Pedestrian Detection and Tracking in Thermal Images from Aerial MPEG Videos

Conference Paper · January 2018

CITATIONS
0

READS
6

5 authors, including:

Rob Haelterman
Royal Military Academy
69 PUBLICATIONS **310** CITATIONS
SEE PROFILE

Zied Chtourou
School of Aeronautical Specialties. Sfax. Tuni…
40 PUBLICATIONS **125** CITATIONS
SEE PROFILE

Geert De Cubber
Royal Military Academy
82 PUBLICATIONS **366** CITATIONS
SEE PROFILE

Rabah Attia
Ecole Polytechnique de Tunisie
128 PUBLICATIONS **185** CITATIONS
SEE PROFILE

Pedestrian Detection and Tracking in Thermal Images from Aerial MPEG videos

Ichraf Lahouli[1,2,3], Robby Haelterman[1], Zied Chtourou[3], Geert De Cubber[1] and Rabah Attia[2]
[1]*Royal Military Academy, Belgium*
[2]*Tunisia Polytechnic School, Tunisia*
[3]*Military Academy of Tunisia, Tunisia*
ichraf.lahouli@rma.ac.be

Keywords: PEDESTRIAN DETECTION, TRACKING, UAV, MPEG MOTION VECTORS, H.264

Abstract: Video surveillance for security and intelligence purposes has been a precious tool as long as the technology has been available but is computationally heavy. In this paper, we present a fast and efficient framework for pedestrian detection and tracking using thermal images. It is designed for automatic surveillance applications in an outdoor environment like preventing border intrusions or attacks on sensitive facilities using image and video processing techniques implemented on-board of Unmanned Aerial Vehicles (UAV)s. The proposed framework exploits raw H.264 compressed video streams with limited computational overhead. Our work is driven by the fact that Motion Vectors (MV) are an integral part of any video compression technique, by day and night capabilities of thermal sensors and by the distinguished thermal signature of humans. Six different scenarios were carried out and filmed using a thermal camera in order to simulate suspicious events. The obtained results show the effectiveness of the proposed framework and its low computational requirements which make it adequate for on-board processing and real-time applications.

1 INTRODUCTION

Pedestrian detection and tracking using thermal imagery is a widely studied field for surveillance purposes. Despite the lack of color and texture information, the use of such images allows working on both day and night-time. Classical image processing techniques have been used to detect the presence of pedestrians in both still and moving images, mainly using stationary cameras. These cameras need powerful back-end computers and/or networks on which the heavy lifting is done. When the problem is shifted to low-performance processing platforms, the applicability of this approach becomes hard. Indeed, new platforms (e.g UAVs or wireless sensors networks) are able to stream, in real time, videos captured by optical or thermal sensors. However, the amount of information transmitted is huge causing more power consumption due to video transmission which affects the mission duration and also causing problems of analysis and storage mostly because of unimportant information or redundancy. This paper proposes an efficient framework to detect and track pedestrians in thermal images for automatic surveillance purposes with low processing requirements. The ROI detection process is based on saliency maps in conjunction with a contrast enhancement technique as a first step to extract Regions Of Interest (ROI)s. Then, the Discrete Chebychev Moments (DCM)s (Karakasis et al., 2014) are used as a global image content descriptor. Finally, a classification step is ensured by a support Vector Machine (SVM) to distinguish between pedestrians and non pedestrians. The proposed ROI detector is evaluated using two public thermal pedestrian datasets: the OTCBVS benchmark -OSU Thermal Pedestrian Database (Davis and Keck, 2005) and the nine thermal videos taken from the LITIV2012 dataset (Torabi et al., 2012). In these two datasets, humans are taken from a relatively high altitude which can simulate images taken from a low altitude UAV. The performance of the proposed ROI detector is compared to the Maximally Stable Extremal Regions (MSER) detector (Matas et al., 2004) in terms of true detections, false positives and calculation time. MSER is a fast and widely used region based detector. Even though it was introduced in 2004, MSER is still up to date and used in many works as a region-based local extractor like recently in (Tun et al., 2017; Sun et al., 2017; Kumar and Gupta, 2017; Khosravi and Hassanpour, 2017;

Alyammahi et al., 2017; Śluzek, 2017). Its popularity is due to its efficiency to extract salient regions and its low complexity which makes it adequate for real-time applications and low-cost embedded systems. The obtained results of comparison betwenn MSER and the proposed ROI detector prove the robustness of the proposed method in terms of true detection rate and its superiority in terms of reducing false alarms and processing time.
Furthermore, in order to test the proposed ROI tracker in the context of outdoor surveillance, we generated our own dataset by carrying out six different scenarios of suspicious events and filmed the scene using a thermal camera. The tracking process is based on the MPEG MVs corresponding to the extracted ROIs. In fact, the different bounding boxes are tracked through time and their intermediate estimated positions are computed. However, we can not rely exclusively on the MPEG MVs due to the estimation errors generated by the codec. In order to compensate these errors, the proposed ROI detector is launched at a re-detection rate to update the positions and correct the small drifts. The proposed framework does not need frame by frame, neither pixel by pixel processing like in (Ma et al., 2016). It relies on some frames for the ROI detection and on some MVs already computed (as an integral part of the MPEG4 AVC (H264 codec)) for tracking, which makes it suitable for real-time applications with low-end computational platforms.

The paper is organized as follows: In Section 2 we review related state of the art works in motion-based segmentation and tracking. In Section 3, the proposed framework, composed of the ROI detector and the ROI tracker, is explained in detail. Experiments and results are presented in Section 4. We start by presenting the different datasets and by setting the re-detection rate. After that, we demonstrate the effectiveness of the proposed framework by presenting the results of comparison between the proposed detector and MSER in terms of accuracy and time consumption and the performance of the tracking process. Finally, Section 5 concludes the present work and exposes our perspectives for the future steps.

2 RELATED WORKS

The majority of the works in motion-based segmentation and tracking commonly uses Optical Flow (OF) and local feature descriptors such as SIFT (Uemura et al., 2008) or SURF (Bay et al., 2008) like recently in (Zhang et al., 2017; Sundari and Manikandan, 2017; Tun et al., 2017). For example, Wang and Schmid worked on action recognition using improved trajectories (Wang and Schmid, 2013). They estimated camera motion by matching feature points using dense OF and SURF descriptors. Then, they removed the corresponding trajectories to compensate the camera motion(they relied on a human detector to remove inconsistent matches). Wu et al. used the dense particle trajectories of the objects based on OF and proposed an optimisation method to distinguish between the trajectories of moving objects and those due to the camera motion (Wu et al., 2011). Nevertheless, OF, whether sparse or dense, is time-consuming and computationally heavy which limits the speed of feature extraction and makes it inadequate for real-time applications and challenges the mission autonomy (duration of the UAV flight). Some works studied the performance of MVs as an alternative to OF for tracking, action recognition and surveillance purposes. Among the first few who worked in this field, are Park et al. in 2003, who proposed a tracking scheme of an object in MPEG compressed domain (Park and Lee, 2003). They estimated the camera motion using a generalised Hough transform and then tracked the centre of the ROI based on the spatial distribution of colors. In 2004, Babu et al. used motion vectors of compressed MPEG video for segmentation (Babu et al., 2004) then proposed MPEG MV based features along with a Hidden Markov Model (HMM) modelling and motion history information for action recognition (Babu and Ramakrishnan, 2004). In 2006, Yeo et al. made use of MV information to capture the salient features of actions which have independent appearances. They then computed frame-to-frame motion similarity with a measure that takes into account differences in both MV's orientation and magnitude (Yeo et al., 2006). Aggarwal et al. proposed a scheme for object tracking using background subtraction and motion estimation in MPEG videos (Aggarwal et al., 2006). However, their method is mainly concerned with video surveillance applications where the cameras are fixed, which makes it not suitable for implementation on a moving platform like an UAV. Furthermore, the selection of the targets is not automatic. The object to track is marked by the user and not the result of any detection algorithm. In 2013, Biswas et al. proposed an approach to classify H.264 compressed videos, by capturing orientation information from the motion vectors (Biswas and Babu, 2013). They compute Histogram of Oriented Motion Vectors (HOMV) in order to define the motion characteristics of space-time cubes that partially overlap. They then used Bag of Features approach (BOF) for classification. Käs and Nicolas present an

approach to trajectory estimation of moving objects using the H264 MVs (Käs and Nicolas, 2009). Their method consists in performing a Global Motion Estimation (GME) based on the MVs extracted from the compressed stream. The generated outlier masks are the input for an object detection stage, followed by an object matching stage in order to estimate the trajectories in the scene. However, the main drawback of their method is that it can not deal with non moving people since the first step of their flowchart is the GME.

In 2014, Kantorov et al. performed activity recognition by computing Histograms of Optical Flow (HOF) and Motion Boundary Histograms (MBH) using the MPEG MVs as local descriptors, Fisher Vector (FV) for coding and Support Vector Machine (SVM) for classification (Kantorov and Laptev, 2014). They made a comparison with the OF and showed that the use of MPEG MVs showed a significant computational speedup ($\simeq$66%) with a small reduction of recognition accuracy. Recently, Zhang et al. proposed a real-time action recognition method using MVs extracted directly during the decoding process instead of OF. In order to boost the recognition performance, they adapted the models of OF Convolutional Neural Network (CNN) to MV CNN (Zhang et al., 2016). Poularakis et al. proposed an efficient motion estimation method for fast recognition of activities of daily living. They replaced OF calculation with block matching randomly initialized or based on the pre-computed MPEG MV (Poularakis et al., 2017). Only data in Motion Boundary Activity Area (MBAA)(Avgerinakis et al., 2013) are analysed which means that full video decoding is not necessary. These works are mostly dedicated for activity recognition by describing short actions and not for tracking people in long videos. In addition, none of them worked with thermal images.

In this work, we consider detecting and tracking pedestrians in UAV videos using thermal cameras for day and night surveillance. We mainly focus on the computational complexity and we use already available MPEG MVs which makes it suitable for low-performance processing algorithms and real-time applications.

3 PROPOSED METHODOLOGY

The proposed framework is mainly based on two algorithms: the ROI detector and the ROI tracker. The first one corresponds to the feature extractor which in our case are pedestrians. Indeed, it consists of a human detector based on saliency maps in conjunction with contrast enhancement techniques, while the discrete Chebyshev moments are used as a global image content descriptor and a linear support vector machine (SVM) is used as a classifier. The second algorithm aims to extract the MVs of the ROIs, drawn directly from the MPEG compressed video. The aforementioned algorithms are combined to offer an efficient framework for pedestrian detection and tracking with low computational cost. It can be integrated within an H.264 codec as it relies on the intermediary data that is used to produce the output stream. We will present consecutively the two algorithms in detail in the following paragraphs.

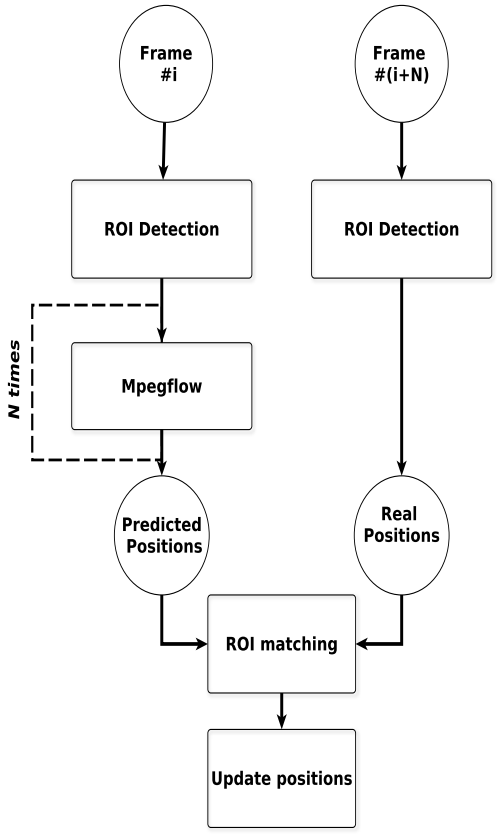


Figure 1: Flowchart of the proposed methodology
i corresponds to the index of the reference frame and *N* corresponds to the re-detection rate (in frames) (see subsection 4.2 for details)

3.1 Proposed ROI Detector

In order to detect pedestrians in thermal images, we extract hot spots assuming that a human is brighter than his background, which is usually suitable for outdoor scenes. We detect ROIs taking into account certain restrictions regarding brightness and target size.

The pedestrian detection part is composed of the following three steps:

1. ROI extraction: A fusion between a saliency map (produced on the basis of Lab colour space) (Achanta et al., 2009) and a wavelet-based contrast enhancement technique (Arodź et al., 2006),
2. Shape description: A global region content descriptor based on the DCMs (Karakasis et al., 2014) up to order 4*4,
3. Classification: SVM in order to distinguish between human and non-human blobs.

Initially, a Gaussian filter is applied to the input image. The resulting output is fed to the saliency map module and to the contrast enhancement module. Then, the two outputs are normalized and fused together using the geometric mean. Finally, the result is converted into a binary image by keeping only hot spot areas, which are further filtered using a size threshold to discard very small/large ROIs.
In this work, the saliency-based map is created fast enough and successfully highlights the included hot spots. The proposed method is kept intentionally simple enough in order to combine efficiency and calculation speed. We further enhance the results of the saliency map by fusing its output with a contrast-enhanced image. Another advantage of the proposed method for ROI selection is that it darkens the surrounding background of hot spots and at the same time highlights them, perfectly preserving their shape. This is very important, since, the shape of objects is used by the Discrete Chebyshev Moment-based descriptor in order to further recognize human objects using a linear SVM. In subsection 4.3, we will present the results obtained after applying this algorithm on the OSU Thermal Pedestrian Database (Davis and Keck, 2005) and the LITIV2012 dataset (Torabi et al., 2012).

3.2 Proposed ROI Tracker

The algorithms of ROI detection and MPEG MV extraction are combined together to ensure the detection and the tracking of pedestrians in thermal images with low computational overhead. Initially, a reference frame is selected as an input image **(Frame #i)**. The detection algorithm is applied to extract the ROIs (pedestrians). The resulting outputs are fed to the MPEG MV extractor module which extracts the MPEG MVs corresponding to the initial ROIs' bounding boxes. Actually, it starts by finding the macroblocks that cover each ROI (H264 codec splits each frame in different macroblocs of $16 * 16$ pixels in our case). Then, during the re-detection period (N times), the algorithm keeps tracking the macroblocs by extracting their relative MPEG MVs and computing the intermediate estimated positions. The figure 1 graphically illustrates the aforementioned process. Driven by the fact that MVs are already computed as an integral part of the H264 codec, we gain a lot in terms of computational cost. Indeed, during this period, we avoid the cost of the frame by frame process since we don't require the image processing techniques (contrast enhancement + saliency map) to detect the pedestrians. However, we cannot rely exclusively on the MPEG MVs due to the errors generated by the estimation steps in the H.264 standard. The proposed solution is to call the aforementioned ROI detector at a re-detection rate N in order to compensate these errors before pursuing the tracking. The choice of N depends on different parameters such as the frame rate and the resolution of the compressed video. An analysis in section 4.2 shows how N is set.

4 EXPERIMENTS & RESULTS

4.1 Presentation of the Different Datasets

In order to validate the proposed ROI detector, we used two different public datasets in outdoor urban environment. Firstly, the OSU Thermal Pedestrian Database (Davis and Keck, 2005), acquired by the Raytheon 300D thermal sensor. It is composed of 10 test collections with a total of 284 frames taken within one minute but not temporally uniformly sampled. The OSU thermal dataset covers a panoply of environmental conditions such as sunny, rainy and cloudy days. Secondly, the nine thermal sequences taken from the LITIV2012 dataset (Torabi et al., 2012). Actually, the dataset is composed of nine pairs of visualthermal sequences.
The public datasets, used to validate the proposed ROI detector, are collections of frames that are not temporally uniformly sampled and don't provide H.264 encoding data which make these datasets not adequate to validate the proposed ROI tracker. In addition, they present pedestrians walking 'normally' in the street. However, our main purpose is to detect suspicious events of pedestrians taken from a thermal sensor onboard of an aerial platform and thus based on the analysis of their trajectories and velocities. Therefore, we have generated our own dataset by carrying out some specific scenarios of suspicious events in an outdoor environment. Video sequences were shot using an MPEG thermal camera.

The different scenarios of abnormal events could be described as follow:

1. *Brutal Turn Back*: two people move slowly in one direction simulating two policemen. A suspicious person walks in the opposite direction and will quickly turn back as soon as he sees them.
2. *Convergence/Divergence*: three people converge at a specific point. They quickly exchange a suspicious object and then diverge and quit the scene.
3. *Velocity Changes*: one person alternates between walking and running.
4. *Occlusion/Non Occlusion*: one person tries to hide behind a car.
5. *Circular Trajectory*: one person moves around a car while focusing on it as if he has some robbery intention.
6. *Rapid Dump of Suspicious Object*: one person is walking and carrying a backpack. When he reached a specific vehicle, he quickly throws his backpack down and continues walking.

The thermal videos were shot using a stationary camera from a relatively high altitude (to simulate an oblique view of an UAV) with a frame rate of 25 frames per second (fps) and an image resolution of 576*704 pixels.

4.2 Setting of the Re-detection Rate *N*

As mentioned before, we can not rely only on the MPEG MVs to track the ROIs. Indeed, a re-detection is used at a specific rate N to avoid the propagation of the estimation errors caused by the extracted MPEG MVs. Choosing this rate is a trade-off between guaranteeing good tracking accuracy and keeping low computational requirements which are our main worry. Firstly, we have taken as reference image **frame #i** on which we applied the detection process. We kept estimating the ROIs' positions in the following frames based on their MPEG MVs exclusively. After that, we computed the overlap between these estimated positions and the ROIs' real positions. This overlap gives an indication of the estimation performance and consequently the errors. The re-detection rate should be as low as possible in order to guarantee the low computational cost of the proposed framework. In deed, starting from the reference frame, we have to increase the number of frames after which we must re-launch the ROI detector process because it is more costly than the MV extraction process. At the same time, we have to ensure the robustness of the MPEG MV based tracking algorithm. We consider an estimation as good if it satisfies this condition:

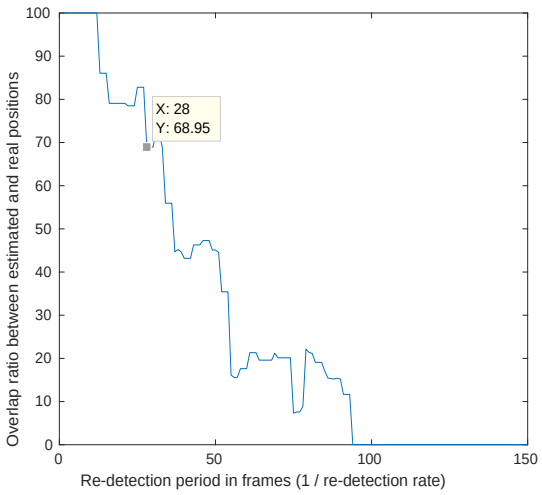


Figure 2: Estimation performance

$$\frac{A_{Detected} \cap A_{Estimated}}{\min(A_{Detected}, A_{Estimated})} \geq 70\% \qquad (1)$$

Where $A_{Detected}$ denotes the area of the detected ROI (at **Frame #(i)**) and $A_{Estimated}$ is the area of the estimated ROI (at **Frame #(i +*N*)**). We choose the same criterion as in (Ma et al., 2016).

We tested the estimation performance on some thermal videos with a frame rate equal to 25fps as mentioned before. The examples are taken from the same video of 676 frames in total. We started by detecting the ROIs at **frame #1** and then launched the tracking process to determine their estimated positions in each successive frame. After that, we computed the overlap between the real positions given by the ROI detector and the estimated ones based on their MPEG motion vectors. Figure 2 shows perfectly how the estimation performance decreases disproportionally to the re-detection cycle. In order to satisfy the condition in equation 1, N should be < 28.

As the frame rate is equal to 25fps, choosing N as 25 leads to a re-detection period of 1 second. In other words, we will launch the ROI extraction algorithm every 1 second. During this period, the tracking is ensured by the computation of the MPEG MVs relative to the detected blobs. For the rest of the experiments, N is set equal to 25.

The remainder of this section is organized as follows. We will present the performance of the proposed ROI detector including the results of the comparison between the proposed ROI extractor and MSER. After that, we will present some results of the proposed ROI tracker using two examples that illus-

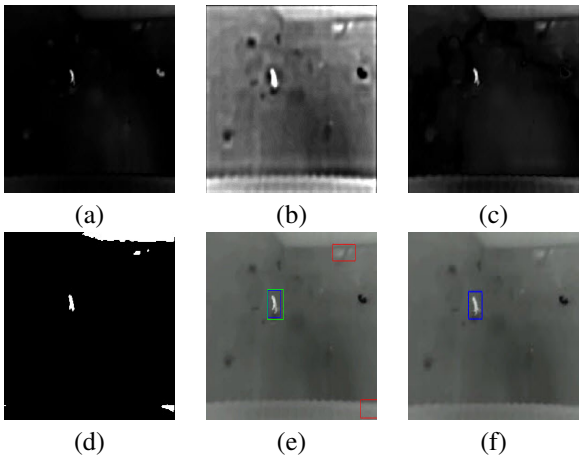


Figure 3: Examples of the different outputs of the proposed ROI detector (LITIV2012 dataset).
(a): Saliency Map (b): Contrast Enhancement (c): geometric mean of SM & CE (d): Binary image (e): Proposed ROI extractor (f): SVM (*green*: Ground Truth ROIs, *blue*: true detections, *red*: false positives)

trate how the trajectories are computed and the how positions are estimated.

4.3 Results of the Proposed ROI Detector

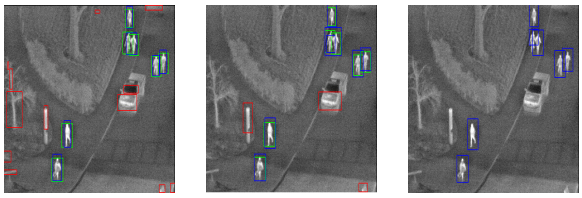
Fig.3 illustrates the outputs of the different modules that compose the proposed ROI detector including SVM. The input is an image from the LITIV2012 dataset. 3.(a) presents the corresponding saliency map and 3.(b) the result of the contrast enhancement technique. These two outputs are combined together based on their geometric mean to obtain 3.(c). At this step, a brightness threshold is applied to obtain the binary image 3.(d) where the Hot Spots are highlighted. After applying a size filter to discard very small/large areas, ROIs are extracted like shown 3.(e). The green bounding boxes correspond to the ground truth presented within the dataset. The blue bounding boxes correspond to the true positives which satisfied the equation (1) while the red ones clearly represent the false positives. The DCMs of the detected ROIs are computed as features vectors and fed to the SVM classifier. Figure 3.(f) shows how the SVM kept only the true positives.

Firstly, the proposed ROI extractor (first stage of the detector before the description and the classification stages) is compared to the MSER detector (Matas et al., 2004) which is a fast and widely used region based detector. The implementation is done on Matlab. Thus, we used the DetectMSERFeatures function available within the Computer Vision System Toolbox. For a reliable comparison, we kept the same set of parameters as for the proposed ROI extractor. The obtained results, shown in Table 1, prove its robustness in terms of true detection with approximately 96% for the OSU Thermal Pedestrian Database and 95% for the LITIV2012 dataset. Furthermore, the proposed ROI extractor presents the advantage of reducing the number of false alarms compared to the MSER detector which is a significant gain and a relevant criterion for surveillance purposes. In addition, it runs about two to three times faster. At this point, the proposed method has not yet been computationally optimized, which means that further gains are possible if it is tweaked accordingly. These two improvements are very important regarding the final purpose which is a real time implementation on a low-performance processing platforms. The UAV should select and then send only pertinent information to the central control station, that does require human attention.

Table 1: Proposed ROI Extractor vs MSER

Criterion	**Proposed ROI Extractor**	**MSER**
OSU Thermal		
True Detection Rate	95.55%	**97.83%**
False Alarms Rate	**29.22%**	51.63%
CPU-time per Image	**0.17 s**	0.46 s
LITIV2012		
True Detection Rate	**95.13%**	85.28%
False Alarms Rate	**26.25%**	39.76%
CPU-time per Image	**0.098s**	0.151s

Once the ROIs are extracted, all of them are resized at their mean size. These boxes are then described using DCMs up to order (4,4) in order to obtain a feature vector of 25 elements for each sample. Half of the samples are reserved for the training while the second half is used for test. Half of the feature vectors are assigned for training and the second half for testing the performance of the classifier. Using the OSU Thermal Pedestrian Database, we obtained 851 human samples and 491 non human samples. We used different kernels of SVM and we found out that the results are quite similar so we keep using a linear SVM. The maximum percentage of true positives is 88%. Concerning the true negatives, all the kernels present rates approaching 100% but this is explainable due to their small number and nature as they are static objects such as a public lighting pole or parked cars. Using the thermal videos from the LITIV2012



(a) (b) (c)

Figure 4: Example of different outputs from the ROI detector (OSU thermal database).
(a): MSER detector (b): Proposed ROI extractor (c): SVM (*green*: Ground Truth ROIs, *blue*: true detections, *red*: false positives)

dataset, we obtained 6237 human samples and 9997 non human samples. The increasing number of the samples leads to a better training thus better classification. For all the kernels, SVM gives a quite similar true positives rate approaching the 98%.

Fig.4 illustrates the difference between the outputs of MSER and the proposed ROI detector. It is shown how the number of false alarms is reduced using the proposed method while keeping a good true detection rate.

4.4 Results of the Proposed ROI Tracker

First of all, we will present how the two algorithms of ROI detection and MEPG MV extraction work. Thus, we start by showing in Fig.5 the outputs of the different algorithms applied between two re-detection times (**Frame #i** & **Frame #(i+*N*)**). In this case, we have selected $\mathbf{i} = 80$ and $\boldsymbol{N} = 25$. The figure 5.(a) shows how the ROI detection algorithm is able to detect the two pedestrians and trace two bounding boxes around them (blue). These boxes define the initial positions for the rest of the framework. The figure 5.(b) illustrates how the estimation algorithm proceeds. The macroblocks (blocs of 16*16 pixels) that cover each ROI are determined, then their relative MPEG MVs are extracted. After that, we follow these macroblocs through time during one cycle (from **Frame #(i+1)** to **Frame #(i+*N*)**) in order to construct the estimated trajectory for each one of them. The estimated positions are then computed based only on the resulting MPEG MVs. The figure 5.(c) presents how the proposed algorithm traces the different estimated trajectories of each macrobloc, by connecting the intermediate positions and traces the final estimated positions of each ROI (red). At this stage, the ROI re-detection process is called in order to avoid the propagation of the estimation errors in the rest of the framework. Indeed, the final estimated position at **Frame #(i+*N*)**) is updated by the real position obtained by the ROI detection algorithm.
Fig.6 shows the effectiveness of the proposed framework to predict the trajectories of three different people in the convergence scenario. Like the first example, blue boxes represent the output of the ROI extractor at the initial **Frame #8**. Red lines present the estimated trajectories of each macrobloc. The green boxes are the output of the ROI extractor at **Frame #(8+25)**. It shows how the framework is able to detect the trajectories of the three people.

5 CONCLUSION & FUTURE WORKS

This paper proposed an efficient framework for pedestrian detection and tracking in thermal images with low computational requirements. The fact that it is not a frame by frame neither pixel level processing and that it relies on MPEG MVs makes it suitable for real-time applications. The results show its effectiveness to detect and track the pedestrians. The proposed framework can be used for automatic surveillance purposes like suspicious behaviour detection. As future works, the performance of the tracking algorithm should be quantitatively measured using the CLEAR MOT metrics (Bernardin and Stiefelhagen, 2008). After that, our perspectives consist on extracting not only trajectories but velocity and acceleration information from MVs. These features (trajectory, velocity and acceleration) would be combined to construct a complete system for action recognition and abnormal event detection. Furthermore, the computational cost can be reduced further if an adaptive update of the re-detection rate parameter is introduced.

REFERENCES

Achanta, R., Hemami, S., Estrada, F., and Susstrunk, S. (2009). Frequency-tuned Salient Region Detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 1597 – 1604.

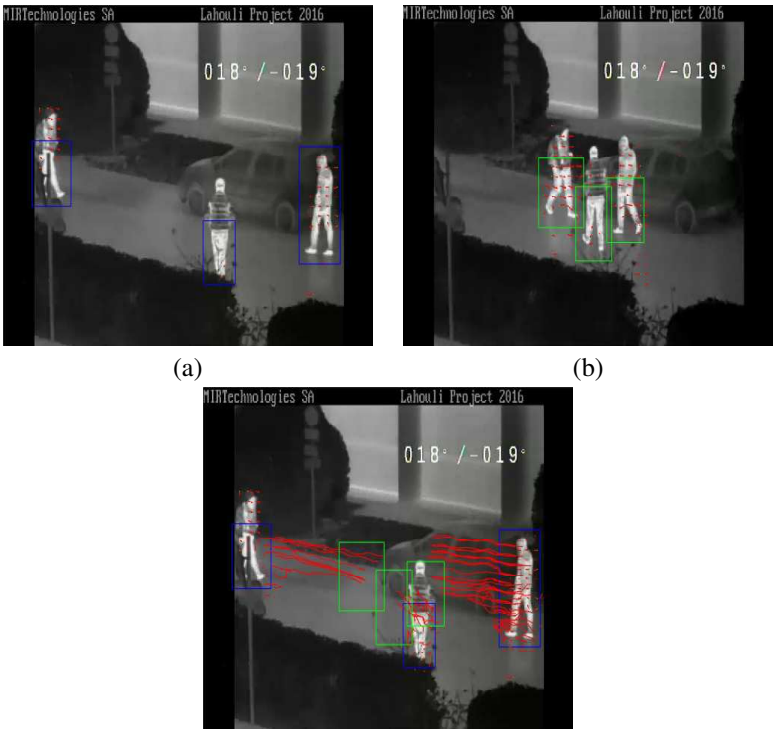
Aggarwal, A., Biswas, S., Singh, S., Sural, S., and Majumdar, A. K. (2006). Object tracking using background subtraction and motion estimation in mpeg videos. In *Asian Conference on Computer Vision*, pages 121–130. Springer.

Alyammahi, S. M. R., Salahat, E. N., Saleh, H. H. M., Sluzek, A. S., and Elnaggar, M. I. (2017). Hardware architecture for linear-time extraction of maximally stable extremal regions (msers). US Patent 9,740,947.

Arodź, T., Kurdziel, M., Popiela, T. J., Sevre, E. O., and Yuen, D. A. (2006). Detection of clustered mi-

(a) (b) (c)

Figure 5: Example of the algorithms' results applied between **Frame #i** and **Frame #(i+*N*)** (i=80, *N*=25).
(a): ROI detection results at **Frame #i** (b): ROI detection & estimation results (c): ROI detection at **Frame #(i)**, estimation & ROI re-detection at **Frame #(i+*N*)** results
(*blue*: Initial ROI positions, *red*: Estimated positions & trajectories, *green*: Real ROI positions)



(a) (b)

(c)

Figure 6: Example of trajectories' estimations of three ROIs (convergence's scenario).
(a): Initial **Frame #8** Initial ROI positions,
(b): Final **Frame #(8+25)** final ROI positions,
(c): Estimated trajectories between **Frame #(8)** and **Frame #(8+25)**.

crocalcifications in small field digital mammography. *Computer Methods and Programs in Biomedicine*, 81(1):56–65.

Avgerinakis, K., Briassouli, A., and Kompatsiaris, I. (2013). Recognition of activities of daily living for smart home environments. In *Intelligent Environments (IE), 2013 9th International Conference on*, pages 173–180. IEEE.

Babu, R. V. and Ramakrishnan, K. (2004). Recognition of human actions using motion history information extracted from the compressed video. *Image and Vision computing*, 22(8):597–607.

Babu, R. V., Ramakrishnan, K., and Srinivasan, S. (2004). Video object segmentation: a compressed domain approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(4):462–474.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359.

Bernardin, K. and Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008(1):246309.

Biswas, S. and Babu, R. V. (2013). H. 264 compressed video classification using histogram of oriented motion vectors (homv). In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 2040–2044. IEEE.

Davis, J. W. and Keck, M. A. (2005). A two-stage template approach to person detection in thermal imagery. In *Application of Computer Vision, 2005. WACV/MOTIONS '05 Volume 1. Seventh IEEE Workshops on*, volume 1, pages 364–369.

Kantorov, V. and Laptev, I. (2014). Efficient feature extraction, encoding and classification for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2593–2600.

Karakasis, E., Bampis, L., Amanatiadis, A., Gasteratos, A., and Tsalides, P. (2014). Digital elevation model fusion using spectral methods. In *2014 IEEE International Conference on Imaging Systems and Techniques (IST) Proceedings*, pages 340–345. IEEE.

Käs, C. and Nicolas, H. (2009). An approach to trajectory estimation of moving objects in the h. 264 compressed domain. *Advances in Image and Video Technology*, pages 318–329.

Khosravi, M. and Hassanpour, H. (2017). A novel image structural similarity index considering image content detectability using maximally stable extremal region descriptor. *International Journal of Engineering-Transactions B: Applications*, 30(2):172.

Kumar, A. and Gupta, S. (2017). Detection and recognition of text from image using contrast and edge enhanced mser segmentation and ocr. *IJOSCIENCE (INTERNATIONAL JOURNAL ONLINE OF SCIENCE) Impact Factor: 3.462*, 3(3):07–07.

Ma, Y., Wu, X., Yu, G., Xu, Y., and Wang, Y. (2016). Pedestrian detection and tracking from low-resolution unmanned aerial vehicle thermal imagery. *Sensors*, 16(4):446.

Matas, J., Chum, O., Urban, M., and Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767.

Park, S.-M. and Lee, J. (2003). Object tracking in mpeg compressed video using mean-shift algorithm. In *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, volume 2, pages 748–752. IEEE.

Poularakis, S., Avgerinakis, K., Briassouli, A., and Kompatsiaris, I. (2017). Efficient motion estimation methods for fast recognition of activities of daily living. *Signal Processing: Image Communication*, 53:1–12.

Śluzek, A. (2017). Mser and simser regions: A link between local features and image segmentation. In *Proceedings of the 2017 International Conference on Computer Graphics and Digital Image Processing*, page 15. ACM.

Sun, X., Ding, J., Dalla Chiara, G., Cheah, L., and Cheung, N.-M. (2017). A generic framework for monitoring local freight traffic movements using computer vision-based techniques. In *Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2017 5th IEEE International Conference on*, pages 63–68. IEEE.

Sundari, V. K. and Manikandan, M. (2017). Real time implementation of surf based target tracking algorithm. *International Journal on Intelligent Electronics Systems*, 11(1).

Torabi, A., Massé, G., and Bilodeau, G.-A. (2012). An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video surveillance applications. *Comput. Vis. Image Underst.*, 116(2):210–221.

Tun, W. N., Tyan, M., Kim, S., Nah, S.-H., and Lee, J.-W. (2017). Marker tracking with ar. drone for visual-based navigation using surf and mser algorithms. , pages 124–125.

Uemura, H., Ishikawa, S., and Mikolajczyk, K. (2008). Feature tracking and motion compensation for action recognition. In *BMVC*, pages 1–10.

Wang, H. and Schmid, C. (2013). Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558.

Wu, S., Oreifej, O., and Shah, M. (2011). Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1419–1426. IEEE.

Yeo, C., Ahammad, P., Ramchandran, K., and Sastry, S. S. (2006). Compressed domain real-time action recognition. In *Multimedia Signal Processing, 2006 IEEE 8th Workshop on*, pages 33–36. IEEE.

Zhang, B., Wang, L., Wang, Z., Qiao, Y., and Wang, H. (2016). Real-time action recognition with enhanced

motion vector cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2718–2726.

Zhang, S., Zhang, L., Gao, R., and Liu, C. (2017). Mobile robot moving target detection and tracking system. In *Proceedings of the 2017 The 7th International Conference on Computer Engineering and Networks. 22-23 July, 2017 Shanghai, China (CENet2017)*.