# Discriminative Tracking by Metric Learning

Xiaoyu Wang[1], Gang Hua[2], and Tony X. Han[1]

[1] Dept. of ECE, University of Missouri
[2] Nokia Research Center, Hollywood
xw9x9@mail.missouri.edu, ganghua@gmail.com, hantx@missouri.edu

**Abstract.** We present a discriminative model that casts appearance modeling and visual matching into a single objective for visual tracking. Most previous discriminative models for visual tracking are formulated as supervised learning of binary classifiers. The continuous output of the classification function is then utilized as the cost function for visual tracking. This may be less desirable since the function is optimized for making binary decision. Such a learning objective may make it not to be able to well capture the manifold structure of the discriminative appearances. In contrast, our unified formulation is based on a principled metric learning framework, which seeks for a discriminative embedding for appearance modeling. In our formulation, both appearance modeling and visual matching are performed online by efficient gradient based optimization. Our formulation is also able to deal with multiple targets, where the exclusive principle is naturally reinforced to handle occlusions. Its efficacy is validated in a wide variety of challenging videos. It is shown that our algorithm achieves more persistent results, when compared with previous appearance model based tracking algorithms.

## 1 Introduction

Appearance based visual tracking has been an active research topic for decades [1, 2, 3, 4, 5, 6, 7, 8]. There are two essential tasks: the *modeling* task builds an appearance model for the visual target; then the *matching* task matches the model with the source visual data to recover the motion of the target objects. Appearance models can roughly be put into two categories: *generative* models [2, 3, 4, 6] and *discriminative* models [7, 9, 8, 5].

Generative models seek a compact model to account for as much visual variations of the appearances as possible. Most often a set of training examples is leveraged either to obtain a subspace model [6, 2, 3] using embedding methods such as principle component analysis (PCA) [6, 3] or Gram-Schmidt decomposition [2], or to learn a Gaussian mixture model [1] using the Expectation-Maximization (EM) algorithm [10].

Discriminative models aim at differentiating the appearances of the visual targets from the background. Most previous works proposed to learn a binary classifier to differentiate the visual target from the background by using, for example, support vector machine (SVM) [7], Boosting [8], linear discriminant analysis [9], and multiple instance Boosting [5]. Compared to generative models, discriminative models may be more desirable for tracking due to the discrimination of foreground and background.

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part III, LNCS 6313, pp. 200–214, 2010.

After the classifier is learnt, most previous works utilize the continuous output of the classification function as the objective for visual matching and tracking. This may be less desirable since the classification functions are trained to be good mainly for making binary decision. In other words, they may not be able to well capture the manifold structure of the discriminative appearances, a vital factor for robust visual tracking.

Given the visual appearance model, different tracking algorithms [11, 12, 13, 14, 15, 16] come with different optimization paradigm for matching. They can largely be classified into two. The first class [11,12] takes a hypothesis generation and observation verification approach by probabilistic information fusion. Seminal works include Kalman filter, probabilistic data association filter (PDAF) [11], and particle filter [12].

However, both Kalman filter and PDAF [11] make the assumption that the visual observations of the target can be obtained in certain ways, which may not be satisfied in many cases. Although particle filter [12] eliminates this assumption by taking a direct verification approach, it needs sufficient number of particle hypotheses, and hence a lot of computation resources for good performance. It is even worse when dealing with high dimensional motions [17, 18]. This is why partitioned sampling [17] and importance sampling [18] are needed to effeciently utilize the limited particle budget.

The second class takes a direct optimization approach, where iterative gradient based search [13,15] is performed, or a linear program [14,16] is solved to obtain the tracking results. Compared to the first class of tracking algorithms, direct optimization [14, 13] usually does not make any additional assumptions about image observations, and the gradient based optimization can be performed efficiently with modern nonlinear program [19]. This renders them to be more applicable when certain assumptions do not hold or the computational resource is constrained.

We propose a unified discriminative visual tracking framework for both appearance modeling and visual matching. It is cast under a discriminative metric learning algorithm proposed by Globerson and Roweis [20]. In our formulation, appearance modeling is to identify a discriminative embedding, and visual matching performs an exemplar based regression on such a manifold w.r.t. the motion parameters. Both steps optimize the same objective function and are performed alternatively by efficient gradient search. Therefore, we achieve two tasks in an unified formulation.

Without requiring any additional efforts, our formulation can naturally deal with the discriminative modeling and visual matching of multiple targets. Due to the mutual discrimination of the multiple appearances, and the joint optimization of multiple motions in our model, our tracking algorithm naturally reinforces the *exclusive principle* [21]. Exclusive principle states that no two visual targets shall account for the same image observations, which is vital to handle cross occlusions, as manifested in [21].

Our unified formulation presents three benefits to previous works: firstly, it presents a unified discriminative formulation where appearances modeling and matching are optimizing the same objective function. Secondly, the unified discriminative formulation gracefully handles visual modeling and tracking of multiple targets where an exclusive principle is naturally reinforced. This makes it to be robust to occlusions occurring among the different visual targets. Thirdly, a principled criterion is derived from it to select the optimal set of visual examples for online learning and matching.

## 2 Discriminative Appearance and Motion Model

### 2.1 A Unified Formulation

We take a unified formulation for joint discriminative appearances modeling and visual matching. More formally, suppose we have a set of labeled training examples $\mathcal{X}_0 = \{\mathbf{x}_i \in \mathbb{R}^N, y_i\}_{i=1}^n$, where $y_i = 1$ means $\mathbf{x}_i$ is among the $n_1$ foreground samples, and $y_i = 0$ implies that $\mathbf{x}_i$ is one of the $n_0+1$ background samples, such that $n_1+n_0+1 = n$. In our experiments, each $\mathbf{x}_i$ is usually a $w \times h$ image patch and $N = w \times h$.

We further denote $\mathbf{I}(\mathbf{m})$ to be the visual target we would like to track where $\mathbf{m} \in \mathbb{R}^L$ is the motion parameters we want to recover. Obviously, the label $y$ of $\mathbf{I}(\mathbf{m})$ is 1, since it represents the visual target. For ease of notation, we denote $\mathbf{x}_0 = \mathbf{I}(\mathbf{m})$. Therefore, our final labeled data set $\mathcal{X} = \mathcal{X}_0 \cup \{(\mathbf{x}_0, y_0 = 1)\}$. Following Globerson and Roweis [20], we propose to learn a Mahalanobis form metric, i.e.,

$$d_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{A} (\mathbf{x}_i - \mathbf{x}_j). \tag{1}$$

where $\mathbf{A}$ is a positive semi-definite (PSD) matrix to be learnt. For each $\mathbf{x}_i \in \mathcal{X}$, define

$$p_{\mathbf{A}}(\mathbf{x}_j|\mathbf{x}_i) = \frac{1}{Z_i} e^{-d_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j)} = \frac{e^{-d_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j)}}{\sum_{k \neq i} e^{-d_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_k)}}. \tag{2}$$

The ideal distribution of the optimal $\mathbf{A}$ shall collapse samples from the same class to be a single point. Specifically, the ideal distribution shall take the following form,

$$p_0(\mathbf{x}_j|\mathbf{x}_i) = \begin{cases} \frac{1}{n_l} & y_i = y_j = l \\ 0 & y_i \neq y_j \end{cases}. \tag{3}$$

where $l \in \{0, 1\}$. Recall that $\mathbf{x}_0 = \mathbf{I}(\mathbf{m})$, we define

$$f(\mathbf{A}, \mathbf{m}) = \sum_{i=0}^{n} KL\left(p_0(\mathbf{x}_j|\mathbf{x}_i) || p_{\mathbf{A}}(\mathbf{x}_j|\mathbf{x}_i)\right) = C + \sum_{y_i = y_j = l} \frac{1}{n_l} \left(d_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j) + \log Z_i\right) \tag{4}$$

where $C = \sum_{y_i = y_j = l} \frac{1}{n_l} \log \frac{1}{n_l}$ is a constant. To have $p_{\mathbf{A}}(\mathbf{x}_j|\mathbf{x}_i)$ to be as close to $p_0(\mathbf{x}_j|\mathbf{x}_i)$ as possible, we only need to proceed to minimize $f(\mathbf{A}, \mathbf{m})$, i.e.,

$$\min f(\mathbf{A}, \mathbf{m}) \tag{5}$$

$$s.t.\ \forall \mathbf{a} \in \mathbb{R}^N, \mathbf{a}^T \mathbf{A} \mathbf{a} \geq 0. \tag{6}$$

where the constraint in Eq. 6 confines $\mathbf{A}$ to be PSD. Solving the above optimization problem would allow us to jointly obtain the optimal discriminative appearance model defined by $\mathbf{A}$, and track the motion of the target visual object, which is defined by $\mathbf{m}$. We solve both by efficient gradient based search, as presented in the following sections.

We shall emphasize here that we present our formulation and optimization in this section with a single visual target for ease of presentation. We will extend the discussion to present more details on how to deal with multiple objects tracking in Sec. 4.

### 2.2 Appearance Model Estimation

In our unified formulation, discriminative appearance modeling refers to identifying the optimal $\mathbf{A}$, which defines the discriminative metric, and thus a discriminative embedding. Assume that the motion parameter $\mathbf{m}$ is fixed, following [20], it is easy to figure out that $f(\mathbf{A}, \mathbf{m})$ is a convex function of $\mathbf{A}$. Taking the derivative of $f(\mathbf{A}, \mathbf{m})$ with respect to $\mathbf{A}$, we have

$$\frac{\partial f(\mathbf{A}, \mathbf{m})}{\partial \mathbf{A}} = \sum_{i,j=0}^{n} (p_0(\mathbf{x}_j|\mathbf{x}_i) - p_{\mathbf{A}}(\mathbf{x}_j|\mathbf{x}_i))(\mathbf{x}_j - \mathbf{x}_i)(\mathbf{x}_j - \mathbf{x}_i)^T. \tag{7}$$

Similar to [20], we take a gradient projection algorithm [22] to obtain the optimal $\mathbf{A}$. Specifically the following two steps are performed:

1. GRADIENT DESCENT:$\mathbf{A} = \mathbf{A} - \epsilon \frac{\partial f(\mathbf{A}, \mathbf{m})}{\partial \mathbf{A}}$, where $\epsilon$ determines the step length for gradient descent.
2. PSD PROJECTION: Compute the eigen-value decomposition of $\mathbf{A}$, i.e., $\{\lambda_k, \mathbf{u}_k\}_{k=1}^N$ such that $\mathbf{A} = \sum_{k=1}^N \lambda_k \mathbf{u}_k \mathbf{u}_k^T$, set $\mathbf{A} = \sum_{k=1}^N \max(\lambda_k, 0) \mathbf{u}_k \mathbf{u}_k^T$.

The first step above performs gradient descent, and the second step reinforces the constraint to make $\mathbf{A}$ to be a positive semi-definite matrix. These two steps are iterated until convergence. Since $f(\mathbf{A}, \mathbf{m})$ is a convex function of $\mathbf{A}$ given $\mathbf{m}$. The iteration of these two steps is guaranteed to find the optimal solution to $\mathbf{A}$.

### 2.3 Motion Parameter Optimization

In this subsection, we fix the discriminative appearance model $\mathbf{A}$, and develop the gradient descent search for the motion parameters $\mathbf{m}$. Not losing any generality, we assume that $\mathbf{m}$ is a linear motion model, i.e.,

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix} \tag{8}$$

where $[x', y']^T$ is the canonical coordinates for the labeled examples, and $[x, y]^T$ is the coordinates in the target video frame. This linear motion model covers a wide variety of motions such as translation, scaling, similarity, as well as full affine motion. We proceed to derive the gradient based search for the full affine motion model.

Recall that $\mathbf{x}_0 = \mathbf{I}(\mathbf{m})$ is the only term that involves the motion parameter $\mathbf{m}$, according to chain rule, we have

$$\frac{\partial f(\mathbf{A}, \mathbf{m})}{\partial \mathbf{m}} = \frac{\partial f(\mathbf{A}, \mathbf{m})}{\partial \mathbf{x}_0} \frac{\partial \mathbf{x}_0}{\partial \mathbf{m}}. \tag{9}$$

With some mathematical manipulation, it can be shown that

$$f(\mathbf{A}, \mathbf{m}) = \frac{1}{n_1} \sum_{y_j=1, j\neq 0} 2d_{\mathbf{A}}(\mathbf{x}_0, \mathbf{x}_j) + \sum_{j=0}^{n} \log Z_j + C(\mathbf{A}) \tag{10}$$

where $C(\mathbf{A})$ is a term which is independent of $\mathbf{x}_0$ and thus independent of $\mathbf{m}$. Therefore, with some more mathematical manipulations, we have

$$\frac{\partial f(\mathbf{A},\mathbf{m})}{\partial \mathbf{x}_0} = \frac{4}{n_1} \sum_{y_j=1, j\neq 0} \mathbf{A}(\mathbf{x}_0 - \mathbf{x}_j) - 2\sum_{j=1}^{n} (p_{\mathbf{A}}(\mathbf{x}_j|\mathbf{x}_0) + p_{\mathbf{A}}(\mathbf{x}_0|\mathbf{x}_j))\mathbf{A}(\mathbf{x}_0 - \mathbf{x}_j). \tag{11}$$

For any parameter $\xi \in \mathbf{m}$, again, applying chain rule, we have

$$\frac{\partial \mathbf{x}_0}{\xi} = \frac{\partial \mathbf{I}(\mathbf{m})}{\partial \xi} = \frac{\partial \mathbf{I}(\mathbf{m})}{x}\frac{\partial x}{\partial \xi} + \frac{\partial \mathbf{I}(\mathbf{m})}{y}\frac{\partial y}{\partial \xi}, \tag{12}$$

where $\frac{\partial \mathbf{I}(\mathbf{m})}{x}$ and $\frac{\partial \mathbf{I}(\mathbf{m})}{y}$ represents the image gradient in the target frame in horizontal and vertical directions, respectively. For ease of notation, we denote them as $\mathbf{I}_x$ and $\mathbf{I}_y$ respectively. Following Eq. 12, we have

$$\frac{\partial \mathbf{x}_0}{\partial a} = \mathbf{I}_x x', \frac{\partial \mathbf{x}_0}{\partial b} = \mathbf{I}_x y', \frac{\partial \mathbf{x}_0}{\partial c} = \mathbf{I}_y x', \frac{\partial \mathbf{x}_0}{\partial d} = \mathbf{I}_y y', \frac{\partial \mathbf{x}_0}{\partial e} = \mathbf{I}_x, \frac{\partial \mathbf{x}_0}{\partial f} = \mathbf{I}_y \tag{13}$$

Therefore, we may easily calculate the gradient of $f(\mathbf{A},\mathbf{m})$ with respect to $\mathbf{m}$ by applying Eq. 9 to Eq. 13. Then we can take a gradient descent step to recover the optimal motion parameter $\mathbf{m}$, i.e.,

$$\mathbf{m} = \mathbf{m} - \eta \frac{\partial f(\mathbf{A},\mathbf{m})}{\partial \mathbf{m}} \tag{14}$$

where the step length $\eta$ could be estimated, for example, by a quasi-Newton method such as L-BFGS [19].

## 3 Online Matching and Model Estimation

One of the main challenges in appearance model based visual tracking is to robustly adapt the model to the visual environment. This adaptation may be indispensable for robust tracking since the target objects may go through drastic visual changes from environmental conditions such as extreme lighting, occlusions, casting shadows, and pose and view changes. The unified formulation we proposed in Eq. 5 enables us to naturally fulfill this task. We proceed to present it in a more formal way.

Extended from the notation of Sec. 2, let $\mathcal{X}^{(t)}$ be the set of $n$ labeled examples we maintain at time instance $t$. We also let $\mathbf{A}_t$ be the current discriminative appearance model, and $\mathbf{m}_t$ be the motion parameters we need to recover. Hence we have $\mathbf{x}_0^{(t)} = \mathbf{I}^{(t)}(\mathbf{m}_t)$. At each time instant $t$, given $\mathcal{X}^{(t)}$ and $\mathbf{A}_t$, we run the gradient descent optimization algorithm outlined in Sec. 2.3 to obtain the optimal motion parameter $\mathbf{m}_t^\star$. This fulfills our visual matching and tracking task. Then we perturb $\mathbf{m}_t^\star$ to generate a set of $\alpha$ negative samples $\mathcal{X}_-^{(t+1)}$ to replace the oldest $\alpha$ negative sample subset $\mathcal{X}_-^{(t)}$ in $\mathcal{X}^{(t)}$. This results in the new labeled examples $\mathcal{X}^{(t+1)}$, i.e.,

$$\mathcal{X}^{(t+1)} = (\mathcal{X}_0^{(t)} \setminus \mathcal{X}_-^{(t)}) \cup \mathcal{X}_-^{(t+1)}. \tag{15}$$

Since $\mathbf{m}_t$ has been recovered, for ease of presentation, we abuse the notation to temporally define $\mathbf{x}_0^{(t+1)} = \mathbf{I}^t(\mathbf{m}_t)$. With $\mathcal{X}^{(t+1)}$ We can then run the gradient projection optimization algorithms outlined in Sec. 2.2 to obtain the optimal $\mathbf{A}_{t+1}$. To proceed with the next matching step to identify the optimal $\mathbf{I}^{t+1}(\mathbf{m}_{t+1})$, with a fixed memory budget, we need to retire one positive examples in the current $\mathcal{X}^{(t+1)}$, we propose a least consistent criterion based on the contribution of each positive examples to the unified cost function $f(\mathbf{A}_{t+1}, \mathbf{m}_t)$. Indeed, fixing $\mathbf{A}_{t+1}$ and $\mathbf{m}_t$, $f(\mathbf{A}_{t+1}, \mathbf{m}_t)$ is a function of $\mathcal{X}^{(t+1)}$, i.e., $f(\mathbf{A}_{t+1}, \mathbf{m}_t) = g(\mathcal{X}^{(t+1)})$. We can similarly define a $g(\cdot)$ function for any subset of $\mathcal{X}^{(t+1)}$ based on Eq. 4. Therefore, for each $\mathbf{x} \in \mathcal{X}^{(t+1)}$, a consistent criterion can be defined as

$$c(\mathbf{x}) = g\left(\mathcal{X}^{(t+1)}\right) - g\left(\mathcal{X}^{(t+1)} \setminus \{\mathbf{x}\}\right). \tag{16}$$

It is easy to understand that the larger $c(\mathbf{x})$ is, the more contribution $\mathbf{x}$ has made to $f(\mathbf{A}_{t+1}, \mathbf{m}_t)$. If the label $y(x) = 1$, a larger $c(\mathbf{x})$ indicates that $\mathbf{x}$ is not very compatible to the rest of the positive samples, and hence should be retired from the sample set. More formally, we select

$$\mathbf{x}^\star = \arg\max_{\mathbf{x} \in \mathcal{X}^{(t+1)}, y(\mathbf{x})=1} c(\mathbf{x}) \tag{17}$$

to retire from $\mathcal{X}^{(t+1)}$. In real operation, we only need to change the numbering of $\mathbf{x}_0^{(t+1)} = \mathbf{I}^t(\mathbf{m}_t)$ to the numbering of $\mathbf{x}^\star$, then we reset $\mathbf{x}_0^{(t+1)} = \mathbf{I}^{t+1}(\mathbf{m}_{t+1})$ which is unknown now to kick off the matching process for the optimal motion parameter $\mathbf{m}_{t+1}$.

The above steps will be repeated from time instant $t$ to time instant $t+1$. Therefore we track the visual target and estimate the discriminative visual appearance model simultaneously in an online fashion, which are all based on efficient gradient based optimization. Most previous approaches resort to heuristics or the oldness of visual samples to select the optimal set of online training examples. While our proposed selection criterion for positive examples in Eq. 17 is derived directly from our unified cost function in a principled fashion, an obvious benefit of our unified formulation.

To initialize the tracking algorithm, we can run an object detector if it applies, such as a face detector [23] or a human detector [24], if we are tracking a face or a person. Or we can request the users to manually specify a tracking rectangle in the first frame. Then the initialized tracking rectangle, either from a detector or manually specified, is perturbed to form the initial set of labeled examples $\mathcal{X}^{(1)}$. More specifically, perturbed rectangles with sufficient overlap with the initial rectangle are regarded as positive examples, while those perturbed rectangles which are deviated too much from the initial rectangles are deemed as negative examples. This bootstraps learning for the optimal discriminative appearance model $\mathbf{A}_2$, which is then adopted to obtain the optimal motion parameter $\mathbf{m}_2$. This processes will be repeated as described above.

Last but not least, when maintaining the labeled example set $\mathcal{X}^{(t)}$, we fix a small set of $\beta$ negative and $\beta$ positive examples extracted from the initialization frame in the set, i.e., we never replace them with new examples. This treatment is very important to keep some invariance to our discriminative appearance model and avoid it to be drifted too drastically in the visual tracking process, a trick which has been adopted also in previous work, such as [8].

## 4 Modeling and Tracking Multiple Objects

Our unified formulation is natural to handle the tracking of multiple targets. To see this, we assume $y_i = 0$ indicates background, and $y_i = 1, \ldots, K$ indicates each of the $K$ visual targets we are intending to track. Let $\mathcal{S}_0 = \{(\mathbf{x}_{0j}, y_{0j} = 0)\}_{j=0}^{n_0}$, and also let $\mathcal{S}_i = \{(\mathbf{x}_{ij}, y_{ij} = i)\}_{j=0}^{n_i}$ for any $i = 1, \ldots, K$, where $\forall i > 0$, $\mathbf{x}_{i0} = \mathbf{I}(\mathbf{m}_i)$ indicates each of the visual targets we want to track in the current frame, where $\mathbf{m}_i$ is represented by $\{a_i, b_i, c_i, d_i, e_i, f_i\}$, as defined in Eq. 8. Following similar steps as we have derived Eq. 4, denote $\mathcal{M} = \{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_K\}$, and $\mathbf{X} = \{\mathbf{x}_{i0}\}_{i=1}^{K}$ we have

$$f(\mathbf{A}, \mathcal{M}) = C + \sum_{i=0}^{K} \sum_{j \neq k=1}^{n_i} \frac{1}{n_i} \left( d_{\mathbf{A}}(\mathbf{x}_{ij}, \mathbf{x}_{ik}) + \log Z_{ij} \right). \tag{18}$$

where $Z_{ij}$ and $d_{\mathbf{A}}(\cdot, \cdot)$ are all defined similar to the corresponding terms defined in Sec. 2.1. Here $\mathbf{A}$ captures the discriminative appearances information for all the $K$ visual targets, and $\mathbf{m}_i$ represents the motion for the $i^{th}$ visual target which, in our experiments, are again the affine motion parameters defined in Eq. 8.

Following similar derivations as in Sec. 2.2 and Sec. 2.3, we can compute

$$\frac{\partial f(\mathbf{A}, \mathcal{M})}{\partial \mathbf{A}} = \sum_{i=0}^{K} \sum_{j=0}^{n_i} \sum_{k=0}^{K} \sum_{l=0}^{n_k} \omega_{ij}(kl)(\mathbf{x}_{kl} - \mathbf{x}_{ij})(\mathbf{x}_{kl} - \mathbf{x}_{ij})^T \tag{19}$$

where

$$\omega_{ij}(kl) = p_0(\mathbf{x}_{kl}|\mathbf{x}_{ij}) - p_{\mathbf{A}}(\mathbf{x}_{kl}|\mathbf{x}_{ij}). \tag{20}$$

With this formula to compute the gradient, we can utilize similar Gradient projection steps outlined in Sec. 2.2 to obtain the optimal $\mathbf{A}$. Notice that here $\mathbf{A}$ captures both the discriminative appearances among all the visual targets, as well as the discriminative information between the visual targets and background. Similarly, we obtain that

$$\frac{\partial f(\mathbf{A}, \mathcal{M})}{\partial \mathbf{x}_{i0}} = \frac{4}{n_i} \sum_{j=1}^{n_i} \mathbf{A}(\mathbf{x}_{i0} - \mathbf{x}_{ij}) - 2 \sum_{k=1}^{K} \sum_{l=0}^{n_k} \beta_{i0}(kl) \mathbf{A}(\mathbf{x}_{i0} - \mathbf{x}_{kl}). \tag{21}$$

where

$$\beta_{i0}(kl) = p_{\mathbf{A}}(\mathbf{x}_{kl}|\mathbf{x}_{i0}) + p_{\mathbf{A}}(\mathbf{x}_{i0}|\mathbf{x}_{kl}) \tag{22}$$

Following Eq. 13, we also have

$$\frac{\partial \mathbf{x}_{i0}}{\partial a_i} = \mathbf{I}_x x_i', \frac{\partial \mathbf{x}_{i0}}{\partial b_i} = \mathbf{I}_x y_i', \frac{\partial \mathbf{x}_{i0}}{\partial c_i} = \mathbf{I}_y x_i', \frac{\partial \mathbf{x}_{i0}}{\partial d_i} = \mathbf{I}_y y_i', \frac{\partial \mathbf{x}_{i0}}{\partial e_i} = \mathbf{I}_x, \frac{\partial \mathbf{x}_{i0}}{\partial f_i} = \mathbf{I}_y. \tag{23}$$

Following chain rules and with Eq. 23, we can easily calculate

$$\frac{\partial f(\mathbf{A}, \mathbf{m}_i)}{\mathbf{m}_i} = \frac{\partial f(\mathbf{A}, \mathcal{M})}{\partial \mathbf{x}_{i0}} \frac{\partial \mathbf{x}_{i0}}{\mathbf{m}_i} \tag{24}$$

With Eq. 24, again, we use L-BFGS [19] to solve the nonlinear optimization problem to obtain each set of motion parameters $\mathbf{m}_i$ for the $i^{th}$ visual target. Based on the above

two gradient based optimization schemes for $\mathbf{A}$ and each $\mathbf{m}_i$, respectively, following similar ideas as outlined in Sec. 3, we can further develop online appearances modeling and updating algorithms and visual matching algorithms for robust visual tracking of multiple objects. We shall not verbose on it since it follows quite similar steps as those outlined in Sec 3.

### 4.1 Discriminant Exclusive Principle

We argue that the proposed joint formulation for multiple object tracking naturally incorporates an exclusive principle [17] in the matching process. Therefore it is robust to handle occlusions among the different visual objects. The exclusive principle states that no two visual tracker shall occupy the same image observation. Our proposed algorithm naturally achieves it because of the joint discriminative appearance model $\mathbf{A}$, which reinforces the mutual discrimination of the appearances between two visual targets $\mathbf{I}(\mathbf{m}_i)$ and $\mathbf{I}(\mathbf{m}_j)$. To see this more clearly, given an optimal $\mathbf{A}$, if $\mathbf{I}(\mathbf{m}_i)$ and $\mathbf{I}(\mathbf{m}_j)$ occupy similar image regions (a.k.a, $\mathbf{m}_j \doteq \mathbf{m}_i$), and thus have similar appearance, the mutual discriminative information encoded in $\mathbf{A}$ would incur a large value for $f(\mathbf{A}, \mathcal{M})$. Therefore, $\mathbf{m}_j \doteq \mathbf{m}_i$ is not an optimal solution to $\mathcal{M}$. In other words, the optimal motion parameter $\mathcal{M}$ is more likely to occur when $\forall 1 \leq i < j \leq K, \mathbf{m}_j \neq \mathbf{m}_i$. Therefore, the exclusive principle among the different visual targets is naturally reinforced.

## 5 Experiments

We dub the name *TUDAMM* to the Tracker with Unified Discriminative Appearance Modeling and Matching (TUDAMM). Comparing with the results of other state-of-the-art trackers [2, 13], we evaluate our TUDAMM using several challenging video sequences including video clips from CAVIAR dataset [25], and other real-world video sequences downloaded from Internet.

### 5.1 Evaluation Criteria

Enlightened by the simplicity and the elegance of the Average Precision (AP) criterion used in the PASCAL grand challenge [26] for object detection evaluation, we define a simple measure for tracker evaluation, namely Average Tracking Precision (ATP). More formally, for each tracking task, a ground truth mask for the object of interest is labeled in each frame $j$. The mask is represented as a point set $\mathcal{G}_j$. The tracking result is represented as a point set $\mathcal{T}_j$ at frame $j$. $(x_i, y_i) \in \mathcal{G}_j$ or $\mathcal{T}_j$ indicates that the pixel at $(x_i, y_i)$ is inside them. For an ideal tracker, $\forall i, \mathcal{G}_j = \mathcal{T}_j$.

For each frame $j$, the tracking precision $r_j$ is defined as: $r_j = |\mathcal{G}_j \cap \mathcal{T}_j| / |\mathcal{G}_j \cup \mathcal{T}_j|$. Noticing that $r_j \in [0, 1]$, the ATP for a tracker of an object in a video clip is defined as:

$$ATP = \frac{1}{N} \sum_{j=1}^{N} r_j = \frac{1}{N} \sum_{j=1}^{N} \frac{|\mathcal{G}_j \cap \mathcal{T}_j|}{|\mathcal{G}_j \cup \mathcal{T}_j|}, \tag{25}$$

where $N$ is the running length of the video clips in frame number. For an ideal tracker, $ATP \equiv 1$. We use it as the exclusive quantitative measure to compare the performance of the TUDAMM with other state-of-the-art trackers.

**Fig. 1.** The sample key frames of the tracking results for CAVIAR dataset. Key frame NO. $443, 455, 467, 488, 501, 772$ are shown from left to right. First row: TUDAMM. Second Row: Meanshift [13]. Third row: Incremental Learning Tracker (ILT) [3].

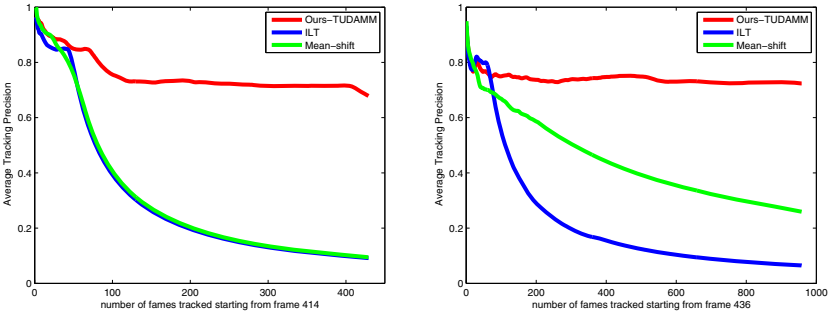
### 5.2 Visual Tracking of Single/Multiple Target(s)

We firstly present the tracking results of TUDAMM for single target on a video sequence from the CAVIAR dataset[1], where three persons are walking in the corridor of a shopping mall in Portugal. We call this video sequence "ThreePerson". We run the proposed tracking algorithm to track one of the three persons individually. The tracking task is challenging in several aspects: 1) the scales of the visual targets change drastically; 2) the three persons walked across each other and thus induced occlusion; 3) some other crossing person occluded the target person.

As shown in the first row of Fig. 1, the TUDAMM tracker successfully tracked the target person from beginning of the sequence to the end of the sequence without any problem, which is more robust than both the mean-shift tracker [13] (second row) and the incremental PCA tracker [3] (third row). Both of these algorithms failed to track the target after the person with red cloth occluded the target person, as displayed in the second and third row of Fig. 1. The robustness of our TUDAMM tracker attributes to our unified discriminative formulation, which makes it more robust to background clutter. For detailed video results, please check out our video demo file "http://vision.ece.missouri.edu/demo/ECCV2010Tracking.avi".

**Quantitative comparisons to other work.** Since the ThreePerson video in the CAVIAR dataset has ground-truth labels of the bounding boxes of the walking persons in the video sequence, we use the ATP criterion presented in Sec. 5.1 to quantitatively evaluate the performances of the proposed TUDAMM tracker, the mean-shift tracker (Meanshift) [13], and the incremental PCA tracker (ILT) [3]. We present two such evaluation results for tracking two different persons in the video in Fig. 2(a) and Fig. 2(b), respectively. It is clear TUDAMM consistently presents more accurate tracking results than

[1] Data set from EC Funded CAVIAR project/IST 2001 37540, downloaded at URL: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.




(a) (b)

**Fig. 2.** (a) The performance comparison for the person tracked in figure 1. (b) The performance comparison for tracking the black person at right to the red person at the starting frame.

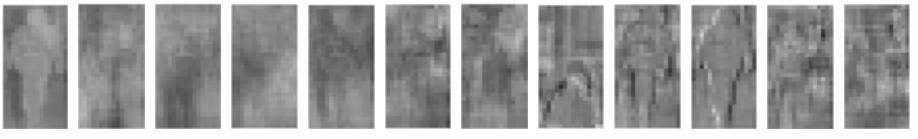


**Fig. 3.** The top 12 eigenvectors (with the descent order from left to right) for the discriminative matrix $A$

the other algorithms, which achieves an average tracking ATP of $75\%$. This demonstrates the good performance of the gradient based matching algorithm to recover the motion parameters.

**Visualizing the appearance model A.** As a matter of fact, the appearance model $\mathbf{A}$ defines a discriminative embedding to differentiate the visual object from the background. Each eigenvector of $\mathbf{A}$ is corresponding to one basis vector of the embedding. To have a better understanding of how the appearance model $\mathbf{A}$ functions, in Fig. 3, we visualize the top $12$ eigenvectors of an optimal $\mathbf{A}$ estimated at frame $436$ when tracking the person in red in the ThreePerson sequence. As we can clearly observe, these eigenvectors focusing on extracting the contour and thus encode the shape information of the target person. They also tend to focus more on features inside the human contour while suppress features outside the human contour. This indicates that our metric learning framework really picks up the discriminative information for tracking.

**Visualizing the gradient optimization processes.** To gain a good understanding of the gradient optimization process of both the discriminative appearance estimation as well as the gradient based optimization process for visual matching, we visualize the evolution of both optimization processes in frame 532 of the ThreePerson sequence, as shown in Fig. 4 and Fig. 5, respectively. The tracking target is the rightmost person in this frame. Fig. 4 visualizes how the tenth eigen-vector of the discriminative model

**Fig. 4.** The evolution of the tenth Eigen vector of $\mathbf{A}$ during gradient optimization in the first 11 steps of gradient descent from left to right

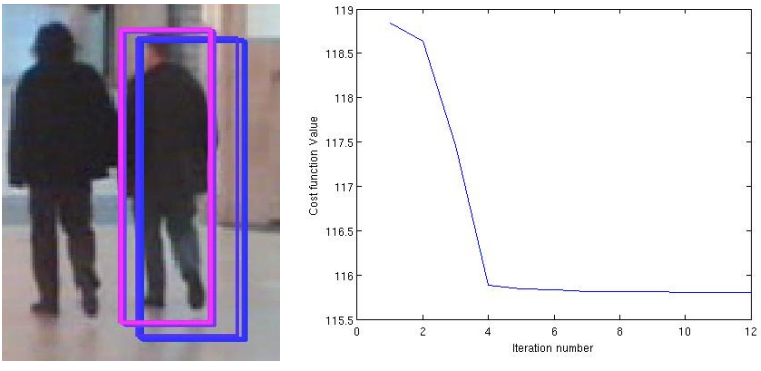



**Fig. 5.** The gradient optimization of objective function w.r.t. the motion parameters in frame 532 in the CAVIAR sequence. The tracking is initialized as the tracking result in frame 512 for better visualization. Red rectangle in the left image is the final converged matching results. The figure on the right displays how the objective function is minimized by gradient descent.

$\mathbf{A}$ evolves in the first 11 iterations. We start the optimization by initializing $\mathbf{A}$ as an identity matrix, so the initialization of the tenth eigen-vector is a unit vector with the tenth element to be one and all the other elements are zero, as shown in the first image in Fig. 4. As we can clearly observe, only after 8 steps of gradient descent the eigen-vector has already been stabilized. From Fig. 5, we can clearly observe the effectiveness of the gradient optimization process in the visual matching step. In only 4 steps of gradient descent, the matching result is already converged. These figures demonstrate the efficiency of the proposed gradient optimization process.

**Tracking under various visual variations.** We have also extensively tested the TUDAMM with other challenging videos used in previous works or downloaded from YouTube with various challenging aspects. We highly recommend to check our demo video for more details of all the tracking results.

More specifically, in Fig. 6, we present the tracking results of a human face from the TUDAMM, the ILT [3], and the Meanshift trackers [13], respectively. The ILT tracker [3] firstly reported results in this video, which is subject to drastic illumination changes and casting shadows. As we can clearly observe in Fig. 6, the TUDAMM

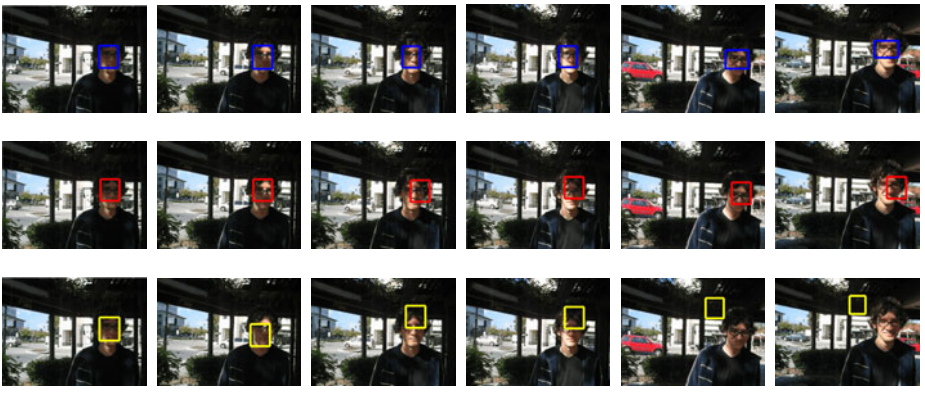


**Fig. 6.** The sample key frames of the tracking results on the challenging face moving under shadow with big illumination change video . Key frame NO. 201, 210, 220, 230, 240, 260 are shown from left to right. First row: TUDAMM. Second Row: Meanshift. Third row: ILT.

**Fig. 7.** The sample key frames of the CrazyCarChasing tracking results of TUDAMM with large scale zooming and camera motion.

robustly tracked the human face despite the dramatic shadows and illumination changes. While both the ILT tracker and the Meanshift tracker failed with the drastic visual variations. The results video contains 71 frames.

In Fig. 7, we report the tracking results of TUDAMM on a car chasing video downloaded from YouTube. The video is subject to large scale change and drastic camera motion since it was taken from a helicopter. Our tracking algorithm successfully tracked the motion of the target car without any problem. The results video contains 578 frames. In Fig. 8, we present the tracking results of a rabbit which underwent a lot of non-rigid motions. TUDAMM successfully tracked the rabbit across the video, which contains 156 frames.

**Tracking multiple targets with cross occlusion.** To demonstrate the ability of TUDAMM in dealing with occlusions in multiple object tracking, we report results in two video sequences, one is the ThreePerson video from the CAVIAR dataset, and the other is a horse racing video downloaded from YouTube. Tracking results in sample video frames are displayed in Fig 9 and Fig 10, respectively. Three people are tracked in the CAVIAR video, while five horse racers are tracked in the horse racing video. As we can clearly observe, despite severe cross occlusion among the different visual targets, our TUDAMM tracked all of them without any problem. This is attributed to the discriminative appearance model induced from our unified discriminative formulation.

**Fig. 8.** The sample key frames of the tracking results by TUDAMM on the RabbitRun video with nonrigid motion

**Fig. 9.** The sample key frames of the tracking results by multiple target TUDAMM on the CAVIAR dataset

**Tracking speed.** Last but not least, with a PC of 2.3-GHz CPU in Windows XP, without any code optimization in our C++ implementation, our tracker runs at 2 frames per second for tracking a single target. It runs at 0.5 frames per second for tracking the three people and 0.2 frames per second for tracking the 5 horses. We expect to have 10 times speed up with reasonable efforts on code optimization.

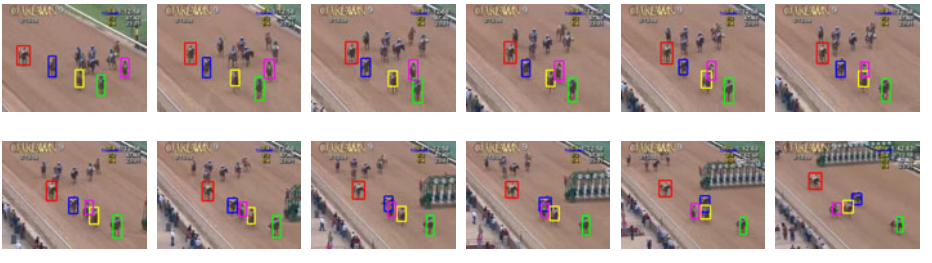


**Fig. 10.** Multiple Tracking results for a horse racing video. The order of the video frame is presented from top-left to bottom-right.

## 6 Conclusion and Future Work

In this paper, we present a unified discriminative framework based on metric learning for robust tracking of either single or multiple targets, where both the appearance modeling and visual matching are optimizing a single objective with efficient gradient based search. Our experimental results validate the efficacy of the proposed tracking algorithm. When tracking multiple targets, our unified formulation encodes an exclusive principle which naturally deals with cross occlusions among the multiple targets. This has also been manifested in our experiments. Future research includes exploring means of integrating our multiple target tracker with state-of-the-art surveillance systems to handle the appearance of new targets and disappearance of old targets.

## References


1. Jepson, A.D., Fleet, D.J., El-Maraghi, T.F.: Robust online appearance models for visual tracking. In: CVPR, vol. 1, pp. 415–422 (2001)
2. Ho, J., Lee, K.C., Yang, M.H., Kriegman, D.: Visual tracking using learned subspaces. In: CVPR, vol. 1, pp. 782–789 (2004)
3. Lim, J., Ross, D., Lin, R.S., Yang, M.H.: Incremental learning for visual tracking. In: NIPS, pp. 801–808 (2005)
4. Yang, M., Wu, Y.: Tracking non-stationary appearances and dynamic feature selection. In: CVPR (2005)
5. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: CVPR (2009)
6. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 484–498. Springer, Heidelberg (1998)
7. Avidan, S.: Support vector tracking. In: CVPR (2001)
8. Avidan, S.: Ensemble tracking. In: CVPR (2005)
9. Collins, R.T., Liu, Y.: On-line selection of discriminative tracking features. In: ICCV, vol. 1, pp. 346–352 (2003)
10. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society Series B 39, 1–38 (1977)
11. Bar-Shalom, Y.: Tracking and data association. Academic Press Professional, Inc., San Diego (1987)
12. Isard, M., Blake, A.: Contour tracking by stochastic propagation of conditional density. In: Buxton, B.F., Cipolla, R. (eds.) ECCV 1996. LNCS, vol. 1065, pp. 343–356. Springer, Heidelberg (1996)
13. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: CVPR, vol. 2, pp. 142–149 (2000)
14. Hager, G.D., Dewan, M., Stewart, C.V.: Multiple kernel tracking with ssd. In: CVPR, vol. 1, pp. 790–797 (2004)
15. Zhao, Q., Brennan, S., Tao, H.: Differential emd tracking. In: ICCV (2007)
16. Wu, Y., Fan, J.: Contextual flow. In: CVPR (2009)
17. MacCormick, J., Isard, M.: Partitioned sampling, articulated objects, and interface-quality hand tracking. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 3–19. Springer, Heidelberg (2000)
18. Wu, Y., Hua, G., Yu, T.: Tracking articulated body by dynamic markov network. In: ICCV, p. 1094 (2003)
19. Zhu, C., Byrd, R.H., Lu, P., Nocedal, J.: Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. ACM Transaction Mathematical Software 23, 550–560 (1997)
20. Globerson, A., Roweis, S.T.: Metric learning by collapsing classes. In: NIPS (2005)
21. MacCormick, J., Blake, A.: A probabilistic exclusion principle for tracking multiple objects. In: ICCV, pp. 572–587 (1999)
22. Rosen, J.B.: The gradient projection method for nonlinear programming. part i. linear constraints. Journal of the Society for Industrial and Applied Mathematics 8, 181–217 (1960)
23. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vision 57, 137–154 (2004)
24. Wang, X., Han, T.X., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: ICCV (2009)

25. Ribeiro, H.N., Hall, D., et al.: Comparison of target detection algorithms using adaptive background models. In: Proc. 2nd Joint IEEE Int. Workshop on Visual Surveillance, pp. 113–120 (2005)
26. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2009 (VOC 2009) Results (2009), http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2009/