INFERRING WHAT THE VIDEOGRAPHER WANTED TO CAPTURE

Yuta Nakashima and Naokazu Yokoya

Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayamacho, Ikoma, Nara, Japan

ABSTRACT

Detecting important regions in videos has been extensively studied for past decades for their wide variety of applications including video summarization and retargeting. Visual attention models draw much attention for this purpose, which find visually salient regions. However, visual attention models ignore intentionally captured regions (ICRs) derived from videographers' intentions, *i.e.*, what the videographers wanted to capture in their videos. This paper proposes a Markov random field-based ICR model for finding them. Observing that a videographer's intention is embedded into camera motion together with objects' motion, our ICR model uses point trajectory-based features to distinguish ICRs from non-ICRs. It also leverages spatial and temporal consistency of ICRs to improve the performance. We have experimentally demonstrated our ICR model's performance and the difference between ICRs and visually salient regions.

Index Terms— Intentionally captured regions, visual attention model, capture intentions, intention map

1. INTRODUCTION

Important region detection in videos plays an essential role in various applications, such as video summarization, retargeting, and compression [1, 2, 3, 4, 5, 6]. For example, video retargeting locates important regions and crops them for adapting the video to small displays [3]. Conventionally, visual attention models [7, 8, 9, 10, 11, 12, 13] have been widely used for detecting important regions. Since they find visually salient regions that viewers may focus on, they can be viewed as important regions from viewers' perspective.

There is another perspective for important region detection, *i.e.*, videographers' one. When a videographer takes a video with a mobile camera, she/he usually has a capture intention [14], which stands for why she/he captures the video, *e.g.*, to record the growth of children or beautiful scenery. Such a capture intention induces regions in the video that are intentionally captured by the videographer as shown in Fig. 1, which are called intentionally captured regions (ICRs). ICRs are essential for the video: If the videographer's capture intention is, *e.g.*, to record the growth of her/his child, hiding the region corresponding to her/his child completely spoils the capture intention. In this sense, we can deem ICRs to be important regions from videographer's perspective.

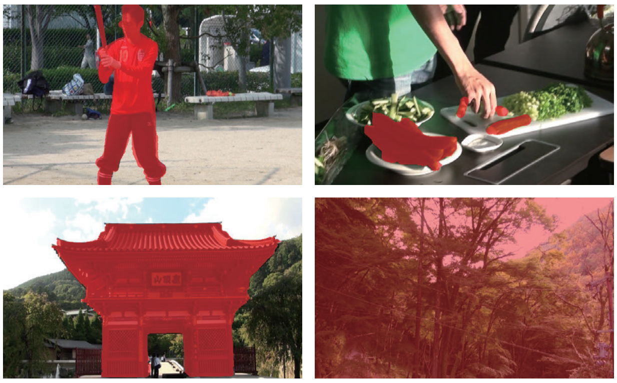

Fig. 1. Examples of ICRs, indicated by pale red regions.

In some applications, ICRs are more reasonable than visually salient regions, especially for videos taken with mobile video cameras, because the main content of such videos is usually determined by the videographers' capture intentions. This strongly motivates us to establish a method to find ICRs. In our previous work [15], a method has been proposed to classify persons in videos into intentionally captured or accidentally framed in persons and used for fully automatic video privacy protection [16, 17, 18]. However, the application of [15] is severely limited due to its incapability of finding ICRs, which can be general objects including even scenery.

This paper proposes to find ICRs through generating an intention map, which represents how likely a pixel belongs to ICRs, as in Fig. 2(d). To generate the intention map, we build an ICR model based on the observation that the capture intentions are embedded in the relationship between the camera and objects' motion. The main challenge of our ICR model is the unavailability of the objects' motion, *i.e.*, ICRs can be arbitrary regions in contrast to [15], which makes explicit detection and tracking of target objects infeasible. We instead use point trajectories obtained by [19] that provides long term trajectories of relatively dense points. The point trajectories are classified into those in ICRs or non-ICRs. The intention map is then generated based on the classification results.

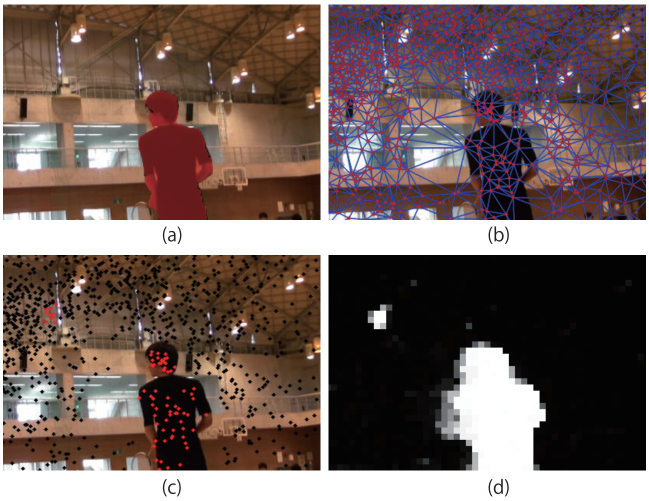

Fig. 2. Example of output for each stage. (a) Original frame with ground truth ICRs. (b) SVM outputs and delaunay triangles. Points in darker red indicate smaller decision value. (c) Probabilities given by Eq. (4). (d) Intention map.

Our ICR model is an extension of the work [20]. To improve the performance of ICR classification, our new ICR model uses the Markov random field (MRF) to leverage the spatial and temporal consistency of ICRs. The main contributions of this paper are summarized as follows: (i) We develop a MRF-based ICR model for classification, which makes use of point trajectories [19] as well as spatial and temporal consistency of ICRs. (ii) We propose to generate intention map so that our ICR model can be used instead of conventional visual attention models. (iii) We experimentally evaluate the performance of our ICR model and compare it with a visual attention model [7] to emphasize the difference between the ICRs and visually salient regions.

2. MRF-BASED ICR MODEL FOR CLASSIFICATION

When a videographer takes a video, she/he preliminarily has a capture intention and determines target objects, which correspond to ICRs in the video. For getting a better view of target objects, the videographer tries to, *e.g.*, arrange them at appropriate positions in the frame by moving the mobile camera. In contrast, the videographer usually does not pay much attention to the other objects in the scene. Therefore, trajectories of each object in the scene can be a cue for distinguishing ICRs and non-ICRs. Since detection and tracking are difficult for an arbitrary object, we instead use point trajectories [19], each of which is composed of the motions of the camera and the object that the point belongs to, for classification.

This classification can be done by the support vector machine (SVM) or other classifiers. Figure 2(b) shows examples of decision values obtained from the SVM, in which the SVM does not correctly find the ICR shown in Fig. 2(a) since it gives large decision values for the most region of the frame.

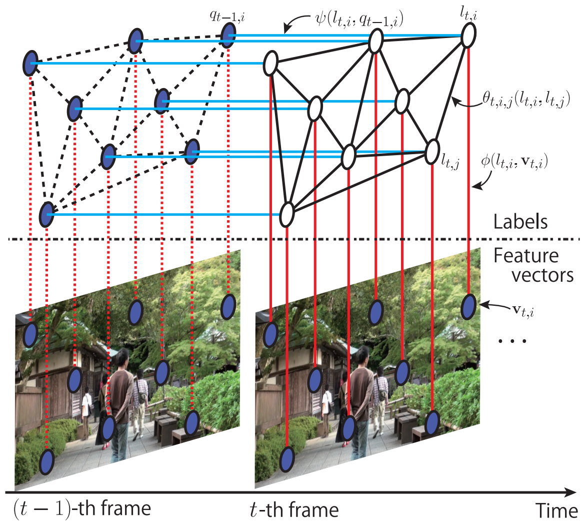

Fig. 3. Example of our ICR model. Shaded vertices represent fixed values when estimating the t-th frame's labels. Solid lines indicate actual links between vertices used in Eq. (3).

For accurate classification, we leverages the following spatial and temporal consistency through a MRF-based ICR model: (i) Since the point trajectories are dense (Fig. 2(b)), the classification results are highly correlated for neighboring points in similar color and motion. (ii) ICRs do not change frequently so that viewers can comprehend what they are seeing.

Features: We firstly calculate point trajectory-based features. Let $\mathbf{x}_{t,i} = (x_{t,i}, y_{t,i})^\top$ be the i-th point in the t-th frame where $\top$ means transpose and $i = 1, \dots, I$, and $\mathbf{X}_{t,i}$ the point trajectory that contains points corresponding to $\mathbf{x}_{t,i}$ in the successive $2T$ frames centered at the frame, *i.e.*,

$$\mathbf{X}_{t,i}^\top = (\mathbf{x}_{t-T,i}^\top, \dots, \mathbf{x}_{t,i}^\top, \dots, \mathbf{x}_{t+T-1,i}^\top). \tag{1}$$

Since $\mathbf{X}_{t,i}$ is a composition of the camera motion and the motion of the object that $\mathbf{x}_{t,i}$ belongs to, we try to find the motion of stationary background objects as a representation of the camera motion. Let $\mathbf{B}_t = (\mathbf{X}_{t,1}, \dots, \mathbf{X}_{t,I'})$ denote a matrix consisting of the point trajectories in the stationary background objects. As the rank of $\mathbf{B}_t$ is three [21], a point trajectory in $\mathbf{B}_t$ can be represented by a linear combination of the three bases of point trajectories. Assuming that the stationary background objects dominate the largest region in the frame, we extract the bases by finding $\mathbf{B}'_t = (\mathbf{X}_{t,j}, \mathbf{X}_{t,k}, \mathbf{X}_{t,m})$ that minimizes the following criterion using RANSAC as in [21]:

$$\sum_i c_{t,i}(\mathbf{B}'_t) = \sum_i \|\mathbf{B}'_t \mathbf{X}_{t,i} - \mathbf{X}_{t,i}\|, \tag{2}$$

where point trajectories in $\mathbf{B}'_t$ are selected from all point trajectories in the frame. We define feature vector $\mathbf{v}_{t,i}$ that consists of all elements of $\mathbf{X}_{t,i}$ and $\mathbf{B}'_t$ as well as $c_{t,i}(\mathbf{B}'_t)$ to indicate how likely $\mathbf{x}_{t,i}$ is on stationary background objects.

MRF-based ICR model: Our classification assigns to $\mathbf{x}_{t,i}$ a label $l_{t,i} \in \{0, 1\}$ representing whether $\mathbf{x}_{t,i}$ belongs to ICRs ($l_{t,i} = 1$) or not ($l_{t,i} = 0$). For this, we develop a MRF-based ICR model, which calculate probability $q_{t,i}$ of $l_{t,i} = 1$ given feature vectors and the previous frame's result.

Figure 3 shows our ICR model. Let $\mathbf{V}_t = \{\mathbf{v}_{t,1}, \ldots, \mathbf{v}_{t,I}\}$ be the set of feature vectors and $\mathbf{q}_{t-1} = (q_{t-1,1}, \ldots, q_{t-1,I})$ a vector of the probabilities of $l_{t-1,i} = 1$ obtained as the previous frame's result. We define the probability of label vector $\mathbf{l}_t = (l_{t,1}, \ldots, l_{t,I})$ given $\mathbf{q}_{t-1}$ and $\mathbf{V}_t$ as $p(\mathbf{l}_t|\mathbf{q}_{t-1}, \mathbf{V}_t) = e^{-E(\mathbf{l}_t, \mathbf{q}_{t-1}, \mathbf{V}_t)}/Z$, where E is an energy function, and Z is a normalizing constant. E is defined as

$$\begin{aligned} E(\mathbf{l}_t, \mathbf{q}_{t-1}, \mathbf{V}_t) &= \sum_i [\phi(l_{t,i}, \mathbf{v}_{t,i}) + \beta\psi(l_{t,i}, q_{t-1,i})] \\ &+ \sum_{(i,j)\in A} \gamma\theta_{t,i,j}(l_{t,i}, l_{t,j}), \qquad (3) \end{aligned}$$

where ϕ, ψ, and θ are data, temporal consistency, and spatial consistency terms, respectively, and A is a set of all adjacent points. β and γ control the contribution of ψ and θ. The marginal probability of $l_{t,i}$ is given by

$$p(l_{t,i}|\mathbf{l}_t \setminus l_{t,i}, \mathbf{q}_{t-1}, \mathbf{V}_t) = \frac{p(\mathbf{l}_t|\mathbf{q}_{t-1}, \mathbf{V}_t)}{\sum_{l_{t,i}\in\{0,1\}} p(\mathbf{l}_t|\mathbf{q}_{t-1}, \mathbf{V}_t)}, \quad (4)$$

and $q_{t,i}$ is given by $q_{t,i} = p(l_{t,i} = 1|\mathbf{l}_t \setminus l_{t,i}, \mathbf{q}_{t-1}, \mathbf{V}_t)$.

Data term $\phi(l_{t,i}, \mathbf{v}_{t,i})$ models the relationship between $l_{t,i}$ and $\mathbf{v}_{t,i}$. We use a SVM for this purpose. The decision value from a trained SVM is denoted by $f(\mathbf{v}_{t,i})$, where a large value of $f(\mathbf{v}_{t,i})$ implies $l_{t,i} = 1$. The data term is given by

$$\phi(l_{t,i}, \mathbf{v}_{t,i}) = \begin{cases} 1 - \varsigma(f(\mathbf{v}_{t,i})) & \text{if } l_{t,i} = 1 \\ \varsigma(f(\mathbf{v}_{t,i})) & \text{otherwise} \end{cases}, \quad (5)$$

where $\varsigma(\cdot)$ is a sigmoid function to make this term in $[0, 1]$.

Temporal consistency term $\psi(l_{t,i}, q_{t-1,i})$ penalizes different labels assigned to corresponding points in successive frames. Considering that a label with a high probability likely to give the same label in the next frame, we define the temporal consistency term as

$$\psi(l_{t,i}, q_{t-1,i}) = \begin{cases} 1 - q_{t-1,i} & \text{if } l_{t,i} = 1 \\ q_{t-1,i} & \text{otherwise} \end{cases}. \quad (6)$$

Spatial consistency term $\theta_{t,i,j}(l_{t,i}, l_{t,j})$ penalizes different labels assigned to adjacent points $\mathbf{x}_{t,i}$ and $\mathbf{x}_{t,j}$, where Delaunay triangulation determines set of adjacent points. Observing if labels $l_{t,i}$ and $l_{t,j}$ agree or not correlates to how much the points differ, we define the spatial consistency term as

$$\theta_{t,i,j}(l_{t,i}, l_{t,j}) = \begin{cases} \min(d_{t,i,j}, \alpha) & \text{if } l_{t,i} = l_{t,j} \\ \alpha & \text{otherwise} \end{cases}. \quad (7)$$

In this term, $d_{t,i,j}$ measures the difference between $\mathbf{x}_{t,i}$ and $\mathbf{x}_{t,j}$ in color, motion, and position. Let $\mathbf{u}_{t,i,j}$ be a vector consisting of the Euclid distances of the colors, motions, and positions between $\mathbf{x}_{t,i}$ and $\mathbf{x}_{t,j}$. Each component in this vector is expected to be small if the labels agrees; therefore, we define $d_{t,i,j}$ as the squared Mahalanobis distance of the vector, *i.e.*, $d_{t,i,j} = \mathbf{u}_{t,i,j}^\top \Sigma^{-1} \mathbf{u}_{t,i,j}$, where the mean is assumed to be $\mathbf{0}$, and the variance Σ is estimated from a training dataset. The term defined in Eq. (7) gives a small value when $d_{t,i,j}$ is small, but the value is bounded by α since a larger value of $d_{t,i,j}$ does not necessarily indicate different labels.

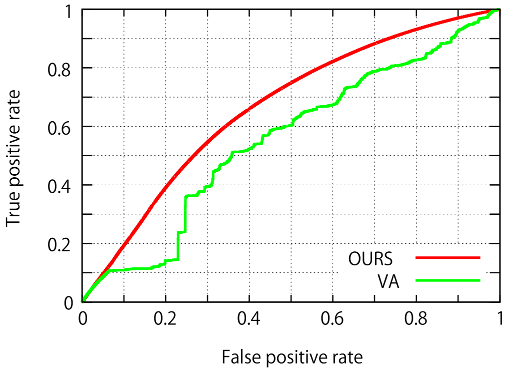

Fig. 4. ROC curves.

A label vector that approximately minimizes the energy can be obtained by the graph cut algorithm [22], which is fed into intention map generation. Figure 2(c) shows $q_{t,i}$'s, which clearly distinguish ICR and non-ICR compared to Fig. 2(b).

3. INTENTION MAP GENERATION

We generate an intention map that indicates how likely each pixel belongs to an ICR. For this, letting $M_t(\mathbf{z})$ be the value of intention map for the pixel at $\mathbf{z} = (x, y)$ in the t-th frame, $q_{t,i}$ is propagated to pixels around $\mathbf{x}_{t,i}$ as

$$M_t(\mathbf{z}) = \frac{1}{K} \sum_{i\in NN_t(\mathbf{z})} q_{t,i}, \quad (8)$$

where $NN_t(\mathbf{z})$ is the set of indices of K nearest neighbor points to $\mathbf{z}$. An example of $M_t(\mathbf{z})$ is shown in Fig. 2(d).

4. EXPERIMENTAL RESULTS

We evaluated our ICR model using a dataset containing 44 videos with ground truth ICRs. The average duration of the videos was 46.2 second at 30 frames per second. We asked the videographers who took the videos to specify the ICRs for every 10 frames since this was a laborious task, and the specified ICRs were used as ground truth.

To evaluate the performance of our ICR model, we employed the false positive rate (FPR) and the true positive rate (TPR). Let TP, TN, FP, and FN be the numbers of true positives, true negatives, false positives, and false negatives, where TP is, for example, defined as the number of points for which $q_{t,i} > TH$ and $\mathbf{x}_{t,i}$ is included in a ground truth ICR.

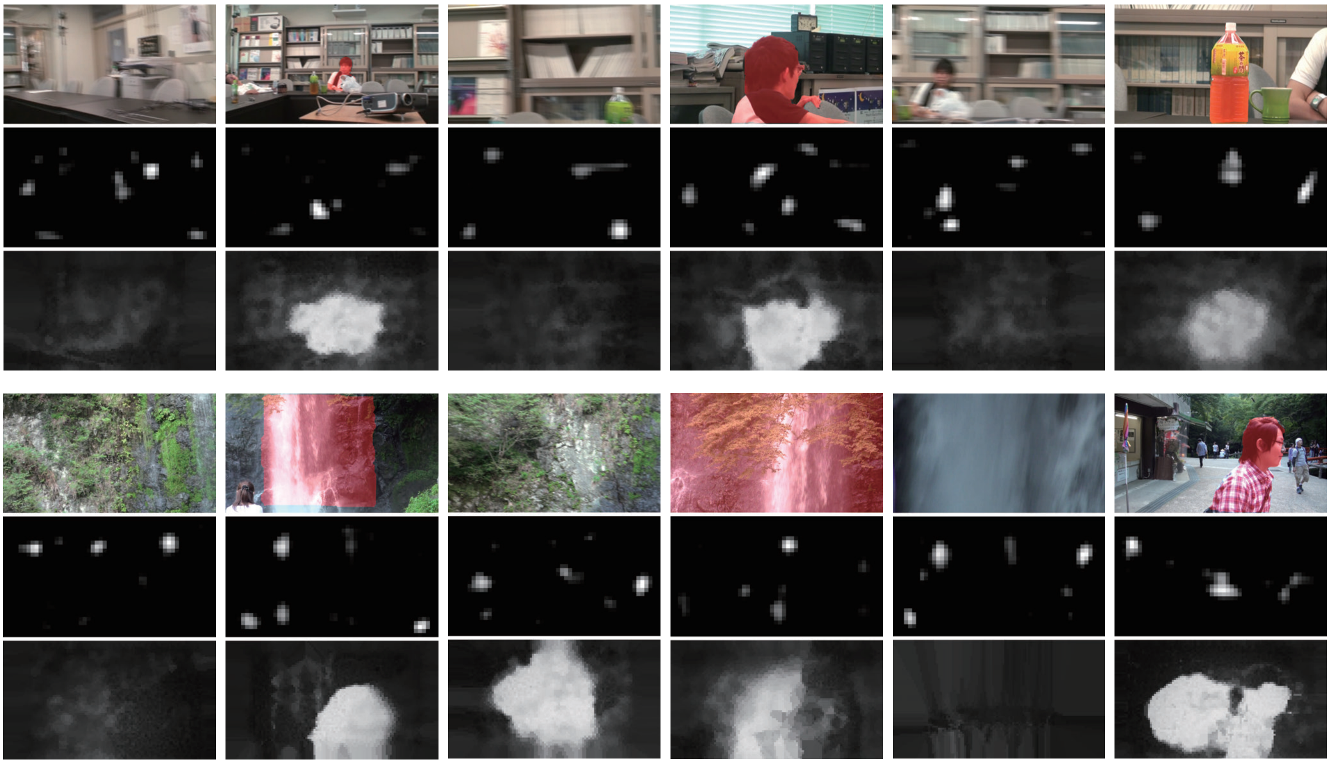

Fig. 5. Example frames. Ground truth ICRs are indicated by pale red regions in top row. Saliency maps and intention maps are given in middle and bottom rows, respectively.

FPR and TPR are defined as

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \quad \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{9}$$

FPR and TPR were calculated via four-fold cross-validation, *i.e.*, we trained the SVM and estimated parameter Σ using 33 videos and calculated FPR and TPR using the remaining 11 videos. Values of T, α, β, and γ were empirically set to 5, 0.2, 5, and 0.8. To show the difference between our ICR model and visual attention models, we generated saliency maps in [7], and calculated FPR and TPR using $\mathbf{x}_{t,i}$'s.

Figure 4 shows the ROC curves for our ICR model (OURS) and the visual attention model (VA). The area under the ROC curves (AUC) were 0.66 for OURS and 0.55 for VA. The low value of VA indicated that the correlation between the visually salient regions and ICRs was low. Our ICR model outperformed the visual attention model, although the performance is not sufficient. The ROC curve for VA was jagged because large values were rare in VA and thus its FPR and TPR were concentrated around $(0, 0)$. Figure 5 shows example frames from two videos in our dataset (the first and fourth rows), the saliency maps (the second and fifth rows), and intention maps (the third and the sixth rows). For intention map generation, we used $K = 3$. To visualize the saliency maps whose values lie in a wide range, we normalized each of them to $[0, 1]$. In these figures, the visual attention model gave responses to visually salient regions (such as a high contrast regions), but such regions do not have a strong correlation with ICRs. Our ICR model, on the other hand, predicted the presence of ICRs except for the third frame in Fig. 5(b). This is because the presence of ICRs are mainly reflected in slow camera motion, and our ICR model successfully captured this cue, although relatively slow camera motion resulted in the failure in the third frame in Fig. 5(b). However, our ICR model did not estimate the shapes of ICRs.

5. CONCLUSION

In this paper, we have introduced a concept of ICRs, which can be viewed as important regions from videographers' perspective, and have proposed to generate an intention map using an ICR model for finding ICRs. Observing that capture intentions are embedded in the camera motion in accordance with object motion, our ICR model uses point trajectories as a cue for distinguishing ICRs from non-ICRs. Our experimental results demonstrated that our ICR model successfully predicted the presence of ICRs, but its performance was still insufficient for estimating the actual shapes of ICRs. Our future work includes determining parameters as they seem to much depend on the video content. Developing new features is another challenge for estimating the shapes of ICRs.

Acknowledgements: This work was partially supported by JSPS KAKENHI 23240024 and 25730115.

6. REFERENCES

[1] L. Itti, “Automatic foveation for video compression using a neurobiological model of visual attention,” *IEEE Trans. Image Processing*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[2] Y.-F. Ma, X.-S. Hua, L. Lu, and H. J. Zhang, “A generic framework of user attention model and its application in video summarization,” *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.

[3] F. Liu and M. Gleicher, “Video retargeting: automating pan and scan,” in *Proc. ACM Int’l Conf. Multimedia*, Oct. 2006, pp. 241–250.

[4] T. Wang, T. Mei, X.-S. Hua, X.-L. Liu, and H.-Q. Zhou, “Video collage: a novel presentation of video sequence,” in *Proc. IEEE Int’l Conf. Multimedia and Expo*, July 2007, pp. 1479–1482.

[5] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua, “Multi-video synopsis for video representation,” *Signal Processing*, vol. 89, no. 12, pp. 2354–2366, Dec. 2009.

[6] P. Wang, J. Wang, G. Zeng, J. Feng, H. Zha, and S. Li, “Salient object detection for searched web images via global saliency,” in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, June 2012, pp. 3194–3201.

[7] L. Itti, C. Koch, and E. Niebur, “A model of saliency based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[8] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, “Learning to detect a salient object,” in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[9] Y. Hu, D. Rajan, and L.-T. Chia, “Attention-from-motion: A factorization approach for detecting attention objects in motion,” *Computer Vision and Image Understanding*, vol. 113, no. 3, pp. 319–331, Mar. 2009.

[10] L. Itti and P. Baldi, “Bayesian surprise attracts human attention,” *Vision Research*, vol. 49, no. 10, pp. 1295–1306, June 2009.

[11] S. Goferman, L. Zelnik-Manor, and A. Tal, “Context-aware saliency detection,” in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, June 2010, pp. 2376–2383.

[12] J. Yang and M.-H. Yang, “Top-down visual saliency via joint CRF and dictionary learning,” in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, June 2012, pp. 2296–2303.

[13] X. Shen and Y. Wu, “A unified approach to salient object detection via low rank matrix recovery,” in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, June 2012, pp. 853–860.

[14] T. Mei, X.-S. Hua, H.-Q. Zhou, and S. Li, “Modeling and mining of users’ capture intention for home video,” *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 66–77, Jan. 2007.

[15] Y. Nakashima, N. Babaguchi, and J. Fan, “Detecting intended human objects in human-captured videos,” in *Proc. IEEE Computer Society Workshop on Perceptual Organization in Computer Vision*, June 2010, 8 pages.

[16] Y. Nakashima, N. Babaguchi, and J. Fan, “Automatically protecting privacy in consumer generated videos using intended human object detector,” in *Proc. ACM Int’l Conf. Multimedia*, Oct. 2010, pp. 1135–1138.

[17] Y. Nakashima, N. Babaguchi, and J. Fan, “Automatic generation of privacy-protected videos using background estimation,” in *Proc. IEEE Int’l Conf. Multimedia and Expo*, 2011, 6 pages.

[18] Y. Nakashima, N. Babaguchi, and J. Fan, “Intended human object detection for automatically protecting privacy in mobile video surveillance,” *Multimedia Systems*, vol. 18, no. 2, pp. 157–173, Mar. 2012.

[19] P. Sand and S. Teller, “Particle video: long-range motion estimation using point trajectories,” in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, June 2006, pp. 2195–2202.

[20] Y. Nakashima and N. Babaguchi, “Extracting intentionally captured regions using point trajectories,” in *Proc. ACM Int’l Conf. Multimedia*, Nov 2011, pp. 1417–1420.

[21] Y. Sheikh, O. Javed, and T. Kanade, “Background subtraction for freely moving cameras,” in *Proc. IEEE Int’l Conf. Computer Vision*, Sept. 2009, pp. 1219–1225.

[22] V. Kolmogorov and R. Zabih, “What energy functions can be minimized via graph cuts?,” *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.