# Minimum Cuts of A Time-Varying Background

David Russell and Shaogang Gong
Department of Computer Science
Queen Mary, University of London
London E1 4NS, UK
{dave,sgg}@dcs.qmul.ac.uk

**Abstract**

Motivated by the demand for an effective background model, robust to non-stationary environmental changes in outdoor scenes, we present a technique using Combinatorial Optimization to extract near-optimal background estimates from blocks of temporally localized frames. Using an existing graph cut technique in conjunction with subspace analysis, we demonstrate a novel background model exhibiting results superior to those achievable with the latter technique alone, and especially suitable for background modelling in outdoor situations where variable lighting conditions prevail.

## 1 Introduction

An effective background model is a crucial first stage in most computer vision applications, especially in outdoor environments. The reliability with which it identifies potential foreground objects directly impacts on the efficiency and performance level achievable by subsequent processing stages such as tracking, recognition and threat evaluation. The nature of such a background is intrinsically statistical. Whilst the concept of statistical scene modelling suggests that there is no exact distinction between what constitutes foreground and background, a useful practical definition for surveillance in a busy urban scene is that people and the objects they cause to move are foreground, whereas buildings, fixtures, trees and permanent objects form the background. The task of the background model in such a setting is to discriminate between the two classes under a potentially wide variety of lighting conditions. Evidently, confusion might still arise, since trees sway in the wind, tending to become foreground, whilst people park their cars, which are eventually subsumed by the background. The most commonly encountered models are based on per pixel techniques such as adaptive Gaussian Mixture Models [9, 10], or subspace analysis based methods [8, 7]. Both approaches have been used with success. However, in typical implementations it is difficult to avoid such background models being contaminated by foreground scene content, eventually resulting in a less discriminative model.

On the other hand, a method detailed in [3] has been shown capable of compiling a 'background' image on a per pixel basis from a short block of input frames by casting the problem as an exercise in optimal labelling. Figure 1 shows an example of how 20 frames from a continuously busy metro ticket hall can lead to a useful background approximation. The background is drawn from parts of any of the input frames which are found to be spatially and temporally consistent. Thus the solution comprises a set of labels or pointers,

one for each pixel in the background image, specifying from which of the 20 input frames each pixel is to be taken. The method described in [3] is an application of Combinatorial Optimization [4] achieving an approximately minimum cost solution using the Minimum Cut/Maximum Flow [6] and Alpha Expansion [2] algorithms. However, prerequisites for this approach to work are: (1) that all of the required background is visible for some of the time, (2) that the required background is more consistently stable than any other foreground pixel intensity, and further (3) that each background pixel is time-independent. These conditions are not always satisfied.

To address the problem, we propose a method whereby objects which are obviously foreground, under a given definition, are eliminated from input frames before allowing those frames to contribute to the construction of a background model. We suggest that such an approach yields a 'purer' representation of the true background, and hence one with heightened sensitivity. Obviously, if this pre-processing stage were totally effective, the task of background segmentation would already have been achieved. In reality, it only offers a useful measure of pre-processing. Our solution thus consists of the pixel-labelling method described above as a stage of pre-processing, operating on short blocks of input frames to produce a temporally localized background estimation per block. These estimates are then used to build an eigenspace model. Such a hybrid approach permits the latter to 'concentrate' on dealing with lighting and shadow changes rather than being contaminated with objects like cars and people which are considered to be foreground.

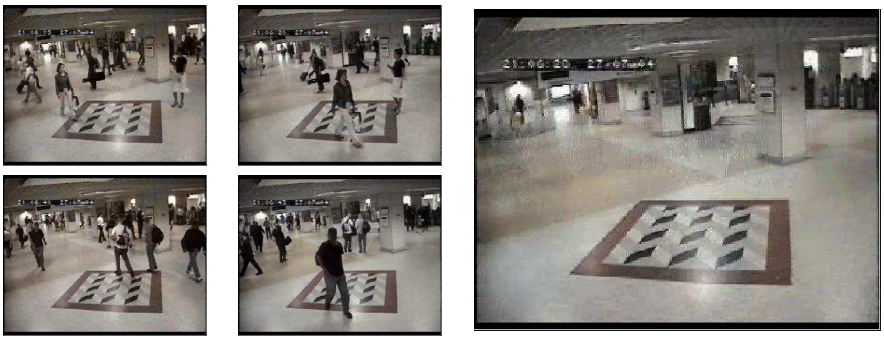

Figure 1: Left: 4 of the 20 input frames. Right: Recovered background.

## 2 Combinatorial Optimization

Given a temporally localized set of $F$ input frames of a scene each of $\mathcal{P}$ pixels, we wish to form an output image $I_{\mathcal{B}}$ to best represent the scene's background at that time. Thus we desire a set of labels $\mathcal{F}$, consisting of one label per pixel, specifying from which input frame that pixel is to be taken. Evidently, the number of possible combinations is large, but finite. In essence, the idea is to assign a cost to each choice of label (1 of $F$) at each pixel, and then solve for the minimum cost over the image as a whole in order to yield the best set of background composition labels. For the algorithm to work, the cost assignment scheme for the pixels has to reflect lower costs for better combinations of labels. This is forced by penalizing poor temporal or spatial correlation between adjacent pixels.

The Ford-Fulkerson algorithm [6] permits exact solution of a combinatorial optimization problem in polynomial time by a minimum graph cut (Min-Cut) in a situation where there are only two class labels. Having defined a suitable costing model, an undirected graph may be constructed for the background image consisting of a node for each pixel, plus two extra nodes known as the *source* and the *sink*, representing these two class labels. The pixel costs become the arc weights on the graph. However, we have $F$ class labels representing our block of input frames, where $F$ might typically be larger than ten or more. Although the exact solution of such a problem is possible, it has been shown to be NP-hard [2]. Instead, an approximate solution can be obtained rather more efficiently by applying the Min-Cut algorithm iteratively, with each class label taking its turn to be the source ($\alpha$), whilst the other $F-1$ class labels become the sink ($\alpha'$), as shown in Figure 2.

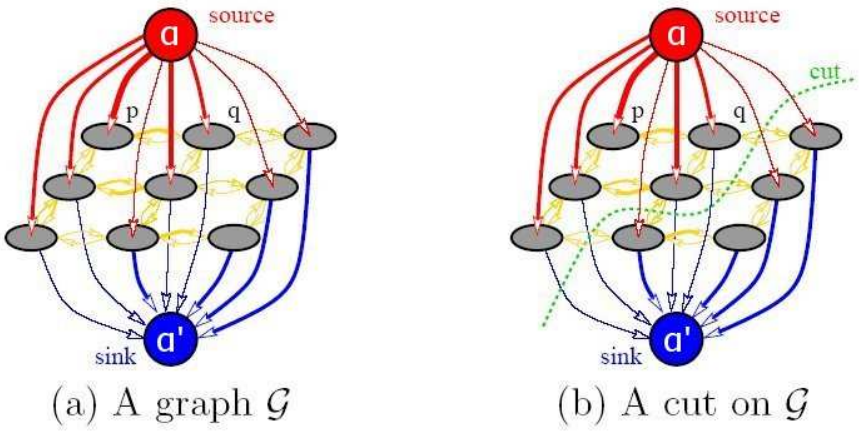


Figure 2: Graph for an array of only 9 pixels, incorporating *source* and *sink* nodes representing the two classes $\alpha$ and $\alpha'$. Weights *between* pixels stem from spatial continuity, whilst those connecting to the *source* and *sink* relate to temporal and motion continuity. The actual graph contains a node for *every* image pixel. Figure taken from [1].

Under this scheme, at any given iteration, a pixel might already belong to the class label which is currently taking its turn at being $\alpha$. In this case, the weight (cost) linking it to $\alpha$ is made infinite, so that the pixel cannot leave the class label at this iteration. The overall result is that as $\alpha$ takes on each class label $F$, pixels from all the other class labels may leave in order to join $\alpha$, but none may leave $\alpha$. This is known as $\alpha$*-expansion* which has been shown by Boykov et al. [2] to lead to an approximately minimum cost labelling solution after a number of cycles of $\alpha$ through $F$. The optimal graph cut at any iteration is then obtained by a process drawing an analogy with network flow, in which arc weights are considered flow capacities, the objective being to achieve maximum flow (Max-Flow) from source to sink. Under this condition, the arcs which are saturated (i.e. have reached their flow capacity) are those which should be cut to achieve the optimal partitioning in the equivalent Min-Cut problem. To arrive at this situation, flow is added to the network incrementally in an iterative fashion until no further addition is possible.

# 3 A Hybrid Pixel-Labelling and Subspace Model

## 3.1 Labelling Cost Functions

Following the notation of [3], a set of input $F$ frames are denoted as $I_1, I_2, \ldots, I_F$, and $I_f(p)$ is a colour intensity vector at pixel $p$ where $p \in \mathcal{P}$ is the set of pixels in an image. A given labelling is defined as $\mathcal{F} = \{f_p\}_{p\in\mathcal{P}}$. The background estimation is formed by taking a pixel intensity vector $p$ from input frame $f_p^*$ such that $I_{\mathcal{B}} = I_{f_p^*}(p)$ where $\{f_p^*\}$ is the set of labels corresponding to minimum cost. The cost of a given labelling $\mathcal{F}$ is the energy function

$$E(\mathcal{F}) = \sum_{p\in\mathcal{P}} D_p(f_p) + \sum_{\{p,q\}\in\mathcal{N}} V_{pq}(f_p, f_q) \tag{1}$$

consisting of terms relating respectively to *temporal smoothness* at pixel $p$, and *spatial smoothness* between pixels $p$ and $q$ in a neighbourhood $\mathcal{N}$ around $p$. The temporal smoothness term $D_p(f_p)$ consists of two parts

$$D_p(f_p) = D_p^S(f_p) + \beta D_p^C(f_p) \tag{2}$$

where $\beta$ controls the balance between $D^S$ and $D^C$. The first $D_p^S(f_p)$ penalizes choice of frames where the local temporal variance, evaluated over $2r$ frames, for a pixel averaged over the three colour components is high, so that

$$D_p^S(f_p) = \min\left(Var_{f_p-r\ldots f_p}(p), Var_{f_p\ldots f_p+r}(p)\right) \tag{3}$$

The second part $D_p^C(f_p)$, known as the consistency cost, penalizes choice of frames in which there is a motion boundary for a pixel. We penalize choice of a frame $f_p$ if, at the pixel in question, it contains significant temporal difference $M_{f_pf} = \left\|I_{f_p} - I_f\right\|_2$ from another frame $f$, but at the same time, the latter contains little spatial difference. A large ratio in the gradients of $M$ and $I$ implies a moving object in frame $f_p$, which we would want to exclude from our background. Using the square of the $L_2$ norm, this ratio is defined

$$\Omega_{f_pf}(p) = \frac{\left\|\nabla M_{f_pf}(p)\right\|^2}{\left\|\nabla I_f(p)\right\|^2 + \varepsilon^2} \tag{4}$$

The small constant $\varepsilon$ prevents the denominator from being zero, and ensures a low cost when there is little gradient in either $M$ or $I$. Confidence about the identification of motion in $f_p$ is gained by averaging $\Omega_{f_pf}$ over all frames

$$D_p^C(f_p) = \frac{1}{F}\sum_{f=1}^{F} \Omega_{f_pf}(p) \tag{5}$$

The spatial continuity cost between two pixels $p$ and $q$ for two input frames $f_p$ and $f_q$ is

$$V_{pq}(f_p, f_q) = \rho\left(\frac{\left\|I_{f_p}(p) - I_{f_q}(p)\right\|^2 + \left\|I_{f_p}(q) - I_{f_q}(q)\right\|^2}{2\times(\text{number of colour planes})}\right) \tag{6}$$

The penalty of choosing $f_p$ and $f_q$ as different source frames for two neighbouring pixels $p$ and $q$ will be small if the frames differ by little in the vicinity of $p$ and $q$, thus encouraging the switch from copying from one frame to another. Such a region is quite likely to represent background in this case. The constant $\rho$ controls the balance between $V$ and the temporal cost $D$.

### 3.2 Subspace Modelling of Min-Cut Labelled Background Pixels

From a sequence of $M$ input frames of size $h \times v$ pixels, we draw overlapping blocks of $F$ frames to which we apply the above background recovery algorithm, yielding $N = M - F + 1$ candidate background frames $I_{\mathcal{B}_1}, I_{\mathcal{B}_2} \ldots, I_{\mathcal{B}_N}$. Thus $I_{\mathcal{B}_1}$ is derived from input frames $1 \ldots F$, $I_{\mathcal{B}_2}$ from frames $2 \ldots F+1$ and so on. The background images are then rasterized to form column vectors $\mathbf{x}_1 \ldots \mathbf{x}_N$ each of length $hv$ elements. The mean vector $\mathbf{m}$ of $\{\mathbf{x}_1 \ldots \mathbf{x}_N\}$ is determined as $\mathbf{m} = \frac{1}{N}\left(\sum_{i=1}^{N} \mathbf{x}_i\right)$. After mean subtraction, the vectors $\{\mathbf{x}_1 \ldots \mathbf{x}_N\}$ are concatenated horizontally to form a matrix $\mathbf{X} = [\mathbf{x}_1 - \mathbf{m}, \mathbf{x}_2 - \mathbf{m} \ldots, \mathbf{x}_N - \mathbf{m}]$. The covariance matrix for the background vectors $\mathbf{x}_n$ where $1 \leq n \leq N$ is then given by the outer product of $\mathbf{X}$ with itself $\mathbf{C} = \mathbf{X}\mathbf{X}^T$ with eigenvectors $\mathbf{v}_i$ and eigenvalues $\lambda_i$ where $1 \leq i \leq N$

$$\mathbf{X}\mathbf{X}^T\mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{7}$$

However, such a matrix would contain $(hv)^2$ elements but only have a rank of at maximum $N$. In this case we take advantage of the low dimensional method in [7], whereby we pre-multiply Equation (7) by $\mathbf{X}^T$ and find that the much smaller matrix $\mathbf{X}^T\mathbf{X}$ of size $N \times N$ has the same eigenvalues as $\mathbf{X}\mathbf{X}^T$ and eigenvectors $\mathbf{u}_i = \mathbf{X}^T\mathbf{v}_i$

$$\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{v}_i) = \lambda_i(\mathbf{X}^T\mathbf{v}_i) \tag{8}$$

Thus we perform eigendecomposition on $\mathbf{C}' = \mathbf{X}^T\mathbf{X}$, and retain the $K$ eigenvectors corresponding to the largest eigenvalues of $\mathbf{C}'$ such that $\frac{\sum_{i=1}^{K}\lambda_i}{\sum_{i=1}^{N}\lambda_i} \geq \gamma$ to form a normalized approximate model $\mathbf{V} = \left[\frac{\mathbf{X}^T\mathbf{v}_1}{|\mathbf{X}^T\mathbf{v}_1|}\frac{\mathbf{X}^T\mathbf{v}_2}{|\mathbf{X}^T\mathbf{v}_2|}\ldots\frac{\mathbf{X}^T\mathbf{v}_K}{|\mathbf{X}^T\mathbf{v}_K|}\right]$ where $\gamma$ represents a given fraction of the original energy. A new image vector $\mathbf{y}$ may then be segmented into foreground and background by projecting into the subspace spanned by $\mathbf{V}$ to determine what parts of it are supported by the model. Re-projecting back into the image space and subtracting from the original image $\mathbf{y}$ leaves the residual image vector $\mathbf{r}$ as

$$\mathbf{r} = (\mathbf{y} - \mathbf{m}) - \mathbf{V}\left(\mathbf{V}^T(\mathbf{y} - \mathbf{m})\right) \tag{9}$$

Thresholding each element $p$ of $\mathbf{r}$ against a constant $\tau$ yields a binary vector $\mathbf{B}$, that may be de-rasterized to the original image aspect ratio to form a binary segmentation mask, which is

$$B_p = \begin{cases} 1 & \text{if } r_p > \tau \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

## 4 Experiment

In order to demonstrate the effectiveness of our scheme, we compared the performance of a subspace model derived from pre-filtered backgrounds obtained by the Min-Cut optimization (the 'Min-Cut + Subspace' method) with that of a subspace model built directly from the $N$ input frames (the 'Direct Subspace' method). We forced both systems to use only 14 eigenvectors, a number which permitted the former to represent 80% of its original covariance energy. In addition, we show the result of using the Min-Cut *alone* on frames taken from the input sequence (the 'Min-Cut Only' method).

Figure 3: Examples illustrating typical level of activity in the chosen urban road scene.

## 4.1 Dataset

For our experiment, we chose a very busy urban scene at a road junction by a metro station containing continuous activity involving both people and vehicles (see Figure 3). Video data was collected over a one hour period in colour at a frame rate of 25Hz, producing 90,000 RGB image frames at a spatial resolution of $720 \times 576$ and 8 bit intensity resolution per colour. For the purpose of model building, every 300th frame was extracted from this to provide a set of $N = 300$ images taken at 12 second intervals.

Using $F = 20$ input frames to evaluate each pre-filtered background, the Min-Cut + Subspace model was constructed using 280 images, whilst the Direct Subspace model used the 300 unprocessed input frames. To accelerate the Min-Cut labelling process, the input frames were sub-sampled to $360 \times 288$ resolution. Although the resultant label set consisted of only this number of elements, the backgrounds were reconstructed using 1 label per 4 pixels in order to preserve the original image resolution. The cost balancing constants for the Min-Cut process were set at $\beta = 1$ and $\rho = 4$, whilst in the consistency cost calculation $\varepsilon = 1$. The threshold for segmentation in both Min-Cut + Subspace and Direct Subspace methods was 20, given that the intensity range for the RGB data was $[0, 255]$. The Min-Cut Only method used 20 frames from the input sequence taken at 3 minute intervals, the binary mask being given by thresholding the difference from the single recovered background. Finally, for all methods, the binary masks were filtered to remove single and small groups of pixels before display.

## 4.2 Results

The graph in Figure 4 illustrates the cumulative distribution of energy (eigenvalues) among the eigenvectors of the covariance matrix for the Min-Cut + Subspace and Direct Subspace models. We note that the former requires considerably fewer eigenvectors to reach a certain energy fraction, thus supporting the idea that the proposed hybrid technique attains a more compact model. The sharp rise of the Min-Cut + Subspace curve for energy fractions above 0.95 here strongly indicates the dominance of a small number of eigenvectors in the model, as intended.

Figure 5 shows typical output from the Min-Cut pre-processing stage. As illustrated by the left image, foreground object removal is not always complete. If the 20 input

frames used to produce this particular background estimation contain the stationary car in most frames, it will be indistinguishable from the background. Although such foreground objects still contaminate the subsequent subspace model, the pre-processing removes so much of the foreground clutter that overall, considerable advantage is gained.

The segmentation mask $B$ for two typical input frames, which were *not* used to build the models, are shown in Figure 6 for all three cases of the experiment. The Min-Cut + Subspace model clearly demonstrates the cleanest segmentation of objects which, for a typical surveillance application, are required to be foreground. For the Direct Subspace model, some of the road markings and shadows from the traffic signals are breaking through into the foreground. The level of foreground clutter contaminating the model and the limited expressive power of having so few available eigenvectors result in desensitization and poor discrimination.

For the Min-Cut Only experiment, the images at the bottom of Figure 6 show problems with shadows. Because no variability is catered for in the single recovered image, the changing shadows at the edges and walls of buildings have not been accommodated well. Since the result of the optimization, a single image compiled from images taken throughout the whole hour of the input video, the chances of a lighting match with a single arbitrary input frame is small. Different parts of the background model will match different lighting conditions.

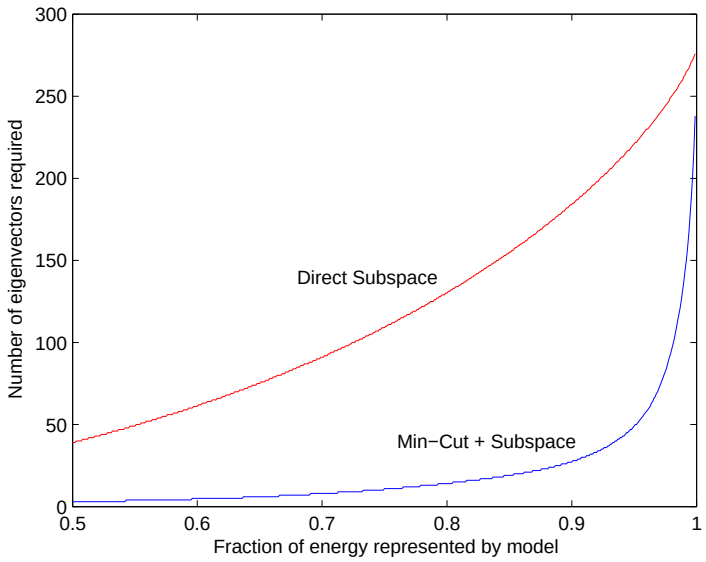


Figure 4: Graph showing that Min-Cut + Subspace consistently requires considerably fewer eigenvectors to retain a certain fraction of energy than the Direct Subspace method.

Figure 5: Typical output from the Min-Cut pre-processing stage. Left: Imperfect object removal. Right: Near optimal background recovery.

# 5 Discussion and Further Work

The success of the hybrid Min-Cut algorithm may be explained by consideration of its two constituent parts separately. The more effectively that we can eliminate foreground objects from the 'background' images by the Min-Cut stage, the more compact becomes the eigenspace model for a given energy fraction. The Min-Cut process can only remove foreground objects if they are not consistently placed in the $F$ source frames. Turning this around, the true background can only be found if it is found to be dominant in relation to the costing rules defined.

There is considerable scope for determining an optimal selection of source frames from real-time incoming video. The imperfect object removal illustrated in Figure 5 is typical of what happens when the choice of source frames is unsuitable. The present method of taking $F = 20$ frames at 12 second intervals is perhaps rather arbitrary and crude. Naturally, the combinatorial optimization will take longer if we decide to use blocks of more than 20 frames, but using less frames might cause some areas of true background never to be discovered.

The optimal sampling interval will depend on the temporal content of the scene. In our example, the activity of people and cars is governed largely by the sequence of the traffic lights on the junction, the cycle time of which was measured to vary between 98s and 116s. Waiting cars accumulating at a red light could, for instance, constitute background if most of the $F$ frames were taken while the cars waited.

An altogether more intelligent way of selecting frames for the optimization stage is required in order to maximize the capability of the pre-processing for elimination of unwanted foreground. One possibility would be to add a further term to the cost function in order to exclude choice of pixels or frames which are too distant from a current version of the model. However, this should be pursued with care, since the resultant system would contain a feedback loop which may invite bootstrapping and instability problems.

Although we chose to use a subspace model for the second stage, possibilities certainly exist for incorporating other techniques. A per pixel model might need less or only one Gaussian if the pre-processing tends to reduce multi-modality in colour space. Dispensing with the Expectation-Maximization stage [5] that usually goes with Gaussian Mixture Models could lead to considerable saving in processing time.

However, we believe that the subspace model as chosen here has the best possibility of success, since it excels in modelling the global linkage of changes between pixels rather than the spatially localized disturbances which the Min-Cut stage tends to attenuate. Such a property makes it ideal for a compact model of daylight variability.

# 6 Conclusion

We have demonstrated that a hybrid background modelling scheme consisting of a pre-processing stage based on the combination of a Min-Cut/Max-Flow algorithm *and* a conventional subspace model shows advantage over the conventional subspace model operating alone. Suitable for application in outdoor environments, we have succeeded in developing a system tolerant of lighting changes, whilst showing robustness to a high level of activity in a complex scene. Although rather computationally intensive, the new algorithm produces useful improvements when running at a sub-multiple of the true frame rate. With refinements in the software architecture, it is believed that the Min-Cut + Subspace method does have a useful rôle to play in practical applications, but in any case is valuable as a vehicle for future research in this direction.

# 7 Acknowledgment

The authors would like to thank Vladimir Kolmogorov for use of his C++ implementation of the MinCut/MaxFlow algorithm which is available at:
*http://www.adastral.ucl.ac.uk/ vladkolm/software.html*.

# References

[1] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in computer vision. *IEEE PAMI*, 26(9):1124–1137, 2004.

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE PAMI*, 23(11):1222–1239, 2001.

[3] S. Cohen. Background estimation as a labeling problem. In *IEEE ICCV*, pages 1034–1041, Beijing, China, October 2005.

[4] W. Cook, W. Cunningham, W. Pulleyblank, and A. Schrijver. *Combinatorial Optimization*. John Wiley and Sons Inc, New York, 1998.

[5] A. Dempster, N.Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[6] L. Ford and D. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8:399–404, 1956.

[7] Y. Li. On incremental and robust subspace learning. *PR*, 37(7):1509–1518, 2004.

[8] N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modelling human interactions. *IEEE PAMI*, 22(8):831–843, August 2000.

[9] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE CVPR*, pages 246–252, Colorado, 1999.

[10] K. Toyama, J. Krumm, B. Brummit, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *IEEE ICCV*, volume 1, pages 255–261, Kerkyra, Greece, 1999.

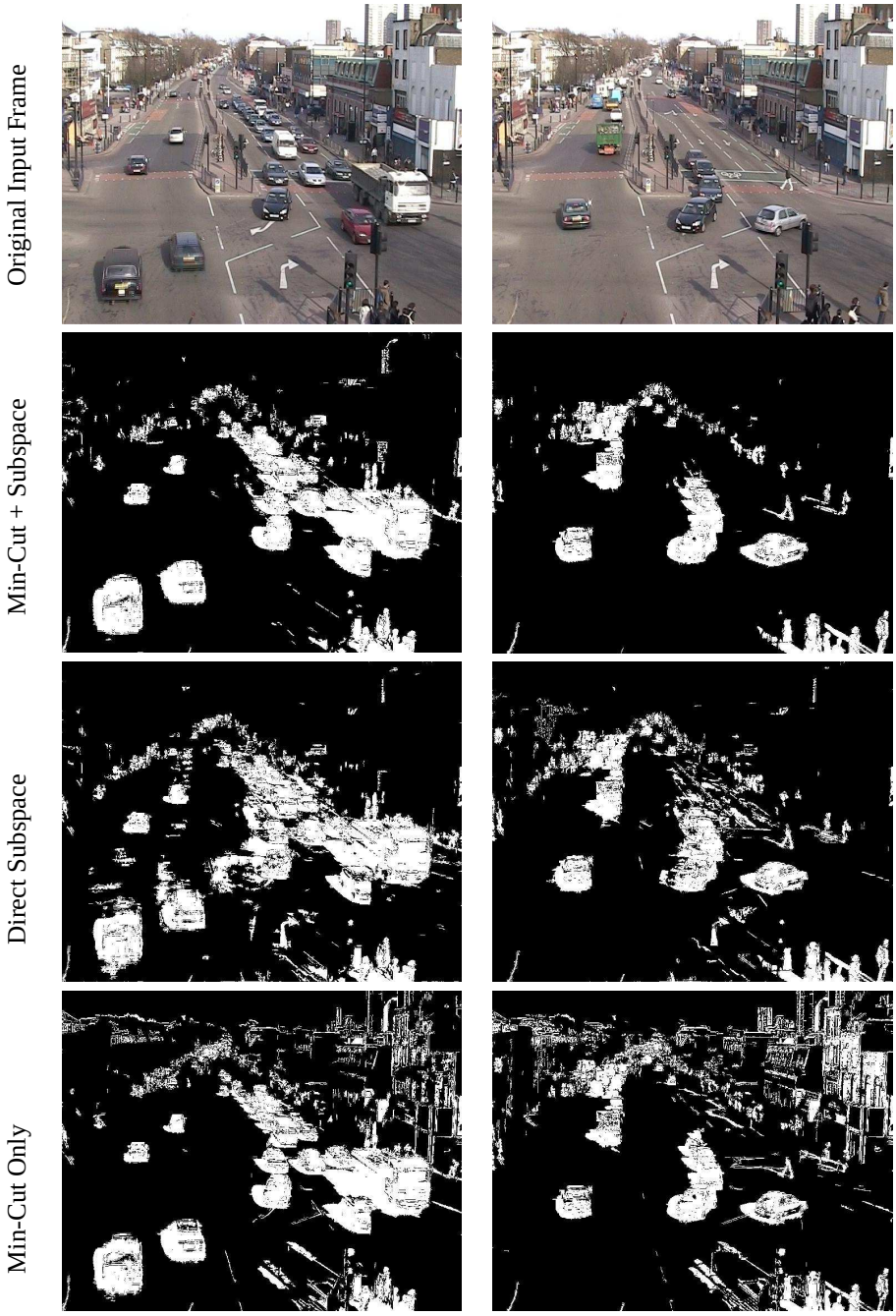


Figure 6: Segmentation of two frames using Min-Cut + Subspace, Direct Subspace, and Min-Cut Only methods. Min-Cut + Subspace shows the best segmentation here.