Tracking Hand-off in Large Surveillance Networks

Alex Cichowski, Christopher Madden, Henry Detmold, Anthony Dick
Anton van den Hengel, Rhys Hill
The Australian Centre for Visual Technologies
School of Computer Science
University of Adelaide
SA, Australia 5005
Email: {alexc,cmadden,henry,ard,anton,rhys}@cs.adelaide.edu.au

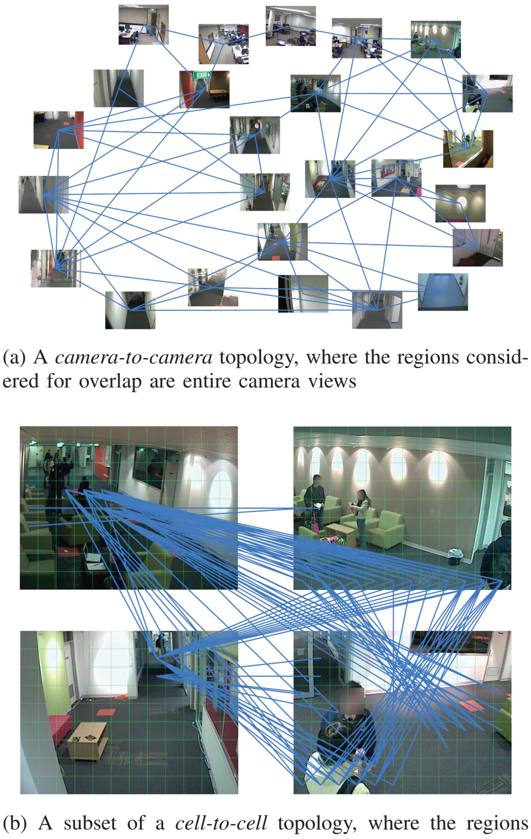
***Abstract*—This paper investigates the use of pairwise camera overlap estimates for supporting target tracking across large networks of surveillance cameras. We compare the use of camera overlap topology information to a method based on matching target appearance histograms, and also evaluate the effect of combining both methods. Tracking accuracy results are reported in terms of precision and recall for a 24 camera network. Camera overlap information is shown to deliver significant advantages for tracking when compared with simply matching target appearance histograms, due to its robustness to low quality imagery. We show empirically that this is the case even for automatically derived overlap estimates containing errors.**

I. Introduction

Automated analysis of video surveillance data is becoming an increasingly important tool for improving public safety and security, supporting both real-time security interventions and post-event analysis. A significant body of research has investigated intelligent analytical software for video surveillance [1], with promising results in the areas of object detection and tracking within single camera views [2]. However, automated analysis is of most use when applied across large networks of cameras, as such networks are infeasible to monitor manually.

This paper investigates the application of automatic methods for tracking individuals across cameras within a surveillance network. Typically, such methods build on existing image analysis techniques for a single camera, including target segmentation [3] and tracking [2]. With the exception of [4], existing work in the area of tracking with multiple cameras focuses on small installations of less than ten cameras [1], whereas modern surveillance systems may have many hundreds or thousands of cameras. We use standard methods for target segmentation and single camera tracking, and instead focus on the task of implementing *hand-off* by joining together the tracks that have been extracted from individual views, as illustrated in Figure 2. We evaluate different approaches to implementing hand-off based on a manually determined *hand-off ground truth*, consisting of 160 randomly chosen test cases, with each case recording the complete set of links to visible instances of the same person in other cameras at the same time.

We consider two main approaches to implementing hand-off: one based on target appearance, and the other on camera overlap topology. Target appearance is represented by an RGB colour histogram, as described in Section II. Appearance based matching has the advantage that it is already used for single camera tracking, and is straightforward to extend to multiple camera tracking. However it has the disadvantage that target appearance can change significantly between viewpoints and due to errors in segmenting the target from the background.


(a) A *camera-to-camera* topology, where the regions considered for overlap are entire camera views

(b) A subset of a *cell-to-cell* topology, where the regions considered for overlap are provided by the cells of a 12x9 grid

Fig. 1: Example topologies for the surveillance data used in this work, estimated using [4], [5]

An *overlap topology* describes the relationships between camera views, and can be used to extend single camera tracking through hand-offs of tracks between cameras that are overlapping. In this work we define a camera overlap topology as the set of links between either pairs of cameras, or pairs of regions within camera views which observe the same area.

978-1-4244-4698-8/09/$25.00 ©2009 IEEE

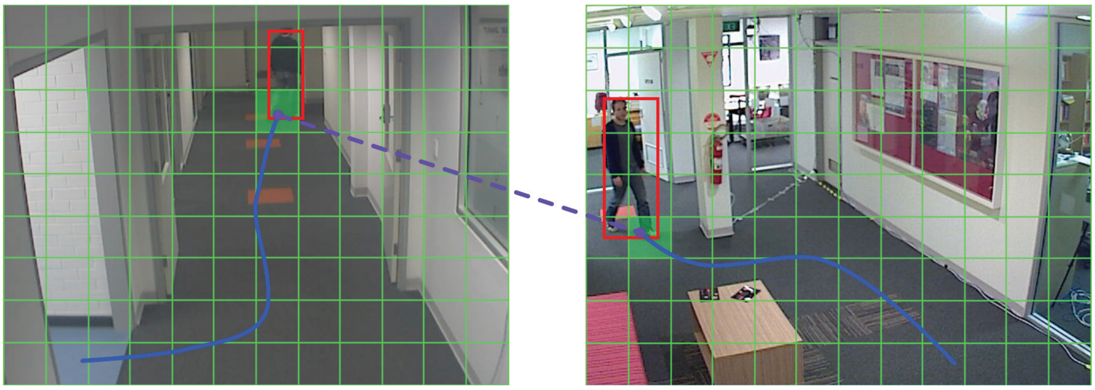


Fig. 2: Example showing a hand-off link (dashed purple line) between two single camera tracks (solid blue lines), as well as the summarisation of individuals to 12x9 grid cells (green shading) as employed in the topology estimation method used

For example, Figure 1 shows an overlap topology estimated using [4], [5] at two different levels of detail. Despite the potential of this information, little attention has been paid to means of obtaining it automatically or to using it for cross camera tracking. For large networks it is difficult to obtain these links reliably, but once obtained topology information is able to dramatically shrink the space to be searched when matching tracks of an individual in different cameras.

In Section II we provide details of the appearance model and appearance comparisons. Creation of a camera overlap topology is detailed in Section III. In Sections IV and V we evaluate the accuracy of tracking hand-off implementations using both appearance and topology in isolation, and then in combination.

II. Appearance Modelling and Matching

Appearance, defined by the colours of the pixels representing a person in an image, is widely used for matching and tracking individuals in video surveillance. Measures based on appearance include correlation of the patch itself, correlating image gradients, and matching Gaussian distributions in colour space. In this paper we use an RGB histogram to represent the appearance of each target. Histograms are a popular choice because they allow for some distortion of appearance: they count only the frequency of each colour rather than its location, and they quantise colours to allow for small variations. They are also compact and easy to store, update and match. Here we use an 8x8x8 RGB histogram that is equally spaced in all three dimensions, totalling 512 bins. Histogram matching is based upon the Bhattacharyya coefficient [6] to determine the similarity of object appearances. If we let i sequentially reference histogram bins, then the similarity of normalised histograms A and B is given by:

$$Similarity = \sum_{i=1}^{512} \sqrt{A_i \cdot B_i} \tag{1}$$

A decision on whether two targets match can then be reached by thresholding this similarity measure. Other similarity measures, such as Kullback-Leibler divergence [7], produced similar results.

III. Topology Estimation Methods

The topology estimation method we use [4], [5] is based on statistical analysis of scene activity to automatically evaluate possible links in the overlap topology. Each camera view is broken into a set of cells, whose occupancy at each frame is defined based upon whether they contain a detected object. In this work we consider two cases: the entire view is considered as a single large cell; or each view is divided into a grid of 12x9 cells as shown in Figure 2. Objects are segmented from the static background using [3], which produces a binary mask defining each object's extent. The location of each object is then summarised as the midpoint of the base of the bounding box of this mask. At each frame the occupancy of cells is correlated across all cameras. This process is performed over time to build up evidence for links based upon the activity within the scene, without the need for matching across cameras or manual intervention.

Overlap estimates can be generated by applying the mutual information measure [8] to cell occupancy data, as it represents the amount of dependence between two given variables [9]. The mutual information, $I(X;Y)$, for two occupancy cells represented as binary random variables X and Y is given by:

$$I(X;Y) = \sum_{y \in Y, x \in X} p_{XY}(x,y) log \left(\frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} \right) \tag{2}$$

High values of $I(X;Y)$ indicate a high dependence, and thus indicate a high probability of the cells overlapping. $p_{XY}(x,y)$, $p_X(x)$ and $p_Y(y)$ are estimated according to the accumulated correlation counts [5]. A distributed implementation of a similar method capable of scaling to many thousands of cameras is outlined in [10].

Pairwise occupancy analysis methods such as this generally require practical measures to address the following assumptions:

1) Cells are either completely coincident or completely disjoint in physical space
2) All cameras are synchronised, ensuring frames are captured at exactly the same time
3) The foreground detection module produces very reliable object segmentation results
4) The ground truth topology does not change over time

Assumption 1 is addressed by accumulating evidence over time. Cells that are not completely coincident will still exhibit a statistical correlation in occupancy. Additionally, a cell is allowed to link to multiple cells in other camera views, catering for poor cell alignment in physical space. We have implemented time padding as suggested in [5], allowing some relaxation of the synchronisation required by assumption 2. Violations of assumption 3 are mitigated by accumulating evidence over time, although these remain a likely source of errors. For the dataset used in this paper, the ground truth camera overlap topology does not change, satisfying assumption 4.

IV. Evaluating Tracking Hand-off

This work evaluates camera overlap topology and appearance comparisons for their ability to provide tracking hand-off between overlapping cameras. The evaluation performed here examines the practical utility of each method for correctly detecting simultaneous tracks of the same individual. The accuracy of each method in itself is *not* evaluated in this paper, although clearly a method's ability to support hand-off will be a product of its underlying accuracy.

Hand-off based on appearance matching can be implemented simply by comparing the current target's appearance descriptor with those of all other potential targets in the system at the given time, and selecting targets meeting a pre-defined appearance similarity threshold. Hand-off based on overlap topology can be implemented by searching for potential targets in each of the regions overlapping that of the current target. For fine grained cell-to-cell overlap topologies, overlaps within a certain distance from the current target may also be considered, as well as potential targets within a certain distance from an overlapping region in another camera.

For large camera networks, the number of false matches resulting from the use of appearance matching alone will generally increase with the number of cameras in the system. In practice, for large networks this will need to be mitigated by applying at least a limited form of camera topology information, and only searching for appearance matches in the same cluster of cameras—for example, in the same building. Hand-off based on appearance matching across the entire network or within clusters may then be further refined by combining it with the use of overlap topology and searching only overlapping regions for targets of matching appearance.

To perform the evaluation, we draw a random sample of detections of people at particular times. For each sample, we manually identify all other simultaneous detections of the same person, in order to obtain a set of ground truth links between detections in different cameras. No restriction is placed on the detections that are sampled in this way; they may occur anywhere within tracks that have been found within a single camera. This reflects the fact that hand-off does not just link the end of one track with the start of another, but rather needs to link arbitrary points from track pairs depending on when a person becomes visible or disappears from another camera.

This is quite a stringent test because it is based solely on instantaneous "snapshots" of the network: no temporal information is used. In reality, information from temporally adjacent frames may be available to assist with deciding on camera hand-off.

A hand-off link is only added to the ground truth where an automatically detectable portion of that object is seen in two cameras. Objects that are very small and might have difficulty being segmented are noted, but not included in the ground truth. Where an object is observed in only one camera, the fact that the case does not form a hand-off link is recorded as part of the ground truth. Cases where object segmentation has produced extremely large errors are discarded, as are cases where the object is almost fully occluded.

This ground truth for tracking hand-off can then be used to evaluate the automatically determined tracking hand-off links. In this paper we present the results as precision-recall (P-R) curves, a standard metric for determining accuracy compared to ground-truth in the information retrieval context [11]. Given a classifier with true positives, TP, false positives, FP, and false negatives, FN, the precision P is given by:

$$P = TP/(TP + FP) \tag{3}$$

with the recall, R, given by:

$$R = TP/(TP + FN) \tag{4}$$

Different values of precision and recall are obtained by varying a threshold appropriate to each method. For appearance matching, the appearance similarity threshold is varied. For topology based matching, the threshold determines the search distance used when searching for overlaps around the current target, and for potentially matching targets around overlapping regions. For hand-off based on both appearance matching and overlap topology, this search distance is set to a constant value and only the appearance similarity threshold is varied.

V. Results

This section outlines the results of experiments to investigate tracking hand-off across a set of 24 surveillance cameras in which there is a degree of overlap between cameras. The dataset is described before the evaluation results are provided for using a topology only, appearance only, and a combination of both. Three overlap topologies are used: manually determined camera-to-camera ground truth; an automatically generated camera-to-camera topology; and an automatically generated 12x9 cell-to-cell topology. With 24 cameras and 12x9 cell grids, the full system consists of over 2500 cells, for which it was too time consuming to manually determine ground truth.

A. 24 Camera Dataset

The 24 camera dataset was captured from a set of digital surveillance cameras placed according to the office floor plan shown in Figure 3. The cameras were placed to provide a high degree of coverage throughout the area, and consisted of a mix of normal and high resolution cameras, as well as wireless cameras. The cameras were not optimised for lighting or quality of the background model they might provide, leading to significant realistic challenges in segmenting foreground objects. The area was recorded for a duration of over four and a half hours, including periods of high and low activity. The floor plan and footage obtained were manually analysed to determine the camera-to-camera ground truth topology, defining which cameras have at least partially overlapping views.

Although the 24 camera network used here does not reach the scale of hundreds or thousands of cameras, it is larger than many typical systems considered, and is used to represent overlap such as that which might be found in real systems at a much larger scale while still being small enough to allow construction of camera-to-camera overlap ground truth. Performance of overlap topology based hand-off is thus expected to be indicative of performance for larger, realistic networks of overlapping cameras. The dataset is also of a size equivalent to a typical cluster of cameras in a larger network, so that performance of appearance matching based hand-off will be indicative of appearance matching performed over clusters of cameras in a larger network.

A set of 500 object detections were randomly chosen from the many hours of footage available. These provided 160 usable test cases where segmentation errors were not extreme, and the individual was not significantly occluded. Whilst such segmentation errors and occlusions often occur in real systems, they are difficult to evaluate based on a single time instant. They are likely to be better evaluated or filtered out using information gained over time during a full-scale single camera tracking process, which is outside the scope of this paper. More cases could be manually extracted, but an increase in the number of usable cases from 100 to 160 did not significantly change the results achieved, suggesting that the sample produces a stable estimate of the performance of each method.

After analysing the results it was found that camera 18, situated with limited overlap in its own room, provided significant erroneous links both within the topology estimation and in the tracking hand-off evaluation. This particular camera was wireless and had significant reception problems. This led to poor frame availability and delays in the video stream. The effects of this poor quality camera have been analysed to determine the problems that can arise with such cameras.

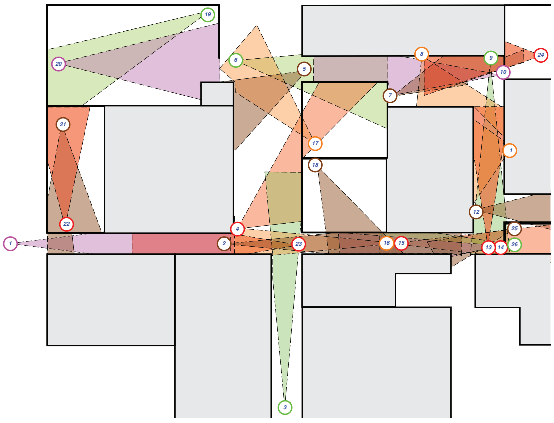


Fig. 3: Floor plan showing the position of the camera views within the dataset

B. Tracking Hand-off

Figure 4 presents the tracking hand-off results in terms of precision and recall, for appearance matching and using each of the three overlap topology estimates. The results clearly show that using the appearance model alone has much lower precision when detecting tracking hand-offs than using any of the topology estimates alone. However, it does have some effect, as it performs better than randomly “guessing” the tracking hand-off links. This low precision could be due to a number of factors that influence appearance measures, such as segmentation errors and illumination effects. Appearance can be dramatically affected by segmentation errors, especially where portions of the background are incorrectly included. Illumination may influence the appearance of an object differently depending upon the direction that an object is observed from. For example, light from a doorway could make the front of an object appear brighter than observations of the same individual made from the rear. Lights creating strong shadows also contribute to illumination and segmentation problems. Additionally, some cameras in the dataset were behind glass windows, which can introduce reflections. This had a minor effect on segmentation, but influenced object appearance. Regardless of these effects, distinguishing between individuals wearing similar clothing is difficult using appearance.

Estimates of tracking hand-off links based on the automatically generated camera-to-camera topology are similar to those based on the camera-to-camera ground truth topology, with the unreliable camera removed. The precision was considerably worse with this camera included, as it introduced a number of false links in the topology. Using overlap at a camera level alone is problematic, however, because objects may be observed in non-overlapping regions of overlapping views and erroneously be considered to be the same object. An increased spatial resolution of overlap was achieved using 12x9 cells with overlap registered between particular portions

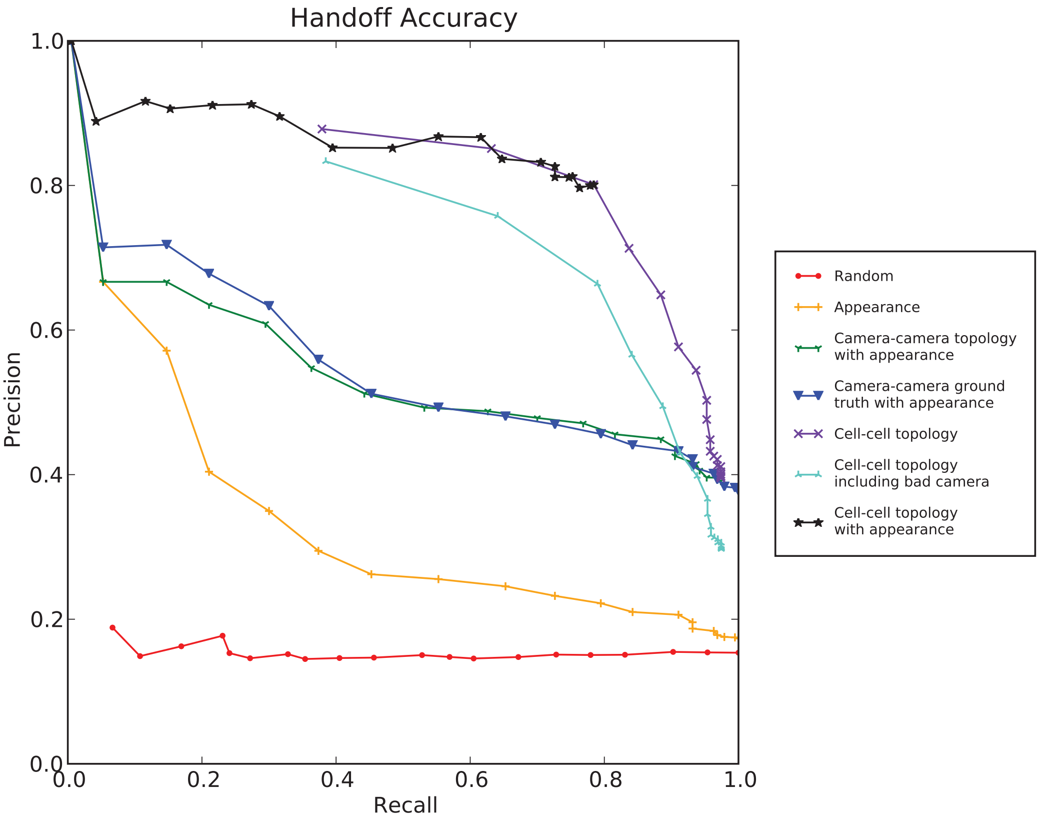


Fig. 4: P-R curves demonstrating tracking hand-off results

of overlapping camera views. The results clearly show that using this finer spatial resolution outperforms the tracking hand-off achieved using even the ground truth camera-to-camera topology. Thus, more cases where different objects are seen in non-overlapping portions of the camera views are correctly excluded in the hand-off search process.

One of the more surprising results is that the combination of the appearance model and a cell-to-cell topology does not significantly increase the precision for a given level of recall. This would suggest that incorrect links which fit the topology model are not correctly excluded by using appearance. A more complex appearance model that accumulated information over time, or improvements in the accuracy of object segmentations may improve the accuracy of appearance; however unless the appearance model or extraction technique can compensate for illumination changing the perceived object appearance, these are still likely to be very difficult cases in real surveillance environments. The accuracy of determining appearance similarity also depends significantly upon the individual clothing that is worn. In practice, many people wear similar clothing, often with a significant amount of black or dark colours. If the appearance model can only capture large differences in appearance and clothing, then it may be difficult to accurately discriminate between individuals; however capturing nuances in appearance can lead to large data structures that can be even more sensitive to illumination and segmentation errors.

By contrast, topology information is derived solely from object detection in each cell, and thus does not depend on the appearance of people in the video. It is less sensitive to these issues, and able to obtain a topology that is accurate enough to be useful for tracking even in environments where the camera struggles to accurately capture the appearance of each person.

The effect of removing the poor quality wireless camera is also indicated in the graph. It is clear that the precision of the topology is reduced by including such cameras, as the time delay allows for evidence to arise supporting overlaps that do not occur. The precision difference for appearance-only tracking was much less and is not shown in the graph. This is because the appearance is not as significantly affected by time delays, so removing the less reliable camera does not have much of an affect.

The 24 camera network evaluated here provides a realistic representation of overlap such as that which might be found in large scale real systems. Typically overlap occurs between two or three cameras, ensuring that the number of correct links in an overlap topology does not significantly grow with the number of cameras and making hand-off based on overlap topology scalable to very large systems. Such characteristics of the clusters of cameras that overlap cannot be exploited by using appearance features alone, and an increase in cameras in the system will likely increase the chances that people of similar appearance will be seen, increasing the likelihood of incorrect matches. A combination of the appearance and overlap topology methods, where appearance features are used to discriminate more accurately between matches based on overlap topology links, could also scale to very large networks, though the results presented here suggest such a system would require a more sophisticated appearance model to go beyond the accuracy afforded by a method based on cell-to-cell topology alone.

VI. Conclusion

Video surveillance is increasingly being used to safeguard the public, and public institutions, from terrorism and other attacks. A key component for the effectiveness of video surveillance is the ability to continuously observe and track the movements of individuals or groups of interest. This task is difficult for operators of small scale systems, and becomes infeasible in large scale systems with hundreds or thousands of cameras, even for groups of highly trained operators, due to the large number of possible transitions between camera pairs.

This paper has evaluated appearance and topology information as options for providing tracking hand-off between cameras, enabling continuous tracking across overlapping cameras. The results indicate that appearance alone does not provide reliable support for tracking hand-off between cameras, for a variety of reasons including segmentation errors, scene illumination differences, and the similarity in appearance of many of the clothes that people typically wear. By contrast, camera-to-camera ground truth topology provides a higher degree of hand-off precision, with an automatically estimated camera-to-camera topology providing a similar level of accuracy. Higher resolution topologies restrict the search space further, leading to improved precision. For the appearance model used and the dataset evaluated, the use of both high resolution (cell-to-cell) topology and appearance features together has a similar or lower level of precision than using the equivalent topology alone, suggesting that appearance has not provided significant additional utility in distinguishing correct hand-off links. Appearance based tracking hand-off is also not likely to scale well to very large camera networks as all targets observed at a time instant will need to be compared. In contrast, hand-off based on overlap topology scales in complexity according to the degree of overlap between cameras, rather than the number of cameras in the system.

Fully automatic network wide tracking in crowded environments remains an unsolved problem. However, current technology can provide significant assistance to operators by suggesting likely tracking hand-offs. For large surveillance systems with many hundreds or thousands of cameras, having a system suggest the next view of an object could assist operators in both live analysis and forensic investigations.

In future work we plan to extend the evaluation to include detections over a time interval, rather than operating only instantaneously. Based on this we hope to improve our estimates of target segmentation and appearance. We also plan to evaluate topology data for use in tracking across a greater variety of camera networks to evaluate its general effectiveness.

References

[1] M. Valera Espina and S. A. Velastin, “Intelligent distributed surveillance systems: A review,” *IEE Proceedings - Vision, Image and Signal Processing*, vol. 152, no. 2, pp. 192–204, 2005.

[2] A. Yilmaz, O. Javed, and M. Shah, “Object tracking: A survey,” *ACM Computer Surveys*, vol. 38, no. 13, pp. 1–45, 2006.

[3] C. Stauffer and W. E. L. Grimson, “Learning patterns of activity using real-time tracking,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[4] A. van den Hengel, A. Dick, and R. Hill, “Activity topology estimation for large networks of cameras,” in *AVSS ’06: Proc. IEEE International Conference on Video and Signal Based Surveillance*, 2006, pp. 44–49.

[5] R. Hill, A. van den Hengel, A. R. Dick, A. Cichowski, and H. Detmold, “Empirical evaluation of the exclusion approach to estimating camera overlap,” 2008, proceedings of the International Conference on Distributed Smart Cameras.

[6] A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions,” *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.

[7] S. Kullback, “The kullback-leibler distance,” *The American Statistician*, vol. 41, pp. 340–341, 1987.

[8] E. Sommerlade and I. Reid, “Cooperative surveillance of multiple targets using mutual information,” in *ECCV Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, October 2008.

[9] A. Cichowski, C. Madden, A. van den Hengel, R. Hill, H. Detmold, and A. Dick, “Contradiction and correlation for camera overlap estimation,” in *AVSS International Conference on Advanced Video and Signal Based Surveillance*, September 2009.

[10] H. Detmold, A. van den Hengel, A. R. Dick, A. Cichowski, R. Hill, E. Kocadag, Y. Yarom, K. Falkner, and D. Munro, “Estimating camera overlap in large and growing networks,” in *2nd IEEE/ACM International Conference on Distributed Smart Cameras*, 2008.

[11] V. Raghavan, P. Bollmann, and G. S. Jung, “A critical investigation of recall and precision as measures of retrieval system performance,” *ACM Trans. Inf. Syst.*, vol. 7, no. 3, pp. 205–229, 1989.