen, J.: Gaze Behaviour of Expert and Novice Microneurosurgeons Differs during Observations of Tumor Removal Recordings, *Proc. Eye Tracking Research & Applications* (*ETRA*) (2012).
[4] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J.: LIBLINEAR: A library for large linear classification, *JMLR*, Vol.9, No.6/1/2008, pp.1871–1874 (2008).
[5] Felzenszwalb, P.F., Girshick, R.B., McAllester, D. and Ramanan, D.: Object Detection with Discriminatively Trained Part-Based Models, *IEEE Trans. Pattern Analysis and Machine Intelligence* (*PAMI*), Vol.32, No.9, pp.1627–1645 (2010).
[6] Harel, J., Koch, C. and Perona, P.: Graph-Based Visual Saliency, *Proc. Conference on Neural Information Processing Systems* (*NIPS*) (2007).
[7] Hoffman, J.E.: Visual Attention and Eye Movements, *Attention*, Pashler, H. (ed.), Vol.31, No.1992, chapter 3, pp.119–153, Psychology Press (1998).
[8] Itti, L., Koch, C. and Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Anaysis, *IEEE Trans. Pattern Analysis and Machine Intelligence* (*PAMI*), Vol.20, No.11, pp.1254–1259 (1998).
[9] Judd, T., Ehinger, K., Durand, F. and Torralba, A.: Learning to Predict Where Humans Look, *Proc. International Conference on Computer Vision* (*ICCV*) (2009).
[10] Pang, D., Kimura, A., Takeuchi, T., Yamato, J. and Kashino, K.: A stochastic model of selective visual attention with a dynamic Bayesian network, *Proc. IEEE International Conference on Multimedia and Expo*, pp.1073–1076 (2008).
[11] Peters, R.J. and Itti, L.: Beyond Bottom-up: Incorporating Task-Dependent Influences into a Computational Model of Spatial Attention, *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) (2007).
[12] Riche, N., Mancas, M. and Culibrk, D.: Dynamic Saliency Models and Human Attention: A Comparative Study on Videos, *Proc. Asian Conference on Computer Vision* (*ACCV*) (2012).
[13] Santella, A., Agrawala, M., DeCarlo, D., Salesin, D. and Cohen, M.: Gaze-based interaction for semi-automatic photo cropping, *Proc. SIGCHI Conference on Human Factors in Computing Systems* (*CHI*) 2006).
[14] Uřičář, M., Franc, V. and Hlávac, V.: Detector of Facial Landmarks Learned by the Structured Output SVM, *Proc. International Conference on Computer Vision Theory and Applications* (2012).
[15] Viola, P. and Jones, M.: Rapid Object Detection Using a Boosted Cascade of Simple Features, *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) (2001).

(Communicated by *Noboru Babaguchi*)