COMBINATION OF THERMAL AND COLOR IMAGES FOR ACCURATE FOREGROUND / BACKGROUND SEGMENTATION IN OUTDOOR ENVIRONMENT

Louis St-Laurent[1], *Donald Prévost*[1], *Xavier Maldague*[2]

[1]National Optics Institute, Quebec City, Quebec, Canada
[2]Department of Electrical and Computer Engineering, Laval University, Quebec City, Quebec, Canada

ABSTRACT

In the context of outdoor video surveillance, this paper attempts to answer the following question: how to combine LWIR and color information in order to optimize foreground / background segmentation accuracy. Starting from an improved state-of-the-art color-based approach, we integrated thermal information into the algorithm with a pixel-level analytical fusion technique. Considering that very few public thermal / color video databases are available, we built our own acquisition platforms to grab numerous co-registered LWIR / color videos in a variety of outdoor conditions. We manually generated the pixel-based ground-truth for a representative selection of these sequences and performed a quantitative performance analysis. We demonstrated, among others, that the combination of thermal and color information proposed outperformed the use of a single spectral band in all tested visibility conditions.

Index Terms— Sensor fusion, change detection algorithms, outdoor video surveillance, thermal imaging, object tracking.

1. INTRODUCTION

Performing accurate and real-time automatic detection and tracking of human and vehicles in outdoor environment is very challenging, especially because of dynamic background, lighting changes and climate factors. Combining thermal and color information is not trivial since the contrast between objects of interest and background strongly varies over time in both sensors. To develop a surveillance system efficient toward most conditions, the fusion algorithm must make the most of one sensor strengths without being affected by weaknesses of the other.

Thermal and visible information may be combined at pixel-level or at object-level. In the latter, extraction of objects of interest is performed independently on each modality and association / fusion rules are applied on the output. In the method proposed by Snidaro *et al.* [6], a confidence measure based on contrast of detected thermal and visible foreground blobs is used to weight the contribution of each sensor. Hence, the position of a blob with a good contrast relatively to the background will have a larger impact on the fused predicted position of the track. A main drawback of object-level fusion is that correspondences between thermal and visible detections must be resolved

Approaches combining information at pixel-level can be separated into two classes: representative and analytical fusion. Image fusion is the expression commonly used to describe representative fusion. Generally, its purpose is to generate a new image more informative or intuitive for a human observer. Waxman *et al.* [12] proposed to use such fused images as input for automated detection and tracking. But it is important to understand that the generation of a new image is not required for automated video monitoring applications. For this reason, analytical fusion, which could be defined as the combination of available information from sensors for a more robust analysis and interpretation of video content, seems more suitable for automated video monitoring applications.

Most motion-based approaches (temporal subtraction, background subtraction, optical flow) are valid for both thermal and visible images. For applications where the acquisition unit is fixed, background subtraction methods are almost always used as a first stage of foreground / background classification. Ó Conaire *et al.* [3] proposed to model the color background with a mixture of Gaussians, and the thermal background with a single distribution. A global threshold is used for thermal detection, and its value is adjusted to maximize the similarity between visible and thermal detections. Pixels classified as foreground in each band are combined with a logical “or”. Finally, foreground blobs not containing at least one pixel detected from thermal and from visible images are eliminated, thus potentially leading to miss detections when the contrast observed by one sensor is very low.

2. OVERVIEW OF THE PROPOSED METHOD

To maximize foreground / background segmentation accuracy, our opinion is that thermal and visible information must be combined at the lowest processing level: the pixel. The method presented in this paper may be categorized as a pixel-level analytical fusion technique.

Like in our preliminary work [8], the non-parametric codebook model proposed by Kim *et al.* [1] is used as

starting point for background modelling. But before integrating thermal information in the codebook model, we meticulously optimized the color-based detection method of [1] for outdoor environment [10].

To combine information at pixel-level, temporal and spatial registration of both sensors is required. In [8], an acquisition platform with a beamsplitter (to superpose optical axis) was used to grab a few registered low resolution image sequences. Cumbersome and limited to narrow field of view, we replaced this unit by a side-by-side camera configuration integrated in a rugged housing [9]. Spatial alignment of images grabbed with a side-by-side camera configuration can be performed by affine homography transformation where the projection matrix *H* is determined from pairs of corresponding features like in the work of Torabi *et al.*[11]. But for a more accurate registration, we developed an in-lab internal / external calibration-based procedure [9].

3. HYBRID CODEBOOK

Instead of modeling thermal and color background independently, we propose to combine data from both sensors in a single hybrid codebook in which every pixel is represented by *L* codewords (*CW*):

$$CW_{k=1\ldots L} = \left\{ \overline{Y}_k, \overline{Co}_k, \overline{Cg}_k, \overline{T}_k, f_k, p_k, q_k, MNRL_k \right\} \quad (1)$$

where $\overline{Y}_k, \overline{Co}_k, \overline{Cg}_k$ and $\overline{T}_k$ are luma, chroma orange, chroma green[1] and thermal values of *CW k*. Parameters *f*, *p* and *q* are respectively the number of matches, the time stamp of the first match, and the time stamp of the last match. *MNRL* (Maximum Negative Run-Length) stores the length of the period (in number of frames) during which a *CW* has not been matched. A threshold on *MNRL* is used to filter out *CW* belonging to moving objects.

At every new frame, every pixel is associated to the first sufficiently similar *CW*. If no codeword can be matched, a new one is created and added in a cache codebook. Codewords from the cache are promoted to permanent background codebook when they are repetitively matched, and codewords from permanent background codebook not matched since a long period of time are deleted.

The diagram of Fig. 1 presents the association rules used to match a pixel value with a background *CW k*. The first condition tests the thermal variation against a global detection threshold ε, while the formula written in the second diamond tests if the color pixel value is enclosed in a spherical association volume. If this last condition is not fulfilled, we verify if the observed value is enclosed in the cylindrical association volume corresponding to cast shadow pixels as proposed in [10]. Because of space constraints, please refer to [10] for full description of parameters related to color information (second and third diamonds of Fig. 1).

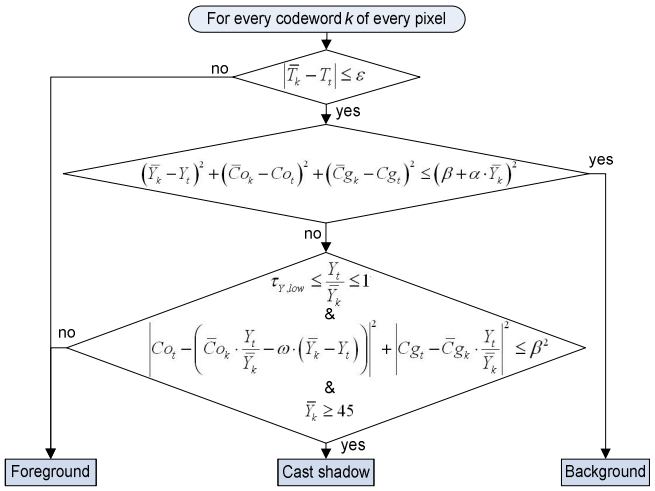

Fig. 1. Flow diagram of the proposed foreground / background / cast shadow classification.

3.1. Semi-automatic tuning of thermal threshold ε

Adjusting the thermal detection threshold in time is particularly important for outdoor scenes because thermal contrast is significantly reduced in presence of rain and wind. In this situation, increasing the camera gain is a common practice, leading to much noisier images. Moreover, most thermal sensors suffer from sudden intensity changes, either caused by AGC compensation when a large highly contrasted object enters in the field of view, or after periodic flat field correction, which is typical to uncooled microbolometers sensors.

To maximize detection accuracy, we propose to update the thermal detection threshold ε at every frame *t* based on a periodically updated standard deviation σ (temporal sensor noise), and on a weighted decay of the intensity variation ΔT measured on previous consecutive thermal frames:

$$\varepsilon_t = \varepsilon_{MIN} + \min\left(1,\ \frac{\max\left(0, \sigma_t - \sigma_{MIN}\right)}{\sigma_{MAX} - \sigma_{MIN}} + \kappa \cdot \Delta\hat{T}_t\right) \cdot \left(\varepsilon_{MAX} - \varepsilon_{MIN}\right) \quad (2)$$

Parameters ε_{MIN}, ε_{MAX}, σ_{MIN} and σ_{MAX} must be determined by the user and correspond to the minimum and maximum allowed detection threshold and sensor noise. κ is a normalizing coefficient used to weight the contribution of the sudden intensity variations. We set κ to 10 for all our experiments.

Having to set the value of four parameters to adjust a single threshold might seem not justified. But throughout more than two years of interaction with an industrial partner using the algorithm, we found that the eq. (2) allows the user to optimize accurately and intuitively the desired detection rate for every application and installation.

[1] For more details on the *YCoCg* color space, please refer to [10].

4. QUANTITATIVE PERFORMANCE ANALYSIS

Measuring the performance of a background / foreground segmentation algorithm is not a trivial task. Several factors may limit the validity of the results: limited quantity and quality of benchmark sequences, inappropriate performance metrics, type of post-processing applied on the detection mask, and non-optimal adjustment of parameters of compared algorithms. In this quantitative performance analysis, a special attention has been addressed to each of these factors.

For more than two years, we extensively tested and analysed the behavior of the proposed algorithm on a huge amount of data. In this section, we present and discuss the results obtained on four sequences grabbed in our parking lot. These co-registered LWIR / color videos are listed in Table 1 and illustrated in Fig. 2. Note that to enhance details, we only display a sub-area of the whole frames. These image sequences with their ground-truth were made publicly available (with many others coregistered thermal-color videos) at www.ino.ca/Video-Analytics-Dataset.

Table 1. Selected videos with ground truth at pixel-level.

Sequence	Frames	Resolution [pixels]	Frames with GT	Compression
ClosePerson (CP)	240	512x184	20	MPEG4
MultipleDeposit (MD)	2400	448x324	15	MPEG4
GroupFight (GF)	1482	452x332	22	MPEG4
ParkingSnow (PS)	2941	448x324	21	MPEG4

Conditions expressed in diagram of Fig. 1 give a preliminary detection mask in which every pixel is classified as background, foreground or shadow. Such preliminary detection masks, with cast shadow pixels printed in gray, are illustrated by columns 3 and 5 of Fig. 2. Typically, some filtering is performed on these masks to remove noise prior to the blob labeling process. Columns 4 and 6 illustrate the enhanced detection masks obtained at the output of these filtering and blob labeling processes. Exactly the same cascade of operations is applied for every algorithm compared in Table 2. These operations consist in spatial filtering of candidate shadow pixels, morphological closure and blob labeling. We also integrated elimination of too small blobs and filling of holes (pixels printed in light gray in the columns 4 and 6 of Fig. 2) in the blob labeling process.

Thanks to the ground truth at pixel-level, the number of true positives (*TP*), false positives (*FP*) and false negatives (*FN*) may be determined for every algorithm and every video. Among existing metrics, we chose the Jaccard coefficient (*J*) used by Rosin and Ioannidis [4]:

$$J = \frac{TP}{(TP + FP + FN)} \qquad (3)$$

To present a more complete comparative study, we also report the detection rate and the false alarm rate:

$$DR = \frac{TP}{TP + FN} \qquad FAR = \frac{FP}{TP + FP} \qquad (4)$$

We present in Table 2 the results of our quantitative performance analysis for the four selected sequences and four different algorithms. For every algorithm, detection thresholds have been set to maximize the Jaccard coefficient. The optimization procedure used was an exhaustive search, and we performed it independently for every sequence. We choose to use a set of optimized parameters for every video instead of a unique set of parameters for all sequences to simulate more closely the fact that detection thresholds of a continuously operating system will be automatically adjusted in time based on illumination and scene conditions.

Table 2. Foreground detection accuracy, quantitative results.

Seq.	Metric	Thermal only	Color only	Hybrid combined	Hybrid independent
CP	***DR***	0.863	0.779	0.898	0.898
	FAR	0.161	0.185	0.123	0.123
	J	0.7405	0.6624	0.7974	0.7978
MD	***DR***	0.728	0.540	0.845	0.849
	FAR	0.207	0.145	0.211	0.225
	J	0.6113	0.4945	0.6892	0.6805
GF	***DR***	0.946	0.661	0.935	0.944
	FAR	0.203	0.206	0.097	0.104
	J	0.7620	0.5639	0.8500	0.8503
PS	***DR***	0.887	0.849	0.947	0.950
	FAR	0.077	0.083	0.059	0.061
	J	0.8255	0.7888	0.8942	0.8950

4.1. Hybrid vs single sensor

A first remark is that the proposed combination of thermal and color data leads to a more accurate detection (higher Jaccard coefficient) for all image sequences.

We can also note that the thermal only algorithm gives better results than the color only version, especially for videos *CP*, *MD* and *GF*. In these three sequences, the presence of dark cast shadows contributes to reduce the *DR* and increase the *FAR* of the color only algorithm.

4.2. Combined vs independent codebook

Thermal and color information may be combined into a single codeword as proposed by eq. (1). The thermal and color background may also be modeled into two independent codebooks. The performances measured with these two alternatives are very similar. A significant difference is only obtained on sequence *MD* (0.6892 vs 0.6805), which is the video with the more dynamic background (oscillating trees). With combined *CW*, a change has only to be detected on either the thermal or the color image to avoid updating the background model with a new observation. The use of combined *CW* is thus less prone to errors than the use of independent *CW* with such dynamic background.

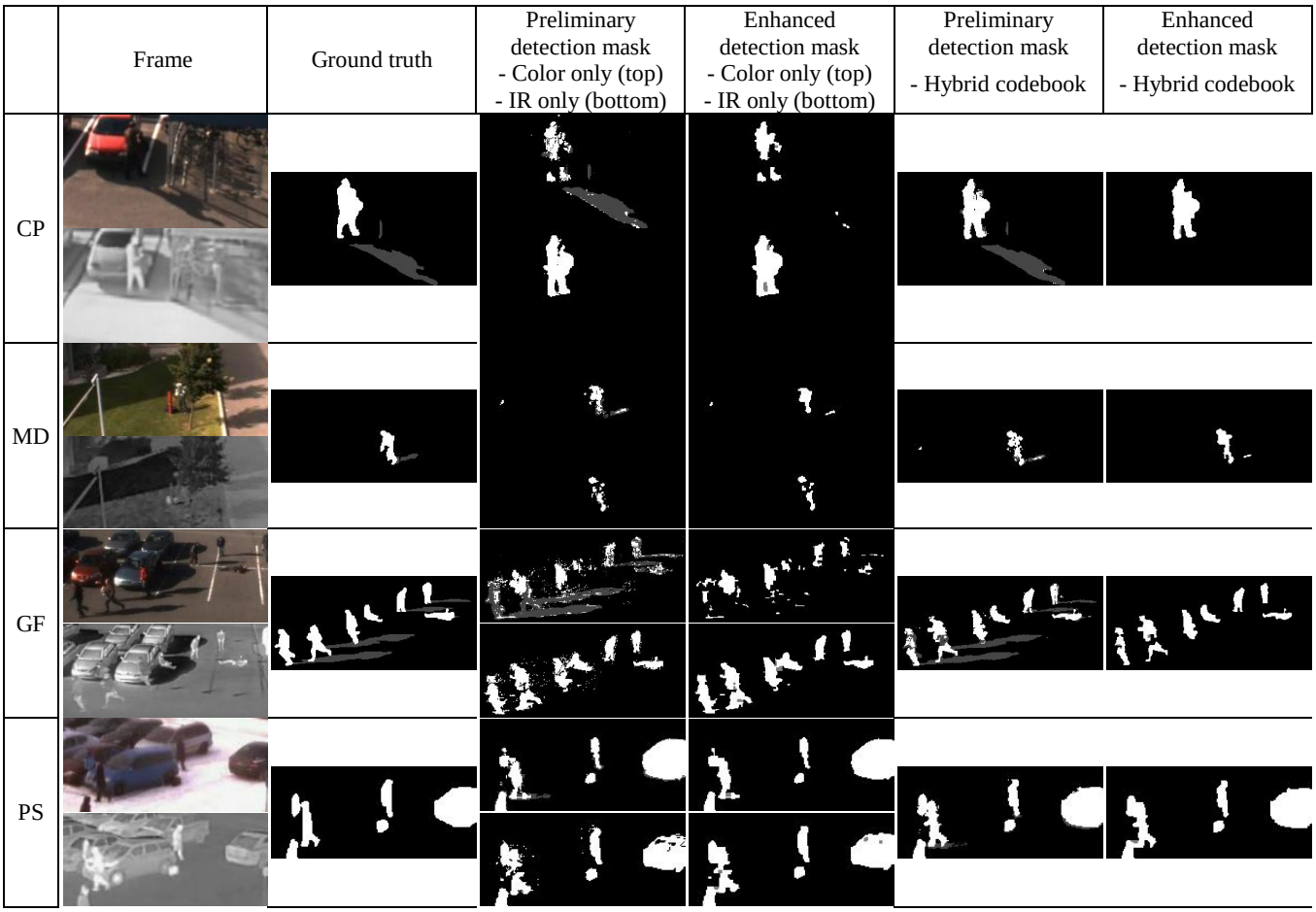

Fig. 2. Examples of preliminary detection masks and enhanced detection masks obtained with every algorithm.

4.3. Memory and processing time

For every video, we report in Table 3 the mean processing time per frame (including filtering and blob labeling processes) in detection mode (initialization mode requires a shorter processing time). Two interesting observations must be reported. First, the combined codebook is 14.3% faster than the independent in average. The redundancy of some operations with independent *CW* explained this difference. Second, the proposed combined algorithm is 32.3% faster, in average, than the summation of thermal and color only.

Table 3. Mean processing time per frame in ms.

Seq.	Thermal only	Color only	Hybrid combined	Hybrid independent
CP	4.64	8.22	9.73	11.16
MD	6.67	11.34	13.61	15.71
GF	6.16	10.57	12.13	13.74
PS	5.99	9.96	11.83	13.57

In terms of memory requirement, the proposed combined codebook requires, in average, 30.5% less memory than with independent codebooks, and 51.4% less memory than the summation of thermal and color only codebooks. This difference is another argument in favour of using a fusion technique at pixel-level rather than at object-level. For this analysis, we limited the permanent background codebook to a maximum of five codewords per pixel.

5. CONCLUSION

It is now accepted that thermal imaging is more suitable than electro-optical sensor for moving object detection and tracking in low-light conditions. However, during day time, when illumination allows contrasted color images, it is not obvious that the combination of thermal and visible information will lead to better results because the addition of weaknesses of both sensors can degrade performances. We demonstrated by the quantitative analysis of section 4 that the proposed fusion method improves the robustness of moving objects extraction in all tested scenarios.

The proposed detection technique, which combines thermal and color information at pixel-level, is suitable for real-time applications. It currently exploits only spectral information, but texture, gradient or spatial constraints could be added to improve segmentation accuracy.

6. REFERENCES

[1] K. Kim, T.H. Chalidabhongse, D. Harwood and L. Davis. Real-time foreground - background segmentation using codebook model", *Real-time Imaging*, vol. 11, no. 3, pp. 172-185, 2005.

[2] X. Maldague et al., "Infrared and Thermal Testing", *Nondestructive Testing Handbook*, vol. 3, 3rd edition, 2001.

[3] C. Ó Conaire, E. Cooke, N. O'Connor, N. Murphy and S. Smeaton, "Fusion of infrared and visible spectrum video for indoor surveillance", *Int. Workshop on Image Analysis for Multimedia Interactive Services*, 2005.

[4] P.L. Rosin and E. Ioannidis, "Evaluation of global image thresholding for change detection", *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2345-2356, 2003.

[5] D. Scribner, P. Warren and J. Schuler, "Extending color vision methods to bands beyond the visible", *IEEE Workshop on Computer Vision Beyond the Visible Spectrum, Methods and Applications*, 1999.

[6] L. Snidaro, G.L. Foresti, R. Niu and K. Varshney, "Sensor fusion for video surveillance", *Int. Conf. on Information Fusion*, 2004.

[7] C. Stauffer and E. Grimson, "Adaptive background mixture models for real-time tracking", *CVPR*, 1999.

[8] L. St-Laurent, D. Prévost and X. Maldague, "Combination of color and thermal sensors for enhanced object detection", *Int. Conf. on Information Fusion*, 2007.

[9] L. St-Laurent, D. Prévost and X. Maldague, "Fast and accurate calibration-based thermal / colour sensors registration", *QIRT*, 2010.

[10] L. St-Laurent, D. Prévost and X. Maldague, "Optimization of color-based foreground/background segmentation for outdoor scenes", *IASTED SPPRA*, 2012.

[11] A. Torabi, G. Massé and G.-A. Bilodeau, "Feedback scheme for thermal-visible video registration, sensor fusion, and people tracking", *CVPR Workshops*, 2010.

[12] A. Waxman, D. Fay, P. Ilardi, P. Arambel and J. Silver, "Active tracking of surface targets in fused video", *Int. Conf. on Information Fusion*, 2007.