Colour Constancy Techniques for Re-Recognition of Pedestrians from Multiple Surveillance Cameras

Alberto Colombo, James Orwell, Sergio Velastin

Colour Constancy Techniques for Re-Recognition of Pedestrians from Multiple Surveillance Cameras

Alberto Colombo, James Orwell, Sergio Velastin

Kingston University, London

Abstract. This paper presents work towards a system for tracking the movements of a pedestrian as they move between the multiple sensors comprising a surveillance system. The colour appearance of the observations is an important cue: it is useful to achieve good color constancy between the color values associated with each camera. A novel method for estimating the appropriate transform between each camera's colour space is proposed, using covariance of the foreground data collected from each camera. Simulations are used to demonstrate that the method only works if the covariance has a sufficiently high ratio between its eigenvalues. The covariance matrices for foreground data collected from 29 surveillance cameras are estimated and shown to have a sufficiently high ratio. The discriminative power of colour-based appearance descriptors is evaluated using several types of colour constancy methods. The proposed method leads to a significant improvement in the simplest and best performing (mean) colour descriptor. It is shown how these descriptors can be integrated into a probabilistic framework for tracking pedestrians from multiple surveillance cameras.

1 Introduction

Automatic analysis of surveillance video provides the capability for particular categories of events to be detected. The most general category is human motion. More specific categories include overcrowding [1], left luggage [2], people standing too close to the platform edge [3], or moving in an unauthorized direction in an airport [4]. The motivation for this capability is usually to alert the operator to the detected event, thereby making this operator more efficient than would be the case, if all video streams had to be monitored purely manually. The performance of the automatic analysis is usually measured in terms of its precision and recall.

Another common task for operators in the CCTV control room is the surveillance of particular individuals as they make their way through the metro system. These individuals may be known to them or else young or vulnerable people, *e.g.* traveling late at night. After an individual has been chosen as the 'target' for this surveillance, the requirement is to maintain a view of them on a display by selecting the camera that shows the appropriate part of the transport network. This camera selection can be a manual process that is performed by the operator. However, this may not always be feasible if the operator has other tasks to

perform simultaneously. By analogy with the automatic detection of events, a sufficiently well-performing automatic system will improve the efficiency of the operators in the CCTV control room.

This paper presents work towards such an automatic system. In Section 2 an overview of the processing stages is presented. In Section 3 the proposed method for improving colour constancy is described. In Section 4 this method is compared with other methods, using cameras from multiple stations in the Torino Metro system. The paper is concluded in Section 5.

2 Process Overview

In this section, the pipeline of data processing is described. It is assumed that all relevant video data is available to the tracking process with correct time-stamp information. A Gaussian Mixture Model (GMM) segmentation is applied to the video data. Only the foreground data is included in subsequent steps. This is then processed to improve colour constancy between cameras, and then descriptors are extracted for each person. These descriptors operate on appearance (colour) and spatio-temporal properties of the observations. Distance measures are then calculated, and compared with expected distributions to derive conditional probabilities that two observations refer to the same person.

The method assumes that the *Network Layout* (a graph describing the relations between the camera views) is available, and that the cameras are calibrated to the ground plane relevant for their field of view. The graph *edges* describe the possible routes between the camera nodes. The edge attributes include, where applicable, standard minimum times for pedestrians to move from one camera area to another. This information is used to calculate *a priori* probability that two observations refer to the same individual. The calibration and minimum time information is provided manually in these experiments, but there are proposed methods for deriving this information automatically [5, 6].

Two sections of the process each require a training procedure. The improvement of colour constancy requires unlabeled samples of foreground data from each camera. The estimate of conditional probabilities from the distributions of distance measures requires samples of labelled observations of people moving through the network.

2.1 Descriptors and Distance Measures

Three appearance descriptors are used in the experiments. Firstly, the *Mean Colour* is the mean (r, g, b) value of the main connected component associated with the observation. Secondly, a vector of features is extracted from each pixel location of the main connected component. These features are the (r, g, b) colour, (x, y) image position and oriented gradient for each channel. The covariance of these features, over a single connected component, is used as the *Covariance* descriptor [7]. The third appearance-based descriptor is the MPEG-7 *Dominant Colour* descriptor [8].

A *Spatio-Temporal* descriptor is also defined. Potential target sightings are tracked using a Kalman filter, deriving the observation vector from the main connected component of the foreground data. This descriptor uses the Kalman state (ground-plane position and velocity, with associated covariance and time-stamp).

A third type of descriptor, called *Topological* descriptor, is defined for those situations where tracking with a Kalman filter, or using a motion model in general, is unsuitable. This is the case *e.g.* when tracking across underground stations, or across areas of the same station connected by a long occlusion. The Topological descriptor is simply the time-stamp and the camera ID of the observation.

Descriptors of the same type can be compared with each other to establish a measure of dissimilarity, referred to here as a *distance measure*. The Mean Colour descriptor uses a Euclidean distance in RGB space. Covariance descriptors use the square log of the eigenvalues of the product between one descriptor and the inverse of the other (it has been shown [7] that this defines a metric in the space of covariance matrices). The distance between two Dominant Color structures is calculated here as the sum of the distances between the individual components, weighted by the product of the proportions of total area occupied by each component.

For two Spatio-Temporal descriptions, the state covariances can be used to calculate the Mahalanobis distance between the two state means, provided that the two states are temporally aligned, *i.e.* calculated for the same frame as each other. However in general the two states will not be temporarily aligned (the tested hypothesis is that the person who was observed in one camera is now observed in another camera, at a later time). To enforce a temporal alignment in these circumstances, the earlier state is moved forward the required number of frames by using the Kalman prediction step an appropriate number of times. The mean Mahalanobis distance (using either covariance matrix) is used as the distance measure.

The distance between two Topological descriptors is the difference between the expected transition time (given by the Network Layout), and the actual transition time (given by the difference between the time-stamps of the descriptors).

Clearly, distances are defined only between descriptors of the same type, and are not comparable with distances of descriptors of different types. The most suitable method for combining information obtained from distance measurements is to express as conditional probabilities, as shown in the following section.

2.2 Integrating Information from Descriptors in a Probabilistic Framework

Using the above analysis, there are three sources of information that contribute to an overall probability that any given observation is the target. Firstly there is the course-scale temporal information that allows observations from different stations or different areas of the same station to be assessed. Under the proposed

framework, this is considered to be the probability of correct association, given only the time-stamps of the two observations (and the graph of minimum timings between camera views). This is written as $P(i{=}j|Z_T)$.

Secondly, there is the fine-scale spatio-temporal information available about the observations from cameras with overlapping, adjacent or nearby cameras. This conditional probability is written $P(i{=}j|Z_{ST})$. Indeed, in any given case only one of Z_T or Z_{ST} will be available (depending on the relationship between the cameras. Thirdly, there is the appearance information from the colour descriptors, which is written as an appearance measurement Z_A, and hence there is available a conditional probability $P(i{=}j|Z_A)$.

To estimate any of the above conditional probabilities, the likelihoods $P(Z|i{=}j)$ and $P(Z|i{\neq}j)$ of these measurements (distance measures) is estimated for both correct and incorrect matches respectively, by accumulating histograms of both types using a representative data set. Bayes' theorem can then be used to invert the expression, given the two prior expressions for $P(Z)$ and $P(i{=}j)$.

2.3 Performance Evaluation

In this context, the purpose of a descriptor measurement Z is to discriminate subsequent observations of the target from observations of other pedestrians. Using the statistics of the distance measures, an expression for $P(Z|i{=}j)$ can be estimated but the inversion requires the prior probability of $P(i{=}j)$, which will vary according to the circumstances of each observation. Therefore, a simple evaluation measure for the descriptor is the percentage uncertainty removed by the measurement Z, given equal prior probability of correct or incorrect match. If the two histograms are identical then this is zero, if they are disjoint then this is 100%. Writing as X, the random variable denoting the match as correct or incorrect, the expression is written as

$$H\left(X;\, Z\right) = \frac{1}{\log 2} \cdot \sum_{X,Z} P\left(x,\, z\right) \cdot \log\left(\frac{P(x,\, z)}{P(x) \cdot P(z)}\right) \tag{1}$$

This corresponds to how seperated the two histograms are, and the metric is used in Section 4 to evaluate the descriptors' performance, following the colour constancy correction outlined in the next Section.

3 Data Pre-Processing for Improved Constancy

To track the path of the target as they move through the metro, appearance descriptions are compared to calculate a probability of correct association. For this comparison to be as effective as possible, any systematic difference in descriptions obtained from two cameras should be identified and removed.

This section describes two approaches that can be adopted to achieve that identification and removal. Both approaches are designed to work in the input signal vector space, for example, the colour space for the video signal, or the

ground-plane space for measurements of position derived from either camera. Both approaches use the statistics of the input signal gathered in a training phase. The two approaches use first- and second-order parameterization of the relevant statistics, as described below, to normalize the input intensity and chromaticity data.

3.1 First-Order Normalization of Chromaticity and Intensity

Systematic variations between cameras of the input signals may be caused by differences in hardware, configuration, illumination and setting. In this work the input data is represented in the YCbCr colour-space, each channel having 8 bits *i.e.* a range of between 0 and 255. Several colour constancy techniques exist, such as [9] and [10], that require a set of reference colours to be given as input. However, given the high number of cameras in the test site (22 per station, on 12 stations), a completely unsupervised method has been preferred. The approach presented in this paper is an adaptation of the 'gray world' algorithm [11], employed in the context of foreground data in common between cameras [12]. It assumes that the set of foreground data from each camera is a fair sample from an underlying distribution of object appearances which need to be rendered as invariant as possible with respect to through which camera they are observed. To normalise the data from two or more cameras, it is assumed that there is available a training set of observations drawn from the same distribution of passengers and poses. The foreground data is then segmented from the background data. Writing each (Y, Cb_i, Cr_i) foreground pixel value from camera i as $\mathbf{c}_i(x, y)$, and the mean value of *all* foreground data from this camera as $\bar{\mathbf{c}_i}$, the first-order corrected values are calculated as

$$\mathbf{c}_i{}'(x, y) = \frac{128}{\bar{\mathbf{c}_i}} \mathbf{c}_i(x, y) \tag{2}$$

This enforces a mean of 128 for each channel of the signal, *i.e.* mid-scale luminance and neutral chromaticity.

3.2 Second Order Normalization of Chromaticity and Intensity

It is also possible to consider the covariance, when equating the per-camera distributions of foreground colour data. If a particular value of colour signal $\mathbf{c}_i(x, y)$ from camera i is represented as a different value $\mathbf{c}_j(x, y)$ in camera j, then the general relationship between these spaces may be written as an affine transform

$$\mathbf{c}_j = R_{ij}\mathbf{c}_i + \mathbf{t}_{ij} \tag{3}$$

where R_{ij} and $\mathbf{t}_{ij}$ are a generalized rotation and a translation between the input spaces of cameras i and j. If there is no mixing between the luminance and two chromaticity channels, then only the diagonal elements of R_{ij} will be nonzero. More generally, there may be mixing between the colour channels, corresponding to some small generalized rotation between the axes of the colour space for each

camera. The normalization process is intended to identify and correct for these differences between the signals from various cameras that presents all input signals in a common representation. If the affine model is valid representation of the differences between the colour response from any two cameras, and there is sufficient structure to the covariance structure of the input statistics, then it is possible to estimate R_{ij} and $\mathbf{t}_{ij}$ from an unlabelled training set of foreground data. This is described in the following section.

Required Covariance Properties If two multivariate random variables, X and Y, are related via a linear transformation $\mathbf{y} = T_{XY}\mathbf{x}$, then it may be possible to estimate T_{XY} from the means μ_X, μ_Y and covariances S_X, S_Y, generated from n samples of both X and Y. Simulations can be used to demonstrate the accuracy of the estimate, by generating a set of random samples from X, transforming them into Y using values of T, and then measuring how closely the two sets of vectors are aligned in some standard common co-ordinate system.

It is more convenient to transform them both onto the same 'whitened' co-ordinate system with zero mean and unit diagonal covariance. If S_X diagonalizes into $E_X \Lambda_X E_X^T$, then the transform to the whitened version $\mathbf{w}_X$ of the vector $\mathbf{x}$ is

$$\mathbf{w}_X = \Lambda_X^{-1/2} E_X^T (\mathbf{x} - \mu_X) \tag{4}$$

If the equivalent process is also applied to the variable Y, to obtain a sample of whitened vectors $\mathbf{w}_Y$, then the accuracy of the estimate can be measured as the expected L_2 distance between the whitened samples, *i.e.* $E[|\mathbf{w}_Y - \mathbf{w}_X|^2]$.

The accuracy of the estimate depends weakly on the number of samples n and strongly on the relative magnitudes of the ellipsoid radii (*i.e.* eigenvalues) of the covariance structure associated with X. For the two-dimensional ellipse, the standard term used to describe these relative magnitudes *eccentricity*, which varies between 0 (circle) and 1 (a line). In this paper, an alternative term is used that is better suited to the investigation and is not limited to two dimensions. The *ellipticality*, ϵ, is defined as the smallest ratio between successive eigenvalues, ordered by size, and can vary between 1 and inf.

Experiments were conducted to measure the reconstruction accuracy, using several test sets consisting of between 200 and 5,000 samples. Values of ϵ from between 1 and 5 were used, creating covariance matrices with the following form:

$$S = \begin{bmatrix} \epsilon^2 & 0 & 0 \\ 0 & \epsilon & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Figure 1 *left* shows how the accuracy of the estimate of S as a function of ϵ for 4 different sample set sizes, and for $\epsilon = 1 \ldots 5$. The graphs show that accuracy increases with increasing values of ϵ, and that the increase speed is higher with bigger sample-set sizes. Below values of $\epsilon = 2.5$, the alignment of the two sets of data failed completely and the accuracy was no better than random. Above values of $\epsilon = 3.0$, the alignment worked very well and the mean squared error

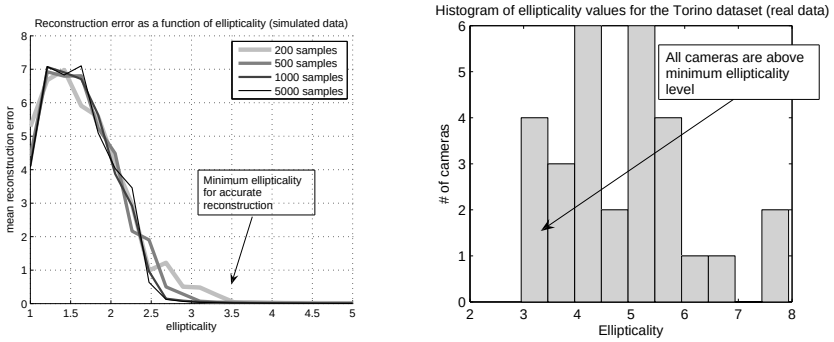

Fig. 1. *Left:* Histogram showing all ϵ values of the Torino dataset. In total 29 videos were used, giving an average ϵvalue of 4.8751 ± 1.6490. *Right:* Accuracy in the reconstruction of the mean μ and covariance S of a normal distribution, as a function of the ϵ of S, for 4 sample-set sizes.

between the two sets of whitened vectors rapidly approached zero. In the range $2.5 < \epsilon < 3$, the alignment process depended on the number of samples. With 5,000 samples the error was low, with 200 samples the error was high. In the next section, the values of ϵ are estimated for the real video signals encountered in a surveillance system.

3.3 Analysis of Input Signal Covariance

The statistics of 29 surveillance video feeds from the Torino Metro system were analysed. Approximately 10 minutes per camera were used; only foreground pixels (extracted using a GMM method) were included. If the eigenvalues of the covariance of these data are sufficiently different, then the simulations suggest that the procedure outlined in the previous section may be applied to improve the alignment of these colour signals.

As shown in Fig. 1 *right*, the values of ϵ are all between 3 and 8. This ensures that we will be able to reconstruct the cameras covariances with good accuracy.

4 Results

The proposed descriptors (along with the proposed colour constancy corrections) were applied to a real set of surveillance video data from multiple stations. The experiments were designed to measure the most effective combination of colour constancy correction, appearance descriptor, and also the relative efficacy of Temporal and Spatio-Temporal descriptors.

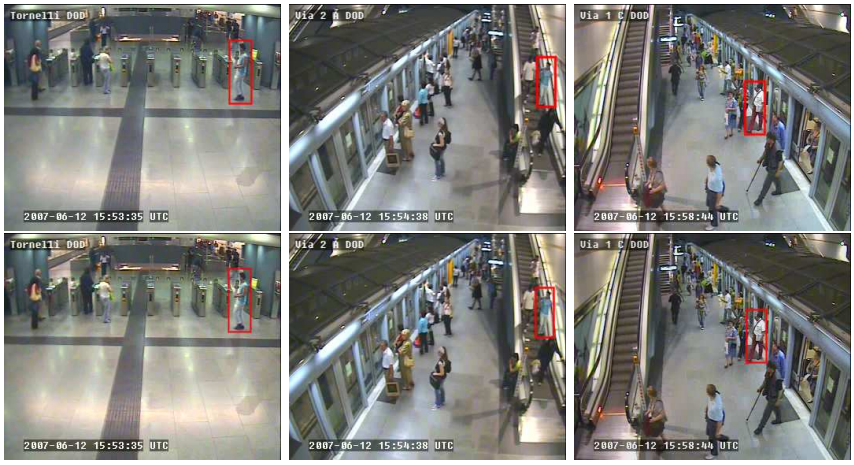

Fig. 2. Some input video files. *Top row:* no colour correction. *Bottom row:* second order colour correction (the difference may not be visible in print). *Left* and *centre*: two views of the same target, yielding a low distance in the Mean-Colour descriptor space. *Right:* a view of a different target, whose descriptor also yields a low distance.

4.1 Input Data

A total of 29 cameras from 2 stations (XVIII Dicembre and Racconigi) in the Torino metro system were used in the experiments. The video format is MPEG-4 Part II, in 2-CIF, *i.e.* 704×288, at 5 fps. The dataset consists of 14 people travelling between the two stations. Each person is observed 10-15 times, giving a total of 178 observations (see Fig. 2 for an example of input data).

All but two of the 29 cameras are calibrated to the ground plane (these two did not have a sufficient number of known points in the view to permit a calibration). Those belonging to the same level of the same station are calibrated with respect to the same world reference frame. For adjacent cameras belonging to different levels or different stations, the average transition (occlusion) time is defined. The cameras are representative of the various situations arising in surveillance scenarios: some are overlapping, some are separated by few occlusions, and some are separated by long occlusions when passengers are walking for some distance between the fields of view. Cameras in different stations are separated by at least the duration of the train journey, which is 2 minutes in this case since two adjacent stations were chosen.

4.2 Experimental Procedure

Each descriptor type was used to generate a description of each observation. Then, for each descriptor type, the distance measures between all observation descriptions are calculated and used to compile two histograms, one for the

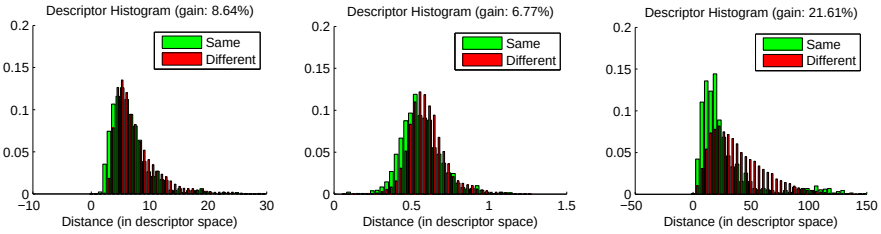

Fig. 3. Same/Different histograms of the Covariance, Dominant Colour and Mean Colour descriptors, after colour normalisation (using second order statistics with a target covariance of 200 for all channels).

case of the same target, and the other for the case of different target. These histograms are plotted in Figure 3: they show the efficacy of the appearance descriptors, applied to the colour-corrected videos (the 2nd order normalisation model was used). The histograms can be used to estimate the likelihood of obtaining a measurement Z_A from a pair of observations i, j given that they are the same (green bins) or different (red bins). The more separated the green and red histograms are, the more effective the descriptor is at discriminating between these cases. It can be seen in the figure that this is the case with the videos corrected with second order normalisation.

Table 1 shows the information gain (percentage) obtained by all the combinations of appearance descriptor and colour constancy methods, assuming equal prior.

Figure 4 shows the corresponding histograms for for Spatio-Temporal and Topological descriptors.

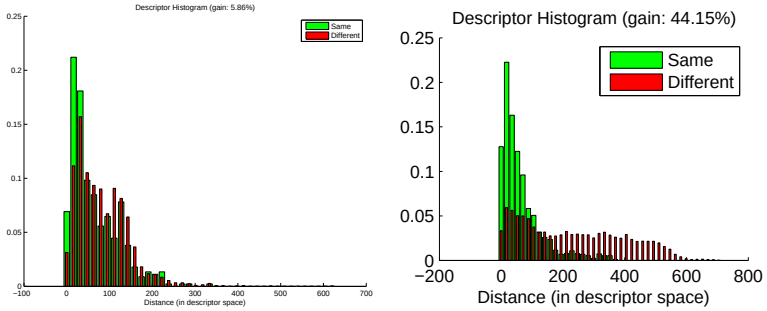

Fig. 4. Spatio-Temporal (*Left*) and Topological (*right*) descriptors.

In order to be able to combine several descriptors together, the problem is cast into a probabilistic framework. Figure 5 shows the process through which

	Colour normalisation				
			2nd order methods		
Descriptor	None	1st order	var. only[a]	global[b]	whitened[c]
Covariance	8.88	8.49	8.75	7.56	8.64
Dominant Colour	9.03	9.67	9.25	7.27	6.77
Mean Colour	11.58	9.39	9.26	8.60	21.61

[a] ignore cross-correlation
[b] force target covariance to match an average, global covariance
[c] force target covariance to be the same on all channels

Table 1. Comparison of the descriptors, in terms of information gain, with respect to colour normalisation.

the histograms of a descriptor are fit to two Negative Binomial distributions that represent $P(Z|i{=}j)$ and $P(Z|i{\neq}j)$, and then inverted using Bayes' Theorem to obtain $P(i{=}j|Z)$ and $P(i{\neq}j|Z)$, that is, the probability that an observation represents the same target or a different one.

5 Discussion and Conclusion

This paper presents work towards the construction of a multi-camera, single-person tracking system. The data processing required prior to tracking includes camera calibration, motion segmentation, colour constancy and description generation. In this paper, the focus was on comparing different colour constancy methods and different descriptors, as well as presenting a Bayesian framework for combining the ouputs of different descriptors. The results, obtained by applying the algorithms to real surveillance videos, can be summarised in two points. Firstly, the application of colour constancy colour constancy does not necessarily improve the performance of appearance descriptors. Secondly, the simplest of the descriptors (Mean Colour) significantly outperformed much more complex descriptors when an appropriate colour constancy method was used. Indeed, the other descriptors gained little, if anything, from colour constancy.

Future work includes therefore the investigation of how to obtain a colour constancy that consistently improves the performance of appearance descriptors. Furthermore, work must address the problem of crowded and cluttered scenarios, where the assumption that the main connected foreground component corresponds to a pedestrian's observation only applies to very specific cases (*e.g.* at the end of an escalator or at entrance/exit turnstyles).

References

1. Velastin, S., Boghossian, B., Lo, B., Sun, J., Vicencio-Silva, M.: Prismatica: toward ambient intelligence in public transport environments. Systems, Man and Cybernetics, Part A, IEEE Transactions on **35** (2005) 164–182

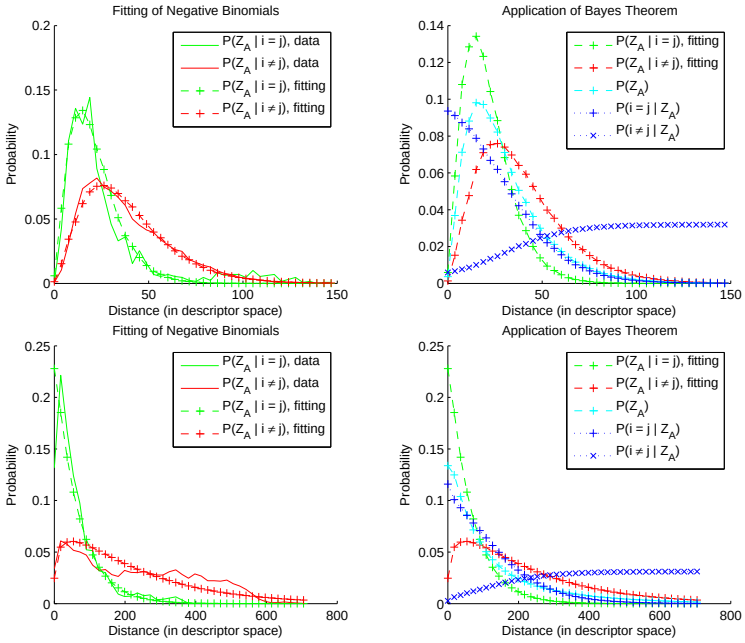

Fig. 5. *Left:* Fitting the histograms of the Mean Colour (*top*) and Topological (*bottom*) descriptors. *Right:* Using Bayes' Theorem to invert the probability density functions.

2. Porikli, F., Ivanov, Y., Haga, T.: Robust abandoned object detection using dual foregrounds. EURASIP J. Adv. Signal Process **2008** (2008) 1–10
3. Schwerdt, K., Maman, D., Bernas, P., Paul, E.: Target segmentation and event detection at video-rate: the eagle project. avss **0** (2005) 183–188
4. Mecocci, A., Pannozzo, M.: A completely autonomous system that learns anomalous movements in advanced videosurveillance applications. In: International Conference on Image Processing. Volume 2. (2005) 586–9
5. Renno, J., Remagnino, P., Jones, G.: Learning surveillance tracking models for the self-calibrated ground plane. Acta Automatica Sinica **29** (2003) 381–392
6.
7. Cohen, I., Ma, Y., Miller, B.: Tracking moving objects across non-overlapping cameras. In: Optics and Photonics for Counterterrorism and Crime Fighting III. Edited by Lewis, Colin. Proceedings of the SPIE. Volume 6741 of Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference. (2007)
8. Sikora, T.: The mpeg-7 visual standard for content description-an overview. Circuits and Systems for Video Technology, IEEE Transactions on **11** (2001) 696–702
9. Porikli, F.: Inter-camera color calibration by cross-correlation model function. In: International Conference on Image Processing. Volume 2. (2003) 133–136
10. Javed, O., Rasheed, Z., Shafique, K., Shah, M.: Tracking across multiple cameras with disjoint views. In: ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision, Washington, DC, USA, IEEE Computer Society (2003) 952
11. Funt, B.: Color constancy in digital imagery. In: International Conference on Image Processing. Volume 3. (1999) 55–59
12. Gilbert, A., Bowden, R. Lecture Notes in Computer Science. In: Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. Springer Berlin / Heidelberg (2006) 125–136