Chapter 7

CUBA: Caltech Unsupervised Behavior Analysis

Up to now, Machine Learning has permeated behavioral biology with methods as the one described in the previous chapter, where the task of classifying fractions of videos into specific behaviors fits perfectly under the supervised learning scenario of learning systems. Yet, unsupervised methods had not been widely applied to this problem. A recently published method analyzes the locomotion of fly larvae using unsupervised methods [69]. The study aims to form a taxonomy of the types of behaviors that can be observed in these larvae. Another method analyzes Caenorhabditis elegans worms and the study introduced the so-called eigen-worm shapes. These shapes constitute a dictionary of behavioral motifs that are able to describe the locomotion of the worms. The description in this space constitutes a fingerprint for each of the worms. The fingerprints are then compared between different mutant genetic lines and wild types of worms.

Nevertheless, these two methods were the only studies found in the literature that apply unsupervised learning methods to the problem of behavioral biology. In this chapter, we study the problem of analyzing behavior in an unsupervised manner. The goal is to be able to analyze automatically large quantities of videos, and draw conclusions in an unbiased way, by letting data speak for itself. The goal is that the method will aid biologists in testing hypotheses, as well as discovering patterns in the data that were previously undetected.

This study led to a system which we call CUBA (Caltech Unsupervised Behavior Analysis). The method extracts basic units that describe the motion of the animals, which we denote as movemes. A moveme, as defined in [5] is the simplest pattern that is associated with a behavior. Movemes are then compounded to form actions. Finally, a concatenation of actions forms an activity. CUBA is able to extract such movemes. The abstraction can then be iterated to form actions. We then compare the behavior of animals in this space, and cluster them into meaningful groups that perform the same activity. With this abstraction, it is possible to understand the patterns that arise in biological

experiments, as well as to use this as a mathematical tool to test hypotheses and discover new patterns. One of the achievements of CUBA is that it was able to discriminate, without supervision, different genetic lines of flies.

We describe CUBA in the following sections. In the first section, we start by defining the problem more concretely, and explain the goal in the context of a few biological experiments where we applied our work. We then describe the various methods used from Machine Learning that constitute CUBA. Finally, we present the results obtained in two different datasets.

7.1 Problem statement

The goal of the method is to analyze behavior of animal videos, using unsupervised machine learning techniques. The term behavior analysis may be ambiguous, unless we specify exactly what we mean by it. Ethology, the branch of biology that studies behavior, is concerned with various aspects of behavior, such as descriptive, causal, genetic, and evolutionary [5]. The first step in the analysis is therefore to be able to have an accurate description of each behavior. With such descriptions it is then possible to test hypotheses about causality, or to find differences between genetic lines, analyze evolutionary changes, etc. Hence, we focus first on the descriptive aspect.

When describing behavior, it is important to answer basic questions such as what is happening, how it is happening, and where it is happening. These answers vary depending on the time scale used in the description. Hence the description of behavior should begin by identifying the simplest meaningful patterns or movemes, and then be able to group movemes into meaningful structured actions. At a larger time scale, actions should be grouped into activities or stories. This hierarchy is analogous to the one used in natural language processing. When we want to understand a sentence, we would like to begin by discovering phonemes, then words, then full sentences, and so on.

With this in mind, our definition of behavior analysis can be decomposed in a few steps. First, given a set of time series that represent trajectories of animals, we want to identify typical coherent, meaningful patterns at various time scale resolutions. Second, we would like to find a hierarchy that allows grouping these patterns into more complex patterns in order to understand behavior at larger time scales. Finally, having abstracted the trajectories in this new space of meaningful patterns, we want to detect common patterns across individuals, detect clusters, find outliers, etc.

To clarify our goal, we describe a biological experiment that was the subject of study in [35]. The experiment aimed to understand if flies actually showed a fear-like response, when an arousing stimulus such as a shadow was repetitively presented. The experiment varied parameters such as the number and frequency of shadows presented, as well as evaluated different conditions such as having the experiment with single and multiple flies at a time, or having an additional attractive resource

such as food in the experiment. Taking this experiment as an example, a successful unsupervised analysis would first identify the movemes which are very basic units that describe the locomotion of the fly, such as slow walking, fast walking, accelerating, resting, hopping, etc. Putting together these movemes, actions can start to emerge. For example the action of escaping can be formed with the movemes of accelerating–flying/hopping–decelerating. An action like feeding could be described by movemes of slow walking–stopping (on food)–slow walking. Finally, at a higher level, an activity or story could describe a fly feeding until shadows pass, then jumping off and moving to the open space, before finally returning to the food after spending some time in the boundary enclosing the arena.

Once we have a full description of what each fly is doing, the abstraction could allow finding similar types of flies, by looking at flies that conform to the same story. Clustering similar flies would help determining what are the typical behaviors of flies in the experiment. Also, it would be possible to identify flies that are behaving differently from the norm, or perhaps identify previously unseen behaviors. This abstraction is therefore useful in summarizing and describing what is occurring in the experiment. Furthermore, the analysis becomes a tool to both formulate and test new hypotheses to understand what is occurring in the experiment.

The above analysis can be done with any type of biological experiment where trajectories of animals can be obtained using a computer vision tracking system. The goal in each experiment might be different, but the descriptive analysis provided by the system will remain the same. Giving biologists the power to summarize and cluster data, as well as to detect patterns in the experiments, is the ultimate goal of the system. Having clarified this goal, we present in the next section the various techniques used in order to construct such a system.

7.2 The method

As described before, the system performs the behavior analysis at three different levels. First it must discover meaningful patterns in the data at the smallest time scale. Then, it must be able to group these patterns or movemes into more complex actions and stories. Finally, the system summarizes the data, clusters trajectories into meaningful groups, detects outliers, etc. We present now the methods used at each of these stages.

7.2.1 Detecting movemes

In order to approach the problem of detecting movemes, we made a simplifying assumption about the animals. We think of them as finite state machines that execute a certain action depending on their internal state. For example, a fly that moves into a “hungry state” would most likely perform actions

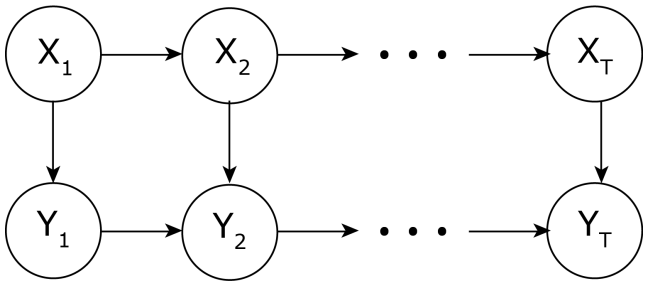


Figure 7.1: Graphical representation of a Hidden Markov Model

of finding food and eating. In a similar way, a fly in an "aroused state" would most likely try to fly and escape. Such model lends itself to a formalization into a Dynamic Probabilistic Graphical Model (PGM), where each node in the model represent a state in the animal in time, and each state has a probability distribution over the possible outcomes of the observed variable. The advantage of this probabilistic approach is that it allows modeling the inherent variability in the locomotion, actions, etc., of individuals, while allowing to group and abstract similar movemes or actions into a state that can have a semantic meaning. This approach has been taken by biologists before, for example, as in [4], where a PGM is used to try and study emotions in animals.

One PGM that lends itself particularly well to our task is a Hidden Markov Model (HMM). An HMM is a model for a stochastic process, in this case the time series. The process is described by hidden states, and at each hidden state a probability distribution determines the observed output, where the output is the time series. Hence, observing the time series allows inferring what the most likely sequence of hidden states is in this process. This fits the description of actions being the result of an internal state of the animal. The sequence of internal states or hidden states that the animal goes through, are reflected by the observed variables such as velocity, acceleration, etc., of the animal. A second important property of the HMM is that it takes into account the temporal relation of the time series. This is the case as the model assumes that the transition between states depends on the previous state of the system. Furthermore, the HMM assumes that the probability of being at any given state is independent of all previous past states given the observed variable at that state and the immediately previous state. This assumption allows us to include the time dependence, but it also simplifies the model considerably, so that inferring the parameters of the model is computationally efficient. Higher-order Markov Models can also be used but introduce a high computational cost. The graphical representation of this model is shown in Figure 7.1.

We introduce some notation that is used to specify the HMM. Let X_t be the hidden state at time t, with $X_t \in \{1, \ldots, Q\}$ where Q is the number of hidden states in the model. Let Y_t be the time series we observe, for time $t = 1, \ldots, T$. X_t can be a value or a vector of observed features,

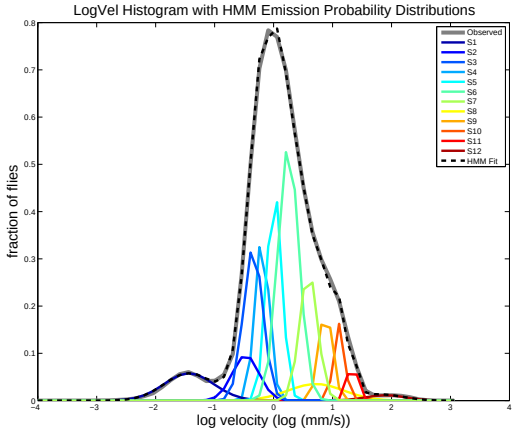


Figure 7.2: Histogram of log-velocities and emission densities for a subset of the "Fear in Flies" dataset

such as velocity, acceleration, wing position, etc. Let $A \in [0,1]^{Q\times Q}$ be the Transition Matrix, where $A_{ij} = P(X_{t+1} = j | X_t = i)$. Let B_q be the Emission Probability of state q, so that $Y_t \sim B_{X_t}$. Finally, let $\pi \in \mathbb{R}^Q$ be the initial state distribution, with $\pi_i = P(X_0 = i)$. Then, a HMM is completely specified by the transition probability matrix, the emission probabilities and the initial distribution, and we denote it by $\lambda(A, B, \pi)$.

Now, in order to learn a HMM, two parameters need to be specified: the number of hidden states Q, and the form of the emission probabilities B_q. As we will show in Section 7.3, we can determine Q through cross-validation. We learn λ for different values of Q using a subset of the data (i.e. training data). Then we use an unseen subset of the data to compute the likelihood of the trajectories given the observed model. As it will become apparent, as Q becomes larger the model no longer increases significantly the likelihood of the data but we start getting diminishing returns by adding complexity to the model. Using an Occam's razor principle, we chose the lowest value of Q that will yield a high likelihood of the data. For small datasets, in fact increasing the number of states will lead to overfitting and hence the likelihood will start decreasing.

The second choice is the parametric form of the emission probabilities. In our experiments, we used Gaussian densities, although other choices are suitable, such as t-distributions, exponential distribution, etc. This choice should be made depending on the data to be analyzed. For example, for data collected in [35], which we will refer to as the "Fear in Flies" dataset, a histogram of the log-velocities of the data yield a distribution that could be approximated by a Mixture of Gaussians, as exemplified by Figure 7.2. In this plot, the gray line indicates the distribution of the data (ignoring the time-stamp of the data), while each colored distribution represents a particular Gaussian

distribution with different parameters. Each of these are the emission probabilities of the estimated HMM for this particular subset of the data. The mixing component of the model, or weights for each component, is given by the steady-state probabilities of the transition matrix of the HMM. The dotted line represents the resulting distribution when these Gaussian components are added together. Hence, the distribution of this particular parameter in log-space is suited for Gaussian emission probabilities, while the distribution for other parameters may require a different type of emission probabilities. Once these choices are made, we proceed to fit the HMM to the data. For all our experiments that require fitting HMMs to data, we used the code in [50].

Once we fit the HMM, $\lambda(A, B, \pi)$, the movemes are the small units in the time series that get assigned to the same hidden state. These movemes are very short, usually in the order of 1 to 10 frames. The hidden states give semantic meaning to each of this movemes. For example, when fly trajectories are analyzed, the hidden states represent the following movemes: stationary, slow walking, slow walking on food, walking, walking on the boundary (usually named thigmotaxis), accelerating, among others. Depending on the number of states chosen for each subset of the data, there are more or less hidden states than the ones described above, so that there is a finer or coarser granularity to the described movemes.

Having extracted coherent meaningful snippets of the time series into what we call movemes, we move on to discover actions.

7.2.2 Detecting actions

In the context of behavior analysis, actions are defined as a combination of movemes that occur in the same order [5], and as in any other context, they can be described by a verb. In order to detect actions, we follow the same approach as the one used for detecting movemes. When movemes are found, the initial observed variables are abstracted into sequences of hidden states. This abstraction groups similar snippets of the time series into the same hidden state, and also transforms the initial continuous time series into a time series of discrete states. Yet, this abstraction does not have to stop at this level. We can further fit a HMM to the new time series in this discrete hidden-state space. The idea once again is to group similar common sequences into single states that will now represent actions.

The details of this process are exactly the same as the one described in the previous subsection, except that now there is no need to choose a parametric form for the emission probabilities. In this case, the emission probabilities are probability mass functions in $[0, 1]^Q$. That is, at state $Q_i^{(2)}$, the probability mass function will determine the probability of observing any one of the 1 to Q states of the first level HMM. Yet there is still a free parameter that needs to be set, which is the number of

hidden states that this new HMM will have. We will call this $Q^{(2)}$. The method for finding $Q^{(2)}$ is the same as the one used for finding Q, via cross-validation. In this case, not only will the likelihood of the data saturate, but it will also be evident that if the emission probabilities for these new states assign all the mass to a single state in the first level HMM, then there is no abstraction being done by the second layer. Rather, we are actually interested in emission probabilities that will have non-zero probabilities of emission in more than one of the states of the first HMM.

To illustrate this, if we fit a HMM to the output of a HMM that is fitted to log-velocity data of flies, a common state that emerges is one in which the emission probability mass function has a large mass for the hidden states that represent the movemes of accelerating/decelerating and hopping. Thus, the concatenation of an acceleration, a hop, and a deceleration are grouped together into a single action. Hence, these new hidden states will be considered as actions.

As it may be clear to the reader, this process can be repeated as many times as desired in order to abstract actions at larger resolutions in time. The experiments we present considered 2-level HMMs, but the method can be applied using as many levels as desired. Now, in order to detect an even higher level degree of abstraction, we move from analyzing single trajectories to finding similarities between trajectories. We explain this concept in the following subsection.

7.2.3 Finding stories

The last level of abstraction that we consider is what we call stories or activities. To do this, rather than simply adding more levels to our HMM model, we compare the full time series of individuals, in order to find clusters of similar individuals. Once a cluster is found, we can represent it by its medoid (the point closest to every other point in the cluster), and the concatenation of actions that constitute this trajectory will constitute a story. As will be shown in Section 7.3, these stories have clear interpretations.

The challenge of clustering the time series lies in quantifying how similar or different two trajectories are. To do this, we begin with the output of the HMM rather than with the original trajectories. The output of the first- or second-level HMMs is easier to deal with as the inherent variability and noise in the observed time series is naturally smoothed out by the HMM. We can now think of the time series in the hidden-state space as vectors in $\{1, \ldots, Q\}^T$. In principle, we could find the distance between two points by using the Euclidean norm of the difference between pairwise vectors. Nevertheless, if we are aiming to abstract a common story between a group of individuals, it is very unlikely that even if two individuals fall into the same story, that they will be at the same state at the same exact time. Not only this might not be the case, but even if two of these vectors follow the same exact sequence of states, it is also possible that the each individual spends different amounts of

time in each state. These two differences may not make the story different, and hence for this reason we want to use a distance that is more flexible than the Euclidean distance of the two vectors.

One common distance that has been used in the analysis of time series is Dynamic Time Warping (DTW) [12], introduced more than twenty years ago. Although technically this is not a distance as it does not satisfy the triangle inequality, it has been used due to the fact that it yields small distances between time series that are slightly distorted in the time axis. The technique has been successfully applied in the speech recognition community [12], in gene alignment as in [1] and [8], in medical applications such as cardiology [22], among others.

In our domain, the advantages of using DTW vs Euclidean distance are similar as those in other time-series applications. We would like to group time series that have small variations in the duration of each of the states, but that in general have a similar sequence of states. Another advantage of DTW is that the algorithm is flexible enough to allow defining a particular cost between each pair of states. In our case, since the numerical value of the hidden states may not be meaningful, we are able to design better cost matrices. The simplest of these approaches sorts the state numbers by the mean of the emission probabilities, so that the order of the states reflects the distance between two states. More sophisticated measures such as the KL distance between the emission probabilities of each pairwise state can be more descriptive.

In short, the DTW distance between two time series is found by comparing at each time step the two time series, and deciding if it is less costly to insert or delete (i.e., warp) frames in one of the time series, in order to make the corresponding cost at each step as low as possible. Dynamic programming is used in order to find this optimal trajectory efficiently, and the cost of finding this distance is $\mathcal{O}(T_1 T_2)$, where T_1 and T_2 are the lengths of each of the time series. Fast approximations to DTW have been proposed in [59] and [3] which reduce the cost to be nearly linear in the length of the trajectories rather than quadratic.

Once distances are computed, we can use any clustering algorithm such as k-means in order to find meaningful groups of trajectories. Since there is no sense for a centroid in this space, that is, an average trajectory, we use the k-medoids algorithm which finds points in the dataset to be the centers of mass of the clusters, and assigns the remaining points to the nearest cluster. As with the k-means algorithm, the solution is only a local minimum for the NP-complete problem of finding the best assignment, so we restart the algorithm many times to obtain a good enough clustering. As with every clustering algorithm, finding the right number of clusters is non-trivial, but we rely on the same methods used widely in practice. For example, plotting the average size (diameter) of the obtained clusters, versus the number of clusters allows choosing a value for k when the average size of the cluster saturates. The resulting clusters will constitute the various stories that can be found in the

data, and the medoids of the cluster will be the representatives of each of these stories.

Having described the various stages of the method, we now present results in two different datasets: the Fear in Flies dataset [35] and the Fly Bowl dataset [20].

7.3 Results on real datasets

The machinery described in the previous section can be used in different ways according to the specific dataset used. In this section we apply CUBA to two datasets that study flies. Each dataset served a different purpose in the biology community and in the same way we use the machinery of CUBA in different ways in order to come to useful conclusions concerning behavior.

7.3.1 CUBA in the Fear in Flies dataset

As described before, the Fear in Flies dataset was the result of a new setup, described in [35], that allowed recording videos of flies in a closed arena, and where a servo-motor controlled paddle produces shadows to stimulate the flies. The idea of the setup was to demonstrate fear-like behavior in fruit files. The dataset consists of recordings of movies that test various conditions of the experiment. It includes arenas with and without food, experiments with a different number of shadows presented, experiments where the interval between shadows or shadow frequency changes, experiments with different genetic lines tested at different temperatures, experiments with different levels of starvation of the flies, among others. Figure 7.3 shows a typical frame of the movies contained in the dataset. This arena shows 10 flies, the food patch in the center (darker region), and the paddle passing on top of the closed arena.

In this section we present results of applying CUBA to subsets of this data. The results are by no means exhaustive but are presented to illustrate how CUBA can be used as a tool for behavior analysis for biologists.

In order to apply CUBA to the dataset, we first need to define what are the observed variables. For this experiment, a Computer Vision tracker algorithm was developed in [35], which outputs the position of each fly at each frame of the movie. Due to the resolution of the movies, as well as to the number of flies in each preparation (usually 10), it is not possible to extract detailed pose features as the ones we describe in Chapter 6. Rather, all the information available is the position of the fly at each time step. A simple feature that can be derived from the position is the velocity of the fly. When we refer to velocity we are referring to the speed in the filmed plane. This feature has the advantage of including temporal correlation between frames, as well as being invariant with respect to the position of the fly. Figure 7.4 shows a fly trajectory as output by the tracker, and the corresponding log-velocity time series. A log-scale is used for the velocity due to the significant difference in speeds a fly can

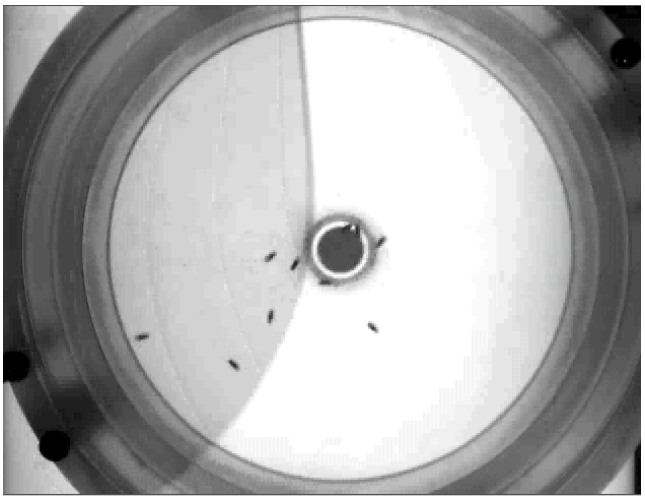


Figure 7.3: Fear in Flies dataset setup. The image corresponds to a frame of the videos captured and depicts the closed arena with a food patch at the center, 10 fruit flies, and the servo-motor controlled paddle that arouses the flies

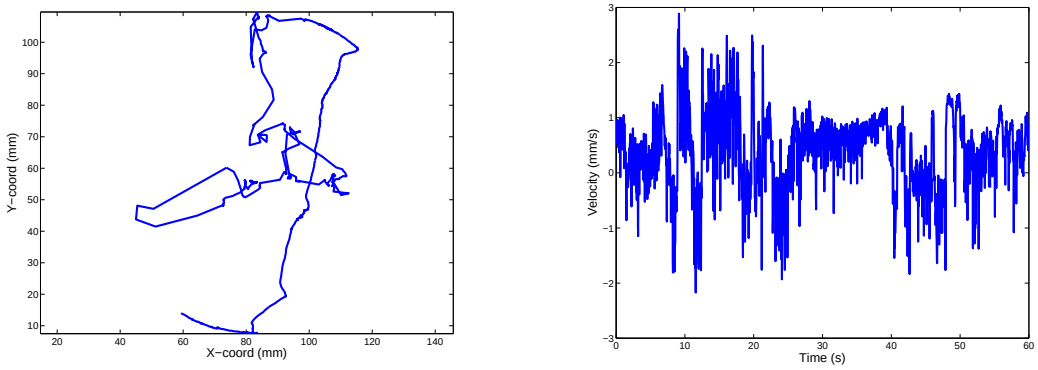
have, for example, between slow walking and flying.

As it is evident from the velocity plot for a single fly, these time series are very noise and therefore hard to understand. For this reason, extracting meaningful patterns or movemes requires the use of the techniques described in the previous section. The first step of CUBA is to learn the movemes using a HMM. Figure 7.5 exemplifies a typical model learned. This model was learned using 150 time series of flies in experiments with food, both for single and 10-fly preparations, as well as for a varying number of shadows.

The number of states was chosen through cross-validation of the likelihood on unseen data. Figure 7.6 shows the likelihood for different models used. As it is clear, the likelihood saturates around 10 to 12 hidden states. 12 states were chosen above 10 in this example, as the 12-state model was able to capture the faster state (small tail in the log-velocity distribution plot) more precisely than the model with 10 states.

With the model learned, the time series of velocities can now be transformed into time series of hidden states, where these states constitute the most probable sequence of hidden states given the model learned. Figure 7.7 shows a few sample trajectories in the hidden-state space. Each state is represented by the same color that was used for the emission densities in Figure 7.5. The plot also color codes the position of the fly by showing white, gray, and black colors on top of the time series to indicate open space, food, and boundary, respectively.

Once the number of states is chosen, it is also possible to give semantic meaning to each of the states. These correspond to movemes. By watching the movies and following the corresponding state



(a) Tracked trajectory of a single fly

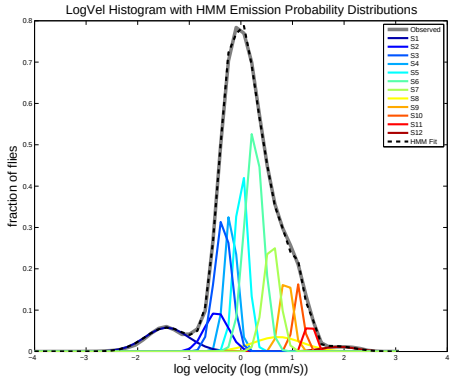
(b) Log-Velocity time series of the fly

Figure 7.4: Trajectory and log-velocity of a single fly as output by the tracker

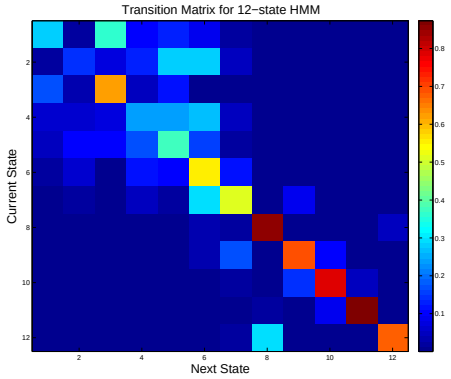
of the flies, it is clear that States 1 to 4 (dark blue states) are given to the fly when there is no visible motion. State 5 is assigned to the fly when it is moving very slowly. State 6 (green) is most frequent when flies are walking very slowly, usually in the food patch. The orange states, 10 and 11, are assigned to the flies when they perform a short fast walk. The dark red state, state 12, is assigned to the flies when they hop/fly. State 8 (yellow) corresponds to an acceleration state, as it is always assigned to flies just before and after a hop/fly event. This explains why the emission probability density function of this state has a very large standard deviation with respect to the other states.

The first major achievement of this analysis, is that it provides automatic classifiers for simple moveme/actions such as hopping or freezing. The work in [35] uses thresholds set by hand in order to detect events of hopping and freezing, but using CUBA, there is no need to fix a subjective threshold value to detect such behaviors. Rather, state 12 represents the hopping bouts, while states 1 and 2 represent the freezing states.

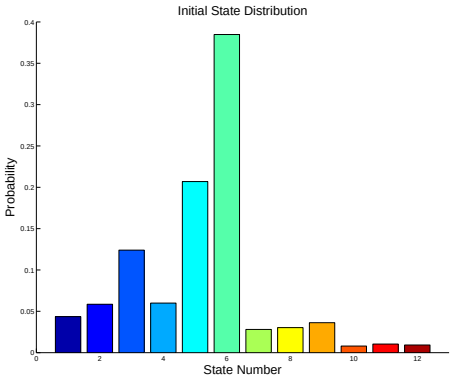
Once the movemes are found, we move on to compounding movemes by running a new HMM on the output of the first-level HMM. A typical outcome is shown in Figure 7.8. The figure shows the emission probability mass function for each of the second-level hidden states. In this example we use $Q^{(2)} = 7$. From the plot, a few actions can be inferred. State 7 has an emission probability that gives most of the weight to the hopping states and the acceleration state. Hence, this action groups acceleration - hop/fly - deceleration into a single action. State 6 has an emission probability that only gives weight to the two fast walking states. Hence, the action of fast walking groups the two previously discovered movemes into a single action. Similarly State 5 groups together the movemes of fast and slow walking. Another interesting action is that of State 3, in which the stationary states and slow walking states of the first level HMM are grouped together. This is a common action that is



(a) Log-velocity histogram of time series and emission densities B



(b) Transition matrix of HMM A



(c) Initial distribution π

Figure 7.5: Typical HMM model $\lambda(A, B, \pi)$ learned for the Fear in Flies dataset

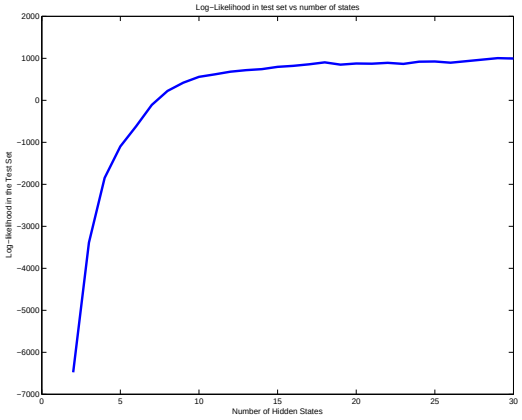


Figure 7.6: Log-likelihood of time series as the number of hidden states in the HMM vary

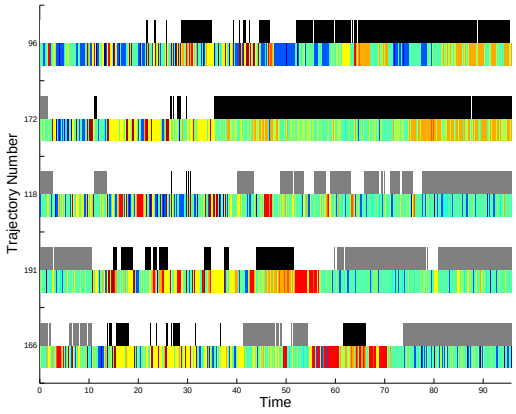


Figure 7.7: Sample time series of flies in the hidden state space with position indication. The color of the hidden states correspond to the colors used in Figure 7.5. The position of the fly is encoded with three colors: white for open space, grey for food patch, and black for the boundary.

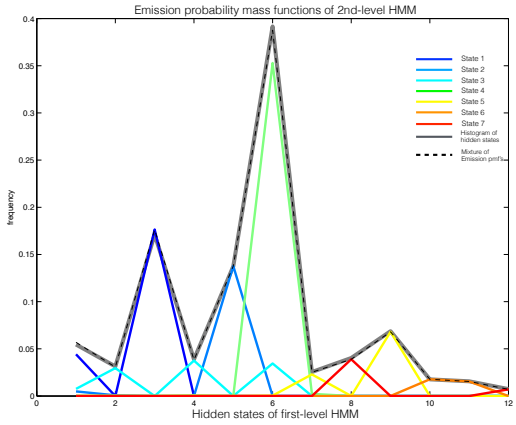


Figure 7.8: Emission probability mass functions of 2nd-level HMM

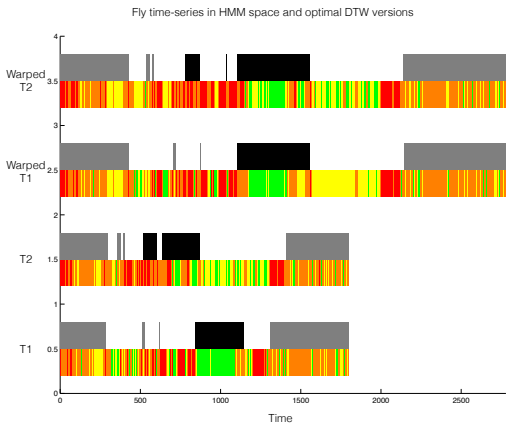


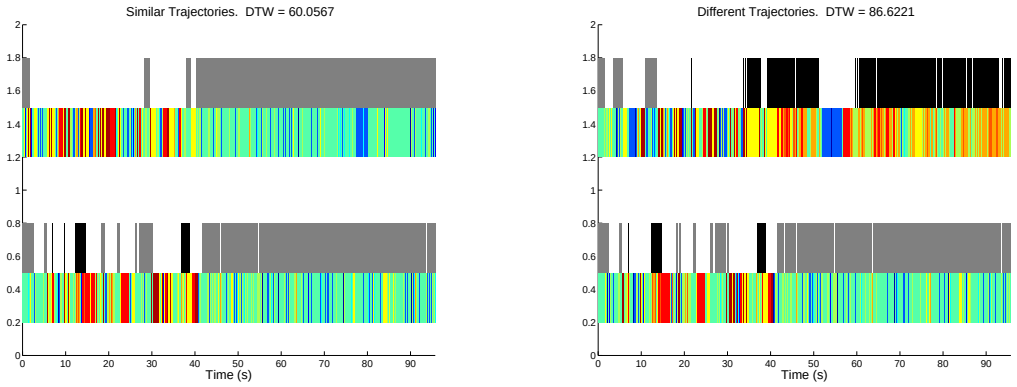
Figure 7.9: Time-series in hidden state space and warped versions found to compute the DTW distance between the pair of time series. Position is overlaid in the figure, but this information is not used in the DTW distance calculation.

mainly visible when the flies are feeding, as they walk slowly and stop continuously while they feed.

We now move on to clustering time series of individuals in order to find stories in this setup. We find the DTW pairwise distances among the individuals we want to analyze. Figure 7.9 shows two time series in a simpler 4-state HMM that are compared for illustration purposes, with the position overlaid, but not taken into account in the DTW distance calculation. The top plots show the warped versions of the time series that yield the minimum cost in a frame to frame comparison of the warped time-series. Once a matrix of distances is computed between each pair of trajectories, we can cluster them using k-medoids. Figure 7.10 shows a pair of similar time series and a pair of dissimilar time-series, when the 12-state HMM is used.

In order to visualize the clusters produced by k-medoids, we use Multidimensional Scaling (MD-scaling) [45]. The technique finds the location of points in a d-dimensional Euclidean space, given a matrix of distances or dissimilarities between points. In our case, the matrix of DTW distances is the input to the MD-scaling algorithm, and we use $d = 2$ in order to visualize the clusters. Figure 7.11 show clusters found in a subset of the data that included 240 time-series of flies on food, coming from setups with 10 flies, and where 10 forward and backward passes of the paddles were presented to the flies. We first examine the clusters by looking at the medoid time-series. Figure 7.11b shows the time series in HMM space of the three medoids of the dataset, with positions overlaid, as well as with vertical lines indicating the time at which the forward and backward passes of the paddle occurred.

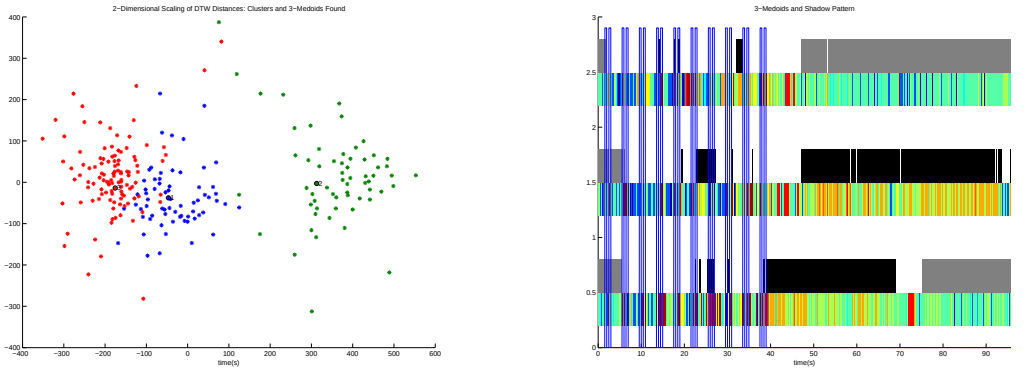
The three medoids can be interpreted as three distinct stories. The first medoid (from bottom to top) indicates a fly walking slowly on food (green state), which gets agitated when shadows pass



(a) Close time-series in DTW sense

(b) Distant time-series in DTW sense

Figure 7.10: Sample close and distant time-series in HMM space according to the DTW distance



(a) 3-medoid clustering in 2-dscaling

(b) Time-series in HMM space of medoids found

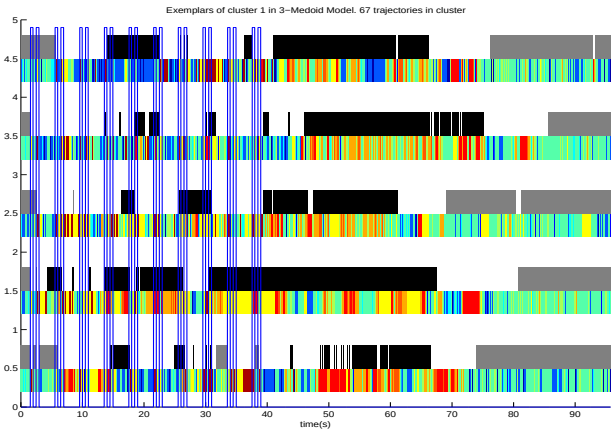
Figure 7.11: Clusters found by k-medoids using DTW distances on a 240-fly subset of the data, visualized using 2-dimensional scaling. The corresponding time-series of the medoids are also shown.

(yellow and red states) and ends up walking quickly (orange states) on the boundary before returning to the food. The second medoid shows a fly that after the shadows pass gets agitated and stays on the boundary walking quickly, as indicated by the predominant orange color of the trajectory. The third story shows instead the fly returning shortly after the shadows pass, walking slowly on food as indicated by the predominant green state. To confirm this observation, we plot five random time-series from each of the clusters. As it is clear visually in Figure 7.12, the medoids are indeed faithful representatives of the time-series that are grouped into the same cluster.

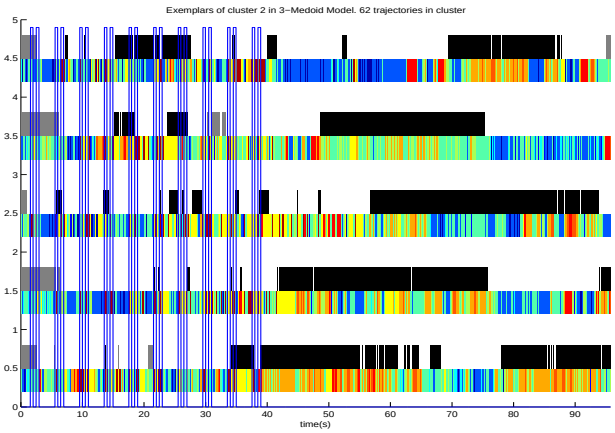
A more detailed analysis of each of the stories can be done by plotting histograms along time in hidden-state space, of each of the stories. Furthermore, as the position of the location of the flies is known, in terms of the three major regions, namely the food patch, open space, and boundary, the histograms can be further separated by location. Figure 7.13 shows these histograms. The histogram for time series in the same story is divided into three, displaying the percentage of flies on food at the top, those in the open space in the middle, and those in the boundary in the bottom. A guide for the mean log-velocity of each state and its standard deviation is plotted for reference at the rightmost part of the plots.

These histograms provide a clear summary of the data, in terms of the velocity of the flies in the experiment. It is clear that before the shadow stimuli are presented, the flies are all walking slowly or stationary on food. The flies then disperse once the shadows are presented, and most flies hop, as it is evidenced by the increase in frequency of the acceleration (yellow) state, and the hopping state (dark red). Another interesting observation is that during the shadow presentations, not only the acceleration and hopping states (yellow and dark red) increase, but it is also the case that the dark blue state increases and decreases periodically with the shadows. This shows evidence of a freezing behavior. As evidenced in other animals such as rodents, when an arousal event such as the presence of a predator occurs, animals freeze before attempting escape. This may be a mechanism to avoid detection, for example. *Yet, before this experiment had been carried out, clear evidence of freezing in flies had not been presented.* CUBA detects evidence of this behavior.

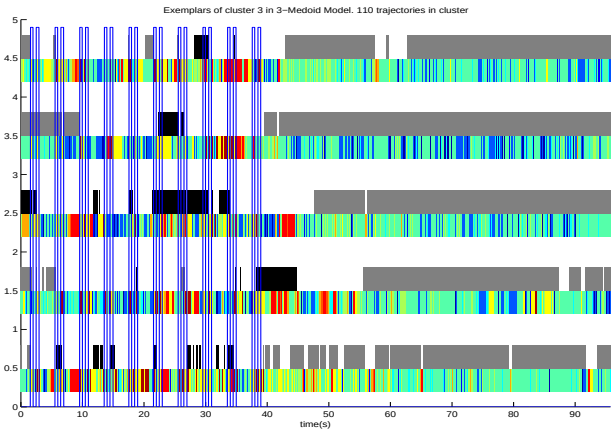
Across all clusters, it is clear that the first two shadows are enough to disperse all flies from the food patch. After the third shadow, it is now common to find flies in the boundary. Three possibilities divide the time series into separate stories. Either flies are very quick to return to food, after settling down in open space, as shown in story 3, and therefore flies will predominantly remain in the slow walking state (green) once they settle down. It may also be the case that flies remain very agitated as shown by the orange states in story 2. Yet, notice that the orange states are mostly predominant when the flies are in the boundary. Therefore this raises an important point, which is that flies walk faster when they are on the boundary. Similarly, when flies are on food, the fast walking states (orange) are



(a) Time-series from cluster 1

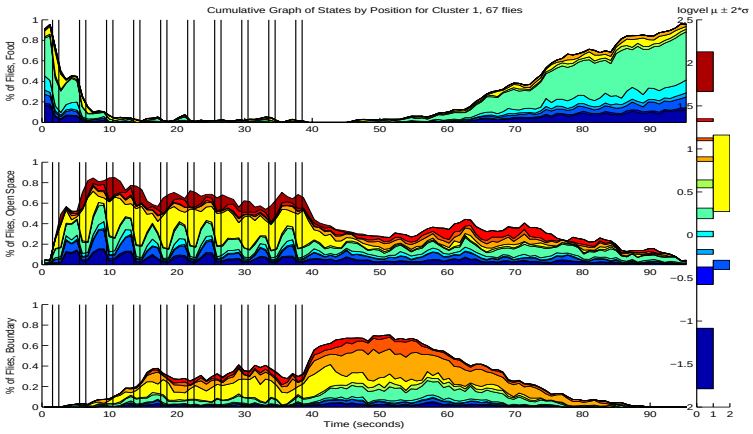


(b) Time-series from cluster 2

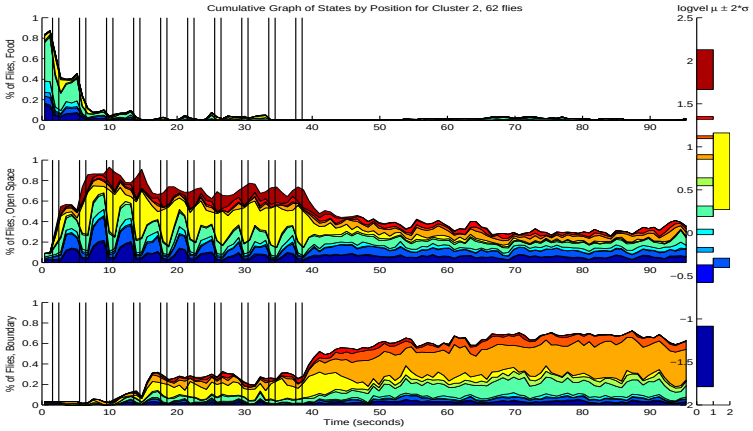


(c) Time-series from cluster 3

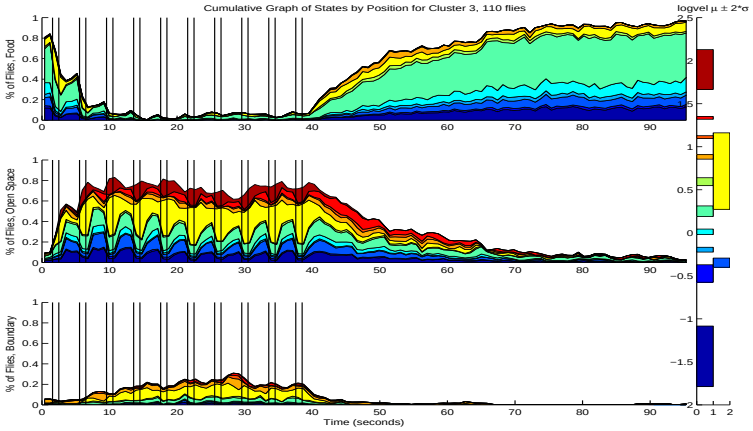
Figure 7.12: Time series in HMM space with overlaid position from the three clusters found



(a) Cluster 1: 67 flies



(b) Cluster 2: 62 flies



(c) Cluster 3: 110 flies

Figure 7.13: Histograms of states occupied by flies by cluster (story) and position (food patch, open space, and boundary)

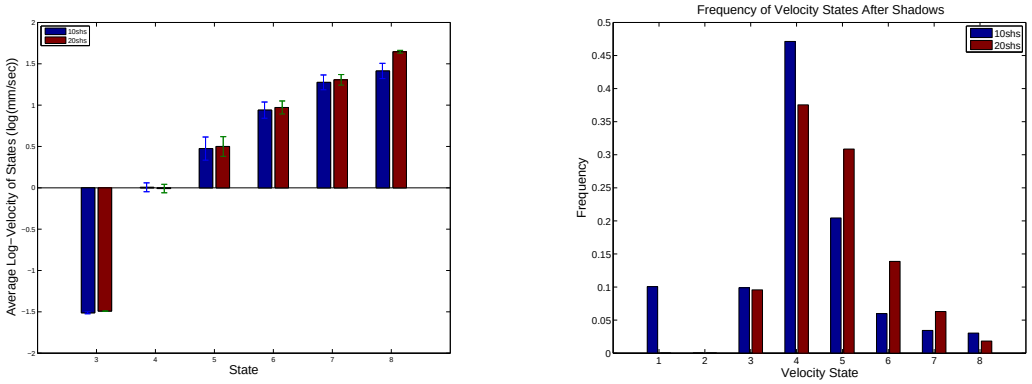
almost not present. This observation could be subject to causal testing, that is, biologists could test if flies walk faster because they are in the boundary, or if it is the case that agitated flies spend more time in the boundary. Finally, there is an intermediate story, in which flies disperse to open space, go to the boundary, but then do return to food during the length of the video.

With this visualization, it is clear how applying CUBA to the dataset allows summarizing and identifying the different behaviors present in the experiment. The detection of movemes as well as automatic classifiers for behaviors like freezing and hopping, present a major advantage as the classifiers are a byproduct of the method that require no manual annotations. Also, clustering trajectories into similar stories allows identifying the different types of behaviors elicited by the stimuli presented, as well as having a concise but detailed understanding of what flies are doing. This methodology, compared to usual methods that compare simple statistical quantities, such as mean or median velocities as done in [35], allows a deeper understanding of the biological phenomena observed. A simple value like the mean can be deceiving in the case of outliers, while the description given by CUBA gives a clear picture of the phenomena observed.

Another application of CUBA to a biological experiment like the Fear in Flies dataset, is that it allows using the mathematical models obtained to test hypotheses. For example, a simple experiment that was carried out in the original investigation was to understand how the number of shadows presented to the flies affected the flies. If more shadows elicited higher levels of agitation this could imply an integration model for "fear" in flies. To illustrate how CUBA can be used to show this, we apply the method comparing two scenarios: groups of 10-flies where 10 shadows are presented, and groups of 10-flies subject to 20 shadows. We now learn two HMM models: one for the 10-shadow data, and another for the 20-shadow experiment. We keep the number of hidden states Q equal for both models.

We ran this experiment in two sets of data: 240 flies subject to 10 shadows, and 240 flies subject to 20 shadows. Figure 7.14 shows the mean of the emission distributions found for both models, with blue bars representing the model for the 10-shadow data and red for the 20-shadow data. The standard deviation of the emission density function means are found by fitting the model multiple times with different splits of the data into training and validation sets. For this subset of the data, the chosen model using cross-validation was $Q = 8$. As it is clear from Figure 7.14a, state number 8, which constitutes the hopping/flying state, has a significantly higher value for the mean in log-velocity space. For the remaining states, there is no significance difference between the mean of the states. This shows clear evidence that the arousal caused by 20 shadows leads to higher observed velocities than in the case of 10-shadow presentations.

Figure 7.14b also shows evidence for the higher arousal in the 20-shadow experiment as it shows



(a) Means of emission densities for the HMM models for the10 vs 20 shadow experiments

(b) Time-series in HMM space of medoids found

Figure 7.14: Comparisson of HMMs learned for the 10 vs 20 shadow experiments

the distribution of flies in each of the hidden states after the shadows are presented. As it is clear from the plot, for the 20-shadow experiment, the flies spend more time in states 5 through 7 than the flies in the 10-shadow experiment, which spend most of their time in the slow walking state (state 4). This again shows evidence of a scalable response by the shadows, as the number of shadows increase.

This simple experiment exemplifies how hypothesis can be tested using the CUBA machinery. We now move on to an application to CUBA to the Fly Bowl dataset, which was created in order to understand interactions of flies in a closed bowl, when the wings have been removed.

7.3.2 CUBA in the Fly Bowl dataset

The Fly Bowl dataset is composed of videos of flies in a planar enclosed arena. There are usually 20 flies in each arena, and the wings of the flies have been removed. The software developed in [20] uses Computer Vision techniques to extract the position of the flies, as well as their orientation. Various techniques are used in order to avoid identity swapping when flies are close to each other. The work presented in that paper detects and analyzes specific behaviors in flies: walk, stop, sharp turn, backup, crab walk, touch, and chase. The high-throughput system developed in that study allowed analyzing various genetic lines. The vectors describing behavior frequency and duration allowed predicting genetic lines as well as gender.

In this context, we use CUBA to analyze the data without previously defining behaviors. For a subset of the data consisting of 415 fly trajectories filmed during 13 minutes, we run the CUBA machinery, once again on the log-velocity time series, and obtain a HMM with 15-states or movemes. We then run a second HMM and choose through cross-validation $Q^{(2)} = 6$ states. The inferred trajectories in HMM space are then clustered, obtaining this time 4 clusters. Figure 7.15 shows the

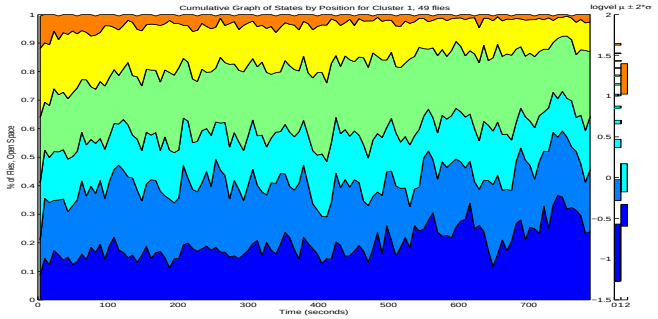
histograms of states in time for each of the clusters. In this case the fly bowl does not allow flies to walk in the border, and since there is no food, there is no spatial differentiation in the flies.

The actions represented by the 6 hidden states of the second-level HMM are the following: state 1 and state 2 (dark blue states) represent stationary states; state 3 (cyan) represents a slow walk–stop–slow walk repetition; state 4 (green state) represents a continuous slow walking state; state 5 (yellow) is a fast walking state; and state 6 (orange) involves accelerating–fast walking–decelerating repetition. The histograms of Figure 7.15 also show plots that indicate which states of the first level 15-state HMM are included in each of the second-level hidden states, with their corresponding means and standard deviation.

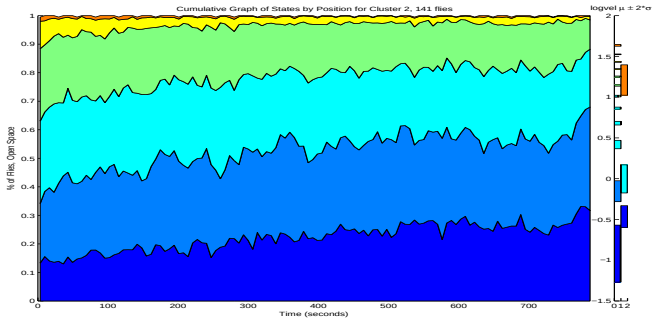
The four clusters indicate roughly the following stories: cluster 1 indicates flies that are transitioning between states and hence are changing velocities throughout the movie, although they tend to become more stationary towards the end; cluster 2 on the other hand indicates flies that are mostly stationary or slow walking but the frequency of states remains mostly constant throughout the movie, indicating mostly constant velocities of the flies; cluster 3 on the other hand indicates flies that are agitated throughout the movie, indicated by the high percentage of state 6 (orange), as well as constant changes in the frequencies of states indicating transition between states; finally cluster 4 is similar to cluster 2, except that the flies do not tend to stop by the end of the movie, but rather maintain a similar speed throughout the movie.

In this case, due to the lack of external stimuli, characterizing behavior in the form of stories is less meaningful as behavior is usually the result of a response to some external or internal stimuli. Nevertheless, the stories obtained become informative if we analyze the percentage of flies in each video that fall in each of the clusters. Figure 7.16 shows the number of flies in each of the clusters, for each of the videos analyzed. Notice that for some videos less than 20 flies are shown. This is the case as some preparations had less flies or because the tracker could not detect a reliable trajectory for some of the flies.

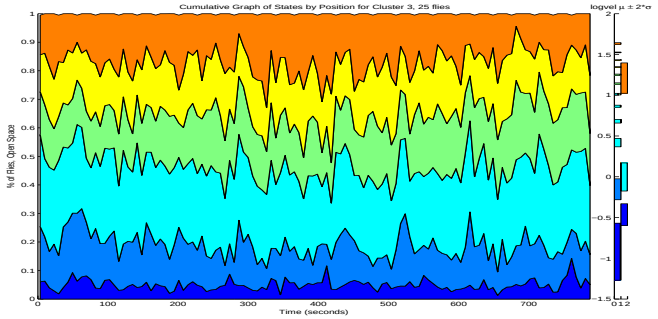
It may seem clear to the reader that the distribution among clusters is very different across the different videos in the dataset. In fact, when we discussed this difference with biologists, they immediately identified that the subset of videos used had different genetic lines of flies. Videos 1 and 2 correspond to one genetic line, videos 3 to 12 correspond to the control genetic line of wild type flies, videos 13 to 15 correspond to another genetic line, and the final six videos correspond to a different line. Clearly, the first two videos corresponding to one genetic line, account for most of the flies that fall into cluster 3, the story of highly agitated flies. The videos in the third group account for most of the flies falling in cluster 1. Finally the remaining videos have a similar distribution of flies in clusters 2 and 4, clusters which are very similar and indicate flies moving at a nearly constant and slow



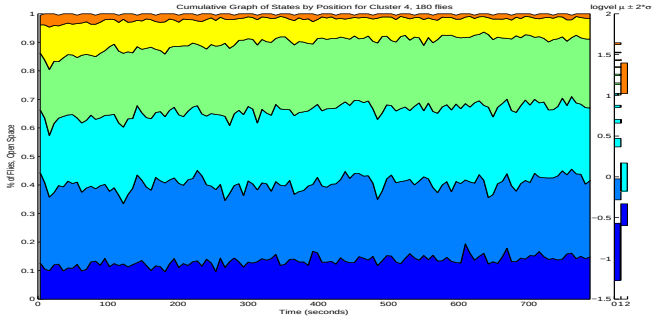
(a) Cluster 1: 49 flies



(b) Cluster 2: 141 flies



(c) Cluster 3: 25 flies



(d) Cluster 4: 180 flies

Figure 7.15: Histograms of states occupied by flies by cluster for a subset of the Fly Bowl dataset

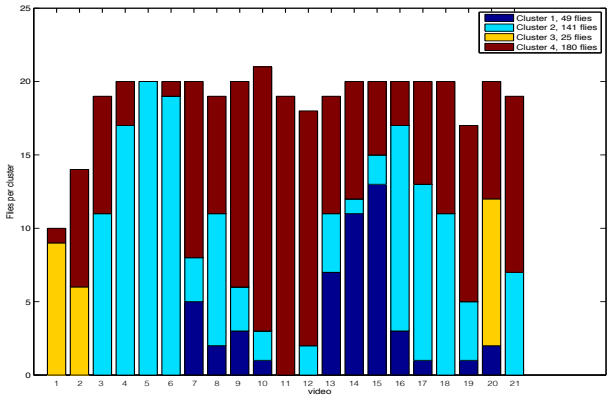


Figure 7.16: Flies per cluster per video in a subset of the Fly Bowl dataset

speed throughout the movie. That is, wild type flies under this setup tend to be "calm and steady" throughout the video, while the other genetic lines are prone to agitation and constant changes in speed. Hence, applying CUBA to this dataset in fact allows us to discriminate between genetic lines, just as the study that used predefined behaviors as descriptors in order to predict gender and genetic lines did.

In fact, CUBA can be modified to be a semi-supervised method, as once DTW distances are computed, we can use MD-scaling in this case not for visualization purposes but to embed the time-series in a high-dimensional Euclidean space. In this space, supervised machine learning methods can be used to identify various groups of flies, such as different genetic lines or genders, if this information is available. In our experiment, this was done using completely unsupervised methods as we were initially unaware the videos contained different genetic lines, and yet the method was able to discriminate these. Hence, applying CUBA to a different dataset illustrates a different use that can be given to the method developed.

The unsupervised method developed in this chapter is able to detect behavioral patterns at different time scales. We can detect movemes by means of a HMM and its hidden states, a technique that allows building classifiers for specific behaviors without the need for annotation. We can detect actions by further grouping movemes using a new HMM whose observed variables are the hidden states of the initial HMM. Furthermore, we can find distances between the observed time-series in this new hidden-state space, by using an appropriate measure as DTW, which allows time deformations between series when finding the comparison. This allows clustering in meaningful groups the time-series, yielding typical stories found in the data. The machinery then allows testing hypotheses, by comparing the models obtained, and allows understanding the data at a much more profound level, compared to the

case where simple statistics like the mean and median are used to summarize the data. Finally, the cluster membership distribution for groups of flies can be discriminative enough to detect different genetic lines, as exemplified in the experiments on the Fly Bowl data. These are just some of the advantages of using a system like CUBA. As biologists begin using this method to analyze their data, it will become clear how much more can be learned on how these computational methods can help advance the study of behavior.