Multitarget region tracking based on short-sight modeling of background and color distribution temporal variation

Julien Mille
http://liris.cnrs.fr/~jmille
Université de Lyon, CNRS
Université Lyon 1, LIRIS, UMR5205
F-69622, France

Jean-Loïc Rose
http://liris.cnrs.fr/~jlrose
Université de Lyon, CNRS
Université Lyon 2, LIRIS, UMR5202,
F-69676, France

We address the problem of joint segmentation and tracking of multiple objects using an energy-minimization based approach using color histograms. As in a few other existing approaches, a single color probability distribution per object and background is handled. In this context, global histograms may be problematic for tracking in real scenes with cluttered backgrounds, where statistical color data is highly scattered, preventing the estimation of reliable color statistics for object/background discrimination. To overcome this limitation, we introduce a *short-sight perception* modeling of background, which concentrates on the vicinity of tracked objects and thus extract more consistent statistical data for accurate separation between objects and background. To account for temporal consistency, our energy is also endowed with a novel data term explicitly based on temporal variation of color distribution within objects and local background regions.

Given current image frame I_t and associated partition P_t into $n+1$ regions, (background Ω_t^0 and n objects $\{\Omega_t^1,...,\Omega_t^n\}$), and next image frame I_{t+1}, we aim at determining the partition P_{t+1} in next frame **maximizing the** *a posteriori* **probability**, assuming conditional independence between image pixels:

$$\begin{aligned} P_{t+1}^* &= \underset{P_{t+1}}{\operatorname{argmax}}\, \mathrm{p}(P_{t+1}|I_t,I_{t+1},P_t) \\ &= \underset{P_{t+1}}{\operatorname{argmax}} \prod_{\mathbf{x}\in D} \underbrace{\mathrm{p}(I_{t+1}(\mathbf{x})|I_t,P_t,P_{t+1})}_{\text{Color likelihood}} \underbrace{\mathrm{p}(P_{t+1}|I_t,P_t)}_{\text{Prior}} \end{aligned} \quad (1)$$

We assume that likelihood $\ell_{t+1}^i(\mathbf{x})$ of observing $I_{t+1}(\mathbf{x})$ depends only on I_t, P_t and the region which $\mathbf{x}$ will belong to at time $t+1$. Taking the negative logarithm of *a posteriori* probability gives the **energy to be minimized**:

$$E[P_{t+1}] = \sum_{i=0}^{n} -\left\{\int_{\Omega_{t+1}^i} \log \ell_{t+1}^i(\mathbf{x})\mathrm{d}\mathbf{x}\right\} - \log \mathrm{p}(P_{t+1}|I_t,P_t) \quad (2)$$

The choice of likelihood functions depends on the assumptions made about temporal consistency of color, whereas prior probability depends on constraints on shape and motion of the tracked objects. For likelihood functions, we rely on **Non-parametric** kernel-based global estimation of color PDFs, which yields **one distribution per region**:

$$\ell_{t+1}^i(\mathbf{x}) = \frac{1}{\left|\Omega_t^i\right|} h_t^i(I_{t+1}(\mathbf{x})) \quad (3)$$

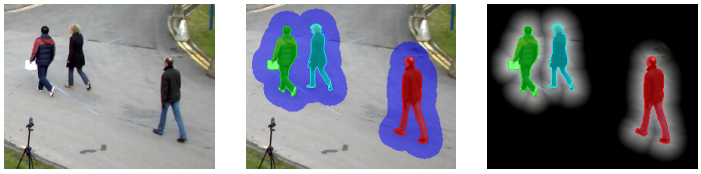
where $h_t^i(\alpha)$ is the **Gaussian kernel-based histogram** of color α. Pixels at time $t+1$ will tend to be included into the best matching region, regarding statistics at time t. As regards the prior term, no prior knowledge regarding shape or motion, hence the length of object boundaries is a natural smoothness term:

$$-\log \mathrm{p}(P_{t+1}|I_t,P_t) = \omega \sum_{i=1}^{n} \left|\partial\Omega_{t+1}^i\right|$$

One of the shortcomings inherent to tracking approaches based on global color PDFs is that color statistics may be unconfident and/or scattered in cluttered backgrounds, containing various objects and static parts with different appearances. If background likelihood gets small for nearly all colors, regions are more likely to include background pixels and *leak* outside actual objects. To overcome this limitation, we head towards a background model based on "short-sight perception", which is related to local modeling approaches [1, 2] and the idea of spatial context brought up in [3]. We consider a **band domain** B^i of width w around each object Ω^i and state that the contribution ψ of background pixel $\mathbf{x}$ to perception of object Ω^i decreases with respect to euclidean distance:

$$\psi_t^i(\mathbf{x}) = 1 - \frac{1}{w} \min_{\mathbf{y}\in\Omega_t^i} \|\mathbf{x}-\mathbf{y}\|$$

Consequently, each object has its **own local perception** of surrounding background. The far background (pixels which do not belong to any band) is intentionnally ignored **concentrate on background pixels close to objects**



Short-sight perceptions of background: original image (left), segmented targets with surrounding bands (center) and background faded to black with respect to its contribution to perceptions (right)

The spatial fuzziness of outer band histogram k and resulting likelihood function q makes the contributions of pixels be weighted with respect to their distance to the target object:

$$k_t^i(\alpha) = \int_{B_t^i} \psi_t^i(\mathbf{y}) K_\sigma(I_t(\mathbf{y})-\alpha) \quad \text{and} \quad \mathrm{q}_{t+1}^i(\mathbf{x}) = \frac{1}{\displaystyle\int_{B_t^i} \psi_t^i(\mathbf{y})\mathrm{d}\mathbf{y}} k_t^i(I_{t+1}(\mathbf{x}))\mathrm{d}\mathbf{x}$$

where α is any considered color. We replace global background likelihood with **local weighted likelihoods over bands**, which gives a relaxed version of minimization problem (2):

$$E_{\mathrm{SS}}[P_{t+1}] = \sum_{i=1}^{n} \left\{ -\int_{\Omega_{t+1}^i} \log \ell_{t+1}^i(\mathbf{x})\mathrm{d}\mathbf{x} - \int_{B_{t+1}^i} \psi_{t+1}^i(\mathbf{x}) \log \mathrm{q}_{t+1}^i(\mathbf{x})\mathrm{d}\mathbf{x} + ... \right.$$

Favouring close pixels prevents the discrepancy from being affected by changes on far background pixels. In case of moving background, gradual changes in local background representations are allowed. The second issue we deal with is that segmentation may be undesirably affected by overlap between color distributions of various neighboring object. To overcome this issue, we propose to explicitly penalize excessive variation of color distributions between successive frames, by introducing **histogram distance** J. The previous energy is endowed with a term favoring small variation of region and band histograms between successive frames:

$$J\left(h_{t+1}^i, h_t^i\right) + J\left(k_{t+1}^i, k_t^i\right) \quad (4)$$

In our experiments, J is chosen as the **symmetrized Kullback-Leibler divergence**. The partition-dependent optimization problem is turned into a textcoloremph-color**discrete labeling problem**, similar to Markov random fields. We introduce a labeling function $\phi : D \to \{0,...,n\}$ and reformulate the energy as a functional of ϕ. Starting from an initial state ϕ^0 taken as the labeling of partition P_{t+1}, labels of pixels located on the interface between regions are switched in order to decrease the energy, leading to a local minimum. Experiments are carried out on synthetic and real videos, such as as pedestrian sequence taken from the PETS2009 database:

Multitarget tracking on a PETS2009 pedestrian sequence

[1] T. Brox and D. Cremers. On local region models and a statistical interpretation of the piecewise smooth Mumford-Shah functional. *International Journal of Computer Vision*, 84(2):184–193, 2009.

[2] S. Lankton and A. Tannenbaum. Localizing region-based active contours. *IEEE Transactions on Image Processing*, 17(11):2029–2039, 2008.

[3] H.T. Nguyen, Q. Ji, and A. Smeulders. Spatio-temporal context for robust multitarget tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):52–64, 2007.