

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/321875560>

Pedestrian Tracking in the Compressed Domain using Thermal Images

Conference Paper · December 2017

CITATIONS

0

READS

66

5 authors, including:



Ichraf Lahouli

Royal Military Academy

6 PUBLICATIONS 0 CITATIONS

SEE PROFILE



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

Some of the authors of this publication are also working on these related projects:



Master Thesis [View project](#)



Video surveillance [View project](#)

All content following this page was uploaded by [Rob Haelterman](#) on 12 January 2018.

The user has requested enhancement of the downloaded file.

Pedestrian Tracking in the Compressed Domain Using Thermal Images

Ichraf Lahouli^{1,2,3}, Robby Haelterman¹, Zied Chtourou², Geert De Cubber¹,
and Rabah Attia³

¹ Royal Military Academy,
Brussels, Belgium

² VRIT Lab, Military Academy of Tunisia,
Nabeul, Tunisia

³ SERCOM Lab, Tunisia Polytechnic School,
La Marsa, Tunisia

Abstract. The video surveillance of sensitive facilities or borders poses many challenges like the high bandwidth requirements and the high computational cost. In this paper, we propose a framework for detecting and tracking pedestrians in the compressed domain using thermal images. Firstly, the detection process uses a conjunction between saliency maps and contrast enhancement techniques followed by a global image content descriptor based on Discrete Chebychev Moments (DCM) and a linear Support Vector Machine (SVM) as a classifier. Secondly, the tracking process exploits raw H.264 compressed video streams with limited computational overhead. In addition to two, well-known, public datasets, we have generated our own dataset by carrying six different scenarios of suspicious events using a thermal camera. The obtained results show the effectiveness and the low computational requirements of the proposed framework which make it suitable for real-time applications and on-board implementation.

1 Introduction

For decades, many works have been done on pedestrian detection and tracking using thermal imagery, especially for surveillance and driver's assistance applications. The reason is that such images allow working on day and night-time even though the texture and the colour information are missing. Nowadays, in parallel with the radar systems, the surveillance of borders, for example, is ensured by new platforms like drones equipped with optical and/or thermal sensors and transmission modules for a real-time video streaming to a central station in the ground commonly named Ground Control Station (GCS). However, the amount of information transmitted to the GCS is huge and full of redundancy and non-pertinent information. Consequently, many problems of storage and analysis are encountered in addition to the challenges caused by the high processing and the high bandwidth requirements for the data streaming.

The motion-based segmentation is widely used to detect and track moving objects like pedestrians. In this context, the majority of the works uses Optical

Flow (OF) and local feature descriptors such as SIFT and SURF. Wu et al. [1] used the OF to compute the dense particle trajectories of the objects and proposed an optimization method to filter the noisy trajectories due to the camera motion. Wang et al. [2] use dense OF and SURF descriptors to match the feature points. They also relied on a human detector to discard the inconsistent matches. Nevertheless, whether the OF is sparse or dense, is still computationally heavy and time-consuming which makes it not suitable for real-time applications. As an alternative, some studies focus on the possibility of exploiting the MVs in the MPEG compressed domain. Park et al. [3] estimated the camera motion using a generalized Hough transform and then tracked the centre of the ROI based on the spatial distribution of colours. Babu et al. [4] used motion vectors of compressed MPEG video for segmentation and a Hidden Markov Model (HMM) and motion history information for action recognition. Yeo et al. [5] used the MV information to capture the salient regions and to compute frame-to-frame motion similarity. Biswas et al. [6] used the orientation information of the MVs to classify the H.264 compressed videos. Käs and Nicolas [7] proposed an approach to estimate the trajectories of the moving objects in the compressed domain. Firstly, a Global Motion Estimation (GME) based on the MVs is performed to generate the masks which are the input of an object detection stage. Secondly, an object matching stage is used for the trajectories' estimation.

In 2014 and in the context of activity recognition, Kantorov et al. [8] used the MPEG MVs as local descriptors, Fisher Vector (FV) for coding and SVM for classification. They prove that, in comparison to the OF, the use of the MPEG MVs present a significant computational speedup ($\simeq 66\%$) while a small reduction of recognition accuracy is noticed ($\simeq 1\%$). Zhang et al. [9] proposed a real-time action recognition method in the compressed domain using the MPEG MVs. To improve the recognition accuracy, they proposed a sort of transferable learning by adapting the models of the OF Convolutional Neural Network (CNN) to the models of the MV CNN. In order to recognize activities of daily living, Poularakis et al. [10] proposed a motion estimation method based on the pre-computed MPEG MV instead of OF. In addition, they did not work on the whole frame but focused only on data in the Motion Boundary Activity Area (MBAA)[11] which also decreased the computational cost.

In this paper, we aim to present an efficient framework for pedestrians' detection and tracking in thermal images with low processing requirements. Concerning the detection process, the first step is to extract the Regions Of Interest (ROI)s using a conjunction between saliency maps and contrast enhancement techniques. Then, feature vectors are generated using the DCMs [12]. Finally, a linear SVM is used to classify the ROIs into pedestrians and non-pedestrians. In order to validate the proposed ROI detector, two public, thermal pedestrian datasets are used: the OTCBVS benchmark -OSU Thermal Pedestrian Database [13] and the nine thermal videos taken from the LITIV2012 dataset [14]. A comparison is carried out between the proposed ROI detection process and the Maximally Stable Extremal Regions (MSER) detector in terms of calculation time and true positives and false positives rates. According to the obtained results, the pro-

posed method is robust in terms of true positives rate and even beats MSER in terms of false positives rates and processing time. Concerning the tracking process, we proposed an approach which is based on the precomputed MPEG MVs of only the ROIs which are previously generated by the detection process. For the experiments, we generated our own dataset by carrying out six different scenarios of suspicious events and filmed the scene using a thermal sensor. The decoding of all the frames is not needed. Globally, the proposed method does not need a pixel by pixel or a frame by frame processing. It relies on some frames to detect the ROIs and on some MVs already computed (as an integral part of the MPEG4 AVC (H264 codec)) for tracking. This makes it adequate for real-time applications and for implementation on low-end computational platforms.

The paper is organized as follows: In Section 2, the proposed framework is presented in details by explaining the two processes of ROI detection and tracking using MPEG MVs. The section 3 is allocated to the experiments and the results, including the comparison between the proposed detector and MSER, the choice of the re-direction rate and the performance of the tracking process. Finally, Section 4 summaries the present paper and exposes some perspectives of future works.

2 Proposed methodology

In this section, we will try to explain the proposed framework in details. Actually, it relies on two main processes: the ROI detection process and the ROI tracking process. The first one extracts the ROIs which correspond to the pedestrians. The second one tracks these ROIs by using its MVs drawn directly from the MPEG compressed video. We will present the two processes consecutively.

2.1 Proposed ROI detection process

Our main purpose is to ensure the surveillance of borders and sensitive facilities using thermal images taken from an airborne platform. As we are in an outdoor environment, we assume that the pedestrians are brighter than their background. This means that our method is part of the '*Hot Spot*' Methods. ROIs are detected according to certain restrictions regarding their brightness and their size. The proposed ROI detection process can be divided into three steps:

1. ROI extraction: A conjunction between a wavelet-based contrast enhancement technique [15] and a saliency map (produced on the basis of Lab colour space) [16],
2. Shape description: DCMs (up to order 4*4) are used as a global region content descriptor [12],
3. Classification: a linear SVM is used to classify the ROIs into humans and non-humans according to their DCM feature vectors.

Firstly, the saliency map and the contrast-enhanced image are computed and fused together using their geometric mean. Then, a brightness threshold is applied to generate a binary image which conserves only the hot spot areas. Finally,

a size threshold allows discarding very small/big ROIs.

2.2 Proposed tracking process using MPEG Motion Vectors

Since the videos transmitted between the remote platforms (drones/cameras) are usually streamed to the central station in a compressed form, we should propose an object tracking approach which avoids the decompression of each frame. In order to save the processing resources and reduce the computational cost, the tracking should be done in the compressed domain. Indeed, the tracking process is based on the motion information and not the visual features such as the shape like in the detection process. After the segmentation of the input image and the extraction of the ROIs, these regions are tracked in the compressed domain based on their motion vectors.

The H.264 video compression standard generates motion vectors that contain motion information between regions in different frames. It is not a pixel level processing. It starts by splitting each frame into macroblocks (usual squares of $8 * 8$ or $16 * 16$ pixels). Then, it estimates the displacements between these areas through time and stored it's as orientation and magnitude information. In our work, we will not extract the MPEG MVs of all the macroblocks. Actually, the algorithm starts by finding the macroblocks that cover each ROI. Then, it keeps tracking these macroblocks through time by computing the intermediate estimated positions based on its relative MPEG MVs. Although these MPEG MVs are useful, we cannot rely exclusively on its due to the noise and the errors generated by the motion compensation step. We propose to compensate these errors by launching the aforementioned ROI detection process at a re-detection rate namely \mathbf{N} . The recall of the human detector will adjust the intermediate estimated positions of the ROIs. The choice of this frequency \mathbf{N} is not fixed but depends on different parameters such as the frame rate and the resolution of the video sequence. An analysis in section 3.3 shows how \mathbf{N} is chosen.

3 Experiments & Results

3.1 Presentation of the different datasets

In order to validate the ROI detection process, two different public, thermal pedestrian datasets were used:

- *OSU Thermal Pedestrian Database*[13]: 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.
- *LITIV2012 dataset*[14]: specifically the nine thermal sequences. Indeed, the dataset is composed of nine pairs of visualthermal sequences.

In order to validate the tracking process, we can not test on the two public datasets previously used to validate the proposed detection process. The reason is that both datasets do not provide H.264 encoding data so we needed to generate our own dataset. Since our main application is the video surveillance of borders and sensitive facilities, we carried out different scenarios of suspicious events in an outdoor environment and filmed the scene using a thermal sensor. Indeed, we recorded thermal videos of pedestrians taken from a stationary camera with an image resolution of 576*704 pixels and a frame rate of 25 frames per second (fps).

Table 1 gives an overview of the six scenarios of suspicious events.

3.2 Validation of the proposed detection process

Firstly, the proposed ROI extractor (first stage of the detector before the description and the classification stages) is compared to the MSER detector [17] which is a fast, widely used and simple region based detector. MSER was introduced in 2004 but is still up to date and widely used a region-based local extractor like recently in [18–23]. The popularity of MSER is due to its efficiency and its low complexity which makes it adequate for real-time applications. The implementation is done on Matlab. Thus, the DetectMSERFeatures function, available in the Computer Vision System Toolbox, is used. The experiments were run under the same set of parameters like size thresholds. Table 2 shows the robustness of the proposed detection process in terms of true detection with approximately 96% for the OSU Thermal Pedestrian Database and 95% for the LITIV2012 dataset. Furthermore, it beats MSER in terms of reducing the false alarms' rate which is a great criterion for surveillance purposes. Concerning the CPU time, the proposed detection process also beats MSER by running about two to three times faster. Regarding the desired application of the proposed framework, these two improvements are pertinent and make the proposed framework suitable for a real-time implementation on a drone for instance, in order to select and then send only true alarms to the GCS.

3.3 Validation of the proposed tracking process

Re-detection rate

In order to compensate the estimation errors caused by the extracted MPEG MVs, the proposed detection process is recalled at a re-detection rate. Choosing this parameter is a trade-off between keeping low computational requirements and guaranteeing good tracking accuracy. In other words, we have to avoid the re-launch of the ROI detector process and at the same time, we have to ensure the robustness of the whole framework. Indeed, the recall of the proposed detection process means the decompression of the frame and the application of image processing techniques that are more computationally costly than the simple extraction of the precomputed MPEG MVs. To set up the re-detection rate, we first applied the proposed detection process on a reference image **frame #i** to generate the ROIs. Then, the positions of these ROIs are estimated based







Scenario	Description	Frame example
<i>Brutal turn back</i>	<i>2 people move in one direction (policemen) + 1 single suspicious person walks in the opposite direction. Once he sees them he will rapidly turn back.</i>	
<i>Convergence/divergence</i>	<i>3 suspicious people converge, quickly exchange an object and then diverge and quit the scene.</i>	
<i>Velocity changes</i>	<i>1 single suspicious person who walks then runs then slows down again + non suspicious people.</i>	
<i>Occlusion/Non Occlusion</i>	<i>1 single suspicious person tries to hide behind a car + non suspicious people.</i>	
<i>Circular trajectory</i>	<i>1 single suspicious person moves around a car while focusing on it (robbery intention) + non suspicious people.</i>	
<i>Rapid dump of a suspicious object</i>	<i>1 single suspicious person walks carrying a backpack then puts it down near a vehicle + non suspicious people.</i>	

Table 1. Suspicious Events' scenarios

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

Table 2. Proposed ROI Detector vs MSER

exclusively on their MPEG MVs. To measure the estimation performance of the tracking using only the MPEG MVs, we computed the overlap between the real and the estimated positions of the ROIs. We consider an estimation as good if it satisfies the condition below:

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

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 [24].

We tested the estimation performance on the video sequences of our own dataset (frame rate=25fps). Fig. 1 illustrates the computed overlap on a thermal video sequence composed of 676 frames in total. We started by detecting the ROIs at **frame #1** and then kept tracking these ROIs based exclusively on its MPEG MVs. At each frame, the overlap between the real positions (given by the ROI detector) and the estimated positions is computed. Fig. 1 shows that the estimation performance decreases disproportionally to the re-detection rate. In order to satisfy the condition in equation 1, N should be < 28 . In other words, to ensure the robustness of the proposed tracking process, the frequency to recall the detection process should be not more than 28 frames. As the frame rate is equal to 25fps, choosing N equal to 25 means a recall of the detection process each 1 second. For the rest of the experiments, $N=25$.

Tracking example

At this stage, we will present an example that illustrates how the proposed approach works well. Fig. 2 shows the effectiveness of the proposed framework to predict the trajectories of three different people in the convergence scenario. Fig. 2.(a) presents the initial **Frame #8** and the outputs of the proposed detection process in Blue. These bounding boxes correspond the initial ROI positions. Fig. 2.(b) presents the **Frame # $(8+25)$** and the target ROI positions in green. Fig. 2.(c) shows the estimated trajectories of the ROIs between the two frames. For each ROI, a trajectory is computed based on the MPEG MVs of the macroblocks

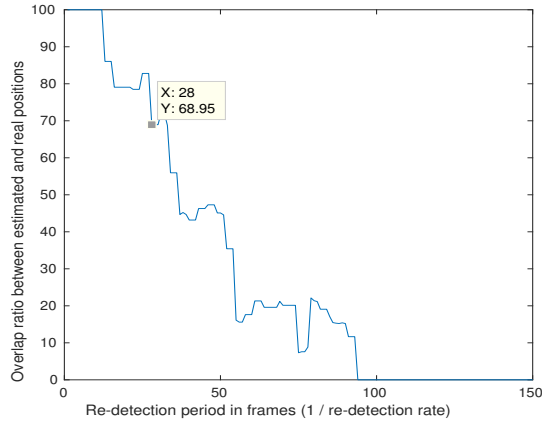


Fig. 1. Estimation performance

that cover it. The example shows how the proposed framework was able to properly estimate the trajectories of the three pedestrians.

4 Conclusion & Future works

This paper proposed an efficient approach for pedestrian detection and tracking in thermal images with low computational requirements. The proposed framework is not a frame neither a pixel level processing and it relies on the MPEG MVs which makes it suitable for real-time applications. The results show its effectiveness to detect and track pedestrians in thermal images even though there is no colour or texture information. As future works, the performance of the tracking algorithm should be quantitatively measured using, for example, the CLEAR MOT metrics [25]. At this stage of work, only the trajectories of the different pedestrians in the scene are extracted. However, in order to construct a complete system for surveillance and abnormal event detection applications, these trajectories need firstly to be described using feature vectors. Then, a machine learning approach should be developed in order to allow the system to autonomously detect the suspicious people. In addition, trajectories alone might not be sufficient but need to be combined with velocity and acceleration information, which might be also computed in the compressed domain.

5 Acknowledgment

The generation of the proposed dataset using thermal cameras is supported by MIRTECHNOLOGIES SA, Chemin des Eysines 51, 1226 Nyon, CH.

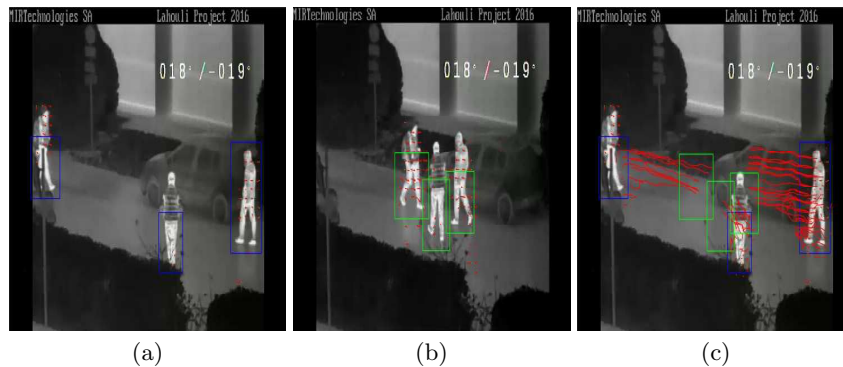


Fig. 2. Example of trajectories' estimations of three ROIs (convergence's scenario).
 (a): Initial **Frame #8** Initial ROI positions,
 (b): Target **Frame #(8+25)** Target ROI positions,
 (c): **Estimated trajectories** between **Frame #(8)** and **Frame #(8+25)**.

References

1. S. Wu, O. Oreifej, and M. Shah, "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*, pp. 1419–1426, IEEE, 2011.
2. H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3551–3558, 2013.
3. S.-M. Park and J. Lee, "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*, vol. 2, pp. 748–752, IEEE, 2003.
4. R. V. Babu, K. Ramakrishnan, and S. Srinivasan, "Video object segmentation: a compressed domain approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 4, pp. 462–474, 2004.
5. C. Yeo, P. Ahammad, K. Ramchandran, and S. S. Sastry, "Compressed domain real-time action recognition," in *Multimedia Signal Processing, 2006 IEEE 8th Workshop on*, pp. 33–36, IEEE, 2006.
6. S. Biswas and R. V. Babu, "H. 264 compressed video classification using histogram of oriented motion vectors (homv)," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 2040–2044, IEEE, 2013.
7. C. Käs and H. Nicolas, "An approach to trajectory estimation of moving objects in the h. 264 compressed domain," *Advances in Image and Video Technology*, pp. 318–329, 2009.
8. V. Kantorov and I. Laptev, "Efficient feature extraction, encoding and classification for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2593–2600, 2014.

9. B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, "Real-time action recognition with enhanced motion vector cnns," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2718–2726, 2016.
10. S. Poularakis, K. Avgerinakis, A. Briassouli, and I. Kompatsiaris, "Efficient motion estimation methods for fast recognition of activities of daily living," *Signal Processing: Image Communication*, vol. 53, pp. 1–12, 2017.
11. K. Avgerinakis, A. Briassouli, and I. Kompatsiaris, "Recognition of activities of daily living for smart home environments," in *Intelligent Environments (IE), 2013 9th International Conference on*, pp. 173–180, IEEE, 2013.
12. E. Karakasis, L. Bampis, A. Amanatiadis, A. Gasteratos, and P. Tsalides, "Digital elevation model fusion using spectral methods," in *2014 IEEE International Conference on Imaging Systems and Techniques (IST) Proceedings*, pp. 340–345, IEEE, 2014.
13. J. W. Davis and M. A. Keck, "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*, vol. 1, pp. 364–369, Jan 2005.
14. A. Torabi, G. Massé, and G.-A. Bilodeau, "An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video surveillance applications," *Comput. Vis. Image Underst.*, vol. 116, pp. 210–221, Feb. 2012.
15. T. Arodź, M. Kurdziel, T. J. Popiela, E. O. Sevre, and D. A. Yuen, "Detection of clustered microcalcifications in small field digital mammography," *Computer Methods and Programs in Biomedicine*, vol. 81, no. 1, pp. 56–65, 2006.
16. R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned Salient Region Detection," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pp. 1597 – 1604, 2009.
17. J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.
18. W. N. Tun, M. Tyan, S. Kim, S.-H. Nah, and J.-W. Lee, "Marker tracking with ar. drone for visual-based navigation using surf and mser algorithms," , pp. 124–125, 2017.
19. X. Sun, J. Ding, G. Dalla Chiara, L. Cheah, and N.-M. Cheung, "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*, pp. 63–68, IEEE, 2017.
20. A. Kumar and S. Gupta, "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*, vol. 3, no. 3, pp. 07–07, 2017.
21. M. Khosravi and H. Hassanpour, "A novel image structural similarity index considering image content detectability using maximally stable extremal region descriptor," *International Journal of Engineering-Transactions B: Applications*, vol. 30, no. 2, p. 172, 2017.
22. S. M. R. Alyammahi, E. N. Salahat, H. H. M. Saleh, A. S. Sluzek, and M. I. Elnaggar, "Hardware architecture for linear-time extraction of maximally stable extremal regions (msers)," Aug. 22 2017. US Patent 9,740,947.
23. A. Śluzek, "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*, p. 15, ACM, 2017.

24. Y. Ma, X. Wu, G. Yu, Y. Xu, and Y. Wang, "Pedestrian detection and tracking from low-resolution unmanned aerial vehicle thermal imagery," *Sensors*, vol. 16, no. 4, p. 446, 2016.
25. K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, p. 246309, May 2008.