TOTAL BREGMAN DIVERGENCE FOR MULTIPLE OBJECT TRACKING

Andrés Romero, Lionel Lacassagne*

Laboratoire de Recherche en Informatique
Bat 650, Université Paris Sud
Email: andres.romero@u-psud.fr
lionel.lacassagne@u-psud.fr

Michèle Gouiffès

Laboratoire d'Informatique pour la Mécanique
et les Sciences de l'Ingénieur
Bat 508, Université Paris Sud
Email: michele.guiffes@u-psud.fr

ABSTRACT

In this paper we propose a multi-target tracking based on the *tracking-by-detection* paradigm. The problem is casted as a discrete association problem where a cost is assigned to each detection-tracklet pair and the evolution of many factors such as position, speed and appearance is observed. As new tracklets enter the scene their appearance is modeled using covariance matrices equipped with the total Bregman divergence to perform the comparisons and robust model updates. Our method provides near-state of the art results in terms of accuracy and is able to execute in real-time.

Index Terms— Total Bregman divergence, tracking in Riemannian manifolds, multiple object tracking.

1. INTRODUCTION

In any surveillance system, being able to do multiple object tracking is one of the most requested capabilities. Without reliable tracking results higher-level algorithms such as trajectory and behavior analysis become very hard. While being considered as classical, multi-target tracking still represents a very challenging problem in computer vision. One of the great difficulties faced by this problem, is the huge state-space cardinality it is forced to deal with. The possible number of trajectories a target can follow is incredibly high (discrete case) or even infinite (continuous case). In addition, when objects suffer considerable changes in scale inside the tracking zone, the state-space cardinality grows even more. These difficulties can be tackled by introducing some dynamical constraints in the linear and angular velocity to reduce the number of possibilities into a physically plausible set of trajectories, locations and scales. Appearance information such as color or texture can be very useful to follow or recover a target after an occlusion, but taking advantage from it is not always straightforward as in many cases: occlusions, scale, illumination and appearance changes and background clutter contamination can drastically reduce the performance of these methods. To overcome this problem, the representation of the target has to be invariant and robust to all such phenomena, but unfortunately most color invariants, although robust against lighting changes, can lead to a reduction in the separability between targets increasing the number of matching ambiguities. In addition, when targets have a non rigid motion or have low textural or structural contents, gradient or corner-based methods, such as the classical KLT [1] or popular key-point matching methods such as SIFT [2] or SURF [3] are not appropriate.

As a consequence of the continuous increase in the performance offered by human detection methods such as the

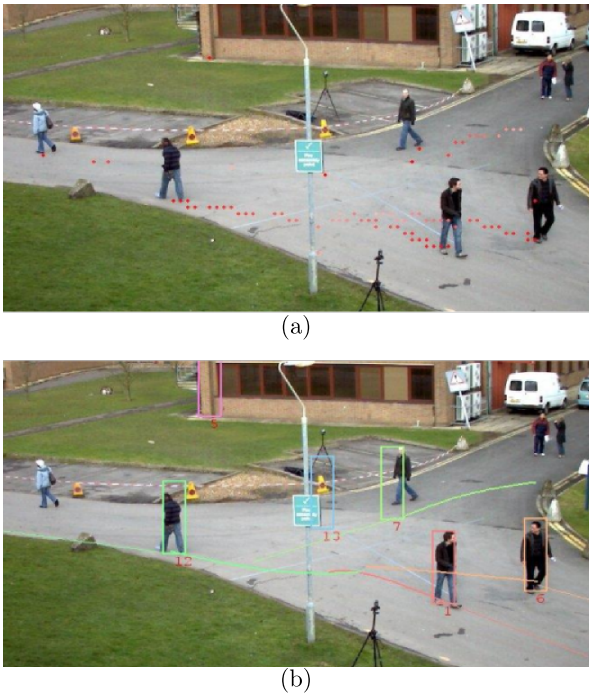

Fig. 1. In every frame t our multi-target tracking receives a series of detections $\{\mathbf{d}_1 \cdots \mathbf{d}_n\}$ and consistently traces the object trajectories $\{\mathbf{p}_1 \cdots \mathbf{p}_m\}$.

*To appear at ICIP 2013, September 15-18, Melbourne Australia.

histograms of oriented gradients (HOG)[4], detection in Riemannian manifolds [5], LBP based detection [6] and relative optical flow (HOF) [7], multi-target tracking methods based in the *tracking by detection* strategy [8] are becoming very popular. This article is particularly influenced by the ideas presented in [9] and [10] where discrete/continuous energy functions summarize the cost associated to a set of individual target locations. More specifically, an overall energy function considers target's dynamical properties such as linear and angular speed, and encourages trajectory persistence while penalizing unexpected target disappearances or sudden creations occurring far from the tracking area borders.

While the possibility of including appearance information inside the energy function is left open in [9], the authors preferred not to do it because in their experiments they found that handling appearance information in the form of color histograms was not discriminant enough. In this article we recover this idea and propose a very efficient and robust technique to handle color and gradients information. In our approach objects are represented by covariance matrices as in [11, 12] and [13]. This type of representations are very compact and allow us to perform multiple comparisons against candidate locations in real-time. The set of covariance matrices lies in the Riemannian manifold formed by the set of symmetric positive definite matrices (SPD), many different metrics and divergences have been proposed in the literature to do element by element comparisons in this set. Between the most popular we have the Riemannian metric in [14] and the Jensen-Bregman LogDet Divergence[13] which is not properly a metric but it is less expensive in computational terms. In [15], multiple observations of the same target are blended together into a general model using the sample mean Riemannian covariance matrix which depends in the Riemannian metric for SPD matrices. One problem of this approach is that it can be seriously affected by the presence of outliers. Because of its computational efficiency and its robustness against outliers, in this work we are interested in the total Bregman divergence (TBD) and its corresponding ℓ_1-norm based center (t-center) defined in [16].

The rest of the paper is structured as follows: in Section 2 we provide an in-depth description of the objects appearance representation in the form of covariance matrices and the geometrical properties of the total Bregman divergence. The discrete global energy function and its individual components is presented in Section 3. The energy minimization strategy is presented in Section 4 and finally we present the experimental results of our approach in Section 5.

2. APPEARANCE MODELING

Given an image I_t of size $W \times H$ a mapping function ϕ is applied to each pixel yielding a $W \times H \times d$ dimensional tensor image $F(x,y) = \phi(I_t, x, y)$. This mapping function ϕ may analyze local image information such as position, color, gradients and filter responses. A rectangular region of n feature points $\{\mathbf{z}_k\}_{k=1\cdots n-1}$ inside $F(x,y)$ is described compactly by

$$\mathbf{C}_R = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{z}_k - \mu)(\mathbf{z}_k - \mu)^T, \tag{1}$$

of size $d \times d$. Here, vector μ is the mean of the feature points $\{\mathbf{z}_k\}_{k=1\cdots n-1}$. In our experiments the feature map

$$\phi(I_t, x, y) = \left[\begin{array}{ccccccc} x & y & R & G & B & I_x & I_y \end{array}\right]$$

was selected to integrate color and directional gradients distribution in the object region.

Covariance matrix representation is advantageous because it preserves spatial and statistical information and allows to compare different sized regions. However, direct arithmetic subtraction fails to compare covariance matrices because this set of matrices does not lie on the Euclidean space. The Riemannian metric defined in [14] is the proper metric for this manifold. Equipped with this metric it is even possible to find the centers of a set covariance matrices with the mean Riemannian covariance (MRC) proposed in [15]. One problem of the Riemannian metric is that it is too expensive to compute, other disadvantage is that the MRC is affected by outliers. The geometric median proposed in [17] offers a solution to this problem but the existence of this geometric median is not guaranteed and similar to the MRC it is expensive to compute. Recent articles such as [13] and [18] propose other dissimilarity measures to alleviate this difficulties. The total Bregman divergence (TBD) proposed in [18] is a robust and efficient dissimilarity measure.

$$\delta_{TBD}(P,Q) = \frac{\log\left(\det\left(P^{-1}Q\right)\right) + tr(Q^{-1}P) - d}{2\sqrt{c + \frac{(\log(\det Q))^2}{4} - \frac{d(1+\log 2\pi)}{2}\log(\det Q)}}, \tag{2}$$

where $c = \frac{3d}{4} + \frac{d^2\log 2\pi}{2} + \frac{(d\log 2\pi)^2}{4}$. While the TBD based ℓ_1-norm center (known as t-center) $\bar{Q}$ is a robustly represents of a set of covariances $\{Q_i\}_{i=1}^n$

$$\bar{Q} = \left(\sum_{i=1}^{n}\frac{w_i^{-1}}{Q_i}\right)^{-1}, \tag{3}$$

where

$$w_i = \frac{\left(2\sqrt{c + \frac{(\log(\det Q_i))^2}{4} - \frac{d(1+\log 2\pi)}{2}\log(\det Q_i)}\right)^{-1}}{\sum_j\left(2\sqrt{c + \frac{(\log(\det Q_j))^2}{4} - \frac{d(1+\log 2\pi)}{2}\log(\det Q_j)}\right)^{-1}}. \tag{4}$$

While equations (2) and (4) may look complicate, they are much more economical than the Riemannian metric based alternatives because they have an analytic form among other reasons.

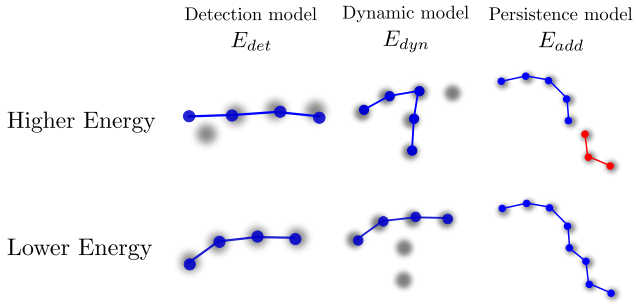

Fig. 2. Decisions are made considering the addition of multiple energy terms. Top row shows the configurations that render higher energies for each individual term, lower energies are exemplified in the lower row. Detection peaks are denoted in gray.

3. DISCRETE MULTI-OBJECT TRACKING

The aim of our multi-tracking algorithm is to consistently detect, identify and trace object locations trough time. Ideally, the number of trajectories by object is exactly one, but inter-object occlusions, disappearances i.e., background occlusions, and false negatives in the detection method may cause trajectory drifts and fragmentations. For every time frame t our method receives a set of detections $D_t = \{\mathbf{d}_1, \cdots, \mathbf{d}_m\}$ and the set of active paths (tracklets) $P_t = \{\mathbf{p}_1, \cdots, \mathbf{p}_n\}$. Each $\mathbf{p}_j$ contains historical information about the object $\forall t \in \{t_{start}, \cdots, t_{end}\}$ where t_{start} is the frame where $\mathbf{p}_j$ was created and t_{end} the last frame where information about $\mathbf{p}_j$ is available. This historical information registers all previous locations $\mathbf{x}_j^t$ and previous appearances descriptions in the form covariance matrices $\mathbf{C}_j^t$. From these appearance descriptions an appearance model $\bar{\mathbf{C}}_j$ is calculated using the TBD t-center of equation (4). A linear fitting function (i.e. least-squares) applied at time t, produces a series of predicted locations $\hat{\mathbf{x}}_j^t, \forall t \in \{t_{end}+1, \cdots, t_{end}+N_f\}$ for every active path $\mathbf{p}_j \in P_t$. Here, N_f defines the length of the predictions. In our experiments it was set to $N_f = 15$.

Frame by frame the algorithm evaluates all possible $\mathbf{d}_i$ and $\mathbf{p}_j$ correspondences and creates new paths as it is required. At the end, the algorithm outputs a dynamic mapping set M_t containing the final set of tuples $M_t = \{(\mathbf{d}_i, \mathbf{p}_j)\}$ that minimize a discrete energy function. This idea is illustrated in Fig. 1-a where a set of points is marking locations of the detections (old detection are marked in a lighter red). In Fig.1-b these detections have been grouped into trajectories (tracklets) tracing the evolution of object locations in the scene. Energy function $E(\mathbf{d}_i, \mathbf{p}_j)$ rewards plausible configurations and penalizes unreasonable ones. Our energy function includes most of the terms presented in [9]: a detection term based on the image detection data E_{det}, a physically motivated term which models the object dynamics (linear and angular velocities) E_{dyn} an appearance term E_{app} that measures the similarity of the location covariance matrix against the covariance model inside $\mathbf{p}_j$, and one last term which penalizes path creations favoring persistent trajectories E_{add}

$$E(\mathbf{d}_i, \mathbf{p}_j) = E_{det} + E_{dyn} + E_{app} + E_{add}. \quad (5)$$

Figure 3 shows how each one of this terms collaborates to detect the more plausible configurations.

3.1. Detection Model

Objects are detected with a sliding window that measures the HOG classifier response [4]. The energy term E_{det} uses the predictions $\hat{\mathbf{x}}_j^t$ inside $\mathbf{p}_j$. Energy is lower when $\hat{\mathbf{x}}_j^t$ passes trough a region of high pedestrian likelihood:

$$E_{det}(\hat{\mathbf{x}}_j^t) = \lambda + \sum_{g=1}^{D(t)} \frac{-c}{\|\hat{\mathbf{x}}_j^t - \mathbf{d}_g\|^2 + c} \quad (6)$$

In equation (6), $D(t)$ represents the number of detection peaks of frame t and $\mathbf{d}_g$ is the location of peak g. The value of λ penalizes predictions $\hat{\mathbf{x}}_j^t$ lacking of image evidence. Its value was fixed to 0.05 in all the experiments.

3.2. Dynamic Model

The motion term E_{dyn} assumes a constant velocity model:

$$E_{dyn} = \alpha \|\mathbf{v}_j^t - \hat{\mathbf{v}}_j^{t+1}\|, \quad (7)$$

where $\mathbf{v}_j^t = \dot{\mathbf{x}}_j^t = \mathbf{x}_j^t - \mathbf{x}_j^{t-1}$ is the current velocity vector of path p_j, and $\hat{\mathbf{v}}_j^{t+1}$ is estimated considering the detection location $\mathbf{d}_i$ as $\hat{\mathbf{v}}_j^{t+1} = \mathbf{d}_i - \mathbf{x}_j^t$.

3.3. Appearance Model

Every detection $\mathbf{d}_i \in D_t$ has covariance descriptor $\hat{\mathbf{C}}_i$ which is compared against the t-center $\bar{\mathbf{C}}_j$ associated to every path $\mathbf{p}_j \in P_t$. In the case of strong similarity the energy function is awarded with a negative value, if not the energy functions grows in proportion to the TBD.

$$E_{app} = \begin{cases} \beta \delta_{TBD}(\hat{\mathbf{C}}_i, \bar{\mathbf{C}}_j) & \text{if } \delta_{TBD}(\hat{\mathbf{C}}_i, \mathbf{C}_j) > \delta_{max} \\ -\gamma & \text{if } x < \delta_{max} \end{cases} \quad (8)$$

3.4. Persistence Model

When the algorithm find correspondences between detections and paths, it also considers the possibility of a new object entering the scene. This hypothesis is penalized in proportion to the Euclidean distance between the detection $\mathbf{d}_i \in D_t$ and the tracking area borders.

4. ENERGY MINIMIZATION STRATEGY

Our energy function is certainly not convex, we tackle this problem considering every pair resulting from the combination of D_t and P_t. This way our problem reduces to a classical combinatorial optimization problem which is solved efficiently by Munkres algorithm [19] (the Hungarian method). Several tracking methods in computer vision have followed this idea [20].

For each detection $\mathbf{d}_i \in D_t$ the minimization algorithm builds a list of candidate paths measuring the distance between $\mathbf{d}_i$ and the predicted location $\hat{\mathbf{x}}_j^t$ of the active path $\mathbf{p}_j \in P_t$. The TBD between the $\mathbf{d}_i$'s covariance descriptor $\hat{\mathbf{C}}_i$ and $\mathbf{p}_j$'s model $\mathbf{c}_j$ may also be considered. Paths whose predictors at frame t are outside a circle of radius r from the $\mathbf{d}_i$ and whose covariance model is extremely dissimilar are considered impossible combinations and an infinite energy is assigned accordingly. In the case of few or no paths close to any of detections, the *null* candidate is added indicating the creation of new path. Once the the algorithm has a complete list of candidates with their respective energies, the algorithm builds a list of all possible combinations $\{M_k\}_{k=0,\cdots,K-1}$ that are not mutually exclusive. Munkres algorithm considers the cost of all these combinations and to selects the less expensive one $M_t = \min(\{M_k\}_{k=0,\cdots,K-1})$. The complete processes is represented compactly in **Algorithm** 1.

Algorithm 1: Multi-target tracking by discrete combinatorial energy minimization.

Data: List of m detections D_t and the list of n paths P_t.
Result: Updated list of paths P_{t+1}.

1. For each detection $\mathbf{d}_i \in D_t$, $i = \{0, \cdots, m-1\}$ build a list of paths p_ℓ containing every $\mathbf{p}_j \in P_t$ whose predictors are located within a radius r from $\mathbf{d}_i$. If there are no paths close to $\mathbf{d}_i$ append c_0 to P_t to evaluate the cost of creating a new path.
2. From all the different p_ℓ's build the list of K possible candidate combinations: $\{M_k\}_{k=0,\cdots,K-1}$.
3. For each combination in $\{M_k\}_{k=0,\cdots,K-1}$, apply equation (5) to evaluate the energy associated to it.
4. Select the optimal combination $M_t = \min(\{M_k\}_{k=0,\cdots,K-1})$ having the minimal energy (using Munkres method).
5. Update the paths position, predictions and appearance model.
6. Create new paths as M_t indicates.

5. EXPERIMENTS

To evaluate our approach we tested it four with widely used real world datasets as well as many other sequences of our own. One of the tested sequences is PETS 2009-S2L1-V1 from the VS-PETS2009 benchmark[1] where only the first viewpoint is used. The other three datasets come from the TUD [2] dataset: TUD-Campus, TUD-Crossings and TUD. Ground truth is available for all these datasets. To evaluate our results we follow the current best practice which are the CLEAR-metrics proposed in [21]. Precision criteria are similar to the ones uses in [9]. False positives, missed targets and identity switches are measured by the Multiple Object Tracking Accuracy (MOTA), the average distance between true object locations and the estimated targets is represented by the Multiple Object Tracking Precision (MOTP). Other metrics such as the number of mostly tracked (MT), partially tracked (PT) and mostly lost (ML) trajectories, and the number of swaps are reported as well. All the videos resulting from these experiments are available[3]. Results are reported in **Table 1**. The execution speed greatly depends on the image size and the number of targets in the sequence all four sequences were executed in real-time with an average speed $\approx$ 20fps.

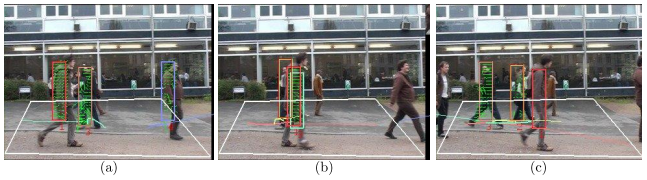

Fig. 3. Target recovered correctly after occlusions in TUD-campus sequence.

Sequence	MOTA	MOTP	MT	PT	ML	Swaps
Pets 2009 S2-L1-V1	70.084%	12.91 px	16	3	0	5
TUD Campus	74.92%	10.12 px	5	2	0	0
TUD	69.73	12.19 px	5	1	1	1
TUD Crossings	74.63%	7.79 px	8	5	0	4

Table 1. Tracking results

6. CONCLUSIONS

We have presented an algorithm that is able to handle multiple targets at the same time robustly in real-time solving a discrete combinatorial energy problem with Munkres method. The use of appearance information in the form of covariance matrices provides a valuable hint to the multi-target problem. Appearance information in the form of covariance matrices is more discriminant when robust *distances* such as the total Bregman divergence are employed. TBD-based t-centers provide a fast and robust average of covariance matrices.

[1] http://www.cvg.rdg.ac.uk/PETS2009
[2] http://www.mis.tu-darmstadt.de/node/428
[3] http://andresromeromier.wikispaces.com/

7. REFERENCES

[1] C. Tomasi and T. Kanade, *Detection and tracking of point features*, School of Computer Science, Carnegie Mellon Univ., 1991.

[2] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer Vision–ECCV 2006*, pp. 404–417, 2006.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, vol. 1, p. 886893.

[5] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on riemannian manifolds," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 10, pp. 17131727, 2008.

[6] Etienne Corvee and Francois Bremond, "Haar like and LBP based features for face, head and people detection in video sequences," in *International Workshop on Behaviour Analysis and Video Understanding (ICVS 2011)*, Sophia Antipolis, France, Sept. 2011, p. 10.

[7] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1030–1037.

[8] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[9] A. Andriyenko and K. Schindler, "Multi-target tracking by continuous energy minimization," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1265–1272.

[10] A. Andriyenko, K. Schindler, and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1926–1933.

[11] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," *Computer VisionECCV 2006*, p. 589600, 2006.

[12] J. Badie, S. Bak, ST Serban, and F. Brémond, "Recovering people tracking errors using enhanced covariance-based signatures," .

[13] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos, "Efficient similarity search for covariance matrices via the jensen-bregman logdet divergence," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2399–2406.

[14] W. Förstner and B. Moonen, "A metric for covariance matrices," *Quo vadis geodesia*, pp. 113–128, 1999.

[15] F. Porikli, O. Tuzel, and P. Meer, "Covariance tracking using model update based on lie algebra," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 1, pp. 728–735.

[16] B.C. Vemuri, Meizhu Liu, S.-I. Amari, and F. Nielsen, "Total bregman divergence and its applications to DTI analysis," *Medical Imaging, IEEE Transactions on*, vol. 30, no. 2, pp. 475 –483, Feb. 2011.

[17] P.T. Fletcher, S. Venkatasubramanian, and S. Joshi, "Robust statistics on riemannian manifolds via the geometric median," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[18] Meizhu Liu, B.C. Vemuri, S.-I. Amari, and F. Nielsen, "Shape retrieval using hierarchical total bregman soft clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 12, pp. 2407 –2419, Dec. 2012.

[19] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial & Applied Mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[20] K. Jungling and M. Arens, "Detection and tracking of objects with direct integration of perception and expectation," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1129–1136.

[21] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The CLEAR 2006 evaluation," *Multimodal Technologies for Perception of Humans*, p. 144, 2007.