Appearance Modeling for Tracking in Multiple Non-overlapping Cameras

Omar Javed
Computer Vision Lab,
University of Central Florida
Orlando, FL, U.S.A
ojaved@cs.ucf.edu

Khurram Shafique
Computer Vision Lab,
University of Central Florida
Orlando, FL, U.S.A
khurram@cs.ucf.edu

Mubarak Shah
Computer Vision Lab,
University of Central Florida
Orlando, FL, U.S.A
shah@cs.ucf.edu

Abstract

When viewed from a system of multiple cameras with non-overlapping fields of view, the appearance of an object in one camera view is usually very different from its appearance in another camera view due to the differences in illumination, pose and camera parameters. In order to handle the change in observed colors of an object as it moves from one camera to another, we show that all brightness transfer functions from a given camera to another camera lie in a low dimensional subspace and demonstrate that this subspace can be used to compute appearance similarity. In the proposed approach, the system learns the subspace of inter-camera brightness transfer functions in a training phase during which object correspondences are assumed to be known. Once the training is complete, correspondences are assigned using the maximum a posteriori (MAP) estimation framework using both location and appearance cues. We evaluate the proposed method under several real world scenarios obtaining encouraging results.

1. Introduction

The problem of estimating the trajectory of an object as the object moves in an area of interest is known as *tracking* and it is one of the major topics of research in computer vision. In most cases, it is not possible for a single camera to observe the complete area of interest because the camera field of view is finite, and the structures in the scene limit the visible areas. Therefore, surveillance of wide areas requires a system with the ability to track objects while observing them through multiple cameras. Moreover, it is usually not feasible to completely cover large areas with cameras having overlapping views due to economic and/or computational reasons. Thus, in realistic scenarios, the system should be able to handle multiple cameras with non-overlapping fields of view.

A commonly used cue for tracking in a single camera is the appearance of the objects. Appearance of an object can be modelled by its color or brightness histograms, and it is a function of scene illumination, object geometry, object surface material properties (e.g., surface albedo) and the camera parameters. Among all these, only the object surface material properties remain constant as an object moves across cameras. Thus, the color distribution of an object can be fairly different when viewed from two different cameras. One way to match appearances in different cameras is by finding a transformation that maps the appearance of an object in one camera image to its appearance in the other camera image. However, for a given pair of cameras, this transformation is not unique and also depends upon the scene illumination and camera parameters. In this paper, we show that despite depending upon a large number of parameters, for a given pair of cameras, all such transformations lie in a low dimensional subspace. The proposed method learns this subspace of mappings (brightness transfer functions) for each pair of cameras from the training data by using probabilistic principal component analysis. Thus, given appearances in two different cameras, and the subspace of brightness transfer functions learned during the training phase, we can estimate the probability that the transformation between the appearances lies in the learnt subspace.

In the following section, we discuss the related research. In Section 3, we show that all BTFs from a given camera to another camera lie in a low dimensional subspace. In Section 4, we present a method to learn this subspace from the training data. In Section 5, we use a probabilistic formulation for tracking in multiple cameras and employ the BTF subspace to determine how likely it is for observations in different cameras to belong to the same object. In Section 6, we present experiments which validate the proposed approach.

2 Related Work

Makris et al. [10] and Rahimi et al. [14] used the information gained from observing location and velocity of objects moving across multiple non-overlapping cameras to determine spatial relationships between cameras. Object correspondences were not assumed to be known in [10], while they were assumed to be known in [14]. Appearance of objects was not used by both methods. In this paper we demonstrate that appearance modelling can supplement the spatio-temporal information for robust tracking.

Huang and Russel [5] presented a probabilistic approach for tracking vehicles across two cameras on a highway. The object appearance was modeled by the mean of the color of the whole object, which is not enough to distinguish between multi-colored objects like people. Inter-camera transition times were modeled as Gaussian distributions and the problem was transformed into a weighted assignment problem for establishing correspondence. Kettnaker and Zabih [9] used a Bayesian formulation of the problem of reconstructing the paths of objects across multiple non-overlapping cameras. Their system required manual input of the topology of allowable paths of movement and the transition probabilities. The appearance of objects was represented by color histograms. Kang et al.[8] presented a method for tracking in overlapping stationary and pan-tilt-zoom cameras. The object appearance was modeled by partitioning the object region into its polar representation. In each partition a Gaussian distribution modeled the color variation. However, in case of non-overlapping cameras, there can be a significant difference in illumination in each of the viewable regions, therefore directly matching the color distributions of objects would not give accurate results. One possible solution to this problem was proposed by Porikli [13]. In his approach, a brightness transfer function (BTF) f_{ij} is computed for every pair of cameras C_i and C_j, such that f_{ij} maps an observed color value in Camera C_i to the corresponding observation in Camera C_j. Once such a mapping is known, the correspondence problem is reduced to the matching of transformed histograms or appearance models. Unfortunately, this mapping, i.e., the BTF, is not unique and it varies from frame to frame depending on a large number of parameters that include illumination, scene geometry, exposure time, focal length, and aperture size of each camera.

3 The Space of Brightness Transfer Functions

In this section, we show that the BTFs for a given pair of cameras lie in a small subspace of the space of all possible BTFs. This subspace is learned from training data and is used for appearance matching of objects during a test phase. Note that a necessary condition, for the existence of a one-to-one mapping of brightness values from one camera to another, is that the objects are planar and only have diffuse reflectance.

Let $L_i(p,t)$ denote the scene radiance at a (world) point p of an object that is illuminated by white light, when viewed from camera C_i at time instant t. By the assumption that the objects do not have specular reflectance, we may write $L_i(p,t)$ as a product of (a) material related terms, $M_i(p,t) = M(p)$ (for example, albedo) and (b) illumination/camera geometry and object shape related terms, $G_i(p,t)$, i.e.,

$$L_i(p,t) = M(p)G_i(p,t). \quad (1)$$

The above given Bi-directional Reflectance Distribution Function (BRDF) model is valid for commonly used BRDFs, such as, the Lambertian model and the generalized Lambertian model [12] (See Table 1). By the assumption of planarity, $G_i(p,t) = G_i(q,t) = G_i(t)$, for all points p and q on a given object. Hence, we may write, $L_i(p,t) = M(p)G_i(t)$.

The image irradiance $E_i(p,t)$ is proportional to the scene radiance $L_i(p,t)$ [4], and is given as:

$$E_i(p,t) = L_i(p,t)Y_i(t) = M(p)G_i(t)Y_i(t), \quad (2)$$

where $Y_i(t) = \frac{\pi}{4}\left(\frac{d_i(t)}{h_i(t)}\right)^2 \cos^4 \alpha_i(p,t) = \frac{\pi}{4}\left(\frac{d_i(t)}{h_i(t)}\right)^2 c$, is a function of camera parameters at time t. $h_i(t)$ and $d_i(t)$ are the focal length and diameter (aperture) of lens respectively, and $\alpha_i(p,t)$ is the angle that the principal ray from point p makes with the optical axis. The fall off in sensitivity due to the term $\cos^4 \alpha_i(p,t)$ over an object is considered negligible [4] and may be replaced with a constant c.

If $X_i(t)$ is the time of exposure, and g_i is the radiometric response function of the camera C_i, then the measured (image) brightness of point p, $B_i(p,t)$, is related to the image irradiance as

$$\begin{aligned} B_i(p,t) &= g_i\left(E_i(p,t)X_i(t)\right) \\ &= g_i\left(M(p)G_i(t)Y_i(t)X_i(t)\right), \end{aligned}$$

i.e., the brightness, $B_i(p,t)$, of the image of a world point p at time instant t, is a nonlinear function of the product of its material properties $M(p)$, geometric properties $G_i(t)$, and camera parameters, $Y_i(t)$ and $X_i(t)$. Consider two cameras, C_i and C_j. Assume that a world point p is viewed by cameras C_i and C_j at time instants t_i and t_j respectively. Since material properties M of a world point remain constant, we have,

$$M(p) = \frac{g_i^{-1}\left(B_i(p,t_i)\right)}{G_i(t_i)Y_i(t_i)X_i(t_i)} = \frac{g_j^{-1}\left(B_j(p,t_j)\right)}{G_j(t_j)Y_j(t_j)X_j(t_j)}. \quad (3)$$

Hence, the brightness transfer function from the image of camera C_i at time t_i to the image of camera C_j at time t_j is given by:

$$\begin{aligned} B_j(p,t_j) &= g_j\left(\frac{G_j(t_j)Y_j(t_j)X_j(t_j)}{G_i(t_i)Y_i(t_i)X_i(t_i)} g_i^{-1}\left(B_i(p,t_i)\right)\right) \\ &= g_j\left(w(t_i,t_j)g_i^{-1}\left(B_i(p,t_i)\right)\right), \quad (4) \end{aligned}$$

where $w(t_i,t_j)$ is a function of camera parameters and illumination/scene geometry of cameras C_i and C_j at time instants t_i and t_j respectively. Since Equation 4 is valid for any point p on the object visible in the two cameras, we may drop the argument p from the notation. Also, since it is implicit in the discussion that the BTF is different for any two pair of frames, we will also drop the arguments t_i and t_j for the sake of simplicity. Let f_{ij} denote a BTF from camera C_i to camera C_j, then,

$$B_j = g_j\left(wg_i^{-1}\left(B_i\right)\right) = f_{ij}\left(B_i\right). \quad (5)$$

In this paper, we use a non-parametric form of the BTF by sampling f_{ij} at a set of fixed increasing brightness values $B_i(1) < B_i(2) < \ldots < B_i(d)$, and representing it as a vector. That is, $(B_j(1),\ldots,B_j(d))$=$(f_{ij}(B_i(1)),\ldots,f_{ij}(B_i(d)))$. We denote the space of brightness transfer functions (SBTF) from camera C_i to camera C_j by Γ_{ij}. It is easy to see that the dimension of Γ_{ij} can be at most d, where d is the number of discrete brightness values (For most imaging systems, $d = 256$). However, the following theorem shows that BTFs actually lie in a small subspace of the d dimensional space (Please see Appendix I for proof).

Theorem 1. *The subspace of brightness transfer functions Γ_{ij} has dimension at most m if for all $a,x \in \mathbb{R}$, $g_j(ax) = \sum_{u=1}^{m} r_u(a)s_u(x)$, where g_j is the radiometric response function of camera C_j, and for all u, $1 \le u \le m$, r_u and s_u are arbitrary, but fixed 1D functions.*

From Theorem 1, we see that the upper bound on the dimension of subspace depends on the radiometric response function of camera C_j. Though the radiometric response functions are usually nonlinear and differ from one camera to another, they do not

Model	M	G
Lambertian	ρ	$\frac{I}{\pi}\cos\theta_i$
Generalized Lambertian	ρ	$\frac{I}{\pi}\cos\theta_i \left[1-\frac{0.5\sigma^2}{\sigma^2+0.33}+\frac{0.15\sigma^2}{\sigma^2+0.09}\cos(\phi_i-\phi_r)\sin\alpha\tan\beta\right]$

Table 1. Commonly used BRDF models that satisfy Equation 1. The subscripts i and r denote the incident and the reflected directions measured with respect to surface normal. I is the source intensity, ρ is the albedo, σ is the surface roughness, $\alpha = \max(\theta_i, \theta_r)$ and $\beta = \min(\theta_i, \theta_r)$. Note that for generalized Lambertian model to satisfy Equation 1, we must assume that the surface roughness σ is constant over the plane.

have exotic forms and are well-approximated by simple parametric models. Many authors have approximated the radiometric response function of a camera by a gamma function [2, 11], i.e., $g(x) = \lambda x^\gamma + \mu$. Then, for all $a, x \in \mathbb{R}$,

$$g(ax) = \lambda(ax)^\gamma + \mu = \lambda a^\gamma x^\gamma + \mu = r_1(a)s_1(x) + r_2(a)s_2(x),$$

where, $r_1(a) = a^\gamma$, $s_1(x) = \lambda x^\gamma$, $r_2(a) = 1$, and $s_2(x) = \mu$. Hence, by Theorem 1, if the radiometric response function of camera C_j is a gamma function, then the SBTF Γ_{ij} has dimensions at most 2. As compared to gamma functions, polynomials are a more general approximation of the radiometric response function. Once again, for a degree q polynomial $g(x) = \sum_{u=0}^{q} \lambda_u x^u$ and for any $a, x \in R$, we can write $g(ax) = \sum_{u=0}^{q} r_u(a)s_u(x)$ by putting $r_u(a) = a^u$ and $s_u(x) = \lambda_u x^u$, for all $0 \leq u \leq q$. Thus, the dimension of the SBTF Γ_{ij} is bounded by one plus the degree of the polynomial that approximates g_j. It is shown in [3] that most of the real world response functions are sufficiently well approximated by polynomials of degrees less than or equal to 10.

To show empirically that the assertions made in this subsection remain valid for real world radiometric response functions, we consider 10 synthetic cameras C_u, $1 \leq u \leq 10$ and assign each camera a radiometric response function of some real world camera/film (These response functions are shown in Figure 1). For each synthetic camera C_i, we generate a collection of brightness transfer functions, from C_1 to C_i, by varying w in the equation 5 and perform the principal component analysis on this collection. The plot of percentage of total variance over the number of components is shown in Figure 2. It can be seen from the results that in most of the cases, 4 or less principal components capture significant percentage of the variance of the subspace and hence, justify the theoretical analysis.

In the next section, we will give a method for estimating the BTFs and their subspace from training data in a multi-camera tracking scenario.

4 Estimation of inter-camera BTFs and their subspace

Consider a pair of cameras C_i and C_j. Corresponding observations of an object across this camera pair can be used to compute an inter-camera BTF. One way to determine this BTF is to estimate the pixel to pixel correspondence between the object views in the two cameras (see Equation 5). However, finding pixel to pixel correspondences from views of the same object in two different cameras is not possible due to self-occlusion and difference in pose. Thus, we employ normalized histograms of object brightness values for the BTF computation. Such histograms are relatively robust to changes in object pose [15]. In order to compute the BTF, we assume that the percentage of image points on the observed object O_i with brightness less than or equal to B_i is equal to the percentage of image points in the observation O_j with brightness less than or equal to B_j. Note that, a similar strategy was adopted by Grossberg and Nayar [3] to obtain a BTF between images taken from the same camera of the same view but in different illumination conditions. Now, if H_i and H_j are the normalized cumulative histograms of object observations O_i and O_j respectively, then $H_i(B_i) = H_j(B_j) = H_j(f_{ij}(B_i))$. Therefore, we have

$$f_{ij}(B_i) = H_j^{-1}\left(H_i(B_i)\right), \tag{6}$$

where H^{-1} is the inverted cumulative histogram.

We use Equation 6 to estimate the brightness transfer function $\mathbf{f_{ij}}$ for every pair of observations in the training set. Let F_{ij} be the collection of all the brightness transfer functions obtained in this manner, i.e., $\left\{\mathbf{f_{(ij)_1}}, \mathbf{f_{(ij)_2}}, \ldots, \mathbf{f_{(ij)_N}}\right\}$. To learn the subspace of this collection we use the probabilistic Principal Component Analysis PPCA [16]. According to this model, a d dimensional BTF, $\mathbf{f_{ij}}$, can be written as:

$$\mathbf{f_{ij}} = \mathbf{W}\mathbf{y} + \overline{\mathbf{f_{ij}}} + \epsilon. \tag{7}$$

Here $\mathbf{y}$ is a normally distributed q dimensional latent (subspace) variable, $q < d$, $\mathbf{W}$ is a $d \times q$ dimensional projection matrix that

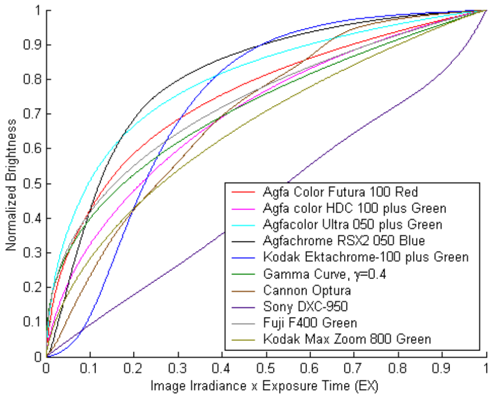


Figure 1. Response Curves assigned to each synthetic camera.

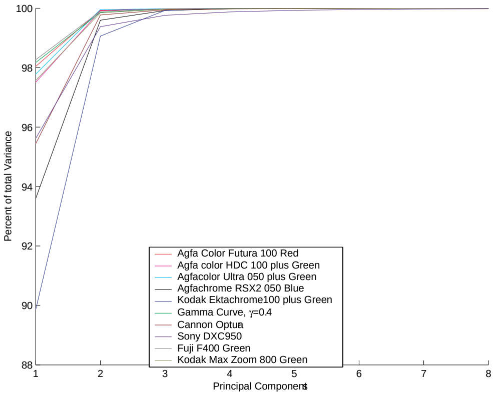


Figure 2. Plots of the percentage of total variance accounted by m principal components (x-axis) of the subspace of brightness transfer functions from synthetic camera C_1 to camera C_i. The plot confirms that a very hight percentage of total variance is accounted by first 3 or 4 principal components of the subspace.

relates the subspace variables to the observed BTF, $\overline{\mathbf{f}_{ij}}$ is the mean of the collection of BTFs, and ϵ is isotropic Gaussian noise, i.e., $\epsilon \sim N(0, \sigma^2 \mathbf{I})$. Given that $\mathbf{y}$ and ϵ are normally distributed, the distribution of f_{ij} is given as

$$\mathbf{f_{ij}} \sim \mathcal{N}(\overline{\mathbf{f_{ij}}}, \mathbf{Z}), \tag{8}$$

where $\mathbf{Z} = \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I}$. Now, as suggested in [16], the projection matrix $\mathbf{W}$ is estimated as

$$\mathbf{W} = \mathbf{U}_q(\mathbf{E}_q - \sigma^2 \mathbf{I})^{1/2}\mathbf{R}, \tag{9}$$

where the q column vectors in the $d \times q$ dimensional $\mathbf{U}_q$ are the eigenvectors of the sample covariance matrix of $\mathbf{F}_{ij}$, $\mathbf{E}_q$ is the $q \times q$ diagonal matrix of corresponding eigenvalues $\lambda_1, \ldots, \lambda_q$, and $\mathbf{R}$ is an arbitrary orthogonal rotation matrix and can be set to an identity matrix. The value of σ^2, which is the variance of the information 'lost' in the projection, is calculated as

$$\sigma^2 = \frac{1}{d-q} \sum_{v=q+1}^{d} \lambda_v. \tag{10}$$

Once the values of σ^2 and $\mathbf{W}$ are known, we can compute the probability of a particular BTF belonging to the learned subspace of BTFs by using the distribution in Equation 8.

Note that till now we have been dealing with only the brightness values of images and computing the brightness transfer functions. To deal with color images we treat each channel, i.e., R, G and B separately. The transfer function for each color channel (color transfer function) is computed exactly as discussed above. The subspace parameters $\mathbf{W}$ and σ^2 are also computed separately for each color channel. Also note that we do not assume the knowledge of any camera parameters and response functions for the computation of these transfer functions and their subspace.

In the next section, we present a formulation of the multi-camera tracking problem, and discuss how the subspace based appearance constraints can be employed along with inter-camera spatiotemporal models to establish correspondence.

5 Formulation of the Multi-Camera Tracking Problem

Suppose that we have a system of r cameras $C_1, C_2, \ldots, C_r$ with non-overlapping views. Further, assume that there are n objects in the environment (the number of the objects is not assumed to be known). Each of these objects are viewed from different cameras at different time instants. Assume that the task of single camera tracking is already solved, and let $O_j = \left\{O_{j,1}, O_{j,2}, \ldots, O_{j,m_j}\right\}$ be the set of m_j observations that were observed by the camera C_j. Each of these observations $O_{j,a}$ is a track of some object from its entry to its exit in the field of view of camera C_j, and is based on two features, appearance of the object $O_{j,a}(app)$ and space-time features of the object $O_{j,a}(st)$ (location, velocity, time etc.). The problem of multi-camera tracking is to find which of the observations in the system of cameras belong to the same object.

For a formal definition of the above problem, we let a correspondence $k_{a,b}^{c,d}$ define the hypothesis that the observations $O_{a,b}$ and $O_{c,d}$ are observations of the same object in the environment, with the observation $O_{a,b}$ preceding the observation $O_{c,d}$. The problem of multi-camera tracking is to find a set of correspondences $K = \left\{k_{a,b}^{c,d}\right\}$ such that $k_{a,b}^{c,d} \in K$ if and only if $O_{a,b}$ and $O_{c,d}$ correspond to successive observations of the same object in the environment. Let Σ be the solution space of the multi-camera tracking problem. We assume that each observation of an object is preceded or succeeded by a maximum of one observation (of the same object). We define the solution of the multi-camera tracking problem to be a hypothesis K' in the solution space Σ that maximizes the a posteriori probability, and is given by:

$$\begin{aligned} K' &= \arg\max_{K \in \Sigma} \prod_{k_{i,a}^{j,b} \in K} \Bigg(P\left(O_{i,a}(app), O_{j,b}(app) | k_{i,a}^{j,b}\right) \\ & \qquad P\left(O_{i,a}(st), O_{j,b}(st) | k_{i,a}^{j,b}\right) P\left(C_i, C_j\right) \Bigg). \end{aligned} \tag{11}$$

If the space-time and appearance probability density functions are known then the posterior can be maximized using a graph theoretic approach. The details of the formulation and the maximization scheme are given in our previous work [7]. We now discuss the choice of appearance and space-time pdfs, i.e., $P\left(O_{i,a}(st), O_{j,b}(st) | k_{i,a}^{j,b}\right)$ and $P\left(O_{i,a}(app), O_{j,b}(app) | k_{i,a}^{j,b}\right)$.

Note that, the training phase provides us the subspace of color transfer functions between the cameras, which models how colors of an object can change across the cameras. During the test phase, if the mapping between the colors of two observations is well explained by the learned subspace then it is likely that these observations are generated by the same object. Specifically, for two observations $O_{i,a}$ and $O_{j,b}$ with color transfer functions (whose distribution is given by Equation 8) $\mathbf{f}_{i,j}^R$,$\mathbf{f}_{i,j}^G$ and $\mathbf{f}_{i,j}^B$,we define the probability of the observations belonging to same object as

$$P_{i,j}(O_{i,a}(app), O_{j,b}(app)|k_{i,a}^{j,b}) =$$

$$\prod\nolimits_{ch\in\{R,G,B\}} \frac{1}{(2\pi)^{\frac{d}{2}}\left|\mathbf{Z}^{ch}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\mathbf{f}_{ij}^{ch}-\overline{\mathbf{f}_{ij}^{ch}}\right)^{T}(\mathbf{Z}^{ch})^{-1}\left(\mathbf{f}_{ij}^{ch}-\overline{\mathbf{f}_{ij}^{ch}}\right)},$$

where $\mathbf{Z} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$. The ch superscript denotes the color channel for which the value of $\mathbf{Z}$ and $\overline{\mathbf{f}_{ij}}$ were calculated. For each color channel, the values of $\mathbf{W}$ and σ^2 are computed from the training data using Equation 9 and Equation 10 respectively.

The Parzen window technique is used to estimate the space-time pdfs between each pair of cameras. Suppose we have a sample S consisting of n, d dimensional, data points $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ from a multi-variate distribution $p(\mathbf{x})$, then an estimate $\hat{p}(\mathbf{x})$ of the density at $\mathbf{x}$ can be calculated using

$$\hat{p}(\mathbf{x}) = \frac{\mathbf{1}}{\mathbf{n}}|\mathbf{H}|^{-\frac{1}{2}}\sum_{\mathbf{i=1}}^{\mathbf{n}}\kappa(\mathbf{H}^{-\frac{1}{2}}(\mathbf{x}-\mathbf{x_i})), \qquad (12)$$

where the d variate kernel $\kappa(\mathbf{x})$ is a bounded function satisfying $\int \kappa(\mathbf{x})d\mathbf{x} = 1$, and $\mathbf{H}$ is the symmetric $d \times d$ bandwidth matrix. The position/time feature vector x, used for learning the space-time pdf's from camera C_i to C_j, i.e., $P(O_{i,a}(st), O_{j,b}(st)|k_{i,a}^{j,b})$, is a vector, consisting of the exit location and entry locations in cameras, exit velocities, and the time interval between exit and entry events. Each time, a correspondence is made during the training phase, the observed feature is added to the sample S. In order to reduce the complexity, $\mathbf{H}$ is assumed to be a diagonal matrix, i.e., $\mathbf{H} = diag[h_1^2, h_2^2, \ldots, h_d^2]$.

6 Results

In this section, we present the results of the proposed method in three different multi-camera scenarios. The scenarios differ from each other both in terms of camera topologies and scene illumination conditions, and include both indoor and outdoor settings. Each experiment consists of a training phase and a testing phase. In both phases, the single camera object detection and tracking information is obtained by using the method proposed in [6]. In the training phase, the correspondences are assumed to be known and this information is used to compute the density of the space-time features (entry and exit locations, exit velocity and inter-camera time interval) and the subspaces of transfer functions for each color channel (red, blue, and green). In the testing phase, these correspondences are computed using the proposed multi-camera correspondence algorithm. The performance of the algorithm is analyzed by comparing the resulting tracks to the ground truth. We say that an object in the scene is tracked *correctly* if it is assigned a single unique label for the complete duration of its presence in the area of interest. The *tracking accuracy* is defined as the ratio of the number of objects tracked correctly to the total number of objects that passed through the scene.

In order to demonstrate the superiority of the subspace based method, this approach is compared to direct color matching for establishing correspondence. Moreover to demonstrate that the appearance matching supplements the spatio-temporal constraints for tracking we also show results for i) only space-time model, ii) only appearance model, and iii) both models. The results of each of these cases are analyzed by using the above defined tracking evaluation measure. The results are summarized in Figures 8(a) and 8(b) and are explained below for each of the experimental setup.

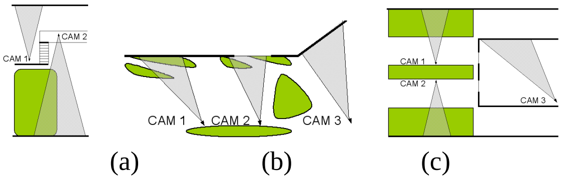


Figure 3. (a)Two camera configuration for the first experiment. The green region is the area covered by grass.(b)Camera setup for sequence 2. All cameras were mounted outdoors. (c) Camera setup for sequence 3. It is an *Indoor/Outdoor Sequence.* Camera 3 is placed indoor while Cameras 1 and 2 are outdoor.

The first experiment was conducted with two cameras, Camera 1 and Camera 2, in an outdoor setting. The camera topology is shown in Figure 3(a). The scene viewed by Camera 1 is a covered area under shade, whereas Camera 2 views an open area illuminated by the sunlight (please see Figure 5). It can be seen from the figure that there is a significant difference between the global illumination of the two scenes, and matching the appearances is considerably difficult without accurate modeling of the changes in appearance across the cameras. The training for the first camera setup was performed by using a five minute sequence. In Figure 4 the transfer functions obtained from the first five correspondences from Camera 1 to Camera 2 are shown. Note that lower color values from Camera 1 are being mapped to higher color values in Camera 2 indicating that the same object is appearing much brighter in Camera 2 as compared to Camera 1. The test phase consisted of a twelve minute long sequence. In this phase, a total of 68 tracks were recorded in the individual cameras and the algorithm detected 32 transitions across the cameras. Tracking accuracy for the test phase is shown in Figure 8(a).

Our second experimental setup consists of three cameras, as shown in Figure 3(a). Testing was carried out on a fifteen minute sequence. A total of 71 tracks in individual cameras were obtained and the algorithm detected 45 transitions within the cameras. All the correspondences were established correctly when both space-time and appearance models were used (see Figure 8).

In the third experiment, three cameras were used for an indoor/outdoor setup Figure 3. Camera 1 was placed indoor while the other two cameras were placed outdoor. Training was done on an eight minute sequence in the presence of multiple persons. Testing was carried out on a fifteen minute sequence. Figure 6 shows some tracking instances for the test sequence. It is clear from Figure 8(a) that both the appearance and space-time models are important sources of information as the tracking results improve significantly when both the models are used jointly.

In Table 2, we show the number of principal components (for each pair of cameras in all three sequences) that account for 99% of the total variance in the inter-camera brightness transfer functions that were computed during the training phase. Note that even

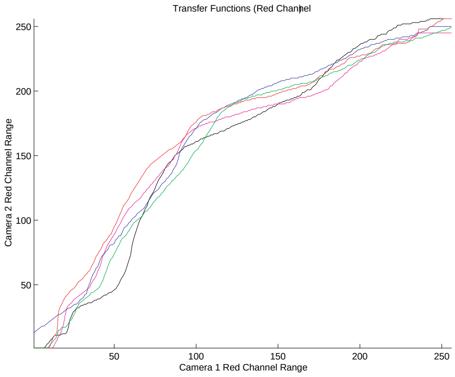


Figure 4. The transfer functions for the Red color channel from Camera 1 to Camera 2, obtained from the first five correspondences from the training data (sequence 1). Note that mostly lower color values from Camera 1 are being mapped to higher color values in Camera 2 indicating that the same object is appearing much brighter in Camera 2 as compared to Camera 1 (as shown in Figure 3(a)).

Seq. #	Camera Pair	# of PCs (Red)	# of PCs (Green)	# of PCs (Blue)
1	1-2	6	5	5
2	1-2	7	7	7
2	2-3	7	7	6
3	1-3	7	6	7
3	2-3	7	7	7

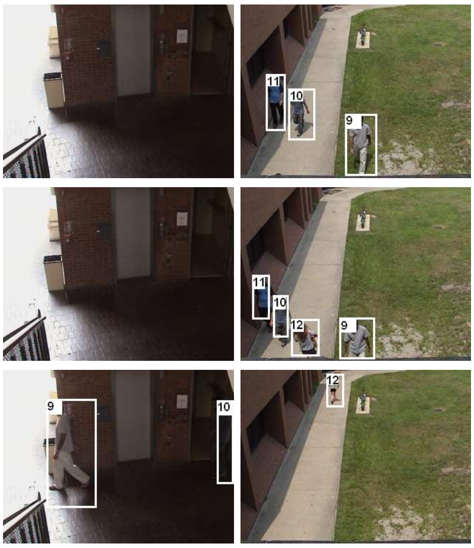
Table 2. The number of Principal Components (PCs) that account for 99% of the variance in the BTFs. Note that for all camera pairs a maximum of 7 principal components were sufficient to account for the subspace of the BTFs.

though the experimental setup does not follow the assumptions of Section3, such as diffuse reflectance or planar objects, the small number of principal components indicates that the inter-camera BTFs lie in a low dimension subspace even in more general conditions.

In order to demonstrate the superiority of the subspace based method we compare it with the direct use of colors for tracking. For direct color base matching, instead of using Equation 12 for the computation of appearance probabilities $P_{i,j}(O_{i,a}(app), O_{j,b}(app)|k_{i,a}^{j,b})$, we define it in terms of the Bhattacharraya distance between the normalized histograms of the observations $O_{i,a}$ and $O_{i,b}$, i.e.,

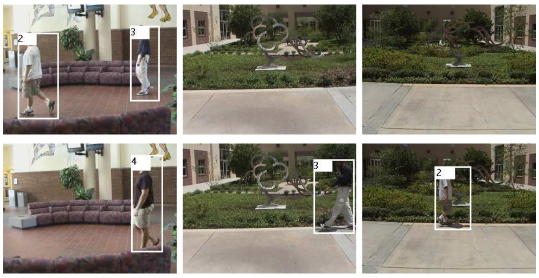
$$P_{i,j}(O_{i,a}(app), O_{j,b}(app)|k_{i,a}^{j,b}) = \gamma e^{-\gamma D(h_i,h_j)}, \qquad (13)$$

where h_i and h_j are the normalized histograms of the observations $O_{i,a}$ and $O_{j,b}$ and D is the modified Bhattacharraya distance [1] between two histograms. The coefficient ranges between zero and one and is a metric.



(a) Cam 1 (b) Cam 2

Figure 5. Frames from sequence 1. Note that multiple persons are simultaneously exiting from camera 2 and entering at irregular intervals in camera 1. The first camera is overlooking a covered area while the second camera view is under direct sun light, therefore the observed color of objects is fairly different in the two views (also see Figure 7). Correct labels are assigned in this case due to accurate color modeling.



(a) Cam 3 (b) Cam 2 (c) Cam 1

Figure 6. Frames from Sequence 3 test phase. A person is assigned a unique label as it moves through the camera views.

Once again, the tracking accuracy was computed for all three multi-camera scenarios using the color histogram based model (Equation 13). The comparison of the proposed appearance modeling approach with the direct color based appearance matching is presented in Figure 8(b), and clearly shows that the subspace based appearance model performs significantly better.

For further comparison of the two methods, we consider two observations, O_a and O_b, in the testing phase, with histograms $H(O_a)$ and $H(O_b)$ respectively. We first compute a BTF, **f**, between the two observations and reconstruct the BTF,

Sequence #	Average BTF Reconstruction Error (Correct Matches)	Average BTF Reconstruction Error (Incorrect Matches)
1	.0003	.0016
2	.0002	.0018
3	.0005	.0011

Table 3. The average normalized reconstruction errors for BTFs between observations of the same object and also between observation of different objects.

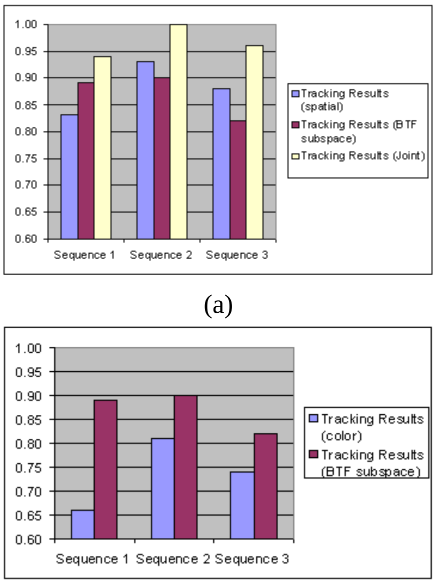
$\mathbf{f}^*$, from the subspace estimated from the training data, i.e., $\mathbf{f}^* = \mathbf{W}\mathbf{W}^T\left(\mathbf{f} - \bar{\mathbf{f}}\right) + \bar{\mathbf{f}}$. Here $\mathbf{W}$ is the projection matrix obtained in the training phase. The first observation O_a is then transformed using $\mathbf{f}^*$, and the histogram of the object O_b is matched with the histograms of both O_a and $\mathbf{f}^*(O_a)$ by using the Bhattacharraya distance. When both the observations O_a and O_b belong to the same object, the transformed histogram gives a much better match as compared to direct histogram matching, as shown in Figure 7 (more results are available in the supplemental file). However, if the observations O_a and O_b belong to different objects then the BTF is reconstructed poorly, (since it does not lie in the subspace of valid BTFs), and the Bhattacharraya distance for the transformed observation either increases or does not change significantly. The aggregate results for the reconstruction error, $\mathbf{f}^*$-Reconstruction Error=$\|\mathbf{f} - \mathbf{f}^*\|/\tau$, where τ is a normalizing constant, for the BTFs between the same object and also between different objects are given in Table 3. The above discussion suggests the applicability of the BTF subspace for the improvement of any multi-camera appearance matching scheme that uses color as one of its components.

7. Conclusions

In this paper, we showed that given some assumptions, all brightness transfer functions from a given camera to another camera lie in a low dimensional subspace. We also demonstrated empirically that even for real scenarios this subspace is low dimensional. The knowledge of camera parameters like focal length, aperture etc was not required for computation of the subspace of BTFs. The proposed system learned this subspace by using probabilistic principal component analysis on the BTFs obtained from the training data and used it for the appearance matching. The appearance matching scheme was combined with space-time cues in a Bayesian framework for tracking. We have presented results on realistic scenarios to show the validity of the proposed approach.

Acknowledgements

This material is based upon work funded in part by the U. S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.



(a)

(b)

Figure 8. (a) Tracking Results. Tracking accuracy for each of the three sequences computed for three different cases. 1. by using only space-time model, 2. by using only appearance model, and 3. both models. The results improve greatly when both the space-time and appearance models are employed for establishing correspondence.(b) Tracking accuracy: comparison of the BTF subspace based tracking method to simple color matching. A much improved matching is achieved in the transformed color space relative to direct color comparison of objects. The improvement is greater in the first sequence due to the large difference in the scene illumination in the two camera views.

Appendix I

Proof:Theorem 1 Let g_i and g_j be the radiometric response functions of cameras C_i and C_j respectively. Also assume that for all $a, x \in \mathbb{R}$, $g_j(ax) = \sum_{u=1}^{m} r_u(a)s_u(x)$, where r_u and s_u are some arbitrary (but fixed) 1D functions, $1 \leq u \leq m$. Let f_{ij} be a brightness transfer function from camera C_i to camera C_j, then according to Equation 5, f_{ij} is given as:

$$\begin{aligned} f_{ij} &= g_j\left(wg_i^{-1}\left(\mathbf{B}_i\right)\right) \\ &= \left[g_j\left(wg_i^{-1}\left(B_i(1)\right)\right) \dots \; g_j\left(wg_i^{-1}\left(B_i(n)\right)\right)\right]^T \end{aligned}$$

Since $g_j(ax) = \sum_{u=1}^{m} r_u(a)s_u(x)$, we may write f_{ij} as follows:

$$\begin{aligned} f_{ij} &= \sum_{u=1}^{m} r_u(w)\left[s_u(g_i^{-1}\left(B_i(1)\right)\right) \dots \; s_u\left(g_i^{-1}\left(B_i(n)\right)\right)]^T \\ &= \sum_{u=1}^{m} r_u(w)s_u\left(g_i^{-1}\left(\mathbf{B}_i\right)\right) \end{aligned}$$

Thus, each brightness transfer function $f_{ij} \in \Gamma_{ij}$ can be represented as a linear combination of m vectors, $s_u\left(g_i^{-1}\left(\mathbf{B}_i\right)\right)$,

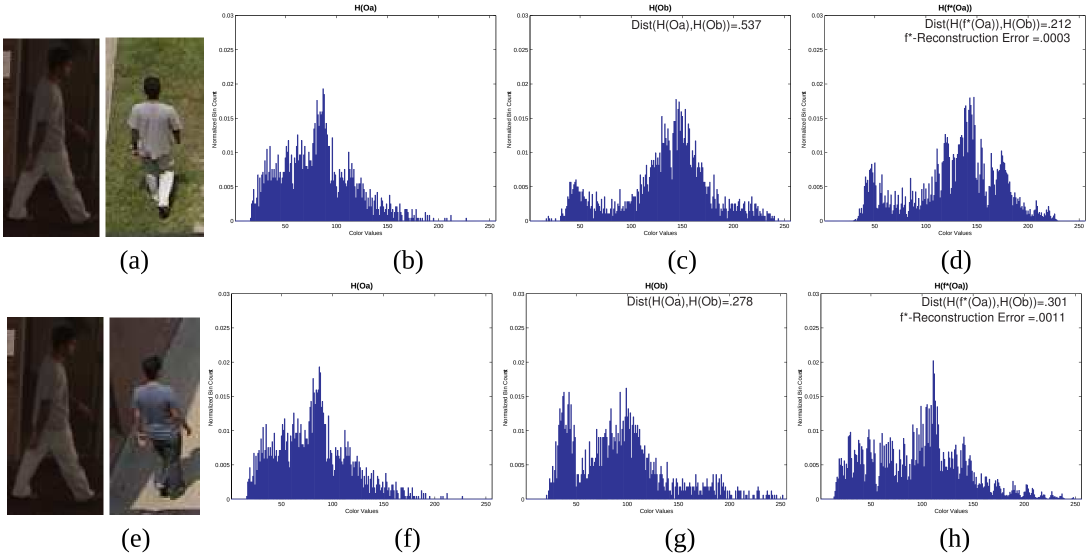


Figure 7. (a) Observations O_a and O_b of the same object from camera 1 and camera 2 respectively from camera setup 1. (b) Histogram of observation O_a (All histograms are of the Red color channel). (c) Histogram of observation O_b. The Bhattacharraya distance between the two histograms of the same object is 0.537. (d) The Histogram of O_a after undergoing color transformation using the BTF reconstruction from the learned subspace. Note that after the transformation the histogram of $(f^*(O_a))$ looks fairly similar to the histogram of O_b. The Bhattacharraya distance reduces to 0.212 after the transformation. (e) Observation from camera 1 matched to an observation from a different object in camera 2. (f,g) Histograms of the observations. The distance between histograms of two different objects is 0.278 . Note that this is less than the distance between histograms of the same object. (h) Histogram after transforming the colors using the BTF reconstructed from the subspace. The Bhattacharraya distance increases to 0.301. Simple color matching gives a better match for the wrong correspondence. However, in the transformed space the correct correspondence gives the least bhattacharraya distance.

$1 \leq u \leq m$. Hence, the dimension of space Γ_{ij} is at most m.

References

[1] D. Comaniciu, v. Ramesh, and P. Meer. “Kernel-based object tracking”. *IEEE Trans. on PAMI*, 25:564–575, 2003.

[2] H. Farid. “ Blind inverse gamma correction”. *IEEE Trans. on Image Processing*, 10(10):1428–1433, Oct 2001.

[3] M. D. Grossberg and S. K. Nayar. “Determining the camera response from images: What is knowable?”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11):1455–1467, November 2003.

[4] B. Horn. *“Robot Vision”*. MIT Press, Cambridge, MA, 1986.

[5] T. Huang and S. Russell. “Object identification in a bayesian context”. In *Proceedings of IJCAI,*, 1997.

[6] O. Javed, Z. Rasheed, O. Alatas, and M. Shah. “Knight-m: A real time surveillance system for multiple overlapping and non-overlapping cameras”. In *IEEE Proc. of ICME*, 2003.

[7] O. Javed, Z. Rasheed, K. Shafique, and M. Shah. Tracking across multiple cameras with disjoint views. In *ICCV*, 2003.

[8] J. Kang, I. Cohen, and G. Medioni. “Continuous tracking within and across camera streams”. In *CVPR*, 2003.

[9] V. Kettnaker and R. Zabih. “Bayesian multi-camera surveillance”. In *CVPR*, pages 117–123, 1999.

[10] D. Makris, T. J. Ellis, and J. K. Black. “Bridging the gaps between cameras ”. In *CVPR*, 2004.

[11] S. Mann and R. Picard. “Being undigital with digital cameras: Extending dynamic range by combining differently exposed pictures”. In *Proc. IS&T 46th Annual Conference*, 1995.

[12] M. Oren and S. K. Nayar. “Generalization of the lambertian model and implications for machine vision”. *International Journal of Computer Vision*, 14(3):227–251, April 1995.

[13] F. Porikli. “Inter-camera color calibration using cross-correlation model function”. In *IEEE Int. Conf. on Image Processing*, 2003.

[14] A. Rahimi and T. Darrell. “Simultaneous calibration and tracking with a network of non-overlapping sensors”. In *CVPR*, 2004.

[15] M. J. Swain and D. H. Ballard. “Indexing via color histograms”. In *ICCV*, 1990.

[16] M. E. Tipping and C. M. Bishop. “Probabilistic principal component analysis”. *Journal of the Royal Statistical Society, Series B*, 61(3):611–622, 1999.