# Mixed State Models for Automatic Target Recognition and Behavior Analysis in Video Sequences

Rama Chellappa, Aswin C. Sankaranarayanan and Ashok Veeraraghavan

Center for Automation Research and Department of Electrical and Computer Engineering, University of Maryland, College Park, MD

## ABSTRACT

Mixed state or hybrid state space systems are useful tools for various problems in computer vision. These systems model complicated system dynamics as a mixture of inherently simple sub-systems, with an additional mechanism to switch between the sub-systems. This approach of modeling using simpler systems allows for ease in learning the parameters of the system and in solving the inference problem. In this paper, we study the use of such mixed state space systems for problems in recognition and behavior analysis in video sequences. We begin with a dynamical system formulation for recognition of faces from a video. This system is used to introduce the simultaneous tracking and recognition paradigm that allows for improved performance in both tracking and recognition. We extend this framework to design a second system for verification of vehicles across non-overlapping views using structural and textural fingerprints for characterizing the identity of the target. Finally, we show the use of such modeling for tracking and behavior analysis of bees from video.

**Keywords:** ATR, Tracking, Behavior Analysis, Mixed state systems, Particle Filtering

## 1. INTRODUCTION

Recognition of people and their activities from videos has always been a topic of considerable interest in the Vision community. There exist a variety of applications ranging from surveillance, motion capture and video indexing and retrieval that stand to gain from research in these topics. In this paper, we summarize some of the representative works in the fields of automatic target recognition (ATR) and behavior analysis. Specifically, we focus on recognition of the target using appearance information and the analysis of behavior encoded in the motion exhibited by the target.

Dynamical systems, in the context of vision applications, form important tools for modeling, analysis and inference. Consider a traditional dynamical system under assumptions of Markovian and conditional independence of observations. Such systems are characterized by three components: (1) prior density, (2) state transition model, and (3) the observation model. For a wide range of vision tasks, ranging from tracking, recognition, structure from motion, activity modeling and gait analysis, the state transition model encodes the motion of the underlying phenomena and the observation model encodes features derived from the captured imagery.

However, in many applications there is a need to model and capture the multiple modes of behavior, motion and/or appearance. To illustrate with an example, consider a ball bouncing on a floor and the problem of tracking its position. For most of the time, the ball exhibits a constant acceleration motion (due to gravity). However, on collision with the ground the motion of the ball is explained by a different model (one that models the collision as a flip in velocity with possibly a loss in energy). This presents us with two distinct choices in the kind of dynamical systems that we choose to use: (1) ONE complicated system that models all the various dynamics providing generality that is of great use, and (2) a dictionary of simple models with a switching mechanism that allows for the generation of complex behavior in the data. In many vision problems, it is much simpler to model the underlying state transition and observation models using a mixture of multiple models allowing for a structured switching between these models as opposed to having one complicated model that explains all the underlying phenomena.

Further author information: E-mail: {rama, aswch, vashok}@umiacs.umd.edu. The authors were partially supported by NSF-ITR 03-24313.

In terms of mathematical modeling, the mixture of models paradigm described above can be easily incorporated in the traditional dynamical system framework by incorporating a discrete state that acts as a switch and chooses the appropriate state transition and observation model. Systems whose state space have one component taking discrete values and the rest taking continuous values are called *mixed state space* systems. In this paper, we show that mixed state space systems are powerful tools for various problems in tracking, recognition and behavior analysis. In problems in automatic recognition of targets, the *switch* encodes the identity of the target, and thereby controls the observation model of the dynamical system so as to match the person's identity and appearance. In behavior analysis, the discrete state governs the motion exhibited by the target, and thereby chooses among a set of motion models that best fits the observed data.

In this paper, we focus on the role and importance of mixed state space systems for problems in vision. In particular, we emphasize the simultaneous track-and-recognize paradigm that naturally arises out of multi-model systems. We show the use of such models in recognizing faces and for verification of vehicles across non-overlapping views by building structural and textural fingerprints. Finally, we discuss the problem of tracking social insects by using dynamical systems that encode the behavior of the insects.

## 2. MIXED STATE SYSTEMS

Consider the mixed state space $\mathscr{X} = \mathbb{R}^d \times \mathscr{D}_N$, where $\mathscr{D}_C = \{1, \ldots, C\}$ defines the discrete part of the state space (cf. Isard and Blake[1]). We defined the state at time $t$ as $\mathbf{X}_t = (\mathbf{x}_t, \theta_t)$, with $\mathbf{x}_t \in \mathbb{R}^d$ and $\theta_t \in \mathscr{D}_C$. As mentioned earlier, the discrete state $\theta$ governs the model selection, defining the specifics of state transition and/or the observation model.

Our main objective is one of *Bayesian Inference*, where given a set of observations $\mathscr{Y}_t = \{y_1, \ldots, y_t\}$ we are interested in estimating the posterior probability density $\pi_t = p(\mathbf{X}_t|\mathscr{Y}_t) = p(\mathbf{X}_t|y_1, \ldots, y_t)$. As mentioned earlier, in this paper we focus on mixed state systems are useful in behavioral analysis and ATR where multiple dynamical systems arise naturally. In such settings, the specific inferences of interest are in estimating the posterior probability of a particular behavior or target identity given the observed data. Simultaneously, we also need to consider the dual problem of tracking the object whose identity or behavior we are interested in. These two inference problems correspond to the marginals of the posterior density $\pi_t$. The marginals of interest are defined as,

$$p(\theta_t|y_{1:t}) = \int p(\mathbf{X}_t|y_{1:t}) d\mathbf{x}_t \tag{1}$$

and

$$p(\mathbf{x}_t|y_{1:t}) = \int p(\mathbf{X}_t|y_{1:t}) d\theta_t \tag{2}$$

However, these two inferences are coupled and in most scenarios we cannot estimate them independently. *This coupling is central to the applications illustrated in this paper**. We use *Bayes' law* to write the expression for $\pi_t$:

$$p(\mathbf{X}_t|y_{1:t}) \propto p(y_t|\mathbf{x}_t\theta_t) \int p(\mathbf{X}_t|\mathbf{X}_{t-1}) p(\mathbf{X}_{t-1}|y_{1:t-1}) d\mathbf{X}_{t-1} \tag{3}$$

The state transition model $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ can be written as

$$p(\mathbf{X}_t|\mathbf{X}_{t-1}) = p(\mathbf{x}_t\theta_t|\mathbf{X}_{t-1}) = p(\mathbf{x}_t|\theta_t\mathbf{X}_{t-1}) p(\theta_t|\mathbf{X}_{t-1}) \tag{4}$$

The dynamics of the system is defined completely by (3) and (4), defining the following sub-components.

- **Multiple Motion/Obs. Models:** For each value of $\theta_t \in \mathscr{D}_C$, we define a dynamical system characterized by the state model $p(\mathbf{x}_t|\mathbf{x}_{t-1}\theta_t)$ and the observation model $p(y_t|\mathbf{x}_t\theta_t)$. Given that $\theta_t$ can take $C$ values, we have $C$ such dynamical systems, each characterizing a different mode of the underlying stochastic process.

*Later in the paper, we refer to this coupling as the tracking-and-recognition framework.

- **Model Transition:** The switching between the $C$ dynamical systems is controlled by the term $p(\theta_t|\mathbf{X}_{t-1})$, the term that concerns the model transition probabilities.

In most vision problems, the models defined above are (in general) non-linear and non-Gaussian, making the evaluation of the integrals in (3),(1) and (2) analytically intractable. However, we can still solve the inference problem by assuming a sample based representation for the posterior probability density $\pi_t$. The sample set $S_t = \{\mathbf{X}_t^{(i)}, w_t^{(i)}\}_{i=1}^N$ forms an approximation $\hat{\pi}_t$ for the posterior $\pi_t$ given by,

$$\hat{\pi}_t = \sum_{i=1}^{N} w_t^{(i)} \delta(\mathbf{X}_t - \mathbf{X}_t^{(i)}) \tag{5}$$

where $\mathbf{X}_t^{(i)}$ are termed *particles* and $w_t^{(i)}$ are the weights associated with the particles. A particle based representation for the posterior density allows for a recursive computation of the posterior (or its approximation). The algorithm for doing this is popularly referred to as particle filtering.[2–4] In this paper, we do not discuss the algorithm in any more detail. The interested reader is referred to the literature on particle filtering algorithms.

In the following sections, we discuss the applicability of such mixed state systems for various problems in recognition and behavioral analysis.

## 3. SIMULTANEOUS TRACKING AND RECOGNITION

Automatic target recognition has long been a topic of research interest. In this section, we first present a mixed state space approach for simultaneous tracking and recognition of human faces. We then extend this approach to identifying vehicles across non-overlapping views using its structure and texture as a *fingerprint* encoding its identity.

### 3.1 Simultaneous Tracking-Recognition for Face Recognition From Videos

Consider the problem of tracking the face of a human from a video. Simultaneously, we are also given a gallery of faces and are interested in identifying the person in the query video. The paradigm of simultaneous tracking and recognition attempts to resolve the uncertainty in the identity along with the uncertainty in the tracking.[5]

In Zhou et al.,[5,6] an affine state space is formulated for the tracking problem. The affine state encodes the deformation of a fixed rectangular[†] template of known dimensions onto the image plane.

A gallery for training is available, and features characterizing each person in the gallery is learnt. Typical features for recognition could be PCA coefficients for each person, or a FLD based classification scheme. The particle filtering framework allows for a wide variety of models to be used. In this paper, we use a single template to characterize each person's identity.

With this we can formally define the models for the tracking-and-recognition problem. Let the number of persons in the gallery be $C$. Let $A_i, i = 1, \ldots, C$ be the appearance model characterizing the identity of each person in the gallery. The mixed state space $\mathscr{X} = \mathbb{R}^6 \times \mathscr{D}_C$ encompasses the six-dimensional affine state space as well the $\mathscr{D}_C$, the discrete state for the target identity. The state transition model is given as,

$$p(\mathbf{X}_t|\mathbf{X}_{t-1}) = p(\mathbf{x}_t\theta_t|\mathbf{x}_{t-1}\theta_{t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\theta_t|\theta_{t-1}) \tag{6}$$

Here, we make the assumption that dynamics exhibited by the object characterized in $\mathbf{x}_t$ is independent of the target identity. This is indeed true for the problem of face recognition, as we do not observe any identity specific motion in humans (allowing for some rare exceptions). We also assume a first order Markov model for the affine deformation $\mathbf{x}_t$ and a random switching model for the identity state.

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) \backsim \mathrm{N}(\mathbf{x}_{t-1}|\Sigma_{\mathbf{x}}) \tag{7}$$

[†]The shape of the template can be chosen arbitrarily depending the domain of the problem. It is important however to know this shape.

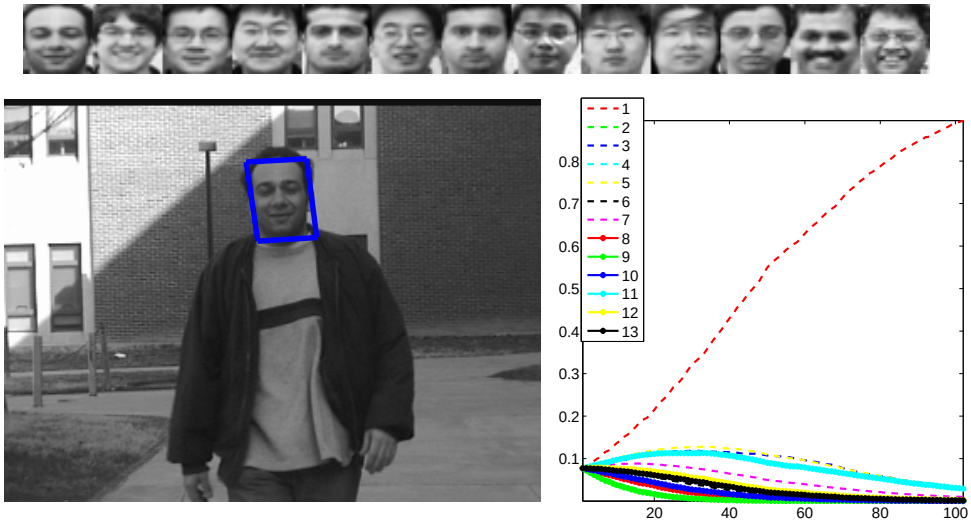


Figure 1. Simultaneous tracking and recognition of faces from a video sequence. (top) Face templates of the 13 persons making the gallery set (left) Snapshot of tracking results (right) Probability scores for recognition, $p(\theta_t|y_{1:t})$. Here the index for identity corresponds to the same as the ordering of faces in the gallery.

$$p(\theta_t|\theta_{t-1}) = c_0\delta(\theta_t - \theta_{t-1}) + \frac{1-c_0}{C-1}\sum_{j=1,j\neq\theta_{t-1}}^{C}\delta(\theta_t - j), 0 < c_0 < 1 \tag{8}$$

The evolution of the identity state $\theta_t$, described in (8), is given by a simple rule. At each time instant, the identity is retained with probability $0 < c_0 < 1$ and randomly switches to another value with probability $(1-c_0)$.

The observation model is defined as:

$$\begin{aligned} z_t &= \mathscr{T}(y_t, \mathbf{x}_t) \\ p(y_t|\mathbf{x}_t\theta_t) &= p(z_t|A_{\theta_t}) \end{aligned} \tag{9}$$

where $\mathscr{T}(y_t, \mathbf{x}_t)$ defines the region on the image $y_t$ defined by the affine deformation $\mathbf{x}_t$, suitably rearranged as a rectangular template. Finally, the $p(z_t|A_{\theta_t})$ defines the matching probability between the tracked region (defined in $z_t$) and the model for recognition that is encoded in the appearance template $A_{\theta_t}$.

### 3.2 Results

We applied the mixed state formulation for recognizing faces from a video sequence. The gallery consisted of 13 faces in a front view. The probe videos had subjects walking towards the camera. Figure 1 summarizes the result for this experiment. Note the plot of $p(\theta_t|y_{1:t})$ (shown in Figure 1). The confidence score of the correct gallery template rises slowly to 1 while the others reduce to 0. Additional results for various tracking and recognition models can be found in Zhou et al.[5,6]

## 4. FINGERPRINTING VEHICLES FOR TRACKING AND RECOGNITION ACROSS NON-OVERLAPPING VIEWS

In the previous section, we showed how mixed state space models could be used to simultaneously track and recognize faces from a video. In this section, we extend this methodology for tracking objects across non-overlapping views. In this particular setting, we need to account for possible changes in *pose* and *illumination* across the views.

To motivate this problem, consider a large area that is observed by a set of cameras. Practical constraints on resources (including power, sensor cost, connectivity) lead to a sparse deployment of the cameras. This creates regions in the sensing area where there is no sensor coverage. Inspite of this temporary loss of coverage tracking

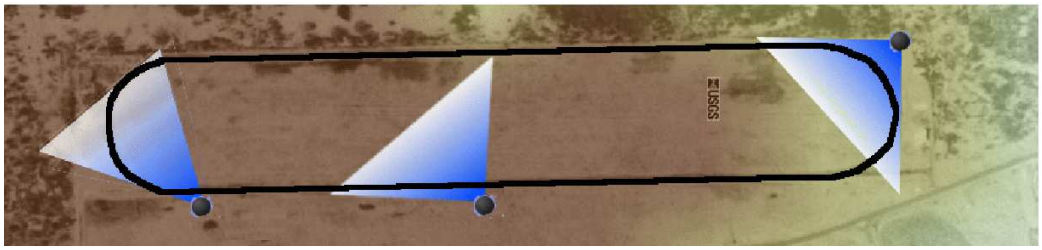


Figure 2. Location of cameras at the data collection held in May 2005, Aberdeen, MD.

algorithms must be able to sense, identify and tag previously encountered targets. Augmented tracking of targets over a sparse network of cameras requires a seamless exchange of sufficient statistics characterizing target identity between cameras.

Consider a sparse distribution of cameras (for example Figure 2) covering a large area with blind regions in the coverage of the target movement. Establishing the identity of the target across these non-overlapping views robustly can be of enormous significance in answering important inference problems. For example, lets suppose a White SUV is seen approaching a checkpoint, and we would like to know if such a vehicle was involved in some critical event in the past. Suppose, a watch-list of such *wanted* vehicles is available with appropriate descriptions, then we could verify the approaching target across the entries in the watch-list. Verification of target identity across non-overlapping views presents two important challenges: pose and illumination. The models built for each target needs to account for possible changes in both. The variations in pose can be addressed explicitly by estimating the 3D structure of the vehicle. Coupling this with statistical models to characterize the appearance of the target, we form the fingerprint representing the target. Estimation of the 3D structure of the target is performed using a factorization approach specifically suited to targets exhibiting planar motion. We embed these models in the mixed state space to simultaneously track the vehicle in an Euclidean scene coordinates and verify its identity.

Estimating structure from observed motion of a set of feature points is a classical vision problem[7,8].[9] The factorization approach to SfM is very powerful and elegant algorithm as it solves for both the structure and motion using just singular value decomposition. In our particular setting, the targets (vehicles) are constrained to move on the ground plane. This introduces the so called *planar motion* constraint that can be used to design a lower rank factorization approach. We use this structure from planar motion approach to estimate the structure of a vehicle observed in a particular view. Further, we can use the observed data to *learn* a statistical model for the associated texture map. This forms the fingerprint for the vehicle. We next describe the models that comprise the mixed state dynamical system for tracking and verification.

### 4.1 System Formulation

We first assume that for each vehicle in our *watch-list*, there is a structure estimate $S^{\ddagger}$ and a texture map $\mathscr{I}$. Given this, the state space for tracking is the location and velocity of the vehicle on a world coordinate system (Euclidean). Let $\mathbf{x}_t = (x_t, y_t, v_t, \alpha_t)$, where $(x_t, y_t)$ is the location on the ground plane, $v_t$ is the speed and $\alpha_t$ is the headings direction in the world coordinate system. Such a world-centric coordinate system allows us to *place* a synthetic target into the image plane. Let the camera projection matrix be given the $3 \times 4$ matrix $P$. We can now define a constant velocity motion model for the vehicle,

$$\mathbf{x}_t = \begin{pmatrix} x_t \\ y_t \\ v_t \\ \alpha_t \end{pmatrix} = \begin{pmatrix} x_{t-1} + v_{t-1} cos(\alpha_{t-1}) \\ y_{t-1} + v_{t-1} sin(\alpha_{t-1}) \\ v_{t-1} \\ \alpha_{t-1} \end{pmatrix} + \omega_t, \omega_t \backsim \mathrm{N}(\mathbf{0}, \Sigma_{\mathbf{x}}) \tag{10}$$

[‡]The representation of $S$ is a key design issue. For the experiments in this paper, we chose a dense point cloud model for $S$. Alternate representation offer different trade-offs between computational requirements and ease in estimation and notation.

The identity state $\theta_t$ follows an evolution governed by the random switching model defined in (8). And the evolution of the tracking states and identity state is assumed independent (as in (6)).

The observation model for this problem involves the interplay between the structural and textural models, suitable projected onto the scene and the output compared with the acquired imagery. As mentioned before, for each identity $\theta \in \{1, \ldots, C\}$, we have $S_\theta$ and $\mathscr{I}_\theta$, the structure and texture respectively. Given a state $\mathbf{x}_t$, we can define a $3 \times 3$ rotation matrix $R(\mathbf{x}_t)$ and a 3D translation $T(\mathbf{x})$ that correspond to a Euclidean transformation of the structure from its reference pose to the location and orientation prescribed by the state. Finally, given the camera projection matrix $P$ , we can compute the overall transformation,

$$P(\mathbf{x}_t) = P \begin{bmatrix} R(\mathbf{x}_t) & T(\mathbf{x}_t) \\ \mathbf{0}^T & 1 \end{bmatrix} \tag{11}$$

Using this transformation, the structure defined in $S_{\theta_t}$ can be rendered onto the image plane by the transformation $P(\mathbf{x}_t)S_{\theta_t}$ (linear in the homogeneous coordinates). Finally, as in the face recognition case, we can define the observation model.

$$p(y_t|\mathbf{x}_t\theta_t) = p(y_t|\mathbf{x}_t, S_{\theta_t}, \mathscr{I}_{\theta_t}) = p(y_t|P(\mathbf{x}_t)S_{\theta_t}, \mathscr{I}_{\theta_t}) \tag{12}$$

### 4.2 Results

The algorithm was tested on a convoy of 5 vehicles comprising two HUMMVs (one truck and one van) and three SUVs of different colors (white, white/gray and blue). This dataset was collected at Aberdeen in May 2005 and the positioning of the video cameras is shown in Figure 1. External parameters of each camera was estimated using the horizon line and a calibration pattern was used to estimate the intrinsic parameters. The test setup was as follows: the fingerprints for the five vehicles were built at the camera at the center of the oval track. These fingerprints were then used for verification at the remaining two cameras.

Figure 3 shows estimates of the 3D structure (with overlaid texture maps) of three vehicles. Figure 4 shows the tracking results of a white SUV. Tracking is performed over the ground plane simultaneously with verification. Superimposed on the frame, in magenta, is the 3D model that matches best. Figure 4 also shows the recognition scores for this experiment. It is seen that the target is recognized correctly as a White SUV and high confidence values are achieved within a second (30 frames).

Figure 3. 3D models from three different targets. (Image courtesy of Sankaranarayanan et al.[10])

## 5. SIMULTANEOUS TRACKING AND BEHAVIOR ANALYSIS

Behavioral research in the study of the organizational structure and communication forms in social insects such as ants and bees has received much attention in recent years.[11,12] Such a study has provided some practical models for tasks like work organization, reliable distributed communication, navigation etc.[13,14] Usually, when an experiment to study these insects is setup, the insects in an observation hive are videotaped. The hours of video data are then manually studied and hand-labeled. This task of manually labeling the video data takes up the bulk of the time and effort in such experiments. In this section, we discuss general methodologies for

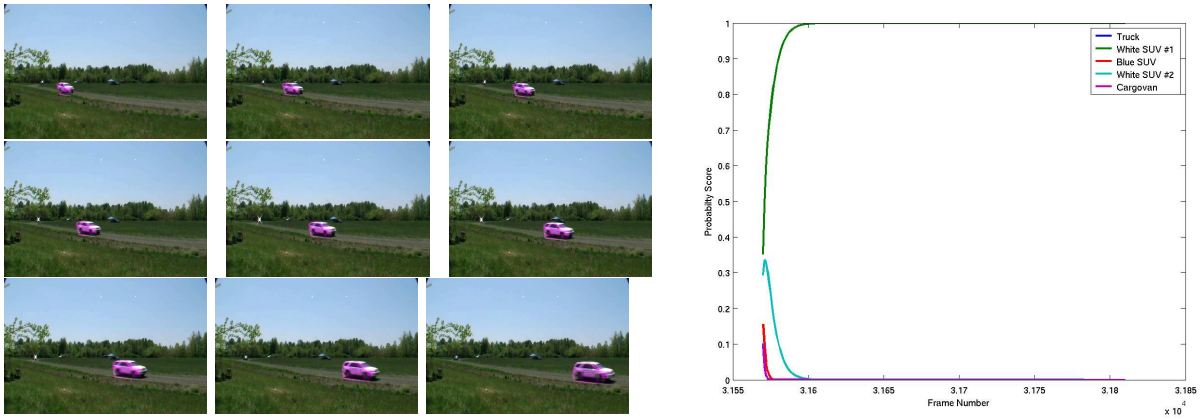


Figure 4. (left) Tracking and Verification of a white SUV. Superimposed in magenta is the projection of the "best" 3D model from watch-list. (right) Confidence plots for recognition associated with tracking results. (Image courtesy of Sankaranarayanan et al.[10])

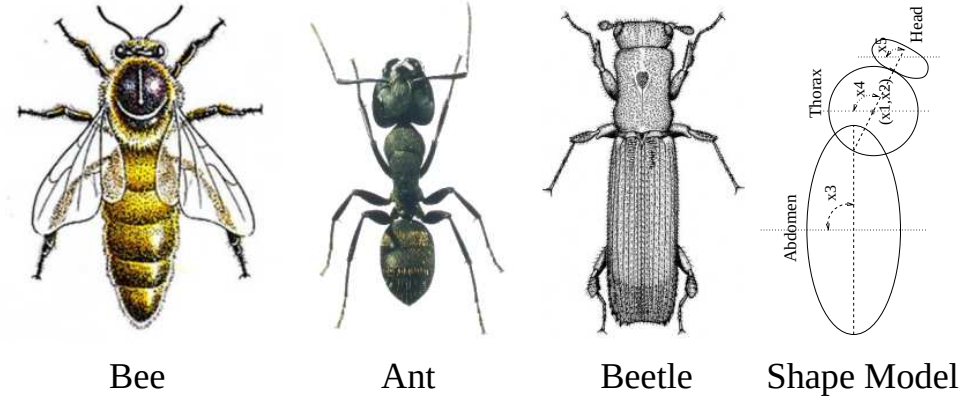


Figure 5. A Bee, an Ant, a Beetle and Shape Model

automatic labeling of such videos and provide an example by following the approach for analyzing the movement of bees in a bee hive. As in the earlier sections, we track and recognize the behavior exhibited by the bee simultaneously. In such a joint approach, accurate modeling of behaviors act as priors for motion tracking and significantly enhances motion tracking while accurate and reliable motion tracking enables behavior analysis and recognition.

### 5.1 Anatomical Modeling and State Space for Tracking

Modeling the anatomy of insects is very important for reliable tracking, because the structure of their body parts and their relative positions present some physical limits on their possible relative orientations. In spite of their great diversity, the anatomy of most insects is rather similar. The body is divided into three main parts: the head, thorax and the abdomen. Figure 5 shows the image of a bee, an ant and a beetle. Though there are individual differences in their body structure, the three main parts of the body are evidently visible. Each of these three parts can be regarded as rigid body parts for the purposes of video based tracking. Most insects also move towards the direction of their head. Therefore, during specific movements such as turning, the orientation of the abdomen usually follows the orientation of the head and the thorax with some lag.

We model the bees with three ellipses, one for each body part. We neglect the effect of the wings and legs on the bees. Figure 5 shows the shape model of a bee. The location of the bee and its parts in any frame can be given by five parameters: the location of the center of the thorax (2 parameters), the orientation of the head, the orientation of the thorax and the orientation of the abdomen (refer Figure 5). Tracking the bee over a video essentially amounts to estimating these five model parameters, $\mathbf{u} = [x_1\ x_2\ x_3\ x_4\ x_5]$ for each frame. This

5−parameter model has a direct physical significance in terms of defining the location and the orientation of the various body parts in each frame.

### 5.2 Behavior Modeling

Insects, especially social insects like bees and ants, exhibit rich behaviors. Modeling such behaviors explicitly is helpful in accurate and robust tracking. Moreover, explicitly modeling such behaviors also leads to algorithms where position tracking and behavior analysis are tackled in a unified framework. Following the premise that the mixed state space model can be used to generate complicated dynamics, we model the basic motions as a vocabulary of local motions. These basic motions are then regarded as states and behaviors are modeled as being Markovian on this motion state space. Once each specific behavior has been modeled as a Markov process, then our tracking system can simultaneously track the position and the behavior of insects in videos.

We model the probability distributions of location parameters $\mathbf{u}$ 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 Gaussian $(p_{m2})$ is used for modeling the turning motion (to accommodate the two possible turning directions).

$$p_{mi}(\mathbf{u}_t|\mathbf{u}_{t-1}) = \mathrm{N}(\mathbf{u}_{t-1} + \vec{\mu_{mi}}, \Sigma_{mi}); \text{ for } i = 1, 3, 4. \tag{13}$$

$$p_{m2}(\mathbf{u}_t|\mathbf{u}_{t-1}) = 0.5\mathrm{N}(\mathbf{u}_{t-1} + \vec{\mu}_{m2}, \Sigma_{m2}) + 0.5\mathrm{N}(\mathbf{u}_{t-1} - \vec{\mu}_{m2}, \Sigma_{m2}) \tag{14}$$

Each behavior $\theta = \{1, \ldots, C\}$ is now modeled as a Markov process of order $K_\theta$ on these motions, i.e.,

$$\mathbf{s}_t = \sum_{k=1}^{K_\theta} A_\theta^k \mathbf{s}_{t-k} \tag{15}$$

where $s_t$ is a vector whose $j^{th}$ element is the probability that the bee is in the motion state $m_j$ and $K_\theta$ is the model order for the behavior indexed by $\theta$. The parameters of each behavior model are made of autoregressive parameters $A_\theta^k$ for $k = 1, \ldots, K_\theta$.

We model three different behaviors: the waggle dance, the round dance and a stationary bee using a first order Markov model. For illustration, we discuss the manner in which the waggle dance is modeled. Figure 6 shows the trajectory followed by a bee during a single run of the waggle dance. It also shows some followers who follow the dancer but do not waggle. A typical Markov model for the waggle dance is also shown in Figure 6.

### 5.3 System Modeling

As before, we address the tracking problem as a problem of estimating the state $\mathbf{X}_t = (\mathbf{x}_t, \theta_t)$ given the image observations $y_{1:t}$. The state $\mathbf{x}_t = (\mathbf{u}_t, \mathbf{s}_t)$ encompasses the basic motion states (encoded in $\mathbf{s}_t$ as well the location of the bee on the image plane encoded in $\mathbf{u}_t = (x_1, \ldots, x_5)_t$. The state transition model characterized by the density $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ is defined as,

$$p(\mathbf{X}_t|\mathbf{X}_{t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1}\theta_t)p(\theta_t|\mathbf{x}_{t-1}\theta_{t-1}) \tag{16}$$

The term $p(\theta_t|\mathbf{x}_{t-1}\theta_{t-1})$ controls the switching of behavior and is learnt from training data. The dynamics of tracking defined in $p(\mathbf{x}_t|\mathbf{x}_{t-1}\theta_t)$ is defined by (13), (14) and (15).

Given the location of the bee in the current frame $\mathbf{u}_t$ and the image observation given by $y_t$, we first compute the appearance $z_t = \mathscr{T}(y_t, \mathbf{u}_t)$ of the bee in the current frame (i.e., the color image of the three ellipse anatomical model of the bee ).

$$p(y_t|\mathbf{x}_t\theta_t) = p(y_t|\mathbf{u}_t) = p(z_t|A_1, \ldots, A_5) \tag{17}$$

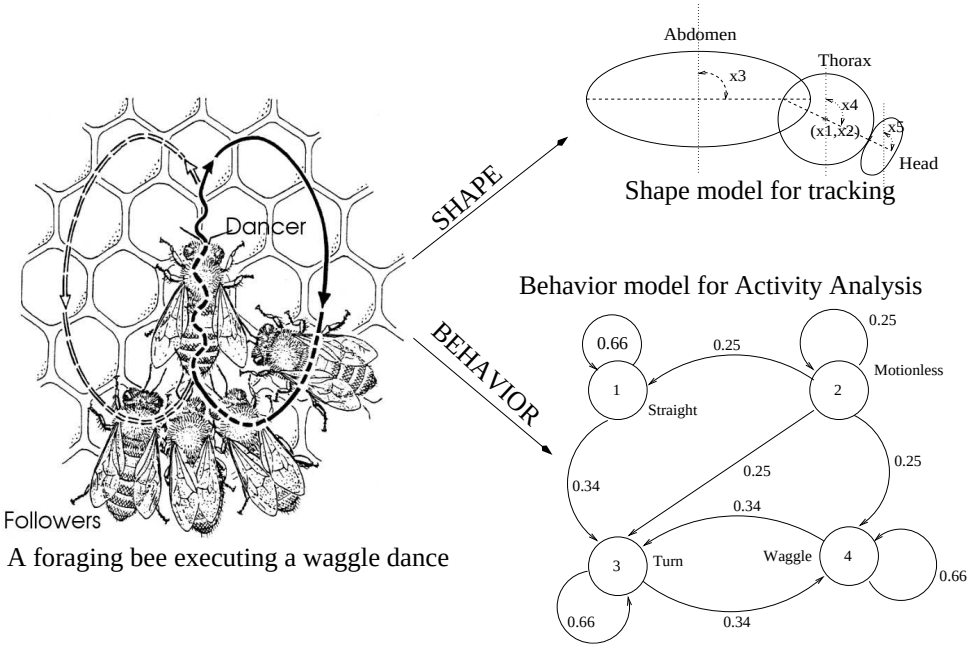


Figure 6. A Bee performing a waggle dance and the behavioral model for the Waggle dance

where the $A_1, \ldots, A_5$ are color-templates that form the exemplars for the appearance of the bee. The RGB components of color are treated independently and identically. The appearance of the bee in any given frame is assumed to be Gaussian centered around one of these five exemplars, i.e.,

$$P(z_t|A_1, \ldots, A_5) = \frac{1}{5}\sum_{i=1}^{i=5} \mathrm{N}(A_i, \Sigma_i) \tag{18}$$

where $N(Z; A_i, \Sigma_i)$ stands for the Normal distribution with mean $A_i$ and covariance $\Sigma_i$.

### 5.4 Results

We conducted tracking experiments on video sequences of bees in a hive. In the videos, the bees exhibited three behaviors: the waggle dance, the round dance and a stationary bee. The video sequences ranged between 50 frames to about 700 frames long. It is noteworthy that when a similar tracking algorithm without a behavioral model was used for tracking, it lost track within 30-40 frames.With our behavior-based tracking algorithm, we were able to track the bees during the entire length of these videos. We were also able to extract parameters like the orientation of the various body parts during each frame over the entire video sequences. We used these parameters to automatically identify the behaviors. We also verified this estimate manually and found it to be robust and accurate.

Figure 7 shows the structural model of the tracked bee superimposed on the original image frame. In this particular video, the bee was exhibiting a waggle dance. As is apparent from the sample frames the appearance of the dancer varies significantly within the video. These images display the ability of the tracker to maintain track even under extreme clutter and in the presence of several similar looking bees. Frames 30-34 show the bee executing a waggle dance. Notice that the abdomen of the bee waggles from one side to another.

We validated a portion of the tracking result by comparing it with a "ground truth" track obtained using manual ("point and click") tracking by an experienced human observer. We find that the tracking result obtained using the proposed method is very close to manual tracking. The mean differences between manual and automated tracking using our method are given in Table 1. The positional differences are small compared the average length of the bee, which is about 80 pixels (from front of head to tip of abdomen).

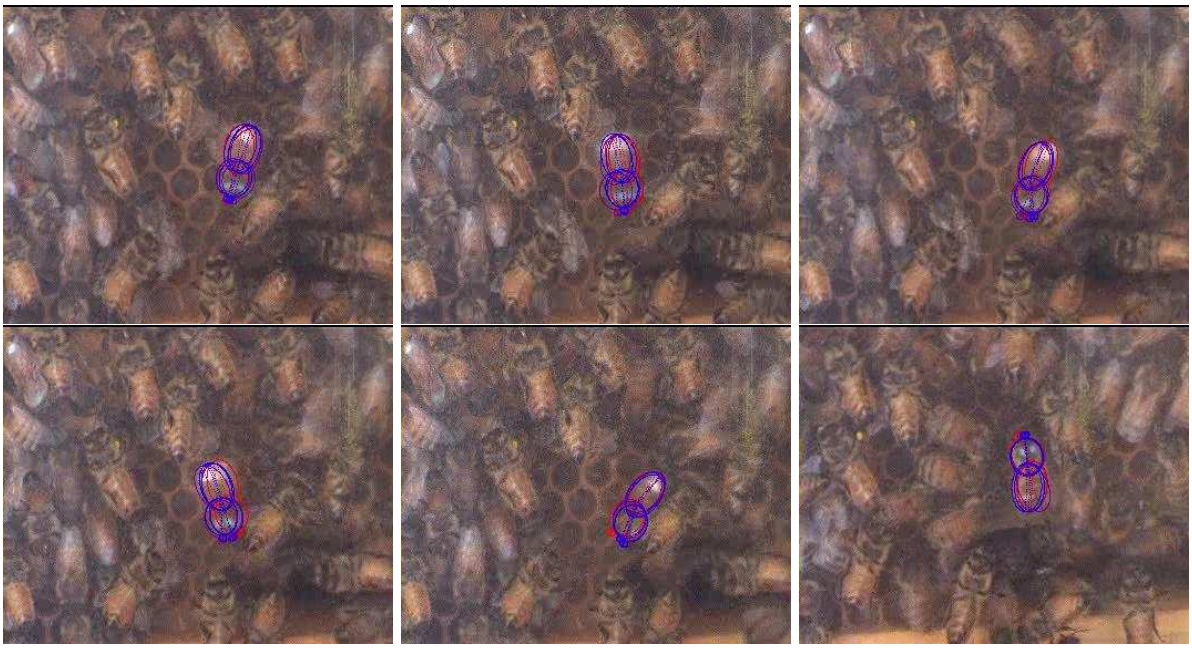

Figure 7. Sample Frames from a tracked sequence of a bee in a beehive. Images show the top 5 particles superimposed on each frame. Blue denotes the best particle while red denotes the fifth best particle. Frame Numbers row-wise from top left :30, 31, 32, 33, 34 and 90. Figure best viewed in color. (Image courtesy of Veeraraghvan et al.[15])

Table 1. Comparison of our tracking algorithm with Ground Truth

| | Average positional difference between Ground Truth and our algorithm |
|---|---|
| Center of Abdomen | 4.5 pixels |
| Abdomen Orientation | 0.20 radians (11.5 deg) |
| Center of Thorax | 3.5 pixels |
| Thorax orientation | 0.15 radians (8.6 deg) |

Figure 8 shows the estimated orientation of the abdomen and the thorax in a video sequence of around 600 frames. The orientation is measured with respect to the vertically upward direction in each image frame and a clockwise rotation would increase the angle of orientation while an anticlockwise rotation would decrease the angle of orientation. The waggle dance is characterized by the central waggling portion which is immediately followed by a turn, a straight run another turn and a return to the waggling section as shown in Figure 6. After every alternate waggling section the direction of the turning is reversed. This is clearly seen in the orientation of both abdomen and the thorax. The sudden change in slope (from positive to negative or vice-versa) of the angle of orientation denotes the reversal of turning direction. During the waggle portion of the dance, the bee moves its abdomen from one side to another while continuing to move forward slowly. The large local variation in the orientation of the abdomen just before every reversal of direction shows the waggling nature of the abdomen. Moreover, the average angle of the thorax during the waggle segments denotes the direction of the waggle axis.

## 6. CONCLUSIONS

In this paper, we discuss the role of mixed state systems for recognition and behavior analysis. In particular, we show the use of such systems for recognition of faces and vehicles from appearance information and analysis of bee dances from their motion patterns. The mixed state dynamical system forms a elegant mathematical framework for coupling the inference problems of tracking and recognition/behavior analysis.

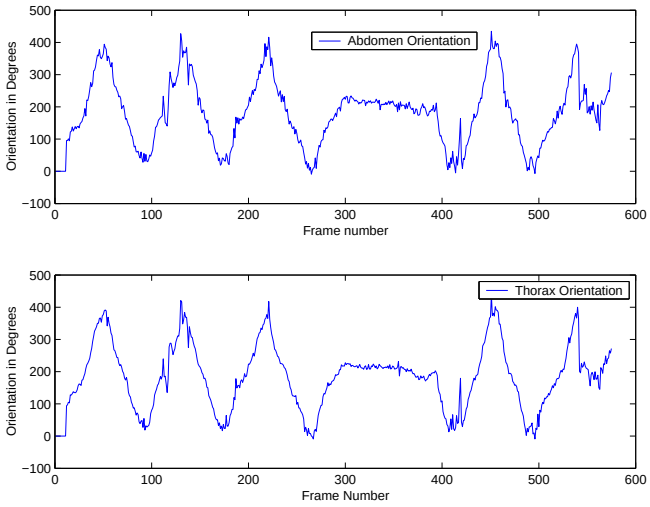


Figure 8. The orientation of the Abdomen and the Thorax of a bee in a video sequence of about 600 frames. (Image courtesy of Veeraghavan et al.[15])

## REFERENCES


[1] Isard, M. and Blake, A., "A mixed-state condensation tracker with automatic model-switching," *Computer Vision, 1998. Sixth International Conference on* , 107–112 (1998).

[2] Doucet, A., Freitas, N. D., and Gordon, N., [*Sequential Monte Carlo methods in practice*] (2001).

[3] Liu, J. S. and Chen, R., "Sequential monte carlo methods for dynamic systems," in [*Journal of American Statistician Association*], **93**, 1032–1044 (1998).

[4] Gordon, N. J., Salmond, D. J., and Smith, A. F. M., "Novel approach to nonlinear/non-gaussian bayesian state estimation," in [*IEE Proceedings on Radar and Signal Processing*], **140**, 107–113 (1993).

[5] Zhou, S., Krueger, V., and Chellappa, R., "Probabilistic recognition of human faces from video," *Computer Vision and Image Understanding* **91**, 214–245 (2003).

[6] Zhou, S., Chellappa, R., and Moghaddam, B., "Visual tracking and recognition using appearance-adaptive models in particle filters," *IEEE Trans. on Image Processing* (November 2004).

[7] Broida, T. J., Chandrashekhar, S., and Chellappa, R., "Recursive 3-D motion estimation from a monocular image sequence," *Aerospace and Electronic Systems, IEEE Transactions on* **26**(4), 639–656 (1990).

[8] Tomasi, C. and Kanade, T., "Shape and motion from image streams under orthography: a factorization method," *International Journal of Computer Vision* **9**(2), 137–154 (1992).

[9] Oliensis, J., "Critique of structure-from-motion algorithms," *Computer Vision and Image Understanding* **80**(2), 172–214 (2000).

[10] Sankaranarayanan, A., Li, J., and Chellappa, R., "Fingerprinting vehicles for tracking across non-overlapping views.," *Army Science Conference* (2006).

[11] V. Frisch, *The Dance Language and orientation of bees.* Cambridge MA:Harvard University Press, 1993.

[12] M. Srinivasan, S. Zhang, M. Lehrer, and T. Collett, "Honeybee navigation en route to the goal: visual flight control and odometry," *Journal of Experimental Biology*, vol. 199, pp. 237–244, 1996.

[13] T. Neumann and H. Bulthoff, "Insect inspired visual control of translatory flight," *Proceedings of the 6th European Conference on Artificial Life ECAL 2001*, pp. 627–636, 2001.

[14] F. Mura and N. Franceschini, "Visual control of altitude and speed in a flight agent," *Proceedings of 3rd International Conference on Simulation of Adaptive Behaviour: From Animal to Animats*, pp. 91–99, 1994.

[15] Veeraraghavan, A., Chellappa, R., and Srinivasan, M., "Shape-and-Behavior Encoded Tracking of Bee Dances," *Transactions on Pattern Analysis and Machine Intelligence* **30**(3), 463–476 (2008).