15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Saliency/non-saliency Segregation in Video Sequences Using Perception-based Local Ternary Pattern Features

K. L. Chan
Department of Electronic Engineering, City University of Hong Kong
83 Tat Chee Avenue, Kowloon, Hong Kong
itklchan@cityu.edu.hk

Abstract

The detection of salient objects in video sequence is an active research area of computer vision. One approach is to perform joint segmentation of objects and background in each image frame of the video. The background scene is learned and modeled. Each pixel is classified as background if it matches the background model. Otherwise the pixel belongs to a salient object. The segregation method faces many difficulties when the video sequence is captured under various dynamic circumstances. To tackle these challenges, we propose a novel perception-based local ternary pattern for background modeling. The local pattern is fast to compute and is insensitive to random noise, scale transform of intensity. The pattern feature is also invariant to rotational transform. We also propose a novel scheme for matching a pixel with the background model within a spatio-temporal domain. Furthermore, we devise two feedback mechanisms for maintaining the quality of the result over a long video. First, the background model is updated immediately based on the background subtraction result. Second, the detected object is enhanced by adjustment of the segmentation conditions in proximity via a propagation scheme. We compare our method with state-of-the-art background/foreground segregation algorithms using various video datasets.

1. Introduction

The detection of salient objects in video has found many applications. Mahadevan and Vasconcelos [1] proposed a center-surround framework for saliency detection. Salient objects are detected via background subtraction. Background pixels are identified when features estimated from the center and surround windows are indiscernible. Tang *et al.* [2] did not perform background modeling. Moving objects are directly detected by clustering of salient motion points with spatial kinetic mixture of Gaussian model. We also formulate the saliency/non-saliency segregation as a background subtraction problem. Background subtraction has various advantages. To learn the background model from the video sequence can result in robust object detection. The detection of salient objects is treated as the complement of background subtraction. This concept correlates well with biological vision. The background, although seen, is ignored so that focus is on the moving targets.

In background subtraction, pixels in each image frame are identified as background if they are similar to the background model. The pixels that are not similar are classified as foreground (saliency). A background subtraction framework contains background modeling, joint background/foreground classification, and background model updating. A recent survey can be found in [3]. The data-driven background subtraction relies on image cues for background representation. State-of-the-art algorithms employ foreground modeling or feedback mechanism to improve moving object detection.

Background representation – The background scene can be represented by statistical model. Pixelwise background color can be modeled by mixture of Gaussians (MOGs) [4]. Zivkovic [5] proposed an algorithm for selecting the number of Gaussian distributions using the Dirichlet prior. Alternatively, Elgammal *et al.* [6] estimated the pdf directly from previous pixels using kernel estimator. Recently, more researches employ background model generated using real observed pixel values. Barnich *et al.* [7] proposed a fast sample-based background subtraction algorithm called ViBe. Background model is initialized by randomly sampling of pixels on the first image frame. Pixel of the new image frame is classified as background when there are sufficient background samples similar to the new pixel. Hofmann *et al.* [8] proposed a similar sample-based method with more tunable parameters. Recent researches show that modeling background by local patterns can achieve higher accuracy. Heikkilä and Pietikäinen [9] proposed to model the background of a pixel by local binary pattern (LBP) histograms estimated around that pixel. Liao *et al.* [10] proposed the scale invariant local ternary pattern (SILTP) which can tackle illumination variations.

Foreground enhancement – Saliency detection can be very difficult under various complex circumstances. Background motions can produce false positive error. Foreground detection can be improved via background model updating or specific foreground model. Many background subtraction methods like [4] update parameters of matched background model with a fixed learning factor. In [8], the foreground decision threshold and model update rate can be adaptively adjusted along the video sequence. In [7], a random policy is employed for updating the background model at the pixel location and its neighbor. Van Droogenbroeck and Paquot [11] inhibited the update of neighboring background model across the background-foreground border. Kim *et al.* [12] proposed a PID tracking control system for foreground segmentation refinement. In [13], MOGs are used to model the color distribution of swimmer pixels. Sheikh and Shah [14] presented a non-parametric density estimation method to model foreground.

2. Perception-based Local Ternary Pattern

A pattern, with multiple pixels, can characterize the local texture more effectively than individual pixel. Biological vision can perceive saliency by local feature contrast. With this concept, we propose a novel perception-based local ternary pattern (P-LTP) which characterizes each pixel based on the perceptual differences with its neighbors. Figure 1 shows a block of 3 x 3 pixels. Each pixel of the block, n_1 to n_8, is compared with the center pixel b. The confidence interval CI of b is defined by (CI_l, CI_u) where CI_l and CI_u are the lower bound and upper bound of CI respectively. If a neighboring pixel n has color within the CI of b, its pattern value t is set equal to 0. If it is above or below CI, its pattern value is set equal to 1 or -1 respectively.

$$t_k = \begin{cases} 0, & CI_l \le n_k \le CI_u \\ 1, & n_k > CI_u \\ -1, & n_k < CI_l \end{cases}, 1 \le k \le 8 \qquad (1)$$

n_1	n_2	n_3
n_4	b	n_5
n_6	n_7	n_8

→

t_1	t_2	t_3
t_4		t_5
t_6	t_7	t_8

Figure 1. Formation of ternary pattern.

The confidence interval of a pixel having a color value b is defined as ($b - d_1$, $b + d_2$). According to Weber's law [15], d_1 and d_2 depend on the perceptual characteristics of b. That is, they should be small for darker color and large for brighter color. Therefore, the confidence interval is defined as ($b - c_1b$, $b + c_2b$). Using peak signal-to-noise ratio (PSNR) measure, b and $b - c_1b$ are just perceptually different from each other if

$$20log_{10}\frac{I_{max}}{b-c_1b} - 20log_{10}\frac{I_{max}}{b} = T_p \qquad (2)$$

where I_{max} is the maximum intensity and T_p is the perceptual threshold. Similarly, b and $b + c_2b$ are just perceptually different from each other if

$$20log_{10}\frac{I_{max}}{b} - 20log_{10}\frac{I_{max}}{b+c_2b} = T_p \qquad (3)$$

Motion picture experts group committee [16] recommends a difference of PSNRs at least 0.5 dB as distinguishable. In our initial background modeling, the perceptual threshold is raised. Assume T_p is 1.0 dB, c_1 = 0.1087 and c_2 = 0.1220.

(a)

63	68	42
64	64	27
61	95	83

→

0	1	0
1		0
0	1	1

65	69	42
63	64	27
60	95	83

→

1	1	0
0		0
0	1	1

130	138	84
126	128	54
120	190	166

→

1	1	0
0		0
0	1	1

(b)

63	68	42
64	64	27
61	95	83

→

0	0	-1
0		-1
0	1	1

65	69	42
63	64	27
60	95	83

→

0	0	-1
0		-1
0	1	1

130	138	84
126	128	54
120	190	166

→

0	0	-1
0		-1
0	1	1

Figure 2. Formation of local pattern: (a) LBP, (b) P-LTP.

Figure 2(a) illustrates the formation of a conventional LBP. The first row shows the formation of LBP for a noise-free image. The second row indicates that LBP is not robust to additive random noise. The third row also shows that LBP cannot keep its invariance against scale transform of intensity. Figure 2(b) illustrates the formation of P-LTP under the same circumstances. It can be seen that P-LTP is robust against random noise and scale transform.

3. Background Model Initialization

A short image sequence is used to generate the initial background model. The pixelwise background model contains two bags of samples. At a given pixel location, colors of all the temporal samples are entered into the background model for that pixel location. Also a block is defined in each initialization image frame centered at that pixel location. The block of pixels is transformed into local ternary pattern. Features are computed from this local pattern. All features, estimated from the spatio-temporal domain, are also entered into the background model for that pixel location.

Temporal color samples – We used invariant color features to represent the color of the pixel. First, we collect temporal samples represented by the normalized color model as defined by [17]. Normalized color varies with a change in object's material and highlights. Our method can correctly classify shaded object region shadow and shadow cast on background. It should be noted that normalized color is indistinguishable along the grey scale of the RGB color space. It is also unstable near the black vertex where normalized color is undefined. Therefore, we also collect the temporal RGB samples. In the background subtraction process, our method will automatically shift to use the temporal RGB samples when {|R – G|, |G – B|, |R – G|} < threshold. The threshold is fixed as 10% of the range of color component value which is 26.

Spatio-temporal local pattern samples – As the computation load for spatio-temporal samples is higher than that of the temporal samples, we need to consider the sampling domain carefully in order to strike a balance between robustness of background model and computation load. We have done experimentation and find that a block size of 3 x 3 pixels is most suitable. In initialization, a very short image sequence may not sufficiently capture the dynamic information of the background while a very long image sequence will demand long computation time and more storage space. We fixed the number of initialization image frames as 30.

127	25	32
118	30	29
40	39	43

→

1	-1	0
1		0
1	1	1

→ 4

Figure 3. Computation of P-LTP feature.

At a given pixel location, a block of pixels is defined. For each color component, a ternary local pattern P-LTP is formed. The pattern codes are summed to form one feature value for the center pixel. Figure 3 illustrates the estimation of one feature value with numbers corre-

sponding to magnitudes of one color component. A homogeneous block will have very small feature values while a block with neighboring pixels perceptually different from the center pixel will have large positive or negative feature values. Therefore, the feature values characterize the texture pattern. Moreover, the pattern can be represented by small amount of feature values (3 for three color components). In contrast, histogram representation will demand more memory for a pattern with p neighboring pixels in the order of 2^p. Liao *et al.* [10] proposed a scale invariant local ternary pattern (SILTP). However, SILTP feature is not rotation invariant. The pattern is not perception-based. The 2 thresholds are fixed as $\pm 0.05b$. The P-LTP feature, as shown in Figure 3, is rotation invariant. The range of P-LTP feature is 17 x 17 x 17 = 4913, which is much larger than that of 256 in histogram representation.

4. Background-Foreground Segregation

Saliency in video is detected by our perception-based saliency detection (PSD). If all feature values of the new pixel match with some temporal color samples or spatio-temporal local binary pattern features of the background model, the pixel is labeled as background. Otherwise, it is labeled as foreground. We propose a novel scheme to estimate the similarity between the pixel and the background model which strikes for balance between efficiency and perceptual accuracy. First, the pixel is compared with the temporal color samples of the background model. The perception-based confidence interval of the new pixel is defined. If two temporal color samples in the background model are found within that confidence interval, the new pixel is labeled as background. In static scene, the background subtraction can be accomplished quickly by this process. In dynamic scene, it may not be possible to find similar color samples. Then, the pixel is compared with the spatio-temporal P-LTP features in the background model. A block with the new pixel at the center is defined. P-LTP feature values for this pixel are calculated using the same method as mentioned in the previous section. Features of the pixel are compared with the features in the background model. We have done experimentation and define a spatio-temporal search space of 7 x 7 pixels x 30 frames centered at the new pixel location. Two patterns are considered similar if the absolute difference of their feature values is <= tolerance. If two patterns in the background model match with the local pattern of the new pixel, the pixel is labeled as background. Otherwise, the pixel is labeled as foreground.

5. Foreground Enhancement

In the background model updating, the total number of color samples and P-LTP features will remain the same. If the new pixel matches with the temporal color features, one temporal color sample will be updated by the following equation

$$c_b^{new} = (1-\alpha_b) c_b^{old} + \alpha_b c_p \qquad (4)$$

where c_p is the color of the new pixel, c_b is the matched temporal color. The updating factor α_b is inversely related to the history of that background sample.

$$\alpha_b = \frac{1}{history_b} \qquad (5)$$

$$history_b = \begin{cases} history_b + 1, & if\ matches \\ history_b, & otherwise \end{cases} \qquad (6)$$

If the P-LTP features of the new pixel match with the P-LTP features of the background model, the P-LTP features will be updated by the following equation

$$f_b^{new} = (1-\alpha_b) f_b^{old} + \alpha_b f_p \qquad (7)$$

where f_p is the feature value of the new pixel, f_b is the matched feature in the background model.

The detected foreground often suffers from distorted shape and holes. To remedy these problems, the confidence interval of the foreground pixel and its neighbors is tightened. Initially, each pixel position has the confidence interval computed by equations (2) and (3) with perceptual threshold T_p equal to 1.0 dB and is saved as *CI* map. Assume that a foreground pixel is likely to have foreground neighbors. The *CI* map is updated by a propagation scheme to adjust the confidence interval of the foreground pixel and its neighbors with T_p reduced to 0.5 dB.

6. Result

We evaluated and compared the performance of PSD with two state-of-the-art background subtraction algorithms ViBe [7] and SILTP [10]. Based on sample consensus, ViBe can achieve very good results with very few tunable parameters. Moreover, the way temporal color samples being used in PSD is similar to ViBe. The comparison of PSD with ViBe demonstrates the significance of the P-LTP features in background modeling. SILTP employs scale invariant local patterns. The comparison of PSD with SILTP can demonstrate the improvement of scale and rotation invariant P-LTP features and spatio-temporal search space in tackling dynamic scenes. All methods are evaluated with a fixed setting and no post-processing on 3 datasets: Wallflower [18], Star [19], ChangeDetection.net [20]. Table 1 shows the F-measure (F1) results on the Wallflower dataset. The best result in a given row is highlighted. PSD can achieve highest F1 on Camouflage and WavingTrees. Overall, SILTP achieves the highest average F1, probably because the image frame size is small. Table 2 shows the F1 results on the Star dataset. PSD can achieve highest F1 on 5 image sequences. Overall, PSD achieves the highest average F1 than ViBe and SILTP. Table 3 shows the weighted average F1 results on the ChangeDetection.net dataset. The videos contain complex backgrounds with larger image frame size. PSD can achieve highest F1 on 5 categories. Overall, PSD achieves the highest average F1 than ViBe and SILTP. The results of SILTP are lower than [10] because no post-processing is applied. Due to page limit, Figure 4 shows some visual results.

Table 1. F1 results on the Wallflower dataset.

Sequence	PSD	ViBe	SILTP
Bootstrap	0.426	0.478	**0.683**
Camouflage	**0.942**	0.931	0.921
ForegroundAperture	0.635	0.644	**0.837**
LightSwitch	0.559	0.159	**0.715**

TimeOfDay	0.085	**0.394**	0.173
WavingTrees	**0.953**	0.933	0.686
Average	0.600	0.590	**0.669**

Table 2. F1 results on the Star dataset.

Sequence	**PSD**	**ViBe**	**SILTP**
AirportHall	0.553	0.496	**0.566**
Bootstrap	0.503	0.514	**0.519**
Curtain	**0.833**	0.775	0.687
Escalator	0.378	**0.445**	0.267
Fountain	**0.535**	0.425	0.237
ShoppingMall	**0.620**	0.522	0.566
Lobby	0.234	0.029	**0.509**
Trees	**0.637**	0.345	0.099
WaterSurface	**0.866**	0.801	0.333
Average	**0.573**	0.483	0.420

Table 3. F1 results on the ChangeDetection.net dataset.

Category	**PSD**	**ViBe**	**SILTP**
baseline	**0.883**	0.874	0.415
dynamic background	**0.517**	0.364	0.031
camera jitter	**0.630**	0.575	0.186
intermittent object motion	**0.580**	0.532	0.397
shadow	0.561	**0.781**	0.366
thermal	**0.710**	0.610	0.282
Average	**0.632**	0.601	0.346

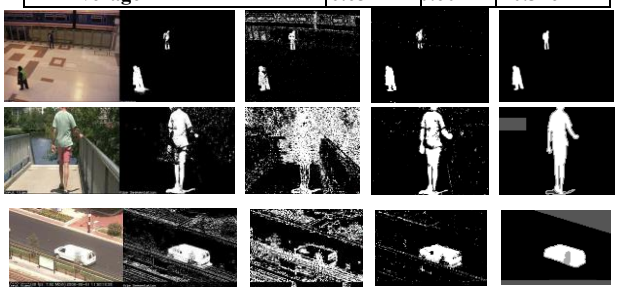

Figure 4. Background subtraction results on the ChangeDetection.net dataset (baseline, dynamic background, camera jitter) – original image frames (first column), results obtained by ViBe (second column), results obtained by SILTP (third column), results obtained by PSD (fourth column), ground truths (last column).

7. Conclusion

We propose a method for saliency detection in video. The non-salient background is modeled by perception-based color and local ternary pattern features. The P-LTP feature is robust to random noise and invariant to scale and rotation transforms. In background-foreground segregation, each pixel of the current image frame is classified as background if it matches with the background model in the spatio-temporal search space. Otherwise, the pixel is classified as foreground. PSD can produce much less false positive errors whilst also keeping false negative errors low. The background model and the segmentation condition are adaptive in order to enhance the saliency detection over a long video sequence.

References

[1] Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. IEEE Trans. on Pattern Analysis and Machine Intelligence Vol. 32, No. 1, 171-177 (2010)

[2] Tang, P., Gao, L., Liu, Z.: Salient moving object detection using stochastic approach filtering. Proc. ICIG 530-535 (2007)

[3] Sobral, A., Vacavant, A.: A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. Computer Vision and Image Understanding Vol. 122, 4-21 (2014)

[4] Stauffer, C., Grimson, W.E.L.: Learning patterns of activity using real-time tracking. IEEE Trans. on Pattern Analysis and Machine Intelligence Vol. 22, No. 8, 747-757 (2000)

[5] Zivkovic, Z.: Improved adaptive Gaussian mixture model for background subtraction. Proc. ICPR 28-31 (2004)

[6] Elgammal, A., Duraiswami, R., Harwood, D., Davis, L.S.: Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. Proc. of IEEE Vol. 90, No. 7, 1151-1163 (2002)

[7] Barnich, O., Van Droogenbroeck, M.: ViBe: a powerful random technique to estimate the background in video sequences. Proc. ICASSP 945-948 (2009)

[8] Hofmann, M., Tiefenbacher, P., Rigoll, G.: Background segmentation with feedback: the Pixel-Based Adaptive Segmenter. Proc. CVPR 38-43 (2012)

[9] Heikkilä, M., Pietikäinen, M.: A texture-based method for modeling the background and detecting moving objects. IEEE Trans. on Pattern Analysis and Machine Intelligence Vol. 28, No. 4, 657-662 (2006)

[10] Liao, S., Zhao, G., Kellokumpu, V., Pietikäinen, M., Li, S.Z.: Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. Proc. CVPR 1301-1306 (2010)

[11] Van Droogenbroeck, M., Paquot, O.: Background subtraction: experiments and improvements for ViBe. Proc. CVPR 32-37 (2012)

[12] Kim, S.W., Yun, K., Yi, K.M., Kim, S.J., Choi, J.Y.: Detection of moving objects with a moving camera using non-panoramic background model. Machine Vision and Applications Vol. 24, 1015-1028 (2013)

[13] Eng, H.-L., Wang, J., Siew Wah, A.H.K., Yau, W.-Y.: Robust human detection within a highly dynamic aquatic environment in real time. IEEE Trans. on Image Processing Vol. 15, No. 6, 1583-1600 (2006)

[14] Sheikh, Y., Shah, M.: Bayesian object detection in dynamic scenes. Proc. CVPR Vol. 1, 74-79 (2005)

[15] Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Pearson/Prentice Hall (2010)

[16] Salomon, D.: Coding for Data and Computer Communications, Springer (2005)

[17] Gevers, T., Smeulders, A.W.M.: Color based object recognition. Pattern Recognition Vol. 32, 453-464 (1999)

[18] Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: principles and practice of background maintenance. Proc. ICCV 255-261 (1999)

[19] Li, L., Huang, W., Gu, I.Y.-H., Tian, Q.: Statistical modelling of complex backgrounds for foreground object detection. IEEE Trans. on Image Processing Vol. 13, No. 11, 1459-1472 (2004)

[20] Goyette, N., Jodoin, P.-M., Porikli, F., Konrad, J., Ishwar, P.: Changedetection.net: a new change detection benchmark dataset. Proc. CVPR 16-21 (2012)