

# Chapter 51

## Multiple Faces Tracking via Statistical Appearance Model

Jie Hou, Yaobin Mao and Jinsheng Sun

**Abstract** Recently, appearance based methods have become a dominating trend in tracking. For example, tracking-by-detection models a target with an appearance classifier that separates it from the surrounding background. Recent advances in multi-target tracking suggest learning an adaptive appearance affinity measurement for target association. In this paper, statistical appearance model (SAM), which characterizes facial appearance by its statistics, is developed as a novel multiple faces tracking method. A major advantage of SAM is that the statistics is a target-specific and scene-independent representation, which helps for further video annotation and behavior analysis. By sharing the statistical appearance models between different videos, we are able to improve tracking stability on quality-degraded videos.

**Keywords** Tracking · Appearance model · Learning

### 51.1 Introduction

Face tracking is a useful computer vision technology in the fields of human computer interface and video surveillance, and has been successfully applied in video conferencing, gaming and living avatars in web applications, etc. Facial

---

J. Hou · Y. Mao (✉) · J. Sun  
School of Automation, Nanjing University of Science and Technology,  
Nanjing, Jiangsu, China  
e-mail: myb\_nust@126.com

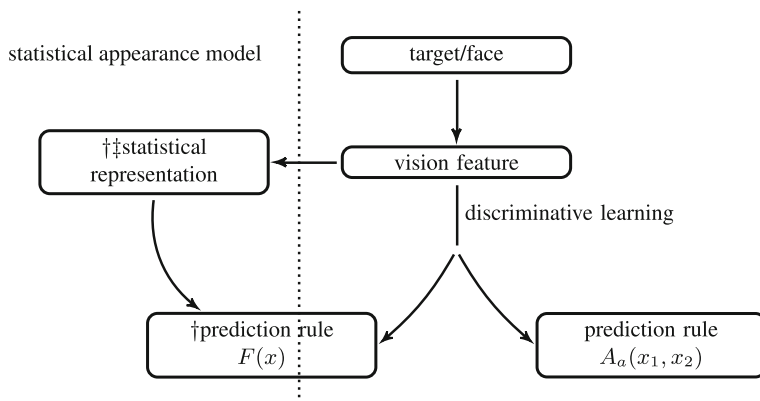
J. Hou  
e-mail: reiase@gmail.com

J. Sun  
e-mail: jssun67@yahoo.com.cn

appearance serves an important role in our daily communications, thus tracking human faces is usually the first step when a vision system tries to understand human beings.

Single face tracking has been well studied by researchers. Vacchetti [1] and Wang [2] use local interest-points matching to track 3D pose of human face. Their methods are robust against illumination and appearance variations, but suffer from error-prone feature matching. Zhou [3] proposed to track faces with Active Appearance Models (AAM), which has advantages of stability and alignment accuracy. One limitation of Zhou's method is that AAM fitting fails when handling large occlusions. Multiple faces problem is also common when performing video annotation and behavior analysis in real-world applications.

In this paper, we model facial appearance with the statistics (mean and variation) of the haar-like vision features, and train each face a discriminative prediction rule. An overview of statistical appearance model is shown in Fig. 51.1. The most significant difference between statistical appearance model and traditional methods is that we split the appearance model into two individual parts: statistical representation and prediction rule. The statistical representation is a target-specific and scene-independent representation. For discriminability, we learn a prediction rule that separating one face from the other faces with separability-maximum boosting (SMBoost). In our previous work [4], SMBoost shows better classification performance than AdaBoost and its online variation on UCI machine learning datasets, and achieves higher accuracy in tracking problems. The major contribution of our paper is a scene-independent appearance model. By reusing the statistical representations, we can apply the facial appearance model trained on one video sequence to other scenes.



**Fig. 51.1** † indicates that the model or the representation is target-specific, and ‡ indicates scene-independence. Prediction rule  $A_a(x_1, x_2)$  is used as global affinity measurement in multi-target tracking [5, 6], and  $F(x)$  is used to model posterior probability  $P(T|x)$  in tracking-by-detection [7]. Our work in this paper is described in the *left* part of the figure

## 51.2 Related Work

Tracking multiple targets that have complex motion traces is a challenging computer vision task, especially considering this problem in cluttered environments. Suppose we are tracking targets  $\{T_i\}_{i=1}^n$  in current frame, we can relocate a target  $T_i$  in the succeeding frame by searching for a candidate  $C$  from the scene that has maximum posterior probability  $P(T_i|C)$ . In tracking-by-detection, the tracker models  $P(T_i|C)$  with a binary classifier. Typically, haar-like feature and online boosting are used to train the classifier, for this combination has been proved to be very discriminative [8]. However, the classifier is scene-specified (see Table 51.1), for resampling of negative samples, which Viola used for generalizing [8], is not included.

In recent years, research on background subtraction and object detection have brought significant improvements to target detection, and lead to a trend of association based tracking, in which the target candidates  $\{C_j\}$  detected in succeeding frames are associated to the targets  $\{T_i\}$  in current frame. For simple scenes, the targets are represented with their motion models. By a precise estimating of motion states (e.g., position, speed and acceleration), we can guess the location of a target by likelihood probability  $P((x, y)|T_i)$ . With that guess, we can pair a candidate target  $C$  with the model that has maximum posterior probability  $P(T_i|C)$ .

However, motion model based association will be defeated if a target was absent from observation for a long time. In such situation, appearance is introduced as a secondary evidence of association. Some early approaches model the observations of a target directly with vision features of good invariance. Recently, Huang [9] proposes perform the association in a more adaptive manner. He joins tracklets into longer tracks by their linking probability:

$$P_{link}(T_i, T_j) = A_m(T_i, T_j) A_a(T_i, T_j) A_t(T_i, T_j) \quad (51.1)$$

where  $T_i$  is the target of the  $i$ th tracklet, and  $A_m$ ,  $A_a$ ,  $A_t$  indicate the motion, appearance and temporal affinities between the tracklets. The linking probability  $A_a(T_i, T_j)$  is a global appearance affinity measurement. A significant advantage of Huang's method is that the appearance affinity could be modeled with discriminative learning. Cheng [5] suggests to represent target appearance with local image descriptors, and use AdaBoost for feature selection. Low-supervised learning, for

**Table 51.1** A comparison of appearance models

		Target-specific	Scene-independent
Tracking-by-detection	Boosting [7]	+	-
	MIL [11]	+	-
Multi-target tracking	Global affinity measurement [5, 6]	-	-
	PIRMPT [12]	+	-
Our method		+	Partly

example, multi-instance boosting, is introduced to handle noisy data [10]. Table 51.1 is a summary of appearance models mentioned in this paper. We focus on whether the method models  $P(T_i|C)$  directly (target-specific), and scene-independence.

### 51.3 Statistical Appearance Model

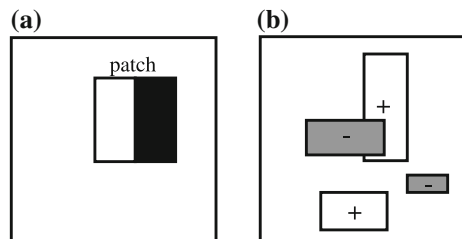
In this section, we present a statistical approach to model object appearance. The key idea of statistical appearance model is characterizing an object with the statistics of vision features (statistical representation, SR). Statistics is not a representation of good invariance. However, we are able to learn a discriminative prediction rule (PR) from the statistics if the vision feature (VF) is choosing properly and learning algorithm (LA) is carefully designed. Vision feature, statistical representation and learning algorithm are three important aspect of statistical appearance model. We will introduce each of them in detail.

#### 51.3.1 Vision Feature

Various vision features has been discussed for modeling appearance, e.g., shape, color histogram and texture (HOG) [12]. However, these features focus on invariance of a target, and might not discriminative enough to classify very similar targets (faces in this paper). Thus, researchers propose extracting these features on pre-defined regions to make the feature more discriminative [5].

Another method of enhancing discriminability is exploring an over-rich weak feature set with boosting. By combining weak classifiers of haar-like features, Viola builds the first practical real-time face detector. Original haar-like features that Viola used in his work are designed to capture within-patch structure patterns (see Fig. 51.2). Babenko uses non-adjacent haar-like feature, which combines 2–4 randomly generated rectangle regions and ignore the adjacency, in their work [10].

**Fig. 51.2** Haar-like feature.  
**a** Within-patch haar-like feature. **b** Cross-patch haar-like feature



### 51.3.2 Statistical Representation of Appearance

In this paper, we use statistical representation for characterizing facial appearance. The statistical representation of a target includes the mean  $E$  and variation  $\Sigma$  of the samples.

Suppose  $\{x_i\}$  is a data sequence, and the expectation of the first  $n$  samples is denoted by  $E[x]^{(n)}$ . When there comes a new sample  $x_{n+1}$ , we update the expectation  $E[x]$  with a learning rate  $\gamma$ .

$$E[x]^{(n+1)} = (1 - \gamma)E[x]^{(n)} + \gamma x \quad (51.2)$$

Variance  $\sigma^2[x]$  can be updated by a subtraction of two expectations:

$$\begin{aligned} \sigma^2[x]^{(n)} &= E[x^2]^{(n)} - (E[x]^{(n)})^2 \\ \sigma^2[x]^{(n+1)} &= (1 - \gamma)E[x^2]^{(n)} + \gamma x^2 - (E[x]^{(n+1)})^2. \end{aligned} \quad (51.3)$$

### 51.3.3 Learning Algorithm

In appearance base tracking, on-line boosting is a common choice for learning the prediction rules. Boosting chooses important features from the given feature pool, so that the prediction rule remains simple as it covers as many features as possible. However, the criterion function of AdaBoost (51.4) is difficult to be estimated on-line, for both the sample set  $\{(x, y)\}$  and the decision function  $F$  change in on-line learning paradigm [4].

$$F^* = \arg \min_F E_{\{(x,y)\}} [L(y, F(x))] \quad (51.4)$$

In our previous work [4], a new online boosting using separability based loss function instead of margin based loss function, is designed. The separability based loss function characterizes the degree that the decision function  $F$  separates the samples of class  $c(\{(x, y)\}_{y=c})$  from the rest  $(\{(x, y)\}_{y=\bar{c}})$ .

$$\hat{L}(c, L, E[x]|_{y=c}, \Sigma[x]|_{y=c}), \quad c \in \{+1, -1\} \quad (51.5)$$

Separability-maximum boosting (SMBBoost) maximizes separability of both classes.

$$\begin{aligned} F^* &= \arg \min_F \Phi(F) \\ &= \arg \min_F E_c [L(c, F, E[x]|_{y=c}, \Sigma[x]|_{y=c})] \end{aligned} \quad (51.6)$$

$E[x]$  and  $\Sigma[x]$  in (51.6) denote the mean and variation of the samples, which are well estimated by (51.2) and (51.3). Within on-line learning paradigm, it is much

easier to optimize SMBoost (51.6), for both  $E[x]$  and  $\Sigma[x]$  are fixed when the algorithm searching for optimal  $F$ .

### 51.4 Multiple Faces Tracking

Faces are similar to each other, thus finding facial appearance differences is much more difficult than separating one face from its surrounding background. The fundamental problem of multiple faces tracking is separating the faces from each other. We build an association based multiple faces tracker with statistical appearance model in the paper. The overview of our tracking framework is shown in Fig. 51.3. Our scheme is similar with [12] except the appearance model. In our tracker, we do not use tracklets for simplicity and real-time performance. We associate face candidates in succeeding frame to the faces in current frame directly. Each step in our framework is described below.

**Face detection** Face detection takes a new frame from the test video sequence and applies a state-of-art face detector to it. Face detection produces face candidates  $\{C_j\}_{j=1}^c$  under association.

**Face classification** Denote by  $\{T_i\}_{i=1}^t$  the statistical representation of the faces already known, also denote by  $\{F_i(x) \in [-1, +1]\}_{i=1}^t$  the prediction rules, where

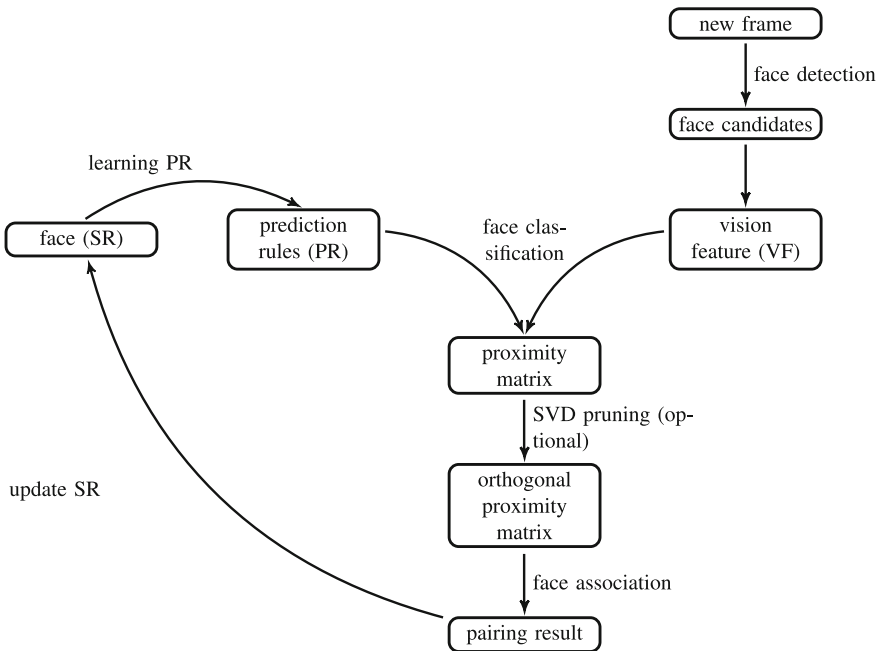


Fig. 51.3 System overview

$F_i(x)$  separate face  $T_i$  from the rest faces. We obtain the proximity matrix  $P$  by applying the prediction rules to the face candidates:

$$P_{i,j} = P(T_i|C_j) = \begin{cases} F_i(C_j), & F_i(C_j) > 0 \\ 0, & \text{otherwise} \end{cases} \quad (51.7)$$

**SVD pruning (optional)** For face association, we need one-against-one pairing, which requires a successful pairing satisfy

$$P_{i,j} = \arg \max_{\delta} P_{\delta,j} = \arg \max_{\delta} P_{i,\delta}$$

Thus we use the method mentioned in [13] to prune the proximity matrix  $P$  into an orthogonal matrix.

**Face association** After pruning, we can associate the face candidates to the faces by maximizing posterior probability  $P(T_i|C_j) = \widehat{P}_{i,j}$ .

**Update statistical representation** For faces that have successfully paired with candidates in succeeding frame, we update their statistical representation with a learning rate  $\gamma$ .

**Learning prediction rule** Suppose  $\{T_i\}_{i=1}^t$  are the faces that we are tracking, we learn each face a discriminative prediction rule. When training the rule, we use statistical representation of  $T_i$  as positive SR, and combine the statistical representations of the rest faces as negative SR  $\overline{T}_i = \sum_{\sigma \neq i} T_{\sigma}$ .

## 51.5 Experiments

In this section, we first perform experiments demonstrating the effectiveness of SAM for multiple faces tracking problems. Then, we present an experiment of sharing statistical representation between two video sequences.

### 51.5.1 Tracking Multiple Faces

We first experiment our method for multiple faces tracking problems. Two test sequences from [14] are used in our evaluation. Association fails when handling face candidates with scaling factor larger than two. Thus we stop the association when one of the detected faces is too large. The tracker will re-catch the faces after scaling. Since we use a motion-free target model, our tracker is robust against large occlusions and missing detecting, and tracks the faces stably. The tracker may assign candidates to wrong faces when facial appearance varies too much, e.g., laughing and rotating. In such situation, we should stop updating the statistical appearance representation. And the faces will switch back when the facial



**Fig. 51.4** Tracking multiple faces on Seq. face\_fast [14]



**Fig. 51.5** Tracking result on Seq. face\_frontal

appearance stops varying. Such failures could be reduced by rejecting sudden large moves (Figs. 51.4, 51.5).

### 51.5.2 Sharing Statistical Representations Between Video Sequences

In this experiment, we share the statistical representations when tracking human faces on different video sequences. Sharing appearance model is useful for video annotation and behavior analysis.

Two test video sequences are used in this experiment. Seq. nokia is a quality-degraded video sequence captured with a Nokia smart phone, which suffers shaking and motion blur (intentionally). Seq. samsung is captured with another Samsung smart phone. Fig. 51.6 shows tracking result of our method on Seq. nokia. And Fig. 51.7 presents tracking result of sharing the models. In Seq. nokia, our tracker fails to assign the candidates to the faces on frame 500 (associating fails), and miss one face in frame 140 and frame 300. By sharing the statistical representation estimated on a clearer observation (Seq. samsung), the stability of tracking result on Seq. nokia got improved, and find correct assignments on the failure frames.



**Fig. 51.6** Tracking multiple faces on Seq. nokia





**Fig. 51.7** Tracking result of sharing statistical representation. The *above row* shows tracking result on Seq. nokia, and the *row below* shows tracking result on Seq. samsung

## 51.6 Conclusion

In this paper, a statistical appearance model, which characterizes facial appearance with statistics, is proposed. SAM captures appearance invariance by exploring an over-rich haar-like feature set, and trains a classifier of good discriminability using separability-maximum booting. A novel framework using SAM is designed for multiple faces tracking. In our framework, we are able to track faces robustly. By sharing the statistical representations, we are able to improve tracking result on quality-degraded video sequence.

**Acknowledgments** This work is supported by the National Natural Science Foundation of China (Grant No. 60974129) and Intramural Research foundation of NJUST (2011YBXM119).

## References

1. Vacchetti L, Lepetit V, Fua P (2004) Stable real-time 3D tracking using online and offline information. *IEEE Trans Pattern Anal Mach Intell* 26:1385–1391
2. Wang Q, Zhang W, Tang X, Shum H-Y (2006) Real-time bayesian 3-D pose tracking. *IEEE Trans Circuits Syst Video Technol* 16:1533–1541
3. Zhou M, Liang L, Sun J, Wang Y (2010) AAM based face tracking with temporal matching and face segmentation. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, 2010, pp 701–708
4. Hou J, Mao Y, Sun J (2012) Visual tracking by separability-maximum online boosting. In: *12th International conference on control automation robotics vision (ICARCV)*, 2012, pp 1053–1058
5. Kuo C-H, Huang C, Nevatia R (2010) Multi-target tracking by online learned discriminative appearance models. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, 2010, pp 685–692
6. Yang B, Nevatia R (2012) Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, 2012, pp 1918–1925
7. Grabner H, Grabner M, Bischof H (2006) Real-time tracking via online boosting. In: *Proceedings of BMVC*, pp 47–56
8. Viola P, Jones M (2002) Robust real-time object detection. *Int J Comput Vis* 57:137–154

9. Huang C, Wu B, Nevatia R (2008) Robust object tracking by hierarchical association of detection responses. In: Forsyth D, Torr P, Zisserman A (eds) *Computer vision—ECCV*. Springer, Berlin, pp 788–801
10. Babenko B, Belongie M-HYS (2011) Robust object tracking with online multiple instance learning. In: *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*
11. Babenko B, Yang M, Belongie S (2009) Visual Tracking with Online Multiple Instance Learning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL
12. Kuo C-H, Nevatia R (2011) How does person identity recognition help multi-person tracking? *IEEE conference on computer vision and pattern recognition (CVPR)*, pp 1217–1224
13. Delponte E, Isgrò F, Odone F, Verri A (2006) SVD-matching using SIFT features. *Gr Models* 68:415–431
14. Maggio E, Piccardo E, Regazzoni C, Cavallaro A (2007) Particle PHD filtering for multi-target visual tracking. Presented at the *IEEE international conference on acoustics, speech and signal processing*