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

Abstract

We address the problem of multitarget region tracking within image sequences. Following recent work on joint segmentation and tracking as well as non-parametric modeling of color statistics, we develop 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.

1 Introduction

A large variety of optimization methods have been proposed and applied to joint object segmentation and tracking in videos. In this context, a partition of each frame into several objects and background is usually sought as the minimizer of an energy functional, enforcing consistency of color statistics and shapes of objects over time and space [6, 14]. Most often, the energy is derived from a Bayesian probabilistic model [1, 9], starting with the maximization of the *a posteriori* probability of current partition given current observation as well as past partitions and images. As is, the *a posteriori* probability is an intractable expression in general, which raises the need to use mathematical transformations based on several simplifying assumptions. The energy derived from such Bayesian framework is generally made up of a data term resulting from the likelihood of image data given the partition and a prior term which embodies soft constraints over object shape and/or motion.

In this work, we are specifically concerned with the energy data term. Shape and motion priors are beyond the scope of this paper, yet they could be added in order to specialize the method towards particular applications requiring more constrained tracking. We aim

BMVC 2011 http://dx.doi.org/10.5244/C.25.44

at tracking objects undergoing simultaneous translation and arbitrary smooth non-rigid deformations. We focus on data terms representing color likelihood of pixels with a global region-based non-parametric model, like in [3]. According to this model, color statistics in a given region, whether this region corresponds to an object or background, are described with a single probability density function (PDF) derived from a kernel-based color histogram. This kind of approach was already addressed for segmentation in single images [7, 10, 13] or videos [3, 5].

Global kernel-based probability estimation raises an issue for tracking in real scenes, especially for the background region, which may be cluttered and contain many non-tracked objects. In such case, statistical color data is highly scattered, so that background distribution may not be confident. To overcome this problem, limiting the spatial range of the energy within a narrow domain may be considered. We propose an approach based on a *short-sight* modeling of background, in which tracked objects deliberately ignore background color data spatially far from them. It allows to obtain consistent indicators for separation between background and object regions. A related principle was recently applied for object segmentation in still images, e.g. [8, 11]. In addition, we propose to model explicitly the temporal variation of color distribution within regions. To account for temporal consistency, we integrate an energy criterion penalizing the variation of histograms between consecutive frames. Our method shares common ideas with probabilistic color tracking [15] or with the mean shift tracker [4], in the extent we seek for regions in which color statistics match references resulting from previous frames. Energy minimization is performed thanks to the recent variational region growing approach developed in [16]. The benefits of using simultaneously *short-sight* background modeling on one hand and color distribution temporal variation on the other hand are demonstrated on synthetic and real image sequences.

2 Region tracking as a Bayesian estimation problem

2.1 Bayesian inference

Our model is defined over continuous space and discrete time. We consider the input video as a sequence of frames $I = \{I_1, ..., I_T\}$. Each frame is a mapping from space domain $D \subset \mathbb{R}^2$ to m-dimensional color domain C (e.g. $m = 3$ for RGB data). Spatio-temporal segmentation aims at providing for each frame I_t a partition P_t of the image domain into $n+1$ regions, i.e. background Ω_t^0 and n objects $\{\Omega_t^1, ..., \Omega_t^n\}$. Tracking is done in a sequential fashion, since next partition P_{t+1} is determined given current frame I_t and partition P_t. In this context, we first rely on the *Maximum A Posteriori* (MAP) framework introduced by Mansouri [9], who uses Baye's theorem and assumes conditional independence between image pixels:

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

Probability $\mathrm{p}(I_{t+1}(\mathbf{x})|I_t, P_t, P_{t+1})$ is the likelihood of observing a particular color at space-time location $(\mathbf{x}, t+1)$ given current image and both current and next partitions. This term will represent our assumptions about color constancy over time, whereas prior probability $\mathrm{p}(P_{t+1}|I_t, P_t)$ models available prior knowledge about object shape and/or motion. A tractable expression is obtained by making the reasonable assumption that the likelihood 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 $t+1$. Moreover, regions are assumed to have distinct color likelihoods, conditioned on their respective current configurations, which leads to:

$$\mathrm{p}(I_{t+1}(\mathbf{x})|I_t,P_t,P_{t+1}) = \mathrm{p}(I_{t+1}(\mathbf{x})|I_t,P_t,\mathbf{x}\in\Omega^i_{t+1}) \tag{2}$$

which we shorten to $\ell^i_{t+1}(\mathbf{x})$ for simplicity, corresponding to the likelihood of color $I_{t+1}(\mathbf{x})$ if $\mathbf{x}$ is inside region Ω^i at time t+1. Hence, the optimal partition maximizes the joint likelihood over every pixel from every region:

$$P^*_{t+1} = \underset{P_{t+1}}{\mathrm{argmax}} \prod_{i=0}^{n} \prod_{\mathbf{x}\in\Omega^i_{t+1}} \ell^i_{t+1}(\mathbf{x})\mathrm{p}(P_{t+1}|I_t,P_t) \tag{3}$$

The MAP estimation of partition P_{t+1} is turned into minimization of energy $E[P_{t+1}]$, taken as the negative log of posterior probability (3):

$$\begin{aligned} E[P_{t+1}] &= -\log\mathrm{p}(P_{t+1}|I_t,I_{t+1},P_t) \\ &= \sum_{i=0}^{n} -\left\{\int_{\Omega^i_{t+1}} \log\ell^i_{t+1}(\mathbf{x})\mathrm{d}\mathbf{x}\right\} - \log\mathrm{p}(P_{t+1}|I_t,P_t) \end{aligned} \tag{4}$$

2.2 Estimation of probability functions

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 estimation of color Probability Density Functions (PDFs). The estimation is global, in the extent that a single distribution is used to describe color statistics in an entire region. In image segmentation, this principle leads for instance to the maximization of histogram entropy [7] or discrepancy between region histograms [10]. More sophisticated color models could be used, however we currently focus on demonstrating the benefits of our approach with a simple color model. Let h^i_t be the kernel-based color histogram of region Ω^i at time t. For a given color α, we have:

$$h^i_t(\alpha) = \int_{\Omega^i_t} K_\sigma(I_t(\mathbf{x})-\alpha)\mathrm{d}\mathbf{x}$$

where K_σ is an m-dimensional isotropic Gaussian kernel with zero mean and standard deviation σ. While normalizing histograms by region areas in order to obtain PDFs, likelihood functions are formulated so that pixels at time $t+1$ will tend to be included into the best matching region, regarding statistics at time t:

$$\ell^i_{t+1}(\mathbf{x}) \approx \frac{h^i_t(I_{t+1}(\mathbf{x}))}{|\Omega^i_t|} = \frac{1}{\int_{\Omega^i_t}\mathrm{d}\mathbf{y}} \int_{\Omega^i_t} K_\sigma(I_t(\mathbf{y})-I_{t+1}(\mathbf{x}))\mathrm{d}\mathbf{y}$$

Estimating color PDFs in this way may be viewed as computing "smoothed" normalized color histograms within regions at time t and assigning color likelihood of a tested pixel $\mathbf{x}$ at time $t+1$ to the value of the corresponding bin in these histograms. To some extent, this approach is a "time-consistent" counterpart of the histogram-based segmentation model of [7]. As regards prior probability, with a generic view, we consider a simple non-temporal smoothness prior. No prior knowledge regarding shape or motion is available. It is thus relevant to consider the length of object boundaries as a regularizer, weighted by user-defined parameter ω controlling its significance:

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

3 Short-sight perception of background

We address one of the shortcomings inherent to tracking approaches based on global kernel-based estimation of color PDFs. The matter here is the lack of confidence of color statistics which might appear in cluttered backgrounds, due to the presence of various objects and static parts with different appearances. To illustrate the undesirable effect it may have on segmentation, consider the curve evolution problem related to the minimization of energy (4). Suppose that a portion of the boundary between background and object Ω^i is described by curve Γ parameterized by arc-length s. Calculus of variations with respect to Γ gives the following gradient flow (see for instance the mathematical framework of the region competition approach [17]):

$$\frac{\partial\Gamma(s)}{\partial\tau} = [\log\ell^0(\Gamma(s)) - \log\ell^i(\Gamma(s))]\mathbf{n}(s) + \ldots$$

where time index t+1 is dropped for simplicity, τ is the algorithmic time and $\mathbf{n}$ is the unit normal vector pointing towards region Ω^i. Regardless of curve implementation, which may rely either on parametric contours or level-sets, the curve will locally expand if color $I_{t+1}(\Gamma(s))$ matches Ω^i's statistical features more than the background's ones, and shrink in the opposite case. Estimating these statistical features over entire regions can be a drawback for tracking in real scenes, especially for the background region distribution, which may be cluttered and contain many non-tracked objects. In such case, statistical color data is highly scattered, so that background color likelihood ℓ^0 may not be confident. If ℓ^0 gets small for nearly all colors, regions are more likely to include background pixels and *leak* outside actual objects. One may consider ignoring background information in the tracking process, removing term over region Ω^0 from energy (4). This would imply to specify a threshold over the log-likelihood of pixels, below which pixels would be rejected from the target region. However, such threshold might be sensitive and thus prevent the method to be easily reproducible. To overcome this limitation and obtain reliable background image data, we head towards a background model based on "short-sight perception". To some extent, we adapt the philosophy of local modeling approaches [2, 8] to the tracking problem. Our approach is also related to the idea of spatial context brought up in [12]. Instead of considering statistical knowledge in the same extent for all background pixels, we attach more importance to background pixels as they are closer to objects, and introduce a relaxed version of the minimization problem (4). To this purpose, let $d(\mathbf{x},\Omega^i)$ be the euclidean distance from any background pixel $\mathbf{x}$ to the nearest pixel in a given object Ω^i:

$$d(\mathbf{x},\Omega^i) = \min_{\mathbf{y}\in\Omega^i} \|\mathbf{x}-\mathbf{y}\|$$

As we wish to decrease the contribution of background pixels to color statistics and energy, we introduce a positive real-valued weighting function ψ, which should be compactly supported and non-increasing with respect to distance d. Given a chosen support width w, several types of weighting functions are relevant, such as steps or piecewise smooth functions:

$$\psi_1(d) = \begin{cases} 1 & \text{if } d \le w \\ 0 & \text{otherwise} \end{cases} \qquad \psi_2(d) = \begin{cases} 1 - \dfrac{d}{w} & \text{if } d \le w \\ 0 & \text{otherwise} \end{cases}$$

Let B^i be the domain around object Ω^i where the distance weight is non-zero for every location. Hence, B^i may be thought of as a band of width w around the object. We consider that background color statistics are relevant only within bands B^i, and thus ignore available knowledge about color appearance in the "far" background $F = \Omega^0 \backslash \bigcup_{i=1}^n B^i$. In our multiple object tracking framework, each object has its own local perception of surrounding back-

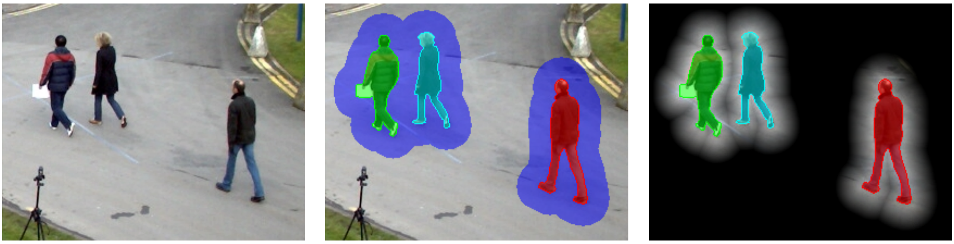

Figure 1: Tracked objects endowed with their own short-sight perceptions of background. Original image (left), segmented targets with surrounding bands (center) and background faded to black with respect to its contribution to perceptions (right)

ground and ignores far background. This principle is depicted in fig. 1. In the right part of the image, background pixels are faded to black proportionally with their distance to the nearest object boundary. Let k be the histogram over the outer band of a given object. It is considered as fuzzy in space, in the extent that contributions of pixels are weighted with respect to their distance to the target object :

$$k_t^i(\alpha) = \int_{B_t^i} \psi(d(\mathbf{x}, \Omega_t^i)) K_\sigma(I_t(\mathbf{x}) - \alpha) \mathrm{d}\mathbf{x}$$

The derived band likelihood function q relates the color of a band pixel to the normalized corresponding bin value in the previous band histogram:

$$\mathrm{q}_{t+1}^i(\mathbf{x}) = \frac{1}{\displaystyle\int_{B_t^i} \psi(d(\mathbf{y}, \Omega_t^i)) \mathrm{d}\mathbf{y}} k_t^i(I_{t+1}(\mathbf{x})) \tag{5}$$

This formulation aims at improving outlining of objects, as discrepancy between background and object colors is efficient mostly in the outer neighborhood of the object. Favouring close pixels advantageously prevents the discrepancy from being affected by changes on background pixel colors as soon as these changes arise far from the object. Moreover, in case of moving background, spatial fuzziness of the likelihood will allow gradual changes in local background representations. Adequacy of color according to statistics in the entire background is replaced by local weighted likelihoods over bands, which leads to the formulation of the *short-sight* energy, to be minimized with respect to candidate partition:

$$E_{\mathrm{SS}}[P_{t+1}] = \sum_{i=1}^{n} \left\{ -\int_{\Omega_{t+1}^i} \log \ell_{t+1}^i(\mathbf{x}) \mathrm{d}\mathbf{x} - \int_{B_{t+1}^i} \psi(d(\mathbf{x}, \Omega_{t+1}^i)) \log \mathrm{q}_{t+1}^i(\mathbf{x}) \mathrm{d}\mathbf{x} + \omega \left| \partial \Omega_{t+1}^i \right| \right\} \tag{6}$$

According to eq. (5) and (6), as background pixels are considered increasingly far from the objects, the contributions of these pixels decrease, both in the estimation of likelihoods at time t and in the energy at time $t+1$. Notice that a single background pixel may be included into more than one band, and thus intervene in the short-sight background representations of several objects. For instance, this is the case for the background part located between the two close persons in fig. 1.

4 Modeling temporal variation of color distribution

The second issue that we focus on is the possible overlapping of color distributions between regions, arising when objects share several colors, e.g the moving disks in fig. 3. Regardless of the perception of background, i.e. short-sighted or global, such overlapping may be especially disturbing for two adjacent objects (or an object and background). The two regions compete with each other with respect to shared colors, and the evolution of pixels is biased towards the region in which the likelihood is the highest. We tackle this problem by explicitly penalizing excessive variation of color distributions between successive frames, using histogram distance. We rely on the reasonable assumption that the distribution of each color within the tracked region and background exhibit small variations between successive frames. Hence, we integrate into energy (6) a term penalizing strong variations of region and band histograms between current and previous times:

$$\begin{aligned} E_{\text{SS-TVCD}}[P_{t+1}] \quad = \quad & \sum_{i=1}^{n} \Big\{ -\int_{\Omega_{t+1}^i} \log \ell_{t+1}^i(\mathbf{x})\mathrm{d}\mathbf{x} - \int_{B_{t+1}^i} \psi(d(\mathbf{x},\Omega_{t+1}^i)) \log \mathrm{q}_{t+1}^i(\mathbf{x})\mathrm{d}\mathbf{x} \\ & +\omega \left|\partial\Omega_{t+1}^i\right| + \lambda \left(J\left(h_{t+1}^i, h_t^i\right) + J\left(k_{t+1}^i, k_t^i\right)\right) \Big\} \end{aligned} \tag{7}$$

where J is a distance measure between histograms. According to this energy formulation, no assumption is made on the static distribution of color, but only on its variation. Moreover, we do not theoretically impose that object and background should have maximally different colors. Commonly employed measures for PDFs are the Kullback-Leibler divergence, the Bhattacharyya coefficient, or the Hellinger distance. In this paper, we chose to use the symmetrized Kullback-Leibler divergence:

$$J_{\text{KL}}(p,q) = \int_C (p(\alpha) - q(\alpha))(\log p(\alpha) - \log q(\alpha))\mathrm{d}\alpha$$

Originally, distances previously cited were designed to compare PDFs instead of histograms directly. Over a given region, the PDF can be estimated by normalizing the histogram by the region area, such that the sum of bins in the PDF is unitary. However, as regards temporal variation of color occurrences, it is relevant to work with non-normalized histograms. An undesirable property of using PDFs to penalize color variation is that adding a pixel $\mathbf{p}$ into a region slightly changes the distribution of all colors within this region. All bin values are modified, regardless of color $I(\mathbf{p})$. This may lead to unwanted behavior of the evolving region, as it can be encouraged not to add a pixel which is actually part of the object, if the energy increase on other bins exceeds the energy decrease on bins near $I(\mathbf{p})$. Using histograms naturally prevents this problem, since the addition/removal of a pixel $\mathbf{p}$ does not change the bin value of colors unrelated to $I(\mathbf{p})$.

If computed on histograms rather than on PDFs, the distance value is meaningful only if the two compared histograms result from regions having similar areas. Indeed, if the tracked object undergoes negligible change in its area between successive frames - i.e. we have $\left|\Omega_{t+1}^i\right| \approx \left|\Omega_t^i\right|$ - the PDF distance can be expressed directly in terms of the histogram distance. Considering the Kullback-Leibler symmetrized divergence, the following simplification applies:

$$J_{\text{KL}}\left(\frac{h_{t+1}^i}{\left|\Omega_{t+1}^i\right|}, \frac{h_t^i}{\left|\Omega_t^i\right|}\right) \approx \frac{1}{\left|\Omega_t^i\right|} J_{\text{KL}}\left(h_{t+1}^i, h_t^i\right) \tag{8}$$

In most studied image sequences, the variation of the proportion of area covered by the object or background is rather small compared to the total area of the image. When comparing histograms instead of PDFs, the range of distance is modified. In this case, the absence of normalizing expression $1/\left|\Omega_t^i\right|$ in eq. (8) can be compensated by changing the weight associated with the temporal variation term.

5 Optimization

Functionals over regions are most often optimized by gradient descent applied on a level set-based reformulation, either of the Euler-Lagrange equation or of the energy itself. Instead of doing so, we minimize energy (7) with the recent variational region growing approach [16], which we embed in a greedy evolution scheme. In addition to its purely algorithmic benefits - direct evolution of a set of pixels instead of a real-valued level set function, no need for time step parameter, *ad hoc* stopping criterion - it avoids to perform continuous calculus of variations. The partition-dependent optimization problem is turned into a discrete labeling problem, similar to Markov random fields. Let $\phi : D \to \{0, ..., n\}$ be the labeling function such that $\phi(\mathbf{x}) = i$ means pixel $\mathbf{x}$ is assigned to region Ω^i. Energy (7) is discretized in space and reformulated as a functional of labeling ϕ (index $t+1$ is dropped for simplicity and δ is the Dirac delta function):

$$\begin{aligned} E_{\text{SS-TVCD}}[\phi] \;=\; & \sum_{i=1}^{n} \Bigg\{ -\sum_{\mathbf{x}\in D} \Big\{ \delta(\phi(\mathbf{x})-i)\log \ell^i(\mathbf{x}) - \delta(\phi(\mathbf{x}))\psi(d^i[\mathbf{x},\phi])\log \mathrm{q}^i(\mathbf{x}) \Big\} \\ & + \lambda \left(J_{\text{KL}}\left(h^i[\phi], h_t^i\right) + J_{\text{KL}}\left(k^i[\phi], k_t^i\right)\right) \Bigg\} + \omega \sum_{(\mathbf{x},\mathbf{y})\in G} \delta(\phi(\mathbf{x}) - \phi(\mathbf{y})) \end{aligned}$$

where background pixels are selected using $\delta(\phi(\mathbf{x}))$. Clique G is the set of couples of neighboring pixels. When considered at time $t+1$, distances to objects d as well as histograms over objects h and bands k are now themselves functionals of the labeling:

$$\begin{aligned} d^i[\mathbf{x},\phi] &= \min_{\{\mathbf{y}\in D \,\mid\, \phi(\mathbf{y})=i\}} \|\mathbf{x}-\mathbf{y}\| \\ h^i[\alpha,\phi] &= \sum_{\mathbf{y}\in D} \delta(\phi(\mathbf{y})-i) K_\sigma(I_{t+1}(\mathbf{y})-\alpha) \\ k^i[\alpha,\phi] &= \sum_{\mathbf{y}\in D} \delta(\phi(\mathbf{y}))\psi(d^i[\mathbf{y},\phi]) K_\sigma(I_{t+1}(\mathbf{y})-\alpha) \end{aligned}$$

On the contrary, since they depend on histograms at time t, likelihood functions ℓ^i and q^i are independent from labeling ϕ. Starting from an initial configuration taken as the labeling of partition P_t, current labeling is evolved according to the region competition principle [17] until it leads to a local minimum of $E_{\text{SS-TVCD}}$. At each iteration, the evolution process considers a set of candidate pixels A, containing pixels having at least one differently labeled neighbor, i.e. located on inner or outer boundaries of objects. For each pixel $\mathbf{x}$ in A, we consider a set of candidate labels $L(\mathbf{x})$ which $\phi(\mathbf{x})$ may be switched to. Most of the time, pixels are on the interface between two regions only, so that $|L(\mathbf{x})| = 1$. Trivially, the set of candidate labels has more elements for pixels located at junctions between more than two regions. The decision of switching labels of candidate pixels is made according to the energy decreasing they lead to. Let $\Delta_{\mathbf{v}}^{l} E[\phi]$ be the induced energy variation of labeling ϕ when a single pixel $\mathbf{v}$ is switched to new label l, provided that $l \in L(\mathbf{v})$:

$$\Delta_{\mathbf{v}}^{l} E[\phi] = E[\tilde{\phi}] - E[\phi] \quad \text{s.t.} \quad \tilde{\phi}(\mathbf{x}) = \begin{cases} l & \text{if } \mathbf{x} = \mathbf{v} \\ \phi(\mathbf{x}) & \text{otherwise} \end{cases} \tag{9}$$

Variation $\Delta_{\mathbf{v}}^{l} E[\phi]$ is computed for all pixels $\mathbf{v} \in A$ and for all $l \in L(\mathbf{v})$. Pixels and associated labels such that $\Delta_{\mathbf{v}}^{l} E[\phi] < 0$ are subsequently sorted by increasing variation. Among all pixels leading to energy decreasing, only the first p pixels are actually switched, in order to maintain stability. In the experiments we made, p was relatively small compared to $|A|$. It should be further noticed that energy variation (9) is not computed as is. Otherwise, due to the nested functionals over ϕ, the algorithmic complexity of a brute-force implementation would be $O(|D|^2)$. In practice, it only requires $O(w^2 + \sigma^m)$ operations, where the w^2 term corresponds to the update of distance d in the circular neighborhood of radius w around $\mathbf{v}$, and the σ^m term is the cost of modifying histogram bins over m-dimensional balls (with radii being functions of σ) around $I_{t+1}(\mathbf{v})$.

6 Experiments and discusssion

We provide experimental results on both synthetic and natural image sequences. Histograms were computed in the initial RGB space quantified to 64 levels per channel. The standard deviation σ for kernel-based estimation was set to 0.75. We considered the segmentation in the first frame as an available input. For each dataset, weights λ and ω were tuned to achieve the best segmentation. In the optimization process, the number p of modified candidate pixels at each iteration was typically set to 20. Reported processing time for a single typical 720×576 frame was in the order of 5s, with a C++ implementation running on an 2.6GHz Intel Core2 Duo architecture.

Ground truth reference segmentations were available for datasets shown in figs. 2 and 3. Accuracy with respect to ground truth was quantified using the Dice similarity index. The 'Toy crocodile' dataset depicted in fig. 2 holds a single moving object and a static background, each one having distinct colors. Fig. 3 depicts the results obtained on the synthetic 'Moving disks' sequence. With this dataset, we aim to evaluate the tracking ability of our model in the difficult case where object and background have strongly overlapping color distributions. In addition to color, the background contains non-tracked moving structures similar to the target object in terms of shape and motion (random translations and rotations). For both sequences with ground truth data, the obtained Dice percentage was around 98%, which corresponds to fairly accurate segmentations.

The purpose of the experiment shown in fig. 4 is to demonstrate the performance of our approach in case of both moving camera and dynamic object. The local and fuzzy representation of background data allows our method to be robust against gradual background changes generated by ego-motion. Fig. 5 depicts comparative results of multitarget pedestrian tracking on a dataset taken from the PETS 2009 benchmark database[1] with different energy configurations, in order to show the improvements made by the short-sight modeling of background and the integration of temporal color histograms variation. The first row shows tracking results obtained with minimization of energy (4), i.e. with likelihood modeling over the entire background and no penalty on temporal variation. In this case, the region competition quickly becomes unable to sufficiently constrain the evolving regions and prevent leaking outside real objects. Without constraints, once the region has included tiny parts of background in a frame, it inevitably propagates in the background in subsequent frames. Replacing the global modeling of background by our short-sight perception managed to increase object/background color likelihood discrepancy and thus to reduce unwanted propagation in the background, as shown in the second row. The integration of histogram temporal variation (third row) allows to advantageously contrain regions within actual objects, as one would obtain by adding shape and motion priors (see for instance results obtained in [3]). Hence, in comparison with shape prior-based approaches, we believe that our method has potential in the sense the tracking process is basically constrained by a data term without additional application-dependent shape priors.

[1] http://www.cvg.rdg.ac.uk/PETS2009

7 Conclusion

We introduced a *short-sight perception* modeling of background for joint segmentation and tracking, focusing on the neighborhood of tracked objects to extract consistent statistical data for accurate separation between objects and background. To account for temporal consistency, we integrated a novel data term explicitly based on temporal variation of color distribution within objects and local background regions. As a possible extension, the method proposed in this paper can be easily modified to incorporate shape and motion priors. In this case, additional terms can be added into the spatio-temporal model. Future work will also include testing and evaluating the proposed model with other relevant color spaces (YUV, Lab, etc.), especially those in which brightness and pure color components are decoupled. We might incorporate motion estimation as well, in order that color variation and optical flow benefit from each other to perform spatio-temporal segmentation.

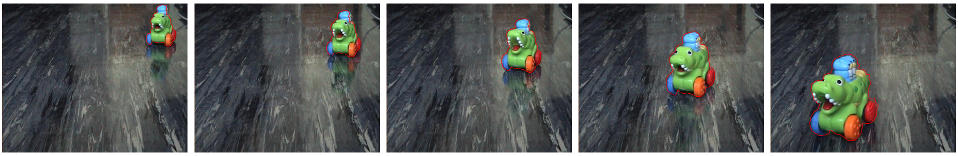

Figure 2: Single target tracking in case of few overlap between static background and object color distribution

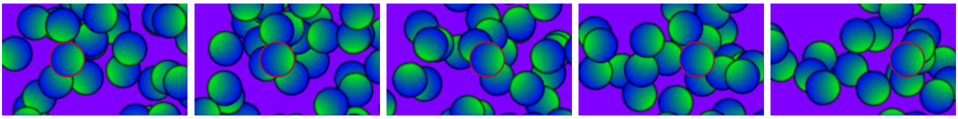

Figure 3: Single target tracking in case of strong overlap between moving background and object color distribution

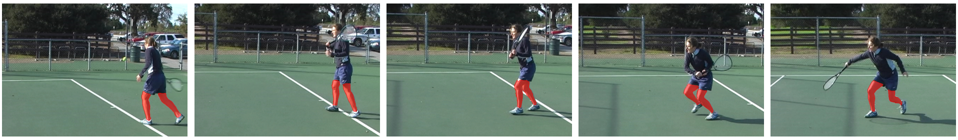

Figure 4: Single target tracking in case of ego-motion

References

[1] C. Aeschliman, J. Park, and A.C. Kak. A probabilistic framework for joint segmentation and tracking. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1371–1378, San Francisco, USA, 2010.

[2] 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.

[3] J. Chiverton, X. Xie, and M. Mirmehdi. Tracking with active contours using dynamically updated shape information. In *British Machine Vision Conference (BMVC)*, pages 253–262, Leeds, UK, 2008.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–575, 2003.

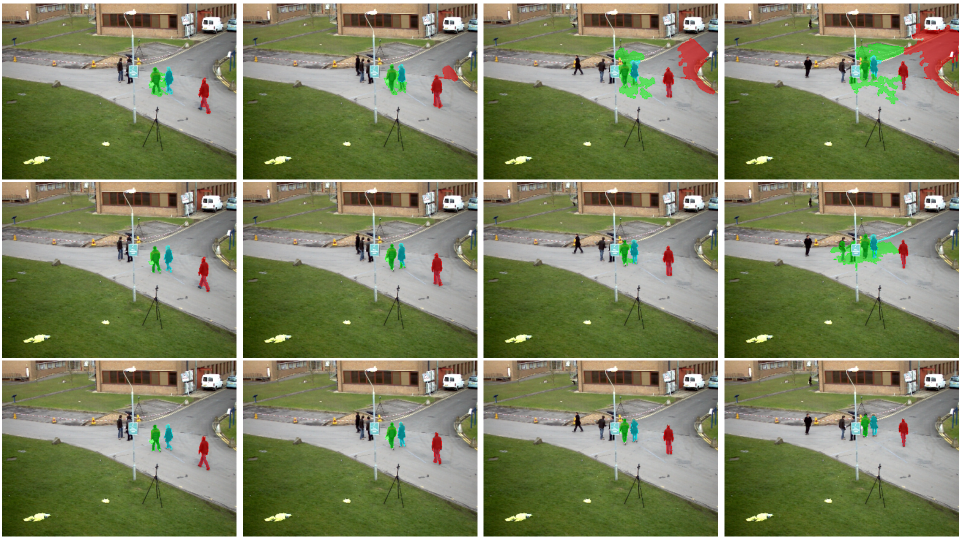

Figure 5: Multitarget tracking with different energy configurations: global background modeling without temporal variation (first row), short-sight modeling of background without temporal variation (second row) and short-sight modeling of background with temporal variation (third row)

[5] D. Freedman and T. Zhang. Active contours for tracking distributions. *IEEE Transactions on Image Processing*, 13(4):518–526, 2004. ISSN 1057-7149.

[6] K. Fundana, N. Overgaard, and A. Heyden. Variational segmentation of image sequences using region-based active contours and deformable shape priors. *International Journal of Computer Vision*, 80(3):289–299, 2008.

[7] J. Kim, J.W. Fisher, A. Yezzi, M. Çetin, and A.S. Willsky. A nonparametric statistical method for image segmentation using information theory and curve evolution. *IEEE Transactions on Image Processing*, 14(10):1486–1502, 2005.

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

[9] A.R. Mansouri. Region tracking via level set PDEs without motion computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):947–961, 2002.

[10] O. Michailovich, Y. Rathi, and A. Tannenbaum. Image segmentation using active contours driven by the Bhattacharyya gradient flow. *IEEE Transactions on Image Processing*, 16(11):2787–2801, 2007.

[11] J. Mille. Narrow band region-based active contours and surfaces for 2D and 3D segmentation. *Computer Vision and Image Understanding*, 113(9):946–965, 2009.

[12] 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.

[13] K. Ni, X. Bresson, T. Chan, and S. Esedoglu. Local histogram based segmentation using the Wasserstein distance. *International Journal of Computer Vision*, 84(1):97–111, 2009.

[14] N. Paragios and R. Deriche. Geodesic active regions and level set methods for motion estimation and tracking. *Computer Vision and Image Understanding*, 97(3):259–282, 2005.

[15] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *European Conference on Computer Vision (ECCV)*, volume 2350 of *LNCS*, pages 661–675, Copenhagen, Denmark, 2002. Springer.

[16] J-L. Rose, C. Revol-Muller, T. Grenier, and C. Odet. Unifying variational approach and region growing segmentation. In 18^{th} *European Signal Processing Conference (EUSIPCO)*, pages 1781–1785, Aalborg, Denmark, 2010.

[17] S. Zhu and A. Yuille. Region competition: unifying snakes, region growing, Bayes/MDL for multiband image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):884–900, 1996.