Foreground Segmentation via Dynamic Tree-Structured Sparse RPCA

Salehe Erfanian Ebadi(✉) and Ebroul Izquierdo

School of Electronic Engineering and Computer Science,
Queen Mary University of London, London, UK
{s.erfanianebadi,e.izquierdo}@qmul.ac.uk

Abstract. Video analysis often begins with background subtraction which consists of creation of a background model, followed by a regularization scheme. Recent evaluation of representative background subtraction techniques demonstrated that there are still considerable challenges facing these methods. We present a new method in which we regard the image sequence as being made up of the sum of a low-rank background matrix and a dynamic tree-structured sparse outlier matrix and solve the decomposition using our approximated Robust Principal Component Analysis method extended to handle camera motion. Our contribution lies in dynamically estimating the support of the foreground regions via a superpixel generation step, so as to impose spatial coherence on these regions, and to obtain crisp and meaningful foreground regions. These advantages enable our method to outperform state-of-the-art alternatives in three benchmark datasets.

1 Introduction

Foreground segmentation plays a critical role in applications such as automated surveillance, action recognition, and motion analysis. Despite the efforts in this field, recent evaluation of state-of-the-art techniques [1,2] showed that there are still shortcomings in addressing all challenges in foreground segmentation. Addressing these challenges, leads to a number of considerations in designing a background model, as well as expected behavior from foreground objects, which in complex real-life applications remains an open problem. The background model can undergo sudden or gradual *illumination changes*, as well as *background motions* such as trees swaying or water rippling in a lake. In addition, *global motion* caused by camera movement or jitter can affect detection of genuine foreground objects. *Noise* is another problematic factor which is interleaved with challenges of *camouflage*. In most cases noise can increase the range of values considered to belong to the background, allowing camouflaged objects to remain undetected. A desirable background model must be able to learn a variety of modes from the video feed, such that it handles variations in the background, *moved objects*, and noise without compromising its ability to detect camouflaged regions.

B. Leibe et al. (Eds.): ECCV 2016, Part I, LNCS 9905, pp. 314–329, 2016.
DOI: 10.1007/978-3-319-46448-0_19

In this paper, we handle all these challenges using an approximated *Robust Principal Component Analysis* (RPCA) based method for background modeling. Given a data matrix containing the frames of a video sequence stacked as its columns, $A \in \mathbb{R}^{m \times n}$, RPCA [3] solves the matrix decomposition problem

$$\min_{L,S} \|L\|_* + \lambda \|S\|_1 \quad s.t. \quad A = L + S, \tag{1}$$

as a surrogate for the actual problem

$$\min_{L,S} rank(L) + \lambda \|S\|_0 \quad s.t. \quad A = L + S, \tag{2}$$

where L is the low-rank component corresponding to the background and S is the sparse component containing the foreground outliers. We are interested in a case where we can decompose the matrix A into three components, namely a low-rank part L that can describe the background of the sequence, along with adaptivity to changes introduced to it, a sparse component S containing only the genuine deforming foreground regions, and a noise component E that collectively contains residual error, noise, and ambiguous pixels:

$$A = L + S + E, \tag{3}$$

meaning that the model does not seek the exact solution of decomposing a scene into background and foreground, but rather the approximate solution $A \approx L + S$ [4–7] whereby the residual error E will have the desired properties described above. λ is a tuning parameter ensuring no genuine foreground regions will be missed. This formulation is still inadequate and we need to introduce some necessary steps to lead to substantially better results.

Background modeling by the low-rank approximation has a number of benefits: firstly, that a robust estimation of the mostly static regions of the image is guaranteed; secondly, that this approximation can in part handle the variations in illumination in the background, such as a tree swaying backwards and forward, or water rippling in a lake, traffic light changes that can be modeled by a few modes, or billboards in a street displaying a few images on repeat during a day. Thirdly, a low-rank approximation of the background can help distinguish between general motion in the scene – which can be due to camera movement – and local varying motions caused by moving objects even in the case of large objects such as a huge truck moving across the scene; since the background regions obey a single highly correlated motion pattern.

Despite the promising effects of using a low-rank approximation for obtaining the background model, a sparse constraint for foreground objects, is far too limited. The foreground regions are usually spatially coherent clusters. Thus, we prefer to detect contiguous regions of various sizes, and then lots of zero entries (regions) in the sparse matrix. With this objective in mind, we propose structured-sparsity inducing norms in the context of a novel dynamic group structure, by which the natural structure of foreground objects in the sparse matrix is preserved. The dynamicity of group structures is derived from the

natural shape of objects in the scene, by selecting clusters of pixels via the SLIC superpixels [8], and dynamically refining the size of these clusters in an iterative process. This proposition, has been proven to be successful in reducing the *foreground aperture* and *camouflage* problems in our experiments.

Because we solve an approximated RPCA problem, it is important to drive the algorithm by means of a knowledge of salient regions and the distribution of outliers, so that the algorithm converges to the correct solution. However, knowledge of the object of interest before even segmenting it seems to make the problem as one of the many chicken-egg problems in computer vision, as we usually need to segment the scene to recognize the objects in it. So, to identify an object and its probable size and location even before segmenting it, we use an intuitive *tandem* initialization step by which the background is encouraged to lean towards the best low-rank approximation of the static parts in the scene, and the sparse part is initialized to take on high probability values for regions of the scene where they exhibit highest statistical *leverage* scores.

In a nutshell contributions of this paper are: inducing structured-sparsity in a novel group structure, namely a dynamic superpixel structure; insensitivity to foreground object size, as a result of using within-patch normalized regularization; assumption of a noise part for discarding false positive pixels (false alarms); low-rank approximation of background to accommodate illumination and small scene changes; a *tandem* algorithm for removal of unwanted ghosting effects that persist in most background subtraction techniques, and targets the unascertained prior knowledge of distribution of outliers; and an exhaustive evaluation using three datasets [9–11] demonstrates top performance in comparison with the state-of-the-art alternatives.

2 Related Work

In the recent years, global models such as principal component analysis (PCA) [3], have gained some popularity due to their computational simplicity and effectiveness in camera shake. However, in those early models the spatial distribution of outliers was not considered. In an effort to incorporate such prior an MRF-based solution [12] has been proposed. But the result of imposing such smoothness constraint is that the foreground regions tend to be over-smoothed; as an example, the details in the silhouette of hands and legs of a moving person is sacrificed in favor of a more compact blob. Our idea is established in the so-called structured-sparsity or group-sparsity measures to incorporate the spatial prior. Structural information about nonzero patterns of variables have been developed and used in sparse signal recovery, and many approaches have been applied to these problems successfully [13]. However, the majority of related methods such as [14] typically assume that the block structure and its location is known or will suffer in *regularization*, *bootstrapping*, or *foreground aperture*. To lift up some of the difficulties the sparsity structure is estimated automatically in [13], however parameter tuning is required to control the balance between the sparsity prior and the group clustering prior for different cases. The authors

of [14] used a two-pass RPCA framework, in which the first pass generates a saliency map that corresponds to locations of the outliers, and then the second pass uses pre-defined salient blocks in the image, to favor spatially contiguous outliers. In another effort [15] used a group sparse structure, in which overlapping pre-defined groups of pixels in a region of an image are used in conjunction with a maximum norm regularization to take into account the spatial connection of foreground regions. In a recent work [16] a superpixel-based max-norm matrix decomposition approach has been proposed, in which homogeneous static or dynamic regions of image are classified as a graph partitioning problem, via Generalized Fused Lasso. In contrast to all the above, our method does not assume a prior size or location or structure for sparsity, and dynamically updates these to best fit the natural object shape in the scene, without a separate training phase or the need for a clean background for background training. In the next section, we introduce our dynamic tree-structured sparsity-inducing norms that leads to substantially better results than other RPCA based methods [4–7] and other state-of-the-art alternatives.

3 Our Algorithm

3.1 Approximated RPCA via Structured-Sparsity Inducing Norms

We propose sparsity-inducing norms that can incorporate prior structures on the support of the errors such as spatial continuity. We essentially consider a special case to the following problem

$$\min_{rank(L)\leq r,S,\tau} \|A\circ\tau - L - S\|_F + \lambda\psi(S) \quad s.t. \quad A\circ\tau = L + S + E, \tag{4}$$

where $\|L\|_F$ is the Frobenius norm of matrix L, defined as $\|L\|_F = \sqrt{\sum_{i,j} L_{ij}^2}$, and λ set at a value that ensures no genuine foreground regions will be missed. We strictly have $rank(L) \leq r < rank(A)$. E is a matrix containing the residual error of the approximation of A by $L + S$. The entries of this matrix can be very large in magnitude, but random and scattered, exhibiting noise-like behavior, and showing no structured shape in the sparsity domain. Therefore, they should neither remain in the foreground as they will trigger many false positives and pollute the foreground model, nor be able to get absorbed into the background model; and the robust low-rank approximation will already ensure the latter case. Most background subtraction methods suffer from this kind of contamination polluting their foreground model, and consequently resort to a final thresholding step or post-processing once the foreground support is calculated. The choice of λ is justified by observations in our experiments, where λ controls a good trade-off between the sparsity of $S + E$ and structured-sparsity of S. We have assumed that the images in matrix $A\circ\tau$ are well aligned, where τ stands for some transformation in the image domain (e.g., 2D affine transformation for correcting misalignment, or 2D projective transformation for handling some perspective change).

The regularizer $\psi(\cdot)$ on S is chosen to be $\|S\|_{2,1}$. $\ell_{2,1}$-norm is a group sparsity inducing norm defined as the ℓ_1-norm of the vector formed by taking the ℓ_2-norm of a matrix. Clearly, the ℓ_1-norm regularization treats each entry (pixel) in S independently. It does not take into account any specific structures or possible relations among subsets of the entries. While in background subtraction scenarios, outliers (objects in the scene) normally have the structural properties of spatial contiguity and locality. Hence, our choice of $\ell_{2,1}$-norm assures selecting the discriminative input features shared across multiple binary predictors.

To induce more diverse and sophisticated sparse error patterns, we consider structured sparsity-inducing norms that involve overlapping groups of variables, motivated by recent advances in structured sparsity [17]. Although it still assumes pre-defined group structures, the overlapping patterns of groups and norms associated with the groups of variables allow to encode much richer classes of structured sparsity. In this work, we consider a tree-structured sparsity-inducing norm. It involves a hierarchical partition of the m variables in S into groups, as shown in Fig. 1(a). The tree is defined in a way that leaf nodes are singleton groups corresponding to individual pixels, and internal nodes/groups correspond to local patches of varying size. Thus each parent node contains a hierarchy of child nodes that are spatially adjacent to each other and constitute a local part in the sparse image S. As illustrated in Fig. 1(a) in the grayed-out regions, when a parent node goes to zero all its descendants in the tree must go to zero. Consequently, the nonzero or support patterns are formed by removing those nodes forced to zero. This is exactly the desired effect of structured error patterns of spatial locality and contiguity.

We can represent a scene using a tree structure by subdivision. In such a tree structure each child node is a subset of its parent node and the nodes of the same depth level do not overlap. Denote $\mathcal{G}$ as a set of groups from the power set of the index set $\{1, \ldots, m\}$, with each group $G \in \mathcal{G}$ containing a subset of these indices. The aforementioned tree-structured groups used in this paper are formally defined as follows: A set of groups $\mathcal{G}$ is said to be *tree-structured* in $\{1, \ldots, m\}$ if $\mathcal{G} = \{\ldots, G_1^i, G_2^i, \ldots, G_{b_i}^i, \ldots\}$ where $i = 0, 1, 2, \ldots, d$, d is the depth of the tree, $b_0 = 1$ and $G_1^0 = \{1, 2, \ldots, m\}$, $b_d = m$ and correspondingly $\{G_j^d\}_{j=1}^m$ are singleton groups. Let G_j^i be the parent node of a node $G_{j'}^{i+1}$ in the tree, we have $G_{j'}^{i+1} \subseteq G_j^i$. We also have $G_j^i \cap G_k^i = \emptyset, \forall i = 1, \ldots, d, j \neq k, 1 \leq j, k \leq b_i$. Similar group structures are also considered in [17]. With the above notation, a general tree-structured sparsity-inducing norm can be written as

$$\psi(S) = \sum_{i=0}^{d} \sum_{j=1}^{b_i} w_j^i \|S_{G_j^i}\|_{2,1}, \tag{5}$$

where $S_{G_j^i}$ is a vector with entries equal to those of S for the indices in G_j^i and 0 otherwise. w_j^i are positive weights for groups G_j^i. It is chosen as $w_j^i = 1/max(A_{G_j^i})$ to overcome sensitivity of the regularization scheme to illumination variance across patches. This within patch normalized regularization is crucial. As we will explain later, using the same regularizing parameter for all the patches

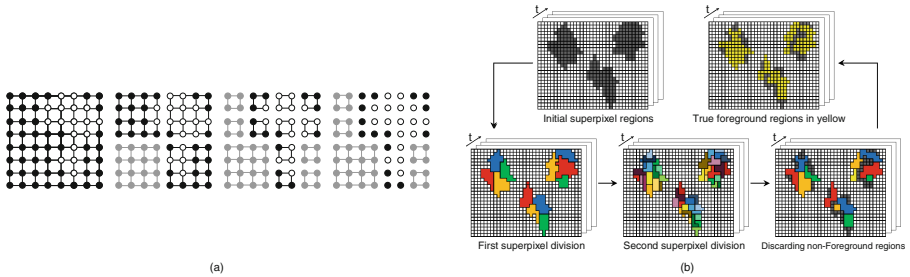

Fig. 1. (a) Tree-structured groups in sparsity induction, division, and discarding procedure. (b) same procedure in superpixel regions where the size and location of groups are not known and change from one frame to next. Grayed-out regions in (a) and (b) are the result of discarding process that is immediately performed on groups that are foreground-absent; thus, saving computation time as they are not processed ever after.

in the scene will usually favor the most prominent features (in this case the illumination variations with largest magnitude). By normalizing each patch with a weight associated with the highest color variation in that patch, this issue is largely subsided; and as such the *camouflaged* objects will have a higher chance of being detected. For the $\ell_{2,1}$-norm, it is the maximum value of pixels in a group that decides if the group is set to nonzero or not, and it does encourage the rest of the pixels to take arbitrary (hence close to maximum) values. Thus, the objective function in the optimization program (3) is modified to the following

$$\min_{rank(L)\leq r,S,\tau} \|A\circ\tau - L - S\|_F + \lambda \sum_{i=0}^{d}\sum_{j=1}^{b_i} w_j^i \|S_{G_j^i}\|_{2,1} \quad s.t. \quad A\circ\tau = L + S + E \tag{6}$$

To solve (6) we use an alternating minimization procedure. This kind of iterative linearization has a long history in gradient algorithms. We first find a good initialization for τ by pre-aligning all frames in the sequences to the middle frame, before the main loops of minimization. Then the linearization of τ is done by the robust multiresolution method proposed in [6,7]. We then proceed by minimizing the function for two parameters L and S one at a time until the solution reaches convergence; that means solving two reduced problems, each being minimized independently form one another

$$L^t = \underset{rank(L)\leq r}{\arg\min} \|A\circ\tau - L - S^{t-1}\|_F^2 \tag{7}$$

$$S^t = \underset{S}{\arg\min} \|A\circ\tau - L^t - S\|_F^2 + \lambda \sum_{i=0}^{d}\sum_{j=1}^{b_i} w_j^i \|S_{G_j^i}\|_{2,1} \tag{8}$$

3.2 Robust Foreground Segmentation via Structured Sparsity

A meaningful structured-sparse solution, is the one that is best able to take into account the natural shape and structure of objects in the scene. There is a need for some mechanism that describes each tree-structured group $\psi(\cdot)$. Each group must take into account connected components belonging to a semantically connected region. For example, a region of pixels with the same color and texture belonging to part of an object (a wheel of a car) must be assigned to a single group. The structured sparse inducing framework defined in the previous section can then be used within the group class to decide whether it belongs to foreground or must be classified as background.

As mentioned before, most block-structured sparse solutions have two limitations. Firstly, the size and location of the blocks need to be set in advance. Secondly, it is hard to see how each block is adapting its shape to the natural structure of objects in the scene. Motivated by these limitations, we propose a new group structure, in which the structure of sparse part is the same as the natural object structure in the scene. In a test image, the scene can be classified into multiple *superpixels*. A good superpixel must obtain perceptually meaningful atomic regions, which can be used to replace the rigid structure of the pixel grid. Moreover, as these results will be used as a pre-processing step in our foreground detection framework, they should be fast to compute, memory efficient, and simple to use. Also, in our segmentation scenario, superpixels should both increase the speed and improve the quality of the results.

We therefore, adopt the *simple linear iterative clustering* (SLIC) algorithm based on the empirical comparison of six state-of-the-art superpixel methods [8]. SLIC adapts k-means clustering to generate superpixels, and is freely available[1]. By default, the only parameters of the algorithm are the desired number of approximately equally-sized superpixels ξ and compactness factor φ controlling adherence of each superpixel region to object boundaries. For our test images, $\xi = 800$ and $\varphi = 20$ are sufficient to adhere well to all object boundaries.

Once the superpixels are obtained, the structured sparsity inducing norms are applied to groups, that are now each superpixel region in the test image. Figure 1(b) shows an example of this procedure. We have adapted SLIC to be able to dynamically divide each superpixel region into approximately equal-sized smaller superpixels that best adhere to object boundaries. If a small superpixel region does not contain any foreground, it is discarded as background immediately and no further processing is performed for this region. If otherwise a region hints presence of foreground, it is divided into several smaller superpixels again. The same process is performed for these smaller regions, and the resulting regions containing foreground are once again divided and put to test. Our experiments have shown that at this depth the classification can be performed without having to perform any further divisions, as the regions are both small enough to safely discard non-foreground regions, and large enough to crisply classify all foreground objects in the scene with fine details correctly. We denote this procedure as sparsity *induction*, *division* and *discarding*. Thus, in the

[1] http://ivrl.epfl.ch/research/superpixels.

general tree-structured sparsity-inducing norm (5) depth of each tree is $d = 3$ and $m = \mathcal{M}$ is dynamically decided by SLIC, since it depends on the natural shape of the objects in the scene. Therefore $\mathcal{G} = \{\dots, G_1^i, G_2^i, \dots, G_{b_i}^i, \dots\}$ where $i = \{0, 1, 2, 3\}$, $b_0 = 1$ and $G_1^0 = \{1, 2, \dots, \mathcal{M}\}$, $b_d = \mathcal{M}$ and correspondingly $\{G_j^d\}_{j=1}^{\mathcal{M}}$ are the smallest superpixel groups.

3.3 Tandem Initialization for Removing Ghosting Effects

In this section we propose the *tandem* approximated RPCA where just like a tandem bicycle the front drive is supported by the back pedaling power. This proposition involves an initialization step before the actual optimization takes place. It is different from algorithms that require a two-pass optimization [14], where the optimization is twice performed to refine results. This is rather expensive in an RPCA framework; instead, we strategically initialize the variables such that we gain even better results. This modification will introduce a prior knowledge of the spatial distribution of the outliers to the model. The direct impact of this modification to the RPCA algorithm is faster convergence. The indirect impact is how it alleviates a persisting problem in background subtraction algorithms, called *"ghosting"* effect. The *ghosts* are either parts of the foreground object that remain in the background model, or parts of the background that leak into the foreground. The main reasons causing these artifacts are: an object moving slowly, or remaining inactive for some period of time, and when the foreground object obscures part of the background during the training period. With current RPCA-based optimizations the ghosts usually persist during the iterative process; this can be seen in Fig. 2. The optimization problems described in Sects. 3.1 and 3.2 are solved by iterative procedures that need to be initialized using starting values of the matrices L, S, and τ. The iterative process is started with a standard (naïve) initialization of $L^0 = A$, $S^0 = 0$, and $\tau^0 = 0$. The rank-r matrix that is the nearest to the matrix A is a low-rank matrix that gives a good first approximation for the static part of the sequence but some parts of the moving objects remain in this rank-r matrix. Hence we propose to construct a matrix S^0 whose columns contain only the more salient part of the difference between A and L^0, where L^0 is the rank-r matrix approximation of the matrix A. This difference matrix $\mathcal{S} = A - L^0$ will contain a sketch of the moving objects in the scene, and therefore is a good initial approximation that contributes to the nonuniformity of the structure of the matrix. We adopt the statistical *leverage* scores to measure the importance of the columns of the difference matrix. These scores can be regarded as a pseudo-motion saliency map.

Let the i-th column of the matrix to be a linear combination of the orthonormal basis given by the left singular vectors of the matrix $\mathcal{S}^i = \sum_{r=1}^{rank} \sigma_r U_r V_r^i$, $i = 1, \dots, \eta$ where U_r is the r-th left singular vector, V_r^i is the i-th coordinate of the r-th right singular vector, and $rank$ is the rank of the matrix $\mathcal{S}$. As the matrices $\mathcal{S}_j$ are approximations of the frames containing the moving objects, they can be considered as approximations to low-rank matrices. One can assume

$$\mathcal{S}_j^i \approx \sum_{r=1}^{\rho} \sigma_r U_r V_r^i, \quad \rho \ll rank \tag{9}$$

Note that any two columns i_1 and i_2 differ only by $\sum_{r=1}^{\rho} V_r^{i_1}$ and $\sum_{r=1}^{\rho} V_r^{i_2}$. Then these terms can be used to measure the importance or contribution of each column to the matrix. The normalized statistical leverage scores [18] of the i-th column of matrix $\mathcal{S}_j$ is defined as

$$\ell_i = \frac{1}{\rho} \sum_{r=1}^{\rho} V_r^i, \quad i = 1, \ldots, \eta, \tag{10}$$

where η is the number of columns of each frame of the sequence. The sub-index j is removed to help understanding this expression. Leverages have been used historically for outlier detection in statistical regression but recently they have been used to give column (or row) order of the amount of motion saliency in a specific part of the image. The vector ℓ_i is a probability vector, i.e. $\sum_{i=1}^{\eta} \ell_i = 1$. Therefore, the columns of each matrix $\mathcal{S}_j$ with leverages greater than $\frac{1}{\eta}$ are the more important columns. So the columns of the initial approximation S^0 contain only the more important columns of the matrices $\mathcal{S}_j$, $j = 1, \ldots, n$. Consequently, the less salient parts of the image are not included in the initialization of the sparse part, making the iterative process faster to converge, yielding more stable results, and increasing the segmentation accuracy.

$$S_j^{0^i} = \begin{cases} \mathcal{S}_j^i, & \ell(\mathcal{S}_j^i) \geq \frac{1}{\eta} \\ 0, & otherwise \end{cases} \tag{11}$$

In Fig. 2 we have shown the effect of the tandem initialization in our model, with comparison to other RPCA-based algorithms. The ghost effects are visible in foreground parts in the forth to sixth columns of this figure, which in turn contaminate the background model in the eighth to tenth columns. Algorithm 1 shows the pseudo-code for our model with tandem initialization and motion parameter estimation.

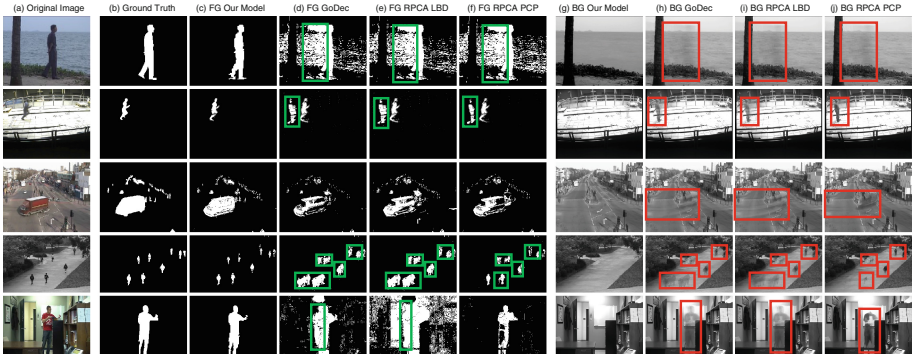

Fig. 2. *Ghosting effects* that persist in RPCA-based methods [19–21]. A contaminated background model in red regions affects the foreground segmentation in green regions. Our tandem model is able to eliminate these artifacts, without post-processing.

Algorithm 1. Pseudo-code with background motion parameter estimation and Tandem initialization

1: **Input:** A, *rank*, λ, ϵ, *maxIter*
2: **Output:** S, L, E, τ
3: *Calculate* $\mathcal{S} \approx \sum_{r=1}^{\rho} \sigma_r U_r V_r, \quad \rho \ll rank$
4: *Tandem initialization:* $\tau^0 = \tau_{pre-align}$, $L^0 = A$, $S^0 = \begin{cases} \mathcal{S}, & \ell(\mathcal{S}) \geq \frac{1}{\eta} \\ 0, & otherwise \end{cases}$
5: **while** $\|A \circ \tau^t - L^t - S^t\|_F^2 / \|A\|_F^2 > \epsilon$ or $t < maxIter$ **do**
   1) Form the matrix $A \circ \tau$ calculating the parameters τ_i^t that infer the mapping that transforms the column vector A_i to the i-th column vector of the matrix $L^{t-1} + S^{t-1}$.
   2) Calculate $L^t = \sum_{i=1}^{rank} \sigma_i U_i V_i^T$ where $svd(A \circ \tau^t - S^{t-1}) = U \Sigma V^T$.
   3) Calculate $S^t = \mathcal{P}_\lambda(\psi(A \circ \tau^t - L^t))$ where $\mathcal{P}_\lambda(x) = sign(x) \max(|x| - \lambda, 0)$.
   4) Calculate the residual noise $E = A - L - S$.
6: **end while**

4 Experiments and Analysis

Our algorithm is implemented and tested in MATLAB on a desktop machine, single core on an Intel Core i7-4770 CPU and 32 GB of RAM. The average processing time on a sequence of 100 RGB frames with resolution 600×800 with image alignment and background motion estimation and superpixel generation step is about 1674 seconds. We perform extensive tests using three datasets [9–11] comprised of a total of 49 videos, allowing us to compare our method to a large number of alternative methods. For all the tests these same set of parameters are used: regularizing parameter $\lambda = 3/\sqrt{\max(m, n)}$, depth of each tree $d = 3$, number of singleton groups $\mathcal{M}$ dynamically chosen by SLIC, number of superpixels per image $\xi = 800$, and compactness factor $\varphi = 20$. All the tests were conducted on the *temporal region of interest* of the sequences, meaning no training stage with clean background was used to obtain the background model. All our results have been reported without refinement or post-processing. We refer to our method as DSPSS short for Dynamic Superpixel Structured-Sparse.

4.1 Qualitative Results

In Fig. 2 we have shown the effect of the tandem initialization in our model, with comparison to other RPCA-based algorithms that suffer from ghosting effects. A contaminated background model visible in red regions, would in turn affect the foreground segmentation in green regions, resulting in high false positive rate. Our algorithm is capable of adapting to slow-moving foreground objects in these sequences, all the while being able to discard non-genuine false-alarm foreground pixels with the robust foreground segmentation via our tree structured sparsity-inducing norms; notice the eliminated water rippling pixels in the foreground segmentation of the first row. Our model is robust to variations in foreground object size; this can be seen in the third row results, where a

Fig. 3. *i2R* results: top row is the original image, second row is the ground truth, and the last row is our unrefined results without post-processing. We used the same frames as [15,16,22–25], for qualitative comparison.

large foreground object is well-segmented simultaneously with small pedestrians in the scene. Our sparsity-inducing norms defined in superpixel regions prove to be effective in obtaining accurate silhouette of foreground regions in all the examples of Fig. 2, specially in the case of first row where the legs of the person walking are *camouflaged* due to similar intensity with the background.

Figure 3 shows segmentation results for the i2R dataset. The top row is the original image, second row is the ground truth, and the last row is our results. We have used the same frames as [15,16,22–25], for qualitative comparison. Figure 3(a) is a scene with pedestrians and cars passing in front of a very dynamic background with trees swaying back and forth and illumination changing rapidly. Our method has been able to crisply detect genuine foreground regions while discarding the dynamic pixels in background. The same scenario applies to Fig. 3(b) and (c), where the fountain and water rippling in the lake make it hard to distinguish genuine foreground regions. The sparsity-inducing norms defined in superpixel regions manifest their effectiveness in adhering well to coherent foreground segmentation, while the tree-structure successfully discards the non-rigid and random foreground alarms caused by the fountain and water turbulence. In (d) a person appears in front of a curtain that moves with wind, and remains there for a period of time, and again walks out of the scene. Our background model has adapted itself to the variations in the scene such that the inactive foreground does not get absorbed into the background. The column (e) is an indoor scene with sudden illumination change; our method suffers a bit from this sudden change, but quickly adapts itself so that the foreground objects would not go undetected. (f) and (g) are simpler to process, except for some camouflage in (f) due to color similarity between some foreground objects and the background, but again good performance is obtained in these scenes. For (h) no training period for background is available; i.e., no foreground-absent frame is seen, but our model is able to obtain a robust background model nonetheless. (i) is a scene with a very fast moving escalator, and people appearing from the end of a hallway that is poorly-lit. Evidently, background modeling with low-rank approximation best proves itself here by adapting well to the repetitive motion of the escalator, by a few modes, and the sparsity-inducing norms are well able

to detect the people moving in the darkness at the back of the scene. Also, the within-patch normalized regularization guarantees insensitivity to foreground object size and illumination invariant performance in all cases.

4.2 Quantitative Results

For quantitative comparison we present the F-measure scores, defined as the harmonic mean of the recall and precision:

$$\text{recall} = \frac{tp}{tp + fn}, \quad \text{precision} = \frac{tp}{tp + fp}, \quad \text{F-measure} = 2\frac{\text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}},$$

where fp is the number of false positives, tn the number of true negatives, etc. The change detection dataset [10] is the largest dataset in our evaluation, and includes a dense ground truth. It also limits parameter tuning, such that a single parameter must be used for all the 31 videos. Video resolution is not great however, often with a low quality de-interlacing algorithm that creates ghosts. The dataset is comprised of six categories, 31 real-world videos (including thermal sequences), totaling over 80,000 frames, to include diverse motion and change detection challenges. The results for these sequences can be seen in Table 1. For the reason of space limit, in each category we compare our model with the top performing methods that have submitted results for that category (readers are referred to [10] and its website for complete list of references and the corresponding performance figures). In addition to this list, we have included the DP-GMM [23] and five RPCA-based methods PCP [21], DECOLOR [12], and very recent 2-pass RPCA [14]. For LSD-GSRPCA [15] and SPGFL [16] only a fraction of the results were reported in their papers, therefore they are included where results are reported. For PCP we use our pre-alignment step for the *camera jitter* sequences and as such we denote it as PCP+Alignment.

The most challenging categories are *intermittent motion* and *thermal*. We advantage at *intermittent motion* category thanks to the tandem initialization to remove the ghosting problem, and the robust low-rank approximation of the background, that can learn multiple modes for the background of a sequence. However, in *thermal* since we do not have a mechanism for handling thermal images our algorithm suffers from artifacts such as heat stamps (e.g., bright spots left on a seat after a person gets up and leaves), heat reflection on floors and windows, and camouflage effects when a moving object has the same temperature as the surrounding regions. In all other categories, our method achieves top performance, thanks to the robust low-rank approximation of the background, that can learn multiple modes for the background of a sequence, to the tandem initialization to remove the ghosting problem, and the pre-alignment step and motion parameter estimation simultaneously with decomposition. Overall, we win on average for the CDnet dataset. This is because our model can handle backgrounds that are complex and dynamic. This ability, in combination with the tree-structured sparsity inducing mechanisms allows it to effectively segment genuine well-outlined foreground regions.

Table 1. *CDNet* [10] dataset: F-measure results for all the categories for the most competitive methods. Table accurate as of March 2016, with results from CDnet http://changedetection.net/. The online chart keeps updating.

Method	Baseline	Camera jitter	Dynamic background	Intermittent motion	Shadow	Thermal	Mean
LSD-GSRPCA [15]	.7173 (9)	-	-	-	-	-	-
SPGFL [16]	.9469 (3)	-	.8519 (3)	.6988 (3)	-	.8156 (3)	-
PCP+Alignment [21]	.9109 (8)	.7218 (7)	.6941 (8)	.5371 (8)	.7885 (7)	.7192 (7)	.7286 (7)
DECOLOR [12]	.9215 (7)	.7776 (5)	.7084 (7)	.5945 (6)	.8317 (4)	.7081 (8)	.7570 (6)
DP-GMM [23]	.9286 (5)	.7477 (6)	.8137 (5)	.5418 (7)	.8127 (5)	.8134 (4)	.7763 (5)
2-pass RPCA [14]	.9281 (6)	.8152 (2)	.7818 (6)	.6826 (4)	.8063 (6)	.7597 (5)	.7956 (4)
SuBSENSE [26]	.9500 (2)	.8150 (3)	.8180 (4)	.6570 (5)	.8990 (2)	.8170 (2)	.8260 (3)
PAWCS [27]	.9397 (4)	.8137 (4)	.8938 (2)	.7764 (2)	.8710 (3)	**.8324 (1)**	.8545 (2)
DSPSS	**.9664 (1)**	**.8662 (1)**	**.9057 (1)**	**.7870 (1)**	**.9177 (1)**	.7328 (6)	**.8626 (1)**

Table 2. *SABS* [9] dataset: F-measure results for nine challenges; only the most competitive algorithms were included.

Method	Basic	Dynamic background	Bootstrap	Darkening	Light switch	Noisy night	Camouflage	No camouflage	H264, 40 kbps	Mean
Barnich [28]	.761 (4)	.711 (3)	.685 (3)	.678 (3)	.268 (4)	.271 (4)	.741 (4)	.799 (4)	.774 (3)	.632 (4)
Zivkovic [29]	.768 (3)	.704 (4)	.632 (4)	.620 (4)	.300 (3)	.321 (3)	.820 (3)	.829 (3)	.748 (4)	.638 (3)
DP-GMM [23]	.853 (2)	.853 (2)	.796 (2)	.861 (2)	**.603 (1)**	.788 (2)	.864 (2)	.867 (2)	.827 (2)	.812 (2)
DSPSS	**.867 (1)**	**.871 (1)**	**.822 (1)**	**.907 (1)**	.570 (2)	**.897 (1)**	**.894 (1)**	**.913 (1)**	**.841 (1)**	**.842 (1)**

Table 3. *i2R* [11] dataset F-measure results. We report our results without parameter tuning, although the dataset allows this.

Method	cam	ft	ws	mc	lb	sm	ap	br	ss	Mean
DECOLOR [12]	.3416 (6)	.2075 (6)	.9022 (5)	.8700 (4)	.646 (6)	.6822 (5)	.8169 (2)	.6589 (4)	.7480 (3)	.6525 (6)
DP-GMM [23]	.7876 (3)	.7424 (5)	.9298 (3)	.8411 (5)	.6665 (5)	.6733 (6)	.5675 (6)	.6496 (5)	.5522 (6)	.7122 (5)
PCP [21]	.5226 (5)	.8650 (3)	.6082 (6)	.9014 (3)	.7245 (4)	.7785 (3)	.5879 (5)	.8322 (3)	.7374 (4)	.7286 (4)
LSD-GSRPCA [15]	.7613 (4)	.8371 (4)	.9050 (4)	.8357 (6)	.7313 (3)	.7362 (4)	.7222 (4)	.5842 (6)	.7214 (5)	.7594 (3)
SPGFL [16]	.8574 (2)	**.9322 (1)**	**.9856 (1)**	**.9744 (1)**	**.8840 (1)**	.8265 (2)	.7739 (3)	.8394 (2)	.8029 (2)	.8751 (2)
DSPSS	**.8993 (1)**	.9105 (2)	.9674 (2)	.9228 (2)	.7680 (2)	**.8499 (1)**	**.8593 (1)**	**.8922 (1)**	**.9163 (1)**	**.8873 (1)**

The SABS dataset [9] presents synthetic image sequences divided into nine categories. As can be seen in the results in Table 2, our algorithm takes the first place in all the scenarios except for *light switch*, since our background model has slowly adapted to changes in this scene.

The i2R dataset [11] dataset results can be seen in Table 3. We achieve top performance again overall for this dataset. We have reported our results without parameter tuning, although the dataset allows this.

5 Conclusion

We have presented a new background subtraction method and validated its efficacy and effectiveness with extensive testing. The method is based on an existing model, namely RPCA, but with new sparsity-inducing norms and group-structured sparsity constraints. Our model surpasses state-of-the-art

performance by taking advantage of the natural shape and structure of objects in the scene, where our sparsity model dynamically evolves to best describe genuine foreground objects in the scene; this gives us a significant advantage when it comes to handling dynamic backgrounds, foreground aperture, and camouflage. Moreover, a novel tandem initialization method is proposed to speed up convergence and remove ghosting effects persisting in RPCA-based methods. Specifically, our model is able to learn a robust background model that can change over time, to cope with a variety of scene changes, in comparison with the existing more heuristic RPCA-based methods. It proves itself to have excellent performance in dealing with heavy noise, thanks to the approximated RPCA model where the residual error is discarded into a third matrix in the decomposition as noise. In addition, estimation of background motion induced by a jittering or moving camera is performed simultaneously with low-rank approximation, that results in excellent performance in shaky videos. Certain improvements can be considered. Our model is yet another batch method, as the frames need to be stored for obtaining a background model. Sudden illumination changes are slowly adapted by the background model, and hence it fails to handle some indoor lighting changes. Furthermore, a more sophisticated model should be able to handle shadows, that are not interesting for later processing. Solutions to these problems could be adapted to our method.

Acknowledgments. The research leading to this paper was fully supported by the LASIE project (http://www.lasie-project.eu/) with funding from the European Unions Seventh Framework Program for research and technological development, grant agreement 607480.

References

1. Bloisi, D.D.: Background Modeling and Foreground Detection for Video Surveillance. CRC Press, Taylor and Francis Group, July 2014
2. Bouwmans, T., Zahzah, E.: Robust PCA via principal component pursuit: a review for a comparative evaluation in video surveillance. Special Isssue on Background Models Challenge, Computer Vision and Image Understanding (2014)
3. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM **58**(3), 11:1–11:37 (2011)
4. Erfanian Ebadi, S., Guerra Ones, V., Izquierdo, E.: Dynamic tree structured sparse RPCA via column subset selection for background modeling and foreground detection. In: 2016 IEEE International Conference on Image Processing (ICIP) (2016)
5. Erfanian Ebadi, S., Guerra Ones, V., Izquierdo, E.: Approximated robust principal component analysis for improved general scene background subtraction. CoRR abs/1603.05875 (2016)
6. Erfanian Ebadi, S., Guerra Ones, V., Izquierdo, E.: Efficient background subtraction with low-rank and sparse matrix decomposition. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 4863–4867, September 2015
7. Erfanian Ebadi, S., Izquierdo, E.: Approximated RPCA for fast and efficient recovery of corrupted and linearly correlated images and video frames. In: 2015 International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 49–52, September 2015

8. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2282 (2012)
9. Brutzer, S., Höferlin, B., Heidemann, G.: Evaluation of background subtraction techniques for video surveillance. In: Computer Vision and Pattern Recognition (CVPR), pp. 1937–1944. IEEE (2011)
10. Wang, Y., Jodoin, P.M., Porikli, F., Konrad, J., Benezeth, Y., Ishwar, P.: CDnet 2014: An Expanded Change Detection Benchmark Dataset. In: IEEE CVPR Change Detection workshop, United States 8 p., June 2014. https://hal-univ-bourgogne.archives-ouvertes.fr/hal-01018757
11. Li, L., Huang, W., Gu, I.Y.H., Tian, Q.: Statistical modeling of complex backgrounds for foreground object detection. IEEE Trans. Image Process. **13**(11), 1459–1472 (2004)
12. Zhou, X., Yang, C., Yu, W.: DECOLOR: moving object detection by detecting contiguous outliers in the low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell. **35**(3), 597–610 (2013)
13. Huang, J., Huang, X., Metaxas, D.: Learning with dynamic group sparsity. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 64–71. IEEE (2009)
14. Gao, Z., Cheong, L.F., Wang, Y.X.: Block-sparse rpca for salient motion detection. IEEE Trans. Pattern Anal. Mach. Intell. **36**(10), 1975–1987 (2014)
15. Liu, X., Zhao, G., Yao, J., Qi, C.: Background subtraction based on low-rank and structured sparse decomposition (2015)
16. Javed, S., Oh, S., Sobral, A., Bouwmans, T., Jung, S.: Background subtraction via superpixel-based online matrix decomposition with structured foreground constraints. In: Workshop on Robust Subspace Learning and Computer Vision, ICCV 2015 (2015)
17. Jia, K., Chan, T.-H., Ma, Y.: Robust and practical face recognition via structured sparsity. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 331–344. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33765-9_24
18. Mahoney, M.W., Drineas, P.: CUR matrix decompositions for improved data analysis. Proc. Nat. Acad. Sci. **106**(3), 697–702 (2009)
19. Zhou, T., Tao, D.: GoDec: Randomized low-rank and sparse matrix decomposition in noisy case. In: Getoor, L., Scheffer, T. (eds.) Proceedings of the 28th International Conference on Machine Learning (ICML-11), ICML 2011, pp. 33–40. ACM, June 2011
20. Guyon, C., Bouwmans, T., Zahzah, E.: Foreground detection based on low-rank and block-sparse matrix decomposition. In: International Conference on Image Processing, ICIP 2012 (2012)
21. Zhou, Z., Li, X., Wright, J., Candès, E.J., Ma, Y.: Stable principal component pursuit. CoRR abs/1001.2363 (2010)
22. Xin, B., Tian, Y., Wang, Y., Gao, W.: Background subtraction via generalized fused Lasso foreground modeling. arXiv preprint arXiv:1504.03707 (2015)
23. Haines, T., Xiang, T.: Background subtraction with dirichletprocess mixture models. IEEE Trans. Pattern Anal. Mach. Intell. **36**(4), 670–683 (2014)
24. Culibrk, D., Marques, O., Socek, D., Kalva, H., Furht, B.: Neural network approach to background modeling for video object segmentation. IEEE Trans. Neural Netw. **18**(6), 1614–1627 (2007)

25. Maddalena, L., Petrosino, A.: A self-organizing approach to background subtraction for visual surveillance applications. IEEE Trans. Image Process. **17**(7), 1168–1177 (2008)
26. St-Charles, P.L., Bilodeau, G.A., Bergevin, R.: Subsense: a universal change detection method with local adaptive sensitivity. IEEE Trans. Image Process. **24**(1), 359–373 (2015)
27. St-Charles, P.L., Bilodeau, G.A., Bergevin, R.: A self-adjusting approach to change detection based on background word consensus. In: 2015 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 990–997. IEEE (2015)
28. Barnich, O., Van Droogenbroeck, M.: Vibe: a powerful random technique to estimate the background in video sequences. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009, pp. 945–948. IEEE (2009)
29. Zivkovic, Z., van der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recogn. Lett. **27**(7), 773–780 (2006)