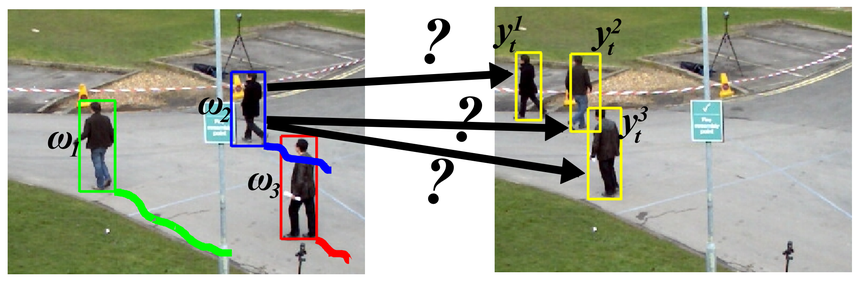
Multi-Target Data Association using Sparse Reconstruction

Andrew D. Bagdanov, Alberto Del Bimbo, **Dario Di Fina**, Svebor Karaman, Giuseppe Lisanti, Iacopo Masi
{dario.difina,svebor.karaman}@unifi.it {bagdanov,delbimbo,lisanti,masi}@dsi.unifi.it

Introduction

- Video surveillance is an important and fundamental application area in computer vision. Visual object tracking is one of the most important tasks related to this.
- Multiple target tracking is to follow targets in an uncontrolled environment while handling problems such as occlusion, similarity in target appearance and crowded scenes.

- The data association (DA) problem is one of the main hurdles to be overcome in multiple target tracking [1].
- It consists of finding the correct assignment between existing tracklets and the set of new observations extracted from the current frame of a sequence.



- **Our approach attempts to reconstruct new observations using a regularized linear combination of tracklets already identified.**
- **It uses ℓ_1-regularized basis expansions to determine the most likely assignment between tracked targets and new observations.**

Related Work

State of the Art for Multi-Target Data Association

- Nearest Neighbor Standard Filter (NNSF): the simple and widely applied approach to multi-target data association.
- JPDAF and Baysian Filters: maintain a statistical model of target motion at each time step.
- Markov Chain Monte Carlo Data Association (MCMCDA): uses random sampling to explore the detection/tracklet association space.

Sparse Methods

- Sparse methods [3, 4] are becoming steadily more popular in the computer vision community.
- These approaches exploit the hypothesis that an arbitrary signal can be reconstructed using a sparse combination of basis vectors.
- Sparse reconstruction has recently been applied to the single-target tracking problem [3].
- In a discriminative classification setting, sparse reconstruction has also been applied to face recognition problems [4].

Isolating the Data Association Problem

Multi-Target Data Association Problems

This work is focused only on the *pure data association* problem. We want:

- a representation that discriminatively models each target through time.
- an accurate rule for discerning each subject from the others in the scene.

Problem Formalization

- We consider a video stream Ψ whose duration is $T \in \mathbb{N}^+$ seconds, and that K different targets moving in the scene can be identified.
- Each $k \in K$ is observable in the time interval $[t_{ks}, t_{ke}] \subset [1, T]$, where t_{ks} is time of the first appearance and t_{ke} is the last appearance or exit time (hence $t_{ks} < t_{ke}$).
- We assume that a perfect detector lets us obtain a set of observations y_t with a cardinality $L \in \mathbb{N}$,

$$Y = \{y_t : t \in [1, T]\}, y_t = \{y_t^i\}_{i=1}^{L}, \forall t.$$

- A tracking algorithm has the aim of defining a set of tracklets: $\Omega = \{\omega_k : k \in [1, K]\}$.
- Each tracklet ω_k will be characterized by a sub-set of observations, where each observation of ω_k belongs to a distinct time instant:

$$\omega_k = \{y_t^i : i \in [1, L], \forall t \in [t_{ks}, t_{ke}]\} \subseteq Y.$$

- An observation y_t^i can only be associated with a single tracklet $\omega_k, \omega_k \cap \omega_j = \emptyset, \forall k, j \in [1, L]$ if $k \neq j$.

Assumptions

- We assume perfect detections and perfect bootstrapping of appearance models in order to isolate data association performance from the complexities of multi-target tracking.
- We assume that n observations have already been associated with the k-th tracked target.

Bibliography

References

[1] Y. Bar-Shalom and T.E. Fortmann. *Tracking and Data Association.* Academic-Press, Boston, 1988.

[2] Michael D. Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(9):1820–1833, September 2011.

[3] Xue Mei and Haibin Ling. Robust visual tracking using l1 minimization. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1436 –1443, 29 2009-oct. 2 2009.

[4] John Wright, Allen Yang, Arvind Ganesh, Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(2), February 2009.

[5] Bo Yang and Ram Nevatia. Online learned discriminative part-based appearance models for multi-human tracking. In *ECCV (1)'12*, pages 484–498, 2012.

Sparse Discriminative Basis Expansion

Discriminative Basis Formation

- Define the sub-basis corresponding to target k as the concatenation of the n feature descriptors of all associated observations: $\mathbf{B}_k = [\mathbf{f}(y_{k,1}), \mathbf{f}(y_{k,2}), \ldots, \mathbf{f}(y_{k,n})]$.
- If there are K targets in the scene, the **discriminative basis B** is obtained by concatenating these sub-bases, which is hence composed of $N = K \cdot n$ feature vectors.
- Our approach is based on solving an ℓ_1-regularized optimization problem:

$$\min_{\alpha} \|\mathbf{f}(y_t^i) - \mathbf{B}\alpha\|_2^2 + \lambda\|\alpha\|_1,$$

  - α is a a sparse projection vector composed of N coefficients that indicate how to reconstruct a new observation $\mathbf{f}(y_t^i)$ using a linear combination of the sample vectors in $\mathbf{B}$.
  - $\lambda \in \mathbb{R}^+$ is used to control the sparsity of α.
  - To define the target *appearance error* we use the residual ε_k^i for each (k, i):

$$\varepsilon_k^i = \|\mathbf{f}(y_t^i) - \mathbf{B}_k\alpha_k\|_2.$$

Spatial Information

- We use the **VOC Score** between tracklets and new observations to incorporate spatial proximity:

$$s_k^i = \frac{A_k \bigcap A_i}{A_k \bigcup A_i}, s_k^i \in [0, 1].$$

- A_k is the bounding box area of the last observation y_τ^l associated with the tracklet ω_k, and A_i is the area of the new observation y_t^i, where $\tau \in [t-5, t-1]$.

Association Error

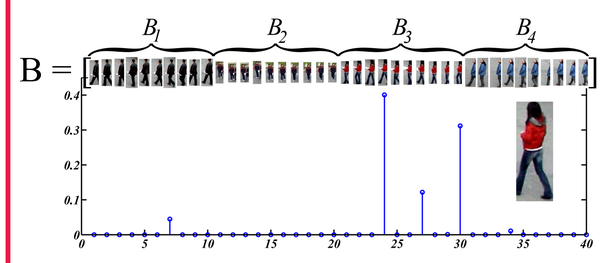
The **total association error** a_k^i is defined as a linear combination of the two errors:

$$a_k^i = (1-\gamma)\varepsilon_k^i + \gamma(1 - s_k^i), \quad \forall (k, i) \in [1, K] \times [1, L].$$

Basis Update

- During the tracking process the discriminative basis may **become outdated** and thus no longer describe well a particular target k.
- **Basis update** is performed by exploiting the associations occurring in a temporal window of W frames.
- For each tracklet we add at most the η best associated observations to the corresponding sub-basis.

Discriminative Basis



An example of a regularized sparse basis expansion and the resulting α projection vector.

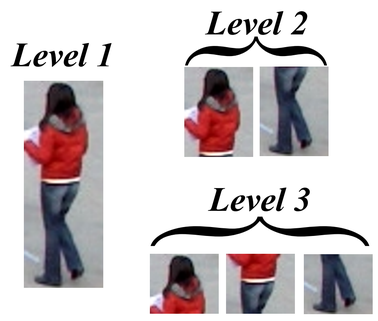
DA Algorithm

Algorithm 1: Data association algorithm

Data: B, Ω, y_t and γ
$\Omega_t = \Omega$: local set of tracklets ;
compute $\mathbf{f}(y_t^i) \forall y_t^i$; $s_k^i, \varepsilon_k^i, a_k^i \forall i, \forall k$;
while $\Omega_t \neq \emptyset \wedge y_t \neq \emptyset$ **do**
  $(\hat{k}, \hat{i}) = \arg\min_{k,i} a_k^i$;
  $\omega_{\hat{k}} = \omega_{\hat{k}} \cup \{y_t^{\hat{i}}\}$;
  $y_t = \{y_t \setminus y_t^{\hat{i}}\}$;
  $\Omega_t = \{\Omega_t \setminus \omega_{t,\hat{k}}\}$;
end while

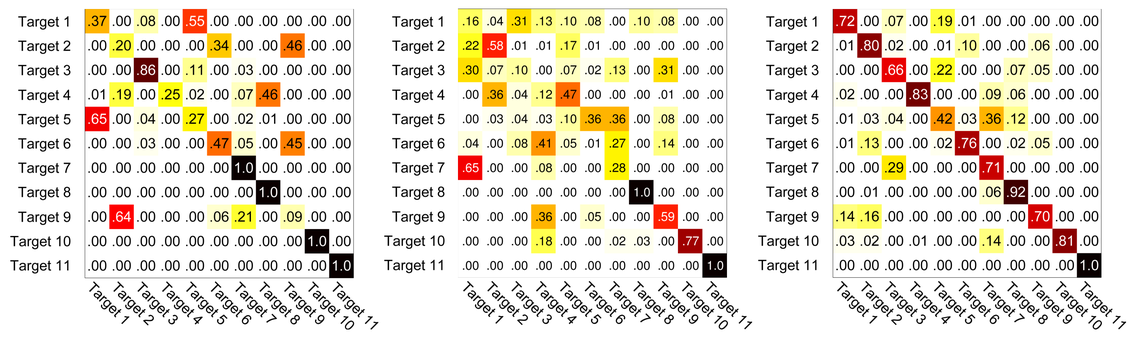
Experimental Results

Feature and Dataset



- We use the *"s2.l1-view01"* sequence of the PETS 2009 dataset to test our approach.
- In the figure on the left, we show three instances of each tracked subject in the PETS dataset sequence.
- In the figure on the right, we show an example of the feature representation we use. We extract an **RGB histogram** for each cell of the three level spatial pyramid shown in figure.

Data Association Performance



Confusion matrices for various parameter settings. Left: $\lambda = 0.7$, $\gamma = 0.5$, no update phase. Center: spatial proximity only with $\gamma = 1$. Right: $\gamma = 0.2$ and $\lambda = 0.1$, basis update with $W = 20$ and $\eta = 3$.

Comparison with the State-of-the-art

Method	MOTA	Recall	Precision	FN Rate	FP Rate	IDS
Yang [5] PM Only	–	92.8%	95.4%	–	–	0
Yang [5] PM + CFT	–	97.8%	94.8%	–	–	0
Breitenstein et al. [2]	79.7%	–	–	–	–	–
Our ℓ_1-DA ($\gamma = 0.2$)	82.8%	82.9%	96.2%	13.9%	0.04%	146
Our ℓ_1-DA ($\gamma = 0.4$)	84.7%	84.8%	98.4%	13.9%	0.02%	60
Our ℓ_1-DA ($\gamma = 0.8$)	80.5%	80.5%	99.9%	19.4%	0%	4

Results on the "s2.l1-view01" sequence of the PETS 2009 dataset.