Background Subtraction via Generalized Fused Lasso Foreground Modeling

Bo Xin, Yuan Tian, Yizhou Wang, Wen Gao
Sch'l of EECS, Peking University, Beijing

Background Subtraction (BS) is one of the key steps in video analysis. Many background models have been proposed and achieved promising performance on public data sets. However, due to challenges such as illumination change, dynamic background etc. the resulted foreground segmentation often consists of holes as well as background noise. In this regard, we consider generalized fused lasso (GFL) regularization [4] to quest for intact structured foregrounds. Together with certain assumptions about the background, we formulate BS as a matrix decomposition problem using regularization terms for both the foreground and background matrices. The optimization was carried out via applying the augmented Lagrange multiplier (ALM) method in such a way that a fast parametric-flow algorithm is used for updating the foreground matrix. Experimental results on several popular BS data sets demonstrate better than state-of-the-arts performance.

We start by introducing our model for the unsupervised model learning problem of BS, where foreground and background coexist in the frames. Given a sequence of n video frames, each frame is denoted as $\mathbf{d}^{(i)} \in \mathbb{R}^p$, $i = 1,...,n$. All data are concatenated into one matrix $\mathbf{D} \in \mathbb{R}^{p\times n}$, which is called the observation matrix. We assume that the observation matrix is the summation of a background matrix $\mathbf{B}$ and a foreground matrix $\mathbf{F}$, both unknown. Therefore, by assuming low-rank of $\mathbf{B}$ and structured sparsity of $\mathbf{F}$, we propose the following matrix decomposition objective,

$$\min_{\mathbf{B},\mathbf{F}} \ \text{rank}(\mathbf{B}) + \lambda \|\mathbf{F}\|_{gfl} \quad \text{s.t. } \mathbf{D} = \mathbf{B} + \mathbf{F}, \tag{1}$$

where $\lambda \geq 0$ is a tuning parameter and $\|\cdot\|_{gfl}$ is the generalized fused lasso regularization defined as

$$\|\mathbf{F}\|_{gfl} = \sum_{k=1}^{n} \{\|\mathbf{f}^{(k)}\|_1 + \rho \sum_{(i,j)\in\mathcal{N}} w_{ij}^{(k)} |f_i^{(k)} - f_j^{(k)}|\}, \tag{2}$$

where $\mathbf{f}^{(k)}$ is the kth foreground vector and $\mathcal{N}$ is the spatial neighborhood set. Due to the l_1 penalties on each pixel as well as each adjacent pair of pixels, solutions of $\mathbf{f}$s tend to be both sparse and spatially connected. Here w_{ij} are introduced to enhance the conventional GFL model such that w_{ij} adaptively encode the strength of the fusion according to the image content. Specifically, $w_{ij}^{(k)} = \exp\frac{-\|d_i^{(k)} - d_j^{(k)}\|_2^2}{2\sigma^2}$, where d is the pixel intensity.

In the situation where pure background frames are given, we explicitly utilized this piece of information by adding constraints to the above optimization. Specifically, we separate the observation matrix $\mathbf{D}$ as $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2]$, where $\mathbf{D}_1$ is the matrix of all pure background frames and $\mathbf{D}_2$ is the matrix containing the rest frames with mixed content. The unknown $\mathbf{B}$ and $\mathbf{F}$ are separated correspondingly. Now we assume $\mathbf{D}_1$=$\mathbf{B}_1$ and thus $\mathbf{F}_1$=$\mathbf{0}$. By applying them to Eq. (1), we have

$$\min_{\mathbf{B},\mathbf{F}} \ \text{rank}([\mathbf{B}_1, \mathbf{B}_2]) + \lambda \|\mathbf{F}_2\|_{gfl} \quad \text{s.t. } \mathbf{D}_2 = \mathbf{B}_2 + \mathbf{F}_2, \tag{3}$$

We further assume that $\text{rank}([\mathbf{B_1}, \mathbf{B_2}]) = \text{rank}(\mathbf{B}_1)$. The idea behind this assumption is that if we have enough pure background frames, the corresponding background vectors fully span the background subspace. By taking this assumption, the columns of the unknown $\mathbf{B}_2$ can be represented using linear combinations of the columns of $\mathbf{B}_1$ (or $\mathbf{D}_1$). Specifically, Eq. (3) becomes

$$\min_{\mathbf{D}_1,\mathbf{S},\mathbf{F}_2} \ \text{rank}(\mathbf{D}_1[\mathbf{I}, \mathbf{S}]) + \lambda \|\mathbf{F}_2\|_{gfl} \quad \text{s.t. } \mathbf{D}_2 = \mathbf{D}_1\mathbf{S} + \mathbf{F}_2. \tag{4}$$

Since $\mathbf{D}_1$ is observed/given, whose rank is irrelevant to the optimization, as before, we assume $\mathbf{D}_1$ to be low-rank, therefore there must exists a sparse coefficient matrix $\mathbf{S}$ (given that $\mathbf{D}_1$ is low-rank). So we can instead propose to solve

$$\min_{\mathbf{S},\mathbf{F}_2} \|\mathbf{S}\|_1 + \lambda \|\mathbf{F}_2\|_{gfl} \quad \text{s.t. } \mathbf{D}_2 = \mathbf{D}_1\mathbf{S} + \mathbf{F}_2, \tag{5}$$

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

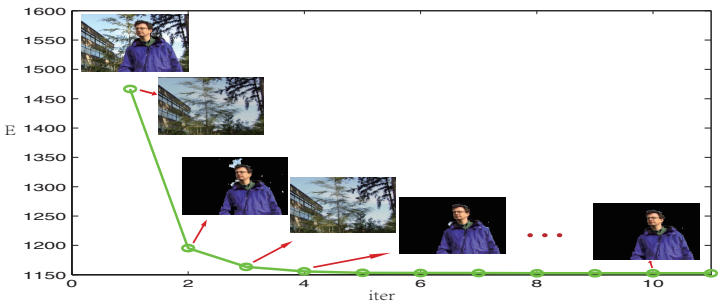

Figure 1: Alternated updating of the background and the foreground. In each iteration (iter) either the background model or the foreground is updated and the objective value (the green plots) keeps decreasing until convergence.

Table 1: Results for on the popular Li[3] data set, given as F-score.

	cam	ft	mr	lb	sc	ap	br	ss
[2]	.7624	.7265	.3871	.6665	.6721	.5663	.6273	.5269
[1]	.5226	.8650	.9014	.7245	.7785	.5879	.8322	.7374
[5]	.8347	.8789	.8995	.6996	.8019	.5616	.7475	.6432
Ours	**.8386**	**.9011**	**.9592**	**.8208**	**.8500**	**.7422**	**.8476**	**.7613**

where $\|\cdot\|_1$ is a convex surrogate for $\|\cdot\|_0$, which counts the number of non-zero entries.

The optimization of both the unsupervised and the supervised case were carried out via applying the augmented Lagrange multiplier (ALM) method in such a way that a fast parametric-flow algorithm is used for updating the foreground matrix. Interestingly, although ALM is a general optimization method, its application to BS helps us to understand how our model alternately pursues and refines the background and the foregrounds. In Figure 1, we visualize the estimation in each iteration of ALM. We observe that the foreground estimation becomes better as the iteration goes on. This is mainly due to the simultaneous estimation of the foreground and the background can reinforce each other. Indeed, experiments show that the proposed model achieves better than state-of-the-art performance on several popular data sets including both natural and synthetic videos. In Table 1, we demonstrate some results on the popular Li data sets, where our model is shown to outperformed the state-of-the-art models. More results, both quantitative and qualitative, can be found in our paper and/or on our webpage. Note also that the algorithm does not take many iterations to converge, and in practice the average number of iterations is about 10-20. Therefore, the major computational cost to pursue structured background and foregrounds in the mid-steps can be eased up by this few iterations. Moreover, since the updating of the foreground are column-wise, the implementation can be highly paralleled in practice. The code can be downloaded on our webpage.

[1] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 2011.

[2] Tom SF Haines and Tao Xiang. Background subtraction with dirichlet processes. In *ECCV*, 2012.

[3] Liyuan Li, Weimin Huang, IY-H Gu, and Qi Tian. Statistical modeling of complex backgrounds for foreground object detection. *Image Processing, IEEE Transactions on*, 13(11):1459–1472, 2004.

[4] Bo Xin, Yoshinobu Kawahara, Yizhou Wang, and Wen Gao. Efficient generalized fused lasso and its application to the diagnosis of alzheimers disease. In *AAAI*, 2014.

[5] Jia Xu, Vamsi K Ithapu, Lopamudra Mukherjee, James M Rehg, and Vikas Singh. Gosus: Grassmannian online subspace updates with structured-sparsity. In *ICCV*, 2013.

http://idm.pku.edu.cn/staff/wangyizhou/