Tracking Multiple People with Illumination Maps

Gloria Zen, Oswald Lanz, Stefano Messelodi, Elisa Ricci
Fondazione Bruno Kessler, FBK-irst, via Sommarive 18, Povo, 38100 Trento, Italy
{*gzen,lanz,messelodi,elericci*}*@fbk.eu*

Abstract

We address the problem of multiple people tracking under non-homogenous and time-varying illumination conditions. We propose a unified framework for jointly estimating the position of the targets and their illumination conditions. For each target multiple templates are considered to model appearance variations due to lighting changes. The template choice is driven by an illumination map which describes the light conditions in different areas of the scene. This map is computed with a novel algorithm for efficient inference in a hierarchical Markov Random Field (MRF) and is updated online to adapt to slow lighting changes. Experimental results demonstrate the effectiveness of our approach.

1 Introduction

Robust visual object tracking in an illumination-varying environment is a classical task in computer vision. Many approaches [1, 2, 6] have been proposed but the problem is still far from being solved. Some methods simply discard the illumination-sensitive information and employ other features considered invariant to illumination ($e.g.$ motion or edges) [2] or use color spaces different from RGB ($e.g.$ YUV or HSI) [6]. However by eliminating the color or the intensity component the ability of the tracker to discriminate the object w.r.t. the background is reduced.

A radically different strategy, which we also adopt in this paper, consists in explicitly modeling target appearance variations due to lighting conditions [1, 4]. We focus our attention on video surveillance single-camera applications and we propose to use particle filters (PFs) for jointly estimating the positions of the targets and their illumination conditions. However differently from [4] where the analysis is limited to a single target, we study the more challenging task of multi-person tracking and we adopt the Hybrid Joint Separable (HJS) PF [5] instead of a MCMC PF as in [1]. We model target appearance variations due to light changes by introducing multiple templates into the HJS algorithm. Roughly inspired by color constancy approaches [3, 7], we compute the templates by applying to a reference image a finite set of diagonal transforms. Interestingly in the PF the template sampling process is constrained by incorporating a prior which allows to choose only among suitable templates. The prior is obtained from the *illumination map* of the scene which we define as the set of beliefs of the nodes in a hierarchical MRF. A main novelty of the paper concerns the inference in the MRF: to perform fast computations we developed a novel message passing scheme extending the work in [8].

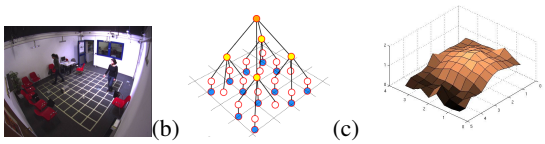

Figure 1. An example of (a) a possible scenario (b) a hierarchical MRF. (c) A plot illustrating the idea of illumination map.

2 Background

HJS PF. In PFs the goal is to approximate the posterior distribution $p(\mathbf{s}_t|o_{1:t})$ where $\mathbf{s}_t$ denotes the hidden state which represents the objects configuration and o_t the associated measurements extracted from the image. The approximation is made by a set of samples (the particles) $\{\mathbf{s}_t^i, w_t^i\}$ each one associated with a weight w_t^i which indicates its "quality". Let $\mathbf{s}_t = \{s^1, .., s^Q\}$ denotes the joint state and s^q the state of the q-th object. Among several methods for multi-person tracking, the HJS algorithm represents a theoretically founded while computationally efficient approach. The main idea of the HJS is to approximate the joint posterior by the outer product of its marginals, *i.e.* $p(\mathbf{s}_t|o_{1:t}) \approx \prod_q p(s_t^q|o_{1:t})$. Then the prediction step is performed independently for each target $p(\mathbf{s}_t|\mathbf{s}_{t-1}) = \prod_q p(s_t^q|s_{t-1}^q)$ (thus with linear cost in the number of targets) while in the update step a joint likelihood $p(o_t|\mathbf{s}_t)$ is considered explicitly taking into account occlusions. Due to lack of space we remind the reader to [5] for details.

Modeling the illumination changes. Light reflected by an object entering a camera is a product of the object reflectance and the illuminant spectrum. In a Lambertian model, the color $\boldsymbol{\rho} = (\rho_0, \rho_1, \rho_2)$ for a point reflectance $S(\lambda)$ with illumination spectrum $E(\lambda)$, illuminated from direction $\boldsymbol{a}$, is given by $\rho_c = \boldsymbol{n} \cdot \boldsymbol{a} \int E(\lambda)S(\lambda)Q_c(\lambda)d\lambda$ where λ is the light wavelength. In this paper we assume that $\boldsymbol{\rho} = \boldsymbol{\Sigma\kappa}$ with $[\boldsymbol{\Sigma}]_{ck} = \boldsymbol{n} \cdot \boldsymbol{a} \int E(\lambda)S_k(\lambda)Q_c(\lambda)d\lambda$, $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \kappa_2)$ $i.e.$ that the reflectance can be expressed by a linear combination of three basis functions $S_k(\lambda)$ [7]. Therefore the function which transforms an image $\boldsymbol{\rho}$ taken under an illuminant with spectrum $E(\lambda)$ to an image $\boldsymbol{\rho}'$ taken under a different illuminant with spectrum $E'(\lambda)$ is $\boldsymbol{\rho}' = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma\rho}$. We also suppose narrowband camera sensors to be approximately spike sensitivities $Q_c(\lambda) = \delta(\lambda - \lambda_c), c = 0, 1, 2$. The results is that $\boldsymbol{\rho}' = \mathbf{D}_{\boldsymbol{\beta}}\boldsymbol{\rho}$ where $\mathbf{D}_{\boldsymbol{\beta}} = diag(\boldsymbol{\beta})$ and $\boldsymbol{\beta} = (\beta^0, \beta^1, \beta^2)$ with $\beta^c = E(\lambda_c)/E'(\lambda_c)$. Thus color change under changing illumination can be described by the so-called diagonal (or von Kries) model [3] according to which each color channel is multiplied by a single factor β^c.

3 Tracking with Illumination Templates

This Section describes the main features of our PF.

Target representation. We modify the HJS algorithm introducing for each target multiple templates [9] representing its appearance under different illumination. To construct the templates set $\mathcal{R}^q = \{r_1^q, \dots r_{\mathcal{L}}^q\}$ we consider an RGB image $\mathbf{P}_1 \in \mathbb{R}^{3\times N}$ extracted from the silhouette of the target q at a specific position $\boldsymbol{x}^q$ in the scene and we compute the associated histogram r_1^q. The other exemplars are obtained applying $\mathcal{L}$ diagonal transforms to $\mathbf{P}_1$ according to von Kries hypothesis, *i.e.* $\mathbf{P}_\beta = \mathbf{D}_\beta\mathbf{P}_1$, with $\boldsymbol{\beta} = (\beta^r, \beta^g, \beta^b)$. From $\mathbf{P}_\beta$ we compute the histograms r_β^q. In practice for each β^c we consider ℓ possible values $i.e.$ $\beta^c \in \{\bar{\beta}_1^c, \dots \bar{\beta}_\ell^c\}$, $\bar{\beta}_i^c \in \mathbb{R}$ for a total of $\mathcal{L} = \ell^3$ templates.

State space. We model a person appearance dividing the body in three parts: head, torso and legs. Each target q is described by the state vector $s^q = (\boldsymbol{x}^q, \boldsymbol{\omega}^q, \boldsymbol{\beta}^q)$ where $\boldsymbol{x}^q$ indicates its position on the floor, $\boldsymbol{\omega}^q$ the shape of the body and $\boldsymbol{\beta}^q = (\beta^{r,q}, \beta^{g,q}, \beta^{b,q})$ describes its illumination conditions.

Observation model. For each body part we extract a $3D$ color histogram: this guarantees a more robust representation w.r.t. three separate histograms. The likelihood $p(o_t|\mathbf{s}_t)$ is computed modeling occlusions as in [5] and evaluating the Bhattacharya distance between the observed histograms and the current references $r_{\boldsymbol{\beta}_t}^q$.

Dynamical model. We define $p(s_t^q|s_{t-1}^q) = \Psi_I(\boldsymbol{x}_t^q)$ $p(\boldsymbol{x}_t^q|\boldsymbol{x}_{t-1}^q)p(\boldsymbol{\omega}_t^q|\boldsymbol{\omega}_{t-1}^q)p(\boldsymbol{\beta}_t^q|\boldsymbol{\beta}_{t-1}^q, \boldsymbol{x}_t^q)$. The MRF prior $\Psi_I(\boldsymbol{x}_t^q) = \prod_{i\neq q}\psi_I(\boldsymbol{x}_t^i, \boldsymbol{x}_t^q)$ penalizes object overlapping with $\psi_I(\boldsymbol{x}_t^i, \boldsymbol{x}_t^q) = e^{-\lambda_I(I_{MAX} - \|\boldsymbol{x}_t^i - \boldsymbol{x}_t^q\|)}$. We define $p(\boldsymbol{\beta}_t^q|\boldsymbol{\beta}_{t-1}^q, \boldsymbol{x}_t^q) = p(\boldsymbol{\beta}_t^q|\boldsymbol{\beta}_{t-1}^q)p(\boldsymbol{\beta}_t^q|\boldsymbol{x}_t^q)$ where $p(\boldsymbol{\beta}_t^q|\boldsymbol{x}_t^q)$ models the likelihood of certain light conditions given a location in the scene. As stated above we assume each color channel to be independent from the others *i.e.* $p(\boldsymbol{\beta}_t^q|\boldsymbol{\beta}_{t-1}^q) = \prod_{c\in\{r,g,b\}} p(\beta_t^{c,q}|\beta_{t-1}^{c,q})$ and $p(\boldsymbol{\beta}_t^q|\boldsymbol{x}_t^q) = \prod_{c\in\{r,g,b\}} p(\beta_t^{c,q}|\boldsymbol{x}_t^q)$. The terms $p(\boldsymbol{x}_t^q|\boldsymbol{x}_{t-1}^q)$, $p(\boldsymbol{\omega}_t^q|\boldsymbol{\omega}_{t-1}^q)$ and $p(\beta_t^{c,q}|\beta_{t-1}^{c,q})$ are modeled as gaussian noise, while $p(\beta_t^{c,q}|\boldsymbol{x}_t^q)$ is defined in the following section. The term $p(\boldsymbol{\beta}_t^q|\boldsymbol{x}_t^q)$ does not depend on the previous target state s_{t-1}^q and we treat it as an additional function in the importance weight ($i.e.$ as multiplicative factors of the likelihood).

4 Creating an Illumination Map

Motivation. We assume that under von Kries hypothesis and on a sufficiently small time interval Δ_t the illumination conditions of two targets in the same location of the scene can be described by the same $\boldsymbol{\beta}$. This assumption is reasonable in our video surveillance applications where targets of small-medium resolution can be sufficiently described by a global model of the illuminant and the reflectance properties of the materials ($i.e.$ the clothes of people) are about the same. Under this premise, the idea is to use the information collected from a set of targets to build an illumination map which aids the PF to track other targets. To this aim we construct a MRF where each observed node contains the information about the illumination condition of a specific region in the scene and we define the PF priors $p(\beta_t^{c,q}|\boldsymbol{x}_t^q)$ from the beliefs of the MRF. We call the set of beliefs the illumination map. In the following we describe the MRF we used and the inference approach we developed to compute its beliefs. To our knowledge the concept of a global illumination map has never been investigated before for visual object tracking.

The hierarchical MRF. We consider a grid discretizing all possible locations in the scene of interest (Fig.1.a) and we construct a hierarchical MRF (Fig.1.b) where each hidden node of level 0 corresponds to a cell in the grid. Each hidden node of level l is connected with P nodes at level $l-1$ (we choose $P = 4$). Latent variables are represented by the illumination coefficients $\boldsymbol{\beta}$. Hidden nodes of level 0 may or not be connected with an observation node. For an observation node i and a target u we collect several color histograms $z_{i,u}$ extracted from the target silhouettes associated to positions $\boldsymbol{x}^u$ inside the i-th cell. Histograms are first collected offline assuming that in this preliminary phase lighting conditions are static. It is worth noting that the targets used in this phase are not necessarily the ones we want to track. To adapt to time-varying illumination we take at each frame the color histograms correspond-

ing to the MMSE tracking estimates and progressively discard older histograms. We also discard histograms if an occlusion has been detected. In practice for each node i we build a buffer of histograms $z_{i,u}$ collected in the temporal interval Δ_t. Then for each target u and each node i we used K-means to compute the cluster centroids $h^k_{i,u}$ and define the set $\{h^1_{i,u} \dots h^{K_u}_{i,u}\}$ as the observation η_i for node i.

Let $\mathcal{G} = (V, E)$ be an undirected graph with a node set V and an edge set E. Let $\mathcal{L}$ be the cardinality of the label set and $\mathcal{H} = \{\eta_1, .., \eta_n\}$. The joint probability function over the entire graph is: $p(\boldsymbol{\beta}_1, .., \boldsymbol{\beta}_n|\mathcal{H}) = \frac{1}{Z}\prod_i \phi_i(\boldsymbol{\beta}_i, \eta_i)\prod_{i,j}\psi_{ij}(\boldsymbol{\beta}_i, \boldsymbol{\beta}_j)$ where $\phi_i(\boldsymbol{\beta}_i, \eta_i)$ is the likelihood for node i, $\psi_{ij}(\boldsymbol{\beta}_i, \boldsymbol{\beta}_j)$ is the pairwise potential between nodes i and j and Z is a normalization constant. We define $\phi_i(\boldsymbol{\beta}_i, \eta_i) = e^{-\lambda_B \sum_{u=1}^{U}\sum_{k=1}^{K_u}\pi^k_{i,u} D_B(h^k_{i,u}, r^u_{\boldsymbol{\beta}_i})}$ and $\psi_{ij}(\boldsymbol{\beta}_i, \boldsymbol{\beta}_j) = e^{-\lambda_E D_E(\boldsymbol{\beta}_i, \boldsymbol{\beta}_j)} e^{-\lambda_n(l) D_E(n_i, n_j)}$ where D_B and D_E denotes the Bhattacharya and the Euclidean distance respectively, n_i indicates the coordinates of node i and $\pi^k_{i,u} = T^k_{i,u}/T_{i,u}$ where $T^k_{i,u}$ is the cardinality of the set of histograms represented by $h^k_{i,u}$ while $T_{i,u}$ is the total number of histograms for the u-th target in the cell i. Pairwise potentials $\psi_{ij}(\boldsymbol{\beta}_i, \boldsymbol{\beta}_j)$ enforce the fact that neighboring nodes should have similar latent variables and, together with K-means, alleviate the effect of noise due to the use of tracking estimates in the observations η_i. The parameter $\lambda_n(l)$ depends on the level l in the hierarchy: it is set to a higher value for potentials connecting leaf nodes with level 1 and decreases going up to the hierarchy. Note that instead of a hierarchical structure we could have adopted a grid graph. However we found the hierarchy sufficiently accurate for our purposes with the advantage of allowing a faster inference. Moreover it naturally deals with nodes with missing observations interpolating between the observed data.

A semi-joint approach for efficient inference. We use sum-product belief propagation (BP) to compute the beliefs of the MRF. Since the set $\mathcal{H}$ is updated online as new tracking estimates are available the beliefs must be efficiently recomputed. While updating the observation potentials is relatively simple (it usually implies running K-means in few nodes and with few data), the computational cost of BP ($\mathcal{O}(|V|\mathcal{L}^2) = \mathcal{O}(|V|\ell^6)$) is prohibitive in our case since we observed that we typically need large graphs ($|V| \approx 100$) and $\ell > 10$. To overcome this difficulty we propose to modify the structure of our graph. We decompose each node i into three nodes r_i, g_i, b_i and modify $\psi_{ij}(\boldsymbol{\beta}_i, \boldsymbol{\beta}_j)$ defining three separate pairwise potentials $\psi^r_{i,j}(\beta^r_i, \beta^r_j)$, $\psi^g_{i,j}(\beta^g_i, \beta^g_j)$, $\psi^b_{i,j}(\beta^b_i, \beta^b_j)$ with $\psi^r_{i,j}(\beta^r_i, \beta^r_j) = e^{-\lambda_\beta D_E(\beta^r_i, \beta^r_j)} e^{-\lambda_n(l) D_E(n_i, n_j)}$. The likelihood is defined as above ($\phi_i(\beta^r_i, \beta^g_i, \beta^b_i) = \phi_i(\boldsymbol{\beta}_i, \eta_i)$) but in the decomposed graph is a clique potential of size 3. The new MRF is a graph with an increased number of nodes but where inference is done more efficiently ($\mathcal{O}(3|V|\ell^3)$). In fact in this *semi-joint* model the message passing consists in inter-node pairwise messages and intra-node triwise messages[1]:

$$
\begin{aligned}
& m^r_{\Psi_{ij}\to j}(\beta^r_j)\alpha \sum_{\beta^r_i} \psi^r_{ij}(\beta^r_i, \beta^r_j) m^r_{i\to\Psi_{ij}}(\beta^r_i) \\
& m^i_{\Phi\to r}(\beta^r_i)\alpha \sum_{\beta^g_i}\sum_{\beta^b_i} \phi_i(\beta^r_i, \beta^g_i, \beta^b_i) \\
& \qquad m^i_{g\to\Phi}(\beta^g_i) m^i_{b\to\Phi}(\beta^b_i)
\end{aligned}
\tag{1}
$$

This model extends the one in [8] where each node in the original graph is decomposed in two nodes. Our decomposition maintains the theoretical properties of the one in [8] (the solution space of the decomposed graph contains that of the original graph) but introduces a main difficulty: the triwise potentials imply computing Eqn.1 that is still demanding ($\mathcal{O}(\ell^3)$). To effectively address this problem we introduce the following result:

Theorem 1 *Let $\phi(\beta^r, \beta^g, \beta^b)$ be a real-valued function having continuous partial derivatives and let R (the remainder term of its first-order Taylor series expansion) be ≈ 0, then Eqn.1 can be computed as:*

$$
m^i_{\Phi\to r}(\beta^r_i) \approx \phi_i(\beta^r_i, \tilde{\beta}^g_i, \tilde{\beta}^b_i) \tag{2}
$$

where $\tilde{\beta}^c_i = \sum_{\beta^c_i} \beta^c_i m^i_{c\to\Phi}(\beta^c_i)$ and $m^i_{c\to\Phi}(\beta^c_i) = \prod_{k\in\mathcal{N}(i)} m^c_{\Psi_{ki}\to i}(\beta^c_i)$.

Proof. *See Supplementary material.*[2]

Using Eqn.2 instead of Eqn.1 the computational cost of message passing is reduced from $\mathcal{O}(3|V|\ell^3)$ to $\mathcal{O}(3|V|\ell^2)$. Regarding the accuracy of the approximation, we must say that despite it is difficult to assess it *a-priori* since it depends on the nature of $\phi(\beta^r, \beta^g, \beta^b)$, our experiments (Table 1) demonstrate its validity.

Table 1. Tracking accuracy (F-measure).

	s1	s2	s3	s4	s5	s6
$n^o targets$	1	1	2	2	3	3
HJS [5]	0.35	0.58	0.62	0.68	0.55	0.36
β-HJS Eqn.1	0.60	0.67	0.70	0.73	0.74	0.76
β-HJS Eqn.2	0.58	0.67	0.69	0.72	0.72	0.77

The illumination map. We used the computed beliefs $b_i(\beta^c_i) = m^i_{\Phi\to c}(\beta^c_i)\prod_{k\in\mathcal{N}(i)} m^c_{\Psi_{ki}\to i}(\beta^c_i)$ to define $p(\beta^{c,q}_t|\boldsymbol{x}^q_t) = \prod_{i\in\mathcal{B}(\boldsymbol{x}^q_t)} b_i(\beta^c_i)$ with $\mathcal{B}(\boldsymbol{x}^q_t) = \{i \in V, i : l(i) = 0, \|n_i - \boldsymbol{x}^q_t\|^2 \leq T\}$ and T a user defined threshold. They represent the *illumination map* of

[1] We reported the expressions of the messages from factor to variable nodes for the equivalent factor graph representation and for the red channel. Similar results apply to the other channels. Φ and Ψ indicate the factor nodes.

[2] http://tev.fbk.eu/people/zen/illumination_map.html

Figure 2. Tracking results without (top) and with the illumination map (bottom).

the scene since they contain information about the most likely β_i^c in the i-th cell of the grid. This concept is exemplified in Fig.1.c where the expected values of $b_i(\beta_i^r)$ corresponding to Fig.1.a are plotted. Fig.1.c clearly shows that illuminated regions correspond to higher values of β^r w.r.t. dark areas.

5 Results and Discussion

We first evaluate the performance of our algorithm on video sequences recorded in our laboratory with non-homogeneous illumination conditions. We build up an illumination map with a set of histograms collected offline in a fully automatic way: we track separately 3 targets freely moving around the room, never occluded by other objects and dressed with colors well discriminative w.r.t. the background. We used the learned map in more challenging situations (6 sequences with multiple targets or targets similar to the background). Despite the illumination conditions are static we also update the map running BP every 50 frames: this allows the tracking to better adapt to the current targets appearance. In Table 1 we compare the performance of: HJS without the illumination model, HJS with the illumination map and inference computed with Eqn.1 and HJS with the map and using Eqn.2. We measure tracking accuracy in terms of average F-measure $F = (2PR)/(P+R)$ where $P = (TS \cap GS)/TS$ is the precision and $R = (TS \cap GS)/GS$ the recall (GS and TS denote the ground truth and the tracking estimate windows). The performance of the PF improve significantly with the proposed illumination model since the targets are tracked even when subject to strong appearance changes. Moreover the results obtained with Eqn.2 are almost as accurate than the ones of Eqn.1 at a much reduced computational cost. For example in our experiments with $|V| = 96$ and $\ell = 26$ Theorem 1 leads to a speed up of about a factor of 1.5. This implies that each iteration of BP is performed in about 100 msecs rather than in 140 msecs.

Fig.2 clearly demonstrates the importance of the map. We consider a sequence with three targets. When

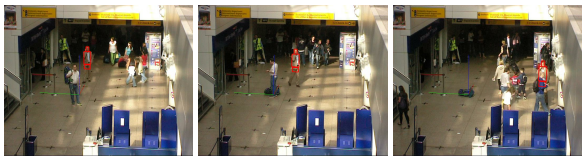

Figure 3. A sequence from PETS 2007.

the HJS is used without an illumination model two of the three targets are lost since their appearances change substantially (see supplementary material[2]). On the other hand, the adoption of the illumination templates without a learned map (*i.e.* $p(\beta_t^{c,q}|\boldsymbol{x}_t^q)$ are uniform distributions) causes the tracker drift (Fig.2(top)). The best results are obtained with the illumination map (Fig.2(bottom)), where all targets are successfully tracked during the entire sequence.

A second series of experiments have been conducted on the public dataset PETS 2007[3]. We selected some sequences with non-homogeneous and time-varying illumination conditions. Fig.3 depicts a challenging scenario with a strong light edge and a target whose color appearance is very similar to the background. In this case it is possible to reduce the risk of drifting by exploiting the information collected by other targets which are tracked correctly. For example in this case we compute the illumination map from the observations collected from one target (the man in white) walking in the room. This map is used successively to track more challenging targets such as the one in Fig. 3 which was not possible to track without.

References

[1] F. Bardet, T. Chateau, and D. Ramadasan. Illumination aware MCMC particle filter for long-term outdoor multi-object tracking and classification. *ICCV*, 2009.

[2] M. Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *IEEE Trans. PAMI*, 22(4):332–336, 2000.

[3] G. Finlayson, M. Drew, and B. Funt. Diagonal transforms suffice for color constancy. *ICCV*, 1993.

[4] A. Kale and C. Jaynes. A joint illumination and shape model for visual tracking. *Proc. CVPR*, 2006.

[5] O. Lanz. Approximate bayesian multibody tracking. *IEEE Trans. PAMI*, 28(9):1436–1449, 2006.

[6] Y. Lee, B. You, and S. Lee. A real-time color-based object tracking robust to irregular illumination variations. *ICRA*, 2001.

[7] D. Marimont and B. Wandell. Linear models of surface and illuminant spectra. *J. Opt. Soc. Am.*, 9, 1992.

[8] A. Shekhovtsov, I. Kovtun, and V. Hlavac. Efficient MRF deformation model for non-rigid image matching. *Proc. CVPR*, 2007.

[9] K. Toyama and A. Blake. Probabilistic tracking with exemplars in a metric space. *IJCV*, 48(1):9–19, 2002.

[3] http://pets2007.net/