Can Humans Fly? Action Understanding with Multiple Classes of Actors

Chenliang Xu[1], Shao-Hang Hsieh[1], Caiming Xiong[2] and Jason J. Corso[1]
[1]Electrical Engineering and Computer Science, University of Michigan, Ann Arbor. [2]Statistics, University of California, Los Angeles.
[1]{cliangxu,shaohang,jjcorso}@umich.edu [2]caimingxiong@ucla.edu

Can humans fly? Emphatically no. Can cars eat? Again, absolutely not. Yet, these absurd inferences result from the current disregard for particular types of actors in action understanding. There is no work we know of on simultaneously inferring actors and actions in the video, not to mention a dataset to experiment with. Our paper hence marks the first effort in the computer vision community to jointly consider various types of actors undergoing various actions. To be exact, we consider seven actor classes (*adult*, *baby*, *ball*, *bird*, *car*, *cat*, and *dog*) and eight action classes (*climb*, *crawl*, *eat*, *fly*, *jump*, *roll*, *run*, and *walk*). Although jointly considering actor and action increases the sample-space complexity, it also leads to more distinct visual signatures in the joint space. For example, although a *bird* and an *adult* can both *eat*, the space-time appearance of a *bird eating* and an *adult eating* are different in significant ways. Our experiments quantitatively demonstrate that inference jointly over actors and actions outperforms inference independently over them.

To support these new actor-action understanding problems, we have created a new dataset, which we call the Actor-Action Dataset or A2D (see Fig. 1), that is labeled at the pixel-level for actors and actions (densely in space over actors, sparsely in time). The A2D has 3782 videos with 99 or more instances per valid actor-action tuple. The trimmed videos have an average length of 136 frames, with a minimum of 24 frames and a maximum of 332 frames. One-third of the videos in A2D have more than one actor performing different actions, which further distinguishes our dataset from most current action classification datasets.

We formulate a general actor-action understanding framework and implement it for three specific problems. These three problems cover different levels of modeling and hence allow us to analyze the new problem thoroughly. Without loss of generality, let $\mathcal{V} = \{v_1, \dots, v_n\}$ denote a video with n voxels in space-time lattice Λ^3 or n supervoxels in a video segmentation [2] represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. We use $\mathcal{X}$ to denote the set of actor labels and $\mathcal{Y}$ to denote the set of action labels. Consider a set of random variables $\mathbf{x}$ for actor and another $\mathbf{y}$ for action. Then the general actor-action understanding problem is specified as a posterior maximization:

$$(\mathbf{x}^*, \mathbf{y}^*) = \underset{\mathbf{x},\mathbf{y}}{\arg\max}\, P(\mathbf{x}, \mathbf{y} | \mathcal{V}) \quad . \tag{1}$$

Single-Label Actor-Action Recognition. Here, $\mathbf{x}$ and $\mathbf{y}$ are simply scalars x and y, respectively, depicting the single actor and action label to be specified for a given video $\mathcal{V}$.

Multiple-Label Actor-Action Recognition. Here, $\mathbf{x}$ and $\mathbf{y}$ are binary vectors of dimension $|\mathcal{X}|$ and $|\mathcal{Y}|$ respectively, x_i takes value 1 if the ith actor is present in the video. We define $\mathbf{y}$ similarly.

Actor-Action Semantic Segmentation. We define the two sets of random variables $\mathbf{x} = \{x_1, \dots, x_n\}$ and $\mathbf{y} = \{y_1, \dots, y_n\}$ to have dimensionality in the number of voxels or supervoxels of a video $\mathcal{V} = \{v_1, \dots, v_n\}$, and assign each $x_i \in \mathcal{X}$ and each $y_i \in \mathcal{Y}$.

We thoroughly analyze empirical performance of both state-of-the-art and baseline methods, including naïve Bayes (independent over actor and action, see Fig. 2 (a)), a joint product-space model (each actor-action pair is considered as one class, see Fig. 2 (b)), and a bilayer graphical model inspired by [1] that connects actor nodes with action nodes (see Fig. 2 (c)). Furthermore, we propose a trilayer model (see Fig. 2 (d)) that has edges linking both actor and action nodes to the actor-action pair nodes. A set of conditional classifiers are explicitly trained for this model and they are the main reason for the increase in performance: separate classifiers for the same action conditioned on the type of actor are able to exploit the characteristics unique to that actor-action tuple. For example, when we train a conditional classifier for action *eating* given actor *adult*, we use all other actions performed by *adult* as negative training samples. Therefore our trilayer model considers all relationships in the individual actor and action spaces as well as the joint product space. In other words, the previous three baseline models are all special cases of the trilayer model.

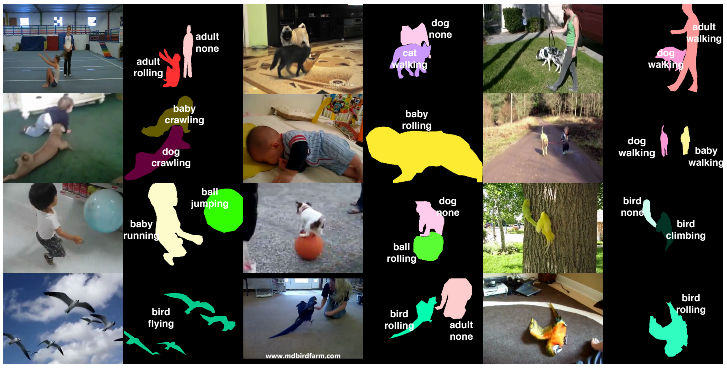

Figure 1: Montage of labeled videos in our new actor-action dataset, A2D. Examples of single actor-action instances as well as multiple actors doing different actions are present in this montage.

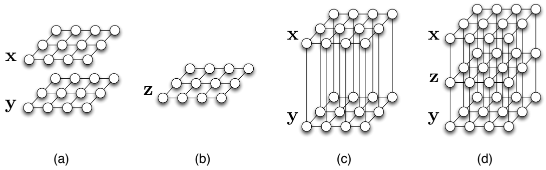

Figure 2: Visualization of different graphical models to solve Eq. 1. The figure here is for simple illustration and the actual voxel or supervoxel graph is built for a video volume.

	Single-Label A-A Recognition		
	Classification Accuracy		
Model	Actor	Action	<A, A>
N-Bayes	70.51	74.40	56.17
JointPS	72.25	72.65	61.66
Trilayer	**75.47**	**75.74**	**64.88**

	Multiple-Label A-A Recognition		
	Mean Average Precision		
Model	Actor	Action	<A, A>
N-Bayes	76.85	78.29	60.13
JointPS	76.81	76.75	63.87
Trilayer	**78.42**	**79.27**	**66.86**

	A-A Semantic Segmentation		
	Average Per Class Accuracy		
Model	Actor	Action	<A, A>
N-Bayes	44.78	42.59	19.28
JointPS	41.96	40.09	21.73
Conditional	44.78	41.88	24.19
Bilayer	44.46	43.62	23.43
Trilayer	**45.70**	**46.96**	**26.46**

Figure 3: Experimental results of the three problems.

Our thorough assessment of all instantiations of the actor-action understanding problem at both the coarse video-recognition level and the fine semantic segmentation level (see Fig. 3) provides strong evidence that the joint modeling of actor and action improves performance over modeling each of them independently. We find that for both individual actor and action understanding and joint actor-action understanding, it is beneficial to jointly consider actor and action. A proper modeling of the interactions between actor and action results in dramatic improvement over the baseline models of the naïve Bayes and joint product space models, as we observe from the bilayer and trilayer models.

Our full dataset including annotations as well as computed features, codebase, and evaluation regimen are released at http://web.eecs.umich.edu/~jjcorso/r/a2d/ to support further inquiry into this new and important problem in video understanding.

[1] L. Ladickỳ, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. H. S. Torr. Joint optimization for object class segmentation and dense stereo reconstruction. *IJCV*, 2012.

[2] C. Xu, C. Xiong, and J. J. Corso. Streaming hierarchical video segmentation. In *ECCV*, 2012.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.