Incremental Modelling of the Posterior Distribution of Objects for Inter and Intra Camera Tracking

Andrew Gilbert, Richard Bowden
CVSSP, University of Surrey, Guildford, GU2 7XH UK
a.gilbert@surrey.ac.uk, r.bowden@surrey.ac.uk

Abstract

This paper presents a scalable solution to the problem of tracking objects across spatially separated, uncalibrated, non-overlapping cameras. Unlike other approaches this technique uses an incremental learning method to create the spatio-temporal links between cameras, and thus model the posterior probability distribution of these links. This can then be used with an appearance model of the object to track across cameras. It requires no calibration or batch preprocessing and becomes more accurate over time as evidence is accumulated.

1 Introduction

CCTV cameras are increasingly used as a deterrent and tool to counteract crime and terrorism. There is, however, heavy constraints upon the operator in managing large scale surveillance systems, therefore, tools to assist the operator's decision process are essential. The goal of this paper is to automatically track objects between cameras (inter camera), often termed, object "handover", where one camera transfers a tracked object to another. To do this we need to learn about the relationships between the cameras, without colour, or spatial pre-calibration. In summary, an ideal system could be described as one that, is able to work immediately upon initialisation, will improve performance as more data becomes available, and is able to adapt to changes in the camera's environment.

The surveillance system needs to be able to link the separate cameras together in such a way that as an object appears on one camera, the system can determine if this is the same object previously tracked on another camera, or a new object. This paper proposes a method to learn the linkage between cameras that does not require pre-calibration or explicit training periods. The system is able to learn incrementally about its environment and the objects that move within it, thus allowing the accuracy of tracking to increase over time. The paper is split into a number of sections, firstly a brief background of inter camera tracking techniques is given. Section 3 shows how the information gained from intra camera tracking is collated and its use in creating the inter camera links is described in section 4. Section 5 explains the sub division optimisation that is carried out to improve the linkage between cameras. Section 6 then describes how the weighted links can be used with an appearance model to track objects inter camera. An experiment with results is then presented in Section 7.

BMVC 2005 doi:10.5244/C.19.39

2 Background

Early tracking algorithms [3][11] required both camera calibration and overlapping fields of view (FOV). These are needed to compute the handover of tracked objects between cameras. Additionally Chang [4] required a 3D model of the environment using epipolar geometry, to allow for the registration of objects across the different overlapping cameras. The requirement that cameras have an overlapping FOV is impractical due to the large number of cameras required and the physical constraint upon their placement.

For a tracking algorithm to be useable in a large environment, it needs to be able to deal with ”blind” spots in the system where a person is not visible to any camera. Most single camera (intra) tracking algorithms rely on smooth motion using the previously observed velocity to predict the future location using methods such as the Kalman filter [15]. However, motion between cameras is rarely smooth. Thus a number of techniques have been developed to handle the ”blind” spots and improve object handover.

Kettnaker and Zabih [12] presented a Bayesian solution to track people across cameras with non-overlapping FOVs. However the system required calibration, with the user providing a set of transition probabilities and their expected duration *a priori*. This means that the environment and the way people move within it must be known. In most surveillance situations this is unrealistic.

Probabilistic or statistical methods have seen some of the greatest focus to solve inter camera tracking. They all use the underlying principle that through accumulating evidence of movement patterns over time it is likely that common activities will be discovered. Huang and Russel [7] use a probabilistic approach for tracking cars on a highway. The appearance of the car is modelled by the mean of the colour and the transition times modelled as gaussian distributions. This approach is very application specific, using only two calibrated cameras with vehicles moving in one direction in a single lane. Javed, et al [8] present a more general system by learning the camera topology and path probabilities of objects using Parzen windows. This is a supervised learning technique where transition probabilities are learnt during training using a small number of manually labeled trajectories. Dick and Brooks [5] use a stochastic transition matrix to describe patterns of motion both intra and inter camera. For both systems the correspondence between cameras has to be supplied as training data *a priori*. The system required an offline training period where a marker is carried around the environment. This would be infeasible for large systems and can not adapt to cameras being removed or added ad hoc without recalibration.

KaewTraKulPong and Bowden [9] or Ellis et al [6] do not require *a priori* correspondences to be explicitly stated, instead they use the observed motion over time to establish appearance and disappearance periods. Ellis learns the links between cameras, using a large number of observed objects to form reappearance period histograms between the cameras. Bowden instead uses appearance matching to build up fuzzy histograms of the reappearance period between cameras. This allows a spatio-temporal reappearance period to be modelled. In both cases batch processing was performed on the data which limits their application.

Colour is often used in the matching process. Black et al [1] use a non-uniform quantisation of the HSI colour space to improve illumination invariance, while retaining colour detail. KaewTraKulPong and Bowden [10] uses a Consensus-Colour Conversion of Munsell colour space (CCCM) as proposed by Sturges et al [14]. This is a coarse quantisation to provide consistent colour representation inter-camera without colour camera calibra-

tion.

3 Object Tracking and Description

The test environment consists of 4 non-overlapping colour cameras in an office environment with the layout shown in Figure 1, the video feeds are multiplexed together to form a time synchronized single video, fed into a P4 windows PC in real time.

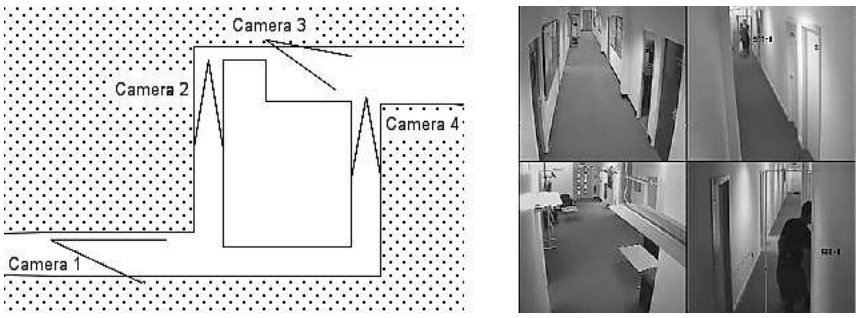

Figure 1: (Left)The top down layout of the camera system, (Right) The tracking environment used.

The object detection tracker [9] used in this work is based on modelling the static background colour distribution and is similar to that originally presented by Stauffer and Grimson [13]. The foreground vs background pixel segmentation is formed using a gaussian mixture model on a per pixel basis, learnt using an online approximation to expectation maximisation. Shadows are identified and removed by relaxing a models constraint on intensity, and the foreground object is formed using connected component analysis on the resulting binary segmentation. Objects are linked temporally to provide movement trajectories within each camera, illustrated in figure 1.

3.1 Colour Similarity

Once the foreground objects have been identified, an object description is formed for inter camera correlation. The colour histogram is used to describe the objects appearance as it is invariant to viewpoint and through quantisation, invariance to illumination can be achieved. Several colour spaces and quantisation levels were investigated including the HSI (8x8x4) approach proposed by Black et al [1] . The Consensus-Colour Conversion of Munsell colour space (CCCM) [14] was selected as it produces marginally superior results, for further details see [2]. CCCM works by breaking RGB colour into 11 basic colours. Each basic colour represents perceptual colour category established through a physiological study of how human's categorise colour variations. This coarse quantisation provides consistent colour representation inter-camera without colour camera calibration provided cameras have similar colour consistency.

4 Building the Temporal links between Cameras

Our approach makes use of the key assumption that, given time, objects will follow similar routes and that repetition will form marked and consistent trends in the overall data. These temporal links can be used to link regions together producing a probabilistic distribution. Thus between each region of a camera a link can be formed estimating the posterior distribution of an objects movement.

Linking all regions to all others is feasible in small scale experimental systems however as the number of cameras increase, the number of possible links required to model the posterior increases exponentially. With only 4 regions in the system there will be 12 links(excluding linking the camera to itself), increasing this to 20 regions, the number of links would be 380. This complexity further increases as multiple entry and exit areas are considered for each camera. With each camera in a system of 20 cameras having 3 entry or exit areas, a total of 3540 links would be required to ensure that all possibilities are covered and as links increase, the amount of data required to learn these relationships also increases. The approach becomes infeasible. However, most of the links between regions are invalid as they correspond to impossible routes. Thus to use the available resources effectively a method is required to distinguish between valid and invalid links. Most solutions to this problem require either, batch processing to identify entry/exit points or hand labeling of the links between regions (impractical in large systems). Both of these approaches are unable to adjust to changes in the environment or camera position. This section proposes a method that is initially coarsely defined but then refines itself over time to improve accuracy as more data becomes available. It has the ability to adjust to any changes that might occur in the environment without a complete system restart.

4.1 Region links

The system starts by identifying links at the basic camera-to-camera level, discarding unused or invalid links. Valid links can then be subdivided to provide a higher level of detail.

The tracking algorithm automatically tracks objects within the camera's FOV and forms a colour appearance model of the person based on the CCCM colour space. The colour histogram $B = (b_1, b_2....b_{11})$ is the median histogram recorded for an object over its entire trajectory within a single camera. All new objects that are detected are compared to previous objects within some allowable reappearance period. The colour similarity is calculated and combined together, to form a discrete probability distribution over time based on this reappearance period T. Thus the frequency f of a bin ϕ is calculated as:

$$f_\phi = \sum_{\forall i,j} H_{ij}; (t_i^{end} - t_j^{start}) < T \tag{1}$$

where t_i^{start} and t_i^{end} are the entry and exit times of object i respectively, T is the maximum allowable reappearance period. H_{ij} is the histogram intersection of objects i and j given by

$$H_{ij} = \sum_{k=1}^{11} min(B_{ik}, B_{jk}) \tag{2}$$

Frequencies are only calculated for an object i that disappears from region y followed by a reappearance in region x ($f_\phi^{x|y}$). Through normalising the total area by $\sum_i^T f_\phi^{x|y}$ we

obtain an estimate to the conditional transitional probability $P(O_{x,t}|O_y)$. An example of $P(O_{x,t}|O_y)$ is shown in Figure 2 where $O_{x,t}$ is object x at time t, using 3 second time intervals. The distinct peak at 3 seconds indicates a link between the two regions. The next section describes how this link can be detected and is used.

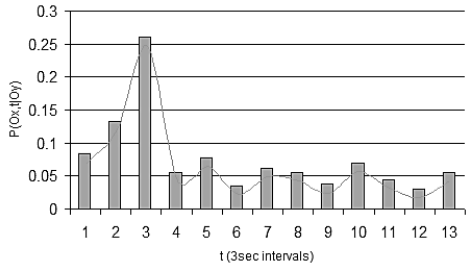

Figure 2: An example of a probability distribution showing a distinct link between two regions

5 Incremental Block sub division and recombination

This section explains how the system identifies which links are valid and therefore when to subdivide the connected blocks. Eventually, adjacent neighbouring blocks can be re-combined to form larger blocks if found to have similar distributions.

The system is based on a rectangular subdivision. Initially, at the top level, the system starts with one block for each of the 4 cameras. This allows tracking to start immediately with a coarse estimate initially uniformly distributed. The 12 links between the blocks are learnt over time using the method described in the previous section. After sufficient evidence has been accumulated, determined by the degree of histogram population, the noise floor level is measured for each link. This is found by taking the median of all the non zero values of the probability distribution. If the maximum peak of the distribution is found to exceed double the noise floor level, there may be a correlation between the two blocks, otherwise it is likely that there is no link and can be safely discarded. The noise level was determined through experimental investigation. In Table 1 different noise floors and time intervals in the probability distribution are shown along with the success rate of correctly identifying links and non-links. The Median x2 was found to provide the best results over the varying time intervals.

If a link is found between 2 blocks, they are both sub divided to each create 4 new equal sized blocks, this process is illustrated in Figure 3. The data previously used to create the links between the original blocks, is then reused to form links in these new split blocks, before the algorithm incorporates new evidence into the distributions.

It is likely that many of the blocks will not form coherent links, and can therefore be removed to minimise the number of links maintained. It is important that links are not removed between blocks that could work with more data. Therefore a link between two blocks is only removed if it has no data in it at all. This cautious method ensures no blocks or links are removed that might be useful in a later iteration. Figure 4 shows how the blocks are removed and sub divided over time. In this example, the maximum

Noise Floor Thres.	Varying Distrib. time intervals with correct linkage shown (%)				
	1sec interval	2sec int	3sec int	4sec int	8sec int
2x Median	69%	82%	75%	94%	82%
3x Median	63%	69%	56%	75%	50%
4x Median	50%	50%	56%	56%	50%

Table 1: Table of % of correctly identified links and non-links between regions, with varying noise floor and time intervals, (max reappearance time 40 secs using 3000 People)

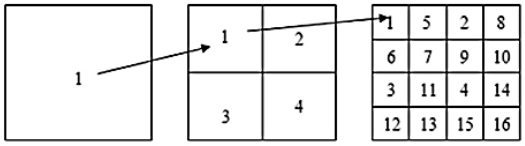

Figure 3: The iterative process of splitting the blocks

number of iterations is 4, after which the amount of data required to sub divide the blocks further exceeds 10000 tracked people (around 5 days of data). More than 5 days data would effect the ability of the approach to adapt to long term changes in the environment.

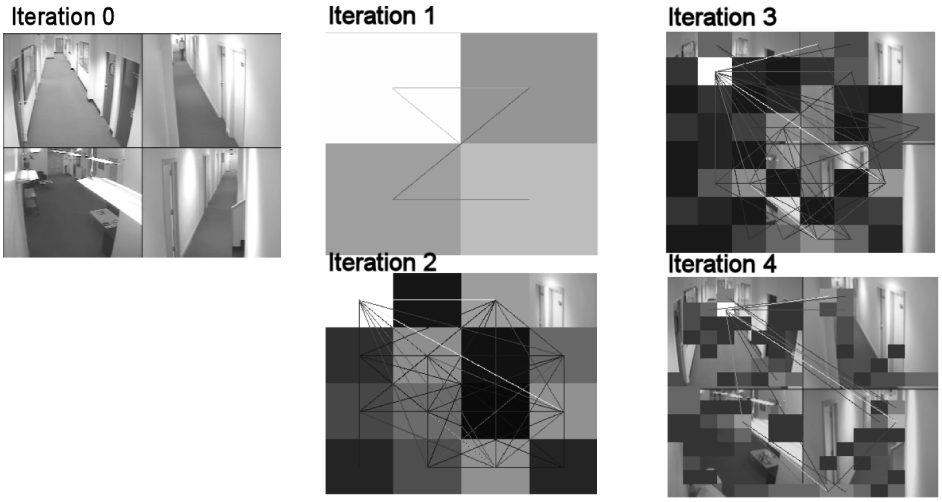

Figure 4: The iterative process of splitting the blocks on the video sequence

Table 2 shows the number of links maintained and dropped at each iteration, along with the amount of data used. It can be seen that with each iteration, the number of possible links increases dramatically, whereas the number of valid links maintained by the system is considerably less. The policy of removing unused and invalid regions improves system scalability. This iterative process can be repeated to further increase the resolution of the blocks.

Iteration	Amount of Data	Number of Blocks	Tot poss Links	Initial links	Dropped links	Kept links
1	1000	4	12	12	0	12
2	5000	16	240	240	45	195
3	10000	60	2540	1631	688	943
4	10000	191	36290	36134	34440	1694

Table 2: Table of number of links maintained and dropped in each iteration

The blocks start to form the entry and exit points of the cameras, Figure 5 (left image) shows the result after 4 iterations. The lighter blocks have a higher importance determined by the number of samples each link contains. As the number of iterations increase,

Figure 5: Left image shows the main identified entry/ exit blocks. Right Image shows the individual blocks that, if similar, are then recombined to form larger more populated blocks, shown by the constant colour areas.

the size of the linked blocks decreases and thus reduce the number of samples detected in each block. This affects the overall reliability of the data used. To counter this, blocks which are found to have similar distributions to neighbouring blocks are combined together to increase the overall number of samples within that region (as illustrated in the right image in Figure 5. In comparing the two images in Figure 5, the effect of combining the neighbouring blocks reduces the number of actual links between blocks, therefore increasing the accuracy of the remaining links.

6 Modelling Posterior Appearance Distributions

This section describes how the weighted links between blocks can be combined to form a single posterior probability distribution. Over time this can be refined to become more detailed as the iterative block splitting process (described previously) takes place.

Given an object which disappears in region y we can model its reappearance probability over time as;

$$P(O_t|O_y) = \sum_{\forall x} w_x P(O_{x,t}|O_y) \tag{3}$$

where the weight w_x at time t is given as

$$w_x = \frac{\sum_i^T f_\phi^{x|y}}{\sum_{\forall y} \sum_i^T f_\phi^{x|y}} \tag{4}$$

This calculated posterior probability is then applied with the appearance model described earlier to enhance the tracking of objects across spatially separated cameras. When a new object is detected, it is compared to all previous objects separately within the time threshold T. The resultant colour similarity is found through histogram intersection using the CCCM colour space and the result weighted by the new object's entry block posterior likelihood to form a weighted hypothesis for the objects previous location. The highest weighted hypothesis determines where an object previously appeared. Values below a nominal threshold are identified as new objects to the system.

7 Results

The 4 camera system described and presented in Figure 1 was used to track and collate a total of 10000 people over a 5 day period. As the data became available the blocks were gradually split and refined. The main entry/exit blocks and links after 4 iterations are shown in Figure 6, along with a spatial map of the blocks.

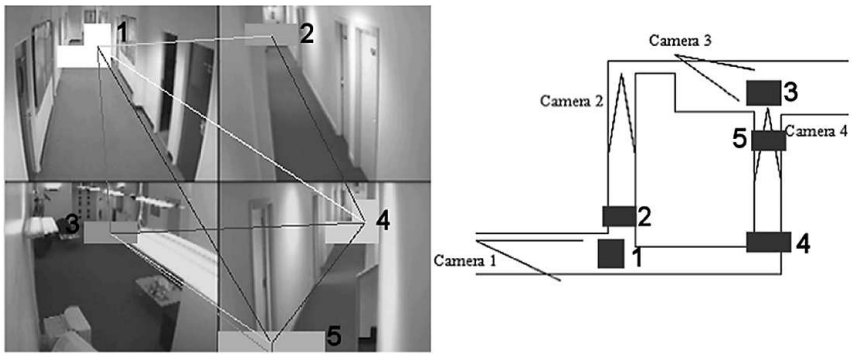

Figure 6: Both the main entry and exit points and a top down layout of the camera system environment with these blocks marked.

An example of the combined posterior probability of Block 1, $P(O_1,t)$ from Figure 6 is shown in Figure 7. The peak which can be seen occurring at 3 seconds indicates the likely time taken for an object to arrive from block 2 which is the most probable block for an object to have arrived from. There is also another peak at 18 seconds, this is the time taken for an object to arrive from block 4 on average.

Table 3 compares the success at tracking people across cameras using only the colour similarity of the objects, against the method proposed in this paper, where the posterior distribution for the regions are used to weight the colour similarity between objects.

Initially, the experiment used only the colour similarity with no weights to track objects. As the available data increased, so did the number of program iterations, each time incorporating new evidence. With each iteration, the linked blocks were split and refined

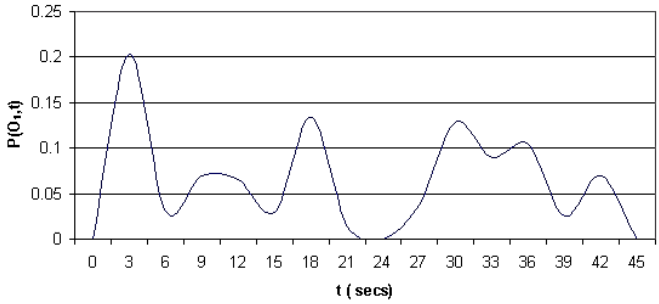

Figure 7: The Posterior Probability distribution of Block 3

Up to Iteration	Success (%)	Additional Data used	Total Data Used
Only Colour Similarity	55%	0	0
1	57%	500	500
2	63%	500	1000
3	68%	4000	5000
4	71%	5000	10000

Table 3: Table of results of using colour similarity and using a weighted colour similarity with varying of iterations to track object cross camera

increasing tracking accuracy. The test video sequence consisted of 45 objects that were tracked for more than 4 seconds. Each new object was compared to other previous objects within a reappearance period of 40 seconds and then either labeled as a new object or as a previously tracked object. The results, compared to ground truth data are presented in Table 3. The final column indicates the amount of data used to create the posterior distribution between regions, starting with no data increasing in stages to 10000 people.

It can be seen that as the amount of data and number of program iterations increases so does the success rate of tracking. This system fulfills the three ideals stated in the introduction. It is not unreasonable to assume that given more data, further improvements could be achieved.

8 Conclusions

We have described a method to automatically derive the main entry and exit areas in a camera probabilistically using incremental learning. This allows for objects to be tracked between spatially separated uncalibrated cameras. The spatio-temporal structure of a surveillance system can be used to weight hypotheses extracted through a coarse colour similarity. The method is completely automatic and unsupervised and able to adapt to changes in the spatio-temporal structure of the system. An experiment shows that using this technique increases tracking accuracy up to 71%. This is a considerable improvement over colour similarity alone, and impressive given the base accuracy (55%). The incremental technique ensures that the system works immediately but will become more

accurate overtime as additional data is acquired.

9 Acknowledgements

This work is supported by an EPSRC/Imaging Faraday partnership industrial CASE award sponsored by Overview Ltd.

References

[1] J. Black, T.J Ellis, and D. Makris. "Wide Area Surveillance with a Multi-Camera Network". *Proc. IDSS-04 Intelligent Distributed Surveillance Systems*, pages 21–25, 2003.

[2] R. Bowden, A. Gilbert, and KaewTraKulPong P. "Tracking Objects Across Uncalibrated Arbitrary Topology Camera Networks, in Intelligent Distributed Video Surveillance Systems". *S.A Velastin and P Remagnino Eds. Chapt 6, IEE, London, to be published*, 2005.

[3] Q. Cai and J. Agrarian. "Tracking Human Motion using Multiple Cameras". *Proc. International Conference on Pattern Recognition*, pages 67–72, 1996.

[4] T. Chang and S. Gong. "Bayesian Modality Fusion for Tracking Multiple People with a Multi-Camera System". *Proc. European Workshop on Advanced Video-based Surveillance Systems*, 2001.

[5] A. Dick and M. Brooks. "A Stochastic Approach to Tracking Objects Across Multiple Cameras". *Australian Conference on Artificial Intelligence*, pages 160–170, 2004.

[6] T.J. Ellis, D. Makris, and J.K. Black. "Learning a Multi-Camera Topology". *Joint IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 165–171, 2003.

[7] T. Huang and S. Russell. "Object Identification in a Bayesian Context". *Proc. International Joint Conference on Artificial Intelligence (IJCAI-97), Nagoya, Japan*, pages 1276–1283, 1997.

[8] O. Javed, Z. Rasheed, K. Shafique, and M. Shah. "Tracking Across Multiple Cameras with Disjoint Views". *Proc. IEEE International Conference on Computer Vision*, pages 952–957, 2003.

[9] P. KaewTrakulPong and R. Bowden. "An Adaptive Visual System for Tracking Low Resolution Colour Targets". *Proc. BMVC'01, Vol 1, Manchester UK*, pages 243–252, 2001.

[10] P. KaewTrakulPong and R. Bowden. "A Real-time Adaptive Visual Surveillance System for Tracking Low Resolution Colour Targets in Dynamically Changing Scenes". *Journal of Image and Vision Computing. Vol 21, Issue 10, Elsevier Science Ltd*, pages 913–929, 2003.

[11] P. Kelly, A. Katkere, D. Kuramura, S. Moezzi, and S. Chatterjee. "An Architecture for Multiple Perspective Interactive Video". *Proc. of the 3rd ACE International Conference on Multimedia*, pages 201–212, 1995.

[12] V. Kettnaker and R. Zabih. "Bayesian Multi-Camera Surveillance". *Proc. IEEE Computer Vision and Pattern Recognition*, pages 253–259, 1999.

[13] C. Stauffer and W.E.L. Grimson. "Learning Patterns of Activity using Real-time Tracking". *PAMI, 22(8)*, pages 747–757, 2000.

[14] J. Sturges and T.W.A. Whitfield. "Locating Basic Colour in the Munsell Space". *Color Research and Application, 20(6):364-376*, 1995.

[15] G. Welch and G. Bishop. "An Introduction to the Kalman Kilter". *Technical Report 95-041,University of North Carolina at Chapel Hill*, 1995.