Detecting Abandoned Objects with a Moving Camera

Hui Kong[1,2], *Member, IEEE,* Jean-Yves Audibert[1,2],and Jean Ponce[1], *Fellow, IEEE*
[1]Willow Team, Ecole Normale Superieure / INRIA / CNRS, Paris, France
[2]Imagine team, Ecole des Ponts ParisTech, Paris, France
Email: tom.hui.kong@gmail.com, audibert@imagine.enpc.fr, ponce@di.ens.fr
http://www.di.ens.fr/willow/ & http://imagine.enpc.fr/

***Abstract*—This paper presents a novel framework for detecting non-flat abandoned objects by matching a reference and a target video sequences. The reference video is taken by a moving camera when there is no suspicious object in the scene. The target video is taken by a camera following the same route and may contain extra objects. The objective is to find these objects. GPS information is used to roughly align the two videos and find the corresponding frame pairs. Based on the GPS alignment, four simple but effective ideas are proposed to achieve the objective: an inter-sequence geometric alignment based on homographies, which is computed by a modified RANSAC, to find all possible suspicious areas, an intra-sequence geometric alignment to remove false alarms caused by high objects, a local appearance comparison between two aligned intra-sequence frames to remove false alarms in flat areas, and a temporal filtering step to confirm the existence of suspicious objects. Experiments on fifteen pairs of videos show the promise of the proposed method.**

***Index Terms*—Abandoned object detection, video matching, geometric and photometric alignment.**

I. Introduction

In recent years, visual surveillance by intelligent cameras has attracted increasing interest from homeland security, law enforcement, and military agencies. The detection of suspicious (dangerous) items is one of the most important applications. These items can be grouped into two main classes, dynamic suspicious behaviors (e.g., a person attempting to attack others) and static dangerous objects (e.g., luggage or bomb abandoned in public places). The scope of this paper falls into the latter category. Specifically, we investigate how to detect non-flat static objects in a scene using a moving camera. Since these objects may have arbitrary shape, color or texture, state-of-the-art category-specific (e.g., face/car/human) object detection technology, which usually learns one or more specific classifiers based on a large set of similar training images, cannot be applied to our scenario. To deal with this detection problem, we propose a simple but effective framework based on matching a reference and a target video sequences. The reference video is taken by a moving camera when there is no suspicious object in the scene, and the target video is taken by a second camera following a similar trajectory, and observing the same scene where suspicious objects may have been abandoned in the mean time. The objective is to find these suspicious objects. We will fulfil it by matching and comparing the target and reference sequences.

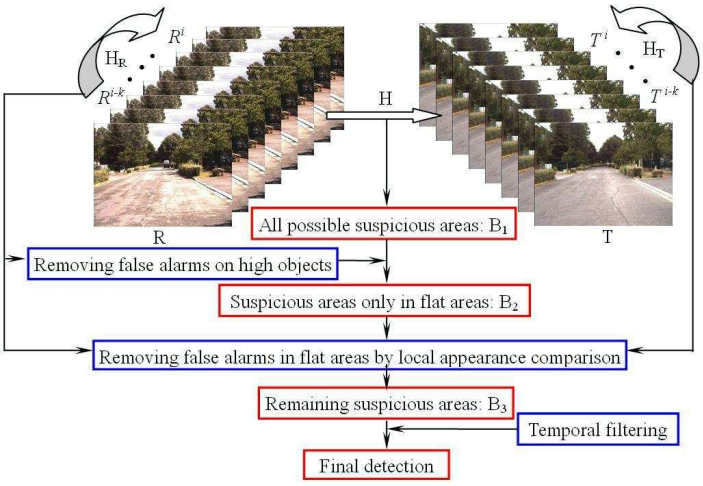

Fig. 1. Flowchart of the proposed framework. H: inter-sequence alignment. H_R: intra-sequence alignment between frames $\mathbf{R}^i$ and $\mathbf{R}^{i-k}$ of $\mathbf{R}$. H_T: intra-sequence alignment between frames $\mathbf{T}^i$ and $\mathbf{T}^{i-k}$ of $\mathbf{T}$. B_1, B_2 and B_3 are the remaining suspicious areas in each step.

To make things efficient, GPS is initially utilized to roughly align the two sequences by finding the corresponding inter-sequence frame pairs. The symbols $\mathbf{R}$ and $\mathbf{T}$ are used throughout this paper to denote the GPS-aligned reference and target video respectively. Based on the GPS alignment, the following four ideas are proposed to achieve our objective (Fig.1): (i) an inter-sequence geometric alignment based on homographies to find all possible suspicious areas, (ii) an intra-sequence alignment (between consecutive frames of $\mathbf{R}$) to remove false alarms on high objects, (iii) a local appearance comparison between two aligned intra-sequence frames to remove false alarms in flat areas (more precisely, in the dominant plane of the scene), and (iv) a temporal filtering step using homography alignment to confirm the existence of suspicious objects. Our experiments demonstrate the effectiveness of the proposed approach even in the presence of large illumination changes between $\mathbf{R}$ and $\mathbf{T}$. **Note:** All figures in this paper are best viewed in color.

II. Related Work

Almost all current methods for static suspicious object detection are aimed at finding abandoned objects using a static camera in a public place, e.g., commercial center,

Fig. 2. Example of GPS-aligned frame pairs. Top row: frames from reference video. Bottom row: frames from target video (an abandoned object appears on the right before the bushes).

metro station or airport hall. Spengler and Schiele propose a tracking/surveillance system to automatically detect abandoned objects and draw the operator's attention to such events [9]. It consists of two major parts: A Bayesian multi-person tracker that explains as much of the scene as possible, and a blob-based object detection system that identifies abandoned objects using the unexplained image parts. If a potentially abandoned object is detected, the operator is notified, and the system provides the operator the appropriate key frames for interpreting the incident. Porikli et al. propose to use two foreground and two background models [7] for abandoned object detection. First, the long- and short-term backgrounds are constructed separately. Thereafter, two foreground models are obtained based on the two background models. The abandoned object can be detected by four hypotheses based on the two foreground and two background models. Guler and Farrow propose to use a background-subtraction based tracker and mark the projection of the center of mass of each object on the ground [3]. The tracker identifies object segmentations and qualifies them for possibility of a "drop-off" event. The stationary object detector running in parallel with the tracker quickly identifies potential stationary foreground objects by watching for pixel regions that consistently deviate from the background for a set duration of time. The stationary objects detected are correlated with drop-off events and the distance of the owner from the object is used to determine warnings and alerts for each camera view. The abandoned objects are correlated within multiple camera views using location information, and a time-weighted voting scheme between the camera views is used to issue the final alarms and eliminate the effects of view dependencies. Smith et al. propose to use a two-tiered approach [8]. The first step is to track objects in the scene using a trans-dimensional MCMC tracking model suited for generic blob tracking tasks. The tracker uses a single camera view, and it does not differentiate between people and luggage. The problem of determining whether a luggage item is left unattended is solved by analyzing the output of the tracking system in a detection process.

All of the above techniques utilize the static cameras installed in some public places, where the background is stationary. However, for some application scenarios, the space to keep a watch on is too large to use static cameras. Therefore, it is necessary to use a moving camera to scan these places. We use in this paper a camera mounted on a moving platform to scan along a specified trajectory for non-flat abandoned objects (e.g., suitcase or bomb abandoned on the road side). This corresponds to a scenario where a scout vehicle "clears" a secure path for a convoy that will follow the same road, perhaps a few hours later. The videos collected by the scout and the following convoy are matched to detect roadside bombs for example. To the best of our knowledge, this is the first attempt at addressing such an issue. The main difficulty is to cope with moving objects (which should be considered as non suspicious in our setting), the presence of shadows, light-saturated areas, appearance changes due to rain and extreme lighting conditions such as those depicted in Fig.2.

III. Proposed Approach

Given a reference video and a target video which are taken by a camera following similar trajectories, GPS is used to roughly align the two videos to reduce computational complexity (the GPS information is obtained every second, which corresponds to roughly every 10 meters). Figure 2 shows three corresponding frame pairs aligned by GPS, where the top and bottom rows are from $\mathbf{R}$ and $\mathbf{T}$ respectively (note the suspicious object in the target sequence).

Generally, it is hard to find the suspicious object if we only compare the GPS-aligned frames, which potentially have large viewpoint variation. This is because GPS alignment can only guarantee that the corresponding inter-sequence frame pair is taken approximately at the same geographical location, but cannot guarantee that the camera has the same view angle for $\mathbf{R}$ and $\mathbf{T}$. In addition, due to speed variation between $\mathbf{R}$ and $\mathbf{T}$, and the different position of the vehicle, alignment using only GPS information may lead to frame pairs separated by as much as 2 meters in 3D real world. Therefore, a fine geometric alignment is necessary. A feature-based alignment method is a better choice than an appearance-based one when the illumination conditions for $\mathbf{R}$ and $\mathbf{T}$ are different [10]. We propose to use 2D homographies [2], [4] for fine alignment. The reasons are that homographies can align two images by registering their dominant planes, and that any non-flat objects (including suspicious and non-suspicious ones) on the dominant plane are deformed while flat objects remain almost unchanged after alignment. We will show that the deformation caused by homography alignment plays a key role in our detection framework (especially when there is an large illumination variation between $\mathbf{R}$ and $\mathbf{T}$). Therefore, we have two assumptions: the suspicious object is a non-flat 3D object[1] (specifically, we are more interested in detecting the abandoned objects which has such a height as a suitcase or gift-boxes etc.), and when it is present in the target sequence, it lies on the ground instead of hanging in the sky, being buried underground or covered by other objects. Note that we do not make the assumption that the route (road) must be flat. In fact, the road can consist of a few flat segments.

[1]Since we do not explicitly infer the 3D scene structure by multi-view geometry method, "non-flat" is defined based on our empirical setup where only the objects with a height of over 4cm and less than 80cm is viewed as "non-flat".

Fig. 3. Examples of alignment based on RANSAC and modified RANSAC (mRANSAC). First column: the reference and target frames. Second column: best inliers obtained from RANSAC and mRANSAC respectively. Third column: the aligned reference frames based on RANSAC and mRANSAC. Fourth column: difference images between the aligned reference frames (by RANSAC and mRANSAC respectively) and the target frames.

With these assumptions, we use inter- and intra-sequence homography alignment as the basis for object detection, where the homographies are computed based on a modified RANSAC (mRANSAC). Figure 1 illustrates the flowchart of the proposed framework. By using the inter-sequence alignment, all possible suspicious areas are highlighted as candidates by setting a suitable threshold on the normalized cross-correlation (NCC) image of the aligned inter-sequence frame pair. By using the intra-sequence alignment on $\mathbf{R}$, we can remove false alarms caused by high objects and other moving vehicles. By using the intra-sequence alignment on both $\mathbf{R}$ and $\mathbf{T}$, most of the false alarms corresponding to the non-suspicious flat areas (e.g., caused by shadow and highly saturated areas) can be removed. Finally, a temporal filtering step is used to remove the remaining false alarms.

Notation: We refer to the sub-figure at the j-th column and the i-th row of a figure as R_iC_j. The i-th frame of $\mathbf{R}$ is denoted by $\mathbf{R}^i$ and the i-th frame of $\mathbf{T}$ is denoted by $\mathbf{T}^i$. We represent the warped $\mathbf{R}^i$ by $\tilde{\mathbf{R}}^i$ and warped $\mathbf{T}^i$ by $\tilde{\mathbf{T}}^i$.

A. Inter-Sequence Geometric Alignment

The SIFT feature descriptor [6] is initially applied to the GPS-aligned frame pairs (we also tried the Harris corner detector, but the result is worse). To reduce the effect of SIFT features of high objects (e.g., trees) on the homography estimation, it is better to apply it only to the image area which corresponds to the ground plane. Therefore, the method proposed in [5] is used to estimate the horizon line passing through the vanishing point of the road. The horizon for straight road can be located at an accuracy of over 96%. For curved road, the vanishing point is detected as the one associated with the main straight part of the road. The performance of vanishing point detection is reported in [5] and the supplemental results for general road images can be found in the section of "General Road Detection from a Single Image" of our project page, http://sites.google.com/site/huikongsite/Home/research. In addition, we emphasize that the homography estimation is insensitive to the accuracy of horizon detection: a detection error of 15 pixels higher or lower than the actual horizon has very little effect on the homography estimation. Only the SIFT features below the vanishing point are viewed as valid. Coarse correspondences between the valid SIFT features of $\mathbf{R}^i$ and $\mathbf{T}^i$ are constructed (step 2 in Table I). Specifically, we first compute a 128-dimensional SIFT descriptor for each keypoint of the reference and target frames (the extraction process just follows Lowe's method). For each descriptor in the reference frame, we search its nearest neighbor in the target frame. Similarly, for each descriptor in the target frame, we search its nearest neighbor in the reference frame. If the two nearest neighbors are consistent, we view them as a match.

Next, the mRANSAC (see Table II) is applied to find the optimal inliers that can be used to compute the homography matrix H_{inter}. Given the locations of the N_p pairs of putatively matched SIFT features - $\mathbb{D} : \{\mathbf{X}_1 \leftrightarrow \mathbf{X}_2\}$, a homogeneous scale factor of 1 is padded as the last coordinate of each data. Then a geometric normalization process (translation + scaling) is performed to transform $\mathbf{X}_1$ to $\breve{\mathbf{X}}_1$ by T_1 and $\mathbf{X}_2$ to $\breve{\mathbf{X}}_2$ by T_2, respectively, so that the means of all the points of $\breve{\mathbf{X}}_1$ and $\breve{\mathbf{X}}_2$ are at origin, and their average lengths equal $\sqrt{2}$, where $T_1 = T^1_{scale} \times T^1_{translate}$ and $T_2 = T^2_{scale} \times T^2_{translate}$. To make things efficient, two loops are adopted where the external loop controls the maximum number of iterations, and the internal loop controls the maximum number of trials allowed to select a non-degenerate[2] point set. In addition, the external loop is also constrained by N, the number of iterations within which mRANSAC never selects a set of n points all of whom are inliers.

In the internal loop, the degeneracy of n randomly sampled

[2]By "degenerate", it means that three or more than three of the randomly selected points are colinear, where the number of randomly selected points, n, is set to "4" in this paper for efficiency.

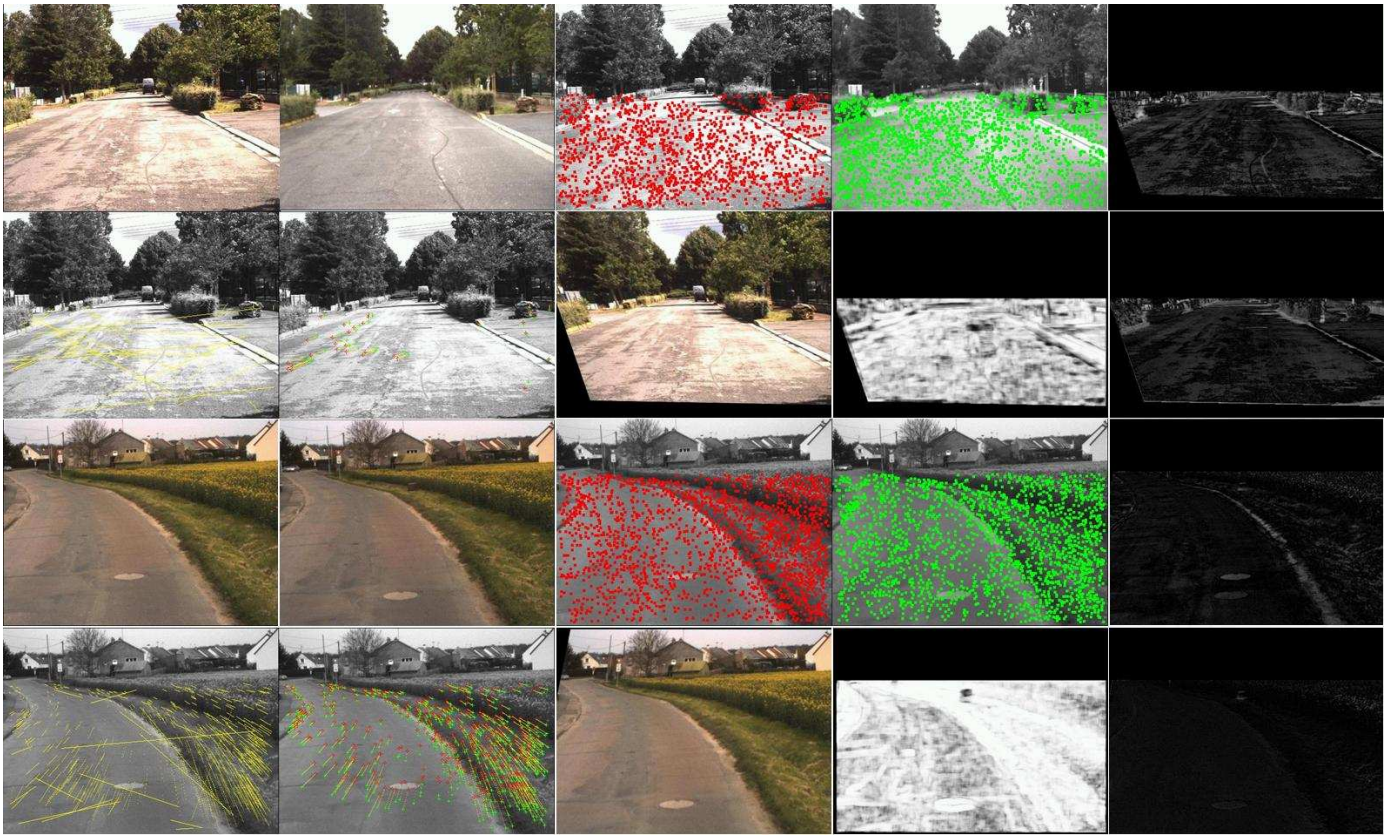

Fig. 4. Examples of inter-sequence alignment. R_1C_1 and R_1C_2 corresponds to the 170^{th} frame of $\mathbf{R}$ and $\mathbf{T}$ respectively. SIFT feature points (below horizon) are shown in R_1C_3 and R_1C_4. R_1C_5 is the grayscale difference image before alignment. The putative correspondences of the SIFT-points are shown in R_2C_1. The inliers from the optimally estimated homography are shown in R_2C_2. R_2C_3 is the warped R_1C_1. R_2C_4 and R_2C_5 are respectively the NCC and difference image between the grayscale R_1C_2 and R_2C_3. Similarly, the bottom two rows show another example.

SIFT feature points is examined, where it involves testing whether any 3 of the n points is colinear, before the n points can be used to fit a homography model H_{temp} (see Table III). If H_{temp} is not empty, $\breve{\mathbf{X}}_1$ is transformed into HX_1 by $HX_1 = H_{temp} \times \breve{\mathbf{X}}_1$ and $\breve{\mathbf{X}}_2$ is transformed into $invHX_2$ by $invHX_2 = H_{temp}^{-1} \times \breve{\mathbf{X}}_2$. The symmetric transfer error of H_{temp} with respect to $\breve{\mathbb{D}}$ (normalized $\mathbb{D}$, see Table II for more details) can be evaluated by $d2 = sum((\breve{\mathbf{X}}_1 - invH\breve{X}_2).^2) + sum((\breve{\mathbf{X}}_2 - H\breve{X}_1).^2)$, where $H\breve{X}_1 = HX_1./HX_1(3,:)$ is the normalized HX_1 and $invH\breve{X}_2 = invHX_2./invHX_2(3,:)$ is the normalized $invHX_2$ (normalization is done by dividing the last coordinate value point-wisely so that the last coordinate of each data is 1).

The number of inliers is calculated as the number of putative matches among $\breve{\mathbb{D}}$ whose symmetric transfer error $d2$ is smaller than the threshold t. Conventionally, homography is usually computed based on the largest lumber of inliers [1]. However, this criterion cannot guarantee good alignment because these inliers may congeal in a small image area. For example, the second image of top row of Fig.3 shows this situation where the yellow arrows link the corresponding SIFT feature points of the reference and target frames (inliers). To overcome this problem, we give the new optimization criterion in mRANSAC which is based on the product of the number of inliers and the scatter of the positions of these inliers. This scatter is measured qualitatively by the largest eigenvalue of a covariance matrix, which is built through the coordinates of the inliers. In this way, we find the best H_{temp} as H_{inter}.

The pseudo codes are listed in Table I to Table III. The second image of bottom row of Fig.3 shows the optimal inliers obtained by mRANSAC. By visually comparing the two images in the last column of Fig.3, mRANSAC gives better alignment than RANSAC. Based on H_{inter}, the reference frame $\mathbf{R}^i$ is warped into $\tilde{\mathbf{R}}^i$ to fit the target frame $\mathbf{T}^i$. We get the NCC image between $\tilde{\mathbf{R}}^i$ and $\mathbf{T}^i$. We binarize the NCC image into B_1 with a threshold t_1 (i.e., those pixels whose values are lower than t_1 are highlighted as possible suspicious areas). We will describe in Section IV how to set the t_1 value. Two inter-sequence alignment examples are shown in Fig.4 where only the image area below horizon is viewed as valid. Examples of detected regions B_1 in two frames of these videos are shown in Fig.5. However, there are many false alarms in them. In the following section, we introduce intra-sequence geometric alignment to remove these false alarms.

B. Intra-Sequence Geometric Alignment

The procedure for intra-sequence geometric alignment is similar to that for inter-sequence alignment. The difference is that both the reference (the frame to be warped) and target frames are from the same video this time. For the intra-sequence alignment on $\mathbf{R}$, we align $\mathbf{R}^i$ and $\mathbf{R}^{i-k}$. For the intra-sequence alignment on $\mathbf{T}$, we align $\mathbf{T}^i$ and $\mathbf{T}^{i-k}$.

TABLE I
MRANSAC-BASED HOMOGRAPHY ESTIMATION

1. Apply SIFT algorithm to get a set of SIFT feature descriptors for $\mathbf{R}^i$ and $\mathbf{T}^i$
2. Find the putative matches, $\mathbb{D} : \{\mathbf{X}_1 \leftrightarrow \mathbf{X}_2\}$, between these SIFT features of $\mathbf{R}^i$ and $\mathbf{T}^i$.
3. Call mRANSAC (Table II) to find the optimal H_{inter}.

Fig. 5. Suspicious object areas B_1 (highlighted) based on the inter-sequence alignment.

Generally, the choice of k depends on the speed of the moving camera. If the camera moves fast, k should be set to a small number, and vice-versa. We take $k = 5$ for our experiments, with the platform moving at an approximate speed of 30 kilometers per hour and the displacement of the camera between i-th and $(i-k)$-th frames being about 10 meters. Since the illumination variation between the intra-sequence reference and target frames is usually small, the intra-sequence alignment generally aligns the dominant planes very well (even when shadow appears in one and disappears in the other, as in the case shown in the top row of Fig.6). To remove false alarms on high objects, as proposed in Section III-B1, we use the results of the intra-sequence alignment for $\mathbf{R}$. This process is illustrated in Fig.6.

1) Removal of False Alarms on High Objects: After applying the intra-sequence alignment to $\mathbf{R}^i$ and $\mathbf{R}^{i-k}$ (as shown in the top row of Fig.6. R_1C_1 and R_1C_2 correspond to the 165^{th} and 170^{th} frames of $\mathbf{R}$. R_1C_3 is the warped R_1C_1 to fit R_1C_2. R_1C_4 and R_1C_5 are the difference images before and after alignment respectively.), we warp $\mathbf{R}^{i-k}$ into $\tilde{\mathbf{R}}^{i-k}$ to fit $\mathbf{R}^i$. Thus we can obtain the NCC image between $\tilde{\mathbf{R}}^{i-k}$ and $\mathbf{R}^i$, shown as R_2C_1 in Fig.6. Intuitively, we can locate the high-object areas in $\mathbf{R}^i$ by setting a suitable threshold t_2 on the NCC image because the high objects in $\tilde{\mathbf{R}}^{i-k}$ are deformed and the NCC scores at these locations are usually low. The pixels whose NCC values are lower than t_2 are treated as possible high-object areas of $\mathbf{R}^i$, denoted by the binary image B_r^{ho}. An example of B_r^{ho} is shown as R_2C_2 in Fig.6. We notice that some lower parts of the road area in R_1C_3 are blurred due to the homography deformation, therefore these areas also have low NCC values, which can explain that some road areas are also highlighted in R_2C_2 although they are flat. To deal with this problem, we have tried two principled ways: one is to blur the reference frame by Gaussian filtering and the other is to deblur the transformed target frame by trilinear interpolation. However, the results are not better than a simple adhoc one, i.e., we remove from B_r^{ho}

TABLE II
MRANSAC

Input: N_p pairs of putative matches - $\mathbb{D} : \{\mathbf{X}_1 \leftrightarrow \mathbf{X}_2\}$.
Hm - the Homography model used to fit the data
n - the minimum number of data required to fit the model (4)
mI - the maximum number of iterations allowed (500)
$mInd$ - the maximum number of trials to select a non-degenerate data set (100).
t - a threshold for determining when a datum fits a model (0.001)
Output: $bestH$ - model which optimally fits the data

1. Normalizing $\mathbf{X}_1$. A homogeneous scale factor of 1 is padded as the last coordinate of $\mathbf{X}_1$. Then move their average to origin by subtracting their mean, and scale them to average length of $\sqrt{2}$. Let $T_1 = T^1_{scale} \times T^1_{translate}$. Similarly normalize $\mathbf{X}_2$ and get T_2. Represent the normalized data by $\breve{\mathbb{D}} : \{\breve{\mathbf{X}}_1 \leftrightarrow \breve{\mathbf{X}}_2\}$.
2. Let $trialcount = 0$; $trialcountND = 0$; $bestH$ =nil;
   Let $N = 1$;$inliers$ =nil;$maxScore$ =0;$bestInliers$ =nil;
3. While $(N > trialcount)$ & $(trialcount \leq mI)$
   $degenerate = 1$; $count = 1$;
   While $(degenerate)$ & $(count \leq mInd)$;
   Randomly sample n pairs of data, $\mathbf{P}_n$, from $\breve{\mathbb{D}}$.
   If $\mathbf{P}_n$ is not degenerate.
   Call H-Fitting (Table III) to get H_{temp} from $\mathbf{P}_n$.
   If H_{temp} is empty
   $degenerate = 1$;
   $count = count + 1$;
   If $(degenerate)$
   $trialcount = trialcount + 1$;
   Break;
   Evaluate H_{temp} by the inliers matching H_{temp}.
   $HX_1 = H_{temp} \times \breve{\mathbf{X}}_1$;
   $invHX_2 = H^{-1}_{temp} \times \breve{\mathbf{X}}_2$
   $H\breve{X}_1 = HX_1./HX_1(3,:)$;
   $invH\breve{X}_2 = invHX_2./invHX_2(3,:)$;
   $d2 = sum((\breve{\mathbf{X}}_1 - invH\breve{X}_2).^2) + sum((\breve{\mathbf{X}}_2 - H\breve{X}_1).^2)$
   $inliers$=find$(d2 < t)$;
   $ninliers = |inliers| \times scatter(inliers)$;
   If $(ninliers > maxScore)$
   $maxScore = ninliers$;
   $bestInliers = inliers$;
   $bestH = H_{temp}$;
   $fracinliers = ninliers/N_p$;
   $pNoOutliers = 1 - fracinliers^n$;
   $N = \log(1-p)/\log(pNoOutliers)$;
   $trialcount = trialcount + 1$;
4. Call H-Fitting (Table III) based on $bestInliers$ to get H;
5. $H = T_2^{-1} \times H \times T_1$;

TABLE III
H-FITTING

Input: Randomly sampled n pairs of data, $\mathbf{P}_n$.
Output: 3×3 Homography matrix $H^T = [\mathbf{h}^1, \mathbf{h}^2, \mathbf{h}^3]$

1. Normalizing $\mathbf{P}_n$ in the way as in Table II and denoting the normalized $\mathbf{P}_n$ as $\breve{\mathbf{P}}_n$
2. For each pair of data in $\breve{\mathbf{P}}_n$, $\breve{\mathbf{p}}^i_1 = [u_1, v_1, w_1]^T$ and $\breve{\mathbf{p}}^i_2 = [u_2, v_2, w_2]^T$, $i = 1, ..., n$, construct $A_i = \begin{pmatrix} \mathbf{0}^T & -w_2 \times (\breve{\mathbf{p}}^i_1)^T & -v_2 \times (\breve{\mathbf{p}}^i_1)^T \\ w_2 \times (\breve{\mathbf{p}}^i_1)^T & \mathbf{0}^T & -u_2 \times (\breve{\mathbf{p}}^i_1)^T \\ -v_2 \times (\breve{\mathbf{p}}^i_1)^T & u_2 \times (\breve{\mathbf{p}}^i_1)^T & \mathbf{0}^T \end{pmatrix}$
3. Stack the n A_i into a $3n \times 9$ matrix $A = [A_1, ..., A_n]^T$.
4. Get the least square solution of $A \times \mathbf{h} = \mathbf{0}$ [3] by SVD of A: $A = \Sigma\Phi\Lambda^T$. The unit vector corresponding to the smallest singular value is the solution $\mathbf{h}$.
5. Reshape $\mathbf{h}$ into H and denormalize H as in Table II

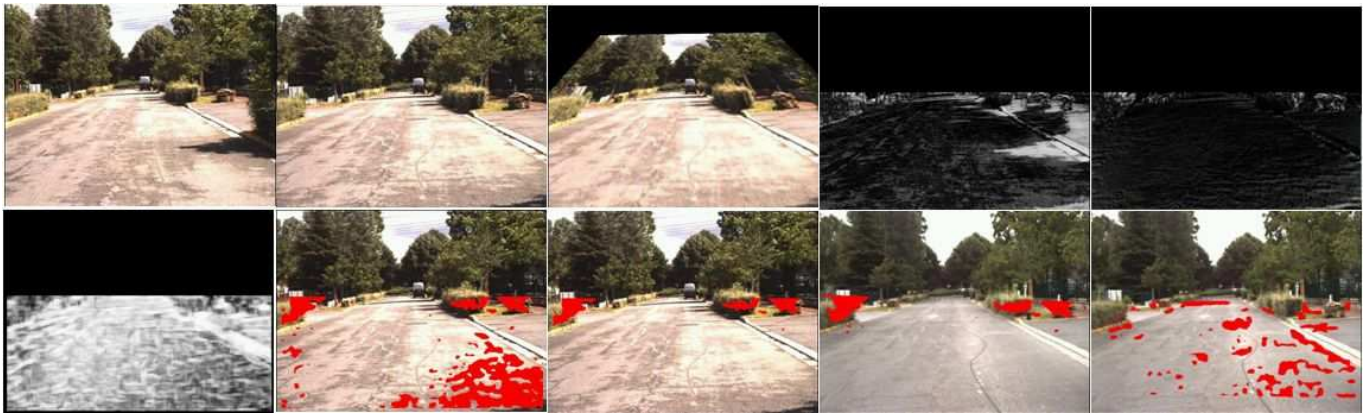

Fig. 6. Illustration of intra-sequence alignment for $\mathbf{R}$ and removing false alarms on high objects based on the alignment results. See text in Section III-B1 for details.

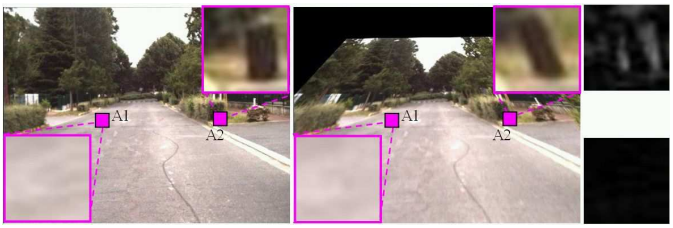

Fig. 7. Illustration of local appearance comparisons. $\mathbf{T}^i$ is shown in the left column and $\tilde{\mathbf{T}}^{i-k}$ is shown in the middle column. A1 and A2 are two exemplar local patches, which are zoomed in at the bottom-left and top-right corners. The last column shows their grayscale intensity difference with the top and bottom one corresponding to A2 and A1 respectively.

the highlighted clusters whose centroids are lower than 40% of the image height. The B_r^{ho} after such a removal is shown as R_2C_3 in Fig.6. Correspondingly, the high-object areas in $\mathbf{T}^i$ are represented by B_t^{ho}, which is obtained by transforming B_r^{ho} based on H_{inter}, shown as R_2C_4. The remaining suspicious areas after removing B_t^{ho} are shown in R_2C_5, which can be represented by

$$B_2 = B_1 - B_1 \cap B_t^{\text{ho}} \tag{1}$$

Based on Eq.(1), there is no risk of removing the true suspicious objects lying in the bottom part of the image by getting rid of the highlighted clusters with low-centroids from B_r^{ho}. Although false alarms in low-end flat regions (road areas) still exist, however, as shown in the next section, these false alarms can be discarded by using the result of intra-sequence alignment for both $\mathbf{R}$ and $\mathbf{T}$. Due to the problem caused by the blurring resulting from the homography transformation, we will seek to use the difference of illumination-normalized aligned frames instead of normalized cross correlation of them again.

2) Removal of False Alarms on the Dominant Plane: Given that any 3D objects lying on top of the dominant plane are deformed after the intra-sequence alignment, while any flat objects remain almost unchanged, we use the difference of grayscale pixel values between $\tilde{\mathbf{T}}^{i-k}$ and $\mathbf{T}^i$, and that between $\tilde{\mathbf{R}}^{i-k}$ and $\mathbf{R}^i$ for removing false alarms on the dominant plane. Figure 7 gives an illustrative example, where obviously the appearances of the patches are quite different in non-flat areas in the target and warped reference frames, while they are very similar in the flat ones. Therefore, we can remove the false alarms in flat areas by setting a threshold on the difference image. Generally, a small threshold should be used if $\mathbf{T}$ is taken at dawn or dusk, while a large one is preferred if $\mathbf{T}$ is taken at noon (especially with strong light). To make the threshold-tuning process easier for general illumination conditions, we normalize the two aligned grayscale intra-sequence frames as follows.

For a 25×25 region S around a given pixel at (x,y), we compute the brightness as the mean pixel value of S, denoted as $mean_S$. To maintain stability in regions with low contrast, we compute the contrast as $C_S = max(max_S - min_S, 20)$, where max_S and min_S are respectively the mean of the 15 maximum and the 15 minimum values over S. The normalized intensity value, $\hat{i}(x,y)$, is computed by $\hat{i}(x,y) = 0.5 + \frac{i(x,y) - mean_S}{C_S}$, with the pixel values falling outside the 0 to 1 range after the process are clipped to lie within the range. We binarize the difference of the two normalized grayscale images with a threshold t_3, which is denoted by B_{intra}^t.

Similarly, we apply the local appearance comparison process to the result of the intra-sequence alignment for $\mathbf{R}$. We binarize the difference of the normalized image pair to be B_{intra}^r with the same threshold t_3 (by tuning t_3 on our training videos, we set it to 0.15). Based on H_{inter}, B_{intra}^r is converted to $\tilde{B}_{\text{intra}}^r$. The left suspicious areas after the local appearance comparison are computed by

$$B_3 = B_2 \cap (B_{\text{intra}}^t - (B_{\text{intra}}^t \cap \tilde{B}_{\text{intra}}^r)) \tag{2}$$

We illustrate the process of removing false alarms in flat areas in Fig.8. The first row shows the intra-sequence alignment for $\mathbf{T}$. R_1C_1 and R_1C_2 are the 165^{th} and 170^{th} frames of $\mathbf{T}$ respectively. R_1C_3 is the warped R_1C_1. R_1C_4 and R_1C_5 are the difference images before and after alignment respectively. In the second row, R_2C_1 is the NCC image between R_1C_2 and R_1C_3. R_2C_2 is the illumination-normalized R_1C_2. R_2C_3 is the illumination-normalized R_1C_3. R_2C_4 is the difference of R_2C_2 and R_2C_3. R_2C_5 shows the highlighted non-flat objects by setting a threshold on R_2C_4. The third row shows the intra-sequence alignment for $\mathbf{R}$. R_3C_1 and R_3C_2 are the 165^{th} and 170^{th} frames of $\mathbf{R}$ respectively. R_3C_3 is the

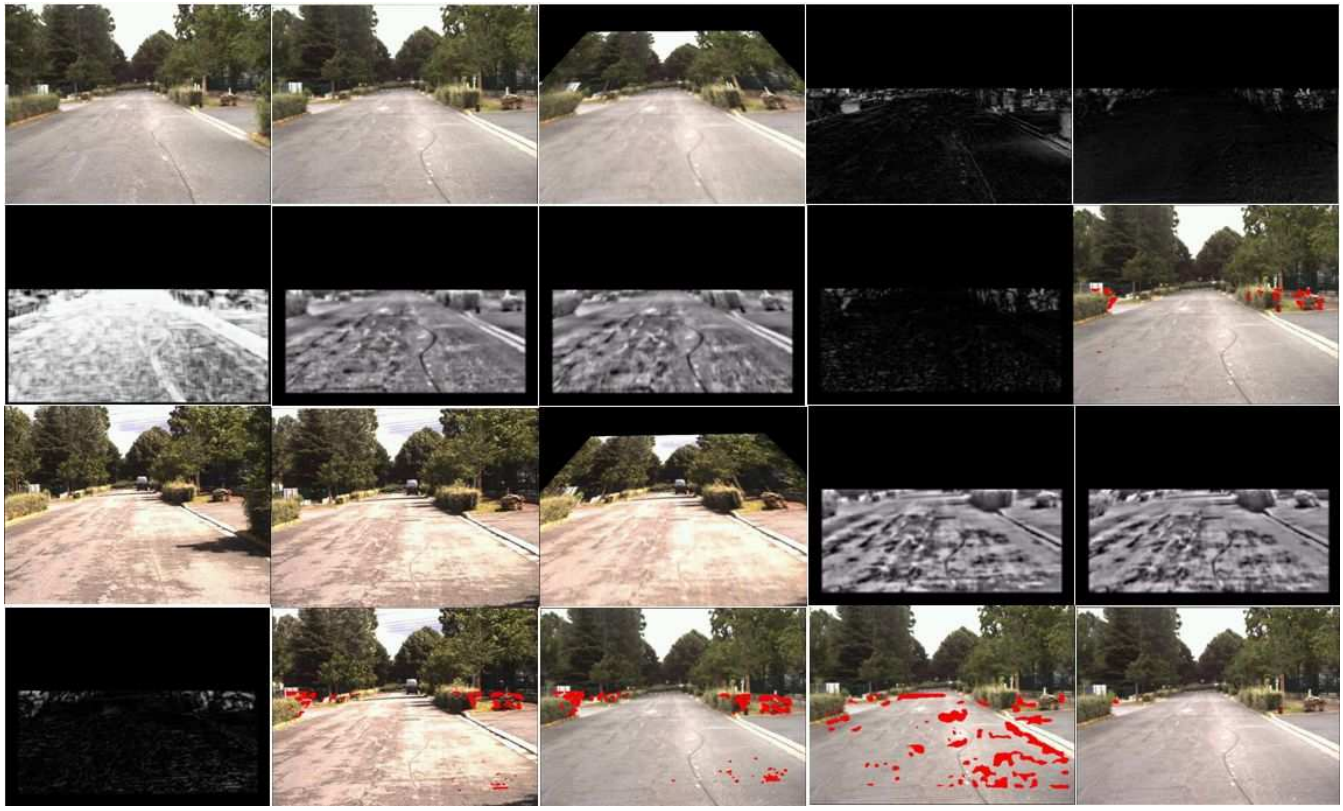

Fig. 8. Illustration of intra-sequence alignment for both $\mathbf{T}$ and $\mathbf{R}$, and removing false alarms in flat areas by local appearance comparison of the alignment results. See text in Section III-B2 for details.

warped R_3C_1. R_3C_4 is the illumination-normalized R_3C_2. R_3C_5 is the illumination-normalized R_3C_3. In the fourth row, R_4C_1 is the difference of R_3C_4 and R_3C_5. R_4C_2 shows the highlighted non-flat objects in $\mathbf{R}^i$ by setting a threshold on R_4C_1. R_4C_3 is obtained by transforming R_4C_2 based on H_{inter}. R_4C_4 is the same as R_2C_4 of Fig.6. R_4C_5 shows the remaining suspicious object areas based on Eq.(2). By checking R_2C_1 (the NCC image of $\tilde{\mathbf{T}}^{i-k}$ and $\mathbf{T}^i$) in this figure, one can see that that it is not a good choice to use NCC for removing false alarms in flat areas (because the NCC values are low in some non-suspicious flat areas).

C. Temporal filtering

We use temporal filtering on B_3 to get our final detection. Let K be the number of buffer frames used for temporal filtering. We assume that $\mathbf{T}^i$ is the current frame, and the remaining suspicious object areas in $\mathbf{T}^i$ after inter- and intra-sequence alignment is denoted by B_3^i. We stack B_3^{i-3}, B_3^{i-2}, B_3^{i-1} and B_3^i into a temporal buffer T_{buffer}. We also stack the homography transformations between any two neighboring frames of the buffer into H_{buffer}. Based on these transformations, B_3^{i-3}, B_3^{i-2} and B_3^{i-1} are respectively transformed to the state which temporally corresponds to the i-th frame, and are intersected with B_3^i respectively. The final detection map is the intersection of these intermediate intersection. We set a threshold for the size of the smallest non-zero cluster in the final detection map (we set 8 for all of our experiments). The details of temporal filtering are listed in Table IV.

TABLE IV
TEMPORAL FILTERING.

Input: Target video sequence, $\mathbf{T}$, buffer size K, and potential suspicious area B_3^i, where i is the frame index.
Output: Detection map: I_{dp}

1. Initialization: Stack the initial $B_3^i, i = 0, ..., K-1$ into the temporal buffer T_{buffer}. Compute the homography matrix H for any two neighboring frames of the initial $K-1$ frames of $\mathbf{T}$ and stack them into H_{buffer}.
2. For an incoming frame, $\mathbf{T}^i, i = K, ..., \infty$, do:
   2.a. Get B_3^i and let $T_{\text{buffer}}(K) = B_3^i$; compute H between $\mathbf{T}^{K-1}$ and $\mathbf{T}^K$, and let $H_{\text{buffer}}(K-1) = H$; let $I_{\text{dp}} = T_{\text{buffer}}(K)$.
   2.b. For $u = 1$ to $K-1$
   $M_1 = T_{\text{buffer}}(u)$;
   For $v = u$ to $K-1$
   $M_2 = H_{\text{buffer}}(v) \times M_1$;
   $M_1 = M_2$;
   $I_{\text{dp}} = I_{\text{dp}} \cap M_1$;
   2.c. Update T_{buffer};
   For $j = 1$ to $K-1$
   $T_{\text{buffer}}(j) = T_{\text{buffer}}(j+1)$;
   2.d. Update H_{buffer};
   For $j = 1$ to $K-2$
   $H_{\text{buffer}}(j) = H_{\text{buffer}}(j+1)$.

IV. EXPERIMENTAL RESULTS

To collect videos, the camera is mounted on the top of a vehicle. The height of the camera (relative to the ground) is about 2.2 meters. In fact, we have tested a couple of heights, including the case where the camera is fixed inside the car. We have observed that there is no significant decadence in

performance as long as the road region covers most of the image area, i.e., the horizon should be guaranteed at the upper half of the image, for which homography transformation between the reference and target frames can be reliably estimated for such cases. For convenience, we adopt the same height of camera for both reference and target videos. We have collected eight reference-target video pairs from three different scenes for training the necessary parameters t_1, t_2 and t_3 (threshold t_1 is used to find all possible suspicious areas based on the NCC image between the aligned inter-sequence frame pair. t_2 is set for the NCC image between the aligned intra-sequence frame pair of $\mathbf{R}$, which is used to remove false alarms on the high objects. t_3 is the threshold set on the difference image between the aligned illumination-normalized intra-sequence frame pair of both $\mathbf{R}$ and $\mathbf{T}$, which is used to remove false alarms in flat region). Among them, five pairs of videos have significant non-global illumination variation between $\mathbf{R}$ and $\mathbf{T}$, the other three have little/global lighting change. By "global" illumination variation, we mean that $\mathbf{R}$ and $\mathbf{T}$ can be photometrically aligned well using only one affine transformation matrix. For example, in the third row of Fig.12, there is little illumination change between $\mathbf{R}$ and $\mathbf{T}$. In the second, fourth and fifth rows, there is a global change, while large local illumination changes can be observed for the first and sixth rows. Figure 9 shows representative frame pairs from three reference-target training video-pairs.

To test the performance of the proposed framework, altogether 15 reference-target test video pairs are collected from another five different scenes. For three scenes, the reference and target videos were taken at different hours in the same day. For the other two scenes, the reference and target videos were taken one just after the other. The original resolution of video frame is 1024×768. In practice, we downsample it to the size of 512×384 to reduce computational load.

We choose different video pairs with varying illumination conditions as training set because we experimentally notice that no fixed t_1 can well adapt to all possible illumination variations. By tuning on the training data, we have set t_1 as follows:

$$t_1 = \begin{cases} 160 & \text{if the lighting change between } \mathbf{R} \text{ and } \mathbf{T} \text{ is non-global} \\ 200 & \text{otherwise} \end{cases} \tag{3}$$

where the NCC image is normalized to the range of 0 to 255, and the precise meaning of "having a non-global lighting change between $\mathbf{R}$ and $\mathbf{T}$" will be given in the next paragraph. The threshold t_2 is set to be smaller than t_1 because the higher t_2 is, the more image regions will be highlighted as high-object areas. Thus, according to Eq.(1), it is possible to remove the true suspicious object with a large t_2. In practice, we set $t_2 = t_1 - 25$.

In testing, once given a pair of videos, several initial frame pairs are aligned and compared to judge whether the illumination variation between $\mathbf{R}$ and $\mathbf{T}$ is a global or non-global change (we also deem it as a global change if there is little variation). The criterion is based on photometric alignment between two frames. Figure 11 illustrates the idea of photometric alignment. After the geometric alignment, we apply traditional

Fig. 9. Example frames of three pairs of training videos. Left column: frames of reference videos. Right column: frames of target videos.

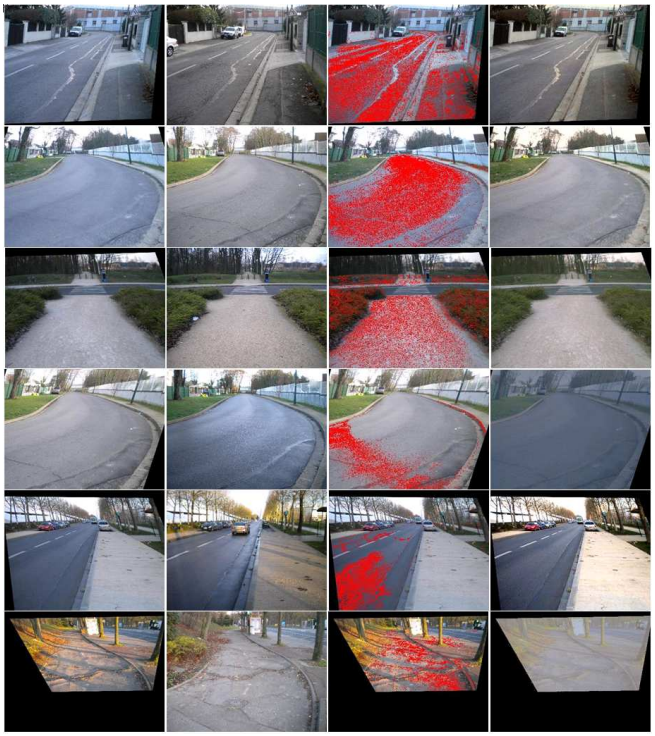

Fig. 10. Examples of photometric alignment for learning t_4 for classifying a global or non-global illumination variation between two corresponding frames. First column: geometrically aligned reference frames. Second column: target frames. Third column: inliers (highlighted with red color) from the optimal affine color transformation. Fourth column: photometrically aligned reference frames based on the optimal affine transformation.

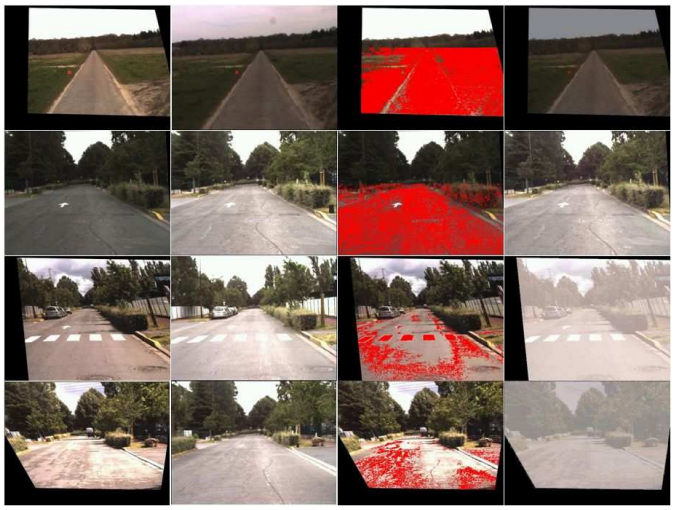

Fig. 11. Examples of photometric alignment for deciding a global or non-global illumination variation between two corresponding frames of the test video-pairs. First column: geometrically aligned reference frames. Second column: target frames. Third column: inliers (highlighted with red color) from the optimal affine color transformation. Fourth column: photometrically aligned reference frames based on the optimal affine transformation.

RANSAC to find an optimal affine transformation between the pixel color values of two corresponding frames. We can find the set of pixels that fit the optimal affine transformation, and call these pixels optimal inliers. We check the percentage P_t, of the number of optimal inliers to the total number valid pixels (only those non-zero pixels below the vanishing point in the aligned reference frame are deemed as valid). If P_t is larger than a threshold t_4, the illumination variation between the two frames is deemed to be global. Otherwise, it is non-global. By photometric alignment based on RANSAC, we tune t_4 on another training data set D_s, which is formed by collecting 45 additional pairs of aligned images (not from the training videos), and finally we set t_4 to be 35%. Figure 10 shows examples of photometric alignment on some frame-pairs of D_s, where the illumination variation in the top three pairs of frames is global, and non-global for the bottom three pairs. Due to the significant appearance change in road area before and shortly after rain, we also view the image-pairs which were taken before and after rain as the case of non-global illumination change even if the real illumination change is global (e.g., cloudy day as shown in the fourth row of Fig.10).

Figure 11 shows examples of photometric alignment on some of the test video frames, where in the first two rows, the illumination change between the aligned reference frame and the target frame can be viewed as global, with P_t being 72% and 59%, while P_t is only about 17% and 15% respectively for the third and fourth rows where the illumination change is large.

In the 15 test video pairs, we have manually checked the total number of suspicious objects. These suspicious objects include test objects and non-test objects (i.e., objects not arbitrarily laid by us, and not in the reference video but in the target video, e.g., parked car). The length of the 15 test videos ranges from 70 to 230 frames. There are about 23 suspicious objects overall and we successfully detect 21 of them. The height of our test objects ranges from about 4 to 80 centimeters. Among the 15 test videos, the test objects are successfully detected in at least 8 frames for 11 videos. Overall, they appear with a corresponding image area of at least 120 pixels in a total of 351 frames, and they are detected in 173 of these frames. Table V lists the number of frames (denoted by N_0) where the test object is detected, and the number of frames (denoted by N) where the image area corresponding to the test object is larger than 120 pixels for the 15 test videos.

TABLE V
NUMBER OF DETECTION FRAMES FOR EACH OF THE 15 TEST VIDEOS. N_0: THE NUMBER OF FRAMES WHERE THE TEST OBJECT IS DETECTED FOR EACH VIDEO. N: THE NUMBER OF FRAMES WHERE THE IMAGE AREA CORRESPONDING TO THE TEST OBJECT IS LARGER THAN 120 PIXELS.

N	15	41	13	41	14	10	9	35
N_0	8	24	8	17	9	3	3	24
N	29	19	13	15	50	29	18	
N_0	17	9	3	10	26	12	0	

Detection fails in only one video where the test object is almost flat. In the test videos where suspicious objects are present, we have only about 49 false alarms out of 2053 total frames (we deem an alarm as false even when the false suspicious object is detected as a true positive in only one frame). To test on videos where no suspicious objects exist, we just swap the reference and test videos and we have about 27 false alarms out of 2053 frames. To avoid detecting moving vehicles as suspicious objects (because the car's movement causes large highlighted area in B^t_{intra}), we have set a threshold of 400 pixels on the area of the largest cluster in B^t_{intra}. Because we have used the results of intra-sequence alignment for $\mathbf{T}$, we can avoid some false alarms caused by some objects which appear in the reference video but disappear in the test video, and can also avoid the false alarms caused by shadows and highly saturated regions (these are demonstrated in some of our demo videos).

We have uploaded the 15 test demo videos to a project webpage (http://sites.google.com/site/huikongsite/Home/research). Figure 12 shows some examples of detected suspicious objects. Since the framework is proposed for detecting any general non-flat static object, it cannot be directly applied for flat object detection. Therefore, our framework fails in cases such as the one in the last row of Fig.12 because the suspicious object is almost flat. Similarly, some suspicious objects cannot be detected when the camera is far from them. This is because the deformation around the object area is too small when it is far away. To deal with this problem, one would have to extend our framework by adding another camera to zoom in the far scenes.

V. CONCLUSION

This paper proposes a novel framework for detecting non-flat abandoned objects by a moving camera. Our algorithm finds these objects in the target video by matching it with a reference video that does not contain them. We use four main ideas: the inter- and intra-sequence geometric alignment, the local appearance comparison, and the temporal filtering based

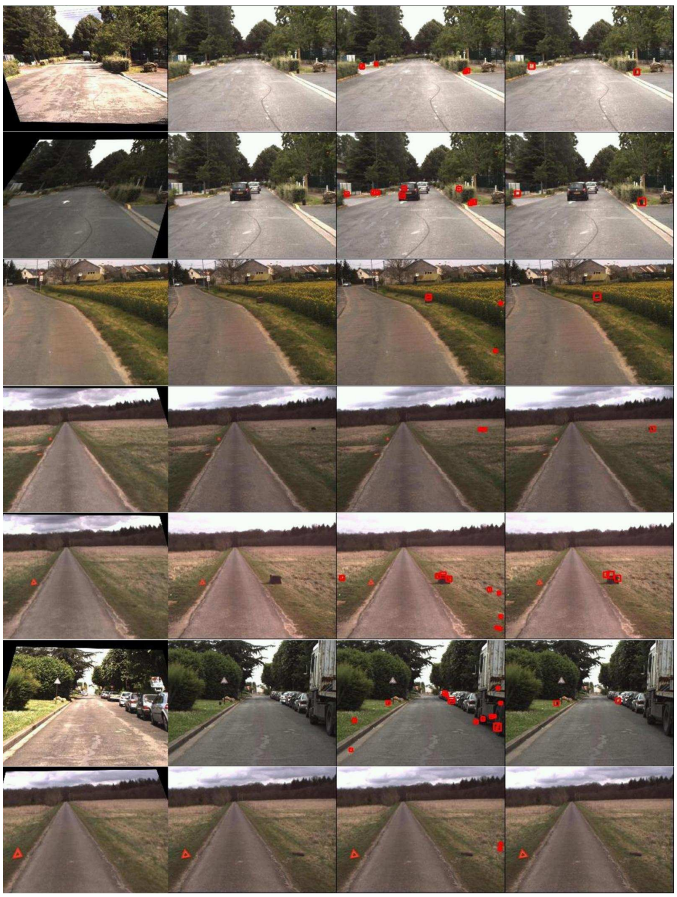

Fig. 12. Examples of detected abandoned objects (note that we fail in detecting the object in the last row). First column: geometrically aligned reference frames. Second column: target frames. Third column: detected objects before temporal filtering. Fourth column: final detection.

on homography transformation. Our framework is robust to large illumination variation, and can deal with false alarms caused by shadows, rain, and saturated regions on road. It has been validated on fifteen test videos.

VI. Acknowledgements

This work was supported by a Postdoctoral Fellowship from Evitech and DGA of France. Thanks to Andrew Zisserman, Josef Sivic, and Oliver White for their helpful discussion. The authors also appreciate Philippe Drabczuk and Pierre Bernas of Evitech Technologies for their efforts in collecting test videos that are used in this paper.

References

[1] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of the ACM*, 1981.
[2] D. Forsyth and J. Ponce. Computer vision: A modern approach. *Prentice Hall*, 2002.
[3] S. Guler and M. K. Farrow. Abandoned object detection in crowded places. *PETS Workshop*, 2006.
[4] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. *2nd edition, Cambridge University Press*, 2003.
[5] H. Kong, J.-Y. Audibert, and J. Ponce. Vanishing point detection for road detection. *CVPR*, 2009.
[6] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.
[7] F. Porikli, Y. Ivanov, and T. Haga. Robust abandoned object detection using dual foregrounds. *EURASIP Journal on Advances in Signal Processing*, 2008.
[8] K. Smith, P. Quelhas, and D. Gatica-Perez. Detecting abandoned luggage items in a public space. *PETS Workshop*, 2006.
[9] M. Spengler and B. Schiele. Automatic detection and tracking of abandoned objects. *PETS Workshop*, 2003.
[10] R. Szeliski. Image alignment and stitching: A tutorial. *Foundation and Trend in Computer Graphics and Vision*, 2006.

PLACE PHOTO HERE

Hui Kong received the PhD degree in computer vision from Nanyang Technological University, Singapore, in 2007. He was a postdoctoral fellow in the Willow team of Ecole Normale Supérieure and the Imagine team of Ecole Nationale des Ponts from 2008 to 2009 in Paris. From 2005 to 2007, he worked as a full-time R&D engineer on digital consumer electronics in the Panasonic Singapore Labs. His research interests lie in computer vision, pattern recognition, and image processing. He is a member of IEEE.

PLACE PHOTO HERE

Jean-Yves Audibert received the PhD degree in mathematics from the University of Paris VI in 2004. Since then, he is a researcher in the Computer Science department at Ecole des Ponts ParisTech. Since 2007, he is also a research associate in the Computer Science department at Ecole Normale Supérieure in a joint INRIA/ENS/CNRS project. His research interest and publications range from Statistics to Computer Vision, including theoretical properties of learning procedures, boosting algorithms, kernel machines, object recognition, image segmentation, content-based image retrieval.

PLACE PHOTO HERE

Jean Ponce received the Doctorat de Troisieme Cycle and Doctorat d'Etat degrees in Computer Science from the University of Paris Orsay in 1983 and 1988. He has held Research Scientist positions at the Institut National de la Recherche en Informatique et Automatique (1981–1984), the MIT Artificial Intelligence Laboratory (1984–1985), and the Stanford University Robotics Laboratory (1985–1989), and served on the faculty of the Dept. of Computer Science at the University of Illinois at Urbana-Champaign from 1990 to 2005. Since 2005, he has been a Professor at Ecole Normale Superieure in Paris, France. Dr. Ponce is the author of over 150 technical publications, including the textbook "Computer Vision: A Modern Approach", in collaboration with David Forsyth. Dr. Ponce is a member of the Editorial Boards of the International Journal of Computer Vision (for which he served as Editor-in-Chief from 2003 to 2008), the SIAM Journal on Imaging Sciences Foundations, and Trends in Computer Graphics and Vision. He was also an Area Editor of Computer Vision and Image Understanding (1994–2000) and an Associate Editor of the IEEE Transactions on Robotics and Automation (1996–2001). He was Program Chair of the 1997 IEEE Conference on Computer Vision and Pattern Recognition and served as General Chair of the year 2000 edition of this conference. He also served as General Chair of the 2008 European Conference on Computer Vision. In 2003, he was named an IEEE Fellow for his contributions to Computer Vision, and he received a US patent for the development of a robotic parts feeder.