Efficient large-scale graph data optimization for intelligent video surveillance

Quanhong Shang¹, Shujun Zhang¹, Yanbo Wang², Chen Sun³, Zepeng Wang⁴ and Luming Zhang⁴

¹State Grid Zhejiang Electric Power Company
²State Grid Zhejiang Electric Power Company Information & Telecommunication Branch
³Zhejiang Huayun Information Technology CO., LTD
⁴Department of CSIE, Hefei University of Technology

Abstract. Society is rapidly accepting the use of a wide variety of cameras Location and applications: site traffic monitoring, parking Lot surveillance, car and smart space. These ones here the camera provides data every day in an analysis Effective way. Recent advances in sensor technology Manufacturing, communications and computing are stimulating The development of new applications that can change the traditional Vision system incorporating universal smart camera network. This Analysis of visual cues in multi camera networks makes wide Applications ranging from smart home and office automation to large area surveillance and traffic surveillance. In addition, dense Camera networks, most of which have large overlapping areas of cameras. In the view of good research, we focus on sparse camera networks. One Sparse camera network using large area surveillance. As few cameras as possible, most cameras do not overlap Each other’s field of vision. This task is challenging Lack of knowledge of topology Network, the specific changes in appearance and movement Track different opinions of the target, as well as difficulties Understanding complex events in a network. In this review in this paper, we present a comprehensive survey of recent studies Results to solve the problem of topology learning, Object appearance modeling and global activity understanding sparse camera network. In addition, some of the current open Research issues are discussed.

1. Introduction
2011 terrorist attacks in the United States in September Lead to review of security measures to see the use of a wide range of surveillance tasks for camera networks become one important research topic. All kinds of valuable science and engineering applications based on the development of effective the use of multi camera network. Under the assumption of a dense With the large angle crossing field camera network (FOV) Camera, early researchers use geometric information Calibrate the camera and reconstruct the shape and trajectory Objects in 3D space 1,2,3,4,5. Despite intensive cameras In recent years, the network has been well studied The network still has a challenging problem as it is It is necessary to locate, track and analyze the target in a wide range Use as many cameras as possible. Camera in a sparse camera the network does not necessarily have overlapping FOVs. An ideal sparsity Camera network monitoring system should generate tracking A sequence of independent moving objects The amount or extent of a camera in a network Overlap between different FOVs [6].

To successfully carry out automatic video surveillance in a sparse camera network, several challenges must be tackled. The first key challenge is track correspondence modeling to meet the requirement of accurate maintenance of person identity across different cameras. The goal of track correspondence modeling is to estimate which tracks result from the same object even though the tracks are captured by
different cameras and at different times. Ideally, by combining the local trajectories of individuals obtained from different cameras, the activities of the individuals can be understood in a global view (Fig. 1). The identification of corresponding tracks is made more difficult by the changes in the appearance of an individual from the FOV of one camera view to that of another. Causes of these changes include variations in illumination, pose and camera parameters.

The second key challenge for a sparse camera network-based video surveillance system is how to learn the relationship between the cameras. Spatial topology refers to distribution and linkage and is utilized to predict the trajectories of targets in a sparse camera network. For example, if a target disappears from the FOV of camera 1, then knowledge of the position of all cameras may allow the inference that the exit from camera 1 is linked to the entrance of the target into the FOV of camera 2 and camera 3. This inference may be supported by similarities in the appearance in the different images. Such inferences depend on the topological relationships between the cameras in the network. As shown in Figure 2, these relationships may be of several different types [6]. If the camera network is dense, there are various methods for multi-camera tracking with overlapping FOVs and, as discussed in the literature, there are various geometric matching methods for manually calibrating the cameras [1], [2], [7], [8], [9], [10]. The methods used in these papers can compute the transformation between 2D images coordinates and 3D spatial coordinates for a ground plane. However, it is unrealistic to expect that the FOVs of all the cameras always overlap one another, and the need for all cameras to be calibrated to the same ground plane usually cannot be satisfied. For instance, a target may appear in the FOV of two different cameras initially and may disappear from one of the FOVs, reappearing after some time in the FOV of a third camera. Several approaches have recently been presented to estimate the spatial topology and linkage in a camera network [11], [12], [13], [14], instead of directly calculating the 3D position in the ground plane, as in [1], [2] and [7], [8], [9], [10].

Figure 1. Tracking targets in a camera network by combining the local trajectories of each individual.

The third challenge is global activity understanding. For an intelligent video surveillance system, it is not enough to only track the targets without further analysis. Through the global automatic and comprehensive analysis of targets in a sparse camera network, the activities of the targets are understood and anomalous events are detected. Several previous surveys, e.g., [15], have discussed activity understanding. A sparse camera network-based video surveillance system usually covers a much larger area, and hence it usually provides more information about the targets trajectory and activities than a single camera or a conventional dense camera network monitoring a small area. However, because the camera network is sparse, it is necessary to carry out global activity understanding from a new point of view.

2. Inter-Camera Tracking Correspondence
Sparse camera networks, the relationship between the two Camera, system, topology and transition
period Camera, is based on the camera identification. In the case of sparse camera network vision monitoring, therefore, it is desirable to determine whether a target the interest of a camera has been observed elsewhere Network. This problem can be easily solved by conventional dense camera network by simply matching the 3D position Target location for each candidate, as camera the network is well calibrated. However, a sparse camera Network based visual surveillance systems typically contain a number. There is no overlap of views because of camera requirements. It is impractical to maintain calibration. The most basic idea Existing methods for non-overlapping camera configurations Target matching between cameras is formulated as recognition. The problem of the system, the target is described Visual appearance cues compared to candidates in video Captured by other cameras in a sparse network. This kind of the recognition problem is often referred to as the “object” Recognize ”, object re identification ’or’ appearance Modeling”. When modeling a person’s goals in an A general assumption for sparse camera networks is Personal clothes remain unchanged on camera. However, even if to accept this assumption, the camera-tracking task remains Challenges due to lighting, pose and camera changes. Parameters in different cameras on the network. * The general camera-tracking framework can be Decomposition into three layers:

• Low level features for visual appearance description: extraction of concise features.
• Integration of low-level features.
• Inter-camera target identification.

2.1. Low Level Features for Visual Appearance Description

The low-level features for visual appearance description can be divided into two categories: global visual features and local visual features. The global visual features encode the target as a whole. The local visual features describe the target as a collection of independent local descriptors, e.g., local patches and Gabor wavelets. A global feature usually contains comprehensive and rich information, so it is very powerful if it is accurate. On the other hand, it is very sensitive to noise, partial occlusions, viewpoint changes and illumination changes. In contrast with global features, local features are less sensitive to these factors; however, as the features are extracted locally, some information, especially spatial information, may be lost.

Global visual features: A global feature encodes the tracking target using a single multidimensional descriptor. For instance, Huang et al. used the size, velocity and mean value of each channel in HSV color space to describe the appearance of vehicles. In many applications, the target is more complex than a vehicle, and thus stronger descriptors are needed. General global features used to represent a target’s appearance include the global histogram GMM and newly defined global descriptors for human reidentification, e.g., panoramic appearance map (PAM) and color position.

a) Global histogram: Histogram-based methods can represent large amounts of information. They are efficient, easy to implement, and have a relatively high tolerance to noise.

One of the earliest attempts to represent targets by histograms was carried out by Kettner and Zabih. Orefej et al. also extracted histograms of HSV color values of pixels in the target area. PCA is applied to 6,000 histograms of HSV color values and HOG descriptors, and the eigenvectors corresponding to the top 30 eigenvalues are extracted as the final representation. Morioka et al. calculated the covariance matrix of color histograms and projected the set of color histograms onto the eigenspace spanned by the eigenvectors corresponding to the first d largest eigenvalues. Lin and Davis proposed a normalized rgs1 color to deal with illumination changes since the independence of chromaticity from brightness in this color space allows using a multivariate Gaussian density function to cope with the differences in brightness. Nakajima et al. described a system that learns from examples to recognize persons in images taken indoors. Images of full-body persons are represented by the RGB color histograms (Fig. 2(a)), the normalized color histograms and the shape histograms.

b) Newly defined global descriptors for human reidentification: Gandhi et al. developed the concept of a PAM (Fig. 2(c)) for performing person re-identification in a multicamera setup. A PAM centering on the person’s location is created with the horizontal axis representing the azimuth angle and the vertical axis representing the height. PAM models the appearance of a person’s body as a convex generalized cylinder. Each point in the map is parameterized by the azimuth angle and height, and the radius of the cylinder is treated as constant. This allows PAM extracts and combines information from all the cameras.
that view the object features to form a single signature. The horizontal axis of the map represents the azimuth angle with respect to the world coordinate system, and the vertical axis represents the object height above the ground plane. Cong et al. proposed a color-position histogram (Fig. 2(b)), in which the silhouette of the target is vertically divided into equal bins and the mean color of each bin is computed to characterize that bin. Compared to the classical color histogram, it consists of spatial information and uses less memory. Moreover, this new feature is a simpler and more reliable measurement for comparing two silhouettes for person re-identification.

![Figure 2. Examples of global representation features: (a) global histogram of RGB color, (b) color position, (c) panoramic appearance map.](image)

Local visual features: a) Local visual feature detection: Examples of local visual features include corners, contour intersections and pixels with unusually high or low gray levels. Contour intersections often take the form of bi-directional signal changes and the points detected at different scales do not move along a straight bisector line; hence, corners are often detected using local high curvature maxima. Lowe proposed localizing points at local scale-space maxima of the difference of Gaussians. Doll’ar applied Gabor filters separately in the spatial and temporal domains. By changing the spatial and temporal size of the neighborhood in which local minima are selected, the number of interest points can be adjusted. Laptev extended the Harris-Stephens corner detector to space-time interest points where the local neighborhood has significant variation in both the spatial and the temporal domains.

b) Local visual feature descriptor: An image can be described using a collection of local descriptors or patches that can be sampled densely or at points of interest. Compared to extracting local descriptors at the interest point, dense sampling retains more information, but at a higher computational cost. A performance evaluation of different local visual feature descriptors has been given by Mikolajczyk and Schmid. We discuss the different types of descriptors as follows.

1) Distribution descriptors: Distributions of intensity, gradient, and shape are widely used as local descriptors. Zheng et al. extracted SIFT features for each RGB channel at each pixel to associate a group of people over space and time in different cameras. These features are clustered and quantized by K-means to build a codebook. The original image is then transformed to a labeled image by assigning a visual word index to the corresponding feature at each pixel of the original image. Wang et al. extracted Histogram of Gradients (HOG) features in the Log-RGB space. They argued that taking the gradient of the Log-RGB space had an effect similar to homomorphic filtering, and made the descriptor robust to illumination changes. Hamdoun et al. utilized a method in known as SURF to detect interest points; it was Hessian-based and used an integral image for efficient computation. The descriptors SURF are 64 dimension vectors which coarsely describe the distribution of Haar-wavelet responses in sub-regions around the corresponding pixels of interest. Laptev and Lindeberg presented a local spatio-temporal
descriptor that included position dependent histograms and the PCA-based dimensionality reduced spatial-temporal gradients around the spatio-temporal interest points.

2) Frequency descriptors: The spatial frequencies of an image carry important texture information and can be obtained by the Fourier transform and the wavelet transforms, such as Haar transform and Gabor transform. In contrast to the Fourier transform and the Haar transform, in which the basis functions are infinite, the Gabor transform uses an exponential weighting function to localize the decomposition of an image region into a sum of basis functions. The region to analyze is first smoothed using a Gaussian filter, and the resulting region is then transformed with a Fourier transform to obtain the time-frequency analysis. Gray et al. used two families of texture filters, taken from Schmid and Gabor respectively, on eight color channels corresponding to the three separate channels of the RGB, YCbCr, and HSV color space, in which the used eight color channels contain only one of the luminance (Y) and V channels.

3. Conclusions
This paper formulated a novel aggregation-based camera network for representing GSP, based on which multiple aesthetic communities can be online discovered according to human visual perception. More specifically, by constructing the GSP of each photo using active learning, salient regions discovery and gaze shifting sequence prediction are incorporated into an aggregation-based deep CNN. Based on this, a so-called deep perception graph is built which effectively measures the aesthetic discrepancy between users, and multiple aesthetic communities can be identified using graph shift. Thereby, an online GMM describes the aesthetic distribution toward each community, based on which a series of media applications (e.g., retargeting and aesthetic assessment) can be tackled in a unified framework.

4. References
[1] S. N. Sinha, M. Pollefeys, and L. McMillan, Camera network calibration from dynamic silhouettes, in Proc. of CVPR, pages: 195–202, 2004.
[2] S. N. Sinha and M. Pollefeys, Multi-view Reconstruction using Photoconsistency and Exact Silhouette Constraints: A maximum-Flow Formulation, in Proc. of ICCV, pages: 349–356, 2005.
[3] R. Szeliski, Rapid Octree Construction from Image Sequences, CVGIP: Image Understanding, 58(1), pages: 23–32, 1993.
[4] J. Starck and A. Hilton, Surface Capture for Performance-Based Animation, IEEE Computer Graphics and Applications, 27(3), pages: 21–31, 2007.
[5] P. R. S. Medonca, K.-Y. K. Wong, and R. Cipolla, Epipolar Geometry from Profiles Under Circular Motion, IEEE T-PAMI, 27(3), pages: 21–31, 2007.
[6] C. Stauffer and K. Tieu, Automated Multi-camera Planar Tracking Correspondence Modeling, in Proc. of CVPR, pages: 259–266, 2001.
[7] D. C. Brown, Close-Range Camera Calibration, Photogrammetric Engineering, 37(8), pages: 855–866, 1971.
[8] G. Champleoux, S. Lavalle, P. Sautot, and P. Cinquin, Accurate Calibration of Cameras and Range Imaging Sensor: The Npbs Method, in Proc. of ICRA, pages: 1552–1557, 1992.
[9] Q. Cai and J. K. Aggarwal, Tracking Human Motion in Structured Environments Using a Distributed-camera System, IEEE T-PAMI, 2(11), pages: 1241–1247, 1999.
[10] R. T. Collins, A. J. Lipton, H. Fujiyoshi, and T. Kanade, Algorithms for Cooperative Multisensory Surveillance, Proc. of IEEE, 89(10), pages: 1456–1477, 2001.
[11] R. Farrell, D. Doermann, and L. S. Davis, Learning Higher-Order Transition Models in Medium-Scale Camera Networks, in Proc. of ICCV, pages: 1–8, 2007.
[12] Y. A. Sheikh and M. Shah, Trajectory Association across Multiple Airborne Cameras, IEEE T-PAMI, 32(2), pages: 361–367, 2008.
[13] D. Makris, T. Elis, and J. Black, Bridging the Gaps between Cameras, in Proc. of CVPR, pages: 205–210, 2004.
[14] O. Javed, Z. Rasheed, K. Shafique, and M. Shah, Tracking across Multiple Cameras with Disjoint Views, in Proc. of ICCV, pages: 952–957, 2003.
[15] O. Williams, A. Blake, and R. Cipolla, Sparse Bayesian Learning for Efficient Visual Tracking, IEEE T-PAMI, 27(8), pages: 1292–1304, 2005.