http://dx.doi.org/10.21611/qirt.2006.065

Thermal imaging for enhanced foreground – background segmentation

by L. St-Laurent*[1], D. Prévost** and X. Maldague*

** Université Laval, Quebec City (Quebec), Canada*
*** National Optics Institute, Quebec City (Quebec), Canada*

Abstract

Foreground – background segmentation is the primary step of most automated video monitoring system aiming at object tracking, event detection or scene interpretation. In uncontrolled environments, with dynamic background and lighting changes, this basic task is very challenging. This work is based on the hypothesis that the combination of LWIR (8-12 μm) and colour cameras can significantly improve the robustness of foreground – background segmentation. An acquisition unit with co-aligned thermal and visible fields of view is used. Starting from a state-of-the-art algorithm for moving objects extraction in colour video, we adapted the method for processing of "*RGBT*" video format. Pros and cons of using thermal imagers in outdoor video monitoring applications are discussed. A preliminary objective performance evaluation of detection accuracy is also presented.

1. Introduction

Over the last decade, we saw the appearance of automated video monitoring in numerous applications. To make apart foreground and background pixels is the first step of most automated video monitoring system aiming at object tracking, event detection or scene interpretation. Still today, this basic task is challenging when the monitoring takes place in uncontrolled environments.

Most solutions proposed for moving object extraction are based on the visible spectrum. Actually, in brightly illuminated scenes, standard colour cameras provide the best information for object segmentation. However, in outdoor applications, darkness and other environmental conditions such as fog, rain and smoke strongly decrease the efficiency of standard cameras. In many applications, achievement of zero miss detection rate is a critical requirement and investment in more powerful imaging systems is justified. This opens the way to video systems combining thermal and colour cameras. This work is based on the hypothesis that the addition of LWIR cameras (8-12 μm) can significantly improve the robustness of foreground – background segmentation in uncontrolled environments.

Only recently, thermal imagery has been exploited in video monitoring applications, and especially for pedestrian detection [1][2][3][4]. Usually, the fact that human skin appears brighter than the background in long-wave thermal imagery is used as the main cue for the detection. However, this constraint is not always satisfied, especially in outdoor scenes. Resorting to others features, like edges or blobs characteristics, is often performed.

[1] : louis.st-laurent.1@ulaval.ca

http://dx.doi.org/10.21611/qirt.2006.065

A few research groups previously addressed the combination of thermal and colour images. As widely known in the field of image fusion, the combination of thermal and visible images is not trivial. For this reason, to combine thermal and colour information (analytical fusion) rather than performing image fusion (representative fusion) is more suitable for automated video monitoring. In [5], objects of interest are first extracted from infrared images along the hypothesis that pedestrian are warmer than background. Selected ROI are then used in both spectrums for contour extraction and fusion. In [6], moving objects are detected and tracked independently in each spectrum. An analysis of object's temporal persistence is used at every frame to select the more reliable sensor.

The work presented in this paper has the distinctive characteristic of combining information from both sensors at pixel level. More specifically, every pixel is classified as foreground or background along its similarity to thermal and colour background model. To provide the required image registration, we developed an approach to align both fields of view at hardware level. Such design can support applications requiring large effective depth-of-field, an advantageous characteristic comparatively to set-ups with parallel or convergent optical axis. As a starting point for background modelling, we selected a state-of-the-art technique developed for colour videos [7]. Our original contribution lies in the adaptation of their algorithm for the processing of augmented "*Red-Green-Blue-Thermal*" (*RGBT*) video format.

In the next section, we first discuss the pros and cons of using thermal imagers in outdoor video monitoring applications. The subject of section 3 is the co-aligned thermal / colour platform used for video acquisition. In section 4, we present the codebook background subtraction algorithm in its original form as proposed by [7]. Then, we describe how we integrated thermal information in the process. For performance evaluation, preliminary quantitative assessments were achieved in order to measure the accuracy improvement obtained from the combination of LWIR and visible spectrums. This is discussed in section 5.

2. Thermal imaging in uncontrolled environment

In this work, the term "uncontrolled environment" is employed to refer to outdoor scenes where illumination and temperature changes occur frequently, and where various atmospheric conditions can be observed. Specific environmental condition will not bear an equivalent impact on thermal and visible imaging because the properties of radiation propagation in the atmosphere vary greatly with wavelength range.

It worth to mention that the LWIR spectral range (8-12 μm) is more suitable than the MWIR range (3-5 μm) for video monitoring of human activity. The main reason is that emitted radiation from objects at ambient temperature (300 K) peaks in this long-wavelength range [8]. This affirmation can be derived from the Planck's law:

$$\lambda_d = \frac{2897.7}{T} \quad (1)$$

where λ_d is the dominant wavelength, 2897.7 is the third radiation constant in μm*K and *T* is the object temperature in Kelvin. From this relation, a human body, at 37°C, have a dominant wavelength of 2897.7 / (37+273.15) = 9.3 μm. Another motivation to select LWIR sensor rather than MWIR is that, along some studies [9][10], the impact of absorption and scattering of fog is less severe in the LWIR waveband.

http://dx.doi.org/10.21611/qirt.2006.065

2.1. Illumination and temperature changes

Obviously, visible cameras are more sensible than LWIR ones to illumination changes. However, since the sunlight warms up exposed surfaces, long-term illumination changes also affect thermal images. And sunlight does not equally warm up all objects of the scene. Those with high thermal inertia, like water, trees and people, will present little changes comparatively to cars or building, which have low thermal inertia. Moreover, dark surfaces with low reflective properties, as asphalt, will be more strongly affected by sun's rays, than brighter objects. It is why a scene observed at 4PM and 4AM appears so differently. In the first case, asphalt and other inanimate objects have been heated all day long and appear warmer than people.

Since the ability to make apart foreground from background is basically related to image contrast, we can state from the preceding remarks that people detection in thermal images will usually be easier during cold and overcast days or during very hot and sunny days than during the middle window where background objects and people appears with the same intensity. From a similar analysis, we can note that accurate extraction of vehicles in infrared imagery is more challenging because they quickly adapt to environment temperature.

2.2. Atmospheric conditions

A first factor that has significant impact on thermal images contrast is the wind, which accelerates warm objects cool down. Hence, the background will tend to have a more uniform temperature and less contrast during strong wind days. This is demonstrated by comparing left and middle column images of figure 1.

Fog, rain and snow are others climate factors that affect thermal wavelength transmission, thus decreasing thermal image contrast. However, along some studies [9][10], and based on our own experience, absorption and scattering in LWIR range seem slightly inferior than in visible range. This can be noticed on right column images of figure 1, where a building can be distinguished in the background of the thermal image, but not in the colour one.

Fig. 1. *LWIR and visible images during slight overcast (left), strong wind and rain (middle) and snow (right).*

http://dx.doi.org/10.21611/qirt.2006.065

3. Image registration

The main distinctive characteristic of our acquisition unit is the hardware registration of thermal and colour images. It provides the major advantage over systems with parallel or convergent optical axis that image registration is valid for a wide depth of field.

We use a glass beamsplitter with ITO coating (Indium-Tin-Oxide). This component, shown in figure 2, reflects 93% of 8-12 µm emitted energy and transmits about 85% of visible waves. The LWIR camera used is the ThermoVision A10, which has a VO_2-based microbolometer detector array of 164 x 128 pixels and a sensitivity of about 80 mK. The Marlin F33C CCD camera (640 x 480 pixels with Bayer filter) is actually used as colour sensor.

To compensate for the distinct cameras resolutions and for the slightly different sensors form factors, the colour image is decimated and the thermal image is interpolated to produce resulting *RGBT* images of 328 x 254 pixels. The bilinear interpolation of the thermal image also acts as high frequency filter, thus removing part of the image noise. Figure 3 illustrates the achieved quality of the registration of thermal and colour images. For visualization purposes, the red channel of the colour image as been replaced by the scaled thermal image.

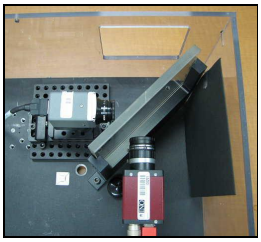


Fig. 2. *Co-aligned LWIR / colour acquisition platform*

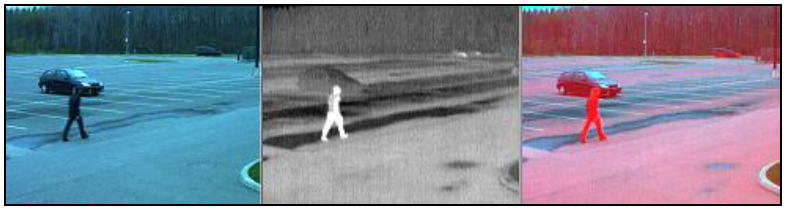


Fig. 3. *Visualization of image registration*

4. Objects extraction

Since moving objects are usually the objects of interest in video monitoring applications, motion-based segmentation techniques are the most suitable for objects extraction. In this work, a still acquisition platform is assumed, thus no camera motion compensation is needed. For this situation, the most efficient objects extraction methods, both in term of accuracy and processing time, are those based on background model subtraction.

http://dx.doi.org/10.21611/qirt.2006.065

4.4. State-of-the-art object extraction methods in colour video

In the last decade, many efficient background modelling and maintenance approaches have been proposed for videos from the visible spectral range. Two of them have brought most attention and have been selected and improved by numerous research groups. The first is the Gaussian Mixture Model of [11] and the second is the non-parametric kernel-based density estimator of [12]. Their interest comes from the fact that they allow modelling of multi-modal background (moving background), like those met in outdoor scenes. They also support the presence of foreground objects in the scene during the background model initialization phase. Recently, a new non-parametric technique inspired from [12] but based on Codebook model [7] has been presented. This one is particularly well suited for real-time applications. Indeed, they demonstrated in their performance analysis that this new method is about three times faster than [11] and [12], with slightly better segmentation quality. It is this algorithm that we selected as a starting point for our "*RGBT*" objects extraction technique.

In short, a quantization/clustering technique is used to generate a compressed form of background model. Every pixel is represented by *L* codewords defined by:

$$CW_{k=1\ldots L} = \left\{ \bar{R}_k, \bar{G}_k, \bar{B}_k, I_k^{\max}, I_k^{\min}, f_k, p_k, q_k, MNRL_k \right\} \quad (2)$$

$\bar{R}_k, \bar{G}_k$, and $\bar{B}_k$ are red, green and blue values, $I_k^{\max}$ and $I_k^{\min}$ are maximum and minimum observed brightness of codeword *k*, and *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. The *MNRL* (*Maximum Negative Run-Length*) is the key parameter. It stores the length of the period (in number of frames) during which a codeword has not been matched. Based on the hypothesis that background pixels are observed periodically, codewords belonging to background will have a small *MNRL* value while codewords generated by the temporary presence of an object of interest will have large *MNRL*. Thus, a threshold on *MNRL* parameter can be used to filter out codewords belonging to objects of interest.

For every new frame, every pixel is associated to the first sufficiently similar codeword in regard to its colour and brightness. 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 are deleted. Typically, mono-modal scene areas will be modelled by only one codeword, while multi-modal areas, like swaying trees, will need more codewords. In our implementation, we limited the permanent background codebook to a maximum of 10 codewords per pixel.

The background model maintenance is achieved by updating red, green and blue values of **the matched codeword** via a weighted average. Here is how the red value of pixel *x* is determined for frame *n*+1 :

$$\bar{R}_{k,n+1}(x) = \frac{\alpha * \bar{R}_{k,n}(x) + R_n(x)}{\alpha + 1} \quad (3)$$

http://dx.doi.org/10.21611/qirt.2006.065

The α coefficient defines the learning rate. A smaller α will bring a faster codeword adaptation, thus leading to a lower rate of false detection when quick illumination changes occur. For more details on specific aspects of the algorithm, as how colour thresholds are defined, please see the original description in [7].

4.2. Integration of thermal information

Thanks to the alignment of thermal and colour images provided by our acquisition platform, we can directly combine information at pixel level. More specifically, we added a parameter $\overline{T_k}$, for thermal intensity, in the codeword representation of every pixel (Eq. (2)). Like for colour parameters, thermal intensity of matched codeword is updated by the weighted average of Eq. (3).

In order to associate a *RGBT* pixel to codeword *k*, a condition on its thermal intensity must also be satisfied :

$$\left|T(x) - \overline{T_k}\right| \leq 5 * \sigma_T \tag{4}$$

where σ_T is the global standard deviation observed on the 10 first thermal frames of the sequence. Contrarily to colour images, we determine gain and offset adjustment for thermal images at the beginning of each sequence. Thus, a higher gain is applied to thermal images of videos grabbed from scenes with poor thermal contrast, leading to higher noise. It is why we decided to relate the thermal intensity threshold to standard deviation.

With our combination of thermal and colour conditions, a pixel is classified as foreground if either its colour **OR** thermal value differs from background codewords. The detection mask obtained leads to a minimum miss detection rate, which is generally a very important requirement for video monitoring applications.

5. Performance analysis

5.1. Accuracy analysis

In this preliminary assessment of our combination approach, we compared the detection accuracy with the one achieved if only visible or thermal information is exploited. For this purpose, we used the *Detection Rate* (*DR)* and *False Alarm Rate* (*FAR*) metrics proposed in [13] :

$$DR = \frac{TP}{TP + FN} \tag{5}$$

$$FAR = \frac{FP}{TP + FP} \tag{6}$$

where *TP*, *FN* and *FP* hold for *True Positive*, *False Negative* and *False Positive* detections. For example, a pixel wrongly classified as foreground (object of interest) is a *False Positive* detection.

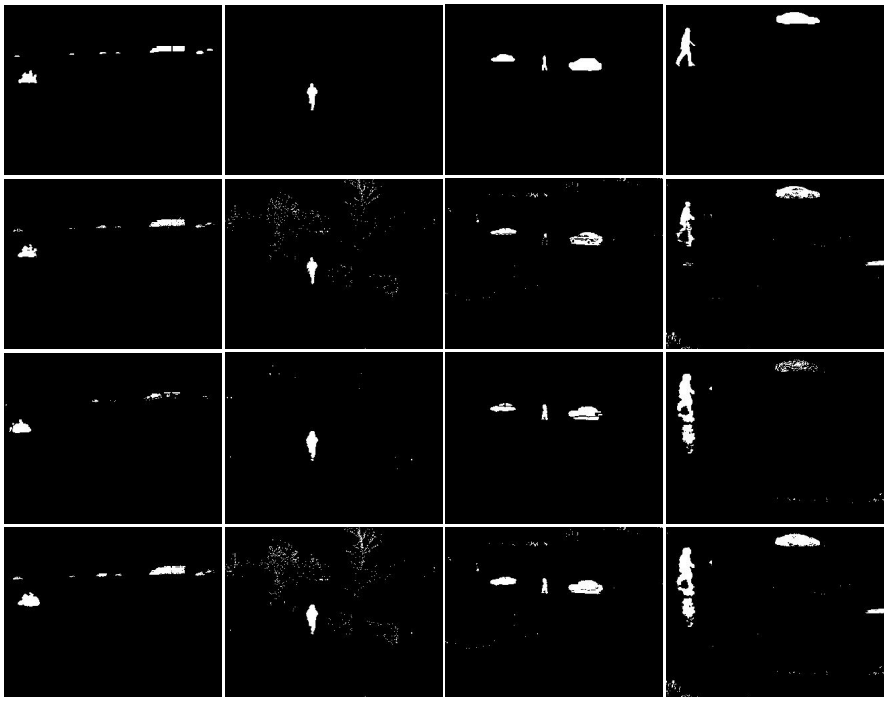
Obviously, these metrics require the availability of references images, commonly referred as *Ground Truth*, in which the correct foreground – background classification is defined. Manual generation of these reference images is a long and laborious task. Thus, the accuracy assessment as been performed on a limited number of four video sequences, and on a sample of 10 regularly space frames in

http://dx.doi.org/10.21611/qirt.2006.065

each of these videos. Test sequences have been selected in order to represent different observation conditions. They are illustrated in figure 4 with examples of *Ground Truth* and segmentation results in figure 5.

Video *A* Video *B* Video *C* Video *D*

Fig. 4. *Video A : sunny afternoon; B : sunrise; C : sunset; D : cloudy and rainy.*



Video *A* Video *B* Video *C* Video *D*

Fig. 5. *Examples of Ground Truth (first row) and detection results from colour (second row), thermal (third row), and from RGBT images (fourth row).*

http://dx.doi.org/10.21611/qirt.2006.065

Table 1 presents cumulated *Detection Rate* and *False Alarm Rate*. For all test sequences, we empirically fixed the learning rate α to 7. It is important to understand that since only the matched codewords are updated, a fast learning rate can be used without causing early integration of motionless objects in the background model. We also fixed the training period to 200 frames, which represent about 25 seconds with our acquisition rate of 7.5 fps.

Table 1. *Comparison of segmentation accuracy.*

Video format	Video *A*		Video *B*		Video *C*		Video *D*	
	DR	***FAR***	***DR***	***FAR***	***DR***	***FAR***	***DR***	***FAR***
RGB	0.82	0.07	0.88	0.78	0.66	0.26	0.94	0.38
Thermal	0.56	0.15	0.89	0.43	0.84	0.19	0.78	0.48
RGBT	0.91	0.17	0.97	0.83	0.94	0.27	0.99	0.45

From results of table 1, we first notice that, for all videos, *Detection Rate* obtained with the combination of thermal and colour information is higher than with any sensor used alone. About the comparison of detection accuracy between thermal and visible videos, we can note that a better *DR* was achieved with colour images for videos *A* and *D.* In sequence *B*, the only moving objects are pedestrians, which are generally characterized by a good thermal contrast. It is why a high thermal *DR* of 89% was measured with video *B*. For sequence *C*, grabbed at sunset, the thermal sensor outperforms the visible one as predicted because of the low illumination.

Similarly to *DR*, *False Alarm Rate* obtained with *RGBT* video format is higher than the *FAR* measured for each sensor alone. This error accumulation comes from the use of a simple logical OR for the combination of thermal and visible detections.

A very high *FAR* of 83% was measured for video B. The main reason behind this is that many bushes and trees are present in the background and a quite strong wind was blowing during the acquisition. Moreover, the small size of the pedestrian in video *B* also contributes to this high *FAR*. When objects of interest are very small, the number of *True Positive* detections will be also very small (*TP* of Eq. (6)), thus leading to high FAR.

We can also remark that *FAR* measured for video *D*, grabbed during rain, are significantly larger than for video *A*. The visible and thermal reflections on wet surfaces are responsible of these numerous false detections, which can be clearly observed in detection masks of figure 5.

Another aspect that must be considered in the interpretation of *FAR* is the fact that no image enhancement has been performed on detection masks for results presented in table 1. For comparison purpose, table 2 indicates *DR* and *FAR* measured after morphological closure and elimination of blobs smaller than 30 pixels.

Table 2. *Segmentation accuracy after mask enhancement.*

Video format	Video *A*		Video *B*		Video *C*		Video *D*	
	DR	***FAR***	***DR***	***FAR***	***DR***	***FAR***	***DR***	***FAR***
RGB	0.81	0.05	0.88	0.47	0.77	0.14	0.91	0.38
Thermal	0.61	0.14	0.92	0.32	0.92	0.19	0.81	0.47
RGBT	0.91	0.16	0.97	0.69	0.97	0.20	0.94	0.45

http://dx.doi.org/10.21611/qirt.2006.065

5.2 Processing time analysis

As mentioned in the introduction, an important requirement for automated video monitoring applications is real-time performance, and the foreground – background segmentation module is generally the processing time bottleneck of that kind of system. In this work, a particular attention has been addressed to this aspect and it is precisely for this reason that the Codebook background model has been selected as a starting point for the foreground – background segmentation algorithm instead of a Gaussian Mixture Model or other non parametric approach.

Table 3 shows the mean processing time measured on each test video sequence. The processor used was a Pentium IV 3 GHz, with 1 GB of RAM. Image resolution is 328 x 254 pixels. Results of table 5 do not include detection mask enhancement. Up to 2 ms per frame can be added for the morphological closure and small blobs deletion.

Table 3. *Comparison of processing time, in ms per frame.*

Video format	Video sequence			
	A	***B***	***C***	***D***
RGB	10.1	12.4	10.2	10.0
Thermal	2.1	2.3	2.0	2.0
RGBT	11.1	13.4	11.0	10.7

We can notice that the processing time of every video is quite similar, except for sequence *B*, which is slightly longer. The high background instability caused by swaying bushes and trees is responsible of this observation. Table 3 also illustrates the interesting fact that the processing time for the *RGBT* format is smaller than the summation of the processing time required for *RGB* and *Thermal* formats individually.

Extending the results to larger images of 640 x 480 pixels would lead to a processing time of about 45 ms per frame. This corresponds to 22 fps, which is largely acceptable for real-time video monitoring applications.

6. Conclusion

Despite the typically poorer resolution and higher noise level of uncooled thermal cameras comparatively to visible spectrum sensors, the addition of thermal information improves sensitivity of detection of foreground – background segmentation algorithms for outdoor video sequences. The proposed approach is based on a state-of-the-art background model, but optimized for *RGBT* video sequences grabbed with our co-aligned LWIR / colour acquisition platform. As demonstrated by the processing time analysis presented in section 5.2, our solution is suitable for real-time video monitoring applications.

The work presented in this paper is the first step through a real-time object tracking system efficient in outdoor environment all day and all year long. The next improvements related to foreground – background aspects will address minimization of false detections. As mentioned in section 5.1, the implemented algorithm leads to a very good detection rate, but also to a large false alarm rate that could be decreased by the analysis detection mask properties. Identification of *False Positive* pixels caused by reflections on wet surfaces is an example.

http://dx.doi.org/10.21611/qirt.2006.065

Performance evaluation of a new foreground – background segmentation algorithm is essential. As mentioned in the previous section, manual generation of *Ground Truth* images is a laborious task, and the particular *RGBT* image format used in our work doesn't allow us to use publicly distributed videos with provided *Ground Truth*. Analysis of others evaluation methods not requiring the availability of references images is planned for our future development phases. This would also enable us to quantitatively evaluate the accuracy detection on a more representative sample volume.

REFERENCES

[1] F. Xu and K. Fujimura. Pedestrian detection and tracking with night vision. In *Proc. IEEE Intelligent Vehicles Symposium*, vol. 1, Paris, France (2002) 21-30.

[2] M. Bertozzi, E. Binelle, A. Broggi and M. Del Rose. Stereo vision-based approaches for pedestrian detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition - Workshops*, vol. 3, San Diego, USA (2005).

[3] C. Dai, Y. Zheng and X. Li. Layer representation for pedestrian detection and tracking in infrared imagery, In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition - Workshops*, vol. 3, San Diego, USA (2005).

[4] Y. Fang, K. Yamada, Y. Ninomiya, B. Horn and I. Masaki. Comparison between infrared-image-based and visible-image-based approaches for pedestrian detection, *IEEE Intelligent Vehicles Symposium*, Columbus, USA (2003) 505-510.

[5] J.W. Davis and V. Sharma. Fusion-based background-subtraction using contour saliency, In Proc. IEEE Conf. on Computer Vision and Pattern Recognition - Workshops, vol. 3, San Diego, USA (2005).

[6] H. Toressan, B. Turgeon, C. Ibarra-Castanedo, P. Hébert and X. Maldague. "Advanced surveillance systems : Combining video and thermal imagery for pedestrian detection", SPIE vol. 5405, *Thermosense XXVI* (2004) 506-515.

[7] 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 (2005) 172-185.

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

[9] D.B. Rensch and R.K. Long. Comparative studies of extinction and backscattering by aerosols, fog, and rain at 10.6µ and 0.63µ, *Applied Optics*, vol. 9, no.7 (1970).

[10] M.A. Naboulsi, H. Sizun and F. de Fornel. Fog attenuation prediction for optical and infrared waves, *Opt. Eng.*, vol. 43, no. 2 (2004) 319-329.

[11] C. Stauffer and E. Grimson, Adaptive background mixture models for real-time tracking, In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 8, Fort Collins, USA (1999) 252-258.

[12] A. Elgammal, R. Duraiswami, D. Harwood and L.S. Davis, Background and foreground modeling using nonparametric kernel density estimation for visual surveillance, *Proc. of the IEEE*, vol. 90, no. 7 (2002) 1151-1163.

[13] G. Medioni. Detecting and tracking moving objects for video surveillance, In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, Fort Collins, USA (1999) 319-325.