Background Subtraction for Freely Moving Cameras

Yaser Sheikh[1] yaser@cs.cmu.edu
Omar Javed[2] omar.javed@objectvideo.com
Takeo Kanade[1] tk@cs.cmu.edu

[1]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213
[2]ObjectVideo Inc., Reston, VA 20191

Abstract

Background subtraction algorithms define the background as parts of a scene that are at rest. Traditionally, these algorithms assume a stationary camera, and identify moving objects by detecting areas in a video that change over time. In this paper, we extend the concept of 'subtracting' areas at rest to apply to video captured from a freely moving camera. We do not assume that the background is well-approximated by a plane or that the camera center remains stationary during motion. The method operates entirely using 2D image measurements without requiring an explicit 3D reconstruction of the scene. A sparse model of background is built by robustly estimating a compact trajectory basis from trajectories of salient features across the video, and the background is 'subtracted' by removing trajectories that lie within the space spanned by the basis. Foreground and background appearance models are then built, and an optimal pixel-wise foreground/background labeling is obtained by efficiently maximizing a posterior function.

1. Introduction

Fundamentally, the objective of background subtraction algorithms is to identify interesting areas of a scene for subsequent analysis. "Interesting" usually has a straightforward definition: objects in the scene that move. The most effective method of isolating these objects is to ensure that motion in the scene *exclusively* translates into motion in video data. This has been achieved by the near ubiquitous assumption in modern surveillance systems of stationary (or nominally stationary) cameras, [11, 34, 4, 5, 27, 2, 12, 26]. The success of these algorithms has led to the growth of the visual surveillance industry, forming the foundation for tracking, object recognition, pose reconstruction, and action recognition. The assumption of camera stationarity, however, severely limits the application of computer vision algorithms — a limitation that is becoming increasingly significant with the growing proliferation of moving camera platforms, like cellular phones, vehicles, and robots. As a larger and larger percentage of video content is produced by moving cameras, the need for foundational algorithms that can isolate interesting areas in such video is becoming increasingly pressing. In this paper, we present an algorithm that takes the definition of "interesting" as movement in the 3D world and extends it to video captured from a freely moving camera. The algorithm takes raw video from a moving camera as input, and outputs a binary mask of moving objects for each observed frame (illustrated in Figure 1).

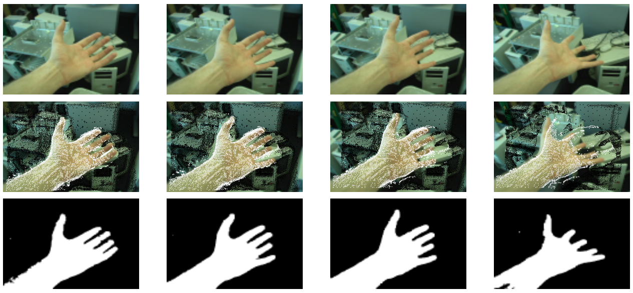


Figure 1. The problem tackled in this paper is to produce pixel-wise segmentations that distinguish between camera induced motion and object induced motion, in general (non-planar) scenes. The top row shows raw video from a moving camera. Points are tracked across the video and segmented into areas at rest and objects in motion (middle row). These sparsely segmented points are used to build background and foreground appearance models, for pixel-wise segmentation (bottom row).

As Palmer notes in [21], image motion is induced by a confluence of camera motion, independent object motion, and the 3D structure of the scene. The fundamental challenge addressed in this paper is the disambiguation of image motion induced by the motion of the camera and image motion influenced by the motion of independently moving objects. The core intuition of our algorithm is that objects in motion can be reliably differentiated from objects at rest at *sparse* locations using geometric constraints, such as the rank constraint for orthographic cameras, [29] or the fundamental polynomial constraint [6] for projective cameras. These sparse locations can then be used to build foreground and background appearance models, which can, in turn, be used to segment independently moving objects.

2. Related Work

The literature on background subtraction and motion segmentation is vast and we have limited this review to major themes. The earliest approach to background subtraction originated in the late 70s with the work of Jain and Nagel [11], who used frame-differencing for the detection of moving objects. Subsequent approaches proposed a progression of probabilistic models for handling uncertainty in background appearance, like the per-pixel Gaussian model of background appearance by Wren *et al.* in [34], Kalman Filters for updating pixel colors in [16] and [14], Gaussian mixture models in [27], non-parametric kernel density estimates by Elgammal and Davis in [2], and the joint spatial-color model by Sheikh and Shah in [26]. In all these approaches, the unifying conceptual theme was the definition of background: areas of the scene that remain at rest. One important variation of this definition was structured dynamism in the background (e.g. waves in water bodies, foliage in the wind, and nominal camera motion), for which various probabilistic models were proposed, (Monnet *et al.* [20], Zhong and Sclaroff [38], and Mittal and Paragios [19]). The definition of background as static (or nominally static) led to a common requisite that the camera remain stationary for the duration of observation.

Research into relaxing this assumption has largely relied on ego-motion compensation, [10], [23], [18], [8], and [22]. A homography or a 2D affine transform is used to compensate for motion and various ideas from conventional background subtraction are applied to detect foreground regions. The scope of these methods is restricted to scenes where the background can be well approximated by a plane or where the camera motion is restricted to pan, tilt, or zoom, i.e. motions where the camera center does not translate. For cases where the camera may translate and rotate, several strategies have been pursued. In the plane+parallax framework ([9, 25, 36]), a homography is first estimated between successive image frames. The registration process removes the effects of camera rotation, zoom, and calibration. The residual pixels, correspond either to moving objects or to static 3D structure with large depth variance (parallax pixels). To estimate homographies, these approaches assume the presense of a dominant plane in the scene, and have been successfully used for object detection in aerial imagery where this assumption is usually valid. Layer-based methods [33, 28, 15, 35] model the scene as piece-wise planar scenes, and cluster segments based on some measure of motion coherency. Yuxin *et al.* [37] use a layer-based approach explicitly for background subtraction from moving cameras but report low performance for scenes containing significant parallax (3D scenes). Finally, motion segmentation approaches like [30, 3, 13, 32] sparsely segment point trajectories based on the geometric coherency of motion.

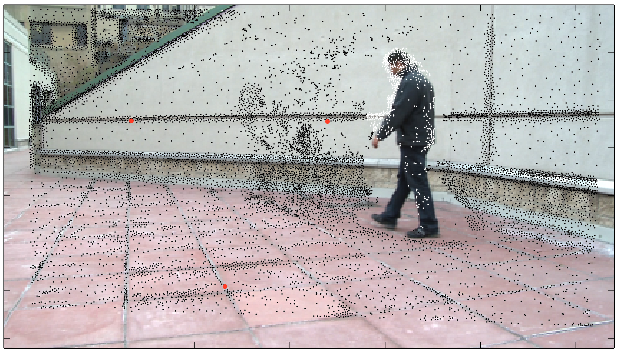


Figure 2. Background trajectories are denoted by black points and the foreground trajectories are denoted by white points. The three red dots denote the three trajectories selected as the background trajectory basis.

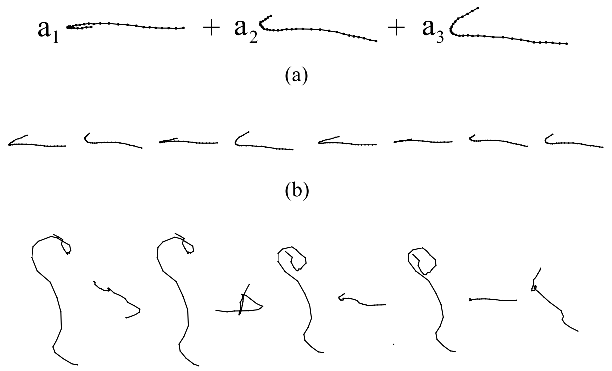
In contrast to all these approaches, the goal of our approach is to extend the conventional definition of background – areas of the scene at rest – to moving cameras and to handle (a) a range of foreground objects sizes; (b) both rigid and non-rigid foreground objects; and (c) fully 3D backgrounds. Our method provides pixel-wise labeling of foreground and background on challenging sequences taken from hand-held cameras.

3. Rank-Constraint for the Background

In an environment at rest, the motion induced in video depends only on the 3D structure of the scene and the motion of the camera. The geometric constraints that this image motion must satisfy are well-understood, [7]. If P salient points have been tracked across a sequence of frames, the trajectory of the i-th point can be represented as $\mathbf{w}_i = [\mathbf{x}_{1i}^T \cdots \mathbf{x}_{Fi}^T] \in \mathbf{R}^{1\times 2F}$, where $\mathbf{x}_{fi} = [u_{fi} \;\; v_{fi}]^T$ in each frame f. The set of these trajectories can be arranged into a registered $2F \times P$ matrix,

$$\mathbf{W}_{2F\times P} = [\mathbf{w}_1^T \cdots \mathbf{w}_P^T]^T = \begin{pmatrix} u_{11} & \dots & u_{1P} \\ v_{11} & \dots & v_{1P} \\ \vdots & & \vdots \\ u_{F1} & \dots & u_{FP} \\ v_{F1} & \dots & v_{FP} \end{pmatrix}. \tag{1}$$

In the noiseless case, and under an assumption of orthographic projection, $\mathbf{W}$ is a rank 3 matrix, [29]. The rank constraint arises from the fact that this matrix can be factored into a $3\times P$ structure matrix of 3D points, and a $2F\times 3$



(a)

(b)

(c)

Figure 3. (a) The background trajectory basis of a 30 frame window for the Hand sequence (shown in Figure 1). In the noiseless case, background trajectories (b) lie in this space, and foreground trajectories (c) do not (with the exception of accidental alignments). In the presence of noise, the projection error measured the probability of association.

orthogonal matrix,

$$\mathbf{W} = \left(\begin{array}{ccc} r_{11} & r_{13} & r_{13} \\ r_{14} & r_{15} & r_{16} \\ \vdots & & \vdots \\ r_{F1} & r_{F2} & r_{F3} \\ r_{F4} & r_{F5} & r_{F6} \end{array} \right) \left(\begin{array}{ccc} X_1 & \dots & X_P \\ Y_1 & \dots & Y_P \\ Z_1 & \dots & Z_P \end{array} \right). \quad (2)$$

A useful way of looking at the rank constraint is that all 2D trajectories projected from areas at rest in the world lie in a subspace spanned by three basis trajectories,

$$\mathbf{w}_i = \sum_{i=1}^{3} a_i \hat{\mathbf{w}}_i, \quad (3)$$

where $\hat{\mathbf{w}}_i$ is the ith basis trajectory. We refer to this as the background trajectory space since the projected trajectory of any stationary point in 3D must lie in this space. Further, from Equation 2, the background trajectory space is of dimension 3. Figure 3 shows the estimated background trajectory basis and exemplars of foreground and background trajectories for a 30 frame window.

If there are independently moving objects in the scene, the rank of $\mathbf{W}$ will, in general, be greater than three. We leverage this constraint to find the subset of columns (and therefore trajectories) that arise from the stationary parts of the scene — all trajectories that are projections of stationary points must lie in the background trajectory space, all those that are projections of independently moving objects will not barring degenerate cases. RANSAC is used to robustly compute the best estimate of the three dimensional trajectory subspace while identifying points that lie within the space.

During RANSAC, a set of three trajectories $\mathbf{w}_i$, $\mathbf{w}_j$, and $\mathbf{w}_k$ is randomly selected. The fitting function that is used to establish consensus is the projection error on the three dimensional subspace spanned by the selected trajectories. The matrix of selected trajectories $\mathbf{W}_3 = [\mathbf{w}_i^T \mathbf{w}_j^T \mathbf{w}_k^T]$, is used to construct a projection matrix,

$$\mathbf{P} = \mathbf{W}_3(\mathbf{W}_3^T \mathbf{W}_3)^{-1} \mathbf{W}_3^T. \quad (4)$$

This projection matrix is used to evaluate the likelihood that a given trajectory $\mathbf{w}_i$ belongs to the background, by measuring the projection error.

$$f(\mathbf{w}_i | \mathbf{W}_3) = \|\mathbf{P}\mathbf{w}_i - \mathbf{w}_i\|_2. \quad (5)$$

If there is enough of a consensus in the data to support the selected trajectories, the routine terminates. Otherwise, another subset of three trajectories are selected, and the process is repeated until a consensus set is found. This process provides the background trajectory basis of the 3 dimensional subspace, an inlier set of n trajectories corresponding to the background, and an outlier set of m trajectories corresponding to the foreground.

Figure 3 shows three trajectories selected from the Hand sequence in Figure 1 for a 30 frame window, and examples of trajectories in the sequences that lie in the subspace spanned by these trajectories (background trajectories) and those that don't (foreground trajectories). Due to occlusion, noise, and varying camera motion, estimated trajectories will typically vary considerably in length. A feature of the proposed approach is that we do not require factorization algorithms like SVD which cannot directly handle missing data. For sustained labeling in longer videos we take a sliding window approach, where the labeling of each frame is computed using trajectories in a temporal window of frames. The trajectories within this window are used to compute the trajectory basis. This ensures that erroneous parts of trajectories (such as their occlusion or exit) do not render the entire trajectory useless.

4. Building Background/Foreground Models

The objective of the algorithm is to produce a binary labeling $\mathcal{L} = [l_1 \cdots l_N]$ for an image with N pixels, given the background and foreground trajectories. We wish to estimate,

$$\mathcal{L}^* = \arg\max_{\mathcal{L}} p(\mathcal{L}|\mathbf{x}). \quad (6)$$

Applying Bayes Theorem and assuming conditional independence, we can factor the term as,

$$p(\mathcal{L}|\mathbf{x}) \propto p(\mathcal{L}) \prod_{i=1}^{N} p(\mathbf{x}_i | l_i). \quad (7)$$

The likelihood $p(\mathbf{x}|\mathcal{L})$ is estimated as,

$$p(\mathbf{x}_i | l_i) = p(\mathbf{x}_i | \psi_b)^{(l_i - 1)} p(\mathbf{x}_i | \psi_f)^{l_i}, \quad (8)$$

Algorithm 1 Given an input video, identify pixels belonging to moving objects

```
Sparse Labeling
Track $P$ points across $F$ frames
$t \Leftarrow 0$
while $t < T$ do
    Randomly select 3 trajectories $[\mathbf{w}_i, \mathbf{w}_j, \mathbf{w}_k]$
    Compute a projection matrix $\mathbf{P}$
    Find inlier trajectories $\{\mathbf{w}_l\}_{1=1}^{l}$
    if $l > d$ then
        Break
    end if
    $t \Leftarrow t + 1$
end while
Pixel-wise Labeling
Create background model $\psi_b$ from inliers
Create foreground model $\psi_f$ from outliers
for $i$ = 1 to $N$ do
    Compute $p(\mathbf{x}_i|\psi_f)$ and $p(\mathbf{x}_i|\psi_b)$
end for
Maximize a posterior using likelihoods $p(\mathbf{x}_i|\psi_f)$ and $p(\mathbf{x}_i|\psi_b)$, with an MRF prior using graph cuts
```

where $p(\mathbf{x}_i|\psi_b)$ is the probability of the pixel belonging to the background and $p(\mathbf{x}_i|\psi_f)$ is the probability of the pixel belonging to the foreground. We use background trajectories and foreground trajectories to create background and foreground appearance models. For a frame, the background trajectory points at that frame are used to construct a background model $\psi_b = [\mathbf{y}_1 \cdots \mathbf{y}_n]$ where each $\mathbf{y}_i$ is a joint color-location vector, i.e. $\mathbf{y}_i = [r_i\ g_i\ b_i\ x_i\ y_i]$. (r, g, b) defines a color in rgb space, and (x, y) is the location of the pixel in the image. The probability of a candidate pixel $\mathbf{x}$ belonging to the background is then evaluated using kernel density estimation,

$$p(\mathbf{x}|\psi_b) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{\mathbf{H}}(\mathbf{x} - \mathbf{y}_i), \tag{9}$$

where $\varphi_{\mathbf{H}}(\cdot)$ is,

$$\varphi_{\mathbf{H}}(\mathbf{z}) = |\mathbf{H}|^{-\frac{1}{2}}\varphi(\mathbf{H}|^{-\frac{1}{2}}\mathbf{z}),$$

$\varphi(\cdot)$ is a kernel function like the Normal or Epanechnikov kernel and $\mathbf{H}$ is a symmetric positive definite bandwidth matrix. Often this bandwidth matrix is estimated adaptively, which vary at various locations in the distribution, i.e.,

$$p(\mathbf{x}|\psi_b) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{\mathbf{H}(\mathbf{x})}(\mathbf{x} - \mathbf{y}_i). \tag{10}$$

The benefit of the adaptive bandwidth estimator is that it removes the need for manual selection of the bandwidth parameters. Furthermore, the bandwidth parameters varies in different areas of space depending on the density of points. This typically leads to gains in classification accuracy, as demonstrated during experiments, primarily because the density of salient points in different areas of the video are different. Similarly, foreground trajectory points are used to construct a foreground model $\psi_f = [\mathbf{x}_1 \cdots \mathbf{x}_m]$ and the probability of a candidate pixel $\mathbf{x}$ belonging to the foreground is then evaluated using,

	Person	Car	Hand
Fixed $\mathbf{H}$ (manual)	[0.61, 0.83]	[.69, .92]	[.96,.97]
Fixed $\mathbf{H}$ (manual) w/ MRF	[0.71, .80]	[.79, .88]	[.92,.97]
Adaptive $\mathbf{H}_i$ (auto)	[.77, .92]	[.50,.95]	[.83,.99]
Adaptive $\mathbf{H}_i$ (auto) w/ MRF	[.80, .95]	[.71,.92]	[.83,.99]

Table 1. Performance table [precision, recall]: Comparing different strategies to compute likelihood and estimate labeling.

$$p(\mathbf{x}|\psi_f) = \frac{1}{m}\sum_{i=1}^{m} \varphi_{\mathbf{H}(\mathbf{x})}(\mathbf{x} - \mathbf{z}_i). \tag{11}$$

A pairwise Markov Random Field is used to enforce smoothness in the labeling as the prior $p(\mathcal{L})$,

$$p(\mathcal{L}) \propto \exp\left(\sum_{i=1}^{N}\sum_{i=1}^{N} \lambda\Big(l_i l_j + (1 - l_i)(1 - l_j)\Big)\right), \tag{12}$$

where λ is a parameter that determines the degree of smoothness imposed by the prior. Combining the prior and likelihood terms, the log-posterior is,

$$\begin{aligned}\log p(\mathcal{L}|\mathbf{x}) &= \left(\sum_{i=1}^{N}\sum_{i=1}^{N} \lambda\Big(l_i l_j + (1 - l_i)(1 - l_j)\Big)\right) \\ &\quad + \sum_{i=1}^{N} \log\left(\frac{p(\mathbf{x}_i|\psi_f)}{p(\mathbf{x}_i|\psi_b)}\right) l_i. \end{aligned} \tag{13}$$

The space of solutions is large (2^N) which precludes an exhaustive search. The globally optimal solution can be efficiently computed using graph-cuts, [1, 17].

5. Results

The algorithm was tested on a variety of sequences with hand held cameras, both indoors and outdoors, containing a variety of nonrigidly deforming objects like hands, faces, and bodies, shown in Figure 5. These sequences are high resolution images with significant frame-to-frame motion — for the hand sequence in Figure 1 the average motion of a background point was 133.90 pixels, for the Car sequence in Figure 5 it was 67.10 pixels, for the Person sequence in Figure 5 it was 90.93 pixels and for the Pedestrian sequence in Figure 5 (a) it was 207.21 pixels. In these sequences,

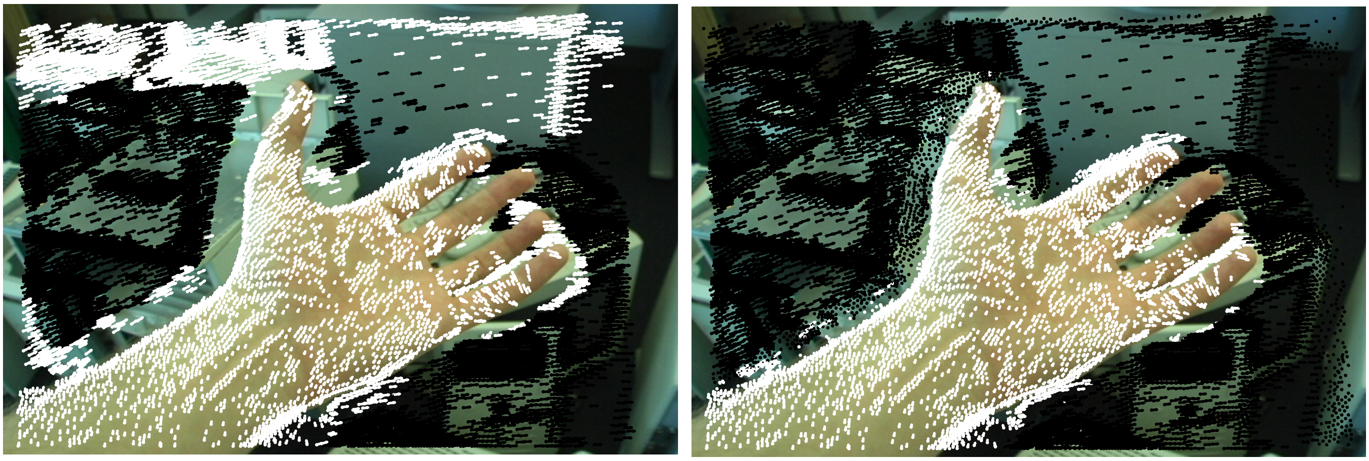


Figure 4. Background trajectory labeling based on a RANSAC fit of a homography (a) and the RANSAC fit using the motion subspace based approach proposed in this paper (b). Points on the background are shown as black arrows, and outliers to the background model are shown as white arrows. Figure 4(a) clearly indicates that a planar model is insufficient to explain the background motion.

there is significant parallax, rendering ego-motion compensation methods like [10] (as shown in Figure 4), and or neighborhood models like those in [2, 26] ineffective. The parameters in the algorithm are the RANSAC threshold t, the number of RANSAC iterations T, the temporal window size s, and the smoothing parameter λ. The values of these parameters remained constant throughout our experiments, $t = 0.01$, $T = 1000$, $s = 30$ and $\lambda = 30$. The likelihood ratio threshold used to produce the results was 1. The trajectories in these sequences were created using the particle video tracker of Sand and Teller [24]. This state-of-the-art algorithm detects a high density of points and provides high quality trajectories across the sequence.

We tested our approach quantitatively by creating ground truth segmentations of the Hand, Person and Car sequences. Table 1 shows precision and recall pairs for the three sequences for labeling using (1) likelihood ratio based labeling with a fixed bandwidth matrix whose parameters were selected manually, (2) maximum a posteriori (with a MRF prior) labeling with a fixed bandwidth matrix whose parameters were selected manually, (3) likelihood ratio based labeling with an adaptive bandwidth matrix whose parameters were selected automatically, and (4) maximum a posteriori (with a MRF prior) labeling with an adaptive bandwidth matrix whose parameters were selected automatically. This table makes two important points. First, that as the distribution of points across the video is not uniform, the selection of a uniform kernel bandwidth is inappropriate. Empirical observation confirms, for foreground regions in particular, the distribution of points can vary substantially in different areas. The adaptive bandwidth approach has the added advantage of being fully automatic and does not require bandwidth tuning. In our experiments, the bandwidth was estimated using a likelihood cross-validation method [31] with weights proportional to the k nearest neighbors, where $k = \sqrt{n}$. The second point of note, is that the MRF prior improves results both qualitatively, by cleaning up isolated pixels, and empirically, improving the precision and recall performance of the method.

6. Summary and Discussion

In this paper, we address the problem of identifying the background, i.e., parts of a scene that are at rest, in videos captured from moving cameras. We leverage the fact that all trajectories corresponding to static areas in the scene lie in a three dimensional subspace to discriminate between background and foreground areas in the scene. RANSAC is used to robustly estimate the background trajectory basis using this rank constraint, and to classify trajectories as inliers (background) or outliers (foreground). Once classified, these trajectories provide a sparse labeling of the video and are used to build background and foreground appearance models. These models are used, with a MRF prior, to estimate the maximum a posteriori pixel-wise labeling of the video by finding the minimum cut of a capacitated graph. The entire algorithm is processed using 2D quantities and although the constraints apply to a fully 3D scene, no explicit 3D reconstruction is required. There are two conceptual assumptions used in this paper: (1) An orthographic camera model is used, and (2) the background is the spatially dominant “rigid” entity in the image.

The primary limitation of the proposed approach is the use of an affine camera model over a more accurate perspective camera model. This model was used because there exists no tractable constraint for perspective cameras that simultaneously constrains motion over more than three frames. The rank constraint on the measurement matrix allows constraints from multiple frames to be used *simultaneously*. This is critical because, in practice, there is no guarantee that camera will move between two or three frames, or that objects move enough between frames. Thus, in the balance, we chose an multi-frame affine method over a two/three-frame perspective method. A convincing demonstration of the efficacy of this model is the results we have shown are on videos captured by real (and therefore perspective) cameras in real 3D scenes with parallax, typical of handheld video and outdoor urban scenes where earlier methods have not been shown to work (so far). In future work, we will investigate techniques that lift the camera model to fully projective while retaining the multi-frame nature of factorization methods for orthographic cameras. The ability to identify moving areas of interest will facilitate higher level research into problems such as action recognition and scene understanding from hand-held cameras.

Acknowledgments

We thank Peter Sand for providing data and trajectories.

References

[1] J. Besag. On the statistical analysis of dirty pictures. *J. Royal Statistical So.*, 1986.

[2] A. Elgammal, R. Duraiswami, D. Harwood, and L. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 2002.

[3] X. Feng and P. Perona. Scene segmentation and 3d motion. *IEEE CVPR*, 1998.

[4] N. Friedman and S. Russell. Image segmentation in video sequences: A probabilistic approach. *Conf. Uncertainty in Artificial Intelligence*, 2000.

[5] I. Haritaogolu, D. Harwood, and L. Davis. W4: Real-time surveillance of people and their activities. *IEEE TPAMI*, 2000.

[6] R. Hartley. Estimation of relative camera positions for uncalibrated cameras. *ECCV*, 1992.

[7] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. *Cambridge University Press*, 2004.

[8] E. Hayman and J. olof Eklundh. Statistical background subtraction for a mobile observer. *IEEE ICCV*, 2003.

[9] M. Irani and P. Anandan. A unified approach to moving object detection in 2d and 3d scenes. *IEEE TPAMI*, 1998.

[10] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *IJCV*, 1992.

[11] R. Jain and H. Nagel. On the analysis of accumulative difference pictures from image sequences of real world scenes. *IEEE TPAMI*, 1979.

[12] O. Javed, K. Shafique, and M. Shah. A hierarchical approach to robust background subtraction using color and gradient information. *IEEE Workshop on Motion and Video Computing*, 2002.

[13] K. Kanatani. Motion segmentation by subspace separation and model selection. *IEEE ICCV*, 2001.

[14] K.-P. Karmann and A. Brandt. Moving object recognition using an adaptive background memory. *Time-Varying Image Processing and Moving Object Recognition*, 1990.

[15] Q. Ke and T. Kanade. A subspace approach to layer extraction. *IEEE CVPR*, 2001.

[16] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russell. Towards robust automatic traffic scene analysis in real-time. *ICPR*, 1994.

[17] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE TPAMI*, 2004.

[18] A. Mittal and D. Huttenlocher. Scene modeling for wide area surveillance and image synthesis. *IEEE CVPR*, 2000.

[19] A. Mittal and N. Paragios. Motion-based background subtraction using adaptive kernel density estimation. *IEEE CVPR*, 2004.

[20] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh. Background modeling and subtraction of dynamic scenes. *IEEE ICCV*, 2003.

[21] S. Palmer. Vision science: Photons to phenomenology. *The MIT Press*, 1999.

[22] Y. Ren, C.-S. Chua, and Y.-K. Ho. Statistical background modeling for non-stationary camera. *Pattern Recogn. Lett.*, 24(1-3), 2003.

[23] S. Rowe and A. Blake. Statistical mosaics for tracking. *IVC*, 14:549–564, 1996.

[24] P. Sand and S. Teller. Particle video: Long-range motion estimation using point trajectories. *IEEE CVPR*, 2006.

[25] H. S. Sawhney, Y. Guo, and R. Kumar. Independent motion detection in 3d scenes. *IEEE TPAMI*, 22, 2000.

[26] Y. Sheikh and M. Shah. Bayesian object detection in dynamic scenes. *IEEE TPAMI*, 2005.

[27] C. Stauffer and E. Grimson. Learning patterns of activity using real-time tracking. *IEEE TPAMI*, 2001.

[28] H. Tao, H. S. Sawhney, and R. Kumar. Object tracking with bayesian estimation of dynamic layer representations. *IEEE TPAMI*, 2002.

[29] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 1992.

[30] P. Torr. Outlier detection and motion segmentation. *Ph.D. Thesis, University of Oxford*, 1995.

[31] B. Turlach. Bandwidth selection in kernel density estimation a review. *Tech. Report, Institut fur Statistik und Okonometrie, Humboldt-Universitat zu Berlin*, 2003.

[32] R. Vidal and Y. Ma. A unified algebraic approach to 2-d and 3-d motion segmentation. *ECCV*, 2004.

[33] J. Wang and E. Adelson. Representing moving images with layers. *IEEE TIP*, 1994.

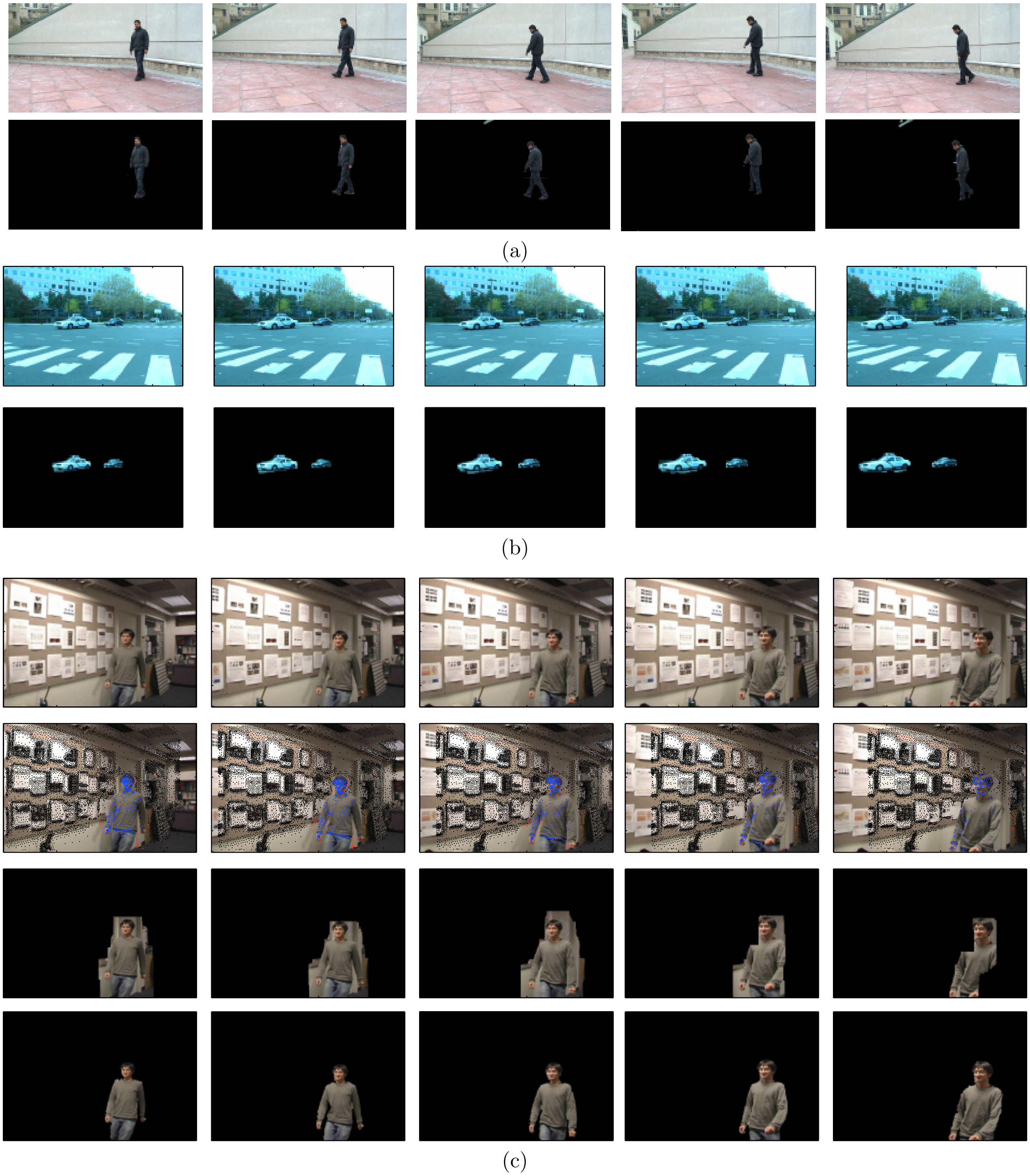
[34] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real time tracking of the human body. *IEEE TPAMI*, 1997.

[35] J. Xiao and M. Shah. Motion layer extraction in the presence of occlusion using graph cuts. *IEEE TPAMI*, 2005.

[36] C. Yuan, G. Medioni, J. Kang, and I. Cohen. Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric constraints. *IEEE TPAMI*, 2007.

[37] J. Yuxin, T. Linmi, D. Huijun, N. Rao, and G. X. Background modeling from a free-moving camera by multi-layer homography algorithm. *IEEE ICIP*, 2008.

[38] J. Zhong and S. Sclaroff. Segmenting foreground objects from a dynamic textured background via a robust kalman filter. *IEEE ICCV*, 2003.



(a)

(b)

(c)

Figure 5. Background subtraction for a freely-moving camera. Results on (a) a pedestrian sequence, (b) the car sequence, and (b) the person sequence. In (a) the average pixel distance moved by the background pixels is 207.21. The image resolution of this sequence is 720 × 1280. In (b) and (c) the image resolution in these sequences is 480 × 720. The average background point motion in the car sequence is 67.10 pixels and the average in the person sequence is 90.93 pixels. In (c) the location of trajectory points in each frame is shown in the second row, with black points denoting background locations, blue points denoting correctly identified foreground pixels, and red points denoting incorrectly identified foreground pixels (false-positives). The estimated labeling images is shown in the third row, and ground truth segmentation is shown in the last row.