A BI-SUBSPACE MODEL FOR ROBUST VISUAL TRACKING

Jialue Fan *Ming Yang* *Ying Wu*

EECS Department, Northwestern University
2145 Sheridan Road, Evanston, IL 60208
{jfa699, mya671, yingwu}@ece.northwestern.edu

ABSTRACT

The changes of the target's visual appearance often lead to tracking failure in practice. Hence, trackers need to be adaptive to non-stationary appearances to achieve robust visual tracking. However, the risk of adaptation drift is common in most existing adaptation schemes. This paper describes a bi-subspace model that stipulates the interactions of two different visual cues. The visual appearance of the target is represented by two interactive subspaces, each of which corresponds to a particular cue. The adaption of the subspaces is through the interaction of the two cues, which leads to robust tracking performance. Extensive experiments show that the proposed approach can largely alleviate adaptation drift and obtain better tracking results.

Index Terms— visual tracking, motion analysis

1. INTRODUCTION

Visual tracking is important for video analysis. It plays a critical role in many emerging applications such as video surveillance and vision-based interfaces. Although many tracking methods have been proposed and investigated [1, 2, 3, 4], there are still enormous challenges in the real situations for long duration tracking in unconstrained environments. One difficulty in practice is that the visual appearance of the target may undergo unpredicted changes for many reasons such as view changes, illumination changes, and partial occlusions. Such non-stationary visual appearances jeopardize visual measurements and lead to tracking failure.

There are two general approaches to overcome this difficulty. One is to find invariant features to the changes [5], and the other is to adapt the tracker to the changes [6]. As the visual invariants are in general very difficult to obtain, adaptation-based methods tend to be more flexible, because the appearance models are adaptive or the features used for tracking can be adaptively selected. However, the risk of adaptation is the model drift, and the appearance model may adapt to the distracters and lose track. This challenge confronts a large variety of adaptation-based visual trackers.

In most existing adaptation schemes, the data used for adapting the model at time t are those that are identified by the old model at time $t-1$ (e.g., use those that are closest to the old model). If no additional constraints are enforced, the model tends to best fit any new data regardless. If the new data used for adaption are from distracters, the model adapts to the distracters and drifts away. In visual tracking, the best match at t found by the appearance model at time $t-1$ does not necessarily to be the target, because of the changes of the visual appearance. Thus, to reduce the risk of adaptation drift, good constraints or supervision that are independent to the old model are needed. For example, positive and negative data that are from segmenting and clustering of the current image can be used to constrain model updating [7].

This paper describes a new model adaptation method based on the interaction between two visual cues or modalities (e.g., color appearance and texture appearance). The visual appearance of the target is represented by two feature subspaces, each of which captures the uncertainty in one visual cue. We propose a bi-subspace model to stipulate the interaction between these two subspaces. The new data identified by one model are used as supervision to update the other, and vice versa. The co-training between the two subspace models exchanges information from one visual cue to the other. It is the coupling of the two feature subspaces that leads to robust model adaptation.

2. SUBSPACE APPEARANCE MODEL

The target is tracked or detected when its visual measurements match the visual appearance model. The visual appearance of an object lies on a manifold in the corresponding feature space. Depending on the features used to describe the target and on the uncertainty of appearances, such a manifold can be quite nonlinear and complex. In addition, in real applications, we may not be able to learn this appearance manifold off-line, when we need a general purpose tracker. Instead, we have to recover and update the appearance manifolds on-the-fly during tracking [8]. Therefore, we make a reasonable assumption that the manifold at a short time interval is linear [6, 9]. The appearances $z \in \Re^m$ lie in a linear subspace $\mathbf{L}$ spanned by r linearly independent basis $\mathbf{A} \in \Re^{m \times r}$. So z is a linear combination of the basis $\mathbf{A}$ and we have $z = \mathbf{A}y$, where y are the components or projections.

The projection of z to the subspace $\Re^r$ is obtained by the least squares solution of $z = \mathbf{A}y$, i.e.,

$$y = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T z = \mathbf{A}^\dagger z \tag{1}$$

where $\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is the pseudo-inverse of $\mathbf{A}$. The reconstruction of the projection in $\Re^m$ is given by:

$$\bar{z} = \mathbf{A}\mathbf{A}^\dagger z = \mathbf{P}z \tag{2}$$

where $\mathbf{P} = \mathbf{A}\mathbf{A}^\dagger \in \Re^{m\times m}$ is the *projection matrix*, which is unique for a subspace regardless of the basis. Its *orthogonal complement subspace* is characterized by $\mathbf{P}^\perp = \mathbf{I} - \mathbf{P}$.

In tracking, we denote by $x(t)$ the location (i.e., the motion) of the target at time t, and by $z(x(t))$ the corresponding visual appearance feature vector of the motion hypothesis $x(t)$. If the appearance model $\mathbf{P}(t)$ is given (and not updating), tracking can be done by finding the best match:

$$\begin{aligned} x^*(t) &= \arg\min_{x(t)} E[\|z(x(t)) - \mathbf{P}(t)z(x(t))\|^2] \\ &= \arg\min_{x(t)} E[\|\mathbf{P}(t)^\perp z(x(t))\|^2] \end{aligned} \tag{3}$$

Similarly, if we know the target's true location $x(t)$, we can find the best appearance model by doing:

$$\mathbf{P}^*(t) = \arg\min_{\mathbf{P}(t)} E[\|\mathbf{P}(t)^\perp z(x(t))\|^2]$$

But when $x(t)$ and $\mathbf{P}(t)$ are both unknown, we cannot estimate them simultaneously by simply doing:

$$\{\mathbf{P}^*(t), x^*(t)\} = \arg\min_{\mathbf{P}(t),x(t)} E[\|\mathbf{P}(t)^\perp z(x(t))\|^2]$$

because for any tracking result, good or bad, one can always find a subspace to update.

To break this dilemma, outside supervision and constraints are necessary. Denote by $z(x^+(t))$ and $z(x^-(t))$ the positive and negative data at time t, respectively. They are given as the outside supervision. The appearance subspace should be updated such that it is close to the old one and the positive data, but far away from the negative data:

$$\begin{aligned} \mathbf{P}^*(t) &= \arg\min_{\mathbf{P}(t)}\{E[\|\mathbf{P}(t)^\perp z(x^+(t))\|^2] \\ &- E[\|\mathbf{P}(t)^\perp z(x^-(t))\|^2] + \alpha\|\mathbf{P}(t) - \mathbf{P}(t-1)\|_F^2\} \end{aligned}$$

where $\alpha > 0$ is a weighting factor. Denote by $\mathbf{C}^+(t) = E[z(x^+(t))z(x^+(t))^T]$ and $\mathbf{C}^-(t) = E[z(x^-(t))z(x^-(t))^T]$. Equivalently, we have:

$$\begin{aligned} \mathbf{P}^*(t) &= \arg\min_{\mathbf{P}(t)}\{tr(\mathbf{P}(t)\mathbf{C}^-(t)) - tr(\mathbf{P}(t)\mathbf{C}^+(t)) \\ &\quad +\alpha\|\mathbf{P}(t) - \mathbf{P}(t-1)\|_F^2\} \end{aligned} \tag{4}$$

where $tr(\cdot)$ denotes the trace of a matrix.

Here we employ an iterative algorithm [7] to solve the above subspace fitting problem. A critical issue is: how to obtain $\{z(x^+(t))\}$ and $\{z(x^-(t))\}$ in the set of unlabeled data in the current image frame?

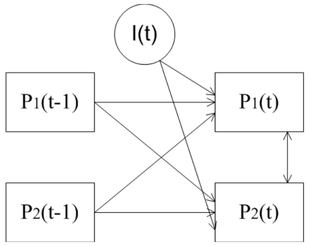

Fig. 1. The bi-subspace model.

3. BI-SUBSPACE APPEARANCE MODEL

3.1. Bi-subspace appearance model

As the appearance changes can be quite dynamic, the appearance subspace can be quite different from time to time. Because the visual appearance has multiple cues, e.g., color and texture, tracking is only feasible when at least one cue is stable for matching. Motivated by this, we propose a bi-subspace appearance model for two interactive cues, where one cue serves as the outside supervision for updating the other, and vice versa. The interaction between the cues gives a robust updating of two appearance subspaces.

We denote the image and the two appearance subspaces at time t by $I(t)$, $\mathbf{P}_1(t)$ and $\mathbf{P}_2(t)$, respectively. The bi-subspace model is illustrated in Figure 1. In tracking, at the current time t, the previous appearance subspaces $\mathbf{P}_1(t-1)$ and $\mathbf{P}_2(t-1)$ are given. They can be used to find matches in the current image frame $I(t)$. Based on $\mathbf{P}_1(t-1)$, $\mathbf{P}_2(t-1)$, and $I(t)$, we need to track the target by estimating $x(t)$, as well as updating the appearance subspaces $\mathbf{P}_1(t)$ and $\mathbf{P}_2(t)$.

The two subspaces are interactive, i.e., the updating on one shall influence the other. Thus, the updating of $\mathbf{P}_1(t)$ is conditioned on $\mathbf{P}_1(t-1)$ for smoothness, on $\{\mathbf{P}_2(t-1), I(t)\}$ that provides outside supervision, and on $\mathbf{P}_2(t)$ that gives interaction. The same thing is applied to $\mathbf{P}_2(t)$.

Even if the two modalities are independent *a priori*, they become correlated *a posteriori* when the observation is given [1]. This conditional dependency suggests the interaction between the two modalities. The interaction between the two subspaces results in the iteration between the updating of $\mathbf{P}_1(t)$ and $\mathbf{P}_2(t)$. This iteration converges to a fixed-point when neither subspace changes.

3.2. Interactions between subspaces

At a given motion hypothesis $x(t)$, we extract two types of appearance features, denoted by $z_1(x(t)) \in \Re^m$ and $z_2(x(t)) \in \Re^n$, respectively. At time t, the old appearance model $\mathbf{P}_1(t-1)$ is used to collect a set of "good matches" in the image $I(t)$:

$$D_1^+ = \{x_1^+(t)|d_1(z_1(x_1^+(t)), \mathbf{P}_1(t-1)) < \theta_1\} \tag{5}$$

where $d_1(z_1(x_1^+(t), \mathbf{P}_1(t-1))$ is a distance between $z_1(x_1^+(t))$ and appearance subspace $\mathbf{P}_1(t-1)$, and θ_1 is a threshold.

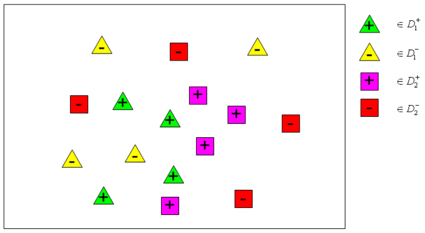

Fig. 2. Illustration of the four training data sets.

Since D_1^+ generates good matches based on cue 1, it can be treated as the set of candidates for positive supervision for updating the other cue. We also collect a set of "bad matches" based on $\mathbf{P}_1(t-1)$, denoted by D_1^-:

$$D_1^- = \{x_1^-(t) | d_1(z_1(x_1^-(t)), \mathbf{P}_1(t-1)) \geq \theta_1\} \qquad (6)$$

D_1^- generates bad matches for cue 1, and it will be used to form negative supervision in updating $\mathbf{P}_2(t)$, as will be described shortly. Similarly, for the second cue, we have:

$$D_2^+ = \{x_2^+(t) | d_2(z_2(x_2^+(t)), \mathbf{P}_2(t-1)) < \theta_2\} \qquad (7)$$
$$D_2^- = \{x_2^-(t) | d_2(z_2(x_2^-(t)), \mathbf{P}_2(t-1)) \geq \theta_2\} \qquad (8)$$

where $d_2(\cdot, \cdot)$ is the distance and θ_2 is the threshold in subspace $\mathbf{P}_2$. We illustrate an example of these positive and negative data in Figure 2. In practice, to reduce the computation, the collection of these four training data sets can be performed in a smaller image region based on motion prediction.

Now we describe how we identify the positive and negative data from these four training data sets. We denote the optimal positive data by $x^{+*}(t)$, then $z_1(x^{+*}(t))$ and $z_2(x^{+*}(t))$ should be very close to the corresponding appearance subspaces, i.e.,

$$\begin{aligned} x^{+*}(t) &= \arg\min_{x(t) \in D_1^+ \cup D_2^+} \{w_1 d_1(z_1(x(t)), \mathbf{P}_1(t-1)) + \\ &\qquad w_2 d_2(z_2(x(t)), \mathbf{P}_2(t-1))\} \end{aligned} \qquad (9)$$

where w_1 and w_2 are weights based on the variance of $\{d_1(z_1(x(t)), \mathbf{P}_1(t-1))\}$ and $\{d_2(z_2(x(t)), \mathbf{P}_2(t-1))\}$.

The optimal negative data should be selected carefully. Because if the negative data are too far from the appearance subspace, they are not quite informative for classification and useless in updating. Hence, we select them based on the "min-max" principle. Denote the optimal negative data for $\mathbf{P}_1(t)$ by $x_1^{-*}(t)$, which gives the closest match in $\mathbf{P}_1(t-1)$ among the bad matches in $\mathbf{P}_2(t-1)$ (i.e., the best one in D_2^-):

$$x_1^{-*}(t) = \arg\min_{x(t) \in D_2^-} d_1(z_1(x(t)), \mathbf{P}_1(t-1)) \qquad (10)$$

This is reasonable because (1) $x_1^{-*}(t) \in D_2^-$, so it is negative data; (2) $x_1^{-*}(t)$ is the most informative among D_2^-. Similarly, we find the optimal negative data $x_2^{-*}(t)$ for $\mathbf{P}_2(t)$:

$$x_2^{-*}(t) = \arg\min_{x(t) \in D_1^-} d_2(z_2(x(t)), \mathbf{P}_2(t-1)) \qquad (11)$$

Then we use $z_1(x^{+*}(t))$ and $z_1(x_1^{-*}(t))$ to update $\mathbf{P}_1(t)$, and use $z_2(x^{+*}(t))$ and $z_2(x_2^{-*}(t))$ to update $\mathbf{P}_2(t)$, using Eq. 4.

In order to obtain a stable solution, the updating process should iterate between two subspaces. We replace $\mathbf{P}_1(t-1)$ and $\mathbf{P}_2(t-1)$ with $\mathbf{P}_1(t)$ and $\mathbf{P}_2(t)$, and obtain the new positive and negative data with respect to $\mathbf{P}_1(t)$ and $\mathbf{P}_2(t)$. We denote the new optimal positive and negative data by $x^{+*(2)}(t)$, $x_1^{-*(2)}(t)$ and $x_2^{-*(2)}(t)$, and denote the new subspace by $\mathbf{P}_1^{(2)}(t)$ and $\mathbf{P}_2^{(2)}(t)$. Similarly, we can obtain $\mathbf{P}_1^{(3)}(t), \mathbf{P}_2^{(3)}(t)$, ... We stop the iteration when

$$\|\mathbf{P}_1^{(i)}(t) - \mathbf{P}_1^{(i-1)}(t)\|_F^2 + \|\mathbf{P}_2^{(i)}(t) - \mathbf{P}_2^{(i-1)}(t)\|_F^2 < \Theta \quad (12)$$

where Θ is a threshold.

The bi-subspace model in our approach is similar in spirit to the co-training idea in [10] for self-supervised learning.

4. EXPERIMENTAL RESULTS

In our experiments, we use brightness pattern and edge as the two cues. A hypothesized image region is normalized to 20 × 20 and rasterized to $z_1(x(t)) \in \Re^{400}$ as the $\mathbf{P}_1$. We obtain the magnitude of the image gradients and rasterize them to $z_2(x(t)) \in \Re^{400}$ to form $\mathbf{P}_2$. The thresholds are determined by the positive data obtained in the previous frame.

We compare the proposed algorithm with the method in [7], where the clustering method is used to form supervision data. We refer to this method as the *clustering method*. In the quantitative study, we have manually annotated a challenging test video, in which a person is drinking, then sits down and hides behind a desk. The ground truth of the target location are manually annotated. The comparison is based on the distance of the ground truth data to the centers of tracking region by various methods. A smaller distance implies a better method. The result is shown in Figure 3.

As shown in Figure 3, the distance curve of our approach is apparently lower than that of the clustering method. This verifies that the bi-subspace training is effective. Some sample frames are shown in Figure 4, where the top row is the results of the proposed method, and the bottom row shows the result of the baseline clustering method. We can see that the clustering method loses track when the head hides behind the desk, as the blue color of the desk dominates the predicted region and is treated as positive data in clustering method. In this case, the result validates that bi-subspace model is more capable in coping with such occlusion scenarios.

Figure 5 shows the results of tracking a moving car under dramatic illumination changes. The appearances of the car under different lighting conditions are significantly different, which makes the tracking very difficult. The experiments show that our approach is robust to illumination changes.

5. CONCLUSION

In this paper, we propose a bi-subspace model to handle the non-stationary appearance tracking problem. In our method,

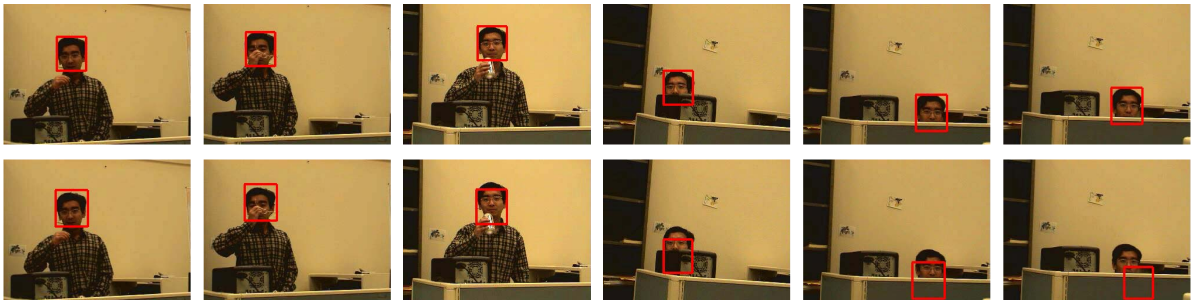

Fig. 4. Tracking a partial occluded target. (top) our method, (bottom) clustering method in [7]

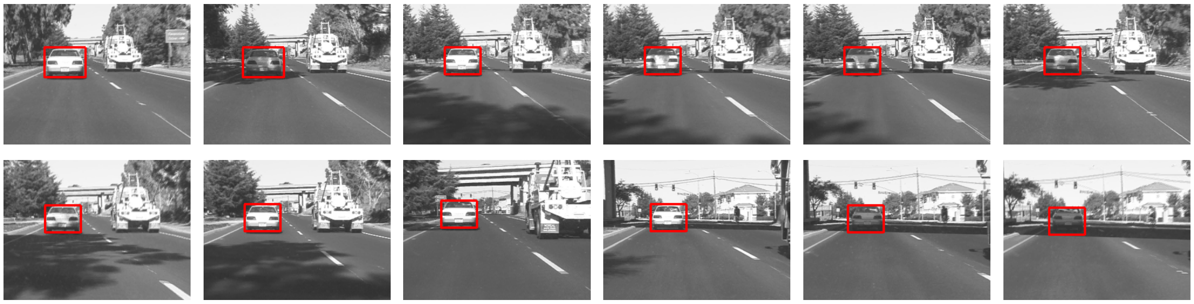

Fig. 5. Tracking a moving car under dramatic illumination changes

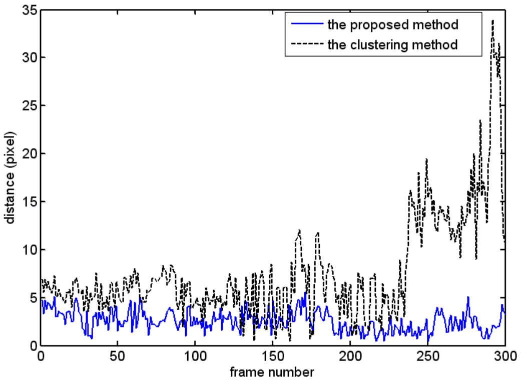

Fig. 3. Comparison of distances of the tracked region centers against the ground truth data.

two appearance subspaces interact with each other in updating, which leads to a more robust tracking performance. In addition, we impose both top-down (i.e., smoothness) and bottom-up (i.e., data-driven) constraints from current observations to make the problem well-posed. Our experimental results show that the proposed algorithm has a better performance than the existing methods.

Acknowledgement This work was supported in part by National Science Foundation Grants IIS-0347877 and IIS-0308222.

6. REFERENCES

[1] Y. Wu and T.S. Huang, "A co-inference approach to robust visual tracking," in *ICCV*, 2001, vol. II, pp. 26–33.

[2] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *CVPR*, 2000, vol. II, pp. 142–149.

[3] G. Hager, M. Dewan, and C. Stewart, "Multiple kernel tracking with SSD," in *CVPR*, 2004, vol. I, pp. 790–797.

[4] A.C. Sankaranarayanan, R. Chellappa, and Q. Zheng, "Tracking objects in video using motion and appearance models," in *ICIP*, 2005, vol. II, pp. 394–397.

[5] S. Avidan, "Support vector tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 8, pp. 1064–1072, August 2004.

[6] J. Ho, K.-C. Lee, M.-H. Yang, and D. Kriegman, "Visual tracking using learned linear subspace," in *CVPR*, 2004, vol. I, pp. 782–789.

[7] M. Yang and Y. Wu, "Tracking non-stationary appearance and dynamic feature selection," in *CVPR*, 2005, vol. II, pp. 1059–1066.

[8] J. Vermaak, P. Perez, M. Gangnet, and A. Blake, "Towards improved observation models for visual tracking: selective adaptation," in *ECCV*, 2002, vol. I, pp. 645–660.

[9] D. Ross, J. Lim, and M.-H. Yang, "Adaptive probabilistic visual tracking with incremental subspace update," in *ECCV*, 2004, vol. I, pp. 215–227.

[10] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *the Workshop on Computational Learning Theory*, 1998.