Zero-shot Recognition of Complex Action Sequences

Jonathan D. Jones* Tae Soo Kim* Michael Peven* Jin Bai
Zihao Xiao Yi Zhang Weichao Qiu Alan Yuille Gregory D. Hager
Johns Hopkins University
3400 N. Charles Street, Baltimore, MD, USA
{jdjones,tkim60,mpeven,jbai12,zxiao10,yzhan286,wqiu7,ayuille1,hager}@jhu.edu

Abstract

Zero-shot video classification for fine-grained activity recognition has largely been explored using methods similar to its image-based counterpart, namely by defining image-derived attributes that serve to discriminate among classes. However, such methods do not capture the fundamental dynamics of activities and are thus limited to cases where static image content alone suffices to classify an activity. For example, reversible actions such as entering and exiting a car are often indistinguishable.

In this work, we present a framework for straightforward modeling of activities as a state machine of dynamic attributes. We show that encoding the temporal structure of attributes greatly increases our modeling power, allowing us to capture action direction, for example. Further, we can extend this to activity detection using dynamic programming, providing, to our knowledge, the first example of zero-shot joint segmentation and classification of complex action sequences in a larger video.

We evaluate our method on the Olympic Sports dataset where our model establishes a new state of the art for standard zero-shot-learning (ZSL) evaluation as well as outperforming all other models in the inductive category for general (GZSL) zero-shot evaluation. Additionally, we are the first to demonstrate zero-shot decoding of complex action sequences on a widely used surgical dataset. Lastly, we show that we can even eliminate the need to train attribute detectors by using off-the-shelf object detectors to recognize activities in challenging surveillance videos.

1. Introduction

When learning activity recognition models using deep neural networks, most approaches assume a fully supervised problem setting where 1) all categories of query actions are known a priori, 2) example instances from such categories are made available during training and 3) the pre-defined closed set of labels are supported by a large and relatively balanced set of examples. Taken together, this has led to an emphasis on ever more advanced regression-style approaches, whereby a neural network model is trained and scored on held-out examples from the same label set in an end-to-end bottom-up fashion. However, many real-world applications do not fit this model because they are naturally “open-set” problems where new labels may be defined at test time, and/or are fine-grained and compositional so that a combinatorial number of possible activity labels may exist, and/or may be data poor so that sufficient labeled training data may not exist for the desired use case. For example, in video surveillance, the goal is often to detect specific unusual activities in a zero-shot manner, e.g. “locate instances where a light brown package is being placed under a car by a man wearing a gray parka.” To successfully answer such a structured query, the ability of a zero-shot system to compose together detectable actor-object relational attributes in an on-demand fashion is highly desired.

In this paper, we present a framework for zero-shot recognition of complex action sequences that models an activity as a sequence of dynamic action signatures. In our framework, an action signature is a particular configuration of visually detectable entities, such as attributes, objects and relations, that describe a temporally local segment of a video. A fundamental observation in our work is that such configurations are often dynamic, rather than static—i.e. an action’s attributes change over time in a characteristic manner. For example, the act of a person entering a vehicle as shown in Figure [1] can be defined as “a person near a vehicle moving into a vehicle”. This can be described as the attribute sequences a person exists followed by a person does not exist and a vehicle exists.

In the remainder of this paper, we show that dynamic action signatures provide a powerful semantic label em-
bedding for zero-shot activity classification and establish a new state-of-the-art zero-shot classification benchmark on a standard zero-shot-learning (ZSL) dataset, Olympic Sports. We also use our methodology to impose constraints on the predicted action sequences themselves, leading to the first zero-shot segmentation results on complex action sequences in a challenging surgical dataset, and establish, for the first time, a zero-shot baseline result that is competitive with end-to-end trained methods.

Finally, in section 4.3 we eliminate any kind of supervised training on the dataset from which unseen (test) cases are drawn by using publicly available, off-the-shelf object detectors to provide action signatures for video surveillance. We combine this with our activity models to provide a true de novo model of an activity. We provide both quantitative and qualitative results of our zero-shot framework using these “on the fly” models on the challenging DIVA dataset, which contains fine-grained human-object interactions under a real world video surveillance setting.

In summary, the main contributions of the paper are:

- A zero-shot classification of complex action sequences with dynamic action signatures which establishes a new state-of-the-art on Olympic Sports dataset. We outperform all other methods for the ZSL evaluation regardless of training assumptions (inductive/transductive).
- To the best of our knowledge, we are the first to demonstrate zero-shot decoding of complex action sequences. We present our results on a surgical dataset, JIGSAWS, to jointly segment and classify fine-grained surgical gestures where we establish an impressive baseline.
- A demonstration of zero-shot classification of fine-grained human-object interactions that requires no supervised training of attributes by leveraging off-the-shelf object detectors in video surveillance.

2. Related Work

Zero-shot recognition approaches aim to compose classifiers for novel unseen categories at inference time given only descriptions provided by either humans or existing knowledge bases without observing any training instances from the test categories. Recent zero-shot action recognition methods focus on how to better learn or establish a mapping between the visual feature and semantic label descriptor spaces. However, in this work, our focus is on designing a better semantic space itself especially geared towards zero-shot recognition of fine-grained activities.

A visual representation of a video is often computed using low-level trajectory based descriptors or deep convolutional neural networks. Either a Fisher vector encoding over trajectory features or an average pooled deep feature serves as an encoding that describes the video as a whole. In comparison, our approach explicitly represents a video as a time series of attribute predictions extracted from temporally local video snippets and does not apply a global aggregation function to the features which allows us explicitly model dynamic state changes of attributes in an activity sequence.

Visual attributes as semantic descriptions of activity labels: The line of work that uses attribute based semantic embedding for zero-shot action recognition followed a pioneering work that originally proposed to categorize novel objects using visual attributes. As a natural extension, manually defined visual attributes have been widely used.
to provide semantic descriptions of human actions for zero-shot learning [2, 5, 7, 21, 24, 46]. Using a fixed collection of attributes, a given sample is embedded as a vector of binary [21, 24] or soft assignments [6, 39] of attribute presence in the input. Manually specified attributes define a powerful semantic embedding space as evidenced by recent attribute based approaches [8, 27, 31, 37] that consistently outperform word-embedding based approaches on zero-shot human action recognition benchmarks [20, 33, 43]. Our approach similarly uses a manually specified action attribute space as our label embedding. However, other methods use a ‘static’ version of action attributes meaning that the semantics are limited to either presence or absence of a particular attribute in a given action sequence. However, we present dynamic attributes where an action attribute can exhibit state changes over time.

**Word embedding as semantic descriptions of activity labels:** The main criticism of attribute based zero-shot learning approaches is that the manual effort involved in defining and associating attributes with activities is not scalable. Alternatively, paired with improvements in natural language processing and parsing [30], a hugely popular Word2vec implementation [29] has attracted many researchers to use word embedding for zero-shot recognition of actions [2, 3, 8, 12, 13, 25, 27, 31, 37, 38, 46, 48, 49]. Alternatively, GloVe [36] provides an alternative text based embedding by constructing a large matrix of co-occurrence statistics between words and contexts. Semantic vectors are obtained such that dot product of co-occurring word and context vectors equals the co-occurrence probability. GloVe is used as the semantic embedding for labels in [15, 50, 51] with promising results. The fundamental assumption of word embedding based approaches is that an activation pattern of the feature layer of the skip-gram model of [29] given an interest-word (text of activity label) as input provides a discriminative representation of novel categories. Text provides an elegant way to circumvent manual definition all together and still provide a reasonable semantic description of activities in a zero-shot manner by leveraging text models trained on large scale text corpora. However, it is at the cost of zero-shot recognition performance as manual attributes consistently outperform Word2vec based representations given the same methodology [26]. The performance gap indicates that there still exists a significant gap between the underlying distributions of visual and text data. Moreover, we believe an important factor in zero-shot activity recognition (as opposed to static applications such as object or scene classification) is to model the temporal evolution of elements in video. A simple text embedding of activity names is not built to describe dynamic elements of a video in full.

**Objects as attributes:** The work of [16] constructs a word embedding augmented by a skip-gram model of object categories in videos. Further, a spatial-aware object based embedding is proposed in [28] for additional zero-shot localization of actions. An approach to learn relations between action-attribute-object in an end-to-end manner using two-stream graph convolutional network is proposed in [8]. We also view objects as a promising source of additional information to provide a rich semantic embedding for zero-shot activity recognition. Our approach allows temporal modeling of object presence and demonstrate that we can compose actor-object interaction detectors in a zero-shot manner using off-the-shelf object detectors in Section 4.3.

**In practice, the distributions of seen and unseen categories are often not well aligned for accurate ZSL inference.** Researchers have identified this problem as the domain shift problem in ZSL, analyzed empirically in [6] and theoretically in [39], which has led to a series of approaches [5, 48, 19, 31, 8, 27] that allow the use of unlabeled instances from the unseen test categories as part of training, defined as the transductive setting for ZSL. In this work, we focus on introducing our new semantic embedding space with dynamic attributes that enables improved zero-shot action recognition rather than a method to better solve the domain shift problem between the seen and unseen categories.

### 3. Methodology

We first establish a basic hierarchy of concepts. At the highest level we have the *activity*—for example, suturing in robotic surgery. Each activity can be decomposed into a sequence of actions \( (y_1, \ldots, y_N) \). Possible examples of actions are “Pushing needle through tissue” in a suturing activity or “throwing javelin” in a sporting event. Zero-shot learning approaches further decompose each action \( y \) into a set of \( K \) elementary attributes (usually taken to be binary-valued) \( y = \{ a_1, \ldots, a_K \} \).

Given a video recording of an activity (represented as a sequence of frames \( X = (x_1, \ldots, x_T) \)), our goal is to map each frame \( x_t \) to its corresponding action \( y_t \) by detecting the presence or absence of each attribute \( \hat{a}(x_t) \) in the video, then choosing the action whose signature \( a(y_t) \) best fits those attributes. In other words, we choose the action with highest score:

\[
\hat{y}(x) = \arg\max_y \text{score}(a(y), \hat{a}(x))
\]  

Our methods focus on defining signatures conveniently, and computing the score efficiently.

#### 3.1. Dynamic Attribute Labeling

Previous work in zero-shot action recognition defines each signature over a set of attributes that are static—i.e. each attribute is presumed to be constant through the duration of the action. However, in many scenarios the actions
of interest are distinguished by their time evolution rather than the presence or absence of static attributes. Take “person entering a vehicle” and “person exiting a vehicle”, for example. Both of these actions share the static attribute “vehicle present”. However, they are differentiated from each other by what happens to the person over time—in an “entering” action the person disappears into the vehicle, but in an “exiting” action the person appears out of it.

In this section we outline a simple and elegant method for implementing dynamic attribute signatures, which also generalizes previous work. Our method is flexible enough to accept a high-level ordering of events, but also permits more temporal information to be provided if it is known. For example, it can implement a signature for “person appears” like “person is absent, then person is present”, or one additionally specifying that a person should be absent for the first 75% of a segment, and present for the remaining 25%. Finally, several existing zero-shot learning datasets are annotated with static attributes, but do not have temporal information. Our framework allows new dynamic signatures to be defined quickly and easily by specifying the temporal evolution of relevant attributes on a per-activity basis.

### 3.2. Activity Signatures

Because they are well-studied, flexible, and easily-composable, we implement our methods using finite-state logic (specifically using the OpenFST toolkit [4]). For a comprehensive overview of finite-state methods and their use in sequence models, see [32].

We define action signatures as finite-state acceptors that implement time-varying rules. Each signature accepts a sequence of attribute detections. In figure 2 we have illustrated state machines implementing the two rules for “Person appears” from the previous section.

At inference time we receive a sequence of attribute detection scores from some black-box system, and need to determine its compatibility with our set of pre-defined attribute signatures. To do so, for each detection sequence, we first instantiate a finite-state transducer that accepts sample indices as input, gives detections as output, and whose weights correspond to attribute detection scores (figure 3 illustrates a minimal example). We then compose that transducer with its corresponding attribute acceptor, giving a machine that measures detection inputs against the attribute rule. Finally, we align the rule to the detection sequence by computing the best path through the state machine, and take the score to be the resulting weight.

### 3.3. Zero-shot reasoning for complex activities

As we established at the beginning of this section, complex activities contain sequences of actions. We can extend the zero-shot classification scenario to perform joint classification and segmentation in a zero-shot manner by defining a sequence-level score function over $M$ hypothesized segments:

$$
score(y, x) = \sum_{i=1}^{M} \phi(y_i, t_i, d_i, x) \tag{2}
$$

In equation $\phi$ implements the segment-level score function of eq. 1 defined over the $i$-th hypothesized segment with start time $t_i$, duration $d_i$, and label $y_i$.

In many cases, these activities have a structure that is known $a$ priori, and which can be exploited to disallow impossible action sequences. For example, the JIGSAWS dataset is composed of surgical suturing videos. In these sequences, only certain gesture sequences are realizable. By adding a pairwise label score, we can impart this knowledge to the system:

$$
score(y, x) = \sum_{i=1}^{M} \phi(y_i, t_i, d_i, x) + \psi(y_i, y_{i-1}) \tag{3}
$$

In practice, we use $\psi(y_i, y_{i-1})$ to implement first-principles knowledge by giving a score of $0$ to a possible transition, and a score of $-\infty$ to an impossible one.

---

$^2$ Note that one could alternatively use the weight of all paths through the machine. Since the best-path and total-path weights can be computed using the same dynamic program by changing the definitions of $(+, \times)$ we focus on the best path without loss of generality.
We can still use the principle in equation 1 when predicting action sequences (i.e., decoding). However, since \( \gamma \) is now a sequence, it is more efficient to compute the argmax using a dynamic program—for example, one of the algorithms presented in [40] or [22].

4. Experiments

We first evaluate our approach on the Olympic Sports [33] dataset in 4.1 which demonstrates the added representational power of dynamic action signatures for zero-shot action classification. In Section 4.2 we then describe our approach for zero-shot decoding of complex action sequences using the JIGSAWS [9] dataset for segmenting a sequence of surgical gestures in a video. Finally in Section 4.3 we demonstrate that fine-grained human-object interactions can be recognized in a zero-shot manner that requires no training at all by harnessing off-the-shelf object detectors.

4.1. Olympic Sports

The Olympic Sports [33] dataset contains 783 videos from 16 sports action categories and we adhere to the 50/50 data splits proposed by [43] to compare our zero-shot experimental results to existing approaches.

Visual features and class embedding: In our experiments, an I3D [1] is trained to predict all attributes from a video segment. Recent methods that use manually labeled attributes as the class embedding (listed in Table 1) often map visual features from the entire clip to an attribute signature in a bag-of-words manner. In contrast, in our approach, we explicitly model the temporal nature of the attribute by capturing their state change through time. In our implementation, we process a video snippet of length 64 and adopt a stride of 32. This leads to a varying number of attribute predictions for a given video sample depending on its temporal duration.

Implementation Details: We implement the following five dynamic attribute patterns defined in Section 3: (0): Absence, (1): Persistence, (2): Start, (3): End and (4): Sometimes. Refer to the supplementary material for all definitions and state machines implemented for each activity using dynamic action signatures. We use an I3D [1] pretrained on Charades [41] and the code including all train settings and model hyperparameters will be made public upon publication.

Experimental Settings: Our approach only requires a correspondence between dynamic action signatures and activities from the seen categories to train the attribute detectors, thus categorizing our method as adhering to the inductive setting. Though our model is naturally inductive, we report a transductive version of the model by introducing an additional auxiliary classifier that is trained to predict whether a sample is from the seen or unseen categories. The purpose of the auxiliary classifier is simply limit the search space of possible actions based on the seen vs. unseen prediction.

Dynamic action signatures improve zero-shot action recognition: As shown in Table 1 our method outperforms both inductive and transductive models by a considerable margin under the ZSL evaluation framework and outperforms other inductive models for the GZSL evaluation. We note that for the transductive setting, we do not modify the original model in any way but introduce an additional auxiliary classifier that is trained to predict whether a sample is from the seen or unseen categories. The auxiliary prediction limits the search space of possible actions and improves performance for the GZSL evaluation as depicted in Table 2.

| Method | Emb | ID / TD | ZSL | GZSL |
|--------|-----|---------|-----|------|
| SJE [2] | W | ID | 28.6 ± 4.9 | 32.5 ± 6.7 |
| IAP [21] | A | ID | 42.3 ± 12.5 | N/A |
| DAP [21] | A | ID | 45.4 ± 12.8 | N/A |
| HAA [24] | A | ID | 46.1 ± 12.4 | 49.4 ± 10.8 |
| SJE [2] | A | ID | 47.5 ± 14.8 | 32.5 ± 6.7 |
| GA [31] | A | ID | 50.4 ± 11.2 | N/A |
| GCN [8] | W | TD | 56.5 ± 6.6 | N/A |
| ECC [37] | W | ID | 59.8 ± 5.6 | N/A |
| **Ours** | A | ID | **71.7 ± 8.5** | **55.7 ± 4.1** |
| GA [31] | W | TD | 41.3 ± 11.4 | 42.2 ± 10.2 |
| MIC [51] | G | TD | 43.9 ± 7.9 | N/A |
| GCN [8] | W | TD | 59.9 ± 5.3 | 50.2 ± 6.8 |
| GA [31] | A | TD | 57.9 ± 14.1 | 52.4 ± 12.2 |
| OD [27] | W | TD | 50.5 ± 6.9 | 53.1 ± 3.6 |
| OD [27] | A | TD | 65.9 ± 8.1 | 66.2 ± 6.3 |
| **Ours** | A | TD | **71.7 ± 8.5** | **72.7 ± 8.1** |

Table 1: Comparison of zero-shot action recognition accuracies on the Olympic Sports dataset. Label embedding: Manually defined attributes (A) or word vector embedding (W); Inductive (ID) or transductive (TD) settings; Standard zero-shot learning (ZSL) or generalized zero-shot learning (GZSL) evaluations.

| Emb | ID / TD | s | u | H |
|-----|---------|---|---|---|
| OD [27] | W | TD | 73.2 | 41.8 | 53.1 |
| OD [27] | A | TD | 71.5 | 61.6 | 66.2 |
| **Ours** | A | ID | 72.0 | 46.0 | 56.1 |
| **Ours** | A | TD | 75.8 | 69.9 | 72.7 |

Table 2: Evaluation of our model under the inductive and transductive settings. The auxiliary seen vs. unseen category classifier significantly improves our model’s ability to accurately classify unseen categories. Seen classes (s), unseen classes (u) and the harmonic mean (H).
Table 3: Comparison of models using static vs. dynamic action signatures. We observe a large gain in performance for ‘bowling’ by allowing attributes such as ‘throw away’ and ‘one-arm swing’ to be dynamic over time. (S): Static version; (D) Dynamic version of our approach.

Table 4: Zero-shot segment classification accuracy on JIGSAWS: comparison of static and dynamic attribute signatures.

4.2. JIGSAWS

We use the JIGSAWS dataset [9] to evaluate our attribute learning and gesture classification methods described above. This is a publicly available dataset containing 39 instances of eight surgeons performing a benchtop simulation training task of suturing in a robot-assisted minimally invasive surgical setting using the da Vinci Surgical System. The dataset contains endoscopic video of the performance as well as motion data for instruments and manipulators that the surgeons control on the system. In this work, we do not use the instrument motion data. Each performance has per-frame gesture class labels for 10 types of actions that occur during the task. JIGSAWS only provides annotations for gestures (activities), and so we use the method described in Section 3.1 to obtain per-frame attribute annotations.

| Categories    | HAA [24] | Ours-S | Ours-D |
|---------------|----------|--------|--------|
| bowling       | 0.0      | 10.6   | 68.1   |
| hammer-throw  | 91.3     | 100    | 87.0   |
| javelin-throw | 88.0     | 87.5   | 62.5   |
| snatch        | 85.7     | 100    | 95.9   |
| mean acc.     | 66.3     | 74.5   | 78.4   |

Figure 4: A visualization of attribute prediction scores over time for a ‘bowling’ sequence. Temporal attribute pattern \((3):End\) provides a better description for ‘open-one-arm’, ‘throw-away’ and ‘one-arm-swing’.

4.2.1 Classification

Although the JIGSAWS dataset is composed of complex surgical activities made up of action sequences, we first evaluate the performance of our method on the traditional action-classification setup. In this experiment we use ground-truth action boundaries to segment each sequence into a set of actions.

Table 4 show the results of an ablation experiment studying the effectiveness of our dynamic attribute signatures on the JIGSAWS dataset. For the static-signature system, we map all dynamic signatures to their nearest static counterparts. Specifically, we map signatures (2) “at beginning” and (3) “at end” to (0) “never”. Notice that accuracy in-
creases by almost 20% when dynamic signatures are used. This is due to an inherent confusion between gestures 4 and 6—the static-signature model overwhelmingly predicts gesture 4 for gesture 6, because these gestures’ signatures differ only in the temporal duration of a single attribute. Allowing dynamic attribute signatures disambiguates between these two gestures, and improves accuracy for gesture 6 from 0% to 83%.

4.2.2 Joint classification and segmentation

We next turn to the task of zero-shot decoding (i.e. joint classification and segmentation) of surgical activity. This task can be performed in a naive way by doing zero-shot classification for individual samples or windows of samples, but frequently practitioners are aware of additional structure that restricts which action sequences are realizable. In this experiment, we compare the performance of a grammar derived from first-principles knowledge of surgical suturing tasks with an unstructured baseline.

More specifically, our grammar describes an ideal execution of the suturing task (see fig. 5 for an illustration). The practitioner begins by reaching for the needle (G1), then moves to the work area (G5), then executes a suture (G2 - G6). At this point they can either transfer the needle from the left to the right hand (G4) and perform another suture, or drop the suture and end the activity (G11). Note that not every sequence in the JIGSAWS dataset conforms to this model—there is a small number of rare states (G8, G9, and G10) that correspond to errors made during the suturing process. Since this work addresses zero-shot applications, we focus on modelling reliable and structured correct cases instead of the more variable incorrect ones.

Table 5 compares the performance of our zero-shot decoding system with the unstructured baseline. Our structured model improves frame-level accuracy over the unstructured one by about 8%. However, the improvement in edit score is much more substantial. By providing information about the basic structure of the sequence derived from what we know about the underlying process, we obtain close to a 100% relative improvement in edit score.

Table 6 compares our method with previous, fully-supervised ones. The first three rows represent unstructured, neural baselines established in [22]. Interestingly, we obtain better edit distance than all of them and better accuracy than two of the three without ever training on gesture-labeled data. Furthermore, our edit distance comes close to that of the segmental spatiotemporal CNN of [22]—a fully-supervised model that also incorporates a grammar.

4.3. DIVA

Both experiments on zero-shot classification of human activities and zero-shot segmentation of surgical gestures involve a supervised training step to obtain attribute detectors using instances from seen categories. In this section, we demonstrate that publicly available, off-the-shelf object detectors can be used to compose a system to classify human-object interactions in a truly supervision-less zero-shot manner. We demonstrate how we encode first-principles temporal logic to define activities using state machines combined with off-the-shelf object detectors.

The DIVA dataset is an untrimmed activity detection dataset that provides both spatial and temporal localization of predefined set of activities. The videos originate from the VIRAT dataset [34] and annotations more suitable for activity detection were collected by the IARPA DIVA
| Method          | edit acc. |
|-----------------|-----------|
| IDT [45]        | 8.5 53.9  |
| VGG [42]        | 24.3 45.9 |
| Spatial CNN [23]| 37.7 74.0 |
| Ours            | 61.7 56.6 |
| Seg-ST-CNN [23] | 66.6 74.7 |
| ST-CNN [23]     | 68.0 77.7 |
| TCN [22]        | 83.1 81.4 |

Table 6: Comparison of our zero-shot method with previous, fully-supervised methods for joint classification and segmentation on JIGSAWS.

| Scene | Entering | Exiting |
|-------|----------|---------|
|       | LOSO     | ALL     | Ours    | LOSO     | ALL     | Ours    |
| 0000  | 9.52 36.8 | 77.8    | 23.8    | 66.7     | 100     |
| 0002  | 37.8 3.85 | 52.6    | 12.5    | 23.5     | 100     |
| 0400  | 42.1 35.3 | 88.9    | 58.8    | 37.5     | 85.7    |
| 0401  | 28.3 33.4 | 81.1    | 7.69    | 31.4     | 100     |
| 0500  | 33.4 0.0  | 100     | 14.3    | 16.7     | 100     |
| Mean  | 30.2 21.87 | 74.6    | 23.4    | 35.16 95.4 |

Table 7: Classification accuracy on the DIVA dataset comparing our approach to fully supervised baselines under the leave-one-scene-out (LOSO) or ALL evaluation settings.

Zero-shot classification without any supervised training: Given detectable objects \{Human, Vehicle\} from [14] and temporal attribute patterns defined in Section 5, we can define a human Entering and Exiting a vehicle with state machines shown in Figure 6.

We compare our zero-shot system to a state-of-the-art end-to-end supervised system [52] in both settings where the system is trained and tested on all five camera locations (ALL), and when it is trained on 4 and tested on a held-out 5th scene (LOSO). The LOSO evaluation tests performance of a supervised system when there exists a large domain gap between training and testing data points which is a reasonable assumption for many practical applications. We observe that the end-to-end baseline generalizes poorly across scenes but our zero-shot approach performs well. For the ALL setting where training conditions are more favorable to an end-to-end model, our zero-shot approach still achieves higher classification performance which suggests that the presented zero-shot approach may serve as a promising alternative when training and the use of end-to-end models is prohibited by data. It shows the value of explicitly injecting first-principles knowledge especially when adequate data for sufficiently training bottom up data-driven models is difficult. The experiments on DIVA shows that using our approach, a practitioner can quickly define a competitive zero-shot action classification system by describing activities with state machines over dynamic action signatures computed using off-the-shelf object detectors.

5. Discussion and Conclusion

As shown in Sections 4.2 and 4.3, zero-shot methods are particularly useful when data conditions are not suitable for data-driven methods. However, there inevitably exists corner cases outside the distribution captured by the zero-shot system. For example, a person exiting a top-less vehicle will always be visible and the state machine would not score such sequence highly. We believe an interesting future work would be to study how to best utilize labeled training data through our framework. Our formulation actually allows the gradients to be propagated all the way back to the attribute detectors so that they can be finetuned under the structure defined by the state machines.

In summary, we presented a framework for modeling fine-grained activities as a state machine of dynamic attributes. We show that temporal attributes define a rich semantic label embedding for zero-shot classification of fine-grained actions and establishes a new state-of-the-art results on the Olympic Sports dataset. Our approach is the first to
establish a competitive baseline for a novel task of zero-shot segmentation of complex surgical gesture sequences. Finally, we show that supervised training can be eliminated entirely by using off-the-shelf object detectors to recognize activities in the surveillance domain.

Acknowledgements. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPUs used for this research. This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00345. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Quo vadis, action recognition? a new model and the kinetics dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4724–4733, 2017.

[2] Zeynep Akata, Scott E. Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2927–2936, 2015.

[3] Ioannis Alexiou, Tao Xiang, and Shaogang Gong. Exploring synonyms as context in zero-shot action recognition. 2016 IEEE International Conference on Image Processing (ICIP), pages 4190–4194, 2016.

[4] Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. Openfst: A general and efficient weighted finite-state transducer library. In Jan Holub and Jan Žďárek, editors, Implementation and Application of Automata, pages 11–23, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[5] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In David Fleet, Tomas Pfajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision – ECCV 2014, pages 584–599, Cham, 2014. Springer International Publishing.

[6] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning multimodal latent attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(2):303–316, Feb 2014.

[7] Chuang Gan, Ming Lin, Yi Yang, Yueting Zhuang, and Alexander G. Hauptmann. Exploring semantic inter-class relationships (sir) for zero-shot action recognition. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI’15, pages 3769–3775. AAAI Press, 2015.

[8] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In AAAI, 2019.

[9] Yixin Gao, S. Swaroop Vedula, Carol E. Reiley, Narges Ahmadi, Balakrishnan Varadarajan, Henry C. Lin, Lingling Tao, Luca Zappella, Benjamin Béjar, David D. Yuh, Chi Chung Grace Chen, René Vidal, Sanjeev Khudanpur, and Gregory D. Hager. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In Modeling and Monitoring of Computer Assisted Interventions (M2CAI) MICCAI Workshop, 2014.

[10] Joshua Gleason, Rajeev Ranjan, Steven Schwarz, Carlos Castillo, Jun-Cheng Chen, and Rama Chellappa. A proposal-based solution to spatio-temporal action detection in untrimmed videos. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 141–150, 2018.

[11] Joshua Goodman. Semiring parsing. Comput. Linguist., 25(4):573–605, Dec. 1999.

[12] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkar- nenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In Proceedings of the 2013 IEEE International Conference on Computer Vision, ICCV ’13, pages 2712–2719, Washington, DC, USA, 2013. IEEE Computer Society.

[13] Meera Hahn, Andrew Silva, and James M. Rehg. Action2vec: A crossmodal embedding approach to action learning. ArXiv, abs/1901.00484, 2019.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. 2017 IEEE International Conference on Computer Vision (ICCV), pages 2980–2988, 2017.

[15] Feiran Huang, Xiaoming Zhang, and Zhoujun Li. Learning joint multimodal representation with adversarial attention networks. In Proceedings of the 26th ACM International Conference on Multimedia, MM ’18, pages 1874–1882, New York, NY, USA, 2018. ACM.

[16] Mihir Jain, Jan C. van Gemert, Thomas Mensink, and Cees G. M. Snoek. Objects2action: Classifying and localizing actions without any video example. In ICCV, pages 4588–4596. IEEE Computer Society, 2015.

[17] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014.

[18] Tae Soo Kim, Yi Zhang, Zihao Xiao, Michael Peven, Weichao Qiu, Jin Bai, Alan L. Yuille, and Gregory D. Hager. Safer : Fine-grained activity detection by compositional hypothesis testing. 2019.

[19] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. pages 2452–2460, 12 2015.

[20] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In Proceedings of the International Conference on Computer Vision (ICCV), 2011.
[21] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In In CVPR, 2009.

[22] Colin Lea, Michael Flynn, Rene Vidal, Austin Reiter, and Gregory Hager. Temporal convolutional networks for action segmentation and detection. pages 1003–1012, 07 2017.

[23] Colin Lea, Austin Reiter, René Vidal, and Gregory D. Hager. Segmented spatiotemporal cnns for fine-grained action segmentation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, Computer Vision – ECCV 2016, pages 36–52, Cham, 2016. Springer International Publishing.

[24] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In CVPR 2011, pages 3337–3344, June 2011.

[25] Kun Liu, Wu Liu, Huadong Ma, Wenbing Huang, and Xiongxiang Dong. Generalized zero-shot learning for action recognition with web-scale video data. World Wide Web, 22(2):807–824, Mar 2019.

[26] Valter Luís, Estevam Junior, Helio Pedrinib, and David Menotti. Zero-shot action recognition in videos: A survey. ArXiv, 2019.

[27] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao. Out-of-distribution detection for generalized zero-shot action recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[28] P. S. M. Mettes and C. G. M. Snoek. Spatial-aware object embeddings for zero-shot localization and classification of actions. In IEEE International Conference on Computer Vision, 2017.

[29] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS’13, pages 3111–3119, USA, 2013. Curran Associates Inc.

[30] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[31] Ashish Mishra, Vinay Kumar Verma, M. Shiva Krishna Reddy, Arulkumar Subramaniam, Piyush Rai, and Anurag Mittal. A generative approach to zero-shot and few-shot action recognition. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 372–380, 2018.

[32] Mehryar Mohri. Weighted Automata Algorithms, pages 213–254. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[33] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, Computer Vision – ECCV 2010, pages 392–405, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[34] Sangmin Oh, Anthony Hoogs, Amittha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In CVPR 2017, pages 3153–3160. IEEE, 2017.

[35] Dan Oneata, Jakob J. Verbeek, and Cordelia Schmid. Action and event recognition with fisher vectors on a compact feature set. 2013 IEEE International Conference on Computer Vision, pages 1817–1824, 2013.

[36] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In In EMNLP, 2014.

[37] Jie Qin, Li Liu, Ling Shao, Fumin Shen, Bingbing Ni, Ji-axin Chen, and Yunhong Wang. Zero-shot action recognition with error-correcting output codes. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[38] Alina Roitberg, Ziad Al-Halah, and Rainer Stiefelhagen. Informed democracy: Voting-based novelty detection for action recognition. In British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018, page 52, 2018.

[39] Bernardino Romera-Paredes and Philip H. S. Torr. An embarrassingly simple approach to zero-shot learning. In Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML’15, pages 2152–2161. JMLR.org, 2015.

[40] Sunita Sarawagi and William W Cohen. Semi-markov conditional random fields for information extraction. In L. K. Saul, Y. Weiss, and L. Bottou, editors, Advances in Neural Information Processing Systems 17, pages 1185–1192. MIT Press, 2005.

[41] Gunnar A. Sigurdsson, Olga Russakovsky, and Abhinav Gupta. What actions are needed for understanding human actions in videos? In International Conference on Computer Vision (ICCV), 2017.

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.

[43] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. CoRR, page 2012.

[44] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV ’15, pages 4489–4497, Washington, DC, USA, 2015. IEEE Computer Society.

[45] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In Proceedings of the 2013 IEEE International Conference on Computer Vision, ICCV ’13, pages 3551–3558, Washington, DC, USA, 2013. IEEE Computer Society.

[46] Qian Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. International Journal of Computer Vision, 124(3):356–383, Sep 2017.
[47] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. pages 5794–5803, 10 2017.

[48] Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive zero-shot action recognition by word-vector embedding. *International Journal of Computer Vision*, 123(3):309–333, Jul 2017.

[49] Xun Xu, Timothy M. Hospedales, and Shaogang Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 343–359, Cham, 2016. Springer International Publishing.

[50] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *ECCV*, 2018.

[51] Chenrui Zhang and Yuxin Peng. Visual data synthesis via gan for zero-shot video classification. In *IJCAI*, 2018.

[52] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2017.