Learning animal social behavior from trajectory features

Eyrun Eyjolfsdottir*, Xavier P. Burgos-Artizzu*, Steve Branson+,
Kristin Branson†, David J. Anderson*, Pietro Perona*
*California Institute of Technology, + University of California, San Diego,
†HHMI Janelia Farm

eeyjolfs,xpburgos@caltech.edu, sbranson@cs.ucsd.edu,
bransonk@janelia.hhmi.edu, wuwei,perona@caltech.edu

1. Introduction

Automatically classifying behavior of humans and animals from video is one of the most interesting and challenging fields of computer vision, [3, 1, 6]. Most of the successful human behavior recognition works use as features for classification information extracted from a direct representation of the scene (visual features), as opposed to indirect representations such as silhouettes, body parts, pose or object positions, which can be very sensitive to viewpoint variation and occlusions in real-world videos [10].

In contrast, indirect representation of the scene is widely used in the case of animals [1, 6, 4]. Animal enclosures allow for a more controlled filming, which reduces viewpoint variations and occlusions, facilitating the extraction of indirect representations of objects in the scene. Moreover, animal bodies are less expressive than humans, therefore causing direct visual features to work worse on animals [2].

One of the most widely used features for animal behavior recognition is the position of animals in time (result of either manual annotations or an object detection+tracking algorithm) [1, 6, 4, 2]. From the positions, usually several trajectory features are computed, such as distance between animals, their direction of movement, velocity, acceleration, etc. These trajectory features are used, together with the behavior labels, to train a supervised classifier that learns the discriminative features across behaviors.

In scenarios where videos are previously segmented [1], such as in KTH [11] or Hollywood2 [8] human datasets, a classic supervised classifier such as SVM [14] or AdaBoost [5] is often used.

In more realistic scenarios, however, the task is to fully segment a continuous video into behavior intervals (behavior category, starting frame, ending frame). Most works on animal behavior recognition fall into this category [1, 6, 4, 2], while recent effort has also been made in the human action recognition community to move in this direction, *e.g.* Virat dataset [9].

In this scenario, more intricate classifiers are needed. In [6] authors use a two layer SVMHMM, while in [2] authors extend Auto-context [13] to video. These classifiers are able to detect behavior classes and at the same time learn behavior transitions, successfully segmenting long, continuous videos into smooth behavior intervals. In this work, we describe a novel extension of Structured SVM and benchmark it against the Auto-context method to measure its robustness and versatility.

The rest of the paper is as follows: Section 2 briefly presents the two methods to be compared, Section 3 presents results on two different datasets (mice and flies) and discusses the results.

[1] Original videos are divided into clips and the classification task consists in determining what action each clip contains.

2. Learning algorithms

We benchmark our Structured SVM approach, which is based on a method proposed in [12] and has been extended in our lab for segmentation of video into actions of fruit flies. We compare it to the method published in [2] alongside CRIM13 dataset.

2.1. Structured SVM

Our approach to detecting actions within a sequence consists of two main components: an *inference algorithm* that takes as input a sequence x and returns y, a segmentation of x into actions, and a *learning algorithm* that learns a model from pairs of sequences and their ground truth segmentations, $\{(x_i, y_i)\}$.

The input x can be any object that evolves over time and contains distinct events of interest. In our experiments, x is a parameterized video of animals interacting, where $x(t)$ is a vector of features representing the tracks of the animals at frame t. The output y can be described as a sequence of labeled intervals ($bouts$), $y = \{y_j\} = \{(b_j, e_j, c_j)\}$, where y_j is the jth interval in the segmentation of x, b_j and e_j mark the beginning and end of the interval and c_j is the class label of the activity that the boundary encloses.

In order to more appropriately represent statistical patterns of temporal motion during an action, the algorithm relies on *bout features*, defined as $\vec{\psi}(x, b, e)$. These can be arbitrary functions over the set of per-frame features $x(t)$, $b \leq t \leq e$. These bout features are multiplied by a vector of learned model weights, computing a score measuring the likelihood that an action occurs at a particular time in the video sequence. These are combined with transition scores that encode the likelihood of two consecutive actions. Collectively, these define a *score function* measuring the likelihood of a segmentation of a video sequence into actions. Inference solves for the segmentation of a video that maximizes this score and is efficiently computable using dynamic programming. Training minimizes a convex upper bound on a customizable *loss function* that measures how much a predicted segmentation disagrees with the ground truth segmentation.

2.2. Auto-context

Recently, a novel method for social behavior recognition based on an extension of Auto-context to video was presented in [2]. Auto-context was first proposed in [13], and has proved to perform well in high-level computer vision problems that benefit from learning a context model.

As part of Auto-context, behaviors are first classified based only on local features (such as trajectory features), and then in subsequent iterations by adding to the feature set a list of temporal context features, computed from confidence levels output of classifiers at previous iteration. Authors propose to generate a large pool of *weak trajectory features* from trajectory information, in a similar way to what is done for object detection in the 2D case [15].

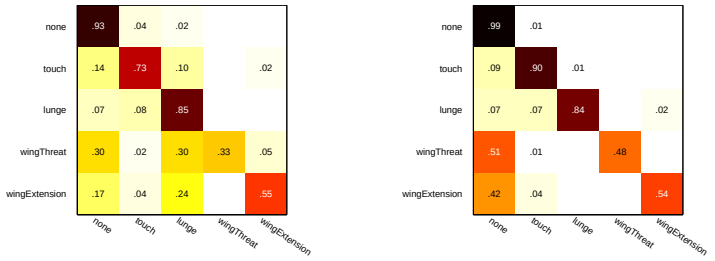
3. Results and Conclusion

For completeness, we benchmark both approaches discussed in Section 2 on two different datasets; Section 3.1 presents the results of both methods on a new dataset collected in our lab, containing videos of fruit flies interacting, and Section 3.2 shows the results on a subset of **CRIM13** mouse dataset [2].

3.1. Flies

In collaboration with biologists, we have collected a new dataset containing videos of fruit flies interacting. In each video, two flies were placed in a 50mm x 40mm chamber, with a patch of food in the middle, meant to encourage aggressive behaviors. The videos are 20 minute long, recorded at 200Hz from a single top down view, such that the length of a 2mm fly is equivalent to 25 pixels.

Each video was hand annotated by a behavioral expert into 4 main behaviors; **touch**, **lunge**, **wing threat**, and **wing extension**, as well as the grab-bag category **other**. The most frequent behavior in these videos



(a) Auto-context 68.0% (b) Structured SVM 75.0%

Figure 1. Confusion matrices and average agreement of the diagonal on our fly dataset.

is **touch** taking up 5% of the frames, and the most infrequent one is **wing extension**, taking up 0.3%.

In order to parameterize the videos we applied fly- detection and segmentation to each frame (using morphological operations and template matching) to obtain five primary features: *body position, body orientation, wing positions,* as well as pixel masks describing *leg pixels*, and on top of the detections we used the Hungarian Algorithm in order to track the individual flies. From the primary features twelve features, invariant of the flies' absolute positions, were computed: *velocity, angular velocity, min wing angle, max wing angle, mean wing length, ratio between body's major and minor axis, distance between flies, angle between flies, facing angle of one fly to the other, minimum distance of a leg to both flies, ratio between foreground and body pixels*, and *contrast*. The resulting features $x(t)$ which we use as input to our learning system are those features and their first two derivatives, a total of 36 features per frame. Expanded into bout-level features, the dimensionality of the learned model was 4650.

Figure 1 shows the results of both methods on our fly dataset. Both methods work reasonably well, and in this case our Structured SVM approach outperforms Auto-context. Surprisingly, both methods make similar mistakes across behaviors, especially when confusing **wing threat** and **wing extension** with **none**. Auto-context struggles more with **touch** behavior.

3.2. Mice

The Caltech Resident-Intruder Mouse dataset (**CRIM13**) consists of 237x2 10 minute videos (recorded with synchronized top and side view) of pairs of mice engaging in social behavior, catalogued into thirteen different actions. A team of behavior experts annotated each video on a frame-by-frame basis for a neurophysiological study of mice [7]. The dataset is publicly available from `www.vision.caltech.edu/Video_Datasets/CRIM13/CRIM13/Main.html` and is thoroughly described in [2].

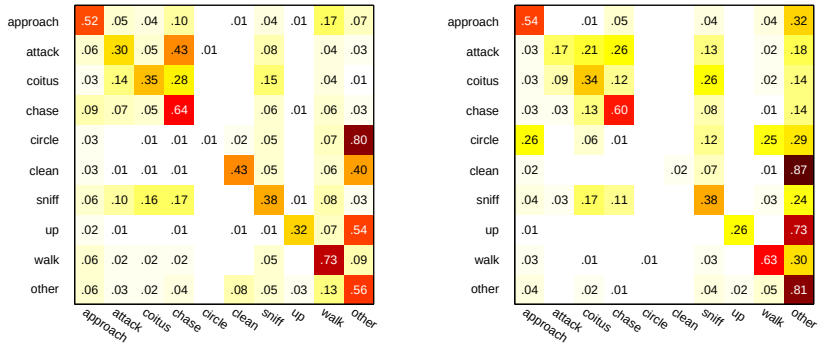
In order to benchmark both classifiers on this dataset, we used as input the original 19 trajectory features proposed by the authors in [2][2]. For time limitation reasons (CRIM13 has 8M frames), we only used a subset of the videos: the *validation* set. This set contains 10 full videos for training and 10 full videos for testing, around 100k frames each, and uses only 10 of the 13 original behaviors.

Figure 2 shows the results of both methods on CRIM13. Both methods perform under 50%, with Auto-context outperforming our Structured SVM approach. The main issue is that the *validation* set is too small; some behaviors have only a few instances in the training set, making it hard to learn a robust classifier. In fact, in the original publication, Auto-context ends up performing at 62% when trained on more data. The main difference between the performance of the two approaches is that Structured SVM seems to struggle more with the **clean** behavior, the one with highest intra-class variation.

3.3. Conclusions

We have benchmarked two different classifiers on two challenging animal behavior datasets, leaving everything else fixed to evaluate their robustness and versatility. Results are encouraging, using each of the

[2]In the case of Auto-context, a set of 2850 weak features are generated from these 19 features.



(a) Auto-context 42.3% (b) Structured SVM 37.2%

Figure 2. Confusion matrices and average agreement of the diagonal on *validation* set of CRIM13 dataset.

methods out of the box on the new dataset performed close the performance as optimized for the original dataset. By optimizing the parameters of the Structured SVM for the mouse dataset we should be able to achieve better results.

References

[1] K. Branson, A. A. Robie, J. Bender, P. Perona, and M. H. Dickinson. High-throughput ethomics in large groups of drosophila. *Nature Methods*, 6(6):451–457, 2009. 1

[2] X. Burgos-Artizzu, P. Dollar, D.Lin, D. Anderson, and P.Perona. Social behavior recognition in continuous videos. In *CVPR*, 2012. 1, 2, 3

[3] J. Candamo, M. Shreve, D. Goldgof, D. Sapper, and R. Kasturi. Understanding transit scenes: A survey on human behavior-recognition algorithms. *Intelligent Transportation Systems, IEEE Transactions on*, 11(1):206 –224, march 2010. 1

[4] F. de Chaumont, R. D.-S. Coura, P. Serreau, A. Cressant, J. Chabout, S. Granon, and J.-C. Olivo-Marin. Computerized video analysis of social interactions in mice. *Nature Methods*, page online, 2012. 1

[5] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55:119–139, 1997. 1

[6] H. Jhuang, E. Garrote, X. Yu, V. Khilnani, T. Poggio, A. Steele, and T. Serre. Automated home-cage behavioural phenotyping of mice. *Nature communications*, 1(6):1–9, 2010. 1

[7] D. Lin, M. P. Boyle, P. Dollar, H. Lee, E. S. Lein, P. Perona, and D. J. Anderson. Functional identification of an aggression locus in the mouse hypothalamus. *Nature*, 470(1):221–227, 2011. 3

[8] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 1

[9] S. Oh and et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, 2011. 1

[10] R. W. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010. 1

[11] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004. 1

[12] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(2):1453, 2006. 2

[13] Z. Tu. Auto-context and its application to high-level vision tasks. In *CVPR*, 2008. 1, 2

[14] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. 1

[15] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 2