Graffiti Detection Using Two Views

Luigi Di Stefano luigi.distefano@unibo.it
Federico Tombari federico.tombari@unibo.it
Alessandro Lanza alanza@arces.unibo.it
Stefano Mattoccia stefano.mattoccia@unibo.it
Stefano Monti stefano.monti3@studio.unibo.it

Department of Electronics Computer Science and Systems (DEIS), University of Bologna
Viale Risorgimento 2, 40136 - Bologna, Italy
Advanced Research Center on Electronic Systems (ARCES), University of Bologna
Via Toffano 2/2, 40135 - Bologna, Italy
www.vision.deis.unibo.it

Abstract

This paper presents a novel video surveillance approach designed to detect vandal acts occurring on the background of the monitored scene, such as graffiti painting on walls and surfaces, public and private property defacing or etching, unauthorized post sticking. The aim of our approach is to detect this class of events rapidly and robustly. We propose to use two synchronized views to deploy synergically depth and intensity information concerning the monitored scene. Our system can work within unstructured environments and with geometrically unconstrained backgrounds.

1. Introduction

Nowadays vandal acts represent a serious problem in urban areas, with thousands of public and private properties being damaged daily all around the world. Costs issued by this problem are huge: e.g. for the problem of graffiti, that relates to the wide range of markings, etchings and paintings that deface public and private properties, an estimate $ 12 billion a year is spent for cleaning and prevention in the United States [1]. Beside the expenses related to repairing, cleaning and/or substituting a vandalized property, indirect costs arise due to the perceived insecurity associated with the occurrence of vandal acts in a certain area. This typically results in a decrease of revenues for commercial activities or services taking place in the area, such as shops, house tenures, and public transport, for which the uncleanness and perceived insecurity lower passenger confidence in the transport system and consequently tend to decrease ridership. Not less important are the social consequences that repetitive vandal acts in a certain urban area imply on the dwellers.

The effort pushed to tackle - or at least control - the diffusion of vandal acts in urban areas worldwide has often resorted to the use of automatic monitoring systems due to the huge amount of public and properties in cities. Commercial products based on audio sensors [11, 3, 2, 15] try to detect graffiti by analyzing the sounds typically occurring during these actions. These devices present notable limitations, since they detect specific sounds and can hardly generalize to different or noiseless vandal acts. Moreover they can be easily tricked by the presence of environmental noise, and they typically need to stand very close to the monitored region.

Due to these reasons, vision-based approaches relying on automatic video analysis have been recently driving increasing attention. On one hand, all the proposed state-of-the-art vision-based systems [14, 1, 5] rely on a single-view approach, that is they try to recognize vandal acts by processing a video sequence obtained from a single camera. On the other hand, two different classes of algorithms can be outlined.

One approach consists in applying behaviour and gesture analysis techniques for recognizing the high-level spatio-temporal pattern corresponding to the person perpetrating the vandal act. For example, detection of graffiti is carried out in [14] by searching for the pattern corresponding to a person writing on a monitored surface. However, such techniques require accurate training of classifiers and generally perform much better when a certain degree of cooperation from the subject can be achieved, which is obviously not the case of vandal acts.

Another approach relies on comparing the current appearance of the monitored object with that of a background model of the scene. Hence, this class of methods can detect only vandal acts which produce visible and stationary changes of the appearance of the monitored scene. We will refer to this class of events as *Stationary Visible Changes* (SVC), which includes paintings on walls and surfaces, public and private property defacing, etching or stealing, unauthorized post sticking. This also concerns other scenarios such as, e.g. for cultural heritage environments or museums, criminal acts such as tearing, dirtying, defacing, stealing of parts of an artwork. Detection of vandal acts by recognizing SVC is carried out in [1, 5].

In particular, in [1] the authors focus on graffiti and propose a low-level approach for SVC recognition based on single-view change detection. This approach inherently suffers from a false positives problem when deployed as vandal acts detectors. In fact, SVC events include most of the common aforementioned vandal acts, but also other frequent events such as people standing still, parked vehicles (such as cars, motorbikes, bicycles), abandoned objects. The same issue arises with method [5], which concerns a fast single-view SVC detector based on the analysis of higher-level events occurring in the monitored scene. This problem is partially dealt with by limiting the detection only on a subset of the camera field-of-view and by assuming that the monitored scene is not crowded. Finally, robustness with regards to sudden illumination changes occurring in the scene is not investigated.

Our idea is to go beyond the visibility and stationarity cues in order to obtain a finer classification of SVC. This should allow for a more effective detection of specific vandal acts based on the recognition of their peculiar effects. To this purpose, we propose to deploy also depth information, so that the class of SVC events can be partitioned into the two following mutually-exclusive sub-classes:

a) *Stationary Appearance Changes* (SAC): stationary visible changes due to variations of the appearance but not the 3D geometry of the scene.
b) *Stationary Geometric Changes* (SGC): stationary visible changes due to variations of the 3D geometry of the scene;

It is clear that most of the vandal acts that previous proposals try to detect as SVC are indeed SAC, for they determine no (or small) variation of the scene 3D geometry, while most false positives have to be ascribed to SGC.

Hence, we propose to effectively detect vandal acts based on the ability of distinguishing between SAC and SGC. In particular, in this paper we present a real-time SAC detection algorithm based on the use of two synchronized views of the monitored scene and on a novel multi-view change detection approach. This deploys on-line intensity information coming from the two image sensors together with knowledge of the 3D structure of the monitored scene, which is obtained once at initialization time by means of a stereo matching process. This enables to detect effectively SAC events even in presence of static subjects that produce SGC. The proposed method can detect graffiti-like events within unstructured environments and does not pose any constraint on the appearance and geometry of the background of the monitored scene. Moreover, by means of a specific stage which is robust with respect to non-linear photometric distortions, our approach can also handle strong sudden illumination changes and shadows. Finally, our system can alert the occurrence of vandal acts while they are being committed.

The paper is organized as follows. Section 2 outlines some basic principles of multi-view change detection, contextually reviewing the state-of-the-art in the field. The proposed vandal acts detection algorithm is presented in Section 3 by describing first the novel multi-view change detector devised for discriminating Appearance Changes (AC) from Geometric Changes (GC), then the simple procedure used to evaluate stationarity of AC. Section 4 presents some experimental results obtained on real scenes under challenging conditions. Finally, Section 5 draws conclusions.

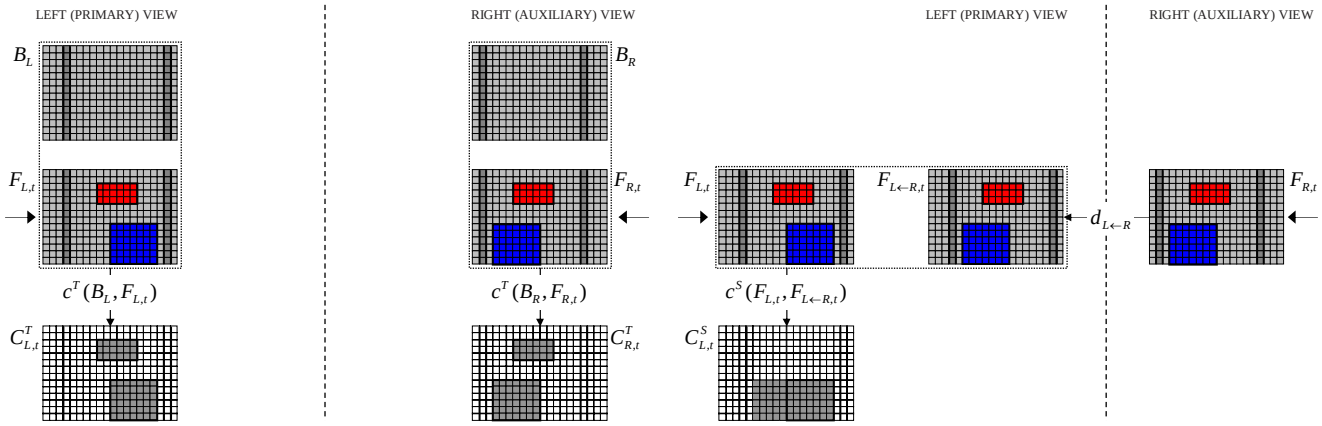
2 Principles of multi-view change detection

The input information to our approach is represented by two synchronized video sequences of a scene characterized by a considerable overlap of field-of-views. Moreover, we assume stationarity of the capturing devices as well as of the scene background geometry, so that geometric registration of the background over the two views, hereinafter denoted as left view (L) and right view (R), can be computed only once at initialization time. Apart from stationarity, no further assumption is made about geometry of the background surface which, in particular, is not constrained to be planar.

The goal of our approach is to compute in one of the two views, referred to as *primary*, a binary mask highlighting the pixels which are sensing a SAC, that is, at the event level, graffiti. To this purpose, we use a novel multi-view change detector to carry out the twofold task of robustly detecting VC and, among these, discriminating between AC and GC. Then, a simple procedure is used to evaluate stationarity of AC. To better illustrate our proposal, in this section we outline some basic principles concerning multi-view change detection and, contextually, review the state-of-the-art in the field.

As regards the way the input information (i.e. the two synchronized video sequences) can be exploited for detecting changes with respect to a reference scene, we define:

a) *temporal consistency constraint*: for a given view-point, the processed frames are images of the



(a) temporal change detection: only the VC super-class can be recognized

(b) spatial change detection: only the GC sub-class can be recognized

Figure 1. Temporal and spatial change detection

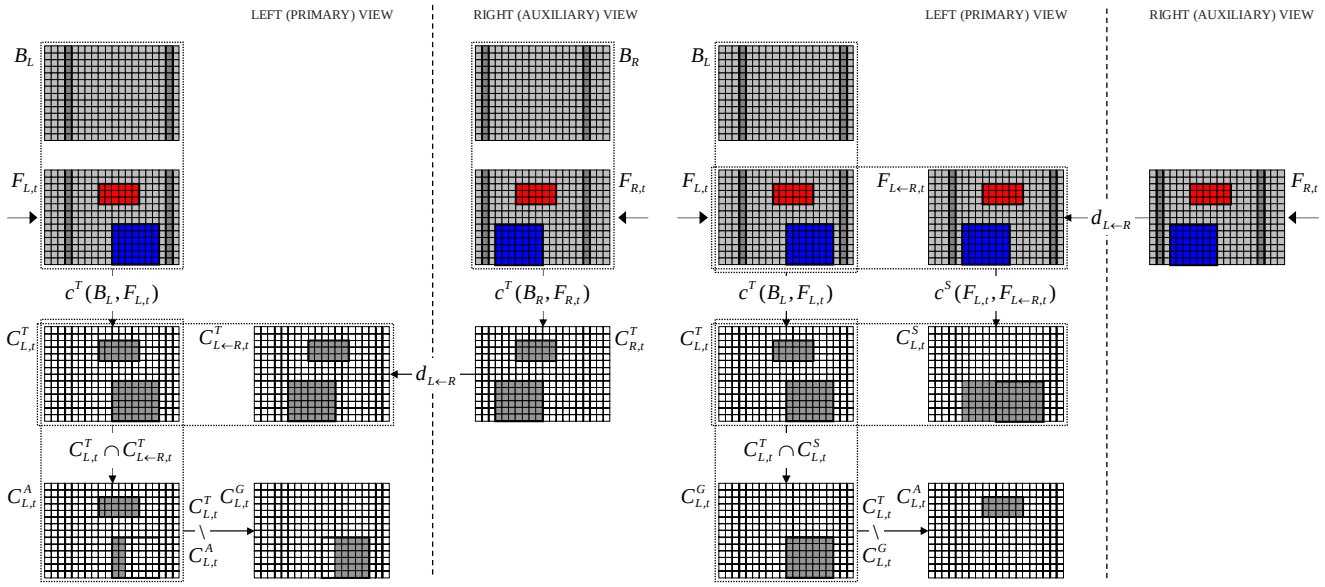
same scene taken at different times;

b) *spatial coherence constraint*: for a given elaboration time instant, the processed frames are images of the same scene taken from different view-points;

The temporal consistency constraint can be exploited to perform a *temporal change detection* independently in each view by a classical background subtraction procedure. That is, at each time t the current frames $F_{L,t}$ and $F_{R,t}$ are compared by a suitable operator $c^T(\cdot,\cdot)$ with as many off-line generated view-dependent appearance models B_L and B_R of the reference scene, that we call *temporal backgrounds*. Two *temporal change masks* are thus obtained, that is two binary masks $C^T_{L,t}$ and $C^T_{R,t}$ comprising the pixels which are currently sensing a violation of the temporal consistency constraint. Temporal change detection is illustrated in Figure 1(a) by means of a toy example consisting of a planar background (light grey with two darker vertical strips), a parallel-axis stereo sensor with the two optical axes perpendicular to the background, and AC (red) as well as GC (blue) events being sensed in the current frames. As one can easily understand and as pointed out in Figure 1(a), generally speaking temporal change detection allows for detecting, independently in each view, the super-class of VC events but not for discriminating between the AC and GC sub-classes. This is due to the fact that recovering depth information from a single view is in principle an ill-posed problem.

Exploitation of the spatial coherence constraint yields the simplest multi-view change detection approach, proposed in [7], that we call *spatial change detection*. The *spatial background*, unlike temporal ones, does not store appearance but geometric information about the monitored scene. In fact, it consists in the disparity map $D_{L\leftarrow R}$ (computed off-line) warping the monitored scene from the auxiliary to the primary view. Spatial background subtraction is thus performed by a background disparity verification. That is, at each time t the auxiliary frame $F_{R,t}$ is warped into the primary view by the background disparity map and then compared by a suitable operator $c^S(\cdot,\cdot)$ with the primary frame $F_{L,t}$. This allows to obtain a *spatial change mask*, that is a binary mask $C^S_{L,t}$ highlighting the pixels which are currently sensing a violation of the spatial coherence constraint. As illustrated in Figure 1(b), only the GC sub-class can be recognized by spatial change detection. In fact, AC events occur on the background surface and, hence, are coherent with respect to the background disparity map. Moreover, the method suffers from an intrinsic false positives problem, called *occlusion shadows*. In fact, the background pixels in the primary view which are occluded by a foreground object in the auxiliary view are inherently detected as changed. To deal with this problem, in [7] the authors propose to exploit more than one auxiliary view and to compute the intersection of the binary masks obtained by comparing the primary with each of the auxiliary views. In [12] the problem is addressed from a sensor planning perspective. In particular, it is shown how occlusion shadows can be removed by using just two views if a suitable sensors configuration is adopted.

The combined exploitation of both the temporal consistency and the spatial coherence constraints is proposed in [8] and [10]. Essentially, both approaches rely on the idea, illustrated in Figure 2(a), of first performing temporal change detection in each view and then carrying out spatial change detection based on the obtained temporal change masks. This should allow to obtain the AC mask $C^A_{L,t}$ and, by subtraction from the temporal change mask, the GC mask $C^G_{L,t}$. However, as pointed out in Figure 2(a) and discussed in detail in Section 3.2, these methods inherently suffer from missed detections in the GC mask and, dually, from false detections in the AC mask.



(a) joint use of temporal and spatial change detection as proposed in [8,10]

(b) joint use of temporal and spatial change detection as proposed in our approach

Figure 2. Joint exploitation of temporal and spatial change detection

3 The proposed algorithm

The proposed graffiti detection algorithm relies on a novel multi-view change detection approach. The novelty consists in a simple yet clever way of combining temporal and spatial change detection so as to perform an effective discrimination between AC and GC. To better illustrate the approach we will distinguish between off-line and on-line elaboration steps. Once the AC events are detected, the proposed procedure for the recognition of stationary AC (SAC) can be regarded as a post-processing step. Hence, it will be described in a separate section together with a simple binary morphology stage applied on the final SAC mask.

3.1 Off-line elaboration

The very first step concerns the calibration of the stereo sensor, which aims at estimating, for each view, the calibration parameters, a set of optical distortion parameters and a rectification homography. This information is condensed into two geometrical transformations $g_L(\cdot)$ and $g_R(\cdot)$ that will be used at each processing time - both off-line and on-line - to compute the undistorted and rectified versions $F_{L,t}$ and $F_{R,t}$ of the captured frames $F^c_{L,t}$ and $F^c_{R,t}$, respectively. In formulas:

$$F_{L,t} = g_L\left(F^c_{L,t}\right) \qquad F_{R,t} = g_R\left(F^c_{R,t}\right) \tag{1}$$

Hence, for each view a short bootstrap sequence of N frames (N in the order of tens) is used to infer an appearance model of the reference scene, i.e. the temporal background:

$$B_L = b\left(F_{L,1}, \ldots, F_{L,N}\right) \qquad B_R = b\left(F_{R,1}, \ldots, F_{R,N}\right) \tag{2}$$

with $b(\cdot)$ denoting a generic, possibly robust, pixel-wise statistical estimator. In the experiments shown in Section 4 we have used the median operator. The two temporal backgrounds are thus fed to a dense stereo matching algorithm so as to compute the disparity map warping the reference scene from the auxiliary (right) to the primary (left) view, i.e. the spatial background:

$$D_{L \leftarrow R} = m\left(B_L, B_R\right) \tag{3}$$

It is worth pointing out here that this operation aimed at obtaining the spatial background needs to be obtained once and for all at initialization time, hence on-line stereo matching is not required by our method. Therefore, with our approach one should deploy an as accurate as possible, even though slow, stereo matching algorithm, so as to maximize the accuracy of the warping function. In the experiments shown in Section 4 we have used the algorithm described in [6].

3.2 On-line elaboration

The main on-line processing steps performed by the proposed algorithm are illustrated in Figure 2(b) by means of the same toy example used in the previous section.

First of all, temporal change detection is performed in the primary view by background subtraction, that is by comparing the current frame $F_{L,t}$ with the off-line generated temporal background B_L, so as to compute the temporal change mask $C^T_{L,t}$:

$$C^T_{L,t} \;=\; c^T\left(B_L, F_{L,t}\right) \tag{4}$$

In particular, to achieve robustness with respect to strong photometric distortions we apply at pixel-level the block-level approach presented in [9]. This algorithm is able to filter-out illumination changes yielding locally order-preserving transformations of pixel intensities.

Spatial change detection is then performed. To this purpose, first of all the auxiliary frame $F_{R,t}$ is warped into the primary view:

$$F_{L\leftarrow R,t} \;=\; d_{L\leftarrow R}\left(F_{R,t}\right) \tag{5}$$

with $d_{L\leftarrow R}(\cdot)$ denoting the operation of warping pixel by pixel the auxiliary frame according to the background disparity map $D_{L\leftarrow R}$. The spatial change mask $C^S_{L,t}$ is then obtained by comparing the primary frame $F_{L,t}$ with the warped auxiliary frame $F_{L\leftarrow R,t}$ according to the operator $c^S(\cdot,\cdot)$:

$$C^S_{L,t} \;=\; c^S\left(F_{L,t}, F_{L\leftarrow R,t}\right) \tag{6}$$

Differently from temporal change detection, here the compared frames are synchronized. Hence, under the assumption of Lambertian surfaces, illumination changes occurring in the monitored scene affect in the same way the amount of radiation incident onto the two sensors. Nevertheless, in general the two sensors can produce different measures (i.e. image intensities) due to the presence of non-Lambertian surfaces, to a different foreshortening of the objects in the two views and different camera parameters (e.g. gain, exposure).

For this reason, also in this case a robust change detection algorithm is desirable. We propose to use a block-based approach and the well-known Normalized Cross-Correlation (NCC) measure, due to its simplicity and its constant complexity. This measure is invariant to linear photometric distortions. It is also worth to point out that the computation of $C^S_{L,t}$ by means of the NCC measure can be efficiently performed using incremental schemes [13, 4], so that complexity turns out independent on block size.

As discussed in the previous section and clearly outlined in Figure 2(b), on one hand the temporal change mask comprises the super-class of pixels sensing a VC, while on the other hand the spatial change mask contains the sub-class of GC pixels and the false positives corresponding to occlusion shadows. Hence, by computing the intersection of the two masks the geometric change mask $C^G_{L,t}$ containing GC pixels can be easily obtained:

$$C^G_{L,t} \;=\; C^T_{L,t} \cap C^S_{L,t} \tag{7}$$

Finally, it is straightforward to compute the appearance change mask $C^A_{L,t}$ by subtracting the geometric from the temporal change mask:

$$C^A_{L,t} \;=\; C^T_{L,t} \setminus C^G_{L,t} \tag{8}$$

Summarizing, in principles this mask should include only those pixels that are currently sensing a change of the background appearance related neither to an illumination change nor to a variation of the background geometry. Given the addressed application domain, such changes can be ascribed to graffiti.

It is worth pointing out that in the method of Figure 2(a) disparity verification is carried out on the binary temporal change masks. As a result, the method will inherently yield false AC in correspondence of the overlapping areas between GC regions found in the primary view and in the warped auxiliary view. Differently, with our approach disparity verification is performed on the original frames, as in 1(b). Hence, for the pixels belonging to the above mentioned overlapping areas a decision is taken based on photometric similarity according to the NCC measure. Since in such areas overlapping between different parts of a foreground object is likely to occur (e.g. the left and right shoulder of a person making graffiti), unless the object is untextured, it is likely that photometric dissimilarity will allow for a correct classification as spatial changes. As a consequence, our method will unlikely yield false AC.

3.3 Stationarity and morphology

Since graffiti yield permanent and static modifications of scene appearance, we can exploit the further constraint that AC detected by the proposed multi-view change detector have to be stationary. To this purpose, we propose to use a simple procedure based on a *post-processing* and *pixel-wise* approach. That is, at each time t a pixel sensing an AC is classified as a SAC if the appearance change is persistent over a given interval of k previous frames. In formulas:

$$C^{SA}_t(\boldsymbol{p}) \;=\; C^A_t(\boldsymbol{p}) \wedge C^A_{t-1}(\boldsymbol{p}) \wedge \ldots \wedge C^A_{t-k}(\boldsymbol{p}) \tag{9}$$

where C^{SA}_t denotes the obtained SAC binary mask and the subscript L is drop for simplicity. Similarly, a persistent absence of AC is required to switch off a SAC pixel in C^{SA}_t.

Finally, to refine the computed SAC binary mask and remove small false positives and false negatives we apply a simple two-steps morphological filtering consisting of an area-opening and a morphological closing. The obtained graffiti blobs are then labelled and their bounding-boxes are extracted.

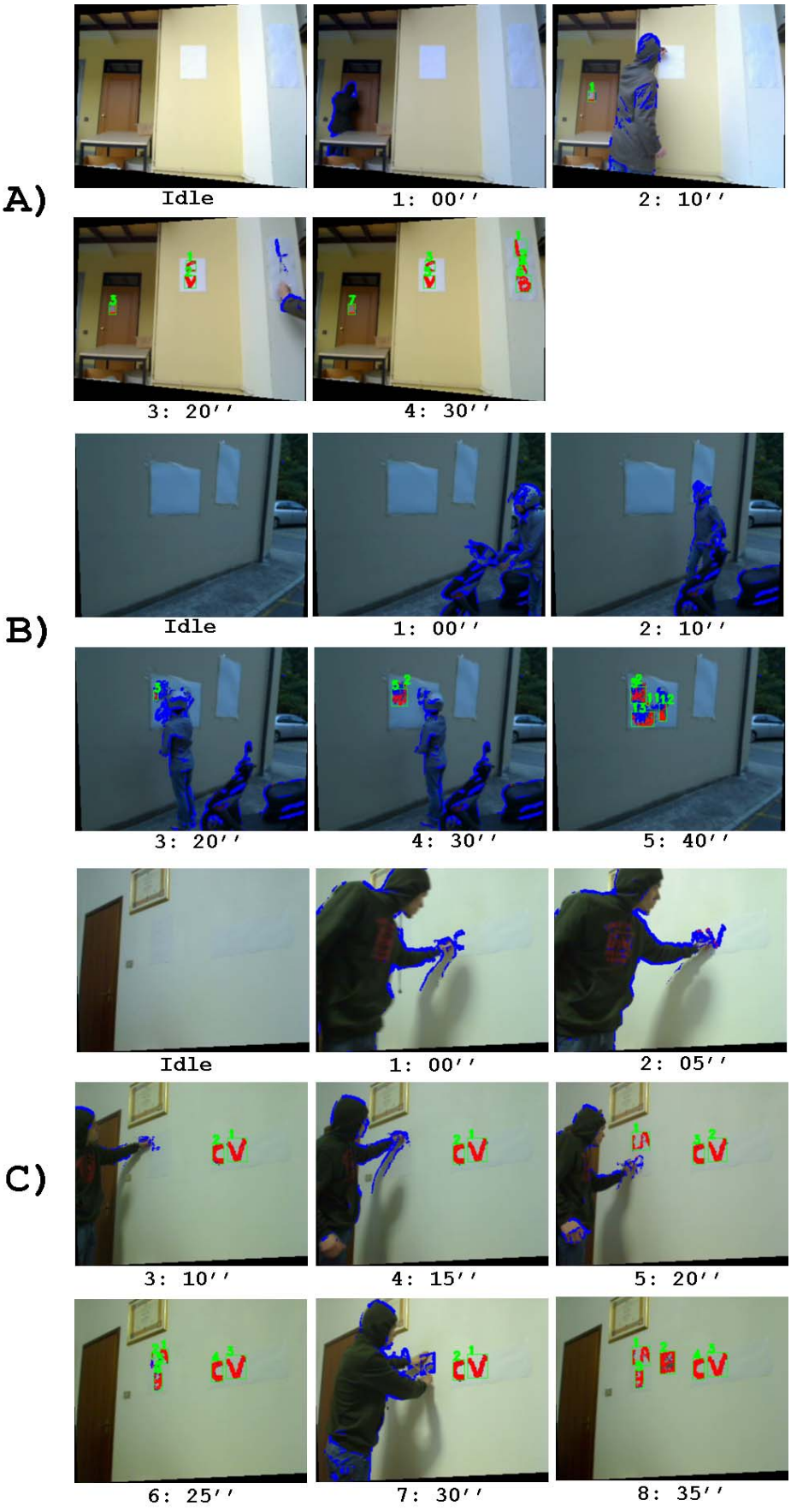


Figure 3. Results dealing with the 3 *Graffiti* sequences (to be viewed in color)

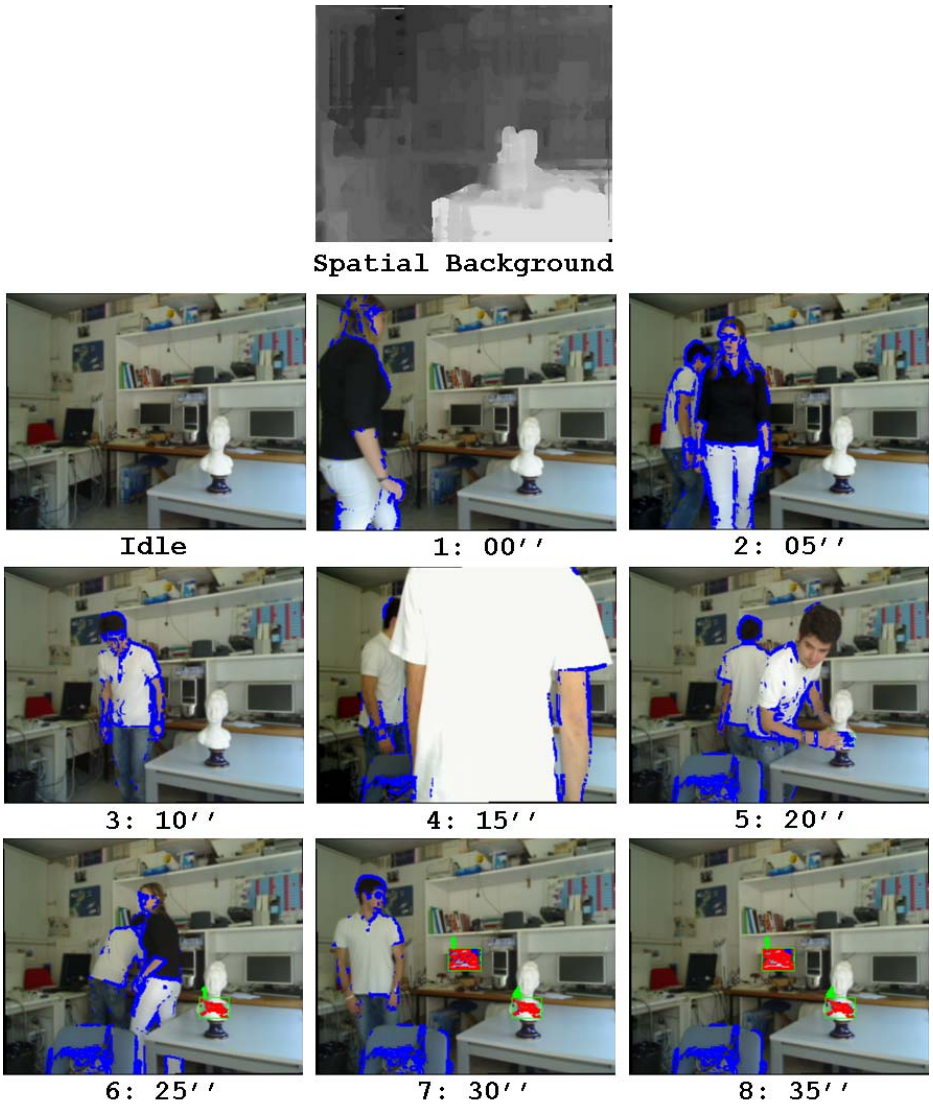


Figure 4. Results dealing with the *Statue* sequence (to be viewed in color)

4 Experimental results

This section presents experimental results aimed at evaluating the capabilities of the proposed approach to detect typical SAC events under real conditions. In particular, we have implemented the proposed algorithm in *C* code using off-the-shelf hardware which includes a PC with an AMD Athlon 2.21 GHz core processor and a very cheap stereo setup represented by two web-cams.

Figure 3 shows the results dealing with three video sequences (*A*, *B* and *C*) concerning graffiti detection. Sequences *A*, *B* refer to an outdoor environment, while sequence *C* refers to an indoor scene. For all sequences, the top left frame shows the idle appearance of the scene (i.e. the background), while remanining frames show the output of the system sampled every 10 or 5 seconds (depending on the dynamics of the event) starting from the beginning of the vandalic action. In particular, the output depicts with blue pixels those points currently detected as GC, while in red those points currently detected as AC. Finally, when a SAC event is detected (i.e. after post-processing) a green bounding box with a numbered label highlights the area where the action is taking place.

In the *Graffiti A* sequence the background is represented by three textureless slanted walls at different depths on which a person posts up a flyer and draws some graffiti, while in the *Graffiti B* sequence the background is mainly composed by a textureless slanted wall. In both sequences it is worth to note that our approach is able to accurately detect the graffiti events at different depths while the action

is occurring. Moreover it is worth pointing out the notable absence of false positives throughout the whole sequences. It is also interesting to note that SGC events are currently discriminated from SAC events (e.g. in *Graffiti B* sequence, the motorbike which is parked in front of the background during the vandal act).

For what concerns the *Graffiti C* sequence, the background is represented by a white slanted wall on which a person draws some graffiti and posts up a flyer. Also in this case, graffiti are correctly and on-line discriminated from GC. Similarly to the previous sequence, false positives are absent along all frames despite the notable presence of shadows on the background. In this case, the adopted robust temporal change detection algorithm allows to reject the majority of shadow points as visible changes (frames 1-5, 7), the remaining ones being discarded by stationarity and morphology (frames 1-5).

Fig. 4 refers to a more general case of vandal acts detection over a complex background. In this case, referred to as *Statue* sequence, the background is constituted by a table and a small statue close to the sensor, plus a variegated group of objects at a further distance. The background models for the two views together with the corresponding disparity maps are shown on the top of the figure. Similarly to the previous cases, frames 1-9 show the output of the system sampled every 5 seconds.

Beside the complex background and the not perfectly synchronized stereo sensor, challenges are also introduced by the events taking place in the scene. That is, different people are moving simultaneously (frames 3, 5-7) even close to the camera (frame 5). Furthermore, a chair is placed in the scene (frames 5-8), this event being correctly not classified as SAC since it represents a SGC. SAC events are represented by defacing of the statue (between frames 6 and 7) and by switching on a monitor (between frames 7 and 8). These events are correctly and accurately detected (frames 7-9). Besides, a person standing still (frame 8) does not produce any false positive since, again, it correspond to a SGC. As for computational requirements, our approach can efficiently process video frames at an average rate of 10 fps.

5 Conclusions

We have presented an original real-time approach for the on-line detection of acts of vandalism, such as in particular graffiti, yielding SAC events in video sequences. Our method relies on two synchronized views and jointly exploits temporal and spatial coherence concerning the monitored scene appearance by means of a novel multi-view change detection algorithm. This allows us to discriminate effectively between events which only change the appearance of the scene, such as graffiti, and those which also affect its geometry. Overall, our method can deal with typically challenging aspects such as crowded scenes, abandoned/removed objects, static intrusions, sudden illumination changes. The experimental results allow us to claim that the proposed algorithm is a robust and accurate solution to detect act of vandalism that yield SAC events in real complex scenes. It is also worth pointing out that our approach can also work with a cheap, common hardware represented by a standard PC and two web cams.

References

[1] D. Angiati, G. Gera, S. Piva, and C. Regazzoni. A novel method for graffiti detection using change detection algorithm. In *Proc. Int. Conf. Advanced Video and Signal-based Surveillance (AVSS'05)*, pages 242–246, 2005.

[2] *http://www.axiumtech.net*. Axium Technologies.

[3] *http://www.broadbanddiscovery.com*. Broadband Discovery Systems Inc.

[4] F. Crow. Summed-area tables for texture mapping. *Computer Graphics*, 18(3):207–212, 1984.

[5] M. Ghazal, C. Vazquez, and A. Amer. Real-time automatic detection of vandalism behavior in video sequences. In *Proc. IEEE Int. Conf. on Systems, Man and Cybernetics (ISIC 2007)*, pages 1056–1060, 2007.

[6] H. Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Proc. Conf. on Computer Vision and Pattern recognition (CVPR 2005)*, volume 2, pages 807–814, 2005.

[7] Y. A. Ivanov, A. F. Bobick, and J. Liu. Fast lighting independent background subtraction. *International Journal of Computer Vision*, 37(2):199–207, June 2000.

[8] S. M. Khan and M. Shah. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *Proc. European Conference on Computer Vision (ECCV'06)*, volume 4, pages 133–146, May 2006.

[9] A. Lanza and L. D. Stefano. Detecting changes in grey level sequences by ML isotonic regression. In *Proc. Int. Conf. Advanced Video and Signal-based Surveillance (AVSS'06)*, pages 1–4, November 2006.

[10] A. Lanza, L. D. Stefano, J. Berclaz, F. Fleuret, and P. Fua. Robust multi-view change detection. In *Proc. British Machine Vision Conference (BMVC'07)*, September 2007.

[11] G. Lerg, A. Devine, D. Roberts, and R. Johnson. Graffiti detection system and method of using the same. US Patent 6600417, July 2003.

[12] S. N. Lim, A. Mittal, L. S. Davis, and N. Paragios. Fast illumination-invariant background subtraction using two-views: Error analysis, sensor placement and applications. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 1071–1078, June 2005.

[13] M. Mc Donnel. Box-filtering techniques. *Computer Graphics and Image Processing*, 17:65–70, 1981.

[14] C. Sacchi, C. Regazzoni, and G. Vernazza. A neural network-based image processing system for detection of vandal acts in unmanned railway environments. In *Proc. Int. Conf. Image Analysis and Processing (ICIAP'01)*, pages 529–534, 2001.

[15] *http://www.traptec.com*. Traptec Inc.