HMM Based Behavior Recognition of Laboratory Animals

Selcuk Sandikci
Brace B.V., Helmond, Netherlands
selcuk.sandikci@brace-automotive.com

Pinar Duygulu
Bilkent University, Dept. of Computer Engineering, Ankara, Turkey
duygulu@cs.bilkent.edu.tr

A. Bulent Ozguler
Bilkent University, Dept. of Electrical and Electronics Engineering, Ankara, Turkey
ozguler@ee.bilkent.edu.tr

Abstract

In pharmacological experiments, a popular method to discover the effects of psychotherapeutic drugs is to monitor behaviors of laboratory mice subjected to drugs by cameras. Automating behavior analysis of laboratory mice saves both time and human labor. In this study, we focus on automated action recognition of laboratory mice from short video clips in which only one action is performed. A two-stage recognition method is designed to address the problem. In the first stage, still actions such as sleeping are separated from other action classes based on the amount of the motion area. Remaining action classes are discriminated by the second stage in which we project 3D action volume onto 2D images by encoding temporal variations of each pixel using discrete wavelet transform (DWT). Resulting images are modeled and classified by hidden Markov models in maximum likelihood sense. We test the proposed action recognition method on a publicly available mice action dataset and achieve promising recognition rates. In addition, we compare our method to well-known studies in the literature.

1. Introduction

In pharmacological experiments involving laboratory mice under the influence of psychotherapeutic drugs, behavior pattern of the mice reveals important clues about physiological effects of the drug. The subject must be monitored and its actions must be annotated in an objective and measurable manner in order to uncover the effects of injected drug. Considering that pharmacological experiments are repeated many times on hundreds of mice for statistical accuracy and consistency, a vision-based action recognition system is highly desirable, since it would save substantial amount of both time and human labor. Another desired specification is that the system should be non-intrusive which is also addressed by a vision-based system.

There are a number of challenges that need to be addressed in order to design a robust action recognition system for mice [7]. Unconstrained motion (i.e. actions in a burst and significant variations in actions), highly deformable blob-like body and small body parts of the mice are the biggest challenges which hinders part-based and template-fitting approaches.

Animal action recognition has attracted less attention compared to human action recognition. Nevertheless, there has been a few remarkable studies for mice action recognition. First of them is performed by Dollar *et. al* [5] who expressed actions as a collection of visual words extracted from spatio-temporal vicinity of interest points in 3D. Visual words are represented by low-level features such as normalized pixel values, brightness gradient vectors and optical flow. Jhuang *et. al* [2] proposed an action recognition method which imitates visual processing architecture of human brain by hierarchical spatio-temporal feature detectors.

Xue ve Henderson [10] constructed affinity graphs using spatio-temporal features to detect Basic Behavior Units (BBUs) in artificially created mice videos. BBUs are assumed to be building blocks for more complex behaviors. They applied Singular Value Decomposition (SVD) to discover BBUs in a given complex behavior. In addition to mice action recognition, there has been also some vision based research on multiple mice tracking based on optical flow, active contours [11], and contour and blob trackers [3].

In this paper, we present a two–stage method for behavior recognition of laboratory mice. First stage of our framework is used to discriminate still actions such as sleeping from the others. We take advantage of the amount of motion area, that is covered by the subject while performing the behavior. The second stage classifies the remaining four actions, namely, drinking, eating, exploring, and grooming.

Inspired by the work of Töreyin *et. al* [1], we utilize Discrete Wavelet Transform (DWT) to analyze temporal characteristics of individual pixels. Then, we form action *summary* images (ASIs) using the amount of temporal fluctuations at each pixel in the video volume. ASIs are transformed into subimage sequences by blockwise raster scanning. We form multidimensional observation sequences by taking intensity histograms of each subimage in the sequence. Hidden Markov models (HMMs) with continuous observation densities are used to model the observation sequences. Classification of action videos with unknown classes is carried out by trained HMMs in the maximum likelihood sense.

The paper is organized as follows: in Section 2, we present the details of our action recognition algorithm. In Section 3, we test our method on a publicly available mice action dataset [5]. We also compare recognition performance of our method with the algorithms in [5] and [2]. Finally in Section 4, we conclude the paper by giving a short summary and providing some future research ideas.

2 Action Recognition Algorithm

2.1 Recognition of Still Actions

We classify sleeping action using a simple method in which we exploit the area spanned by the subject while performing the behavior. The main assumption is that during sleeping the animal is almost still and the spanned area is minimal compared to other behaviors.

In order to determine the spanned area for a given video clip V, temporal standard deviation σ_t of each pixel in the video volume is computed empirically and thresholded with a predefined threshold ϵ. Pixels having standard deviation above the threshold are considered to be moving pixels. This simple method is sufficient to detect moving pixels, since video recording is illumination-controlled in pharmacological experiments. Then, we fit univariate Gaussian distributions to sleeping and non-sleeping behaviors in the training set using the number of moving pixels. We plot the Gaussian distributions learned from training set videos in Figure 1.

Given a test video V_{T} with the number of moving pixels associated with it, we estimate the probability of V_{T} being a sleeping or non-sleeping video using trained Gaussian distributions. Then, V_{T} is classified according to maximum likelihood criterion.

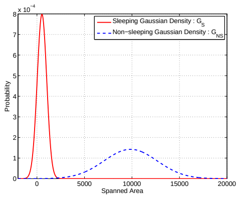


Figure 1. Gaussian distributions for sleeping and non-sleeping actions learned from real data. Spanned area is quantified in terms of number of pixels.

2.2 HMM based Action Recognition

2.2.1 Discrete Wavelet Transform and Feature Extraction

Mice actions can be characterized by a combined motion of different body parts. Although body parts are hard to detect and track, the action can still be characterized by spatial configuration of image regions with different motion energies as seen in Figure 2.

Therefore, we analyze the temporal characteristics of image points by discrete wavelet transform (DWT) [6] applied along the temporal axis. Only high frequency components are considered, since most of the information is carried in them. A simple measure of temporal variations in a pixel is the number of zero crossings in its highband subsignal. Intensity of a pixel in action *summary* image (ASI) is set to the zero crossing number of the corresponding pixel in the original video. Consequently, an ASI has the same resolution as the original video. Some example frames from various actions and their ASIs are illustrated in Figure 2. Small objects in ASIs are assumed to be generated by background clutter noise, thus they are removed. Then, a bounding box image BB is formed such that all of the pixels with nonzero intensity values in ASI are assured to be inside BB.

In order to describe ASIs, we divide the bounding box image BB into a grid of $N_{SI} \times M_{SI}$ overlapping subimages Ω_{SI}. We use an overlap ratio of 75% along both horizontal and vertical directions. Tracing the subimages in a raster scan fashion generates a sequence of overlapping subimages. Tracing scheme is illustrated in Figure 3. After obtaining the subimage sequence, for each subimage Ω_{SI}, an m_{SI} bin histogram is computed based on pixel intensities. Collection of the histograms in the same order with the subimage sequence gives us the observation sequence $\boldsymbol{O} = O_1 O_2 \ldots O_T$ to be used in HMMs. Here, observation symbol O_n is a m_{SI}-dimensional vector and corresponds to the his-

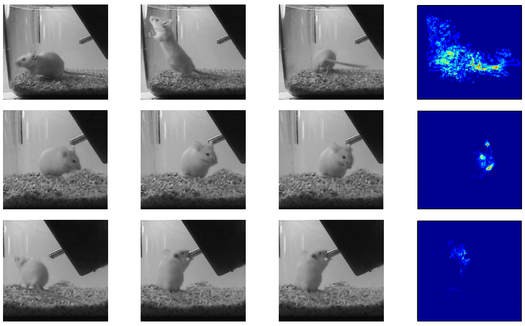


Figure 2. Sample frames from various actions (top: exploring, middle: grooming, bottom: eating) and their corresponding ASIs.

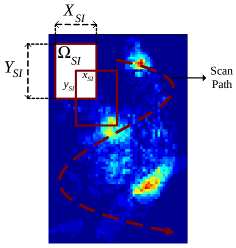


Figure 3. Scanning scheme of BB image.

togram of the n^{th} subimage. Length of the observation sequence is T which is simply the product of N_{SI} and M_{SI} i. e., $T = N_{SI}M_{SI}$.

2.2.2 Modelling of Action *Summary* Images using Hidden Markov Models

Hidden Markov models (HMMs) have been widely used in speech recognition [4], face recognition [8], and action recognition [9]. HMMs are well-known for their applications on modeling time series.

Recall that by exploiting DWT and forming ASI for each action, we were able to reduce the action recognition problem to an image classification problem. In view of the successful applications of HMMs on face recognition, we prefer to follow the work of [8] to model ASIs by HMMs.

We train an HMM model $\Lambda = (A, B, \Pi)$ for each ASI (i.e. for each action video) using the associated observation sequence $\boldsymbol{O} = O_1O_2 \ldots O_T$ with it. Here, Λ includes N hidden states. A is the state transition probability matrix and Π is the initial distribution of states. B is the collection of observation probability distributions for each state, which represents the probability of generating an observation in each state. In our application, observation distributions are modelled by multivariate Mixture of Gaussions with mean vector $\boldsymbol{\mu}_j$, covariance matrix $\boldsymbol{\Sigma}_j$ and weight vector c_j for the j_{th} state.

The model parameters for each HMM are learned by maximizing the probability $p(O|\Lambda)$ using its training observation sequence O. Baum-Welch algorithm [4] is employed to iteratively re-estimate the model parameters such that the probability $p(O|\Lambda)$ achieves its local maximum. To classify a given test video V_{T}, its observation sequence $\boldsymbol{O}_{\mathrm{T}}$ is formed as described in the previous section. Action in V_{T} is assigned to the class of most likely HMM model

$$\operatorname*{argmax}_c p\left(\boldsymbol{O}_{\mathrm{T}} \mid \Lambda_c\right), 1 \leq c \leq \#\text{ of trained HMMs.}$$

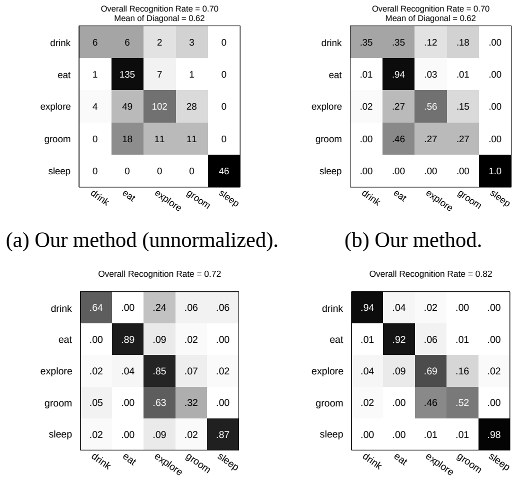
In our method, the training set is the rest of the dataset with the test video omitted. This procedure is repeated for all of the video clips in the dataset and overall recognition accuracy is measured to be the average of all classification runs.

3 Experimental Results

MATLAB implementation of our method is tested on a publicly available mice action dataset recorded by authors of [5]. The dataset consists of short video clips manually cut from seven fifteen–minute videos of the same mouse recorded at different times of a day. In this dataset, there are five action classes, namely `drinking, eating, exploring, grooming`, and `sleeping`. Although there are five classes in the dataset, we notice significant pattern variations among intra-class behaviors. Each video clip corresponds to one action and lasts about 10–15 seconds. Videos are annotated by authors of [5] using advice of veterinarians at the UCSD Animal Care Program.

We apply the first stage of our framework to the mice action dataset to eliminate `sleeping` action and achieve 100% classification accuracy. Then, our HMM-based method (see Section 2.2) is used to classify remaining action classes. Recognition performance is illustrated by confusion matrices in Figure 4 (a) and (b). Our overall recognition rate is 70%, i.e. every 7 out of 10 video clips are recognized correctly.

As seen from Figure 4 (a) and (b), only `eating` action is successfully recognized. 100% success rate for `sleeping` action is inherited from the first stage. Remaining actions are confused with each other. We believe that `eating` action is classified succesfully due to similar appearances of ASIs associated with `eating`, i. e. `eating` action turns out to be unimodal. On the other hand, intra-class variances of ASIs generated by



(a) Our method (unnormalized).

(b) Our method.

(c) Dollar *et. al* [5].

(d) Jhuang *et. al* [2].

Figure 4. Confusion matrices of our method and related studies.

other actions are quite high. We deduce that there are no consistent patterns in other actions to be modeled by HMMs. In other words ASIs generated by other action classes are too random even for HMMs. Besides, it is our belief that the length of the observation sequences is too short for training reliable HMMs. Recall that the length of observation sequences was given as $T = N_{SI}M_{SI}$. One may increase N_{SI} and M_{SI} by dividing bounding box image BB into smaller subimages. However, decreasing the size of subimages gradually disregards spatial relation between pixels of BB, which will eventually decrease the quality of observation vectors and introduce inaccuracies in estimation of HMM parameters.

We compare our method to [5] and [2] on UCSD mice action dataset in Figure 4. Our method outperforms both [5] and [2] for `eating` and `sleeping`. We achieve a similar recognition rate to [5] for `grooming` action. Our performance for `drinking` and `exploring` are lower than both studies. Our overall recognition rate (70%) is very close to that of [5] (72%) and lower than that of [2] (82%).

4 Conclusions

In this paper, we proposed a two–level system to recognize mice actions from short video clips. Designed system is a preliminary work for a general continuous action recognition system which greatly aids pharmacologists in their experiments on mice. The first stage of the system is used to classify still actions such as sleeping, where the second stage is a cascade combination of two subsystems based on DWT and HMMs. We tested our method on a publicly available dataset and achieved an overall recognition rate of 70%. We observed that quantifying the amount of motion is sufficient enough to identify still actions. Accumulating temporal variations of individual pixels all over the time axis discards local temporal information which could be useful in feature extraction. Instead, spatio-temporally windowed wavelet coefficients can be a richer feature representation. Deciding on model complexity and observation symbols is a key issue in HMMs. The observation symbols must be long enough to reliably train HMMs. In order to overcome this shortcoming, multiple observation sequences extracted from multiple videos can improve training of HMMs.

References

[1] B. U. Toreyin, Y. Dedeoglu, A. E. Cetin. Flame detection in video using hidden markov models. *IEEE International Conference on Image Processing*, 2:1230–1233, 2005.

[2] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *IEEE International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[3] K. Branson. *Tracking multiple mice through severe occlusions*. PhD thesis, University of California at San Diego La Jolla, CA, USA, 2007.

[4] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005.

[6] P. P. Vaidyanathan. *Multirate systems and filter banks*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1993.

[7] S. Belongie, K. Branson, P. Dollár, and V. Rabaud. Monitoring animal behavior in the smart vivarium. *Measuring Behavior, Wageningen, The Netherlands*, 2005.

[8] V. V. Kohir and U. B. Desai. Face recognition using a dct-hmm approach. In *Proceedings of the 4th IEEE Workshop on Applications of Computer Vision*, pages 226–231, Washington, DC, USA, 1998.

[9] X. Li. HMM based action recognition using oriented histograms of optical flow field. *Electronics Letters*, 43(10):560–561, 2007.

[10] X. Xue and T. C. Henderson. Video-based animal behavior analysis from multiple cameras. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 335–340, 2006.

[11] Z. Kalafatic. Model-based tracking of laboratory animals. *The IEEE Region 8 EUROCON 2003. Computer as a Tool*, 2:175–178, 2003.