CONTEXT-DRIVEN MOVING OBJECT DETECTION IN AERIAL SCENES WITH USER INPUT

C. Guilmart, S. Herbin

Onera - The French Aerospace Lab
F-91761 Palaiseau, France

P. Perez

Technicolor Research & Innovation
1, avenue de Belle Fontaine
35576 Cesson-Sévigné, France

ABSTRACT

Aerial video sequences are a common source for applications such as intelligence, surveillance or search and rescue. Their off-line analysis however requires a certain level of assistance to reduce the expert's workload. This study focuses on detecting mobile vehicles in such sequences. The proposed approach exploits two types of contextual information: loose user input as tagged areas in a reference frame, and knowledge-based priors to describe specific constraints. Our main contribution is the design of a two-step general framework able to combine these two types of information. The first step is a pixelwise semantic classification labelling each sequence frame structure in *vehicle*, *road* and *background* ; the classifier is based on local motion and appearance features and is organized as an iterative refining process. The second step exploits knowledge-based spatial reasoning to filter out false alarms. A quantitative evaluation on real video sequences demonstrates the usefulness of each level of contextual information.

Index Terms— video analysis, motion detection, aerial video, contextual information

1. INTRODUCTION

Aerial vision is an efficient means for monitoring areas which would be otherwise difficult to observe for safety, cost or accessibility reasons. However, with the ever growing volume of data, automated interpretation tools are needed to help reduce the workload of experts and focus on the small temporal segments which contain some activity. Moving entity detection thus appears as one major objective of a video interpretation chain. This article describes how several levels of contextual information can be used in an activity detection processing chain, specifically in moving vehicle detection. The first level requires light tagging of the scene structure on a reference frame and propagates probabilistically the annotation to the subsequent frames using learned classifiers on local image features from the reference frame. A second level exploits elementary knowledge, namely the fact that "vehicles drive on road and follow its direction", to refine the final activity likelihood and filter out false alarms. Figure 1 displays the successive steps of the framework.

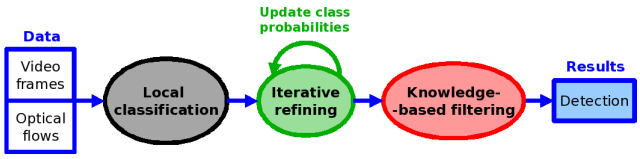


Fig. 1. Diagram of the proposed approach to semi-supervised detection with two levels of context

Our contribution is twofold. First, we introduce an appearance-based learning framework in the field of aerial video interpretation, and show its usefulness compared with pure geometric approaches. Second, we consider two different levels of contextual information management: sample-based and knowledge-based, and show how they complement each other to handle the local and global structure of the scene. This approach is evaluated on different types of aerial videos with various levels of complexity.

2. RELATED WORK

Detecting moving entities using still cameras has been addressed for a long time using background subtraction techniques or a combination of local appearance and motion information. Due to camera motion, specific difficulties arise in aerial videos, such as parallax and occlusion. [1, 2, 3] exploit global motion compensation and geometric modelling to filter those out. In [4, 5, 6], the same problem is tackled along with multi-object tracking. However, in neither case are contextually trained classifiers or knowledge-based priors used as we propose here, which limits spurious detections. Specific classifiers [7], [8] may be used to describe target objects using interest points. It nonetheless requires that they are sufficiently resolved in the images, which is usually not the case in aerial videos. Semantic segmentation exploiting learning techniques has been addressed mainly for image indexing [9] or remote sensing labelling [10]. More recently, several studies [11, 12] have addressed the semantic segmen-

tation of street view videos shot by a still camera or from a car thanks to a specific database [13]. However, those approaches rely on 3D cues which are seldom reliable in aerial videos.

Markov Random field models are a popular probabilistic framework for representing neighborhood dependencies in images. In this huge family of studies, our approach is closest to the "Auto-Context" proposal [14, 15] which exploits an iterative scheme based on neighborhood functions with learned parameters. The main difference lies in the way the current probability map is processed in each iteration.

Context in images, i.e. the use of spatial and semantic relations to improve labelling, has been used in several studies [16], [17]. Most of those methods however rely on a heavy learning phase with databases consisting of thousands of labelled images. Contextual information may also represent groups of objects sharing visual or dynamic characteristics [18, 19], which is useful in situations with such collective appearances or behaviours, e.g. dense traffic, but is not suited to the detection of isolated objects.

3. LOCAL CLASSIFICATION

Our goal is to detect mobile vehicles in a video sequence using local and semantic contextual information. As such, in addition to the two classes *mobile vehicles* (class 1) and *background* (class 2), we add a third one *road* (class 3) which will serve as hidden information. The knowledge-based priors we introduce in the last step of our framework will be expressed through joint detection of *road* and *vehicles*.

We use few temporal features as well as appearance features. The temporal features stem from residual optical flow obtained after dominant affine motion compensation [20]. The original optical flow was computed by the Huber-L1 algorithm [21] as it preserves edges and provides a smooth flow in homogeneous regions. Both types of features are computed as order 1 and 2 moments over local 3×3 patches: among the 6 features for each patch, 4 are related to appearance (moments of intensity and saturation) and 2 are related to motion (moments of the residual flow norm). This norm depends on the global motion through the scale factor of the affine motion: we thus normalize it so that the norm is coherent throughout the sequence.

The multi-class problem is reduced to a set of one-against-one classifiers, the *base learners*. Each base learner is obtained from weak learners through Adaboost. A decision tree [22] generates a pool of weak learners with reasonable error rates from the training samples (6 features and one label for each sample). Each inner node of the tree provides a weak learner in the form of a decision stump, i.e. a binary classifier associated with a single feature and a specific threshold. Based on this pool, Adaboost selects a subset of weak learners and combines them into a strong classifier for the pair of classes under consideration.

The base learners are computed in a training phase. In the test phase, the features are computed for each pixel and the base learners are applied to these features. A symmetric multiple logistic transformation [23] is applied to the outputs of the base learners, 3 in our current system. This results in a semantic probability map with three labels, an example of which is shown in figure 2 (b).

4. ITERATIVE REFINING

This first pixelwise probability map, which has only been inferred from local primitives as described in section 3, is noisy and needs to be refined. We choose an iterative approach as in the auto-context model [14, 15]. The initial set of maps is used as context information, in complement to the original image patches, to train a new set of classifiers, and the algorithm iterates. Contrary to models such as Markov Random Fields (MRFs), Conditional Random Fields (CRFs) or Belief Propagation, there is no need to minimize an explicit energy.

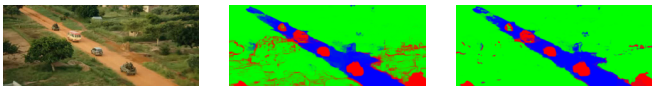
Contextual features are derived from the estimate of the probability maps at the considered iteration. The first n_c features, where n_c is the number of classes (here 3), are averages of the probabilities from the pixels of the 3×3 patch (similar to [14]). To account more completely for the local structure of the image, we add an extra n_c features which are the weighted means of these same patch probabilities: similar pixels should belong to the same class, which may not be true along low-contrast boundaries. The weights associated with the central pixel and a neighbour pixel thus take into account the difference between the two intensities as well as the average angular error between the two residual flow vectors. These n_c features are then normalized. For each pixel, a final vector of $6 + 2n_c$ features is obtained by concatenating the $2n_c$ contextual features to the original appearance and motion ones.

The pixel maps of the class probabilities (one map for each class) are initialized by a uniform distribution on the classes, i.e. $\frac{1}{n_c}$ for each class at each pixel. The corresponding features are not discriminative so they will not be used by the classifier at the first iteration, which is thus equivalent to the local classifier described in Section 3.

In the following iterations, the class probability maps are not uniform anymore, since they are output by the previous classifier. In the training step, the base learners will differ at each iteration. In the test step, they will be applied in the same order. The number of iterations, between 2 and 5, is chosen so as to achieve a trade-off between precision and overfitting. Figure 2 illustrates this refinement step.

5. KNOWLEDGE-BASED CONTEXTUAL FILTERING

In order to detect moving vehicles, two simple criteria can be considered: the distance of a pixel to the road network and the local alignment between the motion of a vehicle and the surrounding road. These criteria are invariant to rotation and



(a) (b) (c)

Fig. 2. Iterative refining of probability maps (a) test frame (b) local classification (iteration 1) (c) iteration 5. Red channel : "mobile vehicle" ; blue channel : "road" ; green channel : "background"

the last one is invariant to scale. However, the quality of the road network representation is critical. In our experiments, the estimated optical flow as well as the the computed road directions were not sharp enough to define a reliable alignment measure. For this reason, we only resort to the car-to-road distance criterion to define knowledge-based rules.

We chose to model the roads by 7 features learned on the training frame : the three RGB color channels as well as $\frac{R}{B}, \frac{R}{G}$ and $\frac{G}{B}$ and intensity. For each feature, we compute the histogram for the training frame (for which the road has been manually tagged) and keep the lower and upper bounds of the bins with sufficient count as thresholds. For each test frame, the binary product of 7 masks (for each feature, 1 for pixels within the associated thresholds and 0 otherwise) gives a rough binary road network. Morphological operations are then carried out on the raw network. It is first closed to smooth the contours, then small and thin regions are considered as noise and filtered out.

After the iterative classification process, we apply the final knowledge-based filters as follows:

- A road network N_{road} is obtained as described above.
- The distance map D_{road} of N_{road} and the subsequent distance score map S_{dist} (soft thresholding of the distance map) are computed.
- The final mobile vehicle score map S_{cont} is computed as the product of S_{dist} and p_1 which is the mobile vehicle probability map after iterative refining.
- Thresholding S_{cont} provides a binary map.
- Classical morphological post-processing is applied to remove the smallest regions and fill the smallest gaps.

6. EXPERIMENTS

The different steps of our system are applied to several videos taken in varying conditions and complexity from documentaries and films. "Blood Diamond" (fig. 2 (a)) is quite simple, with few vehicles and one road. "Are we changing Planet Earth" (fig. 3 (b)) is more complex with several roads and many buildings with appearance similar to roads. The "BBC" sequence is the most challenging sequence (fig. 3 (a)) and presents strong focal changes through time as well as watermarks and strong coding artefacts. The incidence is steep in

(a) (b)

Fig. 3. (a) : "BBC" ; (b) : "are we changing Planet Earth"

Table 1. Equal error rates in %. Figures are given respectly without and with knowledge-based context for the 3 first lines.

	Blood Diamond	are we	BBC
local classification (iter 1)	64 / 82	41 / 52	39 / 47
iteration 2	66 / **83**	48 / 56	40 / **50**
iteration 5	60 / 81	43 / **58**	23 / 30
baseline	50	41	25

all three sequences: the size of the vehicles is much smaller on the top part of the frames.

Equal Error Rates of these curves are used as a convenient means to compare the performances. In addition to such pixel-wise metrics (table 1), object-wise metrics more directly related to the actual application, namely the detection of moving vehicles, can also be used. The connected components of final detection maps are considered as detected objects. Connected component C_k is counted as a hit if its intersection with the nearest ground truth object O_k is large enough relative to the sizes of the objects. Table 2 gives object-based precision (number of true detections related to the total number of detected objects) and recall (number of true detections related to the total number of ground truth objects) for each sequence, for an average precision of 50%.

Our results are compared to a baseline based on minimization of a pairwise MRF. Unary potentials of this MRF are defined as the pixel residual flow norms normalized by the maximum norm of the moving objects on the training frame. Binary clique potentials (for all couples of 8-neighbours) are specified as a weighted means of color differences and average angular errors. Global energy minimization is obtained by min-cut/max-flow approach [24], which yields a binary detection map. Several maps (hence several precision-recall points) are obtained by varying the weight of the unary potentials (the larger this weight, the fewer miss-detections).

The distance-to-road criterion significantly improves the results. For the "are we changing Planet Earth" sequence, the improvement is more marked at iteration 5 than at iteration 2, partially correcting the overfitting (table 1). Some false detections after (appearance and motion based) local classification are due to a wrong estimation of the residual flow caused by compression artefacts ("BBC" sequence) or inexact affine compensation ("are we" sequence). False detections

Table 2. Object-based precision and recall in % for, respectively, the "Blood diamond" sequence (288 ground truth vehicles) / the "are we changing planet Earth" sequence (854 vehicles) / the "BBC" sequence (420 vehicles). Except for the "BBC" baseline, precision is about 50%.

	Precision	Recall
local classification (iteration 1)	51 / 50 / 51	57 / 36 / **39**
iteration 2 with knowledge-based context	50 / 49 / 53	**81** / **55** / 35
baseline	48 / 54 / 43	35 / 16 / 30

due to parallax are partially corrected (buildings or poles in both documentaries as well as the tree in the film). However, 3D structures close to detected road segments will not be filtered. In almost all cases, the whole system yields the best results. That is not true with object-based metrics for the BBC sequence, in which case small vehicles in the top part of the sequence are not detected due to the morphological filtering of small objects.

7. CONCLUSION

In this paper, we have proposed a two-step supervised framework to detect mobile vehicles in aerial video sequences. The scene pixels are first classified in moving vehicles, road and background, using patch-based appearance and motion features. The resulting class probability maps are then iteratively refined. Spurious detections, e.g. due to parallax or compensation errors on strong edges, are finally removed by applying simple knowledge-based priors. Quantitative results, based on both pixel-wise and region-wise metrics, show improvements at each stage. Some false alarms remain though, especially those caused by parallax. The flexibility of our framework offers several directions of research to improve the detection quality, whether they be feature-based or knowledge-based. Finally, temporal consistency and preprocessing of the raw video data have not been taken into account yet.

8. REFERENCES

[1] G. Salgian, J. Bergen, S. Samarasekera, and R. Kumar, *Moving target indication from a moving camera in the presence of strong parallax*, Citeseer, 2006.

[2] H. Yalcin, M. Hebert, R. Collins, and MJ Black, "A flow-based approach to vehicle detection and background mosaicking in airborne video," in *CVPR*, 2005, vol. 2, pp. 1202–vol.

[3] J. Xiao and M. Shah, "Motion layer extraction in the presence of occlusion using graph cuts," *PAMI*, vol. 27, no. 10, pp. 1644–1659, 2005.

[4] Q. Yu, I. Cohen, G. Medioni, and B. Wu, "Boosted markov chain monte carlo data association for multiple target detection and tracking," *ICPR*, vol. 2, pp. 675–678, 2006.

[5] Q. Yu and G. Medioni, "Motion pattern interpretation and detection for tracking moving vehicles in airborne video," in *CVPR*. IEEE, 2009, pp. 2671–2678.

[6] Y. Sheikh, O. Javed, and T. Kanade, "Background subtraction for freely moving cameras," in *ICCV*, 2009, pp. 1219–1225.

[7] J. Xiao, C. Yang, F. Han, and H. Cheng, "Vehicle and person tracking in aerial videos," in *Multimodal Technologies for Perception of Humans*, vol. 4625 of *LNCS*, pp. 203–214. 2008.

[8] B. Ommer, T. Mader, and J.M. Buhmann, "Seeing the objects behind the dots: Recognition in videos from a moving camera," *IJCV 2009*, vol. 83, no. 1, pp. 57–71.

[9] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *CVPR 2008*, pp. 1 –8.

[10] S. Kluckner, T. Mauthner, P. Roth, and H. Bischof, "Semantic classification in aerial imagery by integrating appearance and height information," in *ACCV*, pp. 477–488. 2009.

[11] GJ Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," *LNCS*, vol. 5302, no. PT 1, pp. 44–57, 2008.

[12] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. Torr, "What, where and how many? combining object detectors and crfs," in *ECCV 2010*, vol. 6314 of *LNCS*, pp. 424–437.

[13] G.J. Brostow, J.Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *PRL*, vol. 30, no. 2, pp. 88 – 97, 2009.

[14] Z. Tu, "Auto-context and its application to high-level vision tasks," in *CVPR*, 2008, pp. 1–8.

[15] J. Jiang and Z. Tu, "Efficient scale space auto-context for image segmentation and labeling," in *CVPR*. IEEE, 2009, pp. 1810–1817.

[16] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in Context," in *ICCV*. IEEE, 2007, pp. 1–8.

[17] Geremy Heitz and Daphne Koller, "Learning spatial context: Using stuff to find things," in *ECCV*, vol. 5302 of *LCNS*, pp. 30–43. 2008.

[18] Y. Wu and J. Fan, "Contextual flow," in *CVPR*. IEEE, 2009, pp. 33–40.

[19] J. Fan, J. Xu, and Y. Wu, "Context-aware tracking of small targets in video," in *SPIE*, 2009, vol. 7445, p. 7.

[20] J.M. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models," *JVCIR*, vol. 6, no. 4, pp. 348–365, 1995.

[21] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof, "Anisotropic huber-l1 optical flow," in *BMVC*, 2009.

[22] L. Breiman, JH Friedman, RA Olshen, and CJ Stone, *Classification and regression trees*, Chapman Hall, New York, 1993.

[23] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *The Annals of Statistics*, vol. 28, no. 2, pp. 337–374, 2000.

[24] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *PAMI*, vol. 26, no. 9, pp. 1124–1137, 2004.