11

Object Tracking in Multiple Cameras with Disjoint Views

Pier Luigi Mazzeo and Paolo Spagnolo
Istituto sui Sistemi Intelligenti per l'automazione (CNR)
Italy

1. Introduction

The specific problem we address in this chapter is object tracking over wide-areas such as an airport, the downtown of a large city or any large public area. Surveillance over these areas consists of the search for suspicious behavior such as persons loitering, unauthorized access, or persons attempting to enter a restricted zone. Until now, these surveillance tasks have been executed by human being who continually observe computer monitors to detect unauthorized activity over many cameras.It is known that the attention level drastically sinks after few hours. It is highly probable that suspicious activity would be unregistered by a human operator. A computer vision system, can take the place of the operator and monitor both immediate unauthorized behavior and long-term suspicious behavior. The "intelligent" system would, then, alert a responsible person for a deeper investigation. In most cases, it is not possible for a single camera to observe the complete area of interest because sensor resolution is finite and structures in the scene limit the visible areas. In the realistic application, surveillance systems which cover large areas are composed by multiple cameras with non overlapping Field of Views (FOVs). Multiple camera tracking is important to establish correspondence among detected objects across different cameras. Given a set of tracks in each camera we want to find which of these tracks belong to the same object in the real world. Note that tracking across multiple cameras is a challenging task because the observations are often widely separated in time and space, especially when viewed from non overlapped FOVs. Simple prediction schemes cannot be used to estimate the object position. It could be observed that object's appearance in one camera view might be very different from its appearance in another camera view due to the difference in illumination, pose and camera properties. It is preferable that any multi-view tracking approach does not require camera calibration or complete site modelling. In fact, the benefit of calibrated cameras or site models is unavailable in most situations. Maintaining calibration between a large sensors network is a discouraging task, it is very onerous recalibrate any sensor when it (sensor) slight changes its position.
This chapter presents an algorithm that deals with all described constraints and tracks objects across multiple un-calibrated cameras with non-overlapping FOVs. We investigate different techniques to evaluate intra-camera and inter-camera tracking algorithm based on object appearances. Object appearance can be modelled by its color or brightness, and it is a function of the scene illumination, object geometry, object surface material properties and camera

parameters. Only the object surface, among these variables, remains constant whereas an object moves across cameras.
We compare different methods to evaluate the color Brightness Transfer Function (**BTF**) between non overlapping cameras. These approaches are based on the color histogram mapped among pairs of images of the same person in different FOVs. It is important to point up that our proposed inter-camera appearance models for tracking do not assume:

- Explicit camera calibration,
- A site model,
- Presence of a single ground plane across cameras,
- A particular non-overlapping camera topology,
- Constant illumination,
- Constant camera parameters, for example, focal length or exposure.

It is, definitively, a flexible calibration-free model.

1.1 Related work

Most of the approaches presented in literature suppose the use of calibrated cameras and the availability of the site model. In (Javed et al., 2008) the conformity in the traversed paths of people and car is used to establish correspondence among cameras. The algorithm learns this conformity and hence the inter-camera relationships in the form of multivariate probability density of spacetime variables (entry and exit locations, velocities, and transition times) using kernel density estimation. To handle the appearance change of an object as it moves from one camera to another, the authors demonstrate that all brightness transfer functions, which map one camera in an another, lie in a low dimensional subspace. This subspace is learned by using probabilistic principal component analysis and used for appearance mapping. In (Du & Piater, 2006) particle filters and belief propagation are combined in a unified framework. In each view, a target is tracked by a dedicated particle-filter-based local tracker. The trackers in different views collaborate via belief propagation so that a local tracker operating in one view is able to take advantage of additional information from other views. In (Du & Piater, 2007) a target is tracked not only in each camera but also in the ground plane by individual particle filters. These particle filters collaborate in two different ways. First, the particle filters in each camera pass messages to those in the ground plane where the multi-camera information is integrated by intersecting the target principal axes. This largely relaxes the dependence on precise foot positions when mapping targets from images to the ground plane using homographies. Secondly, the fusion results in the ground plane are then incorporated by each camera as boosted proposal functions. A mixture proposal function is composed for each tracker in a camera by combining an independent transition kernel and the boosted proposal function. Kalman filters are used in (Chilgunde et al., 2004) to robustly track each target shape and motion in each camera view and predict the target track in the blind region between cameras. For multi-camera correspondence matching, the Gaussian distributions of the tracking parameters across cameras for the target motion and position in the ground plane view are computed. Targets matching across camera views uses a graph based track initialization scheme, which accumulates information from occurrences of target in several consecutive video frames. Geometric and intensity features are used in (Cai & Aggarwal, 1999) to match objects for tracking in a multiple calibrated

camera system for surveillance. These features are modelled as multivariate Gaussian, and Mahalanobis distance measure is used for matching. A method to match object appearances over non-overlapping cameras is presented in (Porikli, 2003).In his approach, a brightness transfer function (BTF) is computed for every pair of cameras. Once such mapping function is known, the correspondence problem is reduced to the matching of transformed histograms or appearance models. However, this mapping, i.e., the BTF varies from frame to frame depending on a large number of parameters which include illumination, scene geometry, exposure time, focal length and aperture size of each camera. Thus, a single pre-computed BTF cannot usually be used to match objects for moderately long sequences. An unsupervised approach to learn edge measures for appearance matching between non-overlapping views is presented by (Shan et al., 2005). The probability of two observations from two cameras being generated by the same or different object is computed to perform the matching. The main constraint of this approach is that the edge images of vehicles have to be registered together. Note that this requirement for registering object images could not be applicable for non rigid objects like pedestrians. A Cumulative Brightness Transfer Function (CBTF) is proposed by (Prosser et al., 2008) for mapping color between cameras located at different physical sites, which makes use available color information from a very sparse training set. A bi-directional mapping approach is used to obtain an accurate similarity measure between pairs of candidate objects. An illumination-tolerant appearance representation, based on online k-means color clustering algorithm is introduced in (Madden et al., 2007), which is capable of coping with the typical illumination changes occurring in surveillance scenarios. A similarity measurement is also introduced to compare the appearance representation of any two arbitrary individuals. In (Jeong & Jaynes, 2008) the distortion function is approximated as general affine transformation and the object appearance is represented as mixture of Gaussians. Appearance models are put in correspondence by searching a bijection function that maximizes a metric for model dissimilarity.

A common characteristic of the above related works is that the knowledge of model sites and particular camera positions in various scenarios allow the usage of geometrical and temporal constraints on the entry/exit image areas. In this way the appearance matching among different cameras is carried out on a sets of individuals that are candidate by their positions to be observed by distributed cameras.

2. Wording of disjoint views multi-camera tracking

Let us suppose that we have a system composed by n cameras $C_1, C_2, ..., C_n$ with non-overlapping views. Let us assume that q objects $P_1, P_2, ..., P_q$ are presented in the scene (the number of the objects in the scene is unknown). Each object is viewed from different cameras at different time instants. Let us call $O_j = \{O_{j,1}, O_{j,2}, ..., O_{j,m_j}\}$ the set of m_j observations that were viewed by the camera C_j. Each observation $O_{j,k}$ with $k = 1..m_j$ is generated by a moving object in the FOV of camera C_j. These observations consist of two part: object appearance $O_{j,k}(a)$ and space-time constraints of the object $O_{j,k}(st)$ (position, velocity, time, and so on). Let us assume, furthermore, that both $O_{j,k}(a)$ and $O_{j,k}(st)$ are independent of each other. Multi-camera tracking problem is centered on finding which of the observations in the system of cameras belong to the same object.

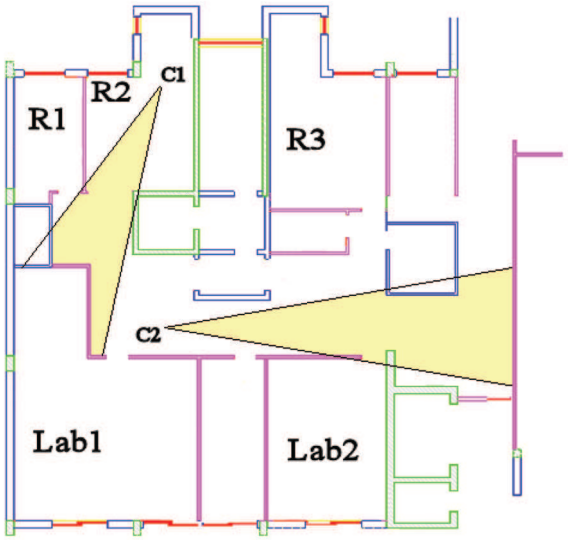


Fig. 1. The cameras configuration inside our office building

Seeing that we have assumed that the single camera tracking problem is solved, the multi-camera tracking task is to link the observations of an object exiting one camera to its observations entering another camera, as the object moves among different camera FOVs.
For a formal problem definition let consider a correspondence among two consecutive observations, i.e., exiting from one camera and entering into another,$O_{i,k}$ and $O_{j,l}$ will be denoted as $\lambda_{j,l}^{i,k}$. Let us consider $\psi(\lambda_{j,l}^{i,k})$ a random variable which is true if and only if $\lambda_{j,l}^{i,k}$ is a valid hypothesis, i.e. $O_{i,k}$ and $O_{j,l}$ are two observations of the same object. We want to find a set of correspondences $\Lambda = \{\lambda_{j,l}^{i,k}, ...\}$ where

$$\lambda_{j,l}^{i,k} \in \Lambda \Longleftrightarrow \beta(\lambda_{j,l}^{i,k}) = true$$

Let Σ be the solution space of the multi-camera tracking problem. If Λ is a candidate solution in Σ, then for all $\{\lambda_{j,l}^{i,k}, \lambda_{q,s}^{p,r}\} \subseteq \Lambda$ where $(j,l) \neq (i,k) \wedge (q,s) \neq (p,r)$. This is because we established that each observation of an object is preceded or succeeded by a maximum of one observation of the same object. Now, let consider Ψ_Λ that is a random variable which is true if and only if Λ represents a valid set of correspondences (all correspondences are correctly established). Ultimately, we want to find a solution in the space Σ of all possible solutions that maximizes the likelihood:

$$\Lambda' = \arg\max_{\Lambda \in \Sigma} P(O|\Psi_\Lambda = true) \tag{1}$$

In other words we can exploit previous equation (eq. 1) obtaining:

$$\Lambda' = \arg\max_{\Lambda \in \Sigma} \prod_{\lambda_{j,l}^{i,k} \in \Lambda} Similarity(O_{i,k}, O_{j,l}) \tag{2}$$

In this way by eq.2 the solution of the multi-camera re-identification problem is defined as the $\Lambda' \in \Sigma$ which maximizes an observation similarity measure. $Similarity()$ is a similarity measure between $O_{i,k}$ and $O_{j,l}$ in the testing data.

3. Appearances changing evaluation across cameras

Here, we want to model the changes in the appearances of an object from one camera to another. Basically, the idea is to learn the changes in the color objects when they move between the cameras, using a set of training data. Based on this set we can extract a function which is able to establish correspondence among object appearance coming from different FOVs. A possible way is proposed in (Porikli, 2003).In his approach, a brightness transfer function (BTF) f_{ij} is estimated for every pair of cameras C_i and C_j such that f_{ij} maps an observed brightness value in camera C_i to the corresponding value in camera C_j. Once that mapping function is known, the correspondence problem is reduced to match the transformed color histogram or the transformed appearance models. It should be noted that, a necessary condition for the existence of the one to one brightness mapping function among two different cameras, is that the object is planar and only has diffuse reflectance. This function, is not unique and it varies from frame to frame depending on different parameters including illumination, scene geometry, exposure time. focal length, aperture size and so on, of each camera. Therefore a single pre-computed mapping cannot normally be used to match objects for any frame sequences. It is possible to show how despite a large number of unknown parameters, all BTFs from a given camera to another one lie in a low dimensional subspace. Moreover, we describe a method to learn this subspace from training data and use this information to determine how, different observations in different cameras belong to the same object. Namely, given observations $O_{i,k}(a)$ and $O_{j,l}(a)$ from cameras C_i and C_j respectively, and given all possible brightness transfer functions (BTFs) from camera C_i to camera C_j we want to compute the probability that the observations $O_{i,k}(a)$ and $O_{j,l}(a)$ belong to the same object.

3.1 The Brightness transfer function space

Let $R_i(p,t)$ be the scene reflectance at a world point $p(x,y,z)$ of an object that is lighted by white light, when it is viewed from camera C_i at time instant t. Assuming that the objects do not have any specular reflectance, we can write $R_i(p,t)$ as a product of a term related to the material $M_i(p,t) = M(p)$ (i.e. albedo) and illumination/camera interaction, geometry and object shape related terms, $S_i(p,t)$, so we have:

$$R_i(p,t) = M(p)S_i(p,t) \tag{3}$$

The above model is used for the description of the Bidirectional Reflectance Distribution Function (BRDF), such as, the Lambertian model and the generalized Lambertian model (Oren & Nayar, 1995) (See Table 1). With the assumption of planarity we have, $S_i(p,t) = S_i(q,t) = G_i(t)$, for all points p and q of a given object. So, we can write eq(3) as

$$R_i(p,t) = M(p)S_i(p) \tag{4}$$

The image irradiance $I_i(p,t)$ is, of course, proportional to the scene radiance $R_i(p,t)$ (Horn, 1986) and we can obtain this:

$$I_i(p,t) = R_i(p,t)Y_i(t) = M(p)S_i(p)Y_i(t) \tag{5}$$

where

$$Y_i(t) = \frac{\pi}{4}\left(\frac{d_i(t)}{h_i(t)}\right)^2 \cos^4 \alpha_i(p,t) \tag{6}$$

is function of some camera parameters at time t. Here we list these intrinsic parameters: $h_i(t)$ is the focal length of the lens; $d_i(t)$ is a lens diameter (aperture); $\alpha_i(p,t)$ is the angle that the light ray from point p makes with the optical axis. However, the sensitivity reduction due to the term $\cos^4 \alpha_i(p,t)$ over an object is consider negligible (Horn, 1986) and can be replaced with a constant c.
Let us now consider $X_i(t)$ which is the time of exposure, and r_i which is the radiometric response function of the camera C_i, then the brightness point measure $B_i(p,t)$, is related to the image irradiance as follow:

$$B_i(p,t) = r_i(I_i(p,t)X_i(t)) = r_i(M(p)S_i(t)Y_i(t)X_i(t)) \tag{7}$$

In other words the image brightness $B_i(p,t)$ of a world point p at time instant t, is a nonlinear function of the product of its materials $M(p)$, its geometric properties $S_i(t)$ and camera parameters, $Y_i(t)$ and $X_i(t)$. Now let consider two cameras, C_i and C_j and let assume that a world point p is observed by both camera C_i and C_j at time instants t_i and t_j, respectively. A material properties M of world point does not change over the time so we have:

$$M(p) = \frac{r_i^{-1}(B_i(p,t_i))}{S_i(t_i)Y_i(t_i)X_i(t_i)} = \frac{r_j^{-1}(B_j(p,t_j))}{S_j(t_j)Y_j(t_j)X_j(t_j)} \tag{8}$$

Therefore the brightness transfer function from the image of C_i camera at time t_i to the camera C_j at time t_j, using equations 7 and 8, become:

$$B_j(p,t_j) = r_j\left(\frac{S_j(t_j)Y_j(t_j)X_j(t_j)}{S_i(t_i)Y_i(t_i)X_i(t_i)} r_i^{-1}(B_i(p,t_i))\right) = r_j(\omega(t_i,t_j) r_i^{-1}(B_i(p,t_i))) \tag{9}$$

where $\omega(t_i,t_j)$ is function of camera parameters and the illumination and scene geometry of cameras C_i and C_j at two different time instant t_i and t_j. So because eq. 9 is valid for all object point p visible by the two cameras, we can eliminate the p argument from the notation. Alike, since it is implicit that the BTF is different for any different pair of frames, we can remove the arguments t_i and t_j for make simpler the equations readability. Let denote with f_{ij} a BTF from camera C_i to camera C_j then starting from eq. 9 we have:

$$B_j = r_j(\omega r_i^{-1}(B_i)) = f_{ij}(B_i) \tag{10}$$

3.2 Inter-camera BTFs estimation

Now let us consider a pair of cameras C_i and C_j. The observations corresponding to the same object across this camera pair can be used to estimate an inter-camera BTF. A way to compute this BTF is to estimate the pixel to pixel correspondence among the object appearance in the two cameras (see eq. 10). However, self occlusion, change of scale and shape, and object

Model	M	S
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 eq. 3. The subscripts i and r denote the incident and the reflected directions measured with respect to normal surface. 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 eq.3, we must assume that surface roughness σ is constant over the plane.

deformation, make finding correspondences among different camera pixels very hard. In order to mitigate these problems, we use normalized histograms of object brightness values for the BTF estimation. The histograms are relatively robust to the changes of the object pose (Swain & Ballard, 1990). Assuming that the percentage of image points on the observed object $O_{i,k}(a)$ with brightness less than or equivalent to B_i is equal to the percentage of image points in the observation $O_{j,l}(a)$ with brightness less than or equivalent to B_j. It should be noted that a similar strategy was adopted in another work to obtain a BTF between images acquired by the same camera in the same FOV but in different illumination conditions (Grossberg & Nayar, 2003). Let be H_i and H_j the normalized cumulative histograms of object oservations I_i and I_j respectively, then

$$H_i(B_i) = H_j(B_j) = H_j(f_{ij}(B_i)) \tag{11}$$

Consequently we obtain,

$$f_{ij}(B_i) = H_j^{-1}(H_i(B_i)) \tag{12}$$

where with H^{-1} we indicate the inverted cumulative histogram. As discussed in the previous subsection, the BTF between two cameras changes instant by instant due to illumination conditions, camera parameters and so on. We apply eq. 12 to estimate the brightness transfer function f_{ij} for every pair of the observations contained in the training set. Also we denote with F_{ij} the collection of all the brightness functions obtained in the described way: $F_{ij} = \{f_{ij}^n\}, n \in \{1, \ldots, N\}$. Note that the discussion has been done dealing with only the brightness values of the images and estimating the brightness transfer functions. To deal with color images we have to compute each channel separately. It should be noted also that the knowledge of any camera parameters and response function for the calculation of these transfer function is not assumed.

3.3 Object color similarity estimation across camera using BTF

It is natural that the observed color of an object can vary widely across multiple non-overlapping camera due to change in scene illumination or of some of the different camera parameters like focal length, CCD gain, and so on. The training phase provides us the set of color transfer functions among the cameras, which models how colors of an object change across cameras. If the mapping function between the colors of two observations is correctly learned, in the test phase it is likely to find the observations generated by the same object. In particular for two observations $O_{i,k}(a)$ and $O_{j,l}(a)$ with color transfer functions f_{ij}^R, f_{ij}^G and f_{ij}^B, we define the probability of the observations belonging to the same object as:

$$P_{ij}(O_{i,k}(a), O_{j,l}(a)|\lambda_{j,l}^{i,k}) = \prod_{ch\in\{R,G,B\}} \gamma e^{-\gamma d(f_{ij}^{ch}(O_{i,k}^{ch}), O_{j,l}^{ch})} \quad (13)$$

where γ is an arbitrary constant and d is a distance between an object appearance in C_j and the transformed one in C_i. The *ch* superscript denote the color channel for which appearance model and brightness transform were calculated.

4. Object appearance modelling

The proposed tracking algorithm models the object appearance using color histogram statistics. The task of finding the same object from the foreground region in current frame can be formulated as follows: the color histogram feature is assumed to have a density function and a candidate region also had a color histogram feature distributed by a certain density. The problem is to find a candidate region which is associated to a density most similar to the target one. A Bhattacharya coefficient measure is used as a similarity metric among the distributions.

4.1 Color histogram extraction

We have implemented and tested different methods to extract the color histogram from the foreground patches in order to remove noise and possible shadow from each object patch. We used various elliptic masks to reach this aim. The ellipse parameters (major and minor axis) are inferred using the patch dimensions and the distance of the object from the camera. Basing on the person position in each FOV, we have assessed, by a mapping function, his body measure in the foreground patches. Our intention is to build the elliptic masks in order to capture more useful information. We want to discard any possible part of the patch that could confuse the histogram distribution. The ellipses are drawn to cover most of the body cutting the head of the person and his eventual shadow (see 2(b), 3(b)). We have compared different combinations of these masks (see pictures 2(b), 2(c), 2(e), 2(f), 3(b), 3(c), 3(e), 3(f)) in order to estimate the potentialities of their performances.

Fig. 2. Six images of two persons in the camera C1. a) Foreground patch extracted of the first person; b) Elliptic mask of the first person; c) Double Elliptic masks of the first person; d) Foreground patch extracted of the second person; e) Elliptic mask of the second person; f) Double Elliptic masks of the second person.

(a) (b) (c) (d) (e) (f)

Fig. 3. Six images of two persons in the camera C2. a) Foreground patch extracted of the first person; b) Elliptic mask of the first person; c) Double Elliptic masks of the first person; d) Foreground patch extracted of the second person; e) Elliptic mask of the second person; f) Double Elliptic masks of the second person.

4.2 Positional histograms

As it should be noted, conventional color descriptors fail in the presence of clothes with the same main colors, but distributed in different way on the whole clothes. So we need to detect a features set able to maintain a level of relationship between global distribution and the displacement of colors on the silhouette. In the presence of well differentiated clothes, conventional histograms perform well, as well as other more refined features, like correlograms, even if this last one is onerous in terms of computational load. Our goal is to detect a feature set able to: perform in a acceptable way (compared with histograms) in presence of easily distinguishable uniforms; outperform histograms in the presence of hardly distinguishable uniforms; maintain a low level of computational load, that allows us to integrate this module in a higher level real team events detection system.

For these reasons we have chosen to work with a modified version of classic histograms, called Positional Histograms. These feature descriptors maintain basic characteristics of histograms (fast evaluation, scale invariance, rotation invariance, and so on); in addition, they introduce a dependance from the position of each point in the image: the global image is partitioned according to a geometrical relationship; the histograms are then evaluated for each region, and concatenated to obtain the final region descriptor.

Formally, the image I is partitioned in n subregions R_i, that satisfy the rules:

$$\bigcup_{i=1}^{n} R_i = I \tag{14}$$

$$R_i \cap R_j = \emptyset \quad \forall i \neq j \tag{15}$$

The first equation guarantees that each point of the image contributes to the final feature set construction, while the second one guarantees that each point gives its contribution just to one partition of histogram. In this way the final feature set contains exactly the same main information, as conventional histograms, but arranged in a different way, maintaining a level of information about the spatial distribution of points in the image.

The geometric rule for the partition should be fixed according to the nature of the problem to be solved. Our experience, and also experimental results we obtained, suggests us to use two main geometrical partitions: the angular sectors and the circular rings, and their fusion version (circular sectors). Polar coordinates allow to easily define the partitions. Each region R_i is composed by points (x,y) that satisfy:

$$R_i = \{(x,y) \| x = r\cos\theta,\ y = r\sin\theta \\ r^i_{MIN} < r < r^i_{MAX},\ \theta^i_{MIN} < \theta < \theta^i_{MAX}\} \quad (16)$$

With this notations, we can now explore details of each partition used in this paper. The starting point of each partition is the center of the image, where reasonably is concentrated the main informative content (a good object detector/tracker is able to maintain the subject in the center of the image).

4.2.1 Angular sectors

In this case each partition is obtained by varying the angle in a given range, according to the desired details level, while the radius ranges in all available values. So, considering D as the main diagonal of the image, and n the number of desired sectors, we have:

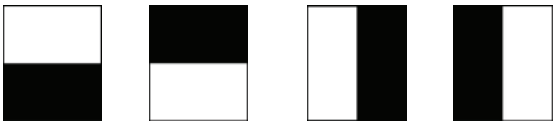
$$r^i_{MIN} = r_{MIN} = 0 \quad (17a)$$

$$r^i_{MAX} = r_{MAX} = D/2 \quad (17b)$$

$$\theta^i_{MIN} = \theta_0 + \frac{2\pi}{n}(i-1) \quad (17c)$$

$$\theta^i_{MAX} = \theta_0 + \frac{2\pi}{n} * i \quad (17d)$$

$$i = 1..n \quad (17e)$$



(a) $n = 2, \theta_0 = 0$ (b) $n = 2, \theta_0 = \pi/2$

(c) $n = 4, \theta_0 = 0$

(d) $n = 8, \theta_0 = 0$

Fig. 4. Plot of some Angular Sectors.

In figure 4 we have plotted some examples of masks for the regions creation in presence of Angular Sectors partitions. In the first rows we have plotted masks for $n = 2$, $\theta_0 = 0$, and $n = 2$, $\theta_0 = \pi/2$. Similarly, in the following rows we propose partitions for $n = 4$ and $\theta_0 = 0$, and $n = 8$ and $\theta_0 = 0$.

4.2.2 Circular rings

Each partition is obtained by varying the radius in a given range, according to the desired details level, while the angle varies in order to cover all possible values between 0 and 2π. So, considering D as the main diagonal of the image, and n the number of desired sectors, we have:

$$r^i_{MIN} = \frac{D * (i-1)}{2n} \tag{18a}$$

$$r^i_{MAX} = \frac{D * i}{2n} \tag{18b}$$

$$\theta^i_{MIN} = \theta_{MIN} = 0 \tag{18c}$$

$$\theta^i_{MAX} = \theta_{MAX} = 2\pi \tag{18d}$$

$$i = 1..n \tag{18e}$$

In figure 5 the masks in presence of Circular Rings partitions with $n = 2$ are plotted.

(a) $n = 2$

Fig. 5. Plot of some Circular Rings.

4.2.3 Circular sectors

The previously exposed partition rules can be combined (overlapped) in order to obtain another set of features that satisfies the conditions of equations 14 and 15. Now radius and angle various simultaneously tracing circular sectors across the image. So it is necessary to define two levels of partitions: the number n_S of desired angular sectors (that influences the range of the angle θ) and the number n_R of desired circular rings (that influences the range of the radius).

$$r^i_{MIN} = \frac{D * (i-1)}{2n} \tag{19a}$$

$$r^i_{MAX} = \frac{D * i}{2n} \tag{19b}$$

$$\theta^j_{MIN} = \theta_0 + \frac{2\pi}{n}(j-1) \tag{19c}$$

$$\theta^{j}_{MAX} = \theta_0 + \frac{2\pi}{n} * j \tag{19d}$$

$$i = 1..n_R \qquad j = 1..n_S \tag{19e}$$

In figure 6 some examples of masks in presence of Circular Sectors partitions are plotted: first row refers to rings obtained for $n_R = 2$, $n_S = 2$ while the second one refers to rings for $n_S = 4$, $n_R = 2$.

(a) $n_R = 2, n_S = 2$

(b) $n_S = 4, n_R = 2$

Fig. 6. Plot of some Circular Sectors.

5. Discussion on tracking in multiple cameras with disjoint views

Starting from two cameras C_1 and C_2 that are localized in different places in our building, we explore some methodology to establish correspondence among these FOVs. As you can see in Figure 1, the cameras' field of views cover different non-overlapping areas. People observed in camera C_2 can take a path across camera C_1 turning right or also turning left into *Lab1* without entering the C_1 field of view. Likewise, people coming from the *Lab1* are observed in camera C_1 without passing through camera C_2. As we described in the previous section the task of the multi-camera tracking algorithm is to establish correspondence across cameras finding which observations, coming from different FOVs, belong to the same object.Because of the cameras' positions, it is not always possible to use space-time constraints among the exits and entrances areas of the different cameras. Obviously people can take many paths across C_1 and C_2 producing different observations of the same objects in the two cameras. The multi-view tracking algorithm we want to investigate *relies just on the object appearances in the two cameras FOV.* It can be noted that the lack of entry/exit constraints make the association task more difficult. We should consider that color distribution of an object can be fairly different when it moves in a single camera FOV. For this reason, matching appearances between different cameras is still more difficult. It is necessary to find the transformation that maps the object's appearance in one camera with its appearance in the other one. In this chapter we consider a training phase in which know objects pass trough both cameras and their appearances are used to estimate a Brightness Transfer Function (BTF). During this phase we tested two different BTFs, ie. the mean BTF (MBTF) and the cumulative BTF (CBTF).In the succeeding phase we test implemented transformation in order to evaluate the object matches that produced the lowest value of the Bhattacharya distance for the appearance of the considered person in one camera compared with the appearances of all the possible persons who had walked through the second camera.

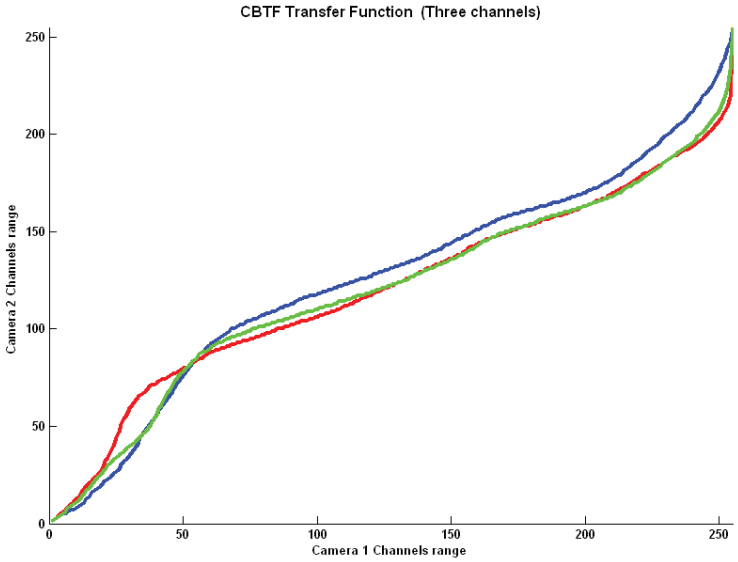


Fig. 7. The transfer functions for the R,G and B color channels from Camera 1 to Camera 2 obtained using Cumulative Brightness Transfer Transform on the objects appearances training set. Note that mostly lower color values from Camera 2 are being mapped to higher color values in Camera 1, indicating that the same object is appearing much brighter in Camera 1 as compared to Camera 2.

5.1 Establish correspondence across multiple cameras

The correspondence problem occurs when an object enters the camera FOV. We need to determine if the object is already being tracked by another camera or it is a new object in the scene. As described in previous sections there are many approaches which are able to estimate the BTFs among different cameras. We compare a mean BTF (see section 21) approach with the cumulative BTF proposed in (Prosser et al., 2008). In figure 2(a) some images of the the tracks of two people in the camera C_1 FOV are showed, while in figure 3(a) the same people are observed in in the camera C_2 FOV. We computed the three channels RGB histograms for each image in the C_1 tracks. We did the same, also, for the C_2 tracks. The histogram were generated using all the 256 bins for each color channel. We want to estimate a BTF f_{12} between the cameras C_1 and C_2 such that, for each couple of objects observations $O_{1,k}(a)$ and $O_{2,l}(a)$ given the brightness value $B_{O_{1,k}}(v)$ and $B_{O_{2,l}}(v)$ we have $B_{O_{2,l}}(v) = f_{12}(B_{O_{1,k}}(v))$ where $v = 0, \ldots, 255$ represents the number of bins, $k = 1, \ldots, M$ represents the number of object appearance in the camera C_1, $l = 1, \ldots, N$, the number of object appearance in the C_2. In order to evaluate the BTF f_{12} we collected a total of $N + M$ histograms obtained on the N object appearances tracked in the camera C_1 and on the same object appearances tracked in the camera C_2. Let be $H_{O_{1,k}}$ and $H_{O_{2,l}}$ the object appearance histograms obtained in the cameras C_1 and C_2 respectively. Now, for each possible different cameras object appearance couple $(O_{1,k}, O_{2,l})$ we want to compute the brightness transfer function using the inverted cumulative histogram (see equation 12) we obtain:

$$f_{O_{1,k}O_{2,l}}(B_{O_{1,k}}) = H^{-1}_{O_{2,l}}(H_{O_{1,k}}(B_{O_{1,k}})) \tag{20}$$

and finally the mean BTF (referred in the following sections as MBTF) $\overline{f}_{12}$

$$\overline{f}_{12} = \sum_{k=1}^{M}\sum_{l=1}^{N} f_{O_{1,k}O_{2,l}} \tag{21}$$

We estimated also a cumulative BTF (CBTF) as described in (Prosser et al., 2008). As it is known, the CBTF generation involves an amalgamation of the training set before computing any BTFs. An accumulation of the brightness values is computed an all the training images of the camera C_1 obtaining a cumulative histogram $\hat{H}_1$. The same is done for all the corresponding training images of the camera C_2 obtaining $\hat{H}_2$. The CBTF $\hat{f}_{12}$ using eq. 12 is

$$\hat{f}_{12}(O_{1,k}) = \hat{H}_2^{-1}(\hat{H}_1(O_{1,k})) \tag{22}$$

also in this case evaluated by using the inverted cumulative histogram. In figure 7 the CBTF obtained on the training set is plotted. It should be noted that mostly lower color values from Camera C_2 are being mapped to higher color values in Camera C_1, indicating that the same object is appearing much brighter in Camera C_1 as compared to Camera C_2.

6. Multi-camera tracking using different BTFs

In order to solve the multi-camera people re-identification problem we have to choose among a set of possible correspondence hypotheses the one that produces the best match. Since, our cameras' configuration allows people to enter into one camera field of view without passing across the other camera's FOV, we consider also the problem of finding a proper method to discard false matches. The common method is to match people appearances by estimating the similarity among color histograms. Once both $O_{i,k}$ and $O_{j,l}$ are converted into the same FOV by eq. 21 and eq. 22 we can compare them directly using the well known Bhattacharya distance measure $D_B()$ and thus the similarity measure from eq. 2 can be defined as follows:

$$Similarity(O_{i,k}, O_{j,l}) = 1 - D_B(O_{i,k}, O_{j,l}) \tag{23}$$

If we denote with H_i and H_j the normalized histograms of the observations $O_{i,k}$ and $O_{j,l}$ the Bhattacharya distance $D_B()$ (Comaniciu et. al, 2003) between two histograms is given as

$$D_B(H_i, H_j) = \sqrt{1 - \sum_{v=1}^{m}\sqrt{H_i(v)H_j(v)}} \tag{24}$$

where m is the total number of histogram bins. The Bhattacharya coefficient ranges between zero and one and is a metric.

Note that in order to compare two color objects, we must apply this process to each of the three RGB channels. Thus the overall similarity measure becomes the mean of the similarity values obtained in all three channels.

Let us, now, consider $\left\{ H_{O_{1,k_1}}, H_{O_{1,k_2}}, \ldots, H_{O_{1,k_{N_k}}} \right\}$ the N_k object appearances histograms of the $k-th$ person in the camera C_1. Suppose that we have P, i.e. $k \in \{1, \ldots, P\}$, people moving

in the camera C_1. When a new observation is taken in the camera C_2 we have to decide either if it could be associated with one among the P individuals moving in the camera C_1 or if it is a new person entering the scene. For each person k, in the camera C_1 with $k \in \{1, \ldots, P\}$, we evaluated the mean color histograms among the N_k appearance observations of the k^{th} person obtaining $\overline{H}_{1,k}$. Anyway the mean histograms cannot be compared with those obtained by the camera C_2 unless the transformation with the BTFs are applied. By using the BTFs described in section 5 we projected the P mean histograms in the new space as follows:

$$\check{H}_k = \overline{f}_{12}(\overline{H}_{1,k}) \tag{25}$$

where $\check{H}_k$ represents the new histogram obtained by using the mean BTF described in eq.21. We, also, have using eq. 22:

$$\tilde{H}_k = \hat{f}_{12}(\overline{H}_{1,k}) \tag{26}$$

which is the histogram transformation using CBTF. Let be $H_{O_{2,l_1}}$ the first observation histogram in the camera C_2. We evaluated the similarity between couple of histogram by using eq. 27. The association is done with the k^{th} person who produces the maximum similarity measure, i.e.

$$\arg\max_k Similarity(H_{O_{2,l_1}}, \tilde{H}_k) \tag{27}$$

(a) Camera 2 Field of view (b) Camera 1 Field of view

Fig. 8. Frames from both camera views. The same person walks from camera 1 to camera 2

7. Obtained results

Different experiments were carried out to test the multi-camera tracking algorithm. The scenario was composed of two cameras located in two different points of our office (see map on figure 1). We used two wireless Axis IP camera with $640x480$ VGA color jpg resolution with an acquisition frame rate of 10 fps. The camera network topology is shown in figure 1, while two images acquired by the two cameras are shown in figures 8(a) and 8(b). Note that the illumination conditions and color quality vary greatly between these views. We divided the experiments into two different parts: in the first part we investigate different kind of method to extract the color histogram from each foreground patch. In the second part we evaluated different approaches to establish correspondence across disjointed views. In both

parts we used the same data set. The data-set consisted of synchronized *mjpeg* videos acquired simultaneously by two different cameras containing eight persons. The patches come from a single camera object detection method described in (Mazzeo et. al, 2008).The data set was obtained by extracting people from whole Field of View of each camera. Note that we did not consider any geometrical constraint on the exiting and entering areas of people moving in the observed scenario. We carried out different experiments using different sets of samples as follows:

- First experiment ten people giving 1456 appearance sample (coming from the same FOV) are used as testing data in order to evaluate the performance of different color histogram extraction methods (See section 4).
- In second experiment we have a training and a testing data set: seven individuals giving 1023 appearance samples in both views were used in the training step, while eight individuals with 1247 appearance samples in both views were used in the testing step (Note that in this case we added one person in the testing phase).

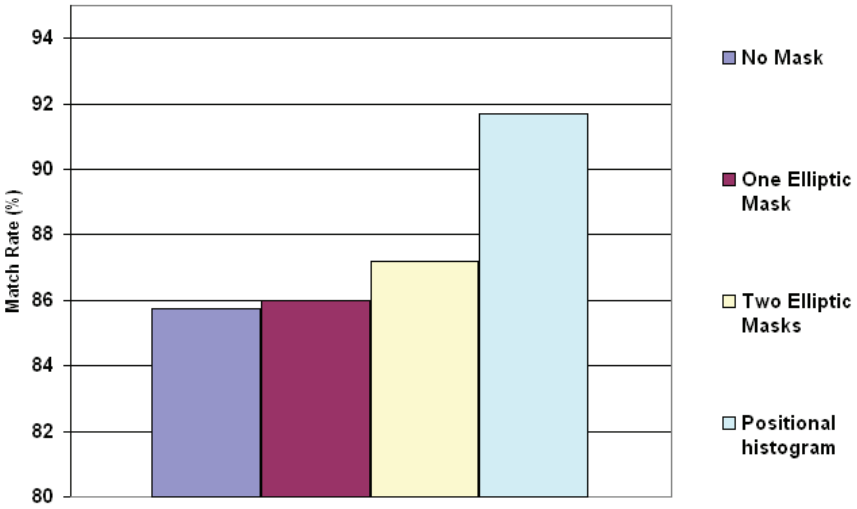


Fig. 9. A matching success comparison in the same FOV (Intra-camera) using different color histogram extraction method.

In the figure 9 are presented the results relative to the intra-camera histogram color tracking. As described in section 4 we evaluated four different approaches to estimate the color histogram from extracted people patches. The similarity between color histogram features belonging to different foreground patches, was measured by means of eq. 27 based on the Bhattacharyya distance (eq. 24). The highest value of eq. 27, among the designated patch and all the possible candidates (seven different people in the same FOV), determines the tracking association. As it is shown in figure 9 it is possible to notice how the positional histogram approach gives better result in term of match rate. By using positional histogram, in fact, it is possible to preserve the color histogram spatial information. In particular we used partition masks based on circular sector with $n_r = 2$ and $n_s = 4$ (see figure 6a). In this context this kind

of masks gave best performance in terms of correct matching rate. In this way, in fact, color histograms of different body parts of each patch are compared with the correspondent parts of the another persons (see subsection 4.2). Results confirm that this color histogram extraction approach discriminates better among the different possible candidates.

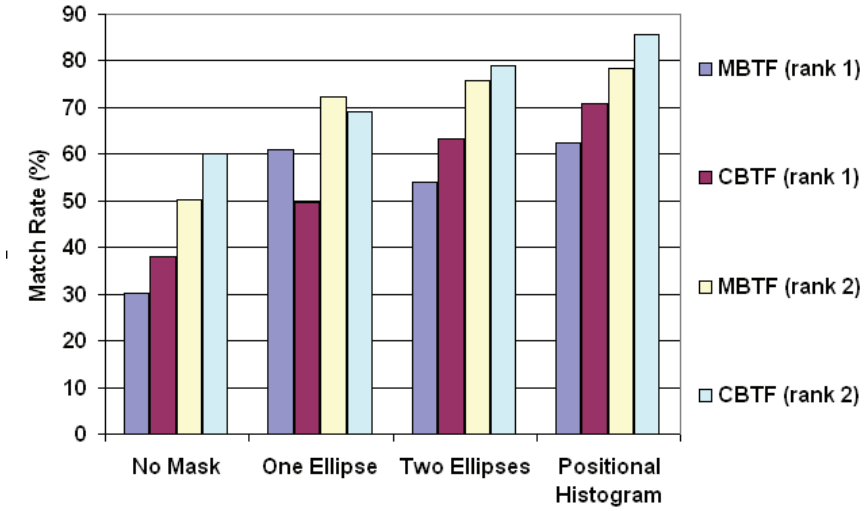


Fig. 10. A matching success comparison in establishing correspondence across cameras using varying color histogram extraction method and different Brightness Transfer Function Estimation.

In figure 10, the results relative to the tracking across different non overlapping cameras, are shown. The experiments consist of a training phase and a testing phase. During the training phase we supposed that the correspondence between the same object in the different cameras'FOV was known and this information was used to estimate the Mean Brightness Transfer Function (MBTF) and the Cumulative Brightness Transfer Function (CBTF). In the testing phase the correspondences between cameras were computed using eq. 27 based on Bhattacharyya distance, and the correct match was associated with the highest value of eq. 27 among all the possible couple of candidates. As it can be noticed, in the testing phase we consider seven people that were present in both views. As described in section 5 we tested two approaches to estimate BTFs among the two cameras: the MBTF and the CBTF. For each people patch we converted his RGB histogram (extracted in the different way described in section 4) into the target region color space (i.e. from camera $C1$ to camera $C2$). They were compared against all individuals observed in this region. In particular we estimated the mean color histogram for the same person in each view and we compared each converted histogram against it. In figure 10 we report both the *rank*1 and *rank*2 results indicating the correct match presence as the highest and the second highest similarity score respectively. As figure shows both transfer functions (MBTF and CBTF) gave quite similar behaviors in term of match rates but it should be noted that the CBTF outperform MBTF in the *rank*1 and *rank*2 results (we matched the histogram also against the mean histogram of the people contained in the training set). Note that, only in the case of the one elliptic mask approach MBTF gives better results than CBTF (the difference is very narrow). However the overall performances confirmed that

CBTF retained more color information than MBTF and produced a more accurate mapping function. The same figure 10 shows a comparison among different color histogram extraction approaches, the match rates confirmed that positional histogram gave the best results. And this is what we expect for the reason explained in the first part of this section. Finally, figure 11 shows the mean and standard deviation of the different color histogram extraction method applied on the patch set used in both part of the experiment. Even these values demonstrate that positional histogram gives the greatest mean score with the lowest standard deviation. It means that the positional histogram has the capability to catch more useful information from the people patches in order to establish correspondences among different same people appearances. Definitely positional histogram method maps the data among the people classes better than the others.

8. Final considerations and future work

This book chapter presented a dissertation on the feasibility of multi-camera tracking algorithms based on the appearance similarity. We considered only two non overlapping cameras located inside an office building. We investigated the reliability of appearance similarity methods to track people in the same FOV and among different FOVs. Then we evaluated different color histogram extraction methods, with different elliptic masks and positional histogram. The obtained results demonstrated that using positional histograms improved overall results in terms of matching rate. Also, we compared two different kinds of Brightness Transfer Function, i.e. the MBTF and the CBTF. The experiments demonstrated quite similar behaviors of the two transferring functions when the simple association problem has to be solved.

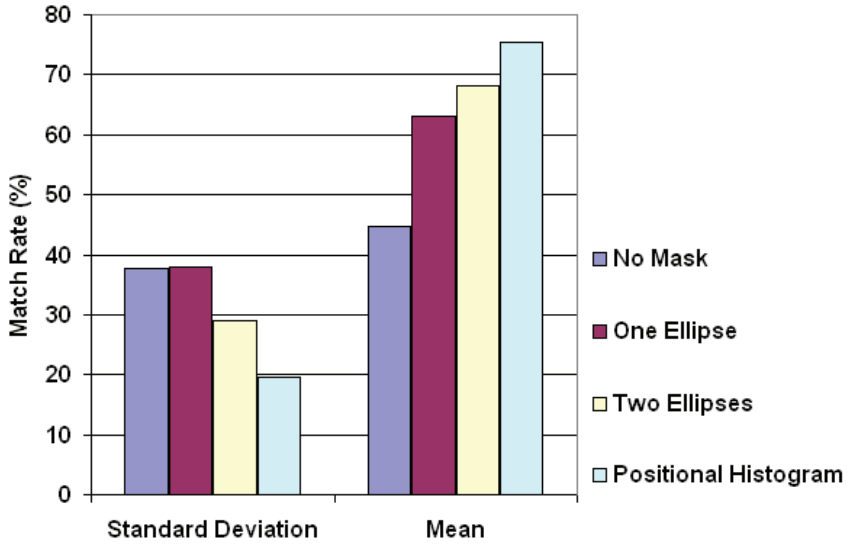


Fig. 11. Mean and standard deviation of the matching rate using different color extraction method to establish the correspondence between the two cameras.

Future work will be addressed on the study of new methodologies for more reliable appearance modelling. As we described in this chapter it is known that people's appearances can be similar in some parts of the body and differ in other parts. So we are thinking to apply some methodologies based on the extraction patch histogram graph. Then, we use different weights in the correspondence matches in order to consider different body parts reliability and highlight only the significant differences among the people appearances.

9. References

O. Javed, K.Safique,Z,Rasheed,M. Shah Modeling inter camera space-time and apperacnce relationships for tracking across non-overlapping views, *Computer Vision and Image Understanding*, Vol.109, 146-162, 2008

W. Du, J. Piater, Data Fusion by Belief Propagation for Multi-Camera Tracking, *The 9th International Conference on Information Fusion*, 2006.

W. Du, J. Piater, Multi-Camera People Tracking by Collaborative Particle Filters and Principal Axis-Based Integration *Asian Conference on Computer Vision*, Springer LNCS 4843 pp. 365-374, 2007.

A. Chilgunde, P. Kumar, S. Ranganath, H. WeiMin, Multi-Camera Target Tracking in Blind Regions of Cameras with Non-overlapping Fields of View, *British Machine Vision Conference BMVC*, Kingston, 7th-9th Sept, 2004.

Q. Cai, J.K. Aggarwal, Tracking human motion in structured environments using a distributed camera system, IEEE Trans. Pattern Anal. Mach. Intell. 2 (11)1241-1247, 1999.

F. Porikli, Inter-camera color calibration using cross-correlation model function, *IEEE Int. Conf. on Image Processing*, 2003.

Y. Shan, H.S. Sahwney, R. Kumar, Unsupervised learning of discriminative edge measures for vehicle matching between nonoverlapping cameras, *IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.

B. Prosser, S. Gong, T. Xiang, Multi-camera Matching using Bi-Directional Cumulative Brightness Transfer Functions, *British Machine Vision Conference*, 2008.

C. Madden, E.D. Cheng, M. Piccardi Tracking people across disjoint camera views by an illumination-tolerant appearance representation, *Machine Vision and Application*, 18, pp 233-247, 2007.

K. Jeong, C. Jaynes Object matching in disjoint cameras using a color transfer approach, *Machine Vision and Application*, 19 , pp 443-455, 2008.

M. Oren, S.K. Nayar Generalization of the Lambertian model and implications for machine vision, *International Journal of Computer Vision*, 14(3) , pp 227-251, 1995.

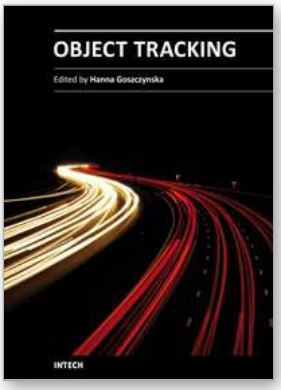
Horn, B.K.P. (1986). *Robot Vision*,MIT Press Cambridge.

Swain, M.J.; Ballard, D.H. (1990). Indexing via color histograms, in *IEEE International Conference on Computer Vision*.

Grossberg, M.D.; Nayar, S.K. (2003). Determing the camera response from images: what is knowable?, in *IEEE Transaction in Pattern Analysis and Machine Intelligence*, 25(11), pp. 1455-1467.

Comaniciu, D.; Ramesh, V.; Meer P. (2003). Kernel-based object tracking, in *IEEE Transaction in Pattern Analysis and Machine Intelligence*, 25, pp. 564-575.

Mazzeo, P.L.; Spagnolo, P.; Leo, M.; D'Orazio T (2008). Visual Player Detection and Tracking in Soccer Matches, in *IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance AVSS*, pp. 326-333.



Object Tracking
Edited by Dr. Hanna Goszczynska

ISBN 978-953-307-360-6
Hard cover, 284 pages
Publisher InTech
Published online 28, February, 2011
Published in print edition February, 2011

Object tracking consists in estimation of trajectory of moving objects in the sequence of images. Automation of the computer object tracking is a difficult task. Dynamics of multiple parameters changes representing features and motion of the objects, and temporary partial or full occlusion of the tracked objects have to be considered. This monograph presents the development of object tracking algorithms, methods and systems. Both, state of the art of object tracking methods and also the new trends in research are described in this book. Fourteen chapters are split into two sections. Section 1 presents new theoretical ideas whereas Section 2 presents real-life applications. Despite the variety of topics contained in this monograph it constitutes a consisted knowledge in the field of computer object tracking. The intention of editor was to follow up the very quick progress in the developing of methods as well as extension of the application.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Pier Luigi Mazzeo and Paolo Spagnolo (2011). Object Tracking in Multiple Cameras with Disjoint Views, Object Tracking, Dr. Hanna Goszczynska (Ed.), ISBN: 978-953-307-360-6, InTech, Available from: http://www.intechopen.com/books/object-tracking/object-tracking-in-multiple-cameras-with-disjoint-views

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821