21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Background Subtraction via Early Recurrence in Dynamic Scenes

Xun Shi and John K. Tsotsos
Department of Computer Science & Engineering, and
Centre for Vision Research,
York University, Toronto, Ontario, Canada
{*shixun, tsotsos*}*@cse.yorku.ca*

Abstract

A biologically motivated model of background subtraction is proposed. The two-step computation borrows the idea from the low-level inhibitive processing of the two-pathway primate visual system. A spatiotemporal representation consistent with the dorsal pathway is computed and refined via center-surround inhibition. This representation catches perceptually salient foreground regions, and is further used to inhibit fine-scale visual features that are confined to the ventral pathway, leading to a high-spatially-accurate representation containing mostly foreground pixels. Output of our work is attached to a state-of-the-art visual saliency model. Results using real dynamic scenes are compared with ground truth, which confirmed that our early recurrent processing can effectively remove background.

1. Introduction

Background subtraction refers to a general process of improving the signal response of a target by removing interference of background pixels. It is a fundamental task in various computer vision and image processing applications. Existing approaches attempt to solve this problem using methods from statistics [4], density estimation [8, 11], feature learning [7, 9], etc.

In this paper we borrow the idea of early recurrent processing from the primate visual system. In its multi-layered visual hierarchy, the brain extracts visual features from input through two main visual pathways [16]. The dorsal pathway computes high-temporal-low-spatial frequency visual signal variations, while the ventral pathway responds mostly to low-temporal-high-spatial visual features. These are often coarsely characterized as computations for motion and for high spatially acuity perception respectively. To do background subtraction, we consider two types of inhibitive mechanisms (Fig. 1). During feature extraction, center-surround (**CS**) inhibition [18] plays an essential role in suppressing signal responses of neighborhood activations via lateral connectivity. Further, recent studies [3, 14, 15] suggest that visual features are calculated in the two pathways with different temporal delays, with features computed in the dorsal pathway significantly earlier than those in the ventral pathway. It is thus possible that dorsal activations play a role in inhibiting ventral computation via early recurrent connections between the two pathways. We term this inhibition as early recurrent (**ER**) inhibition. However, these low-level mechanisms have been generally ignored in the literature of computer vision.

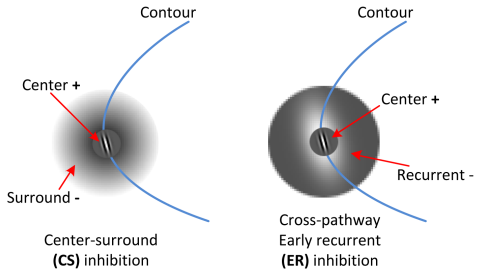

Figure 1. Schematic drawings of the two inhibition mechanisms.

Inspired by the two types of inhibition, a computational model for unsupervised background subtraction is proposed (Fig.2). The model hypothesizes that background in dynamic scenes can be eliminated in two steps. First, spatiotemporal features that are consistent with dorsal analysis are computed. In this representation, activations of foreground and background are mixed. By CS inhibition, a substantial portion of background may be suppressed, leading to a refined spatiotemporal representation containing perceptually salient foreground only. Second, the refined spatiotemporal representation inhibits fine-scale spatial features computed by the ventral pathway via early recurrent connections, such that foreground object features are accurately localized. In its most straightforward man-

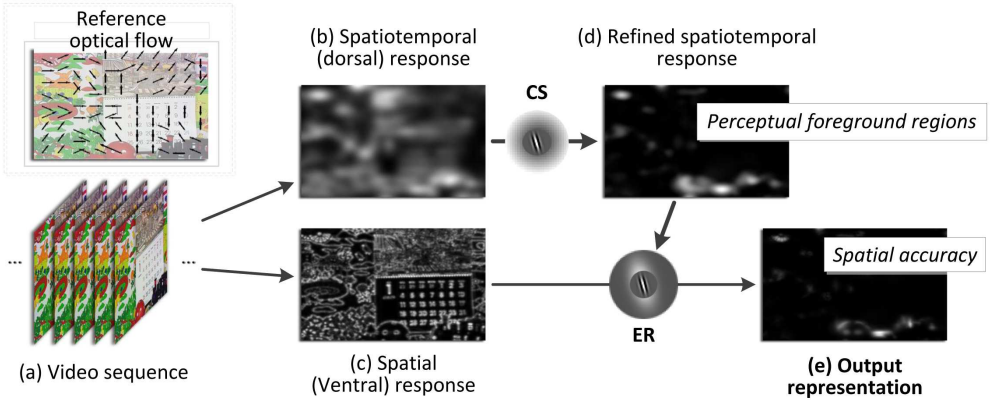

Figure 2. Work-flow of the proposed background subtraction model using Tempete video sequence.

ner, ER inhibition is defined as pixel-wise multiplication.

2. Method of background subtraction

The proposed work simulates the early recurrent processing between the dorsal and the ventral pathway of the primate visual system. The general process can be described as:

$$R = H(E^V \cdot Inh^{ER}), \tag{1}$$

where R denotes output. $H(s) = max(s, 0)$ is a rectification function. E^V denotes energy of ventral (spatial) features. In many existing works, E^V is deemed as output representation. Inh^{ER} denotes ER inhibition between the dorsal pathway and the ventral pathway, which is defined as:

$$Inh^{ER} = H(E^D - \alpha Inh^{CS}), \tag{2}$$

where E^D denotes energy of dorsal (spatiotemporal) features. Inh^{CS} denotes CS inhibition, and α is a constant that weights the CS inhibition.

Formalization of ventral features E^V, dorsal features E^D, CS inhibition Inh^{CS}, and ER inhibition Inh^{ER} are discussed in the rest of this section.

2.1 Visual features computation

Given an image sequence, visual features are extracted by a bank of space-time separable log-Gabor filters. The computation has been shown plausible with physiological properties of the primary visual cortex (V1) [5]. The temporal part of the filter is defined in frequency domain as:

$$F^t(w) = \exp\{\frac{-\log(w/w_0)^2}{2\log(\sigma_t/w_0)^2}\}, \tag{3}$$

where w_0 denotes center frequency. Using different w_0, one can construct a multi-scale representation. σ_t denotes filter bandwidth. $\sigma_t = 0.65w_0$ is used to represent 1.4 octaves. The spatial part of the log-Gabor filter is defined in frequency domain as:

$$F^s(u, v) = \exp\{\frac{-log(u_1/u_0)^2}{2log(\sigma_u/u_0)^2}\} \cdot \exp\{\frac{-v_1^2}{2\sigma_v^2}\}, \tag{4}$$

where $u_1 = u\cos(\theta) + v\sin(\theta)$, $v_1 = -u\sin(\theta) + v\cos(\theta)$, θ denotes orientation, u_0 denotes central spatial frequency, σ_u and σ_v denote spatial bandwidth along u and v axis. In our work $\sigma_u = \sigma_v = 0.55u_0$.

To calculate feature energy, we follow [1] to use a quadrature technique, which computes the square root over filter output that are 90 degrees out of phase as:

$$E_\theta(x, y, t) = \sqrt{S_\theta(x, y, t)^2 + S_{\theta+\frac{\pi}{2}}(x, y, t)^2}, \tag{5}$$

where S_θ denotes filter response of orientation θ.

To compute dorsal and the ventral features, (w, u, v) are varied to catch their neurophysiological properties: dorsal filters are set to high-temporal-low-spatial frequencies (Fig.2b), and ventral filters are set to high-spatial frequency variations (Fig.2c).

2.2 Center-surround inhibition (Inh^{CS})

The center-surround inhibition, Inh^{CS} is defined in an anisotropic manner to self-inhibit the dorsal representation. The process is formalized as a convolution of dorsal energy E^D_θ with a weighting function as:

$$Inh^{CS}_\theta(x, y) = E^D_\theta(x, y) * w^D(x, y), \tag{6}$$

$$w^D(x, y) = \frac{H(DoG_\sigma(x, y))}{\|H(DoG_\sigma(x, y))\|_1}, \tag{7}$$

where $DoG_\sigma(x, y)$ denotes the surround strength of Difference of Gaussian, and $\|.\|_1$ denotes L1 norm. σ_c is the center bandwidth, which is set the same value with σ_u defined in Eq.(4). σ_s denotes the surround bandwidth. We set $\sigma_s = 4\sigma_u$ following [10].

The result of CS inhibition is a refined dorsal energy representation that can perceptually catch image foreground (Fig.2d). Due to low-spatial frequency response profile, this representation lacks precise spatial-acuity.

2.3 Early recurrent inhibition (Inh^{ER})

Via early recurrent connections, the refined dorsal representation, Inh^{ER} inhibits ventral feature E_V. The

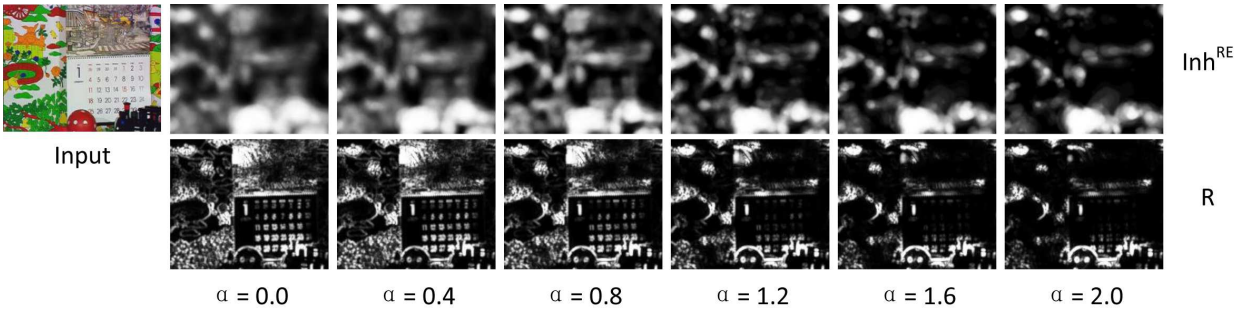

Figure 3. Center-surround inhibition weighting parameter α. Motion patterns included in the input: camera motion (leftward), calendar motion (upward), ball motion(leftward) and train motion (leftward).

inhibition improves feature representation by suppressing inconsistent responses. Based on experiments of orientation-selective neurons [12, 6], we hypothesize the inhibition operates as pixel-wise multiplication:

$$R_\theta(x,y) = H(E^V_\theta(x,y) \cdot Inh^{ER}_\theta(x,y)), \quad (8)$$

where $E^V_\theta(x,y)$ denotes ventral energy extracted by log-Gabor filter of orientation θ. $R_\theta(x,y)$ denotes the output inhibited energy, a representation that has high-spatial feature acuity (Fig.2e). This representation is then ready to be used by higher-level analysis.

3. Experiment

We investigated the effect of center-surround inhibition and early recurrent inhibition in subtracting background. We used real video sequences from [7], which have been widely used for background subtraction applications in the literature. To catch the biolgocial consistency, we follow [14] to set the parameters. For dorsal spatiotemporal features, 4 temporal frequency bands $w_0 \in (3.0, 6.3, 13.2, 27.8)$ and 4 spatial frequency bands $u_0 \in (3.0, 6.3, 13.2, 27.8)$ are used for Eq.(3) and Eq.(4) respectively. For ventral spatial features, 2 fequency bands are used $u_0 \in (3.0, 6.3)$ for high-spatial frequency signal variations for Eq.(4).

3.1 Center-surround inhibition weighting

Fig. 3 illustrates the effect of CS inhibition by setting parameter α to different values in Eq.(2). The first row represents the overall inhibition strength from the dorsal pathway Inh^{ER}, and the second row shows the result of using Inh^{ER} to inhibit ventral features as Eq.(1). When α increases, background pixels (wallpaper and calendar) tend to fade out gradually, while foreground (ball and train) remains mostly unchanged. If continue increasing α, strengths of foreground may also be suppressed. One may easily notice that Inh^{ER} contains targets that are perceptually salient. However, compared with ventral representation, Inh^{ER} is coarse.

3.2 Performance measured by visual saliency

To quantitatively evaluate effect of background subtraction, output of Eq.(1) is attached to a state-of-the-art model, AIM [2] to compute visual saliency. The goal is to determine whether feature maps refined by Inh^{ER} and Inh^{ER} lead to improved saliency representations. It is thus natural to deem AIM output based on original feature maps to provide baseline performance.

Real scene sequences (with ground truth masks) [7] are used. This dataset contains a multitude types of background and spatiotemporal variations, which has been widely used in background subtraction studies.

Fig. 4 compares different output saliency representations. Right column of each video illustrates saliency maps from top to bottom: based on original features (**AIM**), based on features modulated by ER inhibition only (**AIM+ER**), and based on features modulated by both CS and ER inhibitions (**AIM+ER+CS**). High intensity values indicate high saliency. It is clearly shown that there are more similarities between ground truth (top-middle drawing) and salient regions computed by **AIM+ER+CS** than by the other two algorithms.

Saliency performance is measured by mean receiver operating characteristic (ROC) curves over all frames for each sequence. Given ground truth masks, the curve is defined as true positive rate versus false positive rate. It is clearly shown in Fig. 4 that curves produced by **AIM+ER+CS** augment the other two cases significantly. Area under curve (AUC) is calculated. ER+CS inhibitions raise AUC in most tests, which further confirms that the early recurrent interaction is a generally effective method in background subtraction.

4. Conclusion

In this paper, a novel approach of unsupervised background subtraction in dynamic scenes has been proposed, which is inspired by the early recurrent processing of the primate visual system. Representation computed by the dorsal pathway is perceptually consistent with foreground, and representation computed by the ventral pathway, on the other hand, is a high-spatially accurate description with mixed foreground and background variations. The model defines two types of in-

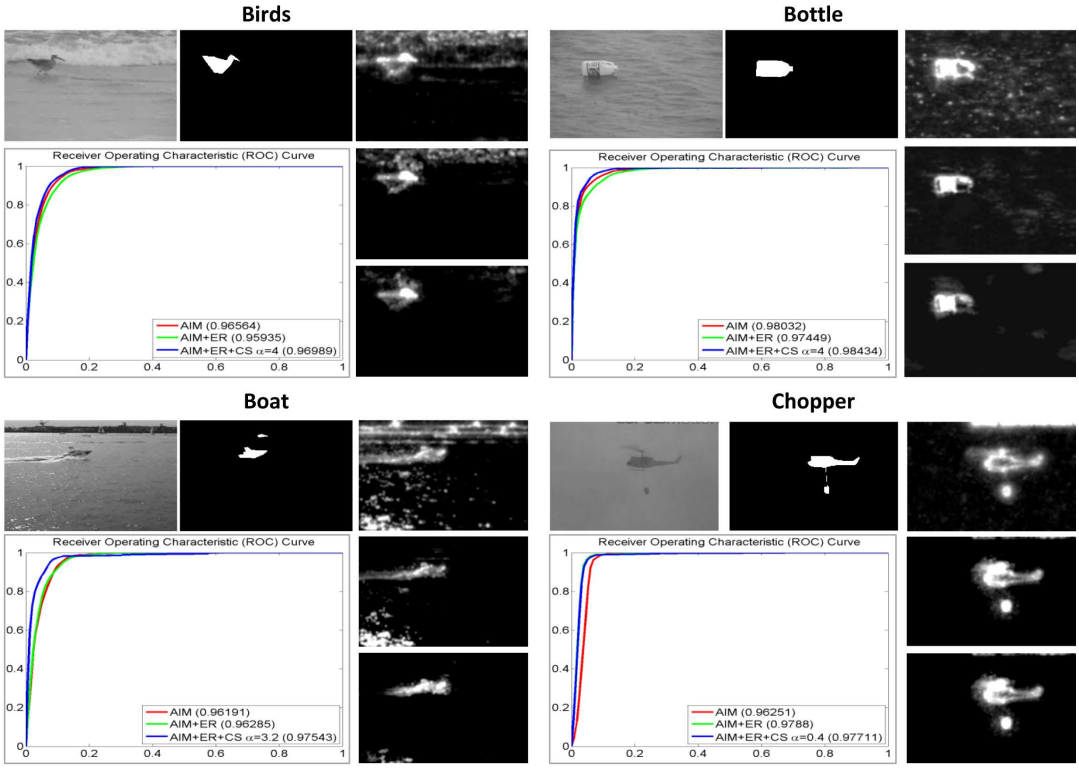

Figure 4. Saliency representation computed by AIM. For each sequence, figures in clock-wise order: input, ground truth, original AIM saliency, AIM+ER, and AIM+ER+CS. Also shown are the mean ROC curves over all frames for each sequence. Area under curve is displayed in the bracket of the legend.

hibition, center-surround (CS) inhibition and (ER) early recurrent inhibition, which improve ventral feature representation by inhibiting its responses to background.

Using a saliency model, we quantitatively evaluated the performance. Results using real scenes clearly conclude that the proposed work is a robust and generally applicable process. In this proposal, we have been mainly focused on applying the inhibitions for background subtraction, possibilities of using this approach in other applications for example object recognition [17] and image retrevial [13] may exist for future work.

References

[1] E. H. Adelson and J. R. Bergen. *J. Opt. Soc. Am. A*, 2(2):284–299, 1985.

[2] N. D. B. Bruce and J. K. Tsotsos. *J. Vision*, 9(3), 2009.

[3] J. Bullier. *Brain Res. Rev.*, 36(2-3):96–107, 2001.

[4] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. *IEEE TPAMI.*, 25(10):1337–1342, 2003.

[5] D. J. Field. *J. Opt. Soc. Am. A*, 4(12):2379–2394, 1987.

[6] F. Gabbiani, H. G. Krapp, C. Koch, and G. Laurent. *Nature*, 420(6913):320–4+, 2002.

[7] D. Gao, V. Mahadevan, and N. Vasconcelos. *J. Vision*, 8(7), 2008.

[8] B. Han, D. Comaniciu, Y. Zhu, and L. Davis. *IEEE TPAMI*, 30(7):1186 –1197, 2008.

[9] B. Han and L. S. Davis. *IEEE TPAMI*, 34:1017–1023, 2012.

[10] E. Kaplan, S. Marcus, and Y. T. So. *J. Physiol.*, 294(1):561–580, 1979.

[11] D.-S. Lee. *IEEE TPAMI*, 27(5):827–832, 2005.

[12] C. J. McAdams and J. H. R. Maunsell. *J. Neurosci.*, 19(1):431–441, 1999.

[13] D. Nistér and H. Stewénius *CVPR*, 2006.

[14] X. Shi, N. D. B. Bruce and J. K. Tsotsos. *CVPRW*, 2011.

[15] X. Shi, and J. K. Tsotsos. *CRV*, 2012.

[16] L. G. Ungerleider and M. Mishkin. *Two Cortical Visual Systems*, Chapter 18:549–586. 1982.

[17] B. Wang, X. Bai, X. Wang, W. Liu, and Z. Tu. *ECCV'10*, V:15–28. 2010.

[18] C. Zeng, Y. Li, and C. Li. *NeuroImage*, 55(1):49–66, 2011.