

Night-time outdoor surveillance with mobile cameras

Anonymous

Keywords: video surveillance, video synchronization, video alignment, change detection.

Abstract: This paper addresses the problem of video surveillance by mobile cameras. We present a method that allows online change detection in night-time outdoor surveillance. Because of the camera movement, *background* frames are not available and must be "localized" in former sequences and registered with the current frames. To this end, we propose a Frame Localization And Registration (FLAR) approach that solves the problem efficiently. Frames of former sequences define a database which is queried by current frames in turn. To quickly retrieve nearest neighbors, database is indexed through a visual dictionary method based on the SURF descriptor. Furthermore, the frame localization is benefited by a temporal filter that exploits the temporal coherence of videos. Next, the recently proposed ECC alignment scheme is used to spatially register the synchronized frames. Finally, change detection methods apply to aligned frames in order to mark suspicious areas. Experiments with real night sequences recorded by in-vehicle cameras demonstrate the performance of the proposed method and verify its efficiency and effectiveness against other methods.

1 INTRODUCTION

In recent years, visual-surveillance systems based on video processing and understanding techniques have attracted increasing interest from urban and building security, military related field and security and video patrolling systems. The principal aim of such systems is to detect potential suspicious items or signs of intrusion, and consequently generate a warning-sign to a human operator. This detection mainly consists of identifying changes between images of the same scene but temporally separated. Most change detection methods proposed in the literature deal with stationary cameras as surveyed in (Radke et al., 2005). This amounts to detect differences using a stationary background. Such approaches reflect minor applicability for wide areas where multiple cameras are required. This drawback can be overcome using a mobile camera to scan these places, but the problem becomes more challenging due to the non-stationary background and varying ambient illumination.

The latter scenario is what we consider in this paper. Specifically, we present a method that helps the video analyst to robustly detect potential and suspicious signs of intrusion by vehicles that repeatedly patrol sensitive areas and private buildings at night-time. This detection cannot rely on specific classifiers mainly due to the following factors: (1) the quality of the acquired images can be significantly degraded at night-time and (2) these signs may be both random

anomalies, such as an intruder, and stationary anomalies, *e.g.* a suspicious suitcase, with arbitrary shapes, color or texture. To this end, we propose an efficient and effective framework to detect potential anomalies exploiting the similarities occurred when one vehicle repeatedly patrols the same ride. This consists of comparing a pair of video sequences recorded by a forward-facing camera attached to the windscreen of the vehicle whose view is what the driver sees. Hence, signs of intrusion or missing objects occurred in the interim between successive rounds can be detected by a background subtraction approach. However, this obviously requires the spatio-temporal alignment of the current sequence with the one captured during the previous round, *i.e.* the video synchronization and the spatial registration of corresponding frames.

Current video synchronization algorithms estimate the temporal relation between two sequences once they have been acquired. However, our goal is to online detect changes at a reasonable rate. Thus, instead of synchronizing the whole sequences solving an offline global optimization problem, the proposed framework counts on a Frame Localization And Registration (FLAR) scheme. In short, given each newly acquired frame during the current ride, we *temporally localize* it against the *background sequence* of the previous ride. This aims in other words at assigning each current frame to a background frame so that their viewpoints ideally are the closest ones. Since efficiency is of major importance in online solutions, the extraction of the corresponding frame relies on

an image retrieval scheme based on the SURF descriptor (Bay et al., 2008). A temporal filter applies to the outcome of the retrieval task in order to handle false positives (outliers). Then, we have to spatially register the corresponding frames into the same image–coordinate system. As the video acquisition takes place at different times, the appearance of corresponding frames varies. To cope with such variations, we adopt the recently proposed ECC image alignment scheme (Evangelidis and Psarakis, 2008) that offers the desired robustness. As a final step, different metrics that count on image differences are applied to detect changes and mark areas of interest. The contribution of this paper is summarized as follows:

- A challenging case for outdoor surveillance by mobile cameras during the night is investigated.
- The proposed FLAR scheme reflects a solution for online surveillance instead of postprocessing after the acquisition of both sequences.
- It incorporates efficient tasks that allows us to envision a real-time execution in GPU-based environment.
- The desired invariance to the motion style of surveillance vehicle (speed, acceleration, backward motion etc) is fulfilled.

1.1 Related Work

The challenging problem of detecting changes between videos acquired by mobile cameras at different times is considerably less tackled than the case of stationary cameras (Radke et al., 2005). Marcenaro *et al.* (Marcenaro et al., 2002) proposed an outdoor-surveillance based on fixed and pan/tilt mobile cameras. In this case, the mobile camera allows to exceed the limitations of the fixed camera that monitors the entire scene. When the position of the mobile camera is unknown, the system fails to detect changes since it needs to know the corresponding background image. Primdahl *et al.* (Primdahl et al., 2005) presented a method that analyzes videos for an automatic navigation of cameras in a specific, well-defined corridor. Sand and Teller (Sand and Teller, 2004) proposed a video matching scheme for two sequences recorded by moving cameras following nearly identical trajectories. Although it allows figures to be compared pixel-wise and detects differences, the key limitation of this algorithm is the computational time of computing a robust image-alignment for several possible pairs of corresponding frames. To make it efficient, Kong *et al.* (Kong et al., 2010) temporally aligned the two sequences using GPS data only and detect abandoned suspicious objects via inter-sequence ho-

mographies whereas temporal filtering was enabled to handle false alarms. In contrast, Soiban *et al.* (Soibam et al., 2009) and Haberdar and Shah (Haberdar, 2010) found manually the visually closest matching frame in the first video for each observed frame of the second one. Finally, Diego *et al.* (?) proposed a video alignment framework based on fusing image-based and GPS observations to spot differences between sequences taken at different times and by independently moving cameras, while Chakravarty *et al.* (Chakravarty et al.,) presented a mobile robot capable of repeating a manually trained route that detect any visual anomalies using stereo-based algorithm; these anomalies are subsequently tracked using a particle filter.

The rest of this paper is organized as follows: Section 2 describes the whole framework and specifically, subsection 2.1 presents the frame localization approach accompanied by a temporal filtering method, while the spatial registration and the change detection tasks are discussed in subsections 2.2 and 2.3 respectively. Experiments to validate the proposed algorithm are presented in Section 3 and results are discussed. Finally, in Section 4, the main conclusions are drawn.

2 FRAME LOCALIZATION AND REGISTRATION

Suppose we are given two video sequences represented as $I^r = \{I_m^r(\hat{\mathbf{x}})\}_{m=1}^M$ and $I^c = \{I_n^c(\mathbf{x})\}_{n=1}^N$, being M, N their number of frames and $\hat{\mathbf{x}} = [\hat{x}, \hat{y}]^t$, $\mathbf{x} = [x, y]^t$ their spatial coordinates respectively. The former, or reference sequence, denotes the sequence taken in a previous ride (background sequence), whereas the latter, or current sequence, is the one being recorded in the current ride following a similar trajectory. Then, the anomalies occurred in the meanwhile between successive rounds can be detected by matching and comparing the current with the reference sequence. That is, thresholding properly a difference image sequence between spatio-temporally aligned sequences allows the detection of missing objects or suspicious signs of intrusion.

To solve the above defined problem we propose a Frame Localization And Registration (FLAR) framework that is shown in Figure 1. Within this framework, the only assumption we make is that the vehicles follow a similar, approximately coincident, route. Initially, the most likely frame of a previous ride is extracted for each newly acquired frame in the current ride (localization step). When the sequences are recorded by on-board cameras in moving vehi-

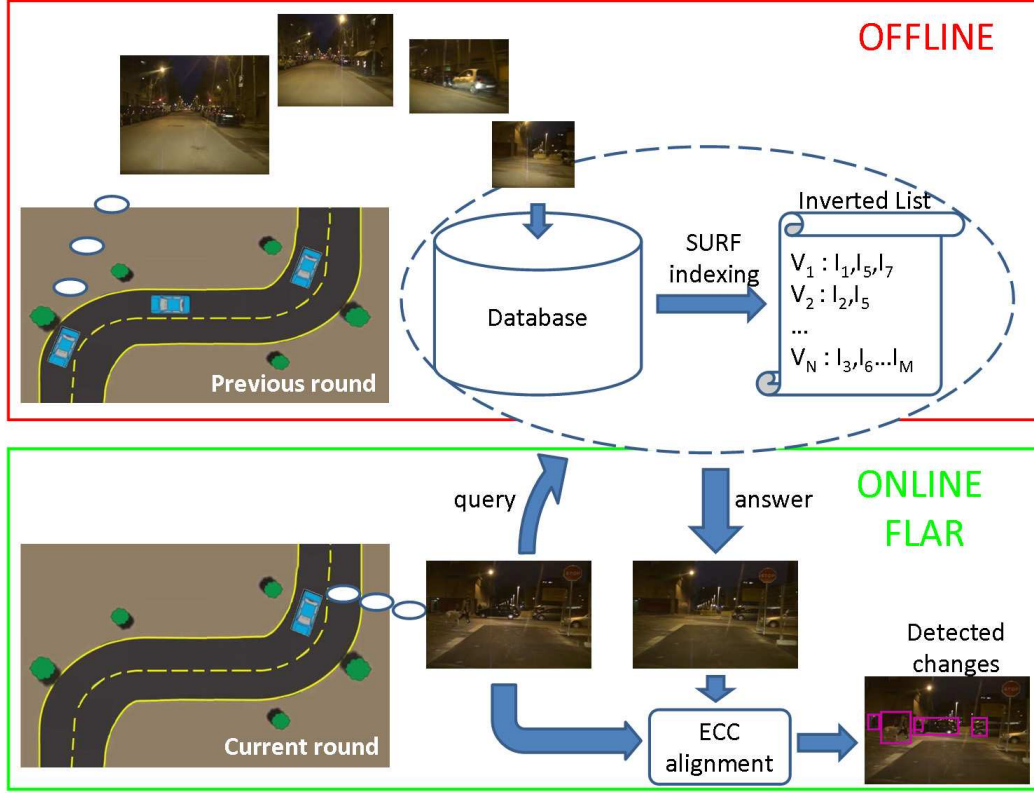


Figure 1: FLAR system for video surveillance with mobile cameras.

cles, this implies a challenging task because of the independently moving cameras and the non-coincident trajectories. As a result, the speed and the position of the cameras vary, while the ambient illumination can be different.

A few different video alignment approaches (Sand and Teller, 2004; Liu et al., 2008; Diego et al., 2010) could be adjusted to our problem. However, none of them is able to estimate the frame correspondence during the acquisition of the current sequence due to their high computational cost. Therefore, we propose an efficient on-line video synchronization algorithm that relies on an image retrieval scheme based on the SURF descriptor (Bay et al., 2008) and a temporal filter. This essentially assigns the n^{th} current frame to the reference index t_n with $t_n \in [1, M]$, thereby providing the desired invariance to the motion style of cameras. Once the correspondence pair (n, t_n) has been found, the dense alignment of the assigned frames is required (registration step) in order to compare them pixel-wise. This kind of comparison is necessary for our endmost goal, that is, the identification of regions that changed in the interim between the records. Now that the solution is roughly formulated, we proceed with the discussion about the details of each step of

the proposed framework.

2.1 Image-retrieval scheme

The image-retrieval scheme based on SURF descriptors (Bay et al., 2008) aims to evaluate efficiently a confidence matrix that measures the similarity of a current frame with all frames in the background, thus allowing the association of the current frame to the most similar reference frame. Our implementation resembles (Sivic and Zisserman, 2009) but disabling the vector quantization step and using only the inverted file. In short, we initially enable the SURF algorithm to localize keypoints in all the background frames and describe the area around them. A visual dictionary is then learned together with building an inverted index list as shown in Fig. 1. Notice that all this implies an offline task since can be done without having the current sequences at our disposal. Given now a current frame of the new ride, we extract its SURF descriptors and look for their closest visual word, thus voting for the assigned reference frames. In order to ignore very frequent visual words that are assigned to many reference frames, we enable the inverse-document-frequency (IDF) weighting scheme (Sivic and Zisser-

man, 2009). Fig. 2 shows an example of the confidence matrix together with two retrieval examples.

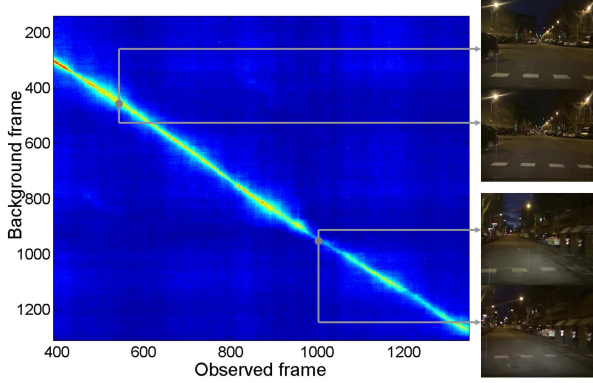


Figure 2: The confidence matrix captures the similarity between current and background frames; two retrieval examples (top frames) are also shown for two queries (bottom frames).

2.1.1 TEMPORAL FILTERING

The above retrieval system provides us with a confidence matrix. To extract the time mapping result, for each observed frame one could simply choose the reference index with the maximum confidence value. However, it might return an erroneous synchronization signal with sharp changes due to isolated points. To avoid sharp transitions, we choose for any query frame the maximum reference index subject to the constraint that lies in a tolerance interval. The latter is defined around the reference index assigned to the previous query frame. By recalling that t_n is the correspondence of the n^{th} , $[t_n - 10, t_n + 10]$ is the tolerance interval of the $(n + 1)^{th}$ current frame for our experiments.

It is obvious that the above smoothness operation rejects only very high frequencies in the time mapping signal. In order to obtain a smoother signal, we propose the use of a filter that applies in the signal t_n . Specifically, such a filter can be described by the standard *difference equation* (Lathi, 1998)

$$T_n = \sum_{i=0}^K b_i t_n - \sum_{j=1}^L a_j T_{n-j} \quad (1)$$

where T_n defines its output. In general, it constitutes an Infinite Impulse Response (IIR) filter, but when $a_j = 0$, it turns into a Finite Impulse Response (FIR) filter of order K (Lathi, 1998). It is important to note that this filter is a causal system and the current output depends only on previous input and output values, being capable for online and real-time solutions. Both type of filters were tested using $K = L = 3$,

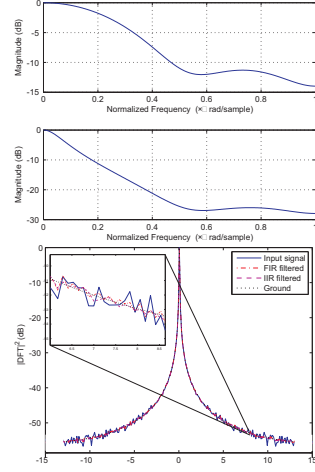


Figure 3: *Left*: Frequency response of the (top) FIR and (bottom) IIR filter. *Right*: The DFT of the input, the outputs and the ground signal (video rate: 25fps).

$b_0 = 0.4$, $b_1 = 0.3$, $b_2 = 0.2$, $b_3 = 0.1$ and $a_1 = 4$, $a_2 = -2$, $a_3 = -1$. In either case, these values of coefficients establish a lowpass filter. IIR provides smoother results because of its higher, theoretically infinite, order. On the other hand, FIR deals better with peaks (outliers) due to its finite order. The frequency response of the filters and the DFT¹ magnitude of the output signals are shown in Figure 3. The input is the time mapping sequence resulted when the proposed method applies to a real nighttime sequence. The ground smooth signal obtained by postprocessing (curve fitting given all input values) is given for comparison. Although both filters behave similarly in low frequencies, IIR output is more close to the ground signal in high frequency band.

2.2 SPATIAL ALIGNMENT

In order to obtain accurate alignment between a reference frame $I^r(\hat{\mathbf{x}})$ and the current observed frame $I^c(\mathbf{x})$, we propose the use of a recently introduced algorithm that is called ECC algorithm (Evangelidis and Psarakis, 2008). This scheme seems to be robust in noise while at the same time is insensitive to global illumination changes. The algorithm uses an enhanced version of the correlation coefficient as an objective function and the goal is its maximization through an iterative scheme.

Let us suppose that the warp $W(\mathbf{x}; \mathbf{p})$ is a 2D mapping based on the standard homography model with eight parameters (Szeliski, 2010), i.e. $\hat{\mathbf{x}} = W(\mathbf{x}; \mathbf{p})$, that provides dense correspondences. Then, ECC algorithm tries to estimate the warp so that the observed

¹Discrete Fourier Transform (Lathi, 1998)

and the warped reference images are similar. In other words, it solves the following maximization problem

$$\max_{\mathbf{p}} \frac{\mathbf{i}_c^t \mathbf{i}_r(\mathbf{p})}{\|\mathbf{i}_c\| \|\mathbf{i}_r(\mathbf{p})\|} . \quad (2)$$

where $\mathbf{i}_c$ and $\mathbf{i}_r(\mathbf{p})$ are the zero-mean vectorized forms of images $I^c(x)$ and $I^r(W(\mathbf{x}; \mathbf{p}))$ respectively. Since the above maximization problem is highly non-linear, the solution of a sequence of secondary problems that follow a closed form solution is proposed in (Evangelidis and Psarakis, 2008). By considering the update rule $\mathbf{p} = \mathbf{p}_0 + \Delta \mathbf{p}$, the vector $\mathbf{i}_r(\mathbf{p})$ can be approximated by $\mathbf{i}_r(\mathbf{p}) \simeq \mathbf{i}_r(\mathbf{p}_0) + J\Delta \mathbf{p}$ using the first-order Taylor expansion formula, where J is the Jacobian of $\mathbf{i}_r(\mathbf{p})$ with respect to $\mathbf{p}$ evaluated at $\mathbf{p}_0$ (see (Evangelidis and Psarakis, 2008) for details). Although after linearization the objective function remains non-linear in $\Delta \mathbf{p}$, it has been proved that the optimum correction vector obeys the following closed form solution

$$\Delta \mathbf{p} = (J^t J)^{-1} J^t \{ \lambda \bar{\mathbf{i}}_o - \bar{\mathbf{i}}_r(\tilde{\mathbf{p}}) \} , \quad (3)$$

with λ being

$$\lambda = \frac{\bar{\mathbf{i}}_p^t P_J \bar{\mathbf{i}}_p}{\bar{\mathbf{i}}_p^t P_J \bar{\mathbf{i}}_p} , \quad (4)$$

where $P_J = I - J(J^t J)^{-1} J^t$ is an orthogonal projection operator and I the identity matrix. Finally, by iteratively following the above parameter update rule, we can obtain an acceptable solution by setting a stopping criterion or fixing the number of iterations. Note that the complexity of this scheme is $O(N_p N_i^2)$ per iteration, where N_p is the number of parameters and N_p is the number of pixels

2.3 CHANGE DETECTION

Although we present a video surveillance algorithm, we do not focus on change detection since this subject has been extensively studied. A nice survey for image detection algorithms can be found in (Radke et al., 2005). Hence, since FLAR provides the corresponding background frame appropriately warped, we use typical methods to detect changes between current and warped reference images. Specifically, we use the Simple Differencing (SD) method by thresholding the image differences, a Minimum Description Length (MDL) model to classify changed and unchanged regions and a statistical method that considers a Gaussian model for the noise (GN) (for details see (Radke et al., 2005)). All these methods return a binary image (mask) at which we apply trivial morphological operations in order to locate bounding boxes in the image of interest.

3 EXPERIMENTAL RESULTS

In this section, we present qualitative and quantitative results to validate the proposed approach. Specifically, we compare the performance of different counterparts of the proposed algorithm with that of the most related works (Diego et al., 2010; Liu et al., 2008; Yang et al., 2007). The evaluation counts on experimenting with six real video sequence pairs recorded by in-vehicle cameras, whose trajectories are approximately coincident. Although we aim at registering nighttime sequences, we consider essential to also test the algorithms with daylight sequences. To this end, we used three sequences of each class denoted as *Night1*, *Night2* and *Night3* (Kong et al., 2010), and as *Day1*, *Day2* and *Day3* (Serrat et al., 2007) respectively. The alignment of these sequences implies a quite challenging task, since the speed of vehicles varies. The average length of night sequences is 2500 frames and the spatial resolution is 720×540 pixels, whereas daylight sequences are shorter in both space and time (200 frames of size 512×384 pixels).

3.1 Synchronization Evaluation

In this section, we evaluate the performance of temporally localize each newly acquired frame during the current ride against the background sequence of the previous ride. To properly assess the quality of the results, we have manually annotated the ground-truth for these datasets, i.e. a narrow reference intervals $[l_n, u_n]$ that each current frame must correspond to, since it is hard to decide the precise corresponding frame visually; the length of these intervals is 3 frames on average. Similar to (Diego et al., 2010), the synchronization error for a candidate pair (n, t_n) is defined as

$$\text{err}(t_n) = \begin{cases} 0 & \text{if } l_n \leq t_n \leq u_n \\ \min(|l_n - t_n|, |u_n - t_n|) & \text{otherwise} \end{cases} \quad (5)$$

The performance of synchronization is quantified through the percentage $1 - \sum_i^N (\text{err}(t_n) > \epsilon) / N$ for $\epsilon = 0, 1$.

In the context of synchronization, we could obviously obtain the temporal matches by exhaustively trying to spatially register each possible pair of frames (one from each sequence) that is clearly not feasible due to its computational cost for background sequences longer than a few seconds at a frame rate of 25–30 frames per second. Thus, the temporal matches could be efficiently obtained by retrieving a short-list (i.e. top-10) of background frames by the image-appearance model proposed in (Diego et al., 2010).

Then, a spatial coherence step using ECC algorithm re-ranks the list w.r.t. the correlation coefficient, thus emerging the closest frame. In the context of retrieval, we also try our scheme by only changing the SURF descriptor with the SIFT one (Lowe, 2004).

Table. 1 shows the synchronization performance achieved by these three methods. We provide results for $\epsilon = 0$ and $\epsilon = 1$ to show the error variance. From the results, we can see that SURF-based method achieves higher synchronization scores than the two other methods across all sequences. It is important to note that the Frame Localization (FL) based on SURF or SIFT descriptors accurately discriminates the background frame by just retrieving the best neighbor. IIR filter provides slightly better scores with both descriptors. However, the contribution of SURF descriptor instead of SIFT is clearly evident especially for night-time sequences. Specifically, SURF-FL (SURF-based FL) remarkably increases the performance of SIFT-FL (SIFT-based FL) by 6% on average while the proposed scheme achieves a 8% better score than that of the exhaustive method. Note that this comparison does not favor our method in the sense that our method does not count on geometric constraints, since we aim at investigating the performance of the net algorithm. However, it is obvious that SURF-FL scheme would be benefitted by such constraints.

By putting aside the exhaustive search method, SURF-FL is more efficient than its SIFT counterpart, since SURF algorithm uses integral images to extract the descriptors. Specifically, given a frame of the nighttime sequences, the extraction of SURF descriptors requires 100 msec, while SIFT algorithm takes more than 1 sec. Moreover, the computational cost of the retrieval part is less for SURF-FL scheme because of the shorter descriptor (64 against 128). In practice, the retrieval time using SURF-FL was 0.88 sec, while *SIFT* required 2.8 sec per frame. The exhaustive method leads to a heavy task since FL takes more than 3 minutes per query. It is very important to note that the whole system is implemented in Matlab (not optimized code), but the the original executable files for SURF and SIFT algorithms are used for extracting descriptors.

3.2 ALIGNMENT AND DETECTION ASSESSMENT

Due to real nature of data, we run a qualitative comparison between the investigated schemes. Since the detection of changes relies on the quality of registration, we present both alignment and detection instances to a more complete comparison. To assess

the alignment, we use a fusion representation, where the *G* channel of the current frame has been replaced by the warped *G* channel of the background corresponding frame in an *RGB* representation. This way, changes are marked by green and pink colors.

In Fig. 4 the corresponding frames obtained by the synchronization methods are shown for various night frames including challenging cases (please view the figure in your screen for details). Given the results of the proposed method (Fig. 4 (bottom)), Fig. 5 presents alignment instances obtained by the SIFT-flow algorithm, the Generalized Dual-Bootstrap version of the ICP algorithm (Yang et al., 2007) (GDB-ICP) and the ECC scheme. Note that the goal of SIFT-flow is a pixel-wise alignment instead of estimating a global geometric transformation as ECC and GDB-ICP do. Note also that GDB-ICP deals well with challenging cases like these in question.

All algorithms behave quite well in the absence of occlusions. As we can see, however, when the scene contains objects visible only in the one sequence, SIFT-flow fails as it creates artifacts or disappears objects. This is probably because it works in a flow (local) basis. On the other hand, ECC and GDB-ICP achieve remarkable results despite the noise and the low information content, with GDB-ICP providing local misalignments in some of two of the depicted frames. Since we are interested in efficiently registering the images, the execution time is of major importance. All algorithms work on half-size images. SIFT-flow requires 29.2 sec/frame, thus setting itself inappropriate for real-time applications. As far as the GDB-ICP scheme is concerned, we must note that the time execution varies with the difficulty level of registration. In our experiments, its average registration time was 42.2 sec/frame, while ECC required 1.7 sec/frame (15 iterations were enough). The ECC algorithm was implemented with OpenCV library, whereas for the others their original executable files were used. As a consequence, given the total average time required for the proposed scheme (FLAR+ECC), we can envision a real-time execution in specific environments (i.e. GPU-based). Notice that the time of ECC can be easily reduced by initializing the homography using feature matching or using only pixels around keypoints (now all image pixels are taken into account).

Although our primary goal is the spatio-temporal matching, we also present results of some change detection methods. As we mentioned in Sec. 2.3, we enabled the SD method as well as the more sophisticated MDL and GN models. Detection results are shown in Fig. 6. Instead of presenting binary masks, we use bounding boxes superimposed in query frames

		Synchronization scores ($\epsilon = 0 \setminus \epsilon = 1$)						
		<i>Night1</i>	<i>Night2</i>	<i>Night3</i>	<i>Day1</i>	<i>Day2</i>	<i>Day3</i>	Average
Exhaustive search		71.5\84.5	61.4\78.8	68.9\83.8	—	93.2\98.6	85.0\99.3	76.0\89.0
SIFT-FL	FIR	67.5\82.9	48.7\68.6	66.9\83.1	70.0\85.0	99.3\100	92.5\96.6	74.2\86.0
	IIR	71.8\86.7	52.2\70.7	77.1\88.8	74.0\93.5	100\100	95.2\100	78.4\90.0
SURF-FL	FIR	72.6\86.3	53.4\71	73.6\87.4	74.0\90.5	99.3\100	88.4\95.2	76.9\88.4
	IIR	78.8\90.6	60.6\76.6	82.6\92.8	96.5\99.5	100\100	100\100	86.4\93.3

Table 1: Synchronization scores (%) obtained by the proposed methods and the competitors for two values of error tolerance ϵ . Symbol “—” means that the exhaustive method totally fails for *Day1* due to repeated patterns in frames.

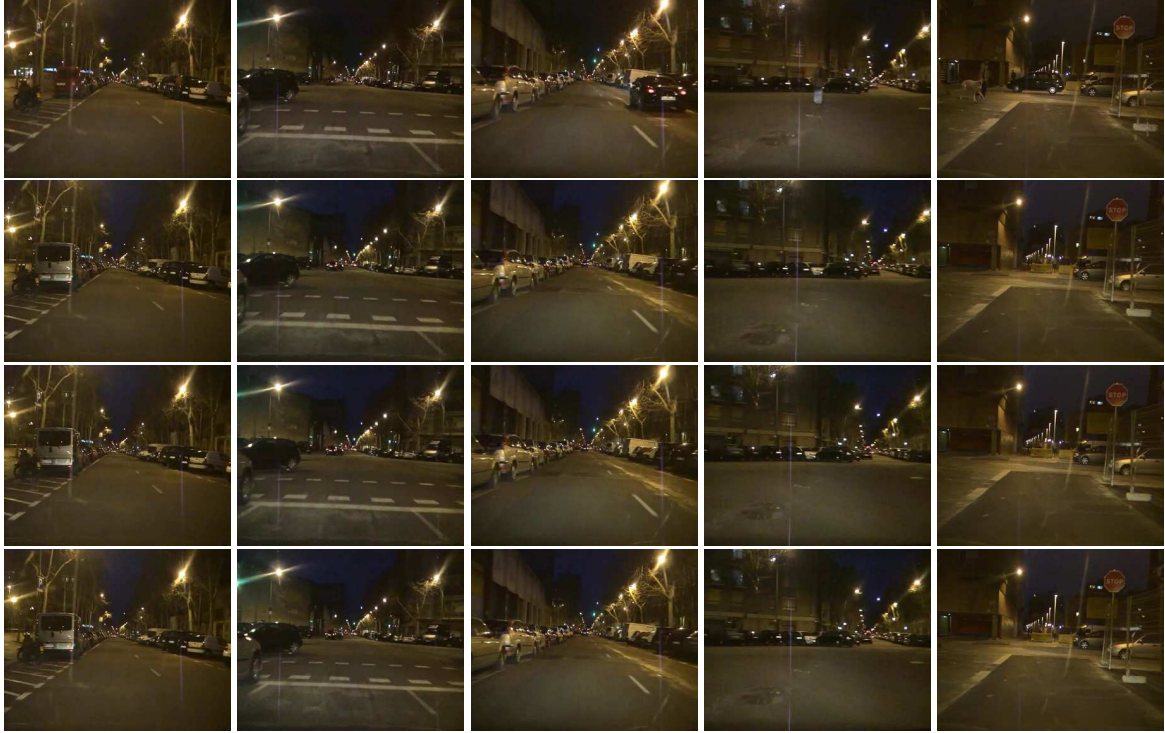


Figure 4: (First row) Query (current) frames of *Night1* sequence and synchronization results obtained by (second row) exhaustive search, (third row) SIFT-based retrieval and (fourth row) SURF-based retrieval.

to annotate detected changes. An “empty” bounding box means that something is missing compared to the background frame (see also the bottom row of Fig. 4). Otherwise, it may be due to local misalignment, different illumination and reflectance, shading etc. We observed that GN method provides slightly better result than MDL and SD methods. We must point out here that, normally, errors in alignment and detection do not happen in successive frames but randomly (see supplemental material). This is helpful for the video analyst who can ignore instant changes. The time required by SD method is meaningless. The complexity of the MDL and GN method is slightly higher, but not prohibitive for real-time applications.

4 CONCLUSIONS AND FUTURE WORK

We presented a novel framework for helping a video analyst to robustly detect changes in night-time outdoor surveillance by mobile cameras. In order to avoid exhaustive cross-frame search of finding background frames, a Frame Localization And Registration (FLAR) is proposed to solve the problem efficiently. The frame localization builds upon retrieving the most similar background frame based on the SURF descriptor together with a temporal filtering applied to the retrieval results to handle outliers. Then, a recently proposed alignment scheme that overcomes appearance variations between frames acquired at different times is used to register the corresponding frames in space; thus applying a simple change



Figure 5: Alignment instances in negative color for (first row) SIFT-flow, (second row) GDB-ICP and (third row) ECC algorithm based on the frame pairs between the top and bottom row in Fig. 4.

detection to aligned frames allows the detection of suspicious areas. Experiments with real night sequences recorded by in-vehicle cameras demonstrate the performance of the proposed method and verify its efficiency and effectiveness against other methods. Moreover, the ability of the proposed scheme to deal with daylight sequences was experimentally verified.

We recall that the proposed system primarily aims at helping the guarding/surveillance staff. A fully automatic system that recognizes objects could work in conjunction with annotated databases in a class level. In order to perform more efficient monitoring, the proposed algorithm can be extended including super-resolution imaging (Kim et al., 2010) and experimental sampling (Wang et al., 2003). The latter dynamically models the evolving attention exploiting the context and past experience information in order to detect and track moving objects in surveillance videos. We leave such approaches for future work.

Supplemental Material: Since it is difficult to illustrate video processing in a written document, we attach video results of the proposed algorithm.

REFERENCES

- Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Comput. Vis. Image Underst. (CVIU)*, 110.
- Chakravarty, P., Zhang, A. M., Jarvis, R., and Klee-man, L. Anomaly detection and tracking for a patrolling robot. In *ACRA'07: Australasian Conference on Robotics and Automation*.
- Diego, F., Ponsa, D., Serrat, J., and Lopez, A. M. (2010). Video alignment for change detection. *IEEE Trans. on Image Processing*, Preprint(99).
- Evangelidis, G. D. and Psarakis, E. Z. (2008). Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Trans. on PAMI*, 30(10):1858–1865.
- Haberdar, H. (2010). Disparity map refinement for video based scene change detection using a mobile stereo camera platform. In *ICPR'10: International Conference on Pattern Recognition*.
- Kim, M., Ku, B., Chung, D., Shin, H., Kang, B., Han, D. K., and Ko, H. (2010). Robust dynamic super resolution under inaccurate motion estimation. In *Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS '10*, pages 323–328, Washington, DC, USA. IEEE Computer Society.
- Kong, H., Audibert, J.-Y., and Ponce, J. (2010). Detecting abandoned objects with a moving camera. *IEEE Transactions on Image Processing (TIP)*, 19(8):2201–2210.
- Lathi, P. (1998). *Signal Processing and Linear Systems*. Berkeley Cambridge Press.
- Liu, C., Yuen, J., Torralba, A., and Freeman, W. T. (2008). Sift flow: dense correspondence across different scenes. In *Proc. of ECCV*.
- Lowe, D. (2004). Distinctive image features from scale invariant keypoints. *IJCV*, 60(2):91–110.
- Marcenaro, L., Marchesotti, L., and Regazzoni, C. (2002). "a multi-resolution outdoor dual camera system for robust video-event metadata extraction". In *Proceedings of the Fifth International*

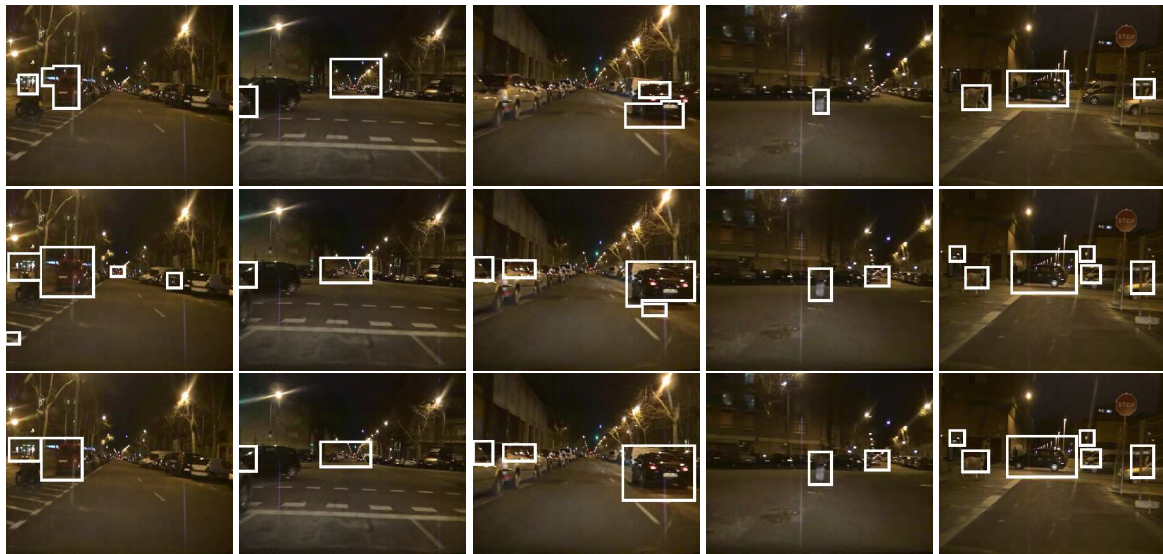


Figure 6: Change detection results using (top) SD, (middle) MDL and (bottom) GN method.

Conference on Information Fusion, 2002., volume 2, pages 1184 – 1189 vol.2.

Primdahl, K., Katz, I., Feinstein, O., Mok, Y. L., Dahlkamp, H., Stavens, D., Montemerlo, M., and Thrun, S. (2005). Change detection from multiple camera images extended to non-stationary cameras. In *In Proceedings of Field and Service Robotics*.

Radke, R. J., Andra, S., Al-Kofahi, O., and Roysam, B. (2005). Image change detection algorithms: A systematic survey. *IEEE Transactions on Image Processing (TIP)*, 14:294–307.

Sand, P. and Teller, S. (2004). Video matching. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 22(3):592–599.

Serrat, J., Diego, F., Lumbreras, F., and Álvarez, J. (2007). Alignment of videos recorded from moving vehicles. *ICIAP'07: 14 th International Conference on Image Analysis and Processing*.

Sivic, J. and Zisserman, A. (2009). Efficient visual search of videos cast as text retrieval. *IEEE Trans. on PAMI*, 31(4):591–606.

Soibam, B., Shah, S. K., Chaudhry, A., and Eledath, J. (2009). Quantitative comparison of metrics for change detection for video patrolling. In *ICCVW'09: IEEE International Workshop on Video-Oriented Object and Event Classification*.

Szeliski, R. (2010). *Computer Vision: Algorithms and Applications*. Springer.

Wang, J., Kankanhalli, M. S., Yan, W., and Jain, R. (2003). Experiential sampling for video surveillance. In *First ACM SIGMM international work-*

shop on Video surveillance, IWVS '03, pages 77–86, New York, NY, USA. ACM.

Yang, G., Stewart, C., Sofka, M., and Tsai, C.-L. (2007). Registration of challenging image pairs: Initialization, estimation, and decision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(11):1973–1989.