Contradiction and Correlation for Camera Overlap Estimation

Alex Cichowski Christopher Madden Anton van den Hengel Rhys Hill
Henry Detmold Anthony Dick

Australian Centre for Visual Technologies
School of Computer Science
The University of Adelaide
http://www.acvt.com.au/research/surveillance/

Abstract

An accurate estimate of camera overlap is a key enabler for efficient network-wide surveillance processing (e.g. inter-camera tracking), especially in large-scale surveillance networks. Techniques based on contradictions in pair-wise occupancy data, such as the exclusion approach, have advantages in robustness and efficiency that make them particularly well suited for large surveillance networks. Correlation techniques share some of these advantages, but have a better understood statistical basis. This paper evaluates a set of contradiction and correlation techniques, using a novel metric, search space precision-recall. This metric reflects the activity-based overlap estimation required for camera handover, such as would be used in inter-camera tracking. Results are reported for a range of networks, including a 24-camera network set up in an office space, where the exclusion estimator showed the best performance.

1. Introduction

Video surveillance networks are increasing in scale with networks of multiple thousands of cameras common. For example, the Washington D.C. police have a network of over 5,000 cameras [4] and Singapore's Local Transport Authority runs a network of nearly 6,000 cameras [5]. The scale of these networks demand software assistance for humans to make sense of the vast amounts of data produced. Assistance is required both for live monitoring, and for forensic analysis of events of interest. Computer vision research has made significant progress in automating the processing of this data on the small scale (see [10] for a survey), but there has been less progress in scaling these techniques to the much larger networks now being deployed.

Recent approaches for analysing large surveillance networks have focused upon providing important network-wide services supporting visual processing. A key example is the estimation of the camera overlap of a surveillance system, or its related activity topology [11], which facilitate processes such as inter-camera tracking. The activity topology is a graph which describes the spatial and temporal relationships between the fields of view of the network's cameras that can be obtained from observed activity in the network. An accurate estimate of an activity topology supports efficient camera handover where activity moves from one camera to another adjacent camera in the system.

An important sub-graph of the activity topology is the activity-based *camera overlap* graph, which includes only edges relating cameras having commonality (i.e. overlap) in their fields-of-view. This sub-graph is easier to obtain by relating activity across cameras. We further consider each field of view as a regular grid of regions, or *cells*, which may overlap [11]. By this we aim to produce a more refined overlap estimate to better facilitate processes such as inter-camera tracking by reducing the search space size required to find a given object in a new camera.

The main contribution of this paper is a novel method for evaluating approaches to estimating camera overlap graphs. We propose a search space-based precision-recall curve to evaluate the likely candidates to improve the estimated overlap between cameras to assist subsequent inter-camera processes. This search space-based evaluation is performed for four estimation methods on a number of networks, including a 24 camera office network.

2. Camera Overlap Estimators

This section describes several methods which have been implemented within a software framework for estimating camera overlap from object activity within a camera network. These approaches are built upon the extraction of moving objects from the background in each camera view [9] and the use of this foreground activity to estimate the relationship between cameras. As in [11], each camera's field of view is considered to be a regular rectangular grid

978-0-7695-3718-4/09 $25.00 © 2009 IEEE
DOI 10.1109/AVSS.2009.62

of cells. A cell is considered to be occupied when it contains the lowest visible point of an object observed in that frame of the camera's video stream, otherwise it is unoccupied. If a camera is inactive at any point, then its cells are considered inactive. The occupancy of cells over time can be analysed using the following methods:

- The exclusion approach [11].
- An approach based on mutual information[8].
- An approach based upon conditional entropy [1].
- The lift operator patented by Intellivid [2].

Similarities between the approaches have been factored into the software of the evaluation framework. Thus it maintains data summarising cell occupancies, and applies a selected method to analyse this data. This is done for each possible pair of cells based upon their joint occupancy. By keeping track of inactive cell states, the framework is robust to camera outages, as periods of camera inactivity do not adversely affect the information that can be derived from the active cells. The implementations reported in this paper use a single centralised implementation of the framework, it is suitable for distribution via a generic partitioning approach, such as that described in [3].

The various approaches can be implemented as either *contradiction* or *correlation* approaches. In contradiction approaches all overlap links are considered possible, with processing resulting in evidence against camera overlap. The accumulation of this evidence leads to the removal of links from the camera overlap graph. In correlation approaches all links are considered non-overlapping until processing results in evidence that they overlap. The accumulation of this evidence leads to the addition of an edge into the camera overlap graph.

For contradiction approaches, it is valid to apply processing for a given pair of cells over an arbitrary subset of the time points. A small or poorly selected subset reduces the probability that contradictions are found; however it does not reduce the evidentiary value of any contradictions that are found. This may lead to extra links in the overlap graph, though true overlaps should not be affected.

Correlation approaches build the overlap graph on evidentiary links and require the set of time points processed to be a statistically valid sample of the complete set of time points. This could be the complete set, if available. Processing only a subset of the available time points will reduce the probability of finding true links in the overlap graph, though there are likely to be less false links.

When cameras are off-line, correlation approaches may be degraded to a greater degree than the available signal, whilst degradation of contradiction approaches are likely to be more proportional to the reduction in signal. This suggests that contradiction approaches could be more robust to missing camera signals than correlation approaches.

2.1. Exclusion Estimator

Exclusion is based on the fact that if, at given points in time, cell i is observed to be occupied and cell j is observed to be unoccupied, then cells i and j do not overlap (*i.e.* there is evidence contradicting overlap). Efficient practical implementation requires several extensions of the basic exclusion principle: i) lowest visible extent applied to foreground blobs to place cells i and j on the same solid surface, ii) accumulation of contradictions over time to improve efficiency and to overcome errors in the occupancy signal output by foreground detection, iii) exploitation of the bidirectional nature of overlap to strengthen the evidentiary value of exclusions and iv) temporal padding of the occupancy signal to overcome clock skew and codec latency effects.

2.2. Mutual Information Estimator

An approach derived from information theory can be based on the mutual information of cell pairs [8]. The mutual information (I) represents the amount of dependence between two given variables. The mutual information, $I(X;Y)$ for two occupancy cells represented as binary random variables X and Y is given by:

$$I(X;Y) = \sum_{y \in Y, x \in X} p_{XY}(x,y) log \left(\frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} \right) \quad (1)$$

In our correlation estimator, high values of $I(X;Y)$ indicate a high degree of dependence, and thus a high probability of overlap.

2.3. Conditional Entropy Estimator

Another approach from information theory is conditional entropy [1]. Conditional entropy quantifies the amount of entropy remaining about a random variable when a second random variable is already known. For the overlap case, one would expect the conditional entropy to be minimal where cells are overlapping. The conditional entropy, $H(X|Y)$ for two occupancy cells represented as binary random variables X and Y is given by:

$$H(Y|X) = - \sum_{y \in Y, x \in X} p_{XY}(x,y) log\, p_{XY}(y|x) \quad (2)$$

Low values of $H(Y|X)$ are expected to indicate high probability of overlap in our correlation estimator.

2.4. Lift Correlation Estimator

Comparing areas with differing traffic levels may lead to occupancy data being less statistically balanced. One could measure the independence of $X = 1$ and $Y = 1$ rather than the independence of the variables X and Y in order to try and observe a greater signal. One operator that attempts this is *lift(X,Y)* [2] which can be described as:

$$lift(X, Y) = \frac{p_{XY}(1, 1)}{p_X(1)p_Y(1)} \quad (3)$$

Values of lift significantly greater than 1.0 provide correlation evidence of overlap.

3. Evaluation: Search Space Precision-Recall

One purpose for a camera overlap estimate is to support processes that analyse activity as it moves from camera to camera, termed *inter-camera processes*. The most obvious such process is tracking targets within a camera view (intra-camera), and using overlap estimates to assist tracking across camera views (inter-camera). A tracking process achieves this by identifying in the overlapping camera the best entry point for the track given its exit point. This is selected by exploring the search space (for example, using appearance descriptors such as [6]) to identify activity that corresponds best to the target.

In practice targets tend to be moving, so the search space for a given point should reflect this by including a neighbourhood around each overlap point. This neighbourhood can be defined as an area within a specified distance of the overlap point in a camera's field of view. In this work we split the camera view into a regular grid of cells, using 12 cells horizontally and 9 cells vertically per camera. We define the neighbourhood of a given cell as including the eight adjacent cells. We propose that the appropriate metrics for evaluating the accuracy of an overlap estimate are *precision* and *recall* of the inter-camera search space, rather than the overlap estimate. This search space better reflects likely candidates for continuing the inter-camera processes.

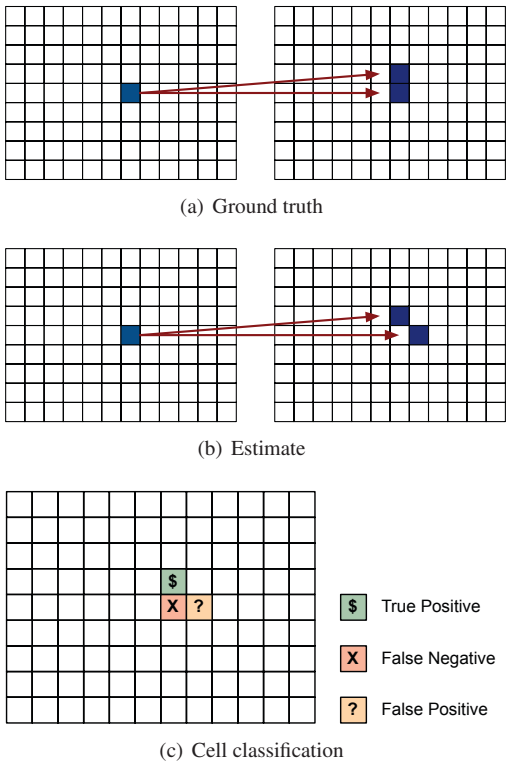
Precision and recall are standard metrics for the accuracy of a classifier, particularly in the information retrieval context [7]. Given a classifier with true positives, TP, false positives, FP, and false negatives, FN, the precision P is given by:

$$P = TP/(TP + FP) \quad (4)$$

and the recall, R is given by:

$$R = TP/(TP + FN) \quad (5)$$

A threshold is applied to select which links are considered overlapping, and is varied to generate the precision-recall (P-R) curve. The thresholds required to obtain the entire P-R range of interest depend on the technique that is being used, and a sensible range of values will often need to be selected by hand.



(a) Ground truth

(b) Estimate

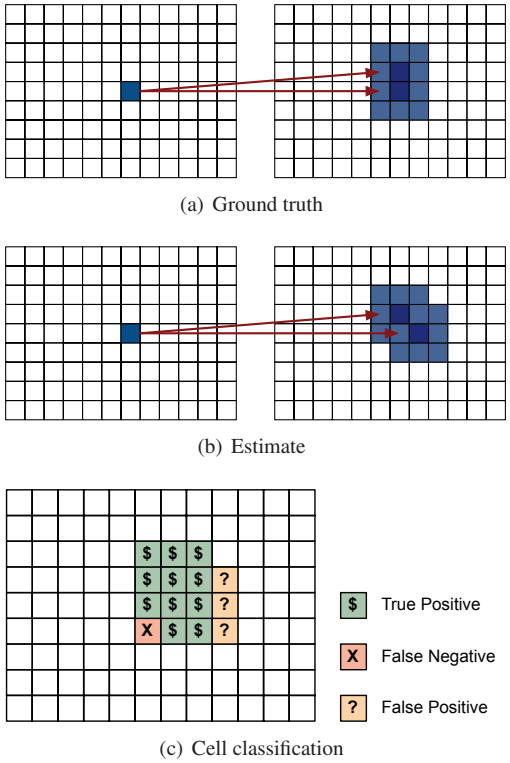
(c) Cell classification

Figure 1. Overlap for a cell in ground truth and estimate.

Figure 1 illustrates a simple case where a cell in one camera overlaps with two cells in another camera. Figure 1(a) shows the ground truth overlap, whereas Figure 1(b) shows the estimated overlap. Figure 1(c) shows the classification of cells in the second camera as true positives, false positives, and false negatives, as determined by comparing the estimated overlap to the ground truth. There is one true positive, one false positive and one false negative, so the overlap precision is 0.5 and the overlap recall is 0.5.

Figure 2 illustrates the same case as Figure 1, but in terms of search space instead of overlap. Similarly, Figure 2(c) shows the classification of cells in the second camera as true positives, false positives, and false negatives, as determined by comparing the estimated search space to the ground truth search space. In both cases the neighbourhood used in searching is a distance of one cell. There are 11 true positives, 3 false positives and 1 false negative. Here, search space precision is $11/14$ and search space recall is $11/12$.

For most processes using camera overlap to define an inter-camera search space, search space recall is more significant than search space precision. This is because anything less than 1.0 for search space recall means that parts



(a) Ground truth

(b) Estimate

(c) Cell classification

Figure 2. Search space for a cell in ground truth and estimate.

of the correct search space for a cell are not being searched, and targets may therefore be lost as they move between cameras. In contrast, search space precision less than 1.0 means merely that the search space is larger than it should be. Thus the processes using that search space are less efficient, but their results are no less accurate than for searching unaided by an activity topology. Notice that for the case illustrated in Figure 2(c), the effect of the error in overlap estimation is less severe on search space recall than on search space precision. This contrasts with the situation in Figure 1(c), where the effects on overlap precision and recall are the same. Thus the search space precision and recall metrics reflect the priority of recall over precision in typical applications using the topology.

4. Results

This section presents the experimental results of each of the methods on three data sets: two five camera data sets obtained in an artificial environment, and a 24 camera data set from a realistic indoor environment. The two five camera data sets were obtained from the same camera setup, and therefore share a ground truth. They differ in their length, and the objects that were recorded moving in their view. The 24 camera data set was obtained from cameras spread throughout the corridors of an office.

4.1. Five Camera Car Data Set

The five camera car data set was obtained from a set of cameras whose overlapping views centre on a common clear ground plane. A detailed ground truth topology of camera overlap was determined by hand for this small data set. This was achieved using markers laid out on the floor which could be matched across the cameras in the observation area. Figure 3 shows the layout of the cameras with a set of coarse grids to demonstrate the overlapping views. The overlap ground truth for a cell was derived by considering any cell in another camera as overlapping where any portion of it was overlapping the given cell. The experimental data used to determine the activity topology was obtained by driving one of two remote controlled cars through the area and capturing over 30 minutes of footage.

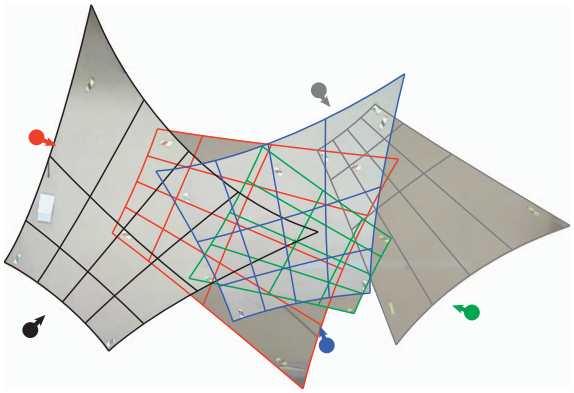
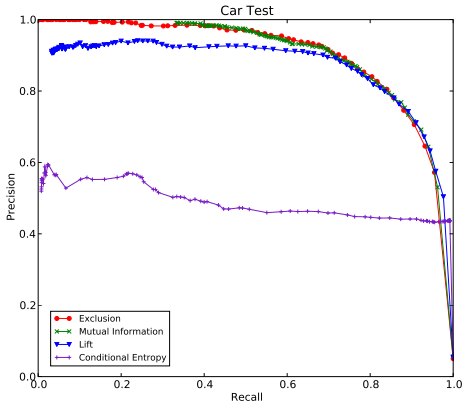


Figure 3. Ground truth overlap for the five cameras

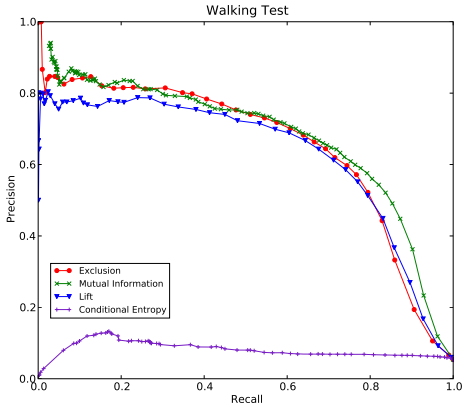
This is the simplest of the data sets, which is reflected in the good precision-recall curves in Figure 4(a). These results show that although the mutual information estimator might slightly outperform the exclusion estimator, they actually produce very similar precision levels for all recall levels. They both outperform the lift estimator for low recall levels, though for a recall greater than 0.7, it becomes indistinguishable from the other estimators. The conditional entropy estimator produces fairly poor precision results, indicating that a significant number of false overlap links are being included.

4.2. Five Camera Walking Data Set

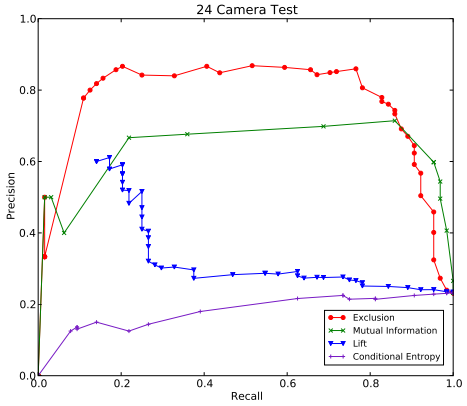
The five camera walking data set was obtained from observations of people walking through the same camera setup shown in Figure 3. This data set therefore utilises the same ground truth, but consists of a separate set of two hours of footage. The people moving through the area had varying directions, speeds, and were sometimes walking in groups. This data is more challenging as the observed objects are larger, can often occlude each other, and can incorrectly



(a) 5 camera car data set



(b) 5 camera walking data set



(c) 24 camera data set

Figure 4. Comparisons of exclusion, mutual information, $lift$, and conditional entropy for the available data sets

be merged together by the foreground detectors into larger blobs. This reduces the quality of the occupancy data extracted, which affects the results of overlap estimation.

The results shown in Figure 4(b) demonstrate the reduced precision that was expected. The mutual information estimator could be considered to have a marginal improvement in performance compared to the exclusion estimator, especially for higher levels of recall. Both of these estimators outperform the lift operator at low recall levels, though not by as much as seen in the car data set. Conditional entropy again performs very poorly, showing a significantly lower precision level than was achieved on the car data. The lower overall precision results are in part because of the difficulty of the data. Some areas have low activity, where it is difficult to observe occupancy and establish overlap links, whilst the occupancy becomes noisier due to incorrectly merged objects. This is an aspect of the activity topology that is difficult to overcome without using a specific calibration design such that activity occurs across the entire camera views.

The lower precision may also occur from the inclusion of links based on *anti-correlation*. This occurs where the occupancy in one camera correlates with the correct link in the other camera, as well as a region around it. Because objects from foreground detection are summarised to its lowest point, adjacent cells will therefore definitely be unoccupied. Links to these adjacent cells will be eliminated by methods that require occupancy in both cells; however their conditional entropy may still be low.

4.3. 24 Camera Office Data Set

The 24 camera data set was obtained from a network of surveillance cameras installed in offices and corridors at the University of Adelaide. They cover a number of corridors inside the building, with a floor plan shown in Figure 5. The cameras recorded many people moving around the area and interacting over a four and a half hour period. This data has a higher degree of difficulty and activity than the previous two data sets. Due to its size, the ground truth was determined using camera to camera links rather than individual cells. The estimated topologies were evaluated with cell links being considered correct when they connected to the appropriately overlapping camera view.

The results for this difficult data set show exclusion outperforming the mutual information estimator. Both of these estimators have significantly higher precision than the lift estimator, whilst conditional entropy again produced very poor results. The difference in results could be due to a number of factors. Firstly, in this larger data set, there is more unstructured activity. Some areas have high traffic for significant periods of time, sometimes with large groups of people, whilst other areas may have little or no activity. This may make it difficult to find statistical links in the data. Exclusion on the seems to be able to better exploit situations

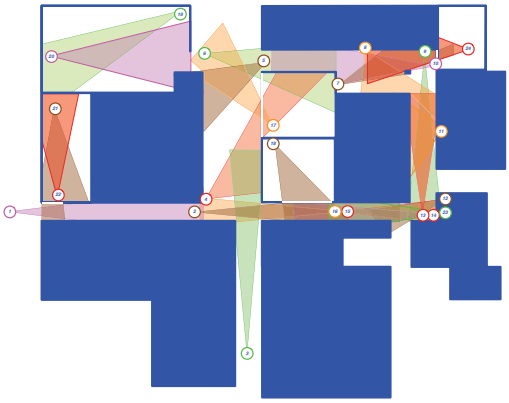


Figure 5. The floor plan showing the position of the cameras within the 24 camera setup

where the data may hold portions demonstrating the non-overlap of some areas. Some techniques may also be more sensitive to the thresholding used and to noise in occupancies, especially when trying to analyse lower traffic areas. The anti-correlation effect may also reduce the precision of the conditional entropy estimator. Another effect that could be of significance is when cameras go off-line during the experiment. This seems to negatively affect the ability of the lift estimator in determining appropriate links for the less reliable areas.

5. Conclusions

This paper has examined a number of techniques for estimating the topology of an overlapping surveillance network. The techniques investigated are based upon the principles of exclusion, mutual information, conditional entropy and the lift operator. These techniques were evaluated using search space-based precision-recall curves. The first two simpler data sets consisted of 5 cameras with views overlapping on an open ground plane, with either a moving remote controlled car, or people walking through the area. In the results from this simpler data set, the mutual information and exclusion methods both produced good results. The lift-based estimator performed slightly worse, whilst the conditional entropy estimator showed a lower precision. This indicates excessive overlap connections, possibly made worse by anti-correlation effects, where occupied cells are linked strongly with unoccupied cells.

The office data set consisted of 24 cameras and had many people walking around in the surveillance area, sometimes in groups, for four and a half hours. Due to the higher activity, this data set has significantly noisier occupancy data. Whilst the lift estimator performed worse than mutual information and exclusion, the conditional entropy results were again particularly poor, possibly due to anti-correlation effects. The mutual information estimator does not perform as well as exclusion, possibly because exclusion can eliminate links in the overlap topology for which it has any contradictory evidence.

These results show that a variety of correlation and contradiction approaches can be used to obtain an overlap estimate within a surveillance network, and that search space-based precision-recall can be an effective way to evaluate them. They have demonstrated that both exclusion and mutual information are viable approaches that can outperform other estimators on a realistic large and noisy data set.

References

[1] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] C. Buehler. Computerized method and apparatus for determining field-of-view relationships among multiple image sensors, 2007. United States Patent 7286157.

[3] 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.

[4] G. Emerling. D.c. police set to monitor 5,000 cameras, 2008. Washington Times.

[5] J. Griffin. Singapore deploys march networks vms solution, 2009. IP Security Watch.

[6] C. Madden, E. D. Cheng, and M. Piccardi. Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Machine Vision and Applications*, 18:233–247, 2007.

[7] 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.*, 7(3):205–229, 1989.

[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] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[10] M. Valera Espina and S. A. Velastin. Intelligent distributed surveillance systems: A review. *IEE Proc. - Vision, Image and Signal Processing*, 152(2):192–204, April 2005.

[11] 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*, pages 44–49, 2006.