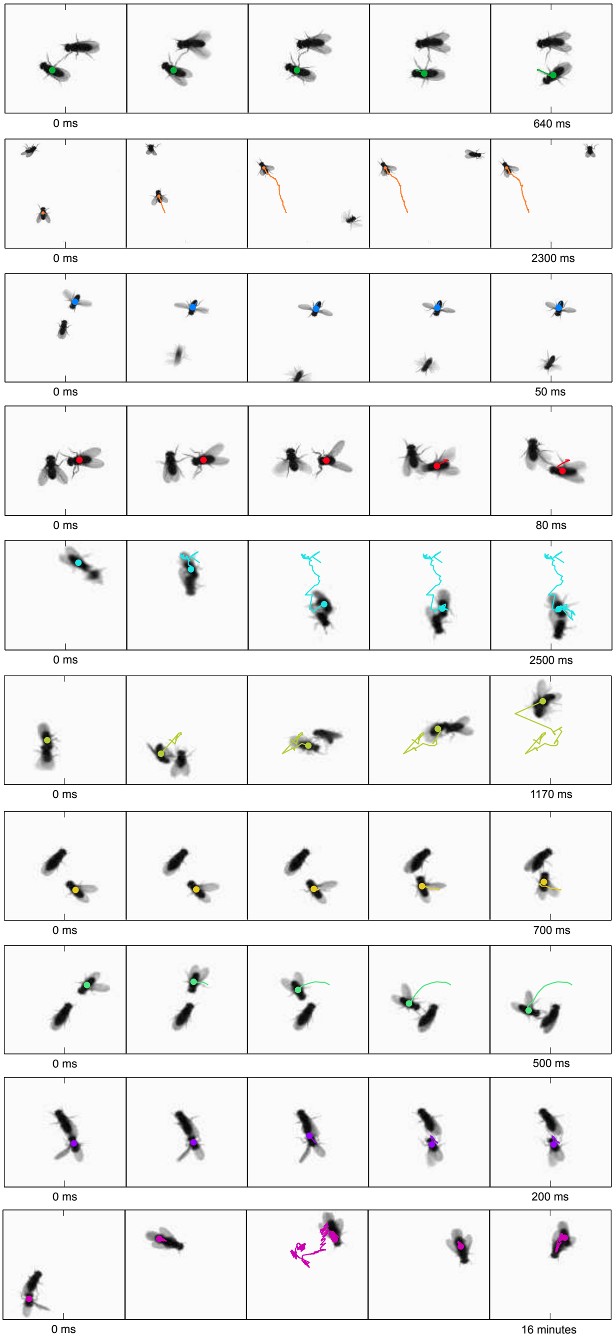
Supplementary Material

Detecting actions of social fruit flies

Actions overview



TOUCH The fly touches the leg, wing, or body of the other fly, and in doing so its gustatory organs sample chemicals that may help identify its gender. *(Neutral)*

WING THREAT The fly extends, and raises, both wings and presents them to the other fly. *(Aggressive)*

CHARGE The fly extends both wings fully and charges towards the other fly. *(Aggressive)*

LUNGE The fly raises itself on its hind legs, then slams down onto (or close to) the other fly's body. *(Aggressive)*

HOLD After lunging, the fly sometimes holds onto the body of the other fly for an extended period. *(Aggressive)*

TUSSLE The two flies lunge at each other repeatedly and tumble around in a hold. *(Aggressive)*

WING EXTENSION The fly extends one wing and vibrates it while presenting it to the other fly. *(Courtship)*

CIRCLE The fly moves along an arc around the other fly while facing it. *(Courtship)*

COPULATION ATTEMPT The fly approaches the other fly from behind, curls its abdomen towards it and tries to copulate, but is unsuccessful. *(Courtship)*

COPULATION The fly approaches the other fly from behind, curls its abdomen towards it and successfully copulates. *(Courtship)*

Figure 1: Action examples and descriptions.

Experimental setup

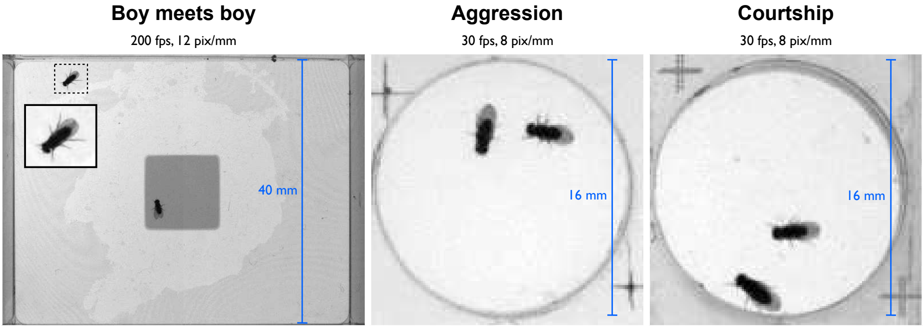


Figure 2: *Boy meets boy* has high temporal and spatial resolution videos and a large chamber with a food patch, *Aggression* and *Courtship* have lower resolution and a much smaller chamber with uniform food surface. The *Courtship* experiments contain one male and one female fly, the others have two male flies.

Bout vs. frame wise performance

A predicted output can have high precision-recall when measured on a per frame basis, but low when measured on a per bout basis, and vise versa. Figure 3 shows examples where this is the case:

i) Under/over segmented bouts → lower bout than frame wise performance
ii) Short missed/false detections → lower bout than frame wise performance
iii) Under/over estimated duration → lower frame than bout wise performance
iv) Offset bout boundaries → lower frame than bout wise performance

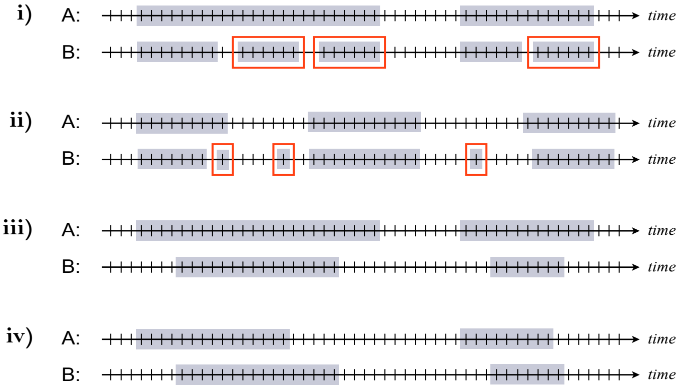


Figure 3: Examples of frame-bout performance discrepancies. Red squares denote missed/false detections, depending on whether A or B is ground truth.

Human vs. human

We trained novice annotators to learn to detect actions in the Fly-vs-Fly dataset, by showing them a subset of annotated movies, having them annotate another subset and providing them with feedback such that they could adjust their detection criteria. Once trained, they re-annotated a large portion of the test data, enough to give an idea about the difficulty of detecting each action. Overall, the trained annotators achieved best performance on the Courtship sub-dataset, which they described as being easier to annotate than the other two sub-datasets, with actions seemingly less ambiguous. Figure 4 shows the bout- and frame wise precision-recall for each action in Fly-vs-Fly, and Figure 5 explains bout-frame performance discrepancies. The human performance is a good indicator for what to expect from automatic detection algorithms; we do not expect perfection, due to action ambiguity and imperfections in ground truth annotations, but ideally they should achieve at least as good performance as humans.

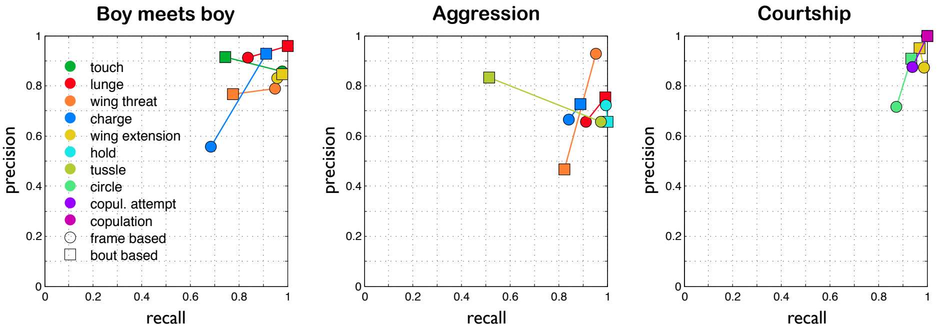


Figure 4: Human performance measured in terms of frame based (circles) and bout based (squares) precision-recall.

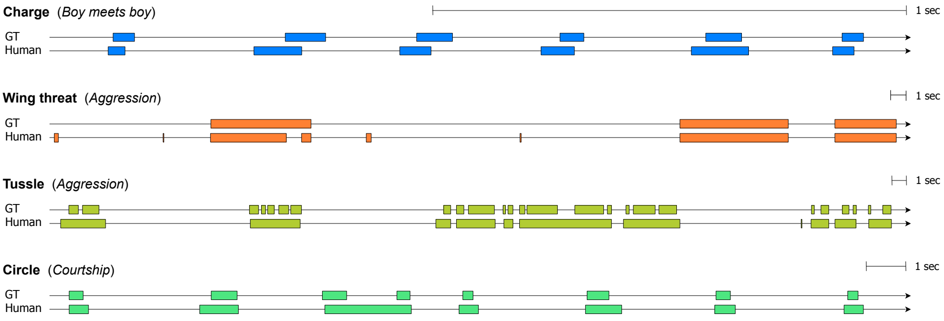


Figure 5: Segmentation samples of actions with high frame-bout performance discrepancy. Here GT refers to experts and Human to trained annotators.

Method performance per action

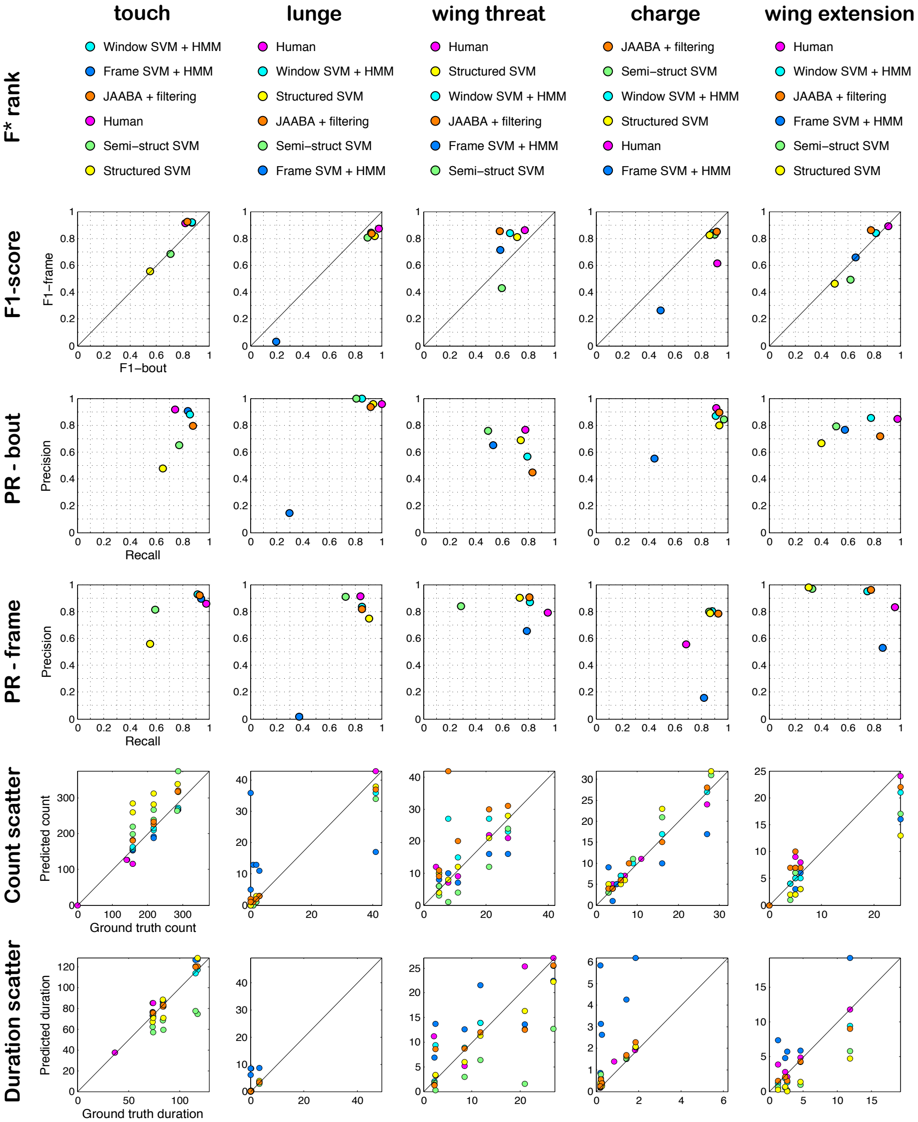


Figure 6: Results: Boy meets boy

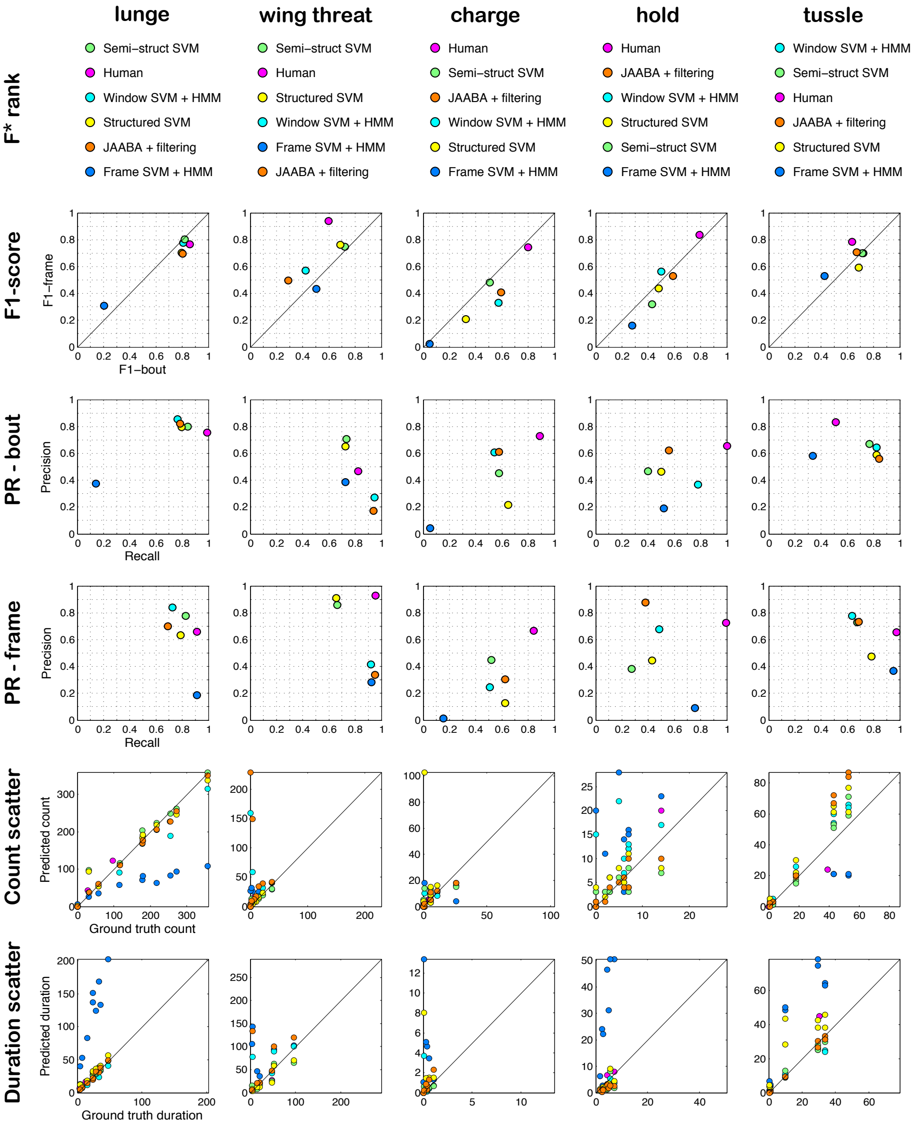


Figure 7: Results: Aggression

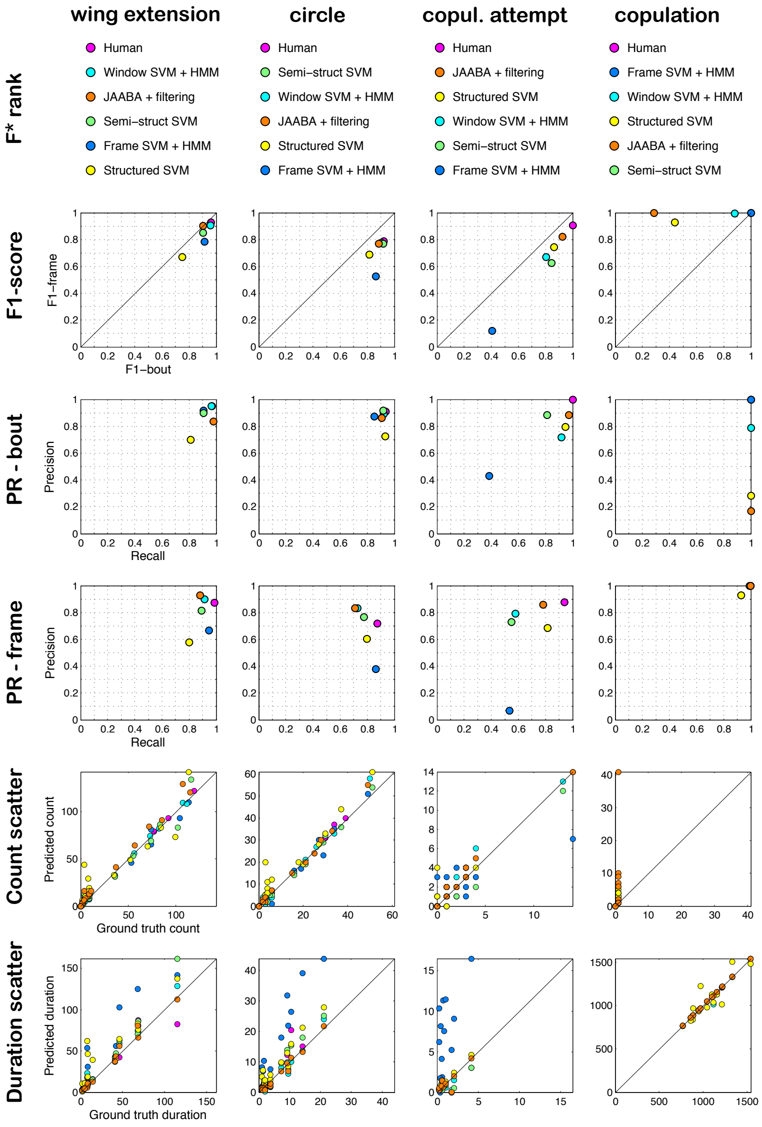


Figure 8: Results: Courtship