SPATIO-TEMPORAL INTERACTION FOR AERIAL VIDEO CHANGE DETECTION

Nicolas Bourdis, Marraud Denis, Hichem Sahbi

SPATIO-TEMPORAL INTERACTION FOR AERIAL VIDEO CHANGE DETECTION

Nicolas BOURDIS, Denis MARRAUD

EADS France
Image Solutions Department
Suresnes, France

Hichem SAHBI

CNRS, LTCI Telecom ParisTech
Paris, France

ABSTRACT

With the growing capacity of video devices, human operators are nowadays overwhelmed by the huge volumes of data generated in different applications including surveillance. Therefore, automatic video processing techniques are required in order to filter out uninteresting data and to focus the attention of operators. However, reliability is still a challenging problem.

In this paper, we show how spatio-temporal redundancy may be exploited to enhance the accuracy of automatic change detection in aerial videos. More precisely, we present an algorithm based on Belief Propagation in order to improve spatio-temporal consistency between successive change detection results. Experiments demonstrate that our method leads to increased accuracy in change detection.

***Index Terms*—** Change detection, aerial videos, spatio-temporal redundancy, belief propagation.

1. INTRODUCTION

With the growing capacity of video devices, frame rates and acquisition resolutions, nowadays more and more applications are dealing with huge amounts of data. As human operators cannot manually process these data streams continuously and indefinitely, automatic video processing techniques are required for various tasks. Change detection is one of these tasks which focuses on detecting abnormal events or areas of potential interest.

In this paper, we focus on the specific problem of change detection in aerial videos, as a way to filter out uninteresting data and focus only on areas containing changes. Change detection [1] refers to the problem of detecting significant and possibly subtle changes between reference and test data (e.g. appearing or disappearing buildings or vehicles), while ignoring insignificant ones, such as environmental changes (illumination, weather, ...) and parallax effects due to camera motion and 3D objects (trees, buildings, relief ...). Most of the current change detection techniques focus on comparison of image pairs, and their extension to video is not straightforward and raises many specific problems. For instance, prior to image comparison, one needs to find reference data which covers the same area as test data. Possible solutions use temporal synchronization of videos acquired along similar trajectories [2, 3] or summarization of reference data into a 3D model [4].

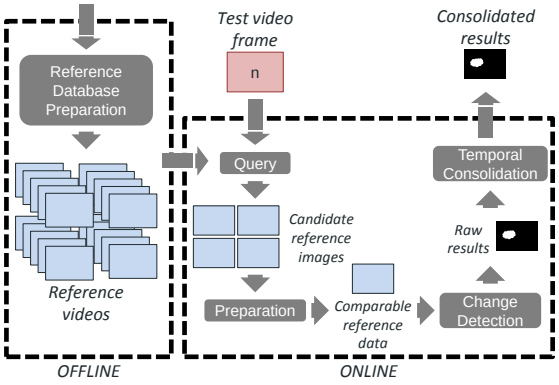


Fig. 1. This drawing presents a work flow visualization of the change detection framework used in our experiments.

Redundancy is another specific problem, which aims to exploit the spatial and temporal coherence of acquisitions through different video frames. Indeed, successive frames cover the same areas but they are taken under different noise conditions, angles, etc., which may affect the precision of change detection if frames are processed independently. In this paper, we present an alternative which exploits the spatial and temporal redundancy, using Belief Propagation. Our approach unifies a pixel-wise change detection criterion with a high order transition term which defines the spatio-temporal interaction between neighbor pixels. Compared to baseline techniques, the gain of our algorithm is clear and consistent. Moreover, the use of Belief Propagation is adequate in the context of aerial video, because it does not require an explicit construction of graphs, hence resulting in improved efficiency.

The remainder of this paper is organized as follows. Section 2 briefly describes our aerial video change detection

framework followed in Section 3 by our main contribution which exploits spatio-temporal redundancy to improve accuracy. Finally, Section 4 presents and discusses the performance of our method.

2. OVERVIEW

This section briefly introduces our change detection framework adapted to spatial and temporal consolidation. In order to speedup the whole process, an off-line preparation of reference data is achieved. Video frames are first spatially indexed by organizing their viewprints using an R-Tree structure. Change detection is then achieved online using only the frames of reference video which cover the same spatial area as the test ones. Finally, candidate reference frames are merged into a mosaic and compared to the current test frame using our algorithm introduced in [5]. The latter has proven to be fast and accurate. As an extension of this previous work, spatial and temporal consolidation is the main focus of this paper as depicted in the flowchart of Fig. 1.

3. SPATIAL AND TEMPORAL INTERACTION

This section describes two approaches for spatio-temporal consolidation of change detection results. The first one is a baseline; it is based on a simple and fast temporal averaging of change detection scores, but suffers from inaccuracies. The second method constitutes our new contribution which addresses the problem using Loopy Belief Propagation [6] and achieves better accuracy while still being computationally efficient.

3.1. Averaging

The baseline method enforces the consistency of the results by averaging the scores of change detection on the same areas through successive frames in a test video. This averaging makes it possible to predict the status of a given pixel as *changed* or *unchanged* depending on the average score. Let $\Omega_t = \{t-T+1, \ldots, t\}$ be a temporal window starting at frame $t-T+1$ and ending at frame t and let $\{b_k(X)\}_{k\in\Omega_t}$ denote a collection of binary variables all referring to an existing physical point X in the scene; here $b_k(X)$ is set to 1 if X is inside the k^{th} frame and 0 otherwise. In the remainder of this section, $\widehat{e}(X)$ stands for a scoring function which allows us to predict the status of X.

As camera sensors may move through successive frames, it is necessary to estimate correspondence between pixels. We address this problem by computing the average score $\widehat{e}(X)$ in an absolute coordinate system, the most natural being the dominant plane of the ground. More precisely, raw change detection scores are computed in the current frame of the test video and mapped to the ground plane, via a projective transformation, where they are averaged with previous observations as follows:

$$\widehat{e}(X) = \frac{1}{N_{obs}(X)} \cdot \sum_{k\ \in\Omega_t} b_k(X)\cdot e_k(H_k^{-1}\cdot X) \qquad (1)$$
$$\text{with } N_{obs}(X) = \sum_{k\in\Omega_t} b_k(X),$$

where $e_k(H_k^{-1}\cdot X)$ refers to the raw change detection score of X in the k^{th} test frame and H_k^{-1} is a projective transformation mapping the coordinates of X from the ground plane to the k^{th} frame. Notice that, when using an infinite temporal window ($T=\infty$), the sums in Eq. 1 may be updated incrementally, which allows us to achieve very efficient and effective consolidation of change detection scores through successive frames. Following evaluation of Eq. 1, physical point X is declared as *changed* (resp. *unchanged*) if and only if the score $\widehat{e}(X)$ is bigger (resp. lower) than a fixed threshold. The latter is adjusted depending on the required false alarm and mis-detection rates (see Section 4).

3.2. Belief Propagation for Spatio-Temporal Interaction

In this section, we use Loopy Belief Propagation in order to predict the status of pixels in test frames. Let $\mathbf{I}_t^{test}$ be the current frame of a test video including $w\times h$ pixels. Let $\{\mathbf{X}_{i,j,t}\}_{i,j}$ and $\{\mathbf{Y}_{i,j,t}\}_{i,j}$ be collections of random variables respectively standing for pixels in $\mathbf{I}_t^{test}$ and their labels (corresponding to the predicted status); here $\mathbf{Y}_{i,j,t}=1$ if pixel (i,j) in $\mathbf{I}_t^{test}$ has changed and $\mathbf{Y}_{i,j,t}=0$ otherwise. Enforcing the spatio-temporal consistency in change detection may be viewed as finding the optimal configuration of labels $\{\mathbf{Y}_{i,j,k}\}_{i\leqslant w, j\leqslant h, k\in\Omega_t}$, best explaining the observations $\{\mathbf{X}_{i,j,k}\}$ obtained on previous test frames. In the following, we denote by $\{\mathbf{I}_k^{test}\}_{k\in\Omega_t})$ and $\{\mathbf{I}_k^{ref}\}_{k\in\Omega_t}$ respectively the k^{th} test and reference frames, for $k\in\Omega_t$. Without any loss of generality, we assume that all these frames are registered with respect to the current test frame $\mathbf{I}_t^{test}$ (see [7] for a survey on image registration methods).

For a fixed t, we model the labeling problem as a 3D Markov Random Field using a non-oriented adjacency graph $\mathcal{G}_t=(\mathcal{V}_t,\mathcal{E}_t)$ where each vertex in $\mathcal{V}_t=\{v_{i,j,k}\}_{i\leqslant w, j\leqslant h, k\in\Omega_t}$ is associated to a pair $\{(\mathbf{X}_{i,j,k},\mathbf{Y}_{i,j,k})\}_{i,j,k}$ and where edges $\mathcal{E}_t=\{e_{\mathbf{n},\mathbf{n}'};\ \mathbf{n}=(i,j,k),\mathbf{n}'=(i',j',k')\}$ are connections between neighbor nodes. More precisely, we consider that neighborhood connections $e_{\mathbf{n},\mathbf{n}'}$ are connections between vertices $v_{\mathbf{n}}$ and $v_{\mathbf{n}'}$, where $\|\mathbf{n}-\mathbf{n}'\|_1=1$ (with $\|.\|_1$ denoting the L_1 norm). This definition characterizes the neighborhood system: each vertex has four spatial and two temporal neighbors. In this context, Loopy Belief Propagation aims to predict the *optimal* configuration of labels $\{\mathbf{Y}_{i,j,k}\}_{i,j,k}$ by minimizing an objective function which trades off unary

potential terms $\{\phi(\mathbf{X}_{i,j,k}, \mathbf{Y}_{i,j,k})\}_{i,j,k}$ and binary interaction terms $\{\psi(\mathbf{Y}_{i,j,k},\ \mathbf{Y}_{i',j',k'})\}$. The unary potential terms $\{\phi(\mathbf{X}_{i,j,k}, \mathbf{Y}_{i,j,k})\}_{i,j,k}$ link the predicted labels to the underlying observations, and each term is defined as:

$$\phi(\mathbf{X}_{i,j,k}, \mathbf{Y}_{i,j,k}) = \begin{cases} f(\mathbf{X}_{i,j,k}) & \text{if } \mathbf{Y}_{i,j,k} = 0 \\ & (\textit{unchanged}) \\ 1 - f(\mathbf{X}_{i,j,k}) & \text{if } \mathbf{Y}_{i,j,k} = 1 \\ & (\textit{changed}) \end{cases} \quad (2)$$

$$\text{where } f(\mathbf{X}_{i,j,k}) = \Big(c_0 \cdot \exp(-c_1 \cdot \Delta(\mathbf{X}_{i,j,k})) - c_0 + 1\Big) \cdot \exp\left(-\frac{e_k(\mathbf{X}_{i,j,k})}{\tau}\right) \quad (3)$$

$$\text{and} \quad \Delta(\mathbf{X}_{i,j,k}) = \Delta_{i,j,k} = \left|\mathbf{I}_k^{test}(i,j) - \mathbf{I}_k^{ref}(i,j)\right| \quad (4)$$

The right-hand side term in Eq. 3, encourages labeling pixels as *changed* (respectively, *unchanged*) if the underlying change detection score is high (respectively, low). This term is weighted by the scale parameter τ which controls false-alarm and mis-detection rates. The left-hand side term in Eq. 3 also controls these detection rates by taking into account image differences weighted by the coefficients c_1, c_0. When tuning these parameters, we found that the best performance is achieved with $c_0 = 0.33$ and $c_1 = \frac{\log(2)}{30}$.
The binary interaction terms $\{\psi(\mathbf{Y}_{i,j,k},\ \mathbf{Y}_{i',j',k'})\}$ exploit the neighborhood system defined earlier in order to enhance the spatio-temporal consistency of labels. These terms are defined as:

$$\psi(\mathbf{Y}_{i,j,k},\ \mathbf{Y}_{i',j',k'}) = \quad (5)$$
$$\begin{cases} 0.95 \cdot \lambda(\Delta_{i,j,k},\ \Delta_{i',j',k'}) + 0.5 \cdot (1 - \lambda(\Delta_{i,j,k},\ \Delta_{i',j',k'})) \\ \qquad \text{if } \mathbf{Y}_{i,j,k} = \mathbf{Y}_{i',j',k'} \\ 0.05 \cdot \lambda(\Delta_{i,j,k},\ \Delta_{i',j',k'}) + 0.5 \cdot (1 - \lambda(\Delta_{i,j,k},\ \Delta_{i',j',k'})) \\ \qquad \text{if } \mathbf{Y}_{i,j,k} \neq \mathbf{Y}_{i',j',k'}, \end{cases}$$

where

$$\lambda(\Delta_{i,j,k},\ \Delta_{i',j',k'}) = \frac{\frac{\pi}{2} - \text{atan}(b \cdot (|\Delta_{i,j,k} - \Delta_{i',j',k'}| - a))}{\frac{\pi}{2} - \text{atan}(-b \cdot a)} \quad (6)$$

In the above definition, $\lambda(\cdot)$ acts as a Radial Basis Function (see plots for different values of a and b in Fig. 2) which influences the interaction between labels depending on the gradient norm $|\Delta_{i,j,k} - \Delta_{i',j',k'}|$. More precisely, large values of this norm are likely caused by the fact that nodes $v_{i,j,k}$ and $v_{i',j',k'}$ belong to independent objects in the scene. Therefore, there should not be any correlation between the underlying labels, this is why the right-hand side terms in Eq. 5, are weighted uniformly. Conversely, low values of the gradient norm are likely caused by nodes $v_{i,j,k}$ and $v_{i',j',k'}$ which belong to the same object, therefore we strongly encourage similar labels for these nodes. Hence, a large weight (0.95) is used in the left-hand side term of Eq. 5 in order to encourage similar labels, while a small weight (0.05) is used for different labels. In practice, we used $a = 25$ and $b = 0.175$, and we chose a temporal window $T = 8$.

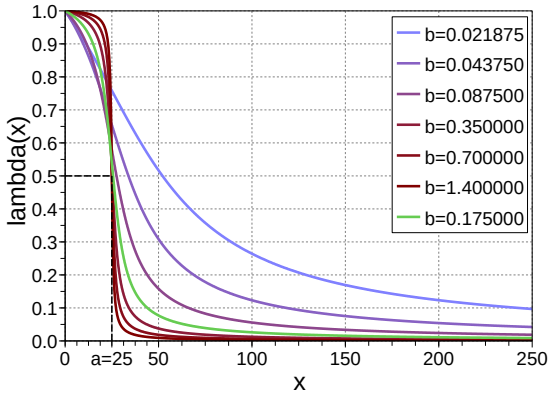


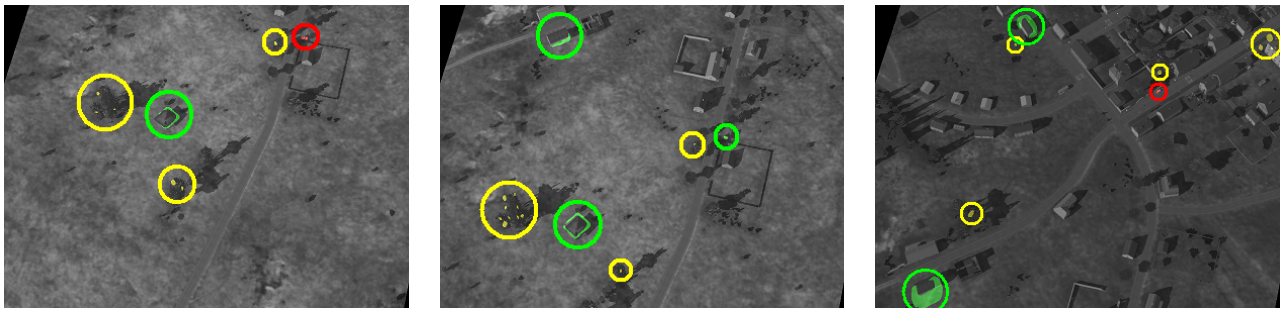
Fig. 2. This figure illustrates the aspect of RBF function λ used in the binary interaction terms. The curve used in our experiments is highlighted in green.

Using these unary and binary terms, Loopy Belief Propagation estimates the probability of change for each node of the graph. This algorithm is iterative and propagates messages carrying the likelihood of each state, from each node to its neighbors (see [6] and [8] for more details). Convergence of this message passing algorithm is not guaranteed in the presence of cycles in the graph, but in practice we observe that this algorithm reaches a good solution in few iterations (corroborated by the enhancement of the change detection performance with respect to the baseline versions, see Section 4 and Fig. 4).

4. EVALUATION

In this section, we compare our approach based on Belief Propagation with respect to the baseline averaging scheme described in Section 3.1 using aerial video sequences for which ground truth changes are known for each frame. Fig. 3 shows visual inspection results obtained using the temporal averaging approach on several video frames, where true detections, false alarms and mis-detections are respectively highlighted in green, yellow and red. In these results, a few false alarms still occur in some locations, especially in areas containing untargeted changes such as waving trees. We also observe in our video results that small changes might not be detected when they enter into the field of view. Nevertheless, temporal consolidation tracks back these mis-detections most of the time a few frames after their appearance. In contrast, important changes are detected as soon as they appear.

Fig. 4 presents the comparison of Precision/Recall performance with respect to the baseline versions (with and with-



Frame 120 Frame 190 Frame 340

Fig. 3. This figure shows typical results of our video change detection algorithm, where true detections, false alarms and mis-detections are respectively highlighted in green, yellow and red.

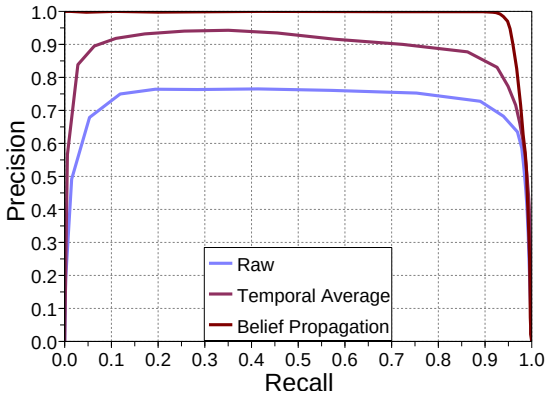


Fig. 4. This figure presents performances obtained by our approach based on Belief Propagation and its comparison with respect to baseline versions (with and without averaging).

out averaging). As expected, for the same values of Recall, spatio-temporal consolidation, based on belief propagation, consistently improves Precision. This results from the optimization of an objective function which accurately models the problem of change detection. This performance enhancement is achieved at the detriment of an increase of processing time, which is still reasonable. On a standard 2.4 GHz computer using a mono-thread implementation, spatio-temporal consolidation of change detection results (800×600 pixels) takes approximately 0.75 seconds for the averaging method and about 15 seconds for the Belief Propagation method. Notice that these execution times may be further improved using hardware acceleration.

5. REFERENCES

[1] R.J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE Transactions on Image Processing*, vol. 14, no. 3, pp. 294–307, 2005.

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

[3] C. Stennett and R.J. Evans, "Visual change detection for route monitoring," in *Proceedings of the 2009 Conference of Electro Magnetic Remote Sensing Defence Technology Centre*, 2009.

[4] T. Pollard and J.L. Mundy, "Change detection in a 3-d world," in *2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '07)*, 2007, pp. 1–6.

[5] N. Bourdis, D. Marraud, and H. Sahbi, "Constrained optical flow for aerial image change detection," in *2011 IEEE International Geoscience and Remote Sensing Symposium*. 2011, IEEE Computer Society.

[6] C.M. Bishop, "Graphical models," in *Pattern Recognition and Machine Learning*, vol. 4, chapter 8, pp. 359—422. Springer, 2006.

[7] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, no. 11, pp. 977–1000, 2003.

[8] P.F. Felzenszwalb and D.R. Huttenlocher, "Efficient belief propagation for early vision," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, 2004, vol. 1, pp. 261–268.