JOINT SPARSITY-BASED ROBUST VISUAL TRACKING

Behzad Bozorgtabar\textsuperscript{1}

\textsuperscript{1}Vision & Sensing, HCC Lab, ESTeM
University of Canberra
Email: Behzad.Bozorgtabar@canberra.edu.au

Roland Goecke\textsuperscript{1,2}

\textsuperscript{2}IHCC, RSCS, CECS
Australian National University
Email: roland.goecke@ieee.org

ABSTRACT
In this paper, we propose a new object tracking in a particle filter framework utilising a joint sparsity-based model. Based on the observation that a target can be reconstructed from several templates that are updated dynamically, we jointly analyse the representation of the particles under a single regression framework and with the shared underlying structure. Two convex regularisations are combined and used in our model to enable sparsity as well as facilitate coupling information between particles. Unlike the previous methods that consider a model commonality between particles or regard them as independent tasks, we simultaneously take into account a structure inducing norm and an outlier detecting norm. Such a formulation is shown to be more flexible in terms of handling various types of challenges including occlusion and cluttered background. To derive the optimal solution efficiently, we propose to use a Preconditioned Conjugate Gradient method, which is computationally affordable for high-dimensional data. Furthermore, an online updating procedure scheme is included in the dictionary learning, which makes the proposed tracker less vulnerable to outliers. Experiments on challenging video sequences demonstrate the robustness of the proposed approach to handling occlusion, pose and illumination variation and outperform state-of-the-art trackers in tracking accuracy.

Index Terms— Particle filter, joint sparsity-based model, iteratively reweighted least squares, adaptive dictionary

1. INTRODUCTION

Although visual tracking has been studied for many years [1, 2, 3], it is still challenging due to inevitable object appearance variation, rotation, cluttered and dynamic backgrounds, and occlusion. A typical setting of the problem is that an object is identified, either manually or automatically, in the first frame of a video sequence and is tracked in the subsequent frames by estimating its likely state (e.g. particle filter [4, 5]) as it moves around. The focus of this paper is on the appearance model since it is usually the most crucial part of any tracking algorithm.

Recently, sparse representation [6, 7] has been successfully applied to visual tracking [8] by modelling the target appearance using a sparse approximation over a template set, which leads to the so-called \( \ell_1 \) trackers as it needs to solve an \( \ell_1 \) norm related minimisation problem. While these trackers show magnificant tracking accuracy, they are computationally very expensive. Although, some recent efforts have been made to speed up this tracking paradigm [9], these methods assume independency between the sparse representations of the particles, which constrain their representation. Consequently, very recent efforts have been made to exploit similarities between particles [10]. These methods implement the multi-task sparse learning approach [11] for visual tracking in a particle filter framework, where learning the representation of each particle is viewed as an individual task. However, these methods highly rely on certain assumptions of the task relationships. For instance, the model might include some outlier particles which do not participate in sharing common structure.

In this paper, we propose a computationally efficient learning approach without making specific structure assumptions. We directly impose a joint sparsity-based regularisation with a mixture of structure and outlier penalties and formulate the objective as an unconstrained convex problem. To derive the optimal solution efficiently, we propose to use an Iteratively Reweighted Least Square (IRLS) method, which is optimised using a conjugate gradient based method that guarantees its convergence.

Finally, we propose an online update scheme robust against outliers thanks to the sparsity nature, so that the tracker can adapt to appearance variations of the target object and the background. We present experimental results on challenging benchmark tracking sequences, which show that the synergistic formulation proposed in this paper results in a more accurate, efficient and robust tracking than the recently proposed sparsity-based trackers.

2. RELATED WORK

Object tracking methods can be categorised as either generative or discriminative. To track objects with appearance change due to various factors, an adaptive appearance model is important. Examples of generative methods are the Incremental Visual Tracker (IVT) [12] with an online subspace algorithm to model target appearance adaptively and Visual Tracking Decomposition (VTD) [13] with multiple motion and observation models to account for appearance variations. Discriminative trackers [14] formulate the tracking problem as a binary classification problem. In this case, the tracker finds the target location that best separates the target from the background. Motivated by its popularity in face recognition, sparse coding techniques have recently migrated over to object tracking [15, 16]. Sparse coding based trackers use a sparse linear representation w.r.t. a set of target and occlusion templates to describe particles sampled in each frame. Particle representations are learned by solving a constrained minimisation problem for each particle independently. Recent work has focused on sparse coding based on Multiple Task Learning (MTL), which makes object tracking more efficient. The key idea of MTL is to explore the hidden relationships among multiple tasks to enhance learning performance.

For another instance, Zhang et al. [17] propose a particle-filter based tracker that exploits the low-rank nature of particle representations and employ an efficient Inexact Augmented Lagrange Multi-
plier (IALM) approach. However, the model commonality is a fairly strong assumption, which is often invalid in real-world applications.

Our proposed method is inspired by the above ideas. We directly impose a regularisation term with a mixture of structure and outlier penalties, which leads to a flexible and robust formulation for robust object tracking.

3. PROPOSED ALGORITHM

In the next section, we first briefly review the Bayesian filtering framework, then we present our object tracker. Finally, we discuss the proposed dictionary learning.

3.1. Bayesian Filtering Framework

Visual tracking is formulated within the Bayesian filtering framework in which the goal is to determine a posteriori probability of the target state. In this paper, we utilize the particle filter as an effective tool in which the observations (gray scale values) are denoted by a Gaussian distribution. The observation model \( p(z_t | s_t) \) can be represented by a Gaussian distribution. The observation model \( p(z_t | s_t) \) reflects the similarity measure for the tracking target. In this paper, the weights of the particles are specified by the difference in contribution of object templates and the background templates, in which a sample with a larger difference score indicates that it is more likely to be generated from the target class. The most likely sample is considered as the tracking result for that video frame.

3.2. Proposed Joint Sparsity-Based Tracker

In our particle filter based tracking method, particles are randomly sampled around the current state of the tracked object according to a zero-mean Gaussian distribution. In the \( t \)th frame, we consider \( N \) particle samples, whose observations (gray scale values) are denoted in matrix form as \( Y = [y_1, \ldots, y_N] \in \mathbb{R}^{m \times N} \). Given a dictionary \( A_t = [a_1, \ldots, a_K] \in \mathbb{R}^{m \times K} \), where \( a_i \) is the \( i \)th dictionary atom. In the noiseless case, each particle \( y_i \) is represented as a linear combination of templates that form a dictionary \( A_t \), such that \( Y = A_t W \). The dictionary columns comprise the target templates that will be used to represent each particle. We denote \( A_t \) with a subscript because the dictionary templates will be moderately updated to incorporate variations in object appearance due to changes in illumination, viewpoint, etc.

Recently, sparse regularisations have been applied to classic problems such as classification based feature selection studies. For example, the multi-task feature learning [18] used the \( \ell_{2,1} \)-norm regularisation to couple feature selection across tasks using a strict assumption that all tasks share a common underlying representation. However, often, the common pattern is shared by many tasks, but not all. Since particles are densely sampled around the current target state, to address this issue, we formulate our tracker as a Joint Sparsity-Based Model to include both \( \ell_{2,1} \)-norm and \( \ell_1 \)-norm regularisations for selecting dictionary templates, which yields a solution that achieves the within and between tasks sparsity (particle representation) simultaneously. Such a formulation makes our proposed tracker more accurate and also notably improves the tracking speed:

\[
\min_W \|Y - A_t W\|_F^2 + \lambda_1 \|W\|_1 + \lambda_2 \|W\|_{2,1}
\]

where \( \|\cdot\|_F \) denotes the Frobenius norm and the constant coefficients \( \lambda_1 \) and \( \lambda_2 \) balance the mixture of the two norms. Such a formulation is shown to be more flexible in terms of handling individual particle’s representation, including both grouped structure norm \( \|W\|_{2,1} \) and outlier detecting norm \( \|W\|_1 \). Here, we propose an efficient least squares based algorithm to solve our objective function in Eq. (2). By choosing \( \lambda_1 = \lambda_1(1 - \gamma) \) and \( \lambda_2 = \lambda_1 \gamma \) and taking the partial derivative with respect to \( w_k (1 \leq i \leq N) \) and setting it to zero, we obtain:

\[
A_t^T y_i - A_t^T A_t w_i + \frac{\lambda}{2} D_t w_i = 0,
\]

where \( w_i \) is the \( i \)th column of matrix \( W \) and \( D_t \) is a diagonal matrix with the \( k \)th diagonal element defined as:

\[
(D_t)_{kk} = \text{diag} \left( (1 - \gamma) \frac{1}{\|w_k^T\|^2_2} + \frac{\gamma}{\|w_{ki}\|^2} \right)
\]

This diagonal element comprises the group impact \( \|w_k\|^2_2 \) by imposing row sparsity on the \( k \)th row of the coefficient matrix \( W \), while the second term \( \|w_{ki}\|^{-2} \) indicates the impact of the \( k \)th dictionary item on the \( i \)th particle. The parameter \( \gamma \) balances the impacts of these two components to the diagonal matrix \( D \). Solving the above equation is equivalent to the following weighted least squares problem:

\[
\arg\min_{w_i} \|y_i - A_t w_i\|_2^2 + \frac{\lambda}{2} \left\| D_t^{-1} w_i \right\|_2^2
\]

To obtain a feasible solution for this high dimensional problem, we rewrite above the equation using [19] as \( H_i^{-1} (A_t^T A_t + \frac{\lambda}{2} D_t^{-1}) w_i = H_i^{-1} A_t^T y_i \), where \( H_i = \text{diag} (A_t^T A_t + \frac{\lambda}{2} D_t) \). Then, we use a gradient based algorithm, e.g. Accelerated Proximal Gradient (APG), to solve this linear system. In our work, we implemented the Preconditioned Conjugate Gradient (PCG), since it is more convenient to our preconditioned linear system. The overall algorithm is summarised in 1. Note that the final updated coefficients are obtained as:

\[
W^{k+1} = \left( A_t^T A_t + \frac{\lambda}{2} D_t^{-1} \right)^{-1} A_t^T y
\]
Algorithm 1: Proposed Algorithm

Input: Initial dictionary $A$ and set of particles $Y$

1. Initialise $k ← 0, W, D^{(0)}$ with the identity matrix;
2. while not converged do
3. Update matrix $H = \text{diag} \left( A^T A + \frac{1}{2} D^{(k)} \right)$;
4. Solve the linear system:
   $$H^{-1} A^T A + \frac{1}{2} D^{(k)} = H^{-1} A^T y;$$
5. Update the matrix $D^{(k+1)}$ using Eq. (4);
6. $k ← k + 1$
end
Output: The updated $W$

4. DICTIORITY UPDATE STRATEGY

The proposed dictionary is initialised by sampling image patches around the initial target position. Furthermore, to attenuate the problem of target drift in tracking, we elevate our dictionary $A_t$ with the background templates, such that $A_t = \{A^{(O)}_t, A^{(B)}_t\}$, where $A^{(O)}_t$ and $A^{(B)}_t$ represent the target object and background templates, respectively. In this paper, we propose an online robust dictionary learning framework. The proposed online dictionary update scheme saves time, while not sacrificing much accuracy. Suppose frame $q$ is the current frame. We define a matrix $Z = [z_{ij}] \in \mathbb{R}^{m \times q}$ in which each column represents the tracking result of one of the $q - 1$ frames processed so far. Inspired by successful work [20], in which online learning was considered as a batch processing method, we devise a dictionary learning by minimising several quadratic functions iteratively using $\ell^1$ fitting function:

$$\min_{A_t, X} \frac{1}{q} \sum_{i=1}^{q} \left\{ \| z_{i} - A_t X_i \|_1^2 + \tau \| X_i \|_1 \right\} \quad \text{(7)}$$

Where we start from initial dictionary and consequently update the dictionary and its robust sparse coding estimates $X$. The dictionary obtained by Eq. (7) is robust against outliers and needs a constant amount of memory as the number of data samples grows. Again, we use Iterative Reweighted Least Squares (IRLS) to estimate each dictionary item:

$$A(j, :) = \arg \min_{h} \sum_{i=1}^{h} w^j_i (z_{ij} - a X_i) \quad \text{(8)}$$

where $w^j_i$ is the $j$th weight of $j$th row of dictionary, $A(j, :)$, and $h$ is the batch size. By taking the derivative and setting it to zero, we obtain the following linear system:

$$B^j = D(j, :) C^j \quad \text{(9)}$$

where $B^j = \sum_{i=1}^{h} w^j_i z_{ij} X_i T^j$ and $C^j = \sum_{i=1}^{h} w^j_i X_i T^j$. We update the statistical parameters $B^j$ and $C^j$ for the next $h$ frames as $C^j = C^{j-1} + \sum_{t=1}^{m} w^j_i X_t T^j$ and $B^j = B^{j-1} + \sum_{t=1}^{m} w^j_i z_{ij} X_t T^j$. With this configuration, we update object templates $A^{(O)}$ in an online manner. To solve the linear system, we employ a conjugate gradient method, taking previous $A^{(O)}$ as initialisation in the current round. For the background templates, we update these in each mini-batch (5 frames in our experiments) from image regions away (e.g. more than 8 pixels) from the current tracking result.

Finally, the tracking result $z_t$ at time instance $t$ is the particle $y_t$ such that

$$i = \arg \max_{y_t=1,\ldots,N} \left( \| A^{(O)}_t W_{O} \|_1 - \| A^{(B)}_t W_B \|_1 \right) \quad \text{(10)}$$

which encourages the tracking result to be represented well by the object and not the background templates.

5. EXPERIMENTS AND DISCUSSION

Experimental Settings We evaluate our algorithm on a number of challenging benchmark sequences used in prior work: Woman from [1], Singer 1 from [13] and Girl from [3]. In addition, Car 11, Faceoc2, David and Caviar are publicly available online.\(^3\) We evaluate our proposed method using the centre location error as well as the overlap rate [21]. The latter is defined as $\frac{\text{area}(R_t \cap R_G)}{\text{area}(R_t \cup R_G)}$, where $R_t$ and $R_G$ denote the bounding box obtained by the ground-truth and a tracker, respectively. The results are shown in Tables 1 and 2, respectively. In Table 1, each row represents the average centre location errors of four algorithms tested on a certain video sequence. Na denotes a tracker lost the target for several frames. In this table, the quantitative evaluation of the SCM Visual tracker [22], MITT tracker [10], FragTrack method [1], VTD tracker [13], $l_1$ tracker [23], MIL tracker [2], IVT tracker[12] and our proposed tracker are presented. The parameters are presented as follows. Note that they are fixed for all sequences. The dictionary size $K$ is 250 including 50 object templates and 200 background templates where the template size is $32 \times 32$. The number of particles sampled in the particle filter is set to 100. $\lambda$ and $\gamma$ in Eq. (2) are set to 0.4, $\tau$ in Eq. (7) is set to 0.15 and the mini-batch size $h$ is 5.

5.1. Severe Occlusion

Occlusion is one of the most general yet critical challenges in object tracking. In fact, several tracker including the MIL tracking algorithm [2], the FragTrack method [1], the $l_1$ tracking method [23] and our tracker are developed to solve this problem. In the Caviar sequence, the target is occluded by two people at times and one of them is similar in color and shape to the target. For most template-based trackers, a simple update with the occluded part often leads to

\(^3\)http://www.cs.toronto.edu/~ross/ivt/

| Table 1. Average centre location errors (in pixels). |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Video Clip      | Frag            | IVT             | MIL             | $l_1$           | VTD             | MITT            | SCM             | Our             |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Singer1         | 22.0            | 8.5             | 15.2            | 4.6             | 4.1             | 3.8             | 1.5             | / Na            |
| Girl            | 18.1            | 48.5            | 32.3            | 62.5            | 21.5            | 4.9             | 9.8             | 5.6             |
| Car 11          | 64.0            | 2.2             | 43.5            | 33.3            | 27.1            | 1.7             | 1.8             | 1.3             |
| Faceoc2         | 15.5            | 10.2            | 14.1            | 11.1            | 10.4            | 8.9             | 4.8             | 4.1             |
| David           | 76.7            | 3.6             | 16.1            | 7.6             | 13.6            | 7.8             | 4.5             | 2.9             |
| Caviar          | 116.1           | 66.0            | 100.2           | 65.7            | 58.3            | 9.4             | 2.7             | 1.6             |
| Woman           | 113.6           | 167.5           | 122.4           | 131.6           | 136.6           | 257.5           | 2.5             | 1.6             |

\(^2\)We update the dictionary every $h = 5$ frames.
drifts. By designing the proposed dictionary learning, which is robust to the outliers, our tracker achieves stable performance throughout the entire sequence. Figure 1 shows tracking results for the Face sequence. Performance on this sequence exemplifies the robustness of our tracker to the heavy occlusion and partial illumination change. Our tracker, SCM [22] and MTT [10], are comparable in tracking the target during the entire sequence. The FragTrack [1] tracking method fails after frame #415. Due to the discriminative sparse nature of our tracking scheme, the confidence score gives higher weights to the candidates considered as positive samples and penalises the others. Therefore, the tracking result will not shift to the background.

The target in the Woman sequence undergoes heavy occlusion and we can see in Figure 1, from frame #112 onwards, the woman is occluded by the car. Our proposed tracker gives the most accurate results. Compared to MTT, the result obtained by our tracker based on the average overlap rate, is 4 times better. That is because unlike MTT, which assumes a common structure for all tasks (particles), we consider both group information as well as outlier penalties in our objective function.

### 5.2. Lighting Condition Changes

In the Car 11 sequence, the contrast between the target and the background is low. While most trackers drift to the cluttered background or other vehicles in the presence of drastic illumination variation, only our tracker and MTT [10] can successfully locate the correct object. In the Singer 1 sequence, while other trackers fail as they are not able to handle the large lighting change, our tracker achieved a better result in this sequence. In the David sequence, a person walks out of a dark conference room and into an area with spot lights. Only a few trackers are able to keep track of the target to the end and our proposed tracker performs with a lower error and higher success rate.

### 6. CONCLUSIONS

In our paper, we propose a joint sparsity-based model in a particle filter framework. Different from related existing methods that ignore the interrelated structures within input particles, we directly imposed a regularisation term with a mixture of structure and outlier norm which led to a flexible and robust formulation and let us to choose more discriminative features. To deal with appearance changes of the tracking target, we implement an adaptive dictionary learning by minimising several quadratic functions iteratively in an online manner. We have demonstrated the potential and the effectiveness of the proposed approach on various challenging video sequences with different tracking scenarios.
7. REFERENCES


