Multi-Target Data Association using Sparse Reconstruction

Andrew D. Bagdanov, Alberto Del Bimbo, Dario Di Fina, Svebor Karaman, Giuseppe Lisanti, and Iacopo Masi*

Media Integration and Communication Center
University of Florence, Florence, Italy
http://www.micc.unifi.it/vim/people/

Abstract. In this paper we describe a solution to multi-target data association problem based on ℓ_1-regularized sparse basis expansions. Assuming we have sufficient training samples per subject, our idea is to create a discriminative basis of observations that we can use to reconstruct and associate a new target. The use of ℓ_1-regularized basis expansions allows our approach to exploit multiple instances of the target when performing data association rather than relying on an average representation of target appearance. Preliminary experimental results on the PETS dataset are encouraging and demonstrate that our approach is an accurate and efficient approach to multi-target data association.

Keywords: Data association, multi-target tracking, sparse methods, video surveillance.

1 Introduction

Computer vision applied to video surveillance applications like abnormal behaviour detection, group interaction analysis, and object tracking has received a lot of attention in the last decade. One of the most important tasks related to these topics is visual object tracking. The task of multiple target tracking is to follow targets in an uncontrolled environment while at the same time handling problems such as occlusion, similarity in the target appearance and crowded scenes.

The data association (DA) problem is one of the main hurdles to be overcome in multiple target tracking and consists of finding the right assignment between the set of tracked targets and the set of new observations extracted from the current frame of a sequence. For each tracked target, the past observations that have already been associated with it compose a tracklet. In figure **??** on the left are shown three tracklets corresponding to three tracked subjects. On the right are three new observations that must be associated to these targets. This task may become difficult in real-world scenarios due to many problems that

* Author names in alphabetical order.

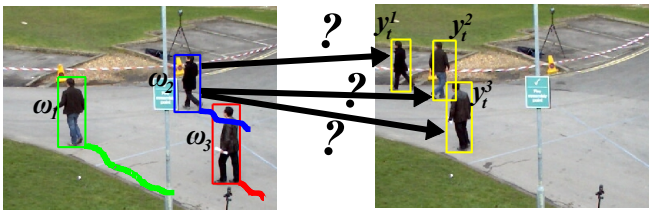


Fig. 1: The data association problem. At each time instant observations y_t^i must be associated with one of the running trackers ω_k.

may arise. One problem is how to create a representation that discriminatively models each target through time, while another is how to build an accurate rule for discerning each subject from the others in the scene. Moreover, if we consider real time constraints, data association must scale well with the number of targets. For these reasons it is usually intractable to solve the data association problem for many targets, and approximations are often applied.

In this article we describe an approach to solving the multi-target data association problem using sparse basis expansion. Our approach attempts to reconstruct new observations using a regularized linear combination of tracklets already identified. Techniques based on sparse basis expansion have lately became popular in the computer vision community. Sparse methods have been applied to tracking [**?**], face recognition [**?**] and vocabulary construction [**?**]. Our approach builds a discriminative basis for each target and uses ℓ_1-regularized basis expansion to determine the most likely assignment between tracked targets and new observations.

In the next section we briefly discuss previous work related to multi-target tracking and sparse methods. In section **??** we describe our technique for using sparse basis expansion to solve the multi-target data association problem. Finally we report some preliminary experiments in section **??** and discuss our ongoing work in section **??**.

2 Related work

In this section we review some of the relevant work on data association and sparse methods. For more complete reviews refer to [**?**, **?**, **?**]. The simple and probably most widely applied approach to multi-target data association is the Nearest Neighbor Standard Filter (NNSF) [**?**]. This method uses the Mahalanobis distance to compute the association error of a running tracker with a given observation at time t. The association is obtained by choosing the smallest Mahalanobis distance between all tracker/measurement pairs, repeating this process in a greedy way until all tracklets have been assigned to an observation. However NNSF is susceptible to integration of incorrect measurements and can produce overconfident estimates over time.

Target tracking is usually performed using a Kalman filter or other Bayesian filter that maintains a statistical model of target motion at each time step. In

this context, one of the most widely used technique is the Joint Probabilistic Data Association Filter (JPDAF) [?,?,?,?,?]. The JPDAF produces a set of hypotheses that associate tracklets with new observations, applies a gate to reduce the set of admissible hypotheses, and then computes a MAP estimate of the Bayes-optimal solution to the data association problem at the current time step. This method does not scale well with the number of targets and observations since many hypotheses can be generated and this can result in a huge increase of the computational complexity of the MAP solution. Another technique widely used in the literature is Markov Chain Monte Carlo Data Association (MCMCDA) [?, ?]. This approach uses a Monte Carlo sampling process defined over "moves" that change associations over short intervals of time. By randomly sampling changes in the association according to a statistical model, the MCMCDA approach is able to efficiently search in a very large space and to find a good approximation to the optimal data association.

Sparse methods 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 (potentially many) basis vectors. Sparse reconstruction has recently been applied to the single-target tracking problem [?]. This approach tries to find the best association between target and observation using a basis composed of past target observations. In a discriminative classification setting, sparse reconstruction has also been applied to face recognition problems [?]. In this work the authors define an approach to face recognition that uses sparse reconstruction of probe images in terms of a dictionary of gallery faces. Classification of an unknown face is performed using the reconstruction error of sparse basis expansions.

3 Data association by sparse reconstruction

This work is focused only on *pure data association* problem, assuming perfect detections and perfect bootstrapping of appearance models in order to isolate data association performance from the complexities of multi-target tracking, leaving to future work the study of a complete tracking framework. In this section we explain and analyze each stage of our method.

3.1 The data association problem

In general, a data association problem is the association of measurements with models (trackers, in our case) at each time step of a sequence. More formally, considering a video stream Ψ whose duration is $T \in \mathbb{N}^+$ seconds, suppose that K different targets moving in the scene can be identified. Now consider a particular target k 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 $t_{ks} < t_{ke}$). For each time instant t, we consider that a perfect detector lets us obtain a set of observations y_t with a cardinality $L \in \mathbb{N}$, such that:

$$\begin{aligned} Y &= \{y_t : t \in [1, T]\}, \\ y_t &= \{y_t^i\}_{i=1}^{L}. \end{aligned} \tag{1}$$

For each instant t the value of L may be different. A tracking algorithm has the aim of defining a set of tracklets:

$$\Omega = \{\omega_k : k \in [1, K]\}. \tag{2}$$

Considering the set Y, 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:

$$\omega_k = \{y_t^i : i \in [1, L], \forall t \in [t_{ks}, t_{ke}]\} \subseteq Y. \tag{3}$$

Note that an observation y_t^i can only be associated with a single tracklet ω_k:

$$\omega_k \cap \omega_j = \emptyset, \forall k, j \in [1, L] \text{ if } k \neq j. \tag{4}$$

In the following, k will refer to a target. A good tracker needs a good data association method to correctly associate observations y_t^i and a tracklets ω_k.

3.2 Sparse discriminative basis expansion for data association

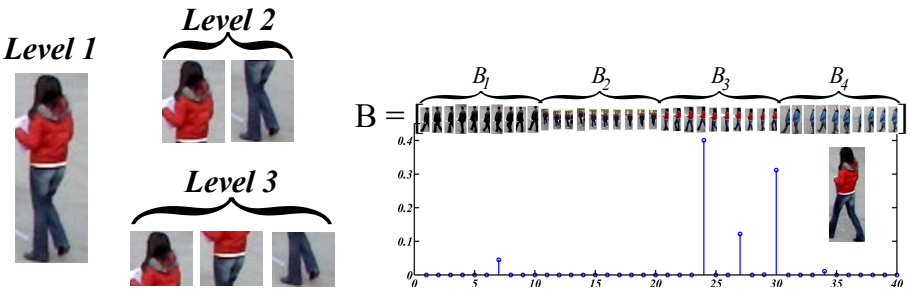
We will first introduce the feature descriptor we use and then we will explain how we build discriminative bases for data association. Using a regularized sparse basis expansion we can then compute a reconstruction error according to the basis of each existing target. Finally, our algorithm solves the data association problem by combining the reconstruction error with spatial proximity information.

Feature adopted For each observation y_t^i, we extract a feature vector $\mathbf{f}(y_t^i)$. In this work, we use a pyramidal color histogram to obtain a multi-level representation of the appearance of each detection. We define a three level pyramid, where the top level corresponds to the full detection window, the second level to two non-overlapping horizontal slices and the third and last level to three horizontal slices (see figure **??**).

Each slice is represented by a RGB color histogram $\mathbf{h}_i$ which is normalized with the ℓ_1 norm, while the whole feature vector $\mathbf{h}$ is normalized with the ℓ_2 norm. This feature maintains multi-level appearence information, that is define as a vector $\mathbf{f}(y_t^i) \in \mathbb{R}^m$ with $m = |\mathbf{h}| = 3072$ bins. It does not rely on complex foreground/background segmentation or part models, and has good illumination invariance and good independence with respect to the quality of the observations.

Discriminative basis construction The key idea behind our approach is the construction and use of a discriminative basis $\mathbf{B}$ that, when used to perform a sparse reconstruction of an unknown target, can be exploited to recover which basis vectors of a tracked target contribute most to the sparse reconstruction. Assuming that n observations have already been associated with the k-th tracked target, we define the sub-basis corresponding to target k as the concatenation of the n feature descriptors of all associated observations:

$$\mathbf{B}_k = [\mathbf{f}(y_{k,1}), \mathbf{f}(y_{k,2}), \ldots, \mathbf{f}(y_{k,n})]\,. \tag{5}$$



(a) (b)

Fig. 2: Left (**??**), the three level pyramid used in our feature representation. Right (**??**), an example of a regularized sparse basis expansion and the resulting α projection vector.

If there are K targets in the scene, the discriminative basis $\mathbf{B}$ is obtained by concatenating these sub-bases, which is hence composed of $N = K \cdot n$ feature vectors. As noted in the beginning of the section, we consider a perfect bootstrapping, inspired by the concept of *reliable tracklet* in [**?**]. We considered for each target a bootstrapping of $n = 25$ frames.

Regularized sparse basis expansion The upper part of figure **??** depicts an example of a discriminative basis. The basis $\mathbf{B}$ is called *discriminative* because each vector of $\mathbf{B}$ is a feature vector $\mathbf{f}(y_t^i)$ associated with a specific target label (i.e. we maintain discriminative information about a each sub-basis).

Solving an ℓ_1-regularized optimization problem:

$$\min_{\alpha} \|\mathbf{f}(y_t^i) - \mathbf{B}\alpha\|_2^2 + \lambda\|\alpha\|_1, \tag{6}$$

lets us obtain a sparse projection vector α. This vector is composed of N coefficients that indicates how to reconstruct a new observation $\mathbf{f}(y_t^i)$ using a linear combination of the sample vectors in $\mathbf{B}$.

The coefficient $\lambda \in \mathbb{R}^+$ in equation (**??**) is used to control the sparsity of α: the larger the value of λ, the lower the ℓ_1 norm of the projection vector α. In the lower part of figure **??** the magnitudes of the reconstruction coefficients α_k^j are depicted for an observation sample.

To estimate which target k should be associated with a new observation $\mathbf{f}(y_t^i)$, we analyze the reconstruction coefficients in α. The vector α can be seen as a concatenation of α_k that are the coeffcients corresponding to each target, such that $\alpha = \left[\alpha_1\ \alpha_2 \cdots \alpha_k \cdots \alpha_K\right]$, where each α_k is composed of n α_k^j. To identify the associated target we define a *reconstruction error* ε_k^i for each (k, i):

$$\varepsilon_k^i = \|\mathbf{f}(y_t^i) - \mathbf{B}_k\alpha_k\|_2. \tag{7}$$

The value ε_k^i corresponds to the reconstruction error when $\mathbf{f}(y_t^i)$ is reconstructed using only those coefficients from α and columns from $\mathbf{B}$ that correspond to

Algorithm 1: Data association algorithm

Data: $\mathbf{B}$, 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

tracked target k. Since the feature is ℓ_2 normalized, $\varepsilon_k^i \in [0, 1]$. The smaller the error ε_k^i, the greater the likelihood that y_t^i represents the target k.

Spatial proximity information We use the VOC Score [**?**] between trackers an new observations to combine spatial proximity and sparse reconstruction error. The spatial proximity score between tracker k and observation i is:

$$s_k^i = \frac{A_k \bigcap A_i}{A_k \bigcup A_i}, s_k^i \in [0, 1], \tag{8}$$

where A_k is the bounding box area of the potential last observation y_τ^l associated with the tracklet ω_k, and A_i is the area of the new observation y_t^i. The VOC score corresponds to the overlap of A_k and A_i normalized by the union of the areas. If the areas are highly overlapping the score s_k^i will tend to one, while its value will tend to zero if the overlap is small. Note that $\tau \in [t-5, t-1]$ i.e. we only compute s_k^i, if there is an association with tracker ω_k in the last 5 frames.

Finally, the VOC Score is used in combination with the reconstruction error ε_k^i introduced in the previous section to define the association error a_k^i:

$$a_k^i = (1-\gamma)\varepsilon_k^i + \gamma(1 - s_k^i), \quad \forall (k, i) \ \in \ [1, K] \times [1, L]. \tag{9}$$

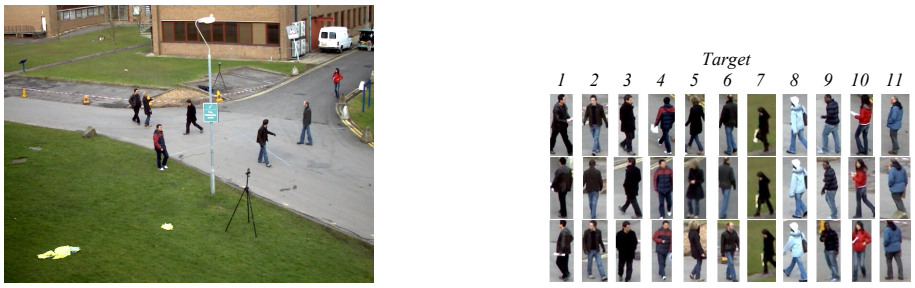
The parameter γ is used to control the tradeoff between spatial and sparse reconstruction in determining the association error.

3.3 Data association algorithm

In this section we put together all of the concepts above to define an algorithm for data association using sparse basis expansions. The purpose of our preliminary study is to determine the potential of sparse methods for data association under ideal conditions. As such, we make a number of simplifying assumptions. Most importantly, we assume that perfect detections are available for all persons appearing in the video stream and that the first n observations of each target can be perfectly associated. This effectively allows us to create the discriminative basis $\mathbf{B}$, and to update it when new targets appear in the scene.

At each time instant t, we compute for each tracklet/observation pair the reconstruction error and spatial proximity that are combined to form the association error, see equation (**??**). A new observation is associated with one of the existing tracklets according to the greedy algorithm **??**.

During the tracking process the information contained in the discriminative basis may become outdated and thus may no longer describe well a particular target k. In our data association algorithm we include an update phase that adds a fixed number of feature vectors for each target. 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 (according to equation (**??**)) to the corresponding sub-basis. With this approach, the discriminative basis size may increase and be different for each person after the update, i.e. we may have $n_k \neq n \ \forall k$. But the ℓ_1-regularization in our reconstruction will always tend to give a sparse projection vector.



(a) (b)

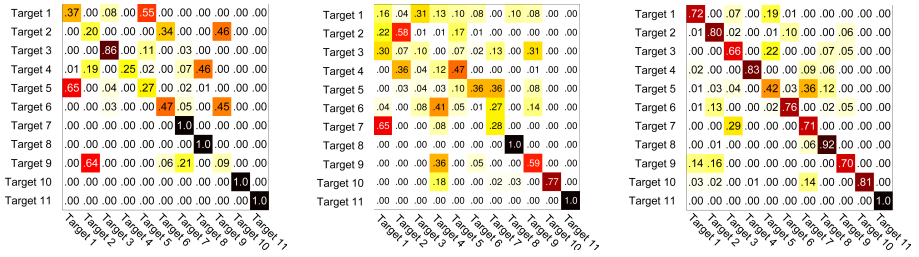
Fig. 3: Example data associations. (a) A sample frame of the dataset. (b) Three instances of each tracked subject in the sequence.

4 Experiments

In this section we report some experimental results obtained with our method varying the parameters λ, γ and with or without the basis update phase. Experiments are performed on the "s2.l1-view01" sequence of the PETS 2009 public dataset (see figure **??**), that is a de-facto standard in the tracking community due to its challenging nature. It is one of the most used sequence in the literature on multi-target tracking [**?**, **?**]. Comparison with state-of-the-art methods are obtained using CLEAR MOT [**?**].

4.1 Data association performance

The results obtained by varying λ and γ are shown as confusion matrices in figure **??**. These matrices give a clear idea of the potential of our method for pure multi-target data association under ideal tracking conditions. The matrix shown in figure **??** was obtained with $\lambda = 0.7$, $\gamma = 0.5$ and without updating the



(a) (b) (c)

Fig. 4: Confusion matrices for various parameter settings. (a) with $\lambda = 0.7$, $\gamma = 0.5$, no update phase. (b) spatial proximity only with $\gamma = 1$. (c) with $\gamma = 0.2$ and $\lambda = 0.1$, basis update with $W = 20$ and $\eta = 3$.

discriminative basis. This value of λ means that the behavior of data association is similar to the NNSF. From this matrix, we can observe nearest neighbor approaches can result in many association errors, in particular for targets 1, 2, 4, 5 and 6. This is principally due to the fact that these targets have similar appearance and frequently occlude each other.

The matrix shown in figure **??** was obtained with $\lambda = 0.1$ and $\gamma = 1$, without updating the discriminative basis. This value of γ forces the proposed association to give more importance to the spatial proximity score. However, as shown in the confusion matrix, the association accuracy is low for most of the targets. The matrix of figure **??** instead is obtained using $\lambda = 0.1$ and $\gamma = 0.2$, with the update of the discriminative basis with $W = 20$ and $\eta = 3$. In this confusion matrix we can observe that this configuration gives the best association results for each target with respect to the other configurations. From these results we can conclude that values of λ which enforce sparsity without reducing the projection vector α to a single non-null value are preferable and also that adding a contribution from spatial proximity improves the results.

4.2 Comparison with the state-of-the-art

In this section we discuss the results obtained with the CLEAR MOT metrics [**?**]. In particular, table **??** reports results with three different configurations of our approach corresponding to $\gamma \in \{0.2, 0.4, 0.8\}$ and some results from state-of-the-art techniques on this sequence [**?**,**?**]. Note that these results are only a byproduct of data-association process and are not directly comparable with the state-of-the-art methods, given that we make many simplifying assumptions about perfect detections and initial discriminative basis construction. This comparison is only intended to give some indication of the potential of our approach, considering that we focused only the *pure data association problem.*

For these experiments we set $\lambda = 0.1$ and update the basis with $W = 20$ and $\eta = 3$. In table **??** we can observe that varying γ mostly affects the number

of identity switches (IDS) and the accuracy (MOTA) of our solution. MOTA indicates the accuracy of the approach in terms of multi-target tracking. The recall is computed as the total number of true positives over the total number of ground truth objects, while the precision is calculated as the total number of true positive over the sum of the number of active tracks over frames. The switch of target identities affects the precision because one id switch induces one less true positive. Moreover, the proposed solution does not deal with target that are no longer present in the scene because they are not removed from our model. This can result in a false negative if a detection is associated to a target that has exited the scene.

From table **??** it is possible to observe that, with a low value of γ, our approach gives more importance to the appearance and this can result in a identity swaps since very little spatial proximity information is considered. However, by increasing the contribution of spatial proximity score in the association score we observe a dramatic decrease in identity switches, though this can also result in lower accuracy. From these results, setting $\gamma = 0.4$ seems to be a good tradeoff between MOTA and IDS. Our method performs competitively with the state-of-the-art, even compared to offline tracking methods that require all detections beforehand to perform association and extract trajectories of each target [**?**].

Method	MOTA	Recall	Precision	FN Rate	FP Rate	IDS
Yang [**?**] PM Only	–	92.8%	95.4%	–	–	0
Yang [**?**] PM + CFT	–	97.8%	94.8%	–	–	0
Breitenstein et al. [**?**]	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

Table 1: Results on the "s2.l1-view01" sequence of the PETS 2009 dataset.

5 Discussion and conclusions

In this paper we propose an approach to multi-target data association that exploits sparse reconstruction and spatial proximity. We show that integrating multiple templates of each target in a discriminative basis helps in the association process, but at the same time spatial constraints are required to obtain good performance for the tracking of multiple targets. With our preliminary results we show that, under very strict ideal detection and initialization hypotheses, our data association approach can be competitive with those used by state-of-the-art tracking methods. Our ongoing work consists in verifying these results with respect to the state-of-the-art in realistic situations through automatic target initialization and update, as well as through long term discriminative basis maintenance over long sequences.