Actionable Saliency Detection:
Independent Motion Detection Without Independent Motion Estimation

Georgios Georgiadis Alper Ayvaci Stefano Soatto
University of California, Los Angeles, 90095, USA
{giorgos,ayvaci,soatto}@cs.ucla.edu

Abstract

We present a model and an algorithm to detect salient regions in video taken from a moving camera. In particular, we are interested in capturing small objects that move independently in the scene, such as vehicles and people as seen from aerial or ground vehicles. Many of the scenarios of interest challenge existing schemes based on background subtraction (background motion too complex), multi-body motion estimation (insufficient parallax), and occlusion detection (uniformly textured background regions). We adopt a robust statistical inference approach to simultaneously estimate a maximally reduced regressor, and select regions that violate the null hypothesis (co-visibility under an epipolar domain deformation) as "salient". We show that our algorithm can perform even in the absence of camera calibration information: while the resulting motion estimates would be incorrect, the partition of the domain into salient vs. non-salient is unaffected. We demonstrate our algorithm on video footage from helicopters, airplanes, and ground vehicles.

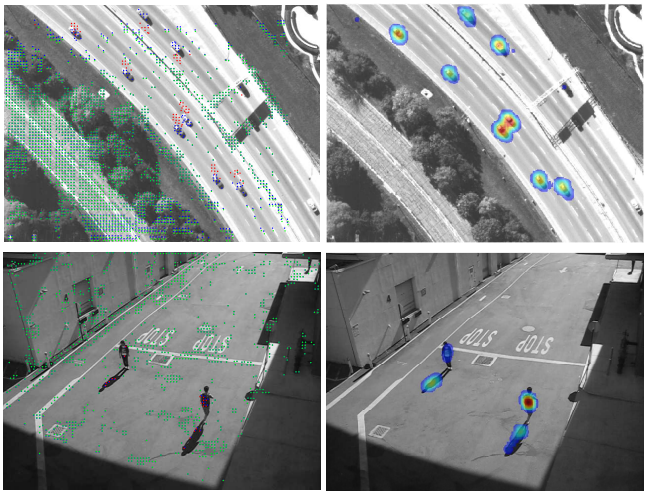

Figure 1: *Detecting salient regions under camera motion: Left: Tracked feature points (blue) are classified as inliers (green) or outliers (red). Right: Estimated salient point density obtained by our algorithm.*

1. Introduction

A subset of a sensing field (e.g. visual) is ordinarily deemed "salient" if it is "sufficiently different" from its surroundings. Saliency is therefore a detection and localization task (illustrated in Fig.1), often motivated by resource constraints: if one can process only a subset of the data, which subset is most "valuable" or "informative"?

Traditionally, saliency detection has been agnostic of the underlying task. More recently, however, several authors have attempted framing saliency detection in an information-theoretic context, by looking at the "most informative" subset of the data, where "information" is measured in the traditional sense of Wiener and Shannon. For instance, Itti and Baldi [11] measure the relative entropy between the prior and the posterior of an image, interpreted as a distribution of pixel values, and use it as a measure of saliency or "surprise".

In this paper, we focus on classes of tasks that involve decisions about the *scene*, rather than about the *image*. These include detection, localization, recognition of objects, events, or spatial locations from images, as well as navigation, manipulation and other spatial control tasks. While often "salient" locations in the image correspond to salient geometric or topological characteristics of the scene (e.g. occluding boundaries), this is not always the case (e.g. material or illumination boundaries). Moreover, whether a salient region of the image does indeed correspond to a geometric or topological characteristic of the scene cannot be positively ascertained from one image alone; therefore, we are interested in saliency detection mechanisms that involve *multiple images*. Of course, because part of the motivation for detecting salient regions is to expedite processing (at the expense of a loss in discriminative power), we are interested in temporally adjacent images (*"small baseline"*), such as two or more temporally consecutive frames of a video.

When the camera is static, as in the case of video surveillance, anything that *moves* is salient. There is a considerable

amount of literature on *background subtraction*, that can be thought of as a form of saliency detection for the specific case of surveillance tasks (see [12] and references therein). However if the camera is moving, then detecting objects that are moving independently is a notoriously difficult problem, for it amounts to detection of independent rigid motions. This involves model selection and regression to find the independently moving objects and their motion. And yet, even when driving, we can easily spot a moving animal in the distance. When flying we can detect another flying vehicle, or vehicles moving on the ground. Several attempts to perform "background subtraction from moving cameras" [15] have improved efficiency compared to multi-rigid motion estimation, that was using algebraic geometric methods [21] or sampling methods that would clearly not be viable for the task of rapid detection of "informative" regions of a video. Moreover, there is no direct link between any of these algorithms and a notion of what "informative" means.

A definition of "information" in the context of visual decision tasks [17], that draws on ideas from Gibson's Ecological Approach to Visual Perception [4], can shed some light on this issue. While the complexity of the image is not necessarily related to its value in a visual decision task, the complexity of the part of the image that would be *discovered* after a finite time interval represents the *"Actionable Information Increment"* provided by the "next image" [17]. It is the decrease in uncertainty about the scene provided by the data. Such a discovery could be due to motion of the viewer, or motion of an object within the scene, or both. In any case, this suggests that *occlusion detection* is a natural form of saliency detection.

Unfortunately, occlusion detection fails to capture important visual phenomena, and indeed even *fails to capture occlusion phenomena* in many cases of practical importance, as we describe next. Therefore, in Sect. 1.2 we propose an alternative scheme for detecting salient regions in videos.

1.1. Occlusion detection fails to detect occlusions

Occlusions are defined as portions of the domain of an image captured, say, at time $t + dt$, that correspond to (are projections of) portions of the *scene* that were occluded from the vantage point where the image at time t was captured. That is, occlusions are something you see in an image but not the other. Unfortunately, such occlusions cannot always be detected in the image: for the examples we mentioned above, if a car is seen from an airplane while traveling on a road that has fairly homogeneous texture, occlusion detection fails. Similarly, a person walking against a white wall can be explained as the person painted on the wall, and deforming with it. This is because occlusion detection from images is based on a hypothesis testing process where the null hypothesis is that portions of two images are *co-visible* when there exists a diffeomorphism ("optical flow") that takes one image onto the other, up to a residual that is statistically simple (white, homoscedastic, and independently distributed) [1].

In formulas, we have that for any given subset Ω of the image domain D, where an image $I : D \to \mathbb{R}$ is measured at each instant of time, the null hypothesis that Ω is *co-visible* between t and $t + dt$ can be written as:

$$H0 = \{\exists \text{ a diffeo } w : \Omega \to D \mid I(x, t + dt) - I(x + w(x), t) \overset{\text{IID}}{\sim} \mathcal{N}\} \quad (1)$$

where the residual $n(x, t) \doteq I(x, t + dt) - I(x + w(x), t)$ is spatially and temporally white, independent and identically distributed according to a simple description, such as a bivariate Normal distribution, $\mathcal{N}$, with diagonal covariance. This means that *co-visible regions are diffeomorphically equivalent up to white noise:* there exists a differentiable and differentially invertible map that takes one image onto the next, except for a white residual. An occlusion is detected as a violation of the null hypothesis, that is when *no* diffeomorphism can be found that can explain the next image using the previous one and the addition of white noise.

Therefore, a car moving on a road (thus generating an occlusion) can be explained as a car painted on the road (generating no occlusion), and the road-car ensemble stretching and compressing to yield images that are indistinguishable from those actually measured. Yet, we can effortlessly detect moving cars from a moving aerial vehicle (Fig. 3,4).

1.2. Key idea and related work

The problem with occlusion detection is that *equivalence up to a diffeomorphism* is too general, and can explain as ordinary (no violation of the null hypothesis) situations that we want to consider salient. We would indeed prefer to detect as salient, any violations of the *rigidity* assumption, but we do not want to perform independent detection of multiple rigid bodies, because that strides with our goal of computational efficiency.

The key idea of this paper is to still pursue saliency detection as violation of co-visibility, but define co-visibility in terms *not* of diffeomorphic equivalence, but rather *epipolar equivalence*. This means that, of all possible diffeomorphisms $w : D \to D$, we only consider those that are compatible with an overall *rigid motion* of the viewer (ego-motion).

In principle, this could be done by computing the "dominant motion", and then detecting outlier regions as salient. However, we do not actually care to even estimate the motion of the viewer; we just want to compute the discriminant for the null hypothesis (1) in the most efficient way, so that it would depend on the smallest possible number of free parameters. As we show in Sect. 2, this number is *two*.

At face value, what we propose looks more complicated than testing for diffeomorphic equivalence $H0$, for we would

have to enforce the additional condition that the diffeomorphism is compatible with a rigid motion. In formulas, after testing $H0$ we would have to test for:

$$H1 = \{\exists\, V \in \mathbb{S}^2, \omega \in \mathbb{R}^3,\ Z : D \to \mathbb{R}^+ \mid w(x) = \pi(\widehat{\omega}\bar{x}Z(x) + V)\} \quad (2)$$

where V is the translational velocity direction, ω is the rotational velocity vector, $Z(x)$ is the depth map, $\bar{x} = [x^T\ 1]^T$ is the homogeneous coordinate of x, and π is a canonical central projection [1]. In words, in order to determine whether a region is salient, we would have to search at each instant for all possible translational directions, rotational velocities and depth maps until *none* of them fits the data up to a white residual. The two hypotheses can be tested simultaneously by substituting the expression of w in (2) into (1). The result would be akin to devising a *robust ego-motion estimation* scheme, whereby one simultaneously tries to find the translational direction V, rotational velocity ω, depth map Z, *and* occluded region Ω. This has been indeed done before in the literature on "dominant motion estimation" [10] and robust motion estimation [16], and relates to robust statistics [8] and outlier rejection in motion estimation [3].

This would already be an improvement on multi-body motion segmentation. If we have a number, say K, of independently moving objects, and N sensors, then multiple motion estimation requires inferring $N + 5K$ unknown parameters [21]. Dominant motion estimation, on the other hand, only requires inferring $N + 5$ parameters in order to build the discriminant for $H0 \cup H1$. Nevertheless, when N is large, this becomes prohibitive. When the calibration of the camera is unknown, in addition to these parameters one would also have to infer 5 additional parameters (optical center $x_0 \in D$, focal length f, aspect ratio s and skew θ).

As we have already anticipated, our goal is not to estimate ego-motion, but to detect salient regions in the image based on violation of rigidity. Therefore, we seek for ways to reduce the discriminant to its minimal form, which we do in Sect. 2. Several algorithms [6, 9, 13, 14, 22] have been proposed to detect salient regions with background subtraction techniques by removing the stationary camera limitation. However, these methods largely rely on estimation of a homography or a 2D affine transform to compensate for the camera motion for the scenes that can be approximated well by a plane and do not work in the case of a complex scene.

Most relevant to our work is Sheikh et al. [15] which also detects independently moving objects by exploiting a geometrical constraint. In their case, they exploit the rank-constraint on trajectories on the background and hence they overcome the restrictions on the scene models shared by others. However, this method requires a large number of frames to operate. This is a limitation of their algorithm since salient regions might appear in very few frames (e.g. Fig. 3,4) and the extracted trajectories might have an even shorter length. On the other hand, our method only requires 2 frames to detect salient regions.

[1] Note that $\hat{\omega} \doteq \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix}$ belongs to the Lie algebra of the skew-symmetric matrices $so(3) \doteq \{S | S^T = -S\}$.

2. Derivation of the discriminant

If we consider the instantaneous motion of the scene relative to the viewer, where the entire scene is moving rigidly, the deformation of the entire domain of the image can be explained as a function of the motion (translational velocity direction V and rotational velocity ω) and the shape of the scene, described by a scalar function from the image domain D to the positive reals, $Z : D \to \mathbb{R}^+$, as described in (2). If we call $y(x) \in \mathbb{R}^2$ the velocity of the projection of the point with coordinates $\bar{x}Z(x) \in \mathbb{R}^3$ onto the image, we have that [18]:

$$y(x) = \mathcal{A}(x)\frac{V}{Z(x)} + \mathcal{B}(x)\omega \quad (3)$$

where:

$$\mathcal{A}(x) \doteq \begin{bmatrix} 1 & 0 & -x_1 \\ 0 & 1 & -x_2 \end{bmatrix} \quad (4)$$

$$\mathcal{B}(x) \doteq \begin{bmatrix} -x_1x_2 & 1+x_1^2 & -x_2 \\ -1-x_2^2 & x_1x_2 & x_1 \end{bmatrix} \quad (5)$$

Traditional dominant motion estimation and robust statistical approaches search for the unknown motion V, ω and range map $Z(\cdot)$ that solve the following optimization problem:

$$\hat{Z}, \hat{V}, \hat{\omega} = \arg\min \int_D \|y(x) - \mathcal{A}(x)\frac{V}{Z(x)} - \mathcal{B}(x)\omega\|_{\mathcal{H}} dx \quad (6)$$

where $\|\cdot\|_{\mathcal{H}}$ denotes a robust norm, for instance a Huber norm [8]. After this is done, one would find the salient regions that violate this model, that is:

$$\Omega \doteq \{x \in D \mid \|y(x) - \mathcal{A}(x)\frac{\hat{V}}{\hat{Z}(x)} - \mathcal{B}(x)\hat{\omega}\| > \epsilon\} \quad (7)$$

where ϵ is related to the regularization parameter in the Huber norm. Note that the region Ω can, and in general will be *multiply-connected*, so even though this is a *binary* classification problem, it enables detecting any number of independently moving objects, each projecting onto a different simply-connected subset of the image domain. Furthermore, when (6) is solved in the continuum, regularization on Z has to be imposed (this is not necessary when (6) is computed at a sparse set of locations). This is laborious, especially because the procedure of finding the motion $\hat{V}, \hat{\omega}$ and the range map $\hat{Z}$ has to be iterated once the outlier set Ω is removed, which in turn changes the motion and range estimates, resulting in a non-convex optimization problem.

Therefore, we resort to a trick introduced by Heeger and Jepson [7], whereby one solves the problem above for the case of the ℓ^2 norm, by exploiting the geometry of Hilbert spaces to "eliminate" the unknown depths $Z(x)$ and unknown rotational velocity ω from (6). This can be done easily since the model (3) is *linear* in $\frac{1}{Z}$ and ω, and therefore one can solve-and-substitute, thereby leaving a set of constraints on the unknown V alone. It has been shown [2] that this can be done without altering the topology of the solution space, in the sense that no spurious solutions are introduced by the algebraic manipulation.

Formally, this can be accomplished (Sect. 2.1) by rewriting the model (3) in terms of an operator $C(V)$ that multiplies all the unknown depths and rotational velocity, then multiplying by the orthogonal projector operator $\hat{C}(V)$ that eliminates the dependency on ω and Z, and leaves constraints on the unknown V only.

2.1. Computation of the optimal discriminant

Following the discussion in Sect. 2, salient points $x \in \Omega$ will be detected as a violation of the hypothesis provided by the model (3). Equivalently, one can seek to infer V in a robust fashion and detect Ω as the outlier set. We focus on a finite number of N sparse measurements, $x_i,\ i = 1, \dots, N$ and introduce a diagonal weight matrix $\mathcal{W} \in \mathbb{R}^{2N \times 2N}$. Ideally, $\mathcal{W}$ should be zero except for points that follow (3). Writing (3) as a system of linear equations for all points and introducing $\mathcal{W}$ as a weight matrix we have:

$$\mathcal{W}Y(X) = \mathcal{W}C(V) \begin{bmatrix} p(x_1) & \dots & p(x_N) & \omega \end{bmatrix}^T \quad (8)$$

where, $p(x_i) \doteq \frac{1}{Z(x_i)}$, $Y \doteq [y(x_1)^T, \dots, y(x_N)^T]^T$, $X \doteq [x_1^T, \dots, x_N^T]^T$, $\mathcal{W} \doteq diag(w_1, w_2, \dots, w_{2N-1}, w_{2N})$,

$$C(V) \doteq \begin{bmatrix} A(x_1)V & & B(x_1) \\ & \ddots & \vdots \\ & A(x_N)V & B(x_N) \end{bmatrix} \quad (9)$$

For any (unknown) V, we can solve for $P \doteq [p(x_1) \dots p(x_N)]^T$ and ω using Least Squares:

$$\begin{aligned} [P, \omega]^T &= (\mathcal{W}C(V))^\dagger \mathcal{W}Y(X) \quad (10) \\ &\doteq (C(V)^T \mathcal{W}^T \mathcal{W} C(V))^{-1} C(V)^T \mathcal{W}^T \mathcal{W} Y(X) \end{aligned}$$

For readability purposes, we will henceforth drop the explicit dependence of C on V and of Y on X. We can then plug the solution of this equation back to the model to get $\mathcal{W}Y = \mathcal{W}C(\mathcal{W}C)^\dagger \mathcal{W}Y$. Rearranging, we get:

$$\hat{C}(V)Y \doteq \left(I - \mathcal{W}C(\mathcal{W}C)^\dagger\right) \mathcal{W}Y = 0 \quad (11)$$

The above constraint is true even in the presence of outliers when the elements of $\mathcal{W}$ corresponding to those equations are 0. In practice, this cannot be achieved though, due to the presence of noise and unmodeled phenomena. Assuming that the error in motion estimation follows a Gaussian distribution the Least Squares estimation of P and ω is optimal. Hence it is important to calculate $\mathcal{W}$ properly so that inference of V is improved. To estimate V we solve the following minimization problem:

$$\underset{V}{\text{minimize}} \quad \psi(V) = \frac{1}{2} ||\hat{C}(V)Y||_2^2 \quad (12)$$

where $V \in \mathbb{S}^2$. To calculate the weight matrix $\mathcal{W}$ we employ a more traditional M-estimator, as customary in robust statistics, that does not explicitly infer $\mathcal{W}$, but instead uses a composite norm residual where the weight of the outliers is reduced. This yields a minimal model, where the only unknowns are the directional coordinates of the translational velocity V, as discussed in the previous section.

Since we expect that most points in the scene will move rigidly, we anticipate that $\hat{C}(V)Y$ is sparse. We would hence want to choose the diagonal elements of $\mathcal{W}$ to enhance sparsity of the residual. In addition, every pair of elements of $\hat{C}(V)Y$, corresponds to the residual for a single point and hence this should also be taken into account when estimating $\mathcal{W}$. The outline of the algorithm is given below.

Algorithm 1: Iterative reweighted subspace minimization (IRWSM).

Initialize $\mathcal{W}^{(1)} = I, V^{(0)} = [1, 0, 0]^T$
foreach $k = 1, 2, 3, \dots,$ *K* **do**
  Solve the following problem initializing with $V^{(k-1)}$:
    $\hat{V}^{(k)} = \arg\min_V \frac{1}{2} ||\hat{C}(V, \mathcal{W}^{(k)})Y||_2^2$
  $e^{(k)} = \hat{C}(\hat{V}^{(k)}, I)Y$
  $\lambda = 1/\text{mean}(||e^{(k)}||)$
  foreach $i = 1, 2, 3, \dots,$ *N* **do**
    $w_{2i-1}^{(k+1)} = w_{2i}^{(k+1)} = \frac{1}{||[e_{2i-1}^{(k)}, e_{2i}^{(k)}]||_2 + \varepsilon}$
  $\mathcal{W}^{(k+1)} = diag(w_1^{(k+1)}, \dots, w_{2N}^{(k+1)})$
  $V^{(k+1)} := \hat{V}^{(k)} / \|\hat{V}^{(k)}\|$

where $\mathbf{1}$ is the indicator function. Note that this is a generalization of the case proposed by [7]. The authors of [7] minimized (12) with $\mathcal{W} = I$ using exhaustive search. In that case the above problem is reduced to minimizing $C^\perp Y \doteq \left[I - C(C^TC)^{-1}C^T\right] Y$. By introducing $\mathcal{W}$, we solve this more general minimization problem to improve outlier rejection. Since the problem is non-convex, we use gradient descent with backtracking line search to estimate V. The details of the computation of the gradient of (12) are provided in the supplementary material[2]. We classify a point as an outlier as follows: define $\hat{e} = [\hat{e}_1 \dots \hat{e}_{2N}]^T \doteq \hat{C}(V^{(K)}, I)Y$

[2] http://vision.ucla.edu/~giorgos/cvpr2012/

and $E_i \doteq [\hat{e}_{2i-1}, \hat{e}_{2i}]^T$ for $i = 1, \ldots, N$. A point i is classified as an outlier when $||E_i||_2$ exceeds ϵ. The threshold ϵ can be determined using various techniques, one of which is explained in Sect. 3.

2.2. Effects of (mis)calibration

The model we have derived assumes that the image coordinates x_i and their corresponding velocities y_i are *calibrated*, that is they are available in metric units relative to the reference frame having origin at the principal point (intersection of the optical axis with the image plane), with the optical axis orthogonal to the image plane and aligned with the spatial Z axis. Most often, however, coordinates and velocities are given in *pixel* units, relative to, say, the top-left corner of the image. One cannot expect, in general, to just be able to plug the latter into the equation and get a sensible answer. Therefore, in this section we explore the effects of miscalibration on outlier detection.

We first show that knowledge of the *principal point* and the *focal length* does not affect the classification of outliers. We introduce the calibration matrix $K \in \mathbb{R}^{3\times3}$ in (3) and rewrite it in homogeneous coordinates:

$$\begin{aligned}\begin{bmatrix} y \\ 0 \end{bmatrix} &= K \begin{bmatrix} \mathcal{A}V & \mathcal{B} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/Z \\ \omega \end{bmatrix} \\ &= \begin{bmatrix} fs_x & fs_\theta & O_x \\ 0 & fs_y & O_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathcal{A}\frac{V}{Z} + \mathcal{B}\omega \\ 0 \end{bmatrix} \end{aligned} \quad (13)$$

From (13) it is obvious that $y(x)$ is independent of (O_x, O_y). On the other hand, also obvious from (13), the focal length and scale do indeed affect the estimation of the velocity. But the focal length does not affect the outlier distribution: writing the expressions of y, from (13), similarly as in Sect. 2.1, we get:

$$\begin{aligned}\mathcal{W}Y(X) &= f\mathcal{W}\mathcal{K}C(V)[P,\omega]^T \\ &\doteq f\mathcal{D}[P,\omega]^T \end{aligned} \quad (14)$$

where $\mathcal{K} \in \mathbb{R}^{2N\times2N}$ is a block diagonal matrix with its block diagonal entries being $\hat{K} \doteq \begin{bmatrix} s_x & s_\theta \\ 0 & s_y \end{bmatrix} \in \mathbb{R}^{2\times2}$. Solving for the same unknowns as before:

$$\begin{aligned}[P,\omega]^T &= (f\mathcal{D})^\dagger \mathcal{W}Y(X) = (f^2\mathcal{D}^T\mathcal{D})^{-1} f\mathcal{D}^T\mathcal{W}Y(X) \\ \mathcal{W}Y(X) &= f\mathcal{D}(f^2\mathcal{D}^T\mathcal{D})^{-1} f\mathcal{D}^T\mathcal{W}Y(X) \\ &= \mathcal{D}(\mathcal{D}^T\mathcal{D})^{-1}\mathcal{D}^T\mathcal{W}Y(X) \end{aligned} \quad (15)$$

Focal length is cancelled out in the expression and hence it is not necessary in order to employ our algorithm. In addition, since scale consists of two positive real numbers (or one number, if the pixels are square, or if the form factor of the pixel is known), one can simply augment the search from two parameters, corresponding to V, to four parameters, corresponding to s_x, s_y . In the following experiments we normalize the pixel coordinates to $[-1, 1]$. Regarding the *skew* of the pixel array, it can be assumed to be zero; that is, the pixels are rectangular, and not generic parallelograms.

3. Empirical evaluation

We tested our algorithm on 15 sequences. The sequences People-1, People-2, Cars-3, Cars-4, Cars-5, Cars-6 shown in this order in Fig. 2 and Cars-2/06 are from the Hopkins 155 motion segmentation dataset [20] and *ground truth* was provided. In addition, the trajectories of feature points are *provided* by the dataset and are available over the whole duration of the video sequence. This makes the dataset appropriate for comparison with Sheikh et al. [15] which requires the trajectories to be present in an extended period of time. These sequences contain objects that move slowly between consecutive frames, they are close to the camera and are moving independently from it.

The sequences Traffic-1,-2,-3,-4 (Fig. 3) were recorded from a helicopter monitoring a traffic jam. The motion of the camera covers a wide variety of translations and rotations. Bridge-1,-2,-3 (Fig. 4) were taken from an airliner approaching Boston Logan airport. People-3 (second row in Fig. 1) is an aerial view of closed distanced objects. These 8 sequences were *manually* annotated. In addition, to extract trajectories in *these* sequences we used the code provided by [19] that yields dense point trajectories. We used the Harris corner detector [5] to eliminate trajectories on textureless regions. Subsequently, an average of 1300 trajectories per frame are left. Since the extracted trajectories are not guaranteed to be present in all frames hence these sequences are not suitable for comparison with [15]. On the other hand, our algorithm is *not limited by the temporal support* of trajectories. Using the resulting trajectories for a pair of frames in a sequence (we use the middle pair), we calculate the optical flow i.e. $y(x_i)$ for $i = 1, \ldots, N$ which is then used as the input to Algorithm 1 to estimate V and determine the salient regions.

To distinguish between inliers and outliers, we calculate $||E_i||_2$ for each point x_i as its residual. We then construct the histogram of the residuals and find the local minimum nearest to the 0 residual bin. The residual value corresponding to this bin is selected as the threshold ϵ.

We successfully detect most of the salient regions in all sequences. In Fig. 1, 2, 3 and 4 we show the tracked regions and the salient regions as classified by our method. In Table 1, we compare the performance of our algorithm to three other methods using the F-measure: (i) RANSAC [3] with epipolar constraint. We fixed the number of iterations to 1000 and varied the threshold for each sequence to obtain the best results, (ii) we implemented the original method proposed by [7] but minimized it with gradient descent rather than exhaustive search i.e. we used $K = 1$ and $\mathcal{W} = I$ as parameters in our algorithm, and (iii) we implemented the

Figure 2: *Sample results from the Hopkins 155 dataset: Odd rows: Images with tracked points. Red and green points show the locations of tracked points as predicted by the model. Points in green are the points that are classified as inliers and in red those that are classified as outliers. Blue dots (not visible for inlier points) are the true positions of tracked points. Even rows: Images showing in color the detected outliers. The color corresponds to a sum of Gaussians centered at each salient point.*

outlier detection method proposed by Sheikh et al. [15] that enforces the rank constraint on trajectories using RANSAC. We also fixed the number of iterations to 1000 and varied the threshold to obtain the best results. This method requires trajectories available in an extended period of time which means it is only possible to compare with it on the Hopkins 155 dataset. For their method, in all experiments, we used either trajectories of length 30 or of the whole video, whichever was the smallest (27 frames on average).

Our algorithm significantly outperforms the other methods in 13 out of 15 sequences and achieves comparable results in the other two, Table 1. Although [15] uses a large

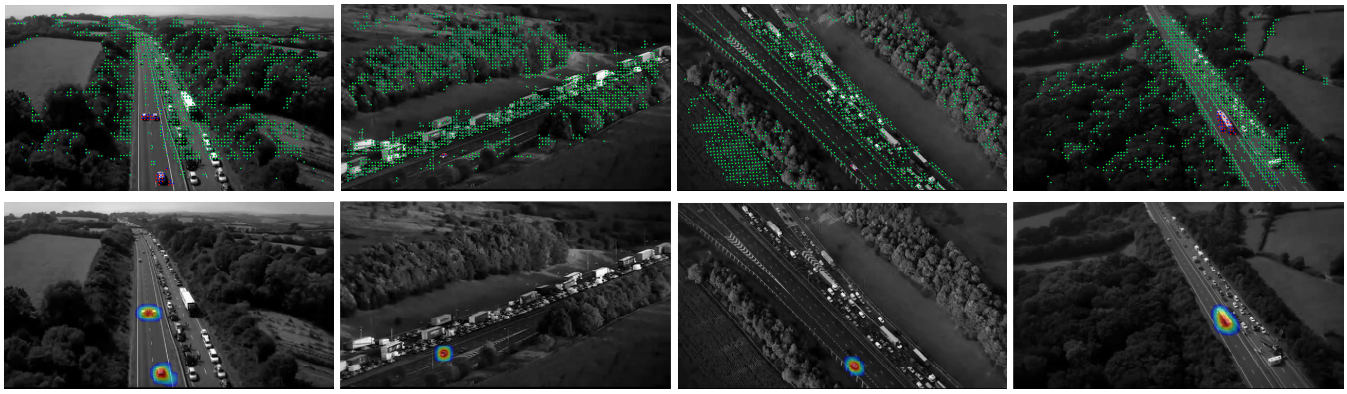

Figure 3: *Four aerial views of a motorway. In all images cars in the right lane are stationary and cars in the left lane are moving. The true outliers in these cases are the moving cars in the left lane. The first row shows the dense tracked points in each image. The second row shows in color the detected outliers. Color convention is the same as in Fig. 2.*

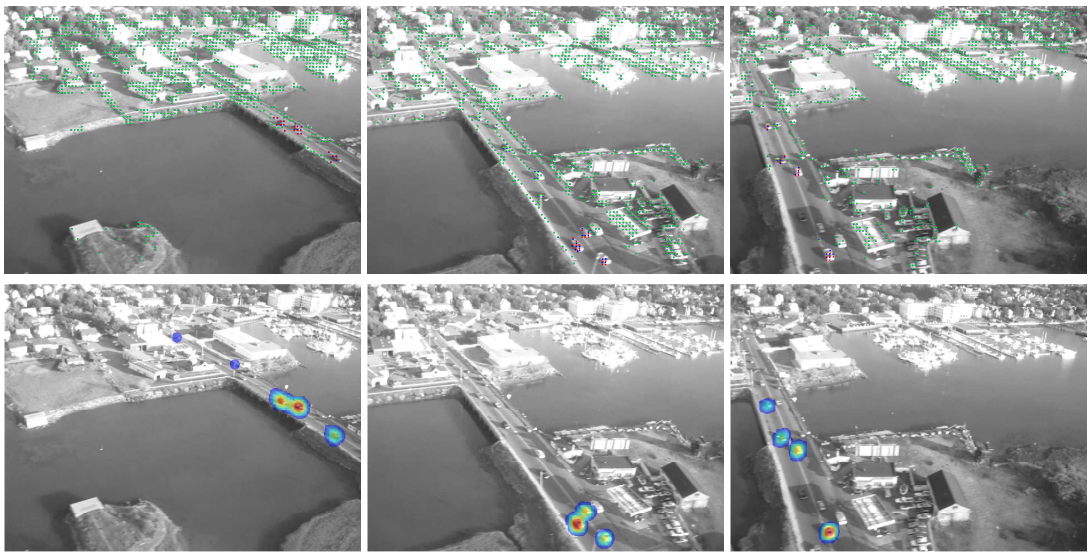

Figure 4: *Three aerial views of a bridge taken from an airliner during a turn approaching Boston's airport. The first row shows the images with the tracked points and the second highlights the salient regions. The color codes are the same as in Fig. 2.*

number of frames to detect outliers, our algorithm still performs better even though we make use of just 2 frames to make a decision. In addition, our method automatically chooses the threshold value whereas for RANSAC and [15] we manually chose the best one. Even so, we still perform significantly better than both of these methods. For the Hopkins 155 dataset, it takes on average 32.6 seconds for a non-optimized MATLAB implementation of our algorithm to converge to a solution, whereas for the rest of the sequences, it takes 68 seconds with an average of 1300 tracked points. We terminate our minimization at each iteration when $||V^{(k)} - V^{(k-1)}||/||V^{(k-1)}|| < 10^{-3}$. The average runtime of RANSAC was 2.3 seconds and that of [15] was 2.6 seconds. Experiments were ran on an Intel 2.4 GHz dual core processor machine.

Failure modes. The most significant failure case of our method is shown in Fig. 4. Moving cars at the far end of the bridge are not detected. This can be accounted to the fact that the outliers that are not detected are far from the camera and they appear stationary due to their relatively small motion.

4. Discussion

We have presented a model for detecting "salient" regions in an image that correspond to objects that are moving in a way that is incompatible with a single rigid motion. Note that, even if the motion is rigid, the deformation it induces on the domain of the image is, in general, as complex as a general diffeomorphism, depending on the shape of the scene, and even more complex if one considers occlusions. Therefore, simple "background subtraction" relative to a small-dimensional parametric motion model (such as an homography) does not work in general. Even occlusion detection, that in principle can be used for testing the co-visibility hypothesis, fails in the presence of objects moving

	People-1	People-2	Cars-2/06	Cars-3	Cars-4	Cars-5	Cars-6	People-3
Ours	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**	**1.00**	**1.00**	**0.91**
Heeger & Jepson [7]	0.35	0.06	0.43	0.24	0.22	0.69	0.11	0.88
RANSAC	0.64	0.77	0.54	0.69	0.21	0.65	0.56	0.69
Sheikh et al. [15]	0.91	0.68	0.95	0.90	0.94	0.93	0.80	-
	Traffic-1	Traffic-2	Traffic-3	Traffic-4	Bridge-1	Bridge-2	Bridge-3	
Ours	**0.78**	0.80	**1.00**	**0.93**	**0.55**	0.52	**0.63**	
Heeger & Jepson [7]	**0.78**	0.80	**1.00**	**0.93**	0.54	0.51	**0.63**	
RANSAC	0.11	**1.00**	0.11	0.35	0.49	**0.60**	0.50	
Sheikh et al. [15]	-	-	-	-	-	-	-	

Table 1: *Comparison on salient point detection performance of our algorithm against [7], RANSAC under epipolar constraint and [15] in terms of the F-measure. We compared the performance on* 15 *sequences. The ground truth and trajectories for the first* 7 *sequences were provided by the Hopkins 155 dataset. The last* 8 *were manually annotated by the authors and trajectories were extracted using [19]. Our algorithm significantly outperforms all other* 3 *methods in almost all sequences.*

on a homogeneous background.

Therefore, we have proposed a scheme to test for violations of co-visibility, relative to an epipolar domain deformation (as opposed to a general diffeomorphic domain deformation) using tools of robust statistics, and a simple expedient to eliminate motion and structure parameters that do not affect the outlier distribution.

We have also shown that accurate calibration of the camera is not necessary: while calibration error clearly affects the motion estimates, we have shown that some calibration parameters (principal point, focal length) do not affect the decision boundary between inlier and outlier, so they can be ignored for the purpose of saliency detection. Scale can either be coarsely calibrated, or estimated as a hidden variable in the regression/classification task.

Failure modes of our algorithm, illustrated in the experiments, include cases where the objects are too small or moving too slowly. As with any classification scheme, there is a dependency on a scalar parameter (detection threshold) that we have chosen using standard guidelines from robust statistics. Our algorithm is currently not operating in real time. However, the problem has significant structure that could be exploited to devise efficient implementations in hardware platforms in the near future.

Acknowledgments. We would like to thank Michalis Raptis and Avinash Ravichandran for valuable discussions. This research was supported by DARPA N66001-11-C-4001, ARO W911NF-11-1-0391 and NSF 0969032.

References

[1] A. Ayvaci, M. Raptis, and S. Soatto. Sparse occlusion detection with optical flow. *IJCV*, 2012. 2

[2] A. Chiuso, R. Brockett, and S. Soatto. Optimal structure from motion:local ambiguities and global estimates. *IJCV*, 2000. 4

[3] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *CACM*, 1981. 3, 5

[4] J. J. Gibson. *The ecological approach to visual perception*. LEA, 1984. 2

[5] C. Harris and M. Stephens. A combined corner and edge detector. *Alvey Vision*, 1988. 5

[6] E. Hayman and J. Eklundh. Statistical background subtraction for a mobile observer. *ICCV*, 2003. 3

[7] D. Heeger and A. Jepson. Subspace methods for recovering rigid motion i. *IJCV*, 1992. 4, 5, 8

[8] P. Huber. *Robust statistics*. Wiley, New York, 1981. 3

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

[10] M. Irani, B. Rousso, and S. Peleg. Detecting and tracking of multiple moving objects using temporal integration. *ECCV*, 1992. 3

[11] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision research*, 2009. 1

[12] Y. Ivanov, C. Stauffer, A. Bobick, and W. Grimson. Video surveillance of interactions. *CVPR*, 1999. 2

[13] Y. Ren, C. Chua, and Y. Ho. Statistical background modeling for non-stationary camera. *PRL*, 2003. 3

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

[15] Y. Sheikh, O. Javed, and T. Kanade. Background subtraction for freely moving cameras. *ICCV*, 2009. 2, 3, 5, 6, 7, 8

[16] D. Skocaj and A. Leonardis. Weighted and robust incremental method for subspace learning. *ICCV*, 2003. 3

[17] S. Soatto. Actionable information in vision. *ICCV*, 2009. 2

[18] S. Soatto, R. Frezza, and P. Perona. Motion estimation via dynamic vision. *ECCV*, 1994. 3

[19] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. *ECCV*, 2010. 5, 8

[20] R. Tron and R. Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. *CVPR*, 2007. 5

[21] R. Vidal. Generalized principal component analysis (gpca). *CVPR*, 2003. 2, 3

[22] 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. *TPAMI*, 2007. 3