

EFFECTIVE MULTI-RESOLUTION BACKGROUND SUBTRACTION

LingFeng Wang, ChunHong Pan

NLPR, Institute of Automation, Chinese Academy of Sciences, {lfwang,chpan}@nlpr.ia.ac.cn

ABSTRACT

In this paper, we propose a novel multi-resolution background subtraction method. We adopt coarse to fine strategy, which is the essence the multi-resolution scheme, to obtain the foreground mask. The rough mask is first gained relied on the *Single Gaussian Model*, which holds minor computation cost. Then, the slightly accuracy mask is calculated by the *Saliency-based Extraction Model*, which contains high accuracy and stability. Finally, *Contour-based Refining Model* is used to refine the mask edge. Our algorithm is evaluated against several video sequences, and experimental results show that the proposed method is suitable for various scenes and is appealing with respect to robustness.

Index Terms— multi-resolution, background subtraction, *Single Gaussian Model*, *Saliency-based Extraction Model*, *Contour-based Refining Model*

1. INTRODUCTION

Background subtraction is a convenient and effective method for detecting foreground objects from a stationary camera. Its mainly depends on the background modeling module. The central idea behind this module is to utilize the visual properties of the scene for building an appropriate representation, that can then be used to classify a new observation as foreground or background.

Existing methods for background modeling can be classified as **predictive** and **statistical**. The **predictive** methods model the scene as a time series and develop a dynamical model to recover the current input based on past observations [1, 2], while the **statistical** methods neglect the order of the input observations and roughly build a probabilistic representation of the observations [3, 4, 5, 6, 7, 8, 9]. A popular **statistical** method is to model each background pixel with a single Gaussian distribution [3]. However, This method does not work well in the case of dynamic natural environments including repetitive motions, i.e. waving vegetation, rippling water, and camera jitter. In [4], the mixture of Gaussians(MoG) approach is proposed to solve these complex, non-static backgrounds. Unfortunately, background with fast variations can not be accurately modeled by just a few Gaussians. To overcome the limitations of parametric methods, i.e. single Gaussian in [3], MoG in [4], a nonparametric technique is developed in [5]. This utilizes a general nonparametric kernel density estimation technique for building a statistical representation of the scene background. [6] uses a codebook to construct a compressed background model. However, both the parametric method [4] and nonparametric method [5, 6] may fail when foreground objects have similar color to background, or even when the illumination variations occur due to sunlight changing outdoor and light switching indoor. The main reason is that, these methods only use the pixel color or intensity information to detect foreground objects. To deal

with this weak description, [7] use a novel and powerful approach based on discriminative texture features represented by **LBP** [10] to capture background statistics. In this paper, each pixel is modeled as a group of adaptive **LBP** histograms that are calculated over a circular region around the pixel. The main limitation of this method is that both memories and computation costs increase greatly with the increasing of the images resolution.

In this paper, we propose a novel multi-resolution, coarse to fine strategy, background subtraction method, which compromises the computation cost with the algorithm robustness. In low-resolution module, we adopt the *Single Gaussian Model* to gain the rough mask. The motivation of using this model relies on the low computation cost. In middle-resolution module, we apply *Saliency-based Extraction Model* to calculate the slightly accurate mask. The motivation of using this model dues to the accuracy of the calculation as well as the scalability of the algorithm. In this paper, we utilize three types of saliency: i.e. intensity saliency, shadow saliency, and contour saliency, which cover contour and texture information. In high-resolution module, we use the *Contour-based Refining Model* to refine the mask edge.

2. MULTI-RESOLUTION BACKGROUND SUBTRACTION

As illustrated in Fig.1, a Gaussian pyramid \mathcal{G} (\mathcal{G}^i is the i -th level) is first constructed for each input frame \mathcal{I} by using the low-pass down-sampling operation

$$\mathcal{G}^i = \downarrow 2[\mathcal{G}^{i-1} \otimes g] \quad (1)$$

where $\downarrow 2[\cdot]$ is the down-sampling operation, \otimes is the convolution operation, g is the two dimensional Gaussian kernel, and $\mathcal{G}^0 = \mathcal{I}$, the original input frame. In this paper, we construct three Gaussian levels specially, which are the low-resolution gray-level layer $\mathcal{G}^l = \mathcal{G}^2$, the middle-resolution color layer $\mathcal{G}^m = \mathcal{G}^1$, and the high-resolution gray-level layer $\mathcal{G}^h = \mathcal{G}^0$. *Single Gaussian Model* is applied on \mathcal{G}^l to generate the low-resolution mask \mathcal{M}^l . Then, the initial middle-resolution mask $\widehat{\mathcal{M}}^m$ is constructed from \mathcal{M}^l by using the nearest up-sampling operation,

$$\widehat{\mathcal{M}}^m = \uparrow 2[\mathcal{M}^l] \quad (2)$$

where $\uparrow 2[\cdot]$ is the up-sampling operation. Based on \mathcal{G}^m and $\widehat{\mathcal{M}}^m$, *Saliency-based Extraction Model* is performed to obtain the middle-resolution mask \mathcal{M}^m . Same as the calculation of $\widehat{\mathcal{M}}^m$, the initial high-resolution mask $\widehat{\mathcal{M}}^h$ is gained from \mathcal{M}^m by using the nearest up-sampling operation. Finally, *Contour-based Refining Model* is utilized to calculate the high resolution mask \mathcal{M}^h , which equals to the final output mask \mathcal{M} .

2.1. Single Gaussian Model

It is often difficult or infeasible to get a sufficiently long video clip without any foreground objects appearing in the scene. To overcome

This work was supported by the National Natural Science Foundation of China (Grant No. 60873161, No. 61005036, and Grant No. 60975037)

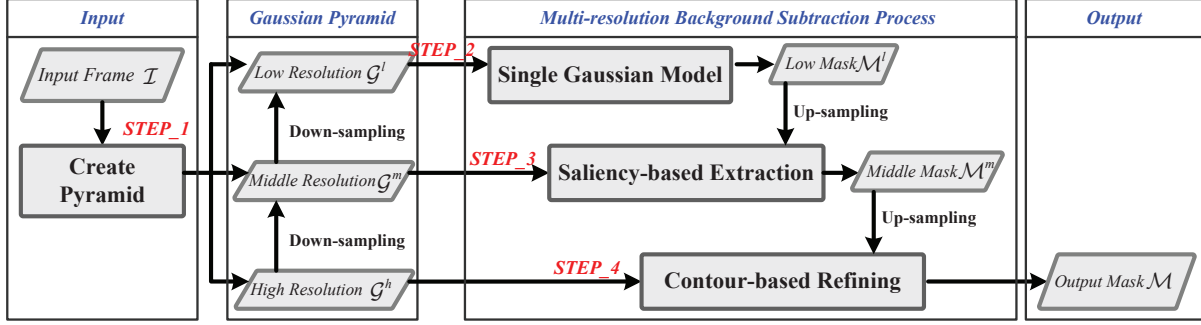


Fig. 1. Overview of the novel background subtraction algorithm.

this problem, we capture first N frames as the learning frames, and create the median image G_{med}^l . Since the objects present in these frames, the statistical background model is created by calculating the weighted single Gaussian model, as follow

$$\mu^l = \frac{\sum_{i=1}^N \omega_i^l \cdot G_i^l}{\sum_{i=1}^N \omega_i^l} \quad \sigma^l = \frac{\sum_{i=1}^N \omega_i^l \cdot (G_i^l - \mu^l)^2}{\frac{N-1}{N} \sum_{i=1}^N \omega_i^l} \quad (3)$$

where the weight ω_i^l is used to minimize the effect of outliers (values far from the median frame G_{med}^l). The weights are computed from a Gaussian distribution centered at G_{med}^l

$$\omega_i^l = \exp \left(-\frac{(G_i^l - G_{med}^l)^2}{2\hat{\sigma}^2} \right) \quad (4)$$

where $\hat{\sigma}$ is the user-settable parameter. Once the statistical model has been constructed, the foreground mask of the new frame is gained by using mean squared Mahalanobis distance, given by

$$\mathcal{M}^l = \begin{cases} 1 & |(\frac{G^l - \mu^l}{\sigma^l})| > \tau^l \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where τ^l is the user-settable threshold.

2.2. Saliency-based Extraction Model

The up-sampled $\widehat{\mathcal{M}}^m$ is first analyzed by extracting all the connected components as regions-of-interest (**ROIs**). Each **ROI** is examined individually in an attempt to extract the foreground mask, which is calculated by combining some types of saliency. In this paper, we use three types of saliency as follows, the intensity saliency \mathcal{S}_I^m , the shadow saliency \mathcal{S}_S^m , and the contour saliency \mathcal{S}_C^m . Note that, other types of saliency can also be added to this model.

Same as the *Single Gaussian Model*, the mean and variance of three color channel $\mu_{r,g,b}^m, \sigma_{r,g,b}^m$ are obtained based on the first N learning frames. The intensity saliency \mathcal{S}_I^m of the current middle-resolution color frame $G_{r,g,b}^m$ is calculated by using the composed mean squared Mahalanobis distance, given by

$$\mathcal{S}_I^m = \sum_{i \in \{r,g,b\}} \left| \frac{G_i^m - \mu_i^m}{\sigma_i} \right| \quad (6)$$

The shadow saliency \mathcal{S}_S^m is represented by the chromatic distortion [11], given by

$$\begin{aligned} \alpha &= \arg \min_{\alpha} \sum_{i \in \{r,g,b\}} \|G_i^m - \alpha \cdot \mu_i^m\| \\ \mathcal{S}_S^m &= \sum_{i \in \{r,g,b\}} \|G_i^m - \alpha \cdot \mu_i^m\| \end{aligned} \quad (7)$$

The contour saliency \mathcal{S}_C^m is represented by the minimum of the normalized current gradient magnitudes and the normalized current-background gradient-difference magnitudes [12],

$$\mathcal{S}_C^m = \min \left(\frac{\|\langle G_x^m, G_y^m \rangle\|}{\text{MAX}}, \frac{\|\langle (G_x^m - \mu_x^m), (G_y^m - \mu_y^m) \rangle\|}{\text{MAX}} \right) \quad (8)$$

where $G_{x,y}^m, \mu_{x,y}^m$ are the x, y gradient magnitudes of the current gray-level middle-resolution frame and gray-level background. And the normalization factors are the respective maximum magnitudes of the input gradients and the input-background gradient-differences (denoted by **MAX** uniformly). The proposed three types of saliency are normalized respectively before the further processing.

Correspondingly, four criterions are proposed to determine a given middle-resolution mask \mathcal{M}^m as follows: 1. mask \mathcal{M}^m should similar to the up-sampled mask $\widehat{\mathcal{M}}^m$; 2. sum of masked \mathcal{S}_I^m by \mathcal{M}^m should as large as possible; 3. sum of masked \mathcal{S}_S^m by \mathcal{M}^m should as small as possible; 4. mask edge \mathcal{M}_E^m should close to \mathcal{S}_C^m . \mathcal{M}_E^m is calculated by the sum of x, y gradient magnitudes, given by

$$\mathcal{M}_E^m = |f_x \otimes \mathcal{M}^m| + |f_y \otimes \mathcal{M}^m| \quad (9)$$

where f_x, f_y are the two differential-filters along x, y directions.

The first criterion is represented by $\|\mathcal{M}^m - \widehat{\mathcal{M}}^m\|$. The second one is represented by $\|\tilde{\mathcal{S}}_I^m\|$, where $\tilde{\mathcal{S}}_{I,i,j}^m = \mathcal{S}_{I,i,j}^m \cdot \mathcal{M}_{i,j}^m$ (Note that $1 \leq i \leq h, 1 \leq j \leq w$, where h, w are the height and width of the matrix). Same as the second criterion, the third one is represented by $\|\tilde{\mathcal{S}}_S^m\|$. The fourth one is represented by $\|\tilde{\mathcal{S}}_C^m\|$, where $\tilde{\mathcal{S}}_{C,i,j}^m = \mathcal{S}_{C,i,j}^m \cdot \mathcal{M}_{i,j}^m$. In order to facilitate the calculation, mask matrixes \mathcal{M}^m are stacked into column vectors $\mathcal{V}^m, \widehat{\mathcal{V}}^m$, saliency matrixes $\mathcal{S}_{I,S,C}$, differential-filters $h_{x,y}$ are arranged into sparse matrixes $\mathcal{H}_{I,S,C}, \mathcal{H}_{x,y}$, and convolution operation is translated into linear operation (detail information are illustrated in Appendix). Accordingly, we transform the four criterions as $\|\mathcal{V}^m - \widehat{\mathcal{V}}^m\|, \|\mathcal{H}_I \mathcal{V}^m\|, \|\mathcal{H}_S \mathcal{V}^m\|$, and $\|\mathcal{H}_C \mathcal{H}_x \mathcal{V}_E^m\| + \|\mathcal{H}_C \mathcal{H}_y \mathcal{V}_E^m\|$, respectively. By combining the above criterions, we can get that,

$$\begin{aligned} \mathcal{V}^m &= \arg \min_{\mathcal{V}^m} \mathcal{F}(\mathcal{V}^m) \\ &= \arg \min_{\mathcal{V}^m} \nu_N \cdot \|\mathcal{V}^m - \widehat{\mathcal{V}}^m\| - \nu_I \cdot \|\mathcal{H}_I \mathcal{V}^m\| + \nu_S \cdot \|\mathcal{H}_S \mathcal{V}^m\| \\ &\quad - \nu_C \cdot \left(\|\mathcal{H}_C \mathcal{H}_x \mathcal{V}_E^m\| + \|\mathcal{H}_C \mathcal{H}_y \mathcal{V}_E^m\| \right) \end{aligned} \quad (10)$$

where ν_N, ν_I, ν_C , and ν_S are the four weighting constants. Thereby,

\mathcal{V}^m can be calculated from $\frac{\partial \mathcal{F}}{\partial \mathcal{V}^m} = 0$, that is,

$$\mathcal{V}^m = \left(\mathcal{J} - \frac{\nu_I}{\nu_N} \mathcal{H}_I^T \mathcal{H}_I + \frac{\nu_S}{\nu_N} \mathcal{H}_S^T \mathcal{H}_S - \frac{\nu_C}{\nu_N} \mathcal{H}_C^T (\mathcal{H}_x^T \mathcal{H}_x + \mathcal{H}_y^T \mathcal{H}_y) \mathcal{H}_C \right)^{-1} \hat{\mathcal{V}}^m \quad (11)$$

where \mathcal{J} is the identity matrix. The middle-resolution mask \mathcal{M}^m is obtained by reshaping the vector \mathcal{V}^m into the matrix.

2.3. Contour-based Refining Model

Same as the *Saliency-based Extraction Model*, contour saliency \mathcal{S}_C^h is calculated from the current frame \mathcal{G}^h and the mean frame μ^h , which is gained from the first N learning frames. The contour saliency \mathcal{S}_C^h is further smoothen with Gaussian g_C^h , that is $\mathcal{S}_C^h = \mathcal{S}_C^h \otimes g_C^h$. The high-resolution mask \mathcal{M}^h is thereby calculated by refining the edge of initial up-sampled mask $\hat{\mathcal{M}}^h$ based on the smoothed contour saliency \mathcal{S}_C^h , given by

$$\mathcal{M}^h = \hat{\mathcal{M}}_i^h + \hat{\mathcal{M}}_o^h \quad (12)$$

where $\hat{\mathcal{M}}_i^h, \hat{\mathcal{M}}_o^h$ are the inner and outer masks calculated as follows,

$$\begin{aligned} \hat{\mathcal{M}}_i^h &= \hat{\mathcal{M}}^h \ominus \mathcal{B} \\ \hat{\mathcal{M}}_o^h &= (\mathcal{S}_C^h > \tau^h) \& (\hat{\mathcal{M}}^h \oplus \mathcal{B} - \hat{\mathcal{M}}^h \ominus \mathcal{B}) \end{aligned} \quad (13)$$

where \ominus, \oplus are the erosion and dilation operation, \mathcal{B} is the corresponding 3×3 square block, and parameter τ^h is the user-settable threshold. From the Eqn.13 we can get that, $\hat{\mathcal{M}}_i^h$ represents the inner part of the initial up-sampled mask $\hat{\mathcal{M}}^h$, while $\hat{\mathcal{M}}_o^h$ is the thresholded edge of the $\hat{\mathcal{M}}^h$ (thresholding by the contour saliency).

3. EXPERIMENTS AND RESULTS

For our method, all the parameters are empirical set as follows: learn frames size $N = 48$, variance $\hat{\sigma} = 10$, threshold $\tau^l = 6$, weighting constants $\nu_N = 0.01, \nu_I = \nu_C = \nu_S = 1$, and threshold $\tau^h = 0.05$. In order to evaluate the effectiveness as well as the robustness of our method, we perform some experiments from three aspects: 1. the detail output of each module; 2. the comparison with the ground truth and the state-of-the-art.

Fig.2 presents the output masks of each module in detail. As shown in this figure, the low-resolution masks only provide the rough position of the moving object, while the middle ones interpret the edge information well, especially shadows are removed. In the high-resolution, the edges are further refined more accurate. The three types of saliency proposed in *Saliency-based Extraction Model* are detailed in Fig.3.

Table 1. Overview of the comparison of all methods (F-score).

Algorithm	GMM	Codebook	LBP	Ours
Bootstrap	0.452	0.665	0.611	0.800
Camouflage	0.979	0.976	0.897	0.982
Time of Day	0.918	0.909	0.760	0.924
Waving Trees	0.893	0.938	0.738	0.945

We compare our approach with the ground truth and state-of-the-arts, such as the **GMM** [4], the **Codebook** [6] and the **LBP** [7].

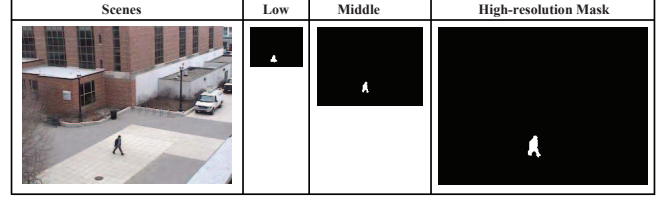


Fig. 2. The overview of output masks of each module. From left to right are the original frame, low-resolution mask, middle-resolution mask, and high-resolution mask.

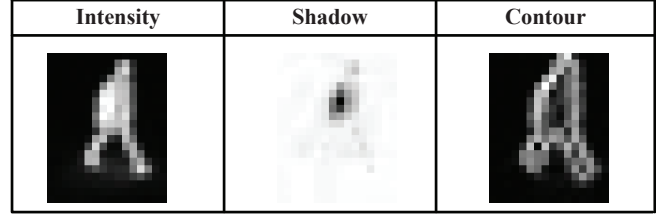


Fig. 3. The overview of three types of saliency: i.e. the intensity saliency, shadow saliency, and contour saliency.

The video sequences are download from the website¹. The results are illustrated in Fig.4. The main challenges of these sequences arise from two aspects, such as the illumination variations and the obvious shadow. As shown in this figure, all the foregrounds can be correctly extracted by our approach. We further use F-score to measure the results. The F-score measures the segmentation accuracy by considering both the recall and the precision, which is defined as

$$F = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \quad (14)$$

where TP, FP, and FN are the true positives (true foreground pixels), false positives (the number of background pixels marked as foreground pixels), and false negatives (the number of foreground pixels that are missed), respectively. Tab.1 gives the numerical comparison of the proposed method with others. We also use another video sequence² to evaluate our method. The result is shown in Fig.5.

4. CONCLUSIONS AND FUTURE WORKS

A effective multi-resolution background subtraction algorithm is proposed in this paper. The main contributions of this work are from two-folds. First, the proposed multi-resolution scheme, coarse to fine strategy, excellently compromise the computation cost with the algorithm effectiveness. Second, the utilized middle-resolution model, that is *Saliency-Based Extraction Model*, provides accuracy of the calculation as well as the scalability of the algorithm. Some future works should be taken into consideration, such as adding other saliency models into the *Saliency-Based Extraction Model* to make the background subtraction results better.

5. APPENDIX

The appendix interprets how to stack the mask matrix into a vector and how to arrange the saliency matrix and filter into the sparse ma-

¹research.microsoft.com/en-us/um/people/jckrumm/wallflower/testimages.htm

²http://cvrr.ucsd.edu/aton/shadow/data/intelligentroom.AVI

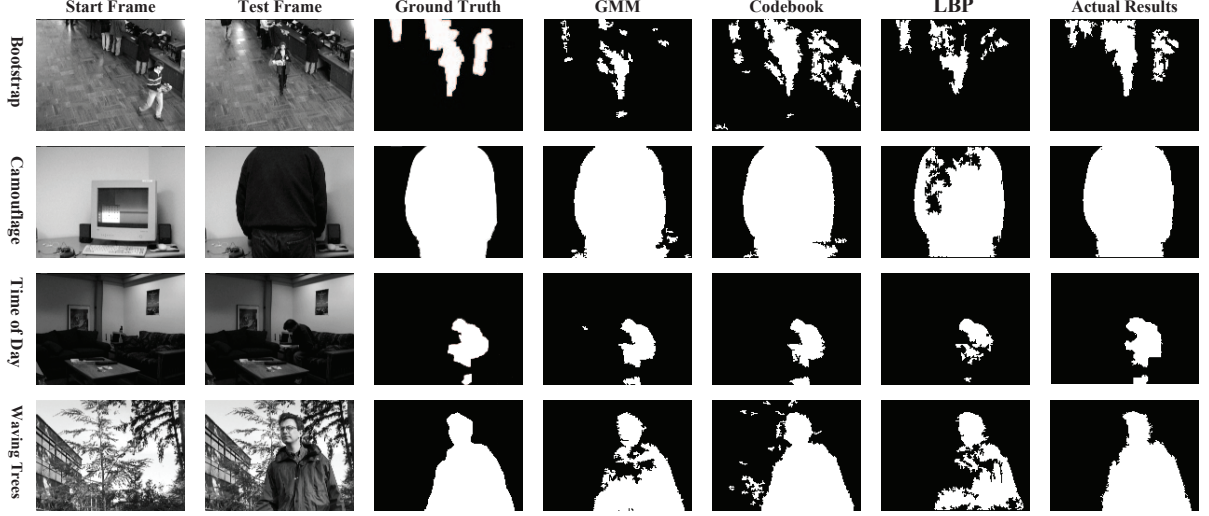


Fig. 4. The comparisons of our method with ground truth and state-of-the-arts, such as the GMM [4], the Codebook [6] and the LBP [7].

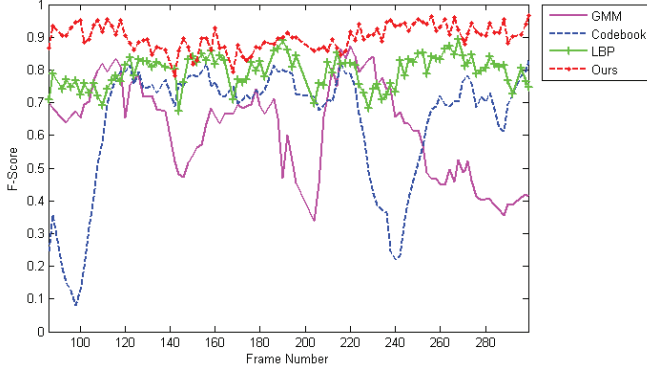


Fig. 5. Comparison results of our method with the traditional methods. The test image sequence exists the obvious shadow where person walking (see the video sequence).

trix. For a mask matrix $\mathcal{M} = \{\mathcal{M}_{i,j} | 1 \leq i \leq h, 1 \leq j \leq w\}$, (h, w are the height and width of the mask matrix), the vector \mathcal{V} (the size of vector \mathcal{V} is hw) is calculated as follow

$$\mathcal{V}_{ij} = \mathcal{M}_{i,j} \quad (15)$$

For a saliency matrix $\mathcal{S} = \{\mathcal{S}_{i,j} | 1 \leq i \leq h, 1 \leq j \leq w\}$, the sparse matrix \mathcal{H} is calculated based on two steps: 1. stacking saliency matrix into a vector \mathcal{V} , that is $\mathcal{V}_{ij} = \mathcal{S}_{i,j}$; 2. arranging vector \mathcal{V} into the sparse diagonal matrix $\mathcal{H} = \{\mathcal{H}_{i,j} | 1 \leq i \leq hw, 1 \leq j \leq hw\}$ (the size of sparse matrix \mathcal{H} is $hw \times hw$), given by

$$\mathcal{H}_{i,j} = \begin{cases} \mathcal{V}_i & i = j \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

For the filter $f_x = [1 \quad -1]$, the sparse matrix $\mathcal{H}_x = \{\mathcal{H}_{x,i,j} | 1 \leq i \leq hw, 1 \leq j \leq hw\}$ (the size of sparse matrix \mathcal{H}_x is $hw \times hw$) is calculated as follow

$$\mathcal{H}_{x,i,j} = \begin{cases} 1 & i = j \\ -1 & i = j + 1 \text{ \& } i \% w \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

where $\%$ is the modulo operation. For the filter $f_y = [1 \quad -1]^T$, the sparse matrix \mathcal{H}_y is calculated as follow

$$\mathcal{H}_{y,i,j} = \begin{cases} 1 & i = j \\ -1 & j = i + w \\ 0 & \text{otherwise} \end{cases} \quad (18)$$

The convolution operation \otimes is translated into linear operation according to the above two translations, given by

$$\mathcal{H}_x \mathcal{V} \Leftarrow f_x \otimes \mathcal{M} \quad \mathcal{H}_y \mathcal{V} \Leftarrow f_y \otimes \mathcal{M} \quad (19)$$

6. REFERENCES

- [1] Dieter Koller, Joseph Weber, and Jitendra Malik, "Robust multiple car tracking with occlusion reasoning," in *European Conference on Computer Vision*(1), 1994, pp. 189–196.
- [2] Nuria M. Oliver, Barbara Rosario, and Alex Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–843, 2000.
- [3] T.J. Darrell C.R. Wren, A. Azarbayejani and A.P. Pentland, "Real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.
- [4] W. Grimson C. Stauffer, "Adaptive background mixture models for real-time tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1999, vol. 2, pp. 246–252.
- [5] D. Harwood A. Elgammal and L.S. Davis, "Nonparametric model for background subtraction," in *European Conference on Computer Vision*, 2000, pp. 751–767.
- [6] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground-background segmentation using codebook model," *Real-Time Imaging In Special Issue on Video Object Processing*, vol. 11, no. 3, pp. 172–185, 2005.
- [7] M. Heikkila and M. Pietikainen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 657–662, 2006.
- [8] Lingfeng Wang, Huaiyu Wu, and Chunhong Pan, "Adaptive elbp for background subtraction," in *Asian Conference on Computer Vision*, 2010, pp. 1866–1877.
- [9] Ruijiang Luo, Liyuan Li, and Irene Gu, "Efficient adaptive background subtraction based on multi-resolution background modelling and updating," in *Advances in Multimedia Information Processing C PCM 2007*, 2007, vol. 4810, pp. 118–127.
- [10] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.
- [11] Thanarat Horprasert, David Harwood, and Larry S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," in *IEEE International Conference on Computer Vision Frame-Rate Workshop*, 1999.
- [12] James W. Davis and Vinay Sharma, "Background-subtraction in thermal imagery using contour saliency," vol. 72, no. 2, pp. 161–181, 2007.