Pattern Recognition in Video

Rama Chellappa, Ashok Veeraraghavan, and Gaurav Aggarwal

University of Maryland, College Park MD 20742, USA,
{rama,vashok,gaurav}@umiacs.umd.edu
WWW home page: http://www.cfar.umd.edu/~rama

Abstract. Images constitute data that lives in a very high dimensional space, typically of the order of hundred thousand dimensions. Drawing inferences from data of such high dimensions soon becomes intractable. Therefore traditionally several of these problems like face recognition, object recognition, scene understanding etc. have been approached using techniques in pattern recognition. Such methods in conjunction with methods for dimensionality reduction have been highly popular and successful in tackling several image processing tasks. Of late, the advent of cheap, high quality video cameras has generated new interests in extending still image-based recognition methodologies to video sequences. The added temporal dimension in these videos makes problems like face and gait-based human recognition, event detection, activity recognition addressable. Our research has focussed on solving several of these problems through a pattern recognition approach. Of course, in video streams patterns refer to both patterns in the spatial structure of image intensities around interest points and temporal patterns that arise either due to camera motion or object motion. In this paper, we discuss the applications of pattern recognition in video to problems like tracking, face and gait-based human recognition, activity recognition and activity based person identification.

1 Introduction

Pattern recognition is the art of being able to look at raw data and cateogarizing it into one of available classes. In order to perform this, we need to first decide on a feature space to represent the data in a manner which makes the classification task simpler. Once we decide the features, we then need to describe each class or cateogary using models. Given data of an unlabelled object, we recognize the class of this object by inferring which of these descriptions best explains the features. This task of detecting, describing and recognizing visual patterns has lead to advances in automating several tasks like optical character recognition, scene analysis, fingerprint identification, face recognition etc.

In the last few years, the advent of cheap, reliable, high quality video cameras has spurred interest in extending these pattern recognition methodologies to video sequences. In video sequences, there are two distict varieties of patterns. Spatial patterns correspond to problems that were addressed in image based pattern recognition methods like fingerprint and face recognition. These challenges

exist in video based pattern recognition also. Apart from these spatial patterns, video also provides us access to rich temporal patterns. In several tasks like activity recognition, event detection/classification, anomaly detection, activity based person identification etc, there exists a temporal sequence in which various spatial patterns present themselves. It is very important to capture these temporal patterns in such tasks. In this paper, we describe some of the pattern recognition based approaches we have employed for tasks including activity recognition, face tracking and recognition, anomaly detection and behavior analysis.

2 Feature Representation

In most pattern recognition (PR) problems, feature extraction is one of the most important tasks. The problem of feature extraction is closely tied to that of pattern representation. It is difficult to achieve pattern generalization without using a reasonably correct representation. The choice of representation not only influences the entire PR approach to a great extent, but also limits the performance of the system, depending upon the appropriateness of the choice. For example, one cannot reliably retrieve the yaw and pitch angles of a face assuming a planar model.

Depending on the problem at hand, the representation itself can manifest in many different ways. Though in the case of still images, only spatial modeling is required, one needs ways to represent temporal information also when dealing with videos. At times, the representation is very explicit like in the form of a geometric model. On the other hand, in a few feature based PR approaches, the modeling part is not so visible. To further highlight the importance of representation, we now discuss the modeling issues related to a few problems in video-based recognition.

2.1 Affine Appearance Model for Video-based Recognition

Recognition of objects in videos requires modeling object motion and appearance changes. This makes object tracking a crucial preceding step for recognition. In conventional algorithms, the appearance model is either fixed or rapidly changing, while the motion model is a random walk model with constant variance. A fixed appearance template is not equipped to handle appearance changes in the video, while a rapidly changing model is susceptible to drift. All these factors can potentially make the visual tracker unstable leading to poor recognition results. In [1], we use adaptive appearance and velocity models to stabilize the tracker and closely follow the variations in appearance due to object motion. The appearance is modeled as a mixture of three different models, viz., (1) object appearance in a canonical frame (first frame), (2) slow-varying stable appearance within all the past observation, and (3) the rapidly changing component characterizing the two-frame variations. The mixture probabilities are updated at each frame based on the observation. In addition, we use a adaptive-velocity

model, where the adaptive velocity is predicted using a first-order linear approximation based on the appearance changes between the incoming observation and the previous configuration.

The goal here is to identify region of interest in each frame of the video and not 3D location of the object. Moreover, we believe that the adaptive appearance model can easily absorb the appearance changes due to out-of-plane pose changes and illumination changes. Therefore, we use a planar template and allow affine transformations only. Figure 1 shows an example where tracker using the described representation is used for tracking and recognizing a face in a video.

Fig. 1. Affine appearance model for tracking

2.2 3D Feature Graphs

Affine model suffices for locating the position of the object on the image, but it does not have the capability to annotate the 3D configuration of the object at each time instant. For example, if the goal is to utilize 3D information for face recognition in video, the described affine representation will not be adequate. Accordingly, [2] uses a cylindrical model with elliptic cross-section to perform 3D face tracking and recognition. The curved surface of the cylinder is divided into rectangular grids and the vector containing the average intensity values for each of the grids is used as the feature. As before, appearance model is a mixture of the fixed component (generated from the first frame) and dynamic component (appearance in the previous frame). Figure 2 shows a few frames of a video with the cylinder superimposed on the image displaying the estimated pose of the face.

Another possibility is to consider using a more realistic face model (e.g., 3D model of an average face) instead of a cylinder. Such detailed 3D representations make the initialization and registration process difficult. In fact, [3] shows experiments where perturbations in the model parameters adversely affect the tracking performance using a complex 3D model, whereas the simple cylindrical model is robust to such perturbations. This highlights the importance of the generalization property of the representation.

2.3 Gait Based Person Identification

Gait is a very structured activity with certain states like heel strike, toe off repeating themselves in a repetitive pattern. Recent research suggests that the gait of an individual might be distinct and therefore can be used as a biometric

Fig. 2. Estimated 3D pose of a face using a cylindrical model for face recognition in videos.

for person identification. We have developed algorithms for gait based person identification [4]. We used either the entire binary image silhouette or a sparses feature description called the width vector to represent the spatial patterns. Models like the Hidden Markov model (HMM) or the Dynamic time warping (DTW) were used to characterize the temporal patterns in gait. The recognition performance of such video based pattern recognition approaches for the problem of gait recognition was better than most contemporay approaches.

2.4 Behavior models for Tracking and Recognition

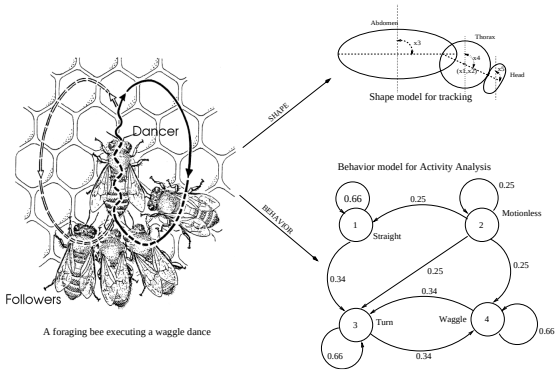


Fig. 3. A Bee performing waggle dance: The Shape model for tracking and the Behavior model to aid in Activity analysis are also shown.

The preceding sections described how to precisely model and represent the shape and the texture the object of interest in a video. These schemes for representation are adept at capturing the spatial patterns. But, if we need to identify patterns in the motion of the objects, then we need a representation that is capable of handling motion patterns explicitly. Statistical modeling of the motion of the objects enables us to achieve this end. Modeling such behaviors explicitly

is helpful in accurate and robust tracking. Moreover, explicitly modeling such behaviors also leads to algorithms where position tracking and activity/behavior analysis are tackled in a unified framework. We believe that the use of behavioral models presents a significant layer of abstraction that is able to capture the variety and complexity of the motions exhibited.

Typically each object could display multiple behaviors. We use Markovian models to represent each behavior of the object. In order to model the composite motion of the object we model the transitions between behaviors using another Markov model. This creates a mixture modeling framework for the motion of the object. For illustration, we will discuss the manner in which we modeled the behavior of insects for the problem of tracking and behavior analysis of insects. We model the probability distributions of location parameters $\mathbf{X}$ for certain basic motions($m_1 - m_4$). We model four different motions- 1) Moving straight ahead, 2) Turning, 3) Waggle, and 4) Motionless. The basic motions, straight, waggle and motionless are modeled using Gaussian pdfs (p_{m1}, p_{m3}, p_{m4}) while a mixture of two Gaussians (p_{m2}) is used for modeling the turning motion. Each behavior B_i is now modeled as a Markov process of order K_i on these motions, i.e.,

$$\mathbf{s}_t = \sum_{k=1}^{K_i} A_{B_i}^k \mathbf{s}_{t-k}; \tag{1}$$

where s_t is a vector whose j^{th} element is $P(\text{motion state} = m_j)$. The parameters of each behavior model are made of autoregressive parameters $A_{B_i}^k$ for $k = 1..K_i$. A typical Markov model for a special kind of dance of a foraging bee called the waggle dance is shown in Figure 3.

3 Particle Filtering for Object Recognition in Video

We have so far dealt with issues concerned with the representation of patterns in video and dealt with how to represent both spatial and temporal patterns in a manner that simplifies identification of these patterns. But, once we choose a certain set of representations for spatial and motion patterns, we need inference algorithms for estimating these parameters. One method to perform this inference is to cast the problem of estimating the parameters as a energy minimization problem and use popular methods based on variational calculus for performing this energy minimization. Examples of such methods include gradient descent, simulated annealing, deterministic annealing and Expectation-Maximization. Most such methods are local and hence are not guaranteed to converge to the global optimum or are too slow to be of practical use. When the state-observation description of the system is linear and Gaussian, estimating the parameters can be performed using the Kalman filter. But the design of Kalman filter becomes complicated for intrinsically non-linear problems and is not suited for estimating posterior densities that are non-Gaussian. Particle filter is a method for estimating arbitrary posterior densities by representing them with a set of weighted particles. We will precisely state the estimation problem first and then show how particle filtering can be used to solve such problems.

3.1 Problem Statement

Consider a system with parameters θ. The system parameters follow a certain temporal dynamics given by $F_t(\theta, D, N)$. (Note that the system dynamics could change with time.)

$$SystemDynamics: \qquad \theta_t = F_t(\theta_{t-1}, D_t, N_t) \tag{2}$$

where, N is the noise in the system dynamics. The auxiliary variable D indexes the set of motion models or behaviors exhibited by the object and is usually omitted in typical tracking applications. This auxiliary variable assumes importance in problems like activity recognition or behavioral analysis 4.3.

Each frame of the video contains pixel intensities which act as partial observations Z of the system state θ.

$$ObservationEquation: \qquad Z_t = G(\theta_t, I, W_t) \tag{3}$$

where, W represents the observation noise. The auxiliary variable I indexes the various object classes being modeled, i.e., it represents the identity of the object. We will see an example of the use of this in Section4.

The problem of interest is to track the system parameters over time as and when the observations are available. Quantitatively, we are interested in estimating the posterior density of the state parameters given the observations i.e., $P(\theta_t/Z_{1:t})$.

3.2 Particle Filter

Particle filtering [5][6] is an inference technique for estimating the unknown dynamic state X of a system from a collection of noisy observations $Z_{1:t}$. The particle filter approximates the desired posterior pdf $p(\theta_t|Z_{1:t})$ by a set of weighted particles $\{\theta_t^{(j)}, w_t^{(j)}\}_{j=1}^{M}$, where M denotes the number of particles. The interested reader is encouraged to read [5][6] for a complete treatment of particle filtering. The state estimate $\hat{\theta}_t$ can be recovered from the pdf as the maximum likelihood (ML) estimate or the minimum mean squared error (MMSE) estimate or any other suitable estimate based on the pdf.

3.3 Tracking and Person identification

Consider a gallery of P objects. Supposing the video contains one of these P objects. We are interested in tracking the location parameters X of the object and also simultaneously recognize the identity of the object. For each object i, the observation equation is given by $Z_t = G(\theta_t, i, W_t)$. Suppose we knew that we are tracking the p^{th} object, then, as usual, we could do this with a particle filter by approximating the posterior density $P(\theta_t/Z_{1:t}, p)$ as a set of M weighted particles $\{\theta_t^{(j)}, w_t^{(j)}\}_{j=1}^{M}$. But, if we didnt know the identity of the object we are tracking, then, we need to estimate the identity of the object also. Let us

assume that the identity of the object remains the same throughout the video, i.e., $I_t = p$, where $p = 1, 2, ...P$. Since the identity remains a constant over time, we have

$$P(X_t, I_t = i/X_{t-1}, I_{t-1} = j) = P(X_t/X_{t-1})P(I_t = i/I_{t-1} = j) \tag{4}$$

$$= \begin{cases} 0 & \text{if } i \neq j; \\ P(X_t/X_{t-1}) & \text{if } i = j; j = 1, 2, ...P \end{cases}$$

As was discussed in the previous section, we can approximate the posterior density $P(X_t, I = p/Z_{1:t})$ using a M_p weighted particles as $\{\theta_{t,p}^{(j)}, w_{t,p}^{(j)}\}^{j=1:M_p}$. We maintain such a set of M_p particles for each object $p = 1, 2, ..P$. Now the set of weighted particles $\{\theta_{t,p}^{(j)}, w_{t,p}^{(j)}\}_{j=1:M_p}^{p=1:P}$ with weights such that $\sum_{p=1:P} \sum_{j=1:M_i} w_{t,p}^{(j)} = 1$, represents the joint distribution $P(\theta_t, I/Z_{1:t})$. MAP and MMSE estimates for the tracking parameters $\hat{\theta}_t$ can be obtained by marginalizing the distribution $P(\theta_t, I/Z_{1:t})$ over the identity variable. Similarly, the MAP estimate for the identity variable can be obtained by marginalizing the posterior distribution over the tracking parameters. Refer to [7] for the details of the algorithm and the necessary and sufficient conditions for which such a model is valid.

3.4 Tracking and Behavior identification

Simultaneous tracking and behavior/activity analysis can also be performed in a similar manner by using the auxiliary variable D in a manner very similar to performing simultaneous tracking and verification. Refer to [8] for details about the algorithm. Essentially, a set of weighted particles $\{\theta_t^{(j)}, w_t^{(j)}, D_t^{(j)}\}$ is used to represent the posterior probability distribution $P(\theta_t, D_t/Z_{1:t})$. Inferences about the tracking parameters θ_t and the behavior exhibited by the object D_t can be made by computing the relevant marginal distribution from the joint posterior distribution. Some of the tracking and behavior analysis results for the problem of analyzing the behaviors of bees in a hive are given in a later section.

4 Pattern Recognition in Video: Working examples

In this section, we describe a few algorithms to tackle video-based pattern recognition problems. Most of these algorithms make use of the material described so far in this paper, in some form or the other.

4.1 Probabilistic Recognition of Human Faces from Video

This work [7] proposes a time series state space model to fuse temporal information in a video, which simultaneously characterizes the motion and identity. As described in the previous section, the joint posterior distribution of the motion vector and the identity variable is estimated at each time instant and then propagated to the next time instant. Marginalization over the motion state variables

yields a robust estimate of the posterior distribution of the identity variable. The method can be used for both still-to-video and video-to-video face recognition. In the experiments, we considered only affine transformations due to the absence of significant out-of-plane rotations. A time-invariant first-order Markov Gaussian model with constant velocity is used for modeling motion transition. Figure 4 shows the tracking output in a outdoor video.

Fig. 4. Example tracking results using the approach in [7].

4.2 Visual Recognition using Appearance-adaptive Models

This work [1] incorporates appearance-adaptive models in a particle filter to perform robust visual tracking and recognition. Appearance changes and changing motion is handled adaptively in the manner as described in Section 2.1. The simultaneous recognition is performed by including the identity variable in the state vector as described in Section 3.3.

4.3 Simultaneous Tracking and behavior analysis of insects

In [8], we aim to build a system that will assist researchers in behavioral research study and analyse the motion and behavior of insects. The system must also be able to detect abnormal behavior and model these abnormal behaviors. Such an automated system significantly speeds up the analysis of video data obtained from experiments and also prevents manual errors in the labeling of data. Moreover, parameters like the orientation of the various body parts of the insects(which is of great interest to the behavioral researcher) can be automatically extracted in such a framework. Each behavior of the insect was modeled as a Markov process on a low-level motion state. The transition between behaviors was modeled as another Markov process. Simultaneous tracking and behavior analysis/identification was performed using the techniques described in Section 3.4. Bees were modeled using the anatomical model described in Section **??**. Three behaviors of bees Waggle Dance, Round Dance and Hovering bee were modeled. Deviations from these behaviors were also identified and the model parameters for the abnormal behaviors were also learnt online. Refer [8] for the details of the approach.

4.4 Activity recognition by modeling shape sequences

Human gait and activity analysis from video is presently attracting a lot of attention in the computer vision community. [9] analyzed the role of two of the most important cues in human motion- shape and kinematics using a pattern recognition approach. We modeled the silhouette of a walking person as a sequence of deforming shapes. We proposed metrics for comparing two sequences of shapes using a modification of the Dynamic time warping algorithm. The shape sequences were also modeled using both AR and ARMA models. The theory of subspace angles between linear dynamical systems was used to compare two sets of models. Figure 5 depicts a graphical visualization of performing gait recognition by comparing shape sequences. Refer to [9] for the details of the algorithm and extended results.

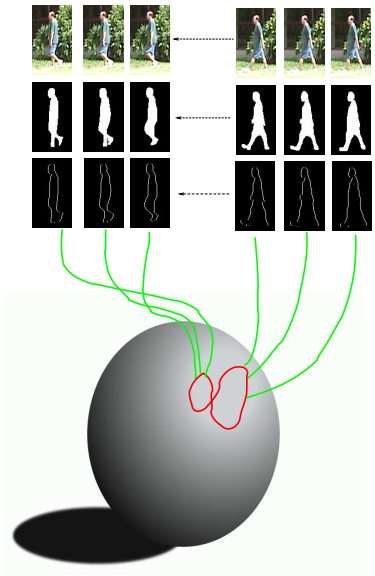
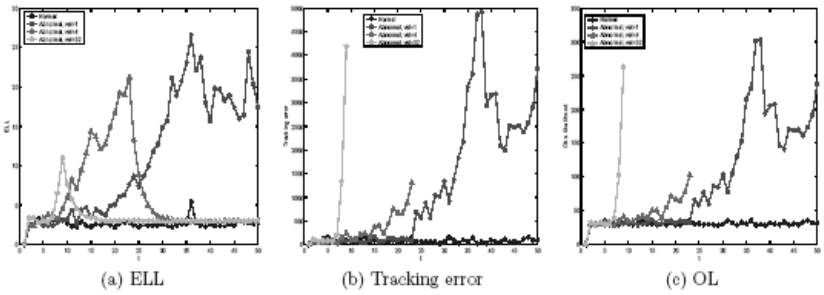


Fig. 5. Graphical illustration of the sequence of shapes obtained during gait

4.5 Activity modeling and anomaly detection

Several approaches including the activity factorization based approach described in the previous subsection. But, when new activities are seen, then we need to develop approaches to detect these anomalous activities. The activity model under consideration is a continuous state HMM. An abnormal activity is defined as a change in the activity model, which could be slow or drastic and whose parameters are unknown. Drastic changes can be easily detected using

the increase in tracking error or the negative log of the likelihood of current observation given past (OL). But slow changes usually get missed. [10] proposes a statistic for slow change detection called ELL (which is the Expectation of negative Log Likelihood of state given past observations) and shows analytically and experimentally the complementary behavior of ELL and OL for slow and drastic changes. We have also established the stability (monotonic decrease) of the errors in approximating the ELL for changed observations using a particle filter that is optimal for the unchanged system. Asymptotic stability is shown under stronger assumptions. Finally, it is shown that the upper bound on ELL error is an increasing function of the rate of change with increasing derivatives of all orders, and its implications are discussed. Figure 6 shows the tracking error, Observation likelihood and the ELL statistic for simulated observation noise.



(a) ELL (b) Tracking error (c) OL

Fig. 6. ELL, Tracking error (TE) and Observation Likelihood (OL) plots: Simulated Obser- vation noise, 2 obs = 9 (3-pixel noise). Notice that the TE and OL plots look alike.

5 Conclusions

We have presented very briefly some of the approaches based on pattern recognition to various problems like tracking, structure from motion, activity modeling, behavior analysis and abnormality detection. The treatment in this paper is not complete and the interested readers are encouraged to read the respective references for details for each of these approaches.

Acknowledgements

The authors would like to thank Shaohua (Kevin) Zhou, Namrata Vaswani, Amit Kale and Aravind Sundaresan for their contributions to some of the material presented in this manuscript.

References

1. Zhou, S., Chellappa, R., Moghaddam, B.: Visual tracking and recognition using appearance-adaptive models in particle filters. IEEE Trans. on Image Processing (2004)
2. Aggarwal, G., Veeraraghavan, A., Chellappa, R.: 3d facial pose tracking in uncalibrated videos. In: Submitted to Pattern Recognition and Machine Intelligence. (2005)
3. Cascia, M.L., Sclaroff, S., Athitsos, V.: Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models. IEEE Trans. on Pattern Analysis and Machine Intelligence **22** (2000) 322–336
4. Kale, A., Rajagopalan, A., Sundaresan.A., Cuntoor, N., Roy Cowdhury, A., Krueger, V., Chellappa, R.: Identification of humans using gait. IEEE Trans. on Image Processing (2004)
5. Doucet, A., Freitas, N.D., Gordon, N.: Sequential Monte Carlo methods in practice. Springer-Verlag, New York (2001)
6. Gordon, N.J., Salmond, D.J., Smith, A.F.M.: Novel approach to nonlinear/non-gaussian bayesian state estimation. In: IEE Proceedings on Radar and Signal Processing. Volume 140. (1993) 107–113
7. Zhou, S., Krueger, V., Chellappa, R.: Probabilistic recognition of human faces from video. Computer Vision and Image Understanding (CVIU) (special issue on Face Recognition) **91** (2003) 214–245
8. Veeraraghavan, A., Chellappa, R.: Tracking bees in a hive. Snowbird Learning Workshop (2005)
9. Veeraraghavan, A., Roy-Chowdhury, A., Chellappa, R.: Role of shape and kinematics in human movement analysis. CVPR **1** (2004) 730–737
10. Vaswani, N., Roy-Chowdhury, A., Chellappa, R.: Shape activity: A continuous state hmm for moving/deforming shapes with application to abnormal activity detection. IEEE Trans. on Image Processing (Accepted for Publication)