Optimal Camera Placement for Multimodal Video Summarization

Vishal Parikh(✉), Priyanka Sharma, Vedang Shah, and Vijay Ukani

Institute of Technology, Nirma University, Ahmedabad, India
{vishalparikh, priyanka.sharma, l5mcei26, vijay.ukani}@nirmauni.ac.in

Abstract. Video Surveillance systems are used to monitor, observe and intercept the changes in activities, features and behavior of objects, people or places. A multimodal surveillance system incorporates a network of video cameras, acoustic sensors, pressure sensors, IR sensors and thermal sensors to capture the features of the entity under surveillance, and send the recorded data to a base station for further processing. Multimodal surveillance systems are utilized to capture the required features and use them for pattern recognition, object identification, traffic management, object tracking, and so on. The proposal is to develop an efficient camera placement algorithm for deciding placement of multiple video cameras at junctions and intersections in a multimodal surveillance system which will be capable of providing maximum coverage of the area under surveillance, which will leads to complete elimination or reduction of blind zones in a surveillance area, maximizing the view of subjects, and minimizing occlusions in high vehicular traffic areas. Furthermore, the proposal is to develop a video summarization algorithm which can be used to create summaries of the videos captured in a multi-view surveillance system. Such a video summarization algorithm can be used further for object detection, motion tracking, traffic segmentation, etc. in a multi-view surveillance system.

Keywords: Multimodal surveillance · Multiview video summarization
Camera placement

1 Introduction

Video surveillance systems deal with monitoring, intercepting or observing activities, behavior, or any other changing information related to people, places or things. Video surveillance systems have evolved over three generations of surveillance systems namely, analog surveillance systems, digital surveillance systems, and smart/intelligent surveillance systems. Nowadays network of various surveillance video sensors/cameras are everywhere. Figure 1 shows an example of a video sensor/camera network, with overlapping as well as non-overlapping fields of view. Multimodal surveillance systems in intelligent transportation systems have a wide application area and has been an emerging field of research. A multimodal surveillance system normally consists of a wireless sensor network of video/image sensors, audio sensors, pressure sensors, thermal sensors and position sensors. Apart from these, some recent advances in sensor hardware includes, an embedded data processing algorithm which is used to process the

P. K. Singh et al. (Eds.): FTNCT 2018, CCIS 958, pp. 123–134, 2019.
https://doi.org/10.1007/978-981-13-3804-5_10

data captured by the sensor and send it to a base station. The placement of different video sensors/cameras are very important. The main idea in multimodal scenario is: weather we get an idea of which camera/sensor has captured an important video content without watching all the videos of all sensors/cameras entirely.

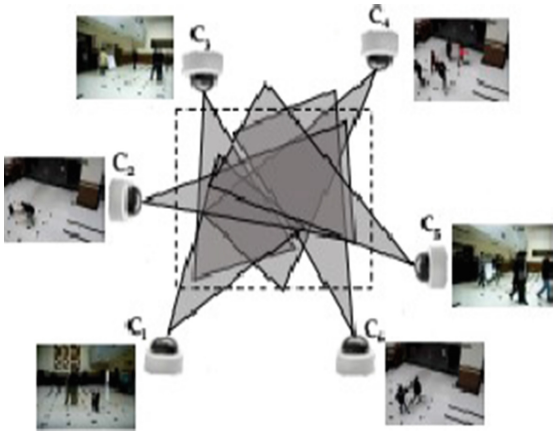

Fig. 1. Illustration of multi-view camera network [3]

Sufficient progress has been made in summarizing a single video. Comprehensive reviews in video summarization can be found in [1] and [2]. Truong et al. in [1] gives two ways in which video can be summarized, a key-frame sequence and a video skim. Nowadays multiview video summarization is gaining popularity as there are number of video sensors/cameras deployed which covers overlapping region. The contribution of our work can be summarized as a video sensor/camera placement strategy for surveillance in intelligent transport system, and also we propose a key-frame based summarization technique to preserve both intra and inter view correlation for MPEG-4 or H.264(AVC) videos for generating video summaries.

2 Related Works

While the goal of many optimal camera placement strategies has been to minimize the overlapping views; with respect to the objective of the proposed system, overlapping views were necessary so as to track the complete path of motion of the subjects under surveillance from multiple views, maximize their visibility and maximize the degree of coverage of the surveillance perimeter. Most summarization algorithms work on single-view video, due to redundancy in multi-view video, multiview video summarization is much more comprehensive.

Zouaoui et al. in [4] have proposed a multimodal system composed of two microphones and one camera integrated on-board for video and audio analytic for surveillance. The system relies on the fusion of unusual audio events detection and/or object detection from the captured video sequences. The audio analysis consists of

modeling the normal ambience and detecting deviation from the trained models during testing, while the video analysis involves classification according to geometric shape and size. However, even though the system succeeds in detecting robust 3D position of objects, it employs only a single camera for surveillance which does not provide a robust multi-dimensional view of the object of interest.

Wang et al. in [5] have presented a system for detecting and classifying moving vehicles. The system uses video sensors along with Laser Doppler Vibrometer (LDVs) a kind of acoustic sensor for detecting the motion, appearance and acoustic features of the moving vehicles - and later on using the data to classify them.

Magno et al. in [6] have proposed a multimodal low power and low cost video surveillance system based on a CMOS video sensor and a PyroelectricInfraRed (PIR) sensor. In order to control the power consumption, instead of transmitting full image, the sensors only transmit very limited amount of information such as number of objects, trajectory, position, size, etc., thus saving a large amount of energy in wireless transmission and extending the life of the batteries. However nothing is done from the point of view of data transmission and power consumption if the targeted object is not detected. In addition to this, this system is used only for detecting an abandoned or removed object from the perimeter under surveillance and hence there is no proper evidence of its usage in a large-scale, dynamically changing environment.

Gupta et al. in [7] have designed a distributed visual surveillance system for military perimeter surveillance. The system is used to detect potential threats and create actionable intelligence to support expeditionary war fighting for the military base camp by using multimodal wireless sensor network. The system employs certain rule-based algorithms for detection of atomic actions from video. Some of the atomic actions that are automatically detected by the system are: a person appearing in a restricted area, tripwire crossing, a person disappearing from a protected perimeter, a person entering or exiting, leave behind action, loiters, take away action, etc. A geodetic coordinate system is used which provides metric information such as size, distance, speed, and heading of the detected objects for high level inference and inter-sensor object tracking.

Prati et al. in [8] have proposed a PIR sensor based multimodal video surveillance system. In this system PIR sensors are used to bring down the cost of deployment of the surveillance systems and at the same time they are combined with vision systems for precisely detecting the speed and direction of the vehicles along with other complex events.

Rios-Cabrera et al. in [9] have presented an efficient multi-camera vehicle identification, detection and tracking system inside a tunnel. In this system a network of non-overlapping video cameras are used to detect and track the vehicles inside a tunnel by creating a vehicle-fingerprint using the haar features of the vehicles despite poor illumination inside tunnel and low quality images.

Lopatka et al. in [10] have proposed a system for detecting the traffic events which uses special acoustic sensors, pressure sensors and video sensors to record the occurrence of audio-visual events. A use-case of detection of collision of the two cars is demonstrated in this paper. The data collected by the multimodal sensors is sent to a computational cluster in real time for analysis of the traffic events. For this purpose a Real Time Streaming Protocol (RTSP) is used in the system.

Wang et al. in [11] have proposed a large scale video surveillance system for wide area monitoring which has capability of monitoring and tracking a moving object in a

widely open area using an embedded component on the camera for detailed visualization of objects on a 2D/3D interface. In addition to this, it is also capable of detecting illegal parking and identifies the drivers face from the illegal parking event.

van den Hengel et al. in [12] have proposed a genetic algorithm for automatic placement of multiple surveillance cameras which is used to optimize the coverage of cameras in large-scale surveillance systems and at the same find overlapping views between cameras if necessary. Yildiz et al. in [13] have presented a bilevel algorithm to determine an optimal camera placement with maximum angular coverage for a WSN of homogeneous and heterogeneous cameras. Zhao et al. in [14] have presented two binary integer programming (BIP) algorithms for finding optimal camera placement and network configuration. Moreover they have extended the proposed framework to include visual tagging of subjects in the surveillance environments. Liu et al. in [15] have presented a Multi-Modal Particle Filter technique to track vehicles from different views (frontal, rear and side view). In addition to this they have also discussed a technique for occlusion handling in surveillance systems.

Denman et al. in [16] have presented a system for automatic monitoring and tracking of vehicles in real time using optical flow modules and motion detection from videos captures by four video cameras. Wang et al. in [17] have proposed an effective foreground object detection technique for surveillance systems by estimating the conditional probability densities for both the foreground and background objects using feature extraction techniques and temporal video filtering. Zheng et al. in [18] have proposed a key-frame selection technique based on motion-feature based approach in which motion information for each key-frame from the traffic surveillance video stream is computed in a GPU based system and key-frames with motion information greater than their neighbors are selected. By implementing GPU based processing capabilities, the authors have shown a significant increase in the accuracy and processing speed of the algorithm.

Panda et al. in [19] have proposed a novel sparse representative selection method for summarizing multi-view videos, that is videos captured from multiple cameras. They have used inter-view and intra-view similarities between the feature descriptors of each view for modelling multi-view correlations. Kuanar et al. in [20] have proposed a bipartite matching method for multi-view correlation of features like visual bag of words, texture, color, etc. and extracting frames for summarization of multi-view videos. In this method the authors have used Optimum-Path Forest algorithm for clustering the intra-view dependencies and removing intra-view redundancies. Liu et al. in [21] have proposed a unique method for visualizing object trajectories in multi-camera videos and creating video summaries of suspicious movements in a building.

3 Optimal Camera Placement in Multimodal Surveillance System

Many large-scale multimodal surveillance systems have used human experts for camera selection and placement, however such a technique is not capable to effectively design a system while considering the multitude of factors. Also a straightforward method to deploy the video cameras would be to deploy them uniformly around the surveillance area. However, in real-world deployment scenarios, such a method of uniform

placement is not practical, since the placement of cameras is restricted by many constraints like costs, availability, visibility, applicability, feasibility, and other factors. This study has investigated the effect of all the factors listed above, and an optimal camera placement strategy has been designed which satisfies all these factors.

Figure 2 gives the coverage of a video camera C in three-dimensional space. With reference to the Fig. 2, point V is the position of the video camera V(x, y, z) and point G indicates the centre of gravity for the video camera V. The four points A, B, C and D are the extreme points in the FOV of V and can be computed using horizontal AOV, vertical AOV and position of the video camera. These points also form the base plane of the rectangular pyramid. Point X is an arbitrary point present in the FOV of video camera V which is to be observed using V.

The volume of the rectangular pyramid formed by the points {V, A, B, C, D} (that is the volume of the coverage area of the camera C) is given by the Eq. 1,

$$V_c = \frac{l * b * h}{3} \tag{1}$$

where h is height of apex from the base.

Now since X is an arbitrary point inside the FOV of V, it forms four tetrahedrons with the four sides of the pyramid and point V as the apex. Volume of each such tetrahedron is given by the Eq. 2,

$$V_i = \frac{\sqrt{2} * Area_{base} * h}{12} \tag{2}$$

where h is height of apex from the base and i = 1 to 4.

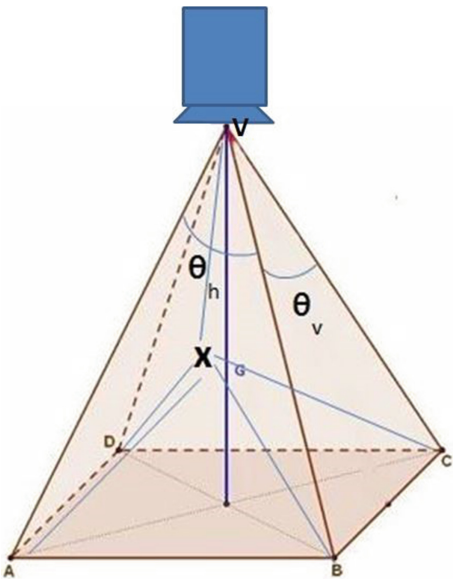

Fig. 2. Coverage of a video camera

Consider V_{total} as the total volume computed by adding volumes of all the four tetrahedrons and pyramid created with X as the apex. V_{total} is given by the Eq. 3,

$$V_{total}^{X} = V_{base} + \sum V_i \tag{3}$$

for all i = 1 to 4 in Eq. 3, V_{base} is the volume of pyramid with point X as apex and points A, B, C, D as the base plane. The Eq. 4 is used to test the presence of a point X within the FOV of a video camera C whose coverage area can be modelled as a rectangular pyramid of volume V, and returns true or false accordingly.

$$FOV(C,X) = \begin{pmatrix} true, & if\ V_c = V_{total}^{X} \\ false \end{pmatrix} \tag{4}$$

Algorithm 1 depicts systematic steps for calculating optimal camera placement in a multimodal surveillance system. This algorithm can be used to decide the placement of multiple cameras at intersections, junctions and crossroads and achieve the best possible coverage of the surveillance area.

Algorithm 1 Optimal Camera Placement

$C_i(x,y,z)$

Require: Surveillance area P of size LxBxH
Cameras C_i, where i = 1 to 4

1: Initialization
2: Find midpoint M_j for each side of the region LxB of P and divide the region LxB of P into four equal regions R_i; where i = 1 to 4
3: **for** i = 1 to 4 **do**
4: For region R_i, midpoints M_j on the adjacent edges of R_i and camera C_i, find $C_i(x,y,z)$ using **function** FOV(C_i,M_j) such that $C_i(x,y,z)$ lies inside $R_i(x,y,z)$
5: **end for**
6: **function:** FOV(C_i,M_j)
7: **Input:** $C_i(x,y,z)$,M_j
8: **Output:** TRUE/FALSE
9: *result* = FALSE
10: For camera C_i find volume of its coverage area V_i
11: For point M_j find individually volume of four tetrahedrons $V_n^{M_j}$ formed by M_j with $C_i(x,y,z)$ as apex and each of the four sides of the coverage area of $C_i(x,y,z)$, where n =1 to 4
12: Find volume $V_{base}^{M_j}$ of the pyramid formed with M_j as apex and the base plane of the coverage area of $C_i(x,y,z)$ as the base plane of the pyramid
13: Find the total volumes of all the tetrahedrons and pyramid created with M_j as: $V_{total}^{M_j} = V_{base}^{M_j} + \sum V_n^{M_j}$, where n = 1 to 4
14: **if** $V_{total} == V_i$
15: $result = TRUE$
16: **end if**
17: **return** *result*
18: **end function**

4 Experimental Results and Discussion

4.1 Simulation Environment and Parameters

The optimal camera placement Algorithm discussed previously was simulated in OMNET++ Network Simulator. Table 1 lists the simulation parameters that were considered while checking the results and validity of the algorithm.

The network topology was configured in a way that each video camera would be placed randomly inside or on the edges of the one-fourth part of the surveillance area as shown in the Fig. 3, since the surveillance area is divided into four equal parts. Also as mentioned in the algorithm, each camera needed to have two midpoints of adjacent sides in their FOV to have the best possible coverage of the surveillance area.

Table 1. Simulation parameters for optimal camera placement algorithm

Parameter	Value
Simulation time	300 s
Surveillance area	25 m × 20 m × Depth of surveillance area (mentioned below)
Depth of surveillance area	10 m to 15 m
Number of video cameras	4
Focal length	4.0 mm
AOV of each camera	90 to 120
Camera deployment (co-ordinates in three dimensional space)	Random

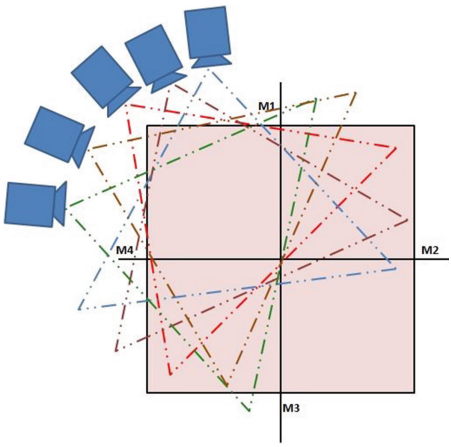

Fig. 3. Surveillance area with various possible camera placements

4.2 Results

During the simulation, the cameras were placed randomly inside each one-fourth part of the surveillance area, near the midpoints and on the vertex joining the two edges. At any moment of time, the AOV and Depth of Field of each camera was fixed and identical, though it was selected randomly from the range specified in Table 1.

The results of the simulation are presented in Table 2. From the results, it can be inferred that the best placement for all the cameras in a multimodal surveillance system is near the vertex joining any two edges of the surveillance area. It can also be derived that higher depth of field and wider angle of view produces better region coverage for a surveillance camera leaving less than 5% of area uncovered. However, since all the cameras have overlapping views, it is possible to achieve better coverage of the whole area leaving very less to zero blind spots. The Fig. 4, shows the best placement for video cameras in a multimodal surveillance system from the result derived using the optimal placement algorithm discussed in Algorithm 1.

Table 2. Simulation results for optimal camera placement algorithm

AOV (degree)	DOF (m)	Camera placement	Area covered (%)
90	10	Random-inside the one-fourth part of surveillance area	29
90	11	Random-inside the one-fourth part of surveillance area	34
100	12	Random-inside the one-fourth part of surveillance area	45
100	13	Random-inside the one-fourth part of surveillance area	50
105	14	Random-inside the one-fourth part of surveillance area	48
100	15	Random-inside the one-fourth part of surveillance area	53
90	10	On the sides of the one-fourth part of surveillance area	43
95	11	On the sides of the one-fourth part of surveillance area	50
110	12	On the sides of the one-fourth part of surveillance area	61
100	13	On the sides of the one-fourth part of surveillance area	70
105	14	On the sides of the one-fourth part of surveillance area	73
120	15	On the sides of the one-fourth part of surveillance area	74
95	10	Near the midpoints	23

(*continued*)

Table 2. (*continued*)

AOV (degree)	DOF (m)	Camera placement	Area covered (%)
110	11	Near the midpoints	28
120	12	Near the midpoints	39
100	13	Near the midpoints	38
100	14	Near the midpoints	27
95	15	Near the midpoints	24
95	10	At the vertices joining the two edges	94
110	11	At the vertices joining the two edges	97
105	12	At the vertices joining the two edges	95
120	13	At the vertices joining the two edges	98
105	14	At the vertices joining the two edges	95
120	15	At the vertices joining the two edges	97

4.3 Video Summarization

The Ko-PER Intersection Dataset [22] that comprises of highly accurate reference trajectories of cars, raw laser scanner measurements, undistorted monochrome camera images has being used. From this dataset, four videos were generated of varying GOP size N = 8, 16, 24 and 30 with a constant frame rate of 30 fps and were encoded using

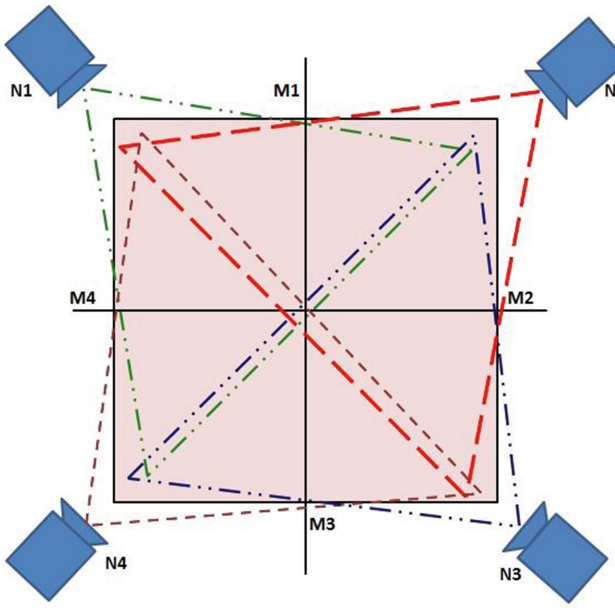

Fig. 4. Optimal camera placement design

the H.264 or MPEG-4 Part 10, Advanced Video Coding (AVC) standard. These different videos were used to check the performance of each frame selection technique individually. The graph presented in Fig. 5 shows the comparison of the total execution times of the proposed three frame selection techniques for video summarization. It is evident from the graph that, with increase in GOP size, the execution time of the algorithms increases. By using any of the technique of frame selection for video summarization, the duration of the final summarized video is same, since no key-frames have been dropped in the process, and hence all the three techniques are suitable to create video summaries without the loss of important contextual information from the video.

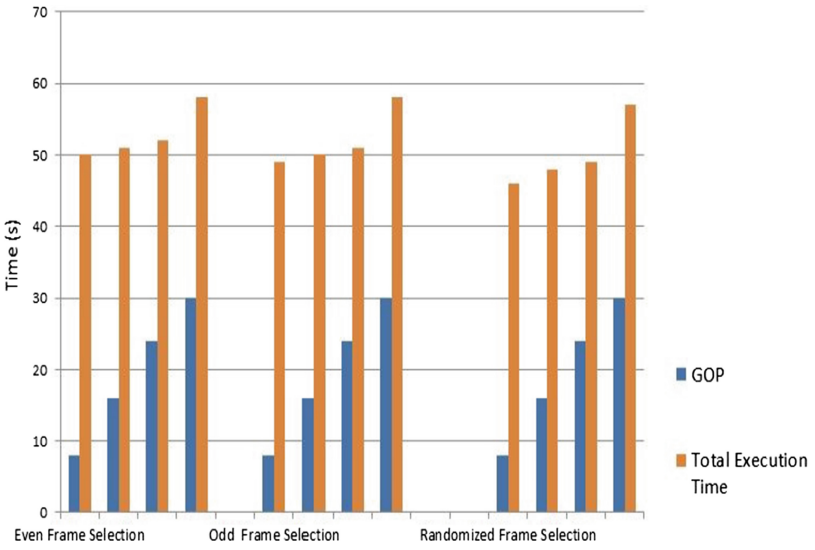

Fig. 5. Comparison of frame selection techniques for video summarization algorithm

5 Conclusion

The proposed optimal camera placement algorithm can be used for deciding the placement of multiple cameras at intersections, junctions and cross-roads without compromising the coverage area of the deployed video cameras and cost of deployment. This optimal camera placement algorithm is capable of providing maximum coverage of the area under surveillance leading to - complete elimination or reduction of the number of blind zones in a surveillance area. It is also maximizing the view of subjects and minimizing occlusions in high vehicular traffic areas. In addition to this, a video summarization algorithm using three different techniques of frame selection for multi-view surveillance systems is presented which can be used to create summaries of large-sized, lengthy video streams of traffic surveillance data and, at the same time reduce the computational processing for creating the video summaries. Collectively,

both the proposed algorithms will be able to reduce the cost of camera deployment, computational cost, power consumption and, provide efficient performance in a multi-view, as well as multimodal surveillance system.

References

1. Truong, B.T., Venkatesh, S.: Video abstraction: a systematic review and classification. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **3**(1), 3 (2007)
2. Money, A.G., Agius, H.: Video summarisation: a conceptual framework and survey of the state of the art. J. Vis. Commun. Image Represent. **19**(2), 121–143 (2008)
3. Panda, R., Roy-Chowdhury, A.K.: Multi-view surveillance video summarization via joint embedding and sparse optimization. IEEE Trans. Multimed. **19**, 2010–2021 (2017)
4. Zouaoui, R., et al.: Embedded security system for multi-modal surveillance in a railway carriage. In: Proceedings of SPIE, January 2016
5. Wang, T., Zhu, Z.: Multimodal and multi-task audio-visual vehicle detection and classification. In: 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 440–446, September 2012
6. Magno, M., Tombari, F., Brunelli, D., Stefano, L.D., Benini, L.: Multimodal abandoned/removed object detection for low power video surveillance systems. In: Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2009, pp. 188–193, September 2009
7. Gupta, H., Yu, L., Hakeem, A., Choe, T.E., Haering, N.: Multimodal complex event detection framework for wide area surveillance. In: CVPR 2011 Workshops, pp. 47–54, June 2011
8. Prati, A., Vezzani, R., Benini, L., Farella, E., Zappi, P.: An integrated multi-modal sensor network for video surveillance. In: Proceedings of the Third ACM International Workshop on Video Surveillance & Sensor Networks, pp. 95–102 (2005)
9. Rios-Cabrera, R., Tuytelaars, T., Gool, L.V.: Efficient multi-camera vehicle detection, tracking, and identification in a tunnel surveillance application. Comput. Vis. Image Underst. **116**, 742–753 (2012)
10. Lopatka, K., Kotus, J., Szczodrak, M., Marcinkowski, P., Korzeniewski, A., Czyzewski, A.: Multimodal audio-visual recognition of traffic events. In: 2011 22nd International Workshop on Database and Expert Systems Applications, pp. 376–380, August 2011
11. Wang, Y.K., Fan, C.T., Huang, C.R.: A large scale video surveillance system with heterogeneous information fusion and visualization for wide area monitoring. In: 2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), pp. 178–181, July 2012
12. van den Hengel, A., et al.: Automatic camera placement for large scale surveillance networks. In: 2009 Workshop on Applications of Computer Vision (WACV), pp. 1–6, December 2009
13. Yildiz, E., Akkaya, K., Sisikoglu, E., Sir, M.Y.: Optimal camera placement for providing angular coverage in wireless video sensor networks. IEEE Trans. Comput. **63**, 1812–1825 (2014)
14. Zhao, J., Cheung, S.C., Nguyen, T.: Optimal camera network configurations for visual tagging. IEEE J. Sel. Top. Signal Process. **2**, 464–479 (2008)
15. Liu, L., Xing, J., Ai, H.: Multi-view vehicle detection and tracking in crossroads. In: The First Asian Conference on Pattern Recognition, pp. 608–612, November 2011

16. Denman, S., et al.: Multi-view intelligent vehicle surveillance system. In: 2006 IEEE International Conference on Video and Signal Based Surveillance, p. 26, November 2006
17. Wang, K., Liu, Y., Gou, C., Wang, F.Y.: A multi-view learning approach to foreground detection for traffic surveillance applications. IEEE Trans. Veh. Technol. **65**, 4144–4158 (2016)
18. Zheng, R., Yao, C., Jin, H., Zhu, L., Zhang, Q., Deng, W.: Parallel key frame extraction for surveillance video service in a smart city, vol. 10, pp. 1–8, August 2015
19. Panda, R., Dasy, A., Roy-Chowdhury, A.K.: Video summarization in a multi-view camera network. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 2971–2976, December 2016
20. Kuanar, S.K., Ranga, K.B., Chowdhury, A.S.: Multi-view video summarization using bipartite matching constrained optimum-path forest clustering. IEEE Trans. Multimed. **17**, 1166–1173 (2015)
21. Liu, S., Lai, S.: Schematic visualization of object trajectories across multiple cameras for indoor surveillances. In: 2009 Fifth International Conference on Image and Graphics, pp. 406–411, September 2009
22. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: The ko-per intersection laserscanner and video dataset. In: 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 1900–1901, October 2014