Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity

Andrew Gilbert and Richard Bowden

CVSSP, University of Surrey, Guildford,
GU2 7XH, England
{a.gilbert, 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 model both the colour variations and posterior probability distributions of spatio-temporal links between cameras. These operate in parallel and are then used with an appearance model of the object to track across spatially separated cameras. The approach requires no pre-calibration or batch preprocessing, is completely unsupervised, and becomes more accurate over time as evidence is accumulated.

1 Introduction

The aim of this paper is to automatically track objects between cameras (inter camera). This is often termed object "handover", where one camera transfers a tracked object or person to another camera. To do this we need to learn about the relationships between the cameras, without colour, or spatial pre-calibration. In summary, an ideal tracking system could be described as one that, upon initialisation is able to work immediately, as more data becomes available will improve performance, and is adaptable to changes in the camera's environment.

To achieve this the system needs to be able to learn both the spatial and colour relationships between non-overlapping cameras. This allows the system to determine if a newly detected object has previously been tracked on another camera, or is a new object. The approach learns these spatial and colour relationships, though unlike previous work it does not require pre-calibration or explicit training periods. Incremental learning of the object's colour variation and movement, allows the accuracy of tracking to increase over time without supervised input.

The paper firstly gives a brief background of inter camera tracking and calibration. With section 3 describing the intra camera tracking and its use in creating the inter camera links is described in section 4. Sections 5 and 6 explain the spatial block subdivision to improve the representation of links and how the links and an object appearance model is used to track inter camera. Incremental camera colour calibration is explained in section 7, with experiments and results that combine both approaches presented in Section 8.

2 Background

Early tracking algorithms [1][2] required both camera calibration and overlapping fields of view (FOV). These are needed to compute the handover of tracked objects between cameras. Additionally Chang [3] 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 constraints upon their placement.

Kettnaker and Zabih [4] 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 [5] presented a probabilistic approach to tracking cars on a highway, modelling the colour appearance and transition times 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* [6] 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 [7] 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 [8] or Ellis *et al* [9] do not require *a priori* correspondences to be explicitly stated, instead they use the observed motion over time to establish reappearance 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* [10] use a non-uniform quantisation of the HSI colour space to improve illumination invariance, while retaining colour detail. KaewTraKulPong and Bowden [11] uses a Consensus-Colour Conversion of Munsell colour space (CCCM) as proposed by Sturges et al [12]. This is a coarse quantisation based on human perception and provides consistent colour representation inter-camera. Most multi camera surveillance systems assume a common camera colour response. However, even cameras of the same type will exhibit differences which can cause significant colour errors. Pre-calibration of the cameras is normally performed with respect to a single known object, such as the 24 main colour GretagMacbeth [13] ColorCheckerTM chart used by Ilie and Welch [14]. Porikli [15] proposes a distance metric and model function to evaluate the inter camera colour response. It is based on a correlation matrix computed from three 1-D quantised RGB colour histograms and a model function obtained from the minimum cost path traced within the correlation matrix. Joshi [16] similarly proposes a RGB to RGB transform between images. By using a 3x3 matrix, inter channel effects can be modelled between the red, green, and blue components.

3 Object Tracking and Description

The test environment consists of 4 non-overlapping colour cameras in an office building, with the layout shown in Figure 1. The area between cameras contains doors and corners removing smooth motion inter camera. The video feeds are multiplexed together to form a time synchronized single video, fed into a P4 windows PC in real time. To detect objects the static background colour distribution is modelled [11] in a similar fashion to that originally presented by Stauffer and Grimson [17]. A gaussian mixture model on a per-pixel basis is used to form the foreground vs background pixel segmentation, learnt using an online approximation to expectation maximisation. Shadows are identified and removed by relaxing a models constraint on intensity but not chromaticity, and the foreground object is formed using connected component analysis on the resulting binary segmentation. Objects are linked temporally with a Kalman filter to provide movement trajectories within each camera, illustrated in Figure 1.

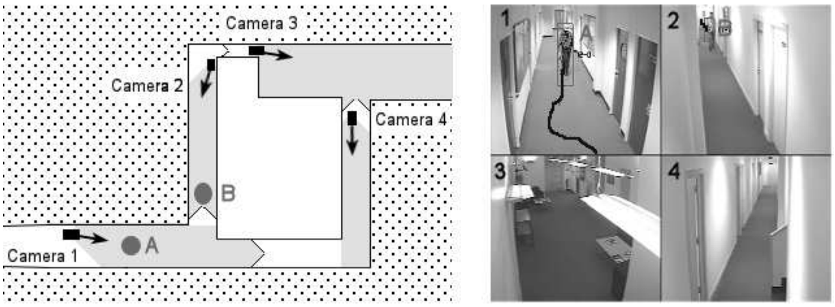

Fig. 1. (Left)The top down layout of the camera system, (Right) The tracking environment used.

3.1 Colour Similarity

Once the foreground objects have been identified, an object descriptor is formed for inter camera correlation. The colour histogram is used to describe the objects appearance as it is spatially invariant and through quantisation, some invariance to illumination can be achieved. Several colour spaces and quantisation levels were investigated including the HSI (8x8x4) approach proposed by Black *et al* [10], the Consensus-Colour Conversion of Munsell colour space (CCCM) [12] and differing levels of conventional RGB quantisation. Without calibrating camera colour responses, CCCM produced marginally superior results and was selected for initial object correlation, for further details see [18]. 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. This coarse quantisation provides a consistent colour representation inter-camera prior to quantisation. With calibration, quantised RGB performs best as will be seen in Section 7.

4 Building the Temporal links between Cameras

To learn the spatial links between cameras, we make use of the key assumption that, given time, objects (such as people) will follow similar routes inter camera and that the repetition of the routes will form marked and consistent trends in the overall data. These temporal links inter camera can be used to link camera regions together, producing a probabilistic distribution of an objects movement between cameras.

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 each camera in a system of 20 cameras having 3 entry or exit regions, a total of 3540 links would be required to ensure that all possibilities are covered. As links increase, the amount of data required to learn these relationships also increases and 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 for the object or person. The colour histogram $B = (b_1, b_2....b_n)$ 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 a set time window, T. 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:

$$f_\phi = \sum_{\forall i}\sum_{\forall j} \begin{cases} H_{ij} & (t_i^{end} - t_j^{start}) < \phi \\ 0 & otherwise \end{cases} \qquad \forall\phi < T \tag{1}$$

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 $H_{ij} = \sum_{k=1}^{11} min(B_{ik}, B_{jk})$. Frequencies are only calculated for an object i that disappears from region y followed by a reappearance in region x ($f^{x|y}$). Normalising the total area by $\sum_i^T f_{\phi=0}^{x|y}$, an estimate to the conditional transition probability $P(O_{x,t}|O_y)$ is obtained. An example of $P(O_{x,t}|O_y)$ is shown in Figure 2 where $O_{x,t}$ is object x at time t. The distinct peak at 6 seconds indicates a link between the two regions.

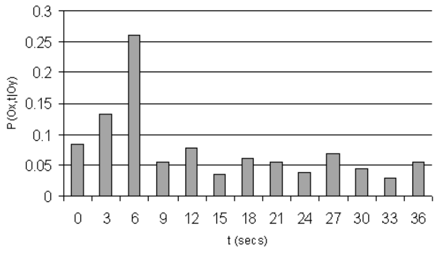

Fig. 2. An example of a probability distribution showing a distinct link between two regions

5 Incremental Block subdivision and recombination

This section explains how the system identifies valid links and therefore when to subdivide the connected blocks. Eventually, adjacent neighbouring blocks can be recombined 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 four cameras. This allows tracking to start immediately with links initially uniformly distributed. The twelve links (ignoring self transitions) 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 could be determined with statistical methods such as the average and standard deviation, however, through experimentation, double the Median of all the values of the probability distribution was found to provide consistent results. If the maximum peak of the distribution is found to exceed the noise floor level, this indicates a possible correlation between the two blocks (eg Figure 2).

If a link is found between two blocks, they are both subdivided to each create four new equal sized blocks. The previous data is then reused and incorporated with future evidence to form links in the newly subdivided blocks. It is likely that many of the blocks will not form coherent links, and if a link has no data in it, it is removed to minimise the number of links maintained. Figure 3 shows how the blocks are removed and subdivided over time. Table 1 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 are considerably less. The policy of removing unused and invalid regions improves system scalability.

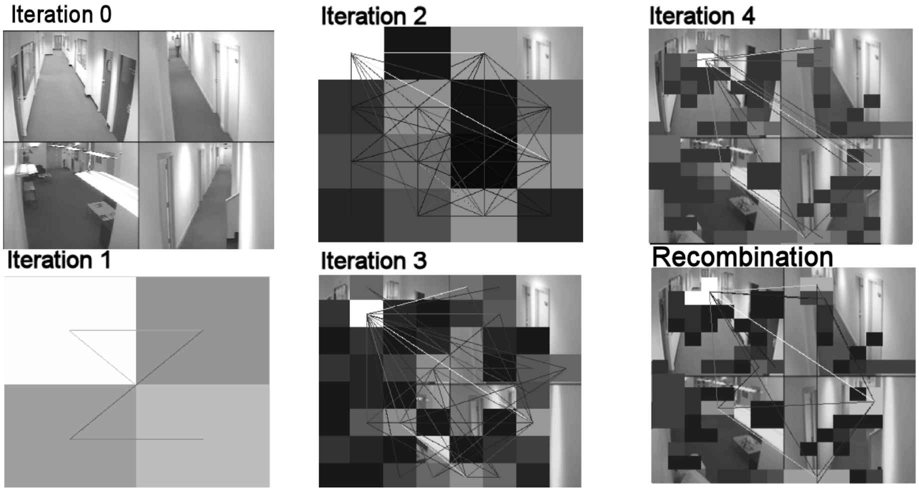

Fig. 3. The iterative process of splitting the blocks on the video sequence

As the process proceeds the blocks start to form the entry and exit points of the cameras, Figure 3 (interation 4) shows the result after 4 subdivisions. The lighter blocks have a higher importance determined by the number of samples each link contains. As the number of iterations increase, the size of the linked blocks decrease and thus reduce the number of samples detected in each block. Low numbers of samples result in unreliable distributions. To counter this,

Table 1. Table of number of links maintained and dropped in each split

Iteration	Amount of Data	Total Possible Blocks	Total Possible Links	Number of Blocks maintained	Total Possible Links	Initial links	Dropped links	Links maintained
1	1000	4	12	4	12	12	0	12
2	5000	16	240	16	240	240	45	195
3	10000	64	4032	60	2540	1631	688	943
4	10000	256	65280	191	36290	36134	34440	1694

blocks which are found to have similar distributions to neighbouring blocks are combined together to increase the overall number of samples within that block (as illustrated in the right image in Figure 3(recombination)). This reduces the number of blocks and therefore links maintained, and increases the accuracy of those links. Should new evidence be identified in previously discarded blocks, eg if a door is suddenly opened, the affected blocks can be recombined to the previous level of subdivision.

6 Calculating Posterior Appearance Distributions

This section describes how the weighted links between blocks can be used to weight the observation likelihood of tracked people. Over time the posterior becomes more accurate 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;

$$P(O_t|O_y) = \sum_{\forall x} w_x P(O_{x,t}|O_y) \tag{2}$$

where the weight w_x at time t is given as

$$w_x = \frac{\sum_{i=0}^{T} f_\phi^{x|y}}{\sum_{\forall y}\sum_{i=0}^{T} f_\phi^{x|y}} \tag{3}$$

This probability is then used to weight the observation likelihood obtained through colour similarity to obtain a posterior probability of a match, across spatially separated cameras. Tracking objects is then achieved by maximising the posterior probability within a set time window.

7 Modelling Colour Variations

The CCCM colour quantisation descriptor used in the previous section, assumes a similar colour response between cameras. However this is seldom the case. Indeed the cameras of Figure 1 show marked difference in colour response even to the human eye. Therefore, a colour calibration of these cameras is proposed that can be learnt incrementally as with the distribution previously discussed.

The system uses the initial CCCM colour descriptor to form posterior distributions, in parallel to forming the colour transformation matrices between cameras. Novelly, the tracked people are automatically used as the calibration objects, and a transformation matrix is formed incrementally to model the colour changes between cameras. However, the people used are not identical sizes, therefore a point to point transformation is unavailable. We therefore use the colour descriptor matched between regions in different cameras to provide the calibration. Equation 4 shows the transformation matrix between image I and the transformed image T using 2 bin RGB quantisation in this simple example.

$$\begin{bmatrix} I_{r_1} & I_{r_2} & I_{g_1} & I_{g_2} & I_{b_1} & I_{b_2} \end{bmatrix} * \begin{bmatrix} t_{r_1r_1} & t_{r_1r_2} & t_{r_1g_1} & t_{r_1g_2} & t_{r_1b_1} & t_{r_1b_2} \\ t_{r_2r_1} & t_{r_2r_2} & t_{r_2g_1} & t_{r_2g_2} & t_{r_2b_1} & t_{r_2b_2} \\ t_{g_1r_1} & t_{g_1r_2} & t_{g_1g_1} & t_{g_1g_2} & t_{g_1b_1} & t_{g_1b_2} \\ t_{g_2r_1} & t_{g_2r_2} & t_{g_2g_1} & t_{g_2g_2} & t_{g_2b_1} & t_{g_2b_2} \\ t_{b_1r_1} & t_{b_1r_2} & t_{b_1g_1} & t_{b_1g_2} & t_{b_1b_1} & t_{b_1b_2} \\ t_{b_2r_1} & t_{b_2r_2} & t_{b_2g_1} & t_{b_2g_2} & t_{b_2b_1} & t_{b_2b_2} \end{bmatrix} \simeq \begin{bmatrix} T_{r_1} & T_{r_2} & T_{g_1} & T_{g_2} & T_{b_1} & T_{b_2} \end{bmatrix} \tag{4}$$

t_{xy} is the term that specifies how much the input from colour channel x contributes to the output of colour channel y. Transformation matrices are formed between the four cameras. Six transformations along with their inverses provide the twelve transformations required to transform objects between the four cameras. As camera calibration is refined the illumination changes that affected the success of the original correlation methods investigated in [18] and section 3, are reduced. This allows other less coarse quantisation (such as RGB) to be used with improved performance as will be shown.

The six transformation matrices for the four cameras are initialised as identity matrices assuming a uniform prior of colour variation between camera. When a person is tracked inter camera and is identified as the same object, the difference between the two colour descriptors, is modelled by the transform matrix t from Equation 4. The matrix t is calculated by computing the transformation that maps the person's descriptor from the previous camera I to the person's current descriptor T. This transformation is computed via SVD. The matrix t is then averaged with the appropriate camera transformation matrix, and repeated with other tracked people to gradually build a colour transformation between cameras. This method will introduce small errors, however it is in keeping with the incremental theme of the paper. Allowing the system to continually update and adapt to the colour changes between cameras as more data becomes available.

To form the transform matrices a number of different quantisations were examined. A 3x3 matrix of the median colour of a person, was found to be too coarse, losing too much colour information. The 11 bin CCCM quantisation used to create the posterior distributions is an arbitrary labeling, not metric and therefore cannot be represented by a linear transformation. However it is more accurate than RGB without calibration. With calibration RGB performs better.

A number of RGB quantisations were investigated with varying accuracy, however a parzen window gives a stable accuracy of 77% over a range of quantisation levels.

8 Results

The final system starts uncalibrated with uniform priors for all distributions and identity matrices for colour transforms. It uses no supervised learning of its environment, instead automatically adding information as it becomes available. This section demonstrates the performance of the incrementally constructed spatio-temporal weights, the inter camera colour calibration and the result of combining both approaches. The data used consisted of 10,000 objects tracked over a period of 72 hours of continuous operation. Evaluation was performed using an unseen ground-truthed 20 minute sequence with 300 instances of people tracked for more than 1 second

Initially, the experiment has no *a priori* information of the environment, using only the CCCM colour similarity between objects to correlate inter camera. The posterior probability of the object match is gained by multiplying the colour similarity by the reappearance probability (3). At each refinement the accuracy increases as indicated in Table 2. After 5 days and 10,000 tracked objects each camera has been split 4 times resulting in a possible 64 regions per camera. At this point accuracy has increased from the base 55% of colour similarity alone to 73%. Equally our incremental learning scheme for colour calibration can be applied. Again as additional objects are added into the colour transformation matrices the accuracy of colour similarity for RGB increases from 42% to 67%.

Table 2. Table of results of using CCCM colour similarity alone, colour calibration alone, posterior distribution weighting of CCCM similarity and a combination of all three. With an increasing number of refinements of the blocks

		Accuracy:		
Block split	Total Data Used	Posterior Distrib Weights	4 bin RGB Colour Calib alone	Combined weight + colour model
CCCM Colour only	0	55%	42%	55%
1	500	60%	55%	68%
2	1000	63%	60%	69%
3	5000	68%	60%	76%
4	10000	73%	67%	78%

Obviously it would be beneficial to combine both of these methods to further increase performance. The first level of block refinement and reappearance period estimation is constructed and the posterior appearance of objects used for colour calibration. This provides a boost in performance as apposed to using colour similarity alone. Once a colour transformation is available, a transformed RGB colour descriptor can be used in learning the second level of block refinement. This process can be repeated where colour calibration can further increase the accuracy of block refinement and vice versa. This is indicated in Table 2 where using this interative scheme raises detection performance from 55% to 78%.

Of course this process can be continued until performance converges to a stable level. Table 3 shows a further 3 iterations without additional data or block refinement providing a final accuracy of 81% which is a significant improvement upon colour similarity alone. This is the stable point for this system without more data being added.

Table 3. Looking at iterations of the colour calibration to further improve accuracy

		Accuracy:		
Iteration	Total Data Used	Posterior Distrib Weights	4 bin RGB Colour Calib alone	Combined weight + colour model
Inital results from block splitting	10,000	73%	67%	78%
1	10,000	73%	69%	80%
2	10,000	73%	70%	81%
3	10,000	73%	70%	81%

The graph in Figure 6, shows how the accuracy increases both over block splits (shown in Table 2), and program iterations (shown in Table 3). The greatest overall increase in accuracy is in the combination of both posterior distribution weights and colour calibration of the cameras. The increase in accuracy allows the system to fulfill the three ideals stated in the introduction, of working immediately, improving performance as more data is accumulated, and can adapt to changes in its environment.

The main entry/exit blocks and links after 4 iterations are shown in Figure 5, along with a spatial map of the blocks.

9 Conclusions

We have described an approach to automatically derive the main entry and exit areas in a camera probabilistically using incremental learning, while simultaneously the colour variation inter camera is learnt to accommodate inter-camera colour variations. Together these techniques allow people to be tracked between spatially separated uncalibrated cameras with up to 81% accuracy, importantly using no *a priori* information in a completely unsupervised fashion. This is a considerable improvement over the baseline colour similarity alone of 55%. The spatio-temporal structure of the surveillance system can be used to weight the observation likelihood extracted through the incrementally calibrated colour similarity. The incremental colour calibration and posterior distribution weighting

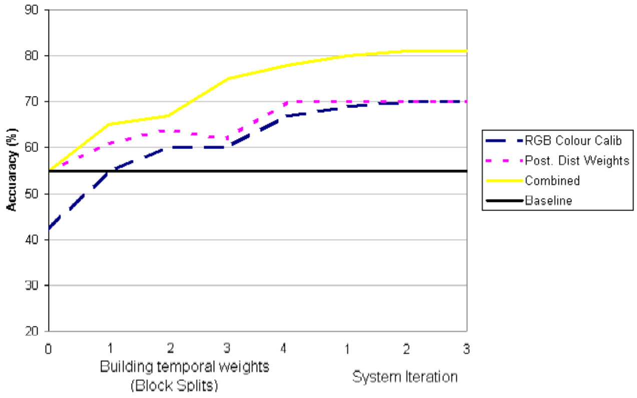

Fig. 4. Comparing the accuracies of; the baseline colour CCCM similarity, colour calibration alone, posterior distributions weights alone (space) and the combination of spatio-temporal weighted colour calibration over a number of program iterations

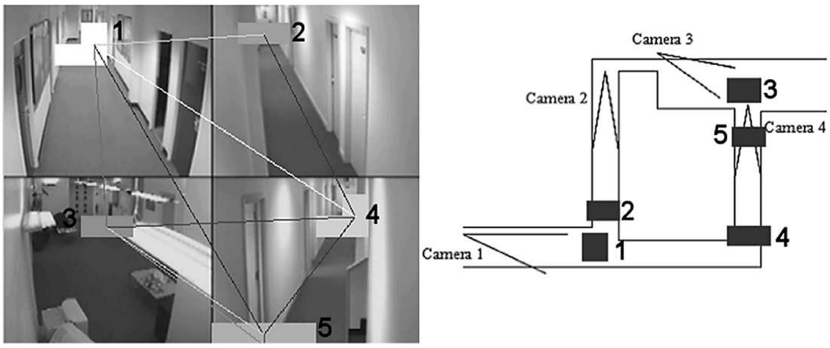

Fig. 5. Both the main entry and exit points and a top down layout of the camera system environment with these blocks marked

are both completely automatic, unsupervised and able to adapt to changes in the environment. The incremental technique ensures that the system works immediately but will become more accurate over time as additional data is acquired.

10 Acknowledgements

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

References

1. Cai, Q., Agrarian, J.: "Tracking Human Motion using Multiple Cameras". Proc. International Conference on Pattern Recognition (1996) 67–72

2. Kelly, P., Katkere, A., Kuramura, D., Moezzi, S., Chatterjee, S.: "An Architecture for Multiple Perspective Interactive Video". Proc. of the 3rd ACE International Conference on Multimedia (1995) 201–212
3. Chang, T., Gong, S.: "Bayesian Modality Fusion for Tracking Multiple People with a Multi-Camera System". Proc. European Workshop on Advanced Video-based Surveillance Systems (2001)
4. Kettnaker, V., Zabih, R.: "Bayesian Multi-Camera Surveillance". Proc. IEEE Computer Vision and Pattern Recognition (1999) 253–259
5. Huang, T., Russell, S.: "Object Identification in a Bayesian Context". Proc. International Joint Conference on Artificial Intelligence (IJCAI-97), Nagoya, Japan (1997) 1276–1283
6. Javed, O., Rasheed, Z., Shafique, K., Shah, M.: "Tracking Across Multiple Cameras with Disjoint Views". Proc. IEEE International Conference on Computer Vision (2003) 952–957
7. Dick, A., Brooks, M.: "A Stochastic Approach to Tracking Objects Across Multiple Cameras". Australian Conference on Artificial Intelligence (2004) 160–170
8. KaewTrakulPong, P., Bowden, R.: "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 (2003) 913–929
9. Ellis, T., Makris, D., Black, J.: "Learning a Multi-Camera Topology". Joint IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS) (2003) 165–171
10. Black, J., Ellis, T., Makris, D.: "Wide Area Surveillance with a Multi-Camera Network". Proc. IDSS-04 Intelligent Distributed Surveillance Systems (2003) 21–25
11. KaewTrakulPong, P., Bowden, R.: "Towards Automated Wide Area Visual Surveillance: Tracking Objects Between Spatially Separated, Uncalibrated Views". In Proc. Vision, Image and Signal Processing, Vol 152, issue 02 (2005) 213–224
12. Sturges, J., Whitfield, T.: "Locating Basic Colour in the Munsell Space". Color Research and Application, 20(6):364-376 (1995)
13. : Gretagmacbeth Color Management Solutions. (www.gretagmacbeth.com)
14. Ilie, A., Welch, G.: "Ensuring Color Consistency across Multiple Cameras". Techincal Report TR05-011 (2005)
15. Porikli, F.: "Inter-Camera Color Calibration by Cross-Correlation Model Function". IEEE International Conference on Image Processing (ICIP),Vol. 2, (2003) 133–136
16. Joshi, N.: "Color Calibrator for Arrays of Inexpensive Image Sensors". MS Thesis, Stanford University Department of Computer Science, (2004)
17. Stauffer, C., Grimson, W.: "Learning Patterns of Activity using Real-time Tracking". PAMI, 22(8) (2000) 747–757
18. Bowden, R., Gilbert, A., 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)