Multi-View Face Tracking with Factorial and Switching HMM

Peng Wang , Qiang Ji
Department of Electrical, Computer and System Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180

Abstract

Dynamic face pose change and noise make it difficult to track multi-view faces in a cluttering environment. In this paper, we propose a graphical model based method, which combines the factorial and the switching Hidden Markov Model(HMM). Our method integrates a generic face model with general tracking methods. Two sets of states, corresponding to appearance model and generic face model respectively, are factorized in the HMM. The measurements on different states are fused in a probabilistic framework to improve the tracking accuracy. To handle pose change, model switching mechanism is applied. The pose model with the highest probabilistic score is selected. Then pose angles are estimated from those pose models and propagated during tracking. The factorial and switching model allows to track small faces with frequent pose changes in a cluttering environment. A Monte Carlo method is applied to efficiently infer the face position, scale and pose simultaneously. Our experiments show improved robustness and good accuracy.

1 Introduction

Visual tracking has very important applications in many fields, such as surveillance, human computer interface and entertainment. Face tracking could be the link between face detection and other recognition tasks in surveillance. In this paper, we propose a probabilistic framework to track multi-view face and estimate the pose.

There are already intensive research on tracking. A very concise and comprehensive formalization of tracking is the state-space model. In this model, the purpose of tracking is to infer unknown states of the objects, such as position, scale and other properties, given model assumption and visual observations. There are three fundamental elements in the state-space model: "state", "observation model" and "inference strategy".

The "state" could be any concise descriptions of objects, such as shape, position and velocity. The state transition models is used to describe how those states dynamically change over times. Kalman filtering generally assumes the linear motion model and Gaussian noise. Sequential Monte Carlo sampling methods are applied to handle nonlinear dynamic systems [10].

Observation models relate hidden states with visible observations. In visual tracking, most used observation models are based on image, data such as color and shape [6, 1]. For complex objects tracking in cluttering environment, it is more desirable to combine multiple cues to improve robustness [17].

Given models and observation, true states are estimated from the inference strategy. Recursive Bayesian filtering provides a powerful formalization for inference with sequential data. In the case that the exact inference is intractable, sampling methods and variational approximation find their usage.

The most difficulty for multi-view face tracking raises from cluttering environments and pose changes. The error caused by noises and accumulated over frames will eventually lead tracking to get lost.

Fusion of multiple cues and tracking methods is claimed to be more robust to clutters [7]. In [7], different measurements are combined together in a probabilistic framework while different state spaces are related with each other by transformation. Color and shape cues are combined in "co-inference" to ease the clutters caused by each of them, therefore improve the whole performance [17]. Those tracking algorithms use features from images data, including color, edge and shape. In fact, there are already many efforts to combine face detectors in tracking. A simple way is to use face detection results to initialize the face tracking. More elegantly, the face detector is integrated in tracking as a measurement. Two face detectors, one for frontal faces and another for profile faces, are used in particle filtering to probabilistically propagate face detection in video [14]. In [2], the confidence from multiple modalities, including edge, color and face appearance, are integrated in Bayesian network.

In this paper, we present a graphical model based method to combine general object tracking with generic face model. The novelty is that we factorize the multi-view face tracking into two sub-inference problems in HMM with minimum

assumption.

Another problem in multi-view face tracking is the pose change. Since the profile face and frontal face have very distinct appearances, it is difficult to handle all the views with a single face model. In this paper, we incorporate the switching state-space model to handle pose changes and estimate pose angles.

As the results, we provide an integrated probabilistic framework for multi-view face tracking in this paper. The paper is organized as follows. In section 2, the face tracking is formalized with a factorial HMM model. Switching multi-view face tracker is proposed in Section 3. Section 4 introduces the Monte Carlo sampling method we used. Results and conclusion are given in the last two sections.

2 Combine Generic Face Model in Factorial HMM

2.1 Tracking with Graphical Model

Visual tracking is shown to be an inference problem in state-space model. Defining state X and observation model $P(Z|X)$, tracking is equivalent to inferring $P(X|Z)$ from given observation Z by Bayes' rule:

$$P(X|Z) = \frac{P(Z|X)P(X)}{P(Z)} \tag{1}$$

Graphical model is a concise representation of probabilistic inference. Hidden Markov Model(HMM) is a simple but powerful Dynamic Bayesian network (DBN), which can naturally handle tracking problems from sequential data[4].

Due to the Markov property, the posterior probability in HMM can be estimated from the observation model and propagated priors by Bayes' rule, as (2).

$$P(X_t|Z_{1:t}) = \frac{1}{C}P(Z_t|X_t)P(X_t|Z_{1:t-1}) \tag{2}$$

where C is the normalization constant.

For complex visual tracking, there could be a set of states from multiple cues. That is, $X = \{X^1, ..., X^M\}$, where M is the number of states. For example in a complex scenario, many factors such as shape, color, illumination condition and occlusion can influence the observation. Those states could be coupled, therefore the temporal relationship between those states become very complex and intractable. Suppose each state can take one of K discrete values, then the dimension of the state transition matrix is $K^M \times K^M$. The dimension will explode for multi-modal observation model after propagation.

2.2 Factorial HMM for Face Tracking

Factorial HMM assumes uncoupled state transitions [3]. Each of the multiple sets of states, $X = \{X^1, ..., X^M\}$, has its own transition, shown as Fig. 1. The total state transitions are factorized into the product of the independent transition for each subset of the states, as (3).

$$P(X_t|X_{t-1}) = \prod_{j=1}^{M} P(X_t^j|X_{t-1}^j) \tag{3}$$

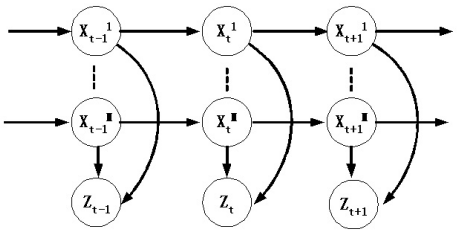


Figure 1: Factorial Hidden Markov Model

The advantage of factorization is the reduction of transition matrix dimension. After factorization, only M matrices with dimension of $K \times K$ are needed to model the state transition.

In [17], Wu and Huang proposed co-inference strategy. Shape and color are identified as two important states whose sampling distributions are coupled in Sequential Importance Sampling so that it is called "co-inference". It is claimed that such co-inference can ease the clutter from either shape or color. However such co-inference is also based on image data and suffers from the observation noises. In this paper, we propose another way of factorization, in which the states correspond to both image observation and generic model.

Carefully looking at the face tracking problem again, we find that it actually involves two purposes. Firstly, it tracks a specific object based on the similarity measurement, the same as other object tracking tasks. Secondly, the object in face tracking belongs to the particular "face" pattern, which should be consistent with our common concept about face. Therefore the face tracking problem is related with not only the specific object template, but also the generic face model. As the natural extension of discussion above, we factorize the face tracking problem in the graphical model and introduce face detectors.

In our factorial HMM, all the the hidden states are factorized into two subsets, X^o and X^f. X^o is the state set related with the the specific face. X^o can be position, scale, velocity and any other discrete or continuous properties about specific objects. X^f is the state related with the generic face model. It takes discrete value $\{+1, -1\}$, where $+1$ indicates the face and -1 for non-face.

The probabilistic formula for the inference in the factorial HMM can be written as (4).

$$P(X_t|Z_{1:t}) = P(X_t^o, X_t^f|Z_{1:t}) \quad (4)$$

The equation (4) can be further decomposed by chain rule as (5). It shows that the face tracking is a combination of two sub inference problems in the factorial HMM.

$$\begin{aligned} P(X_t|Z_{1:t}) &= P(X_t^o, X_t^f|Z_{1:t}) \\ &= P(X_t^f|Z_{1:t})P(X_t^o|Z_{1:t}, X_t^f) \end{aligned} \quad (5)$$

Equation (5) clearly explain the factorial HMM. $P(X_t^f|Z_{1:t})$ is the posterior probabilities of face or non-face given observation. It could be the probabilistic output of any face detector.

Inferring another component in (5), $P(X_t^o|Z_{1:t}, X_t^f = 1)$, is the same as other object tracking methods except the assumption that current object is a face. For those methods that do not integrate generic face model, the assumption behind them is that the specific object state is independent with generic face model so that $P(X_t^o|Z_{1:t}, X_t^f) = P(X_t^o|Z_{1:t})$. In our method, we adopt such assumption so that all the previous work can be well incorporated in our framework.

The advantage of such factorization is that we can easily obtain posterior probability $P(X_t^f|Z_{1:t})$ from the well trained face detector. We will show the benefits of combining generic face model in tracking at the following section.

2.3 Generic Face Model

To apply the generic face model in tracking, it is necessary to learn a face detector. The face detector should be able to output the probabilisties, $P(X^f|Z)$ and $P(Z|X^f)$. A survey on face detection methods can be found in [19].

It should be mentioned here that the face detector in tracking can be much "weaker". It should not achieve the perfect detection rate and false rate as in the face detection. In our experiments, 90% detection rate and 10% false rate are acceptable.

In our method, we choose support vector machine (SVM) to model the face. SVM has been widely applied in face detection and other pattern classification problems [11]. It has excellent generalization capability, which is very desirable for tracking in cluttering environments. However, some other face detection methods can also be integrated in our framework only if they can output probabilities.

Generally SVM does not output probabilities. The numeric result of SVM is just "the distance" from feature point to separate hyperplane, as $f(Z)$ in (6), where y_i is the class label of support vectors z_i, and K is the kernel function.

$$f(Z) = \sum_{i=1}^{N} \alpha y_i K(z_i, Z) + b \quad (6)$$

There are some efforts to interpret SVM from the probabilistic point of view. A direct method is to fit the output of SVM into posterior probability [12]. By assuming the distance output by SVM as a Gaussian distribution, the posterior probability can be directly fit with the sigmoid function as shown in (7). Notation is changed in the way $y_i = X^f$ to make it consistent with previous discussion.

$$P(X^f|Z) = \frac{1}{1 + exp(A_i f(Z) + B_i)} \quad (7)$$

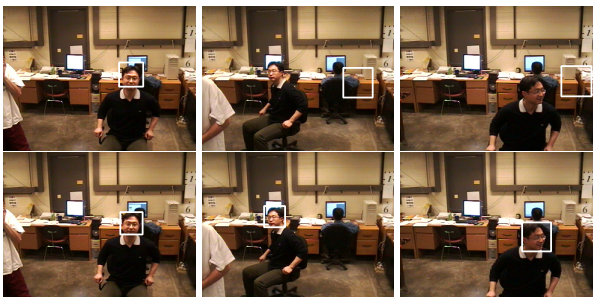


Figure 2: Face tracking with and without generic model. Top row: tracking results only using appearance. Bottom row: tracking results with factorial HMM.

Incorporating the face detector can eliminate the distraction from the cluttering environment. Fig. 2 shows multi-view face tracking with position, scale and pose changes. Color histogram is used to represent the face appearance [1]. The object template is updated at the rate of 10% per frame. Other implementation details are as in Section 5. Fig.2 shows that when the face pose, scale and position change simultaneously, the tracker gets lost because of background distraction and frequent pose changes when only appearance model is used. Our proposed method can successfully track the multi-view face all through the video. It shows greatly improved robustness by combining the generic face model with a simple appearance model.

3 Model Switching and Pose Estimation

In a real video, face appearance changes dynamically with pose changes. Sometimes the difference across multiple views is even larger than the difference between face and non-face. A single model to handle all the views may not

be accurate enough. In this section, we introduce switching mechanism to choose specific pose model and estimate pose in a video sequence.

In visual tracking, switching models are introduced to handle multi-modal targets and changing environments. The multi-modal motion can be naturally modeled with mixed-state in Condensation [5], where the switching variables are also states mixed with other model parameters. In [16], multiple observation models are learned and switched at different cluttering environments. In [18], three view templates are switched to handle multi-view appearance and occlusion. However, those view templates should be learned in advance for specific persons.

To naturally tracking the face pose changes, our method switches the generic face models across different views, which needs a multi-view face detector. There are already many methods to detect multi-view faces [9, 8]. Those methods all focus on detect faces in still images. In [14], the poses are estimated from interpolation of the frontal and profile face detectors. In this section, we construct one generic face model for each pose and switch them during tracking to select the best model.

In factorial HMM, the generic face model is already separated from specific object states. Now we redefine the X^f as the switching state, i.e. $X^f = M^i (i = 1, ..., N)$, where the multi-view face model is represented by M^i, and N is the number of poses.

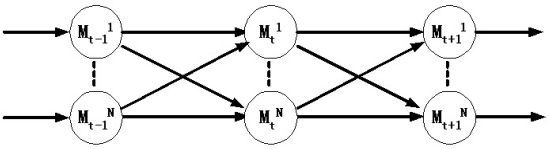


Figure 3: Switching HMM for multiple views

Actually, the pose models here aim to approximate the out-of-plan rotation, not to accurately estimate 3-D pose. The 3-D pose estimation needs stereo, not a topic of our paper. From our experiment, rough face pose models work well for tracking . The training of each pose model is the same as the generic face model in Section 2.3.

The multi-view face models have their own dynamics. Given the pose transition probabilities $P(i, j) = P(X^f = M^i | X^f = M^j)$ and likelihood model for each pose $P(Z|X^f = M^i)$, we can infer the posterior probabilities of face pose over time, as (8).

$$P(M_t^i|Z_{1:t}) \propto P(Z_t|M_t^i) \sum_{M_{t-1}^j} P(M_t^i|M_{t-1}^j) P(M_{t-1}^j|Z_{1:t-1}) \quad (8)$$

The switch state at X_t^f time t is estimated from the tracking results of each poses:

$$X_t^f = arg \max_{M^i} P(M^i|Z_{1:t}) \quad (9)$$

The advantage of switching model is that we can improve the tracking accuracy. Since it is difficult to accurately model all the poses in one face model, we can obtain better measurement on X^f by choosing the appropriate pose model.

Face pose angle can be estimated from those models. Assume that each view is associated with a typical pose angle, $\{\theta_i : M_i\}$, the pose angle can be estimated from multi-view face detector:

$$\bar{\theta}_t = E[\theta_t] = \sum_i \theta_i P(M_i|Z_{1:t}) \quad (10)$$

Fig.4 shows the multi-view face tracking and pose estimation results. All the parameters are set as in Section 5. The half arrow as in Fig.4.(b) indicates half profile faces while large arrow indicates large profile faces. The pose angles estimated from switching HMM are shown in Fig.4.(i). Beginning at frontal position, the face moves to left, then turns to right and finally stays at frontal position. The estimated poses clearly indicate the face moving.

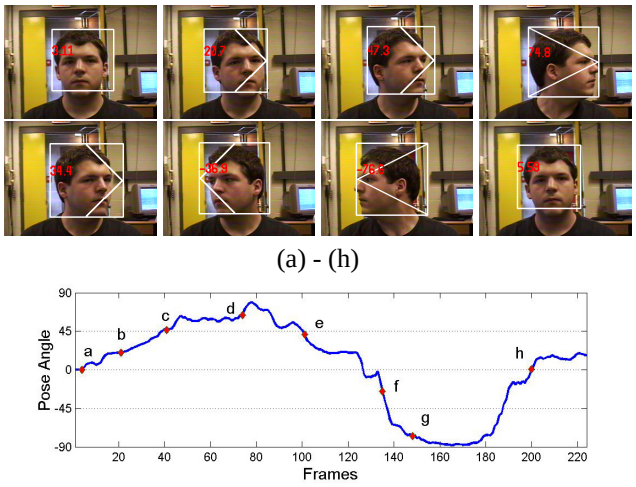


Figure 4: Multi-view face tracking and pose estimation. (a)-(h): tracking results in sequence, arranged from left to right and top to bottom. (i): pose estimated from switching model

4 Monte Carlo Sampling Method

Most close-form inference methods assume models with Gaussian or other tractable distributions, which may not be necessarily satisfied. Sequential sampling methods are simple but powerful tools to avoid such problems [10]. Their

combinations with the recursive Bayesian filtering are very popular in computer vision, such as particle filtering [6].

For multi-view face tracking, the state set related with specific object includes position X^p, scale X^s and velocity X^v to handle approximately linear motion and scale change, i.e. $X^o = (X^p, X^s, X^v)$. For generic face model, X^f can be one of the face poses $M^i, i = 1, ..., N$. Each of them only takes two values, 1 for faces and -1 for non-faces. The samples are represented as $\{x_t^o(k), x^f(k), \pi_t(k)\}, k = 1, ..., N$.

In sampling-based tracking, the posterior probability usually converges at only a few dominant samples. It will cause degeneracy if the future samples are only drawn from those dominant samples. Usually the re-sampling step will draw samples from those dominant samples with an additive Gaussian distribution to prevent sample impoverishment. In our case, since the generic face model is introduced to eliminate the clutters, the posterior distribution is so converged at several particles $x_t(i)$ that the posterior distribution could be approximated as a mixture of Dirac functions $\sum_i \pi_t(i)\delta(x - x_t(i))$. The re-sampling function is actually the spread function with the center at those peak particles. It enables us to predict the states with relatively small number of particles, not above 100 in our experiments.

From (5), we have

$$\pi_t(k) = P(x^o(k)|Z)P(x^f(k)|Z) \tag{11}$$

In our factorial inference, introducing generic face model does not need extra samples because the measurements of face and non-face can be simultaneously obtained from the face detector.

Based on the estimation of posterior probabilities, we can estimated the state as (12).

$$\begin{aligned} \bar{X}_t^o &= \sum_k x^o(k)\pi_t(k) \\ X_t^f &= arg\max_{M^i} \sum_{k:X_t^f(k)=M_i(k)} \pi_t(k) \end{aligned} \tag{12}$$

5 Experiment Results

In our implementation, the face tracker is automatically initialized by the face detector. To improve the detection and tracking robustness, the training data in our method is more like "head" because it includes more features as hair and ears. We apply a combined method to detect multi-view faces, which enables us to detect small and blurred faces in a relatively long distance [15].

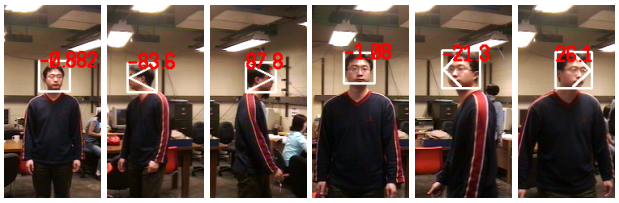
Multi-view faces are modeled with 5 poses, $\{M^i, i = 1, ..., 5\}$. Each pose represents a typical rotation angle θ_i. θ_i ranges from -90^o to 90^o, which represents left profile face to right profile face. The five poses are:

$$[\theta_1, \theta_2, \theta_3, \theta_4, \theta_5] = [-90^o, -45^o, 0^o, 45^o, 90^o] \tag{13}$$

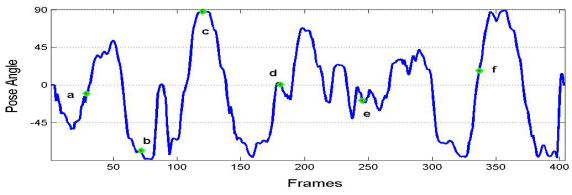
By assuming the progressive change of face poses, the transition between these states can be modeled as follows using the finite state machine (FSM), which can be set up empirically. In implementation, at most 3 face models are kept at each frame. Pruning pose model will save computation load without deteriorating performance.

5.1 Tracking Dynamic Pose Change

We apply proposed method to track multi-view faces in different environments. The first experiment is to track intensive pose change.



(a)-(f)



(g)

Figure 5: Dynamic face tracking and pose estimation. (a)-(f): tracking results, arranged from left to right and top to bottom. (g): estimated pose angle

In the sequence shown as Fig.5, the face undergoes intensive pose changes. At the first frame, the face candidate is detected and used to initialize the face tracker. To obtain robust tracking result, all the five pose model are kept in tracking at the first several frames. After that, we apply pruning strategy to keep only three pose models at each frame. The experiment shows that our method can successfully track pose as well as position and scale.

5.2 Multiple Faces Tracking in Clutters

Our face tracking method can be applied to track multiple people in cluttered environments. The face detection and tracking methods are combined with motion detection. Background modeling and subtraction method is incorporated with face detection [13]. Faces are only detected in motion regions to improve the efficiency and reduce false detections.

Two experiment results are shown here. Fig.6 shows multiple faces tracking result at an indoor environment. The face in this video is with low resolution and dynamic pose

change. The small size, poor illumination condition and cluttering background bring difficulties for tracking. Due to combination of object appearance and generic face model, multiple faces are successfully tracked.

Figure 6: Multiple face tracking in an indoor scene

Figure 7: Tracking results in an outdoor scene.

Experiment results at the outdoor scene is shown at Fig. 7. The camera is set up on campus. Unlike indoor environments, most faces are profile and with scale changes. Our system can quickly capture and successfully tracking the faces.

6 Conclusion

In this paper, we propose a novel formalization for multi-view face tracking based on the graphical model. We find that face tracking can be factorized into two sub inference problems. One is for specific object tracking, and another is based on generic face model. Factorial HMM combines the two models and provides more robust results. For dynamic multi-view face tracking in clutters, switching mechanism is introduced to select most suitable pose model for current tracking. The integrated framework allows us to track intensive pose change and estimate pose angle. We show improved robustness and good accuracy by experiments.

Acknowledgments

The work described in this paper is supported in part by a TSWG grant N41756-03-C-4028.

References

[1] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer, *Kernel-based object tracking*, IEEE. Trans on Pattern Analysis and Machine Intelligence **25** (2003), no. 5, 564–577.

[2] Liu Fang, Lin Xueyin, S.Z. Li, and Shi Yuanchun, *Multi-modal face tracking using bayesian network*, Analysis and Modeling of Faces and Gestures, IEEE International Workshop on, 2003, pp. 135–142.

[3] Z. Ghahramani and M.I. Jordan, *Factorial hidden markov models*, Machine Learning **29** (1997), 245–273.

[4] Zoubin Ghahramani, *An introduction to hidden markov models and bayesian networks*, International Journal of Pattern Recognition and Artificial Intelligence **15** (2001), no. 1, 9–34.

[5] M. Isard and A. Blake, *A mixed-state condensation tracker with automatic model-switching*, Computer Vision, IEEE International Conference on, Jan. 1998, pp. 107–112.

[6] Michael Isard and Andrew Blake, *Condensation:conditional density propagation for visual tracking*, International Journal of Computer Vision, **29** (1998), no. 1, 5–28.

[7] I. Leichter, M. Lindenbaum, and E. Rivlin, *A probabilistic framework for combining tracking algorithm*, Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, vol. 2, June 2004, pp. 445–451.

[8] S.Z. Li, L. Zhu, Z.Q. Zhang, A. Blake, H.J. Zhang, and H. Shum., *Statistical learning of multi-view face detection*, ECCV, May 2002.

[9] Yongmin Li, Shaogang Gong, and H. Liddell, *Support vector regression and classification based multi-view face detection and recognition*, Automatic Face and Gesture Recognition, IEEE International Conference on, 2000, pp. 300–305.

[10] Jun S. Liu and Rong Chen, *Sequential monte carlo methods for dynamic systems*, Journal of American Statistical Association **93** (1998), no. 443, 1032–1044.

[11] E. Osuna, R. Freund, and F. Girosi, *Training support vector machines:an application to face detection*, CVPRComputer Vision and Pattern Recognition, IEEE Computer Society Conference on (1997), 130–136.

[12] J. Platt, *Advances in large margin classifiers*, ch. Probabilistic outputs for support vector amchines and comparison to regularized likelihood methods, MIT Press, Cambridge, MA, 2000.

[13] Chris Stauffer and W.E.L Grimson, *Adaptive background mixture models for real-time tracking*, CVPRomputer Vision and Pattern Recognition, IEEE Computer Society Conference on, vol. 12, 1999, pp. 246–252.

[14] Rigini Choudhury Verma, Cordelia Schmid, and Krystian Mikolajcayk, *Face detection and tracking in a video by propagating detection probabilities*, IEEE Trans. on Pattern Analysis and Machine Intelligence **25** (2003), no. 10, 1216–1228.

[15] Peng Wang and Qiang Ji, *Multi-view face detection under complex scene based on combined svms*, International Conference on Pattern Recognition, vol. 4, 2004, pp. 179–182.

[16] Ying Wu, Gang Hua, and Ting Yu, *Switching observation models for contour tracking in clutter*, Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, vol. 1, June 2003, pp. 295–302.

[17] Ying Wu and Thomas S. Huang, *Robust visual tracking by integrating multiple cues based on co-inference learning*, International Journal of Computer Vision **58** (2004), no. 1, 55–71.

[18] Ying Wu, Ting Yu, and Gang Hua, *Tracking appearances with occlusions*, Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, vol. 1, June 2003, pp. 789–795.

[19] Ming-Hsuan Yang, David J. Kriegman, and N. Ahuja, *Detecting faces in images: A survey*, IEEE Trans. on Pattern Analysis and Machine Intelligence **24** (2002), no. 1, 34–58.