Learning Visual Actions Using Multiple Verb-Only Labels

Michael Wray
michael.wray@bristol.ac.uk

Dima Damen
dima.damen@bristol.ac.uk

University of Bristol, UK

Abstract

This work introduces verb-only representations for both recognition and retrieval of visual actions, in video. Current methods neglect legitimate semantic ambiguities between verbs, instead choosing unambiguous subsets of verbs along with objects to disambiguate the actions. We instead propose multiple verb-only labels, which we learn through hard or soft assignment as a regression. This enables learning a much larger vocabulary of verbs, including contextual overlaps of these verbs. We collect multi-verb annotations for three action video datasets and evaluate the verb-only labelling representations for action recognition and cross-modal retrieval (video-to-text and text-to-video). We demonstrate that multi-label verb-only representations outperform conventional single verb labels. We also explore other benefits of a multi-verb representation including cross-dataset retrieval and verb type (manner and result verb types) retrieval.

1 Introduction

With the rising popularity of deep learning, action recognition datasets are growing in number of videos and scope [6, 15, 17, 21, 24, 34], leading to an increasingly large vocabulary required to label them. This introduces challenges in labelling, particularly as more than one verb is relevant to each action. For example, consider the common action of opening a door and the verbs that could be used to describe it. *"Open"* is a natural choice, with *"pull"* being used to specify the direction as well as *"grab"*, *"hold"* and *"turn"* characterising the method of turning the handle. Any action can thus be richly described by a number of different verbs. This is contrasted with current approaches which use a reduced set of semantically distinct verb labels, combined with the object being interacted with (e.g. *"open-door"*, *"close-jar"*). Sigurdsson *et al.* [35] show that using only single verb labels, without any accompanying nouns, leads to increased confusion among human annotators. Conversely, verb-noun labels cause actions to be tied to instances of objects, *i.e.* opening a fridge is the same interaction as opening a cupboard. Additionally, using a single verb doesn't explore the complex overlapping space of verbs in visual actions. In Fig. 1(a), four 'open' actions are shown, where the verb 'open' overlaps with four different verbs: 'push' and 'pull' for two doors, 'turn' when opening a bottle and 'cut' when opening a bag.

In this paper, we deviate from previous works which use a reduced set of verb and noun pairs [5, 6, 10, 21, 23, 34, 37, 38], and instead propose using a representation of multiple
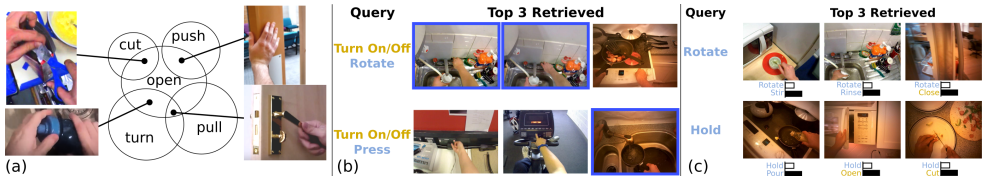
Figure 1: (a) Open can be performed in multiple ways leading to complex overlaps in the space of verbs. Using multiple verb-only representations allows: (b) Retrieval that distinguishes between different manners of performing the action. Actions of turning on/off by *"rotating"* are separated from those of turning on/off by *"pressing"*; (c) Verbs such as *"rotate"* and *"hold"* can be learned via context from multiple actions. [supp. video]

verbs to describe the action. The combination of multiple, individually ambiguous verbs allow for an object-agnostic labelling approach, which we define using an assignment score for each verb. We investigate using both hard and soft assignment, and evaluate the representations on two tasks: action recognition and cross-modal retrieval (*i.e.* video-to-text and text-to-video retrieval). Our results show that hard-assignment is suitable for recognition while soft-assignment improves cross-modal retrieval.

A multi-verb representation presents a number of advantages over a verb-noun representation. Figure 1(b) demonstrates this by querying the multi-verb representation with *"turn-off"* combined with one other verb (*"rotate"* vs. *"press"*). Our representation can differentiate between a tap turned off by rotating (first row in blue) from one turned off by pressing (second row in blue). Both single verb representations (i.e. 'turn-off') and verb-noun representations (i.e. 'turn-off tap') cannot learn to distinguish these different interactions. In Fig 1(c), verbs such as *"rotate"* and *"hold"*, describing parts of the action, can also be learned via context of co-occurrence with other verbs.

To the best of the author's knowledge, (**our contributions**): (i) this paper proposes to address the inherent ambiguity of verbs, by using multiple verb-only labels to represent actions, keeping these representations object-agnostic. (ii) We annotate these multi-verb representations for three public video datasets (BEOID [42], CMU [58] and GTEA Gaze+ [10]). (iii) We evaluate the proposed representation for two tasks: action recognition and action retrieval, and show that hard assignment is more suited for recognition tasks while soft assignment assists retrieval tasks, including for cross-dataset retrieval.

2   Related Work

Action Recognition in Videos.   Video Action Recognition datasets are commonly annotated with a reduced set of semantically distinct verb labels [5, 6, 7, 10, 15, 17, 34, 36]. Only in EPIC-Kitchens [6], verb labels are collected from narrations with an open vocabulary leading to overlapping labels, which are then manually clustered into unambiguous classes. Ambiguity and overlaps in verbs has been noted in [22, 42]. Our prior work [42] uses the verb hierarchy in WordNet [27] synsets to reduce ambiguity. We note how annotators were confused, and often could not distinguish between the different verb meanings. Khamis and Davis [22] used multi-verb labels in action recognition, on a small set of (10) verbs. They jointly learn multi-label classification and label correlation, using a bi-linear approach, allowing an actor to be in a state of performing multiple actions such as *"walking"*

and *"talking"*. This work is the closest to ours in motivation, however their approach uses hard assignment of verbs, and does not address single-verb ambiguity, assuming each verb to be non-ambiguous. Up to our knowledge, no other work has explored multi-label verb-only representations of actions in video.

Action Recognition in Still Images.   Gella and Keller [12] present an overview of datasets for action recognition from still images. Of these, four use a large (>50) number of verbs as labels [4, 13, 33, 45]. ImSitu [45] uses single verb labels, however, they note ambiguities between verbs — with each verb having an average of 3.55 different meanings. They report top-1 and top-5 accuracies to account for the multiple roles each verb plays. The HICO dataset [4] has multi-label verb-noun annotations for its 111 verbs, but the authors combine the labels into a powerset to avoid the overlaps between verbs. Two datasets, Verse [13] and COCO-a [33], use external semantic knowledge to disambiguate a verb's meaning. For zero-shot action recognition of images, Zellers and Choi [47] use verbs in addition to attributes such as global motion, word embeddings, and dictionary definitions to solve both text-to-attributes and image-to-verb tasks. While their labels are coarse grained (and don't describe object interactions) they demonstrate the benefit of using a verb-only representation.

Although works are starting to use a larger number of verbs, the class overlaps are largely ignored through the use of clustering or assumption of hard boundaries. In this work we learn a multi-verb representation acknowledging these issues.

Action Retrieval.   Distinct from recognition, cross-modal retrieval approaches have been proposed for visual actions both in images [14, 30, 48] and videos [9, 18, 43]. These works focus on instance retrieval, *i.e.* given a caption can the corresponding video/image be retrieved and vice versa. This is different from our attempt to retrieve similar actions rather than only the corresponding video/caption. Only Hahn *et al*. [19] train an embedding space for videos and verbs only, using word2vec as the target space. They use verbs from UCF101 [37] and HMDB51 [23] in addition to verb-noun classes from Kinetics [21]. These are coarser actions (*e.g. "diving"* vs. *"running"*) and as such have little overlap allowing the target space to perform well on unseen actions.

As far as we are aware, ours is the first work to use the same representation for both action recognition and action retrieval.

Video Captioning.   Semantics and annotations have also been studied by the recent surge in generating captions for images [1, 2, 3, 8, 25] and videos [39, 40, 41, 44, 46, 49]. These works assess success by the acceptability of the generated caption. We differ in our focus on learning a mapping function between an input video and verbs in the label set, thus providing a richer understanding of the overlaps between verb labels.

3   Proposed Multi-Verb Representations

In this section, we introduce types of verbs used to describe an action and note their absence from semantic knowledge bases (Sec. 3.1). We then define the single and multi-verb representations (Sec. 3.2). Next, we describe the collection of the multiple verb annotations (Sec. 3.3) and detail our approach to learn the representations (Sec. 3.4).

3.1   Verb Types

From the earlier example of a door being opened, we highlight the different types of verbs that can be used to describe the same action, and their relationships. Firstly, as with standard

multi-label tasks, some verbs are related semantically; e.g. *"pull"* and *"draw"* are synonyms in this context. Secondly, verbs describe different linguistic types of the action: *Result verbs* describe the end state of the action, and *manner verbs* detail the manner in which the action is carried out [16, 20]. Here, *"open"* is the result verb, whilst *"pull"* is a manner verb. This distinction of verb types for visual actions has not been explored before for action recognition/retrieval. Finally, there are verbs that describe the sub-action of using the handle, such as *"grab"*, *"hold"* and *"turn"*. These are also result/manner verbs. We call such verbs *supplementary* as they are used to describe the sub-action(s) that the actor needs to perform in order to complete the overall action. Furthermore, they are highly dependent on context.

While these verb relationships can be explicitly stated, they are not available in lexical databases and are hard to discover from public corpora. We illustrate this for the two commonly used sources of semantic knowledge: WordNet [27] and Word2Vec [26].

Verb Relationships		WordNet	Word2Vec
synonyms	(e.g. *"pull"*-*"draw"*)	✓	✗
result-manner	(e.g. *"open"*-*"pull"*)	✗	✗
supplementary	(e.g. *"open"*-*"hold"*)	✗	✗

WordNet [27].   WordNet is a lexical English Database with each word assigned to one or more synsets, each representing a different meaning of the word. WordNet defines multiple relationships between synsets, including synonyms and hyponyms, but does not capture the other two types of relationships (result/manner, supplementary). Moreover, using WordNet requires prior knowledge of the verb's synset.

Word2Vec [26].   Word2Vec embeds words in a vector space, where cosine distances can be computed. These distances are learned based on co-occurrences of words in a given corpus. For example, using Wikipedia, the embedded vector of the verb *"pull"* has high similarity to the embedded vector of *"push"* even though these actions are antonyms. Co-occurrences depend on the corpus used, and do not cover any of the relationships noted above. (*"pull"* and *"draw"*, for example do not frequently co-occur in Wikipedia). Even using a relevant corpus, such as TACoS [32], *"push"* is closest to *"crush"*, *"rough"* and *"rotate"* as the three most semantically similar verbs.

As these relationships cannot be discovered from semantic knowledge bases, we opt to crowdsource the multi-verb labels (see Sec. 3.3).

3.2   Representations of Visual Actions

Single-Verb Representations.   We start by defining the commonly-used single-label representation of visual actions. Each video $x_i \in X$ has a corresponding label $\mathbf{y}_i \in Y$ where $\mathbf{y}_i$ is a one-hot vector. In **Single-Verb Label (SV)**, this will be over verbs $V = \langle v_1, ..., v_N \rangle$ with v_j representing the jth verb. For example, in Fig. 2 the verbs *"pour"* and *"fill"* are both relevant but only one can be labelled, and more general verbs like *"move"* won't be labelled either. Additionally, we define a **Verb-Noun Label (VN)** representation over $\langle v_j, n_k \rangle$ equivalent to the standard approach where a one-hot action vector is used.

Multi-Verb Representations.   We now propose two multi-verb representations. First, **Multi-Verb Label (MV)**, $\mathbf{y}_i = \langle y_{i,j} \in \{0,1\} \rangle$ which is a binary vector over V (*i.e.* hard assignment). Multiple verbs can be used to describe each video, e.g. *"pour"* and *"fill"*, allowing for manner and result verbs as well as semantically related verbs to be represented. Hard label assignment though can be problematic for supplementary verbs. Consider the verbs

Figure 2: Comparisons of different verb-only representations for *"Pour Oil"* from GTEA+.

"hold", *"grasp"* in Fig. 2. These do not fully describe the object interaction yet cannot be considered irrelevant labels. Including supplementary verbs in a multi-label representation would cause confusion between videos where the whole action is *"hold"*, e.g. *"hold button"*.

Alternatively, one can use a **Soft Assigned Multi-Verb Label (SAMV)**, to increase the size of V while accommodating supplementary verbs. Soft assignment assigns a numerical value to each verb, representing its relevance to the ongoing action. In SAMV, each video x_i will have a label vector $\boldsymbol{y_i} = \langle y_{i,j} \in [0,1] \rangle$ over V. For two verb scores $y_{i,j} > y_{i,k} > 0$ when the verb v_j is more relevant to the action x_i, while v_k is still a valid/relevant verb. This ranking makes this representation suitable for retrieval, and allows the set of verbs, V, to be larger due to the restrictions of the previous representations being removed (Fig. 2).

3.3 Annotation Procedure and Statistics

To acquire the multi-verb representations we now describe the process of collecting multi-verb annotations for three action datasets: BEOID [5] (732 video segments), CMU [38] (404 video segments) and GTEA+ [10] (1001 video segments). All three datasets are shot with a head mounted camera and include kitchen scenes. First, we constructed the list of verbs, V, that annotators could choose from, by combining the unique verbs from all available annotations of the datasets [5, 7, 10, 42], giving a list of 90 verbs - we exclude {rest, read, walk} to focus on object interactions. Table 1 includes the manual split of verbs between manner and result. We found that the datasets included an almost even spread of manner and result verbs with 47 and 43 respectively. We then chose one video per action, and asked 30-50 annotators from Amazon Mechanical Turk (2,939 total) to select all verbs which apply. We normalise the responses by the number of annotators, so that for each verb, the score lies in the range of $[0,1]$.

Figure 3 shows samples of the collected scores, for two videos, one from BEOID (left) and one from GTEA (right). Annotators agree on selecting highly-relevant verbs (high score), and agree on not selecting irrelevant verbs (low scores). These represent both result and manner verb types. Additionally, supplementary verbs align with annotators disagreements: These are not fundamental to describing the action, but some annotators did not consider them irrelevant.

In Fig. 4 we present annotation statistics. Figure 4(a) shows the frequency of annotations

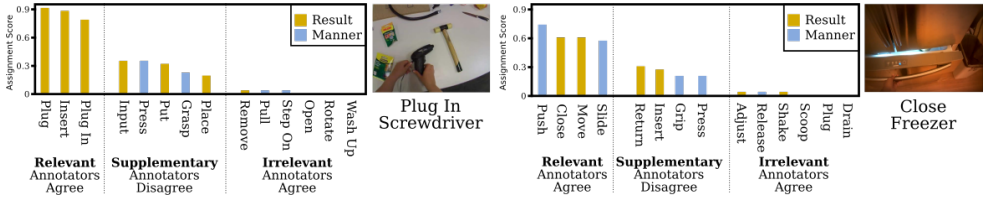| Manner	Result
carry, compress, drain, flip, fumble, grab, grasp, grip, hold, hold down, hold on, kick, let go, lift, pedal, pick up, point, position, pour, press, press down, pull, pull out, pull up, push, put down, release, rinse, reach, rotate, scoop, screw, shake, slide, spoon, spray, squeeze, stir, swipe, swirl, switch, take, tap, tip, touch, twist, turn	activate, adjust, check, clean, close, crack, cut, distribute, divide, drive, dry, examine, fill, fill up, find, input, insert, mix, move, open, peel, place, plug, plug in, put, relax, remove, replace, return, scan, set up, spread, start, step on, switch on, transfer, turn off, turn on, unlock, untangle, wash, wash up, weaken

Table 1: Manual split of the original verbs.

Figure 3: When annotators agree, the verbs are highly relevant or certainly irrelevant. When annotators disagree, verbs are supplementary, for both *result* and *manner* verb types.
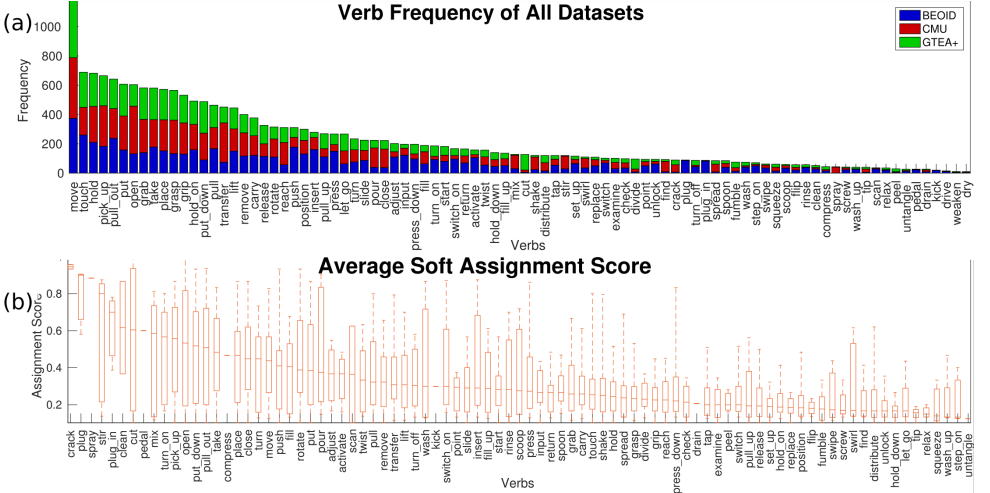

Figure 4: **Annotation Statistics**. Figure best viewed in colour - (a) Number of annotations per verb. (b) Average soft assignment score per verb when chosen.

for each verb in each of the datasets. We observe a long tail distribution with the top-5 most commonly chosen verbs being *"move"*, *"touch"*, *"hold"*, *"pick up"* and *"pull out"*. Figure 4(b) sorts the verbs by their median soft assignment score over videos. We see that *"crack"* has the highest agreement between annotators when present, whereas a generic verb such as *"move"* is more commonly seen as supplementary.

From these annotations, we calculate the representations in Sec. 3.2. For SAMV, the soft score annotations are kept as is. For MV, the scores are binarised with a threshold of 0.5. Finally, SV is assigned by the majority vote.

3.4   Learning Visual Actions

For each of the labelling approaches described in Sec. 3.2, we wish to learn a function, $\phi : \mathcal{W} \mapsto \mathbb{R}^{|V|}$ which maps a video representation $\mathcal{W}$ onto verb labels. For brevity, we define $\hat{\mathbf{y}}_i = \phi(x_i)$, where $\hat{y}_{i,j}$ is the predicted value for verb v_j of video x_i and $y_{i,j}$ is the corresponding ground truth value for verb v_j of video x_i. Typically, single label (SL) representations are learned using a cross entropy loss over the softmax function σ which we use for both SV and VN:

$$L_{SL} = -\sum_i \mathbf{y}_i \log(\sigma(\hat{\mathbf{y}}_i)) \tag{1}$$

For our proposed multi-label (ML) representations, both MV and SAMV, we use the sigmoid binary cross entropy loss as in [28, 51], where S is the sigmoid function:

$$L_{ML} = -\sum_i \sum_j y_{i,j} \log(S(\hat{y}_{i,j})) + (1 - y_{i,j}) \log(1 - S(\hat{y}_{i,j})) \qquad (2)$$

This learns SAMV as a regression problem, without any independence assumptions. We note that due to SAMV having continuous values within the range [0,1], Eq. 2 will be non-zero when $y_{i,j} = \hat{y}_{i,j}$, however the gradient will be zero. We consciously avoid a ranking loss as it only learns a relative order and does not attempt to approximate the representation.

4   Experiments and Results

We present results on the three datasets, annotated in Sec. 3.3, using stratified 5-fold cross validation. We train/test on each dataset separately, but also include qualitative results of cross-dataset retrieval. The results are structured to answer the following questions: How do the labelling representations compare for (i) action recognition, (ii) video-to-text and (iii) text-to-video retrieval tasks? and (iv) Can the labels be used for cross-dataset retrieval?

Implementation. We implemented ϕ as a two stream fusion CNN from [11] that uses VGG-16 networks, pre-trained on UCF101 [57]. The number of neurons in the last layer was set to $|V| = 90$. Each network was trained for 100 epochs. We set a fixed learning rate for spatial, 10^{-3}, and used a variable learning rate for temporal and fusion networks in the range of 10^{-2} to 10^{-6} depending on the training epoch. For the spatial network a dropout ratio of 0.85 was used for the first two fully-connected layers. We fused temporal into the spatial network after ReLU5 using convolution for the spatial fusion as well as both 3D convolution and 3D pooling for the spatio-temporal fusion. We propagated back to the fusion layer. Other hyperparameters are the same as in [11].
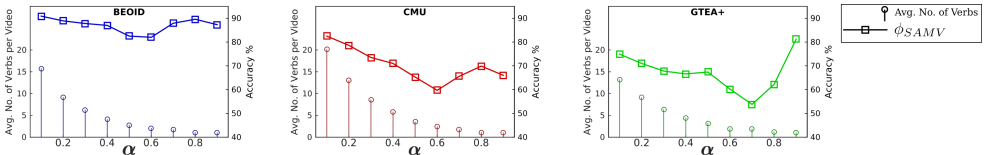
4.1   Action Recognition Results

We first present results on action recognition, comparing the standard verb-noun labels, as well as single verb labels to our proposed multiple verb-only labels.

Evaluation Metric. We compare standard single-label accuracy to multi-label accuracy as follows: Let V_i^L to be the set of ground-truth verbs for video x_i using the labelling representations L, and $\hat{V}_i^L$ the top-k predicted verbs using the same representations, where $k = |V_i^L|$. For single-labels $L \in \{SV, VN\}$, $k = 1$, however for multi-labels, k differs per video, based on the annotations. The accuracy is then evaluated as:

$$A(L) = \frac{1}{|X|} \sum_i \frac{|V_i^L \cap \hat{V}_i^L|}{|V_i^L|} \qquad (3)$$

Where $|X|$ is the total number of videos. $A(\{SV, VN\})$ is the same as reporting top-1 accuracy. For $L \in \{MV, SAMV\}$, given a video with 3 relevant verbs, if the model is able to predict two correct verbs in top-3 predictions then the video will have an accuracy of $0.66 = \frac{2}{3}$. This allows us to compare the recognition accuracy of all models. For SAMV, we threshold the annotations at 0.3 assignment score (denoted by α), and consider these as V_i^{SAMV}.

	BEOID				CMU				GTEA+			
	ϕ_{SV}	ϕ_{VN}	ϕ_{MV}	ϕ_{SAMV}	ϕ_{SV}	ϕ_{VN}	ϕ_{MV}	ϕ_{SAMV}	ϕ_{SV}	ϕ_{VN}	ϕ_{MV}	ϕ_{SAMV}
No. of verbs*	20	40	42	90	12	29	31	89	15	63	34	90
Accuracy	78.1	**93.5**	**93.0**	87.8	59.2	**76.0**	74.1	73.5	59.2	61.2	71.9	**72.9**

Table 2: Action recognition accuracy results (Eq. 3). *for ϕ_{VN}: number of classes in the public dataset.

Figure 5: Action recognition accuracy of ϕ_{SAMV} varying α with avg. no. of verbs per video.

Results.   Table 2 shows the results of the four representations for action recognition. We also show the number of classes in each case. Over all three datasets we see that ϕ_{MV} performs comparatively to ϕ_{VN} on BEOID (-0.5%), worse on CMU (-1.9%) and significantly better on the largest dataset GTEA+ ($+10.5\%$) due to higher overlap in actions during the dataset collection i.e. the same action is applied to multiple objects and vice versa. ϕ_{SV} is consistently worse than all other approaches reinforcing the ambiguity of single verb labels. Additionally, ϕ_{MV} outperforms ϕ_{SAMV} on all datasets but GTEA+, where it is comparable, suggesting it is more applicable for action recognition. We note that the GTEA+ has a higher number of overlapping actions compared to the other two datasets (*i.e.* the number of unique objects per verb is much higher). However, ϕ_{SAMV} is attempting to learn a significantly larger vocabulary.

In Fig. 5, we vary the threshold α on the annotation scores for ϕ_{SAMV}. Note that for $\alpha > 0.5$, some videos in the datasets have $V_i^{SAMV} = \emptyset$ and are not counted for accuracy leading to abrupt changes. The figure shows resilience in recognition suggesting the representation can correctly predict both relevant and supplementary verbs.

4.2   Action Retrieval Results

In this section, we consider the output of ϕ as an embedding space, with each verb representing one dimension. We also evaluate whether the verb labels we have collected are consistent across datasets and provide qualitative cross-dataset retrieval results.

Evaluation Metric.   We use Mean Average Precision (mAP) as defined in [29].

Video-to-Text Retrieval.   First, we evaluate ϕ for video-to-text retrieval. That is, given a video, can the correct verbs be retrieved and in the correct order? Figure 6 compares the representations on each of the labelling methods' ground truth. Results show that ϕ_{SAMV} is the most generalisable, regardless of the ground-truth used, even when not specifically trained for that ground-truth, whereas ϕ_{SV} and ϕ_{MV} only perform well on their respective ground-truth. This shows the ability of ϕ_{SAMV} to learn the correct order of relevance.

Figure 7 compares the three labelling approaches. We also present qualitative retrieval results from ϕ_{SAMV} in Fig. 8, separately highlighting top-3 result and manner verbs. We find that ϕ_{SAMV} learns manner and result verbs equally well with an average root mean square error of 0.094 (for manner verbs) and 0.089 (for result verbs) respectively.
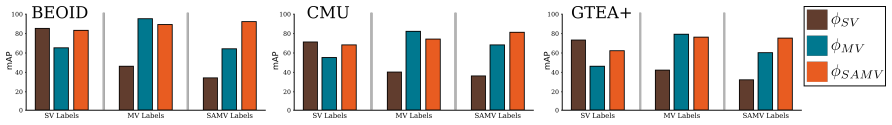
Figure 6: Video-To-Text retrieval results. We show that ϕ_{SAMV} is able to perform well on the single verb labels and multi-verb labels but ϕ_{SV} and ϕ_{MV} underperform on the SAMV labels.

Figure 7: Qualitative results comparing $\phi_{\{SV,MV,SAMV\}}$. Green and red denote correct and incorrect predictions. Orange shows verbs ranked significantly higher than in GT.

Figure 8: Qualitative results of the top 3 retrieved result (yellow) and manner (blue) verbs.

Text-to-Video Retrieval. In the introduction we stated that a combination of multiple individually ambiguous verb labels allows for a more discriminative labelling representation. We test this by querying the verb representations with an increasing number of verbs for text-to-video retrieval, testing all possible combinations of n co-occurring verbs for $n \in [1,5]$. We show in Fig. 9 that mAP increases for ϕ_{SAMV} when the number of verbs rises for both BEOID and GTEA+ significantly outperforming other representations. This suggests that ϕ_{SAMV} is better able to learn the relationships between the different verbs inherent in describing actions. There is a drop in accuracy on CMU due to the coarser grained nature of the videos and thus a much higher overlap between supplementary verbs. The representation still outperforms alternatives at $n = 4$ and $n = 5$.

Cross Dataset Retrieval. We show the benefits of a multi-verb representation across all three datasets using video-to-video retrieval (Fig. 10). We pair each video with its closest
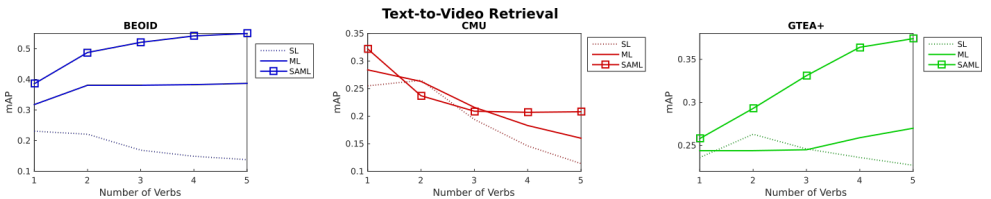

Figure 9: Results of text-to-video retrieval of $\phi_{\{SV,MV,SAMV\}}$ across all three datasets using mAP and a varying number of verbs in the query.
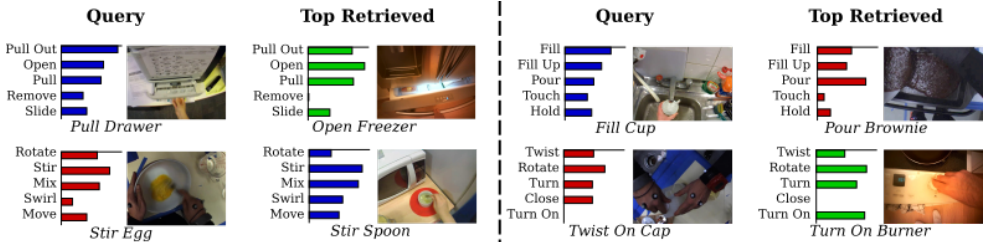
Figure 10: Examples of cross dataset retrieval of videos using either videos or text. Blue: BEOID, Red: CMU and Green: GTEA+.
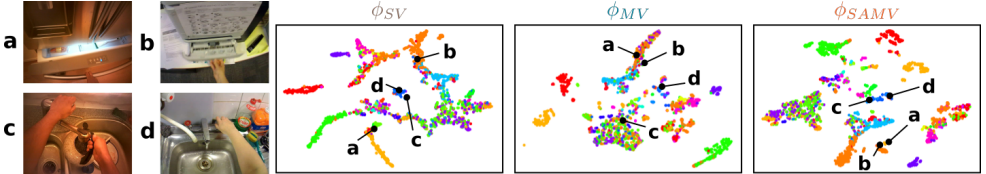

Figure 11: t-SNE representation of $\phi_{\{SV,MV,SAMV\}}$ for all three datasets.

predicted representation from a different dataset. *"Pull Drawer"* and *"Open Freezer"* show examples of the same action with different types of verbs (manner vs. result) on different objects yet their ϕ_{SAMV} representations are very similar. Note the object-agnostic facet of a multi-verb representation with *"Stir Egg"* and *"Stir Spoon"* [see Supp. Video].

In Fig. 11 we show the t-SNE plot of the representations $\phi_{\{SV,MV,SAMV\}}$ of all three datasets. Each dot represents a video coloured by the majority verb. We highlight four videos in the figure. (a) and (b) are separated in ϕ_{SV} due to the majority vote (*"pull"* vs *"open"*) but are close in ϕ_{MV} and ϕ_{SAMV}. Due to the hard assignment, (c) and (d) are far in ϕ_{MV} as the actions require different manners (*"rotate"* vs. *"push"*) but are closer in ϕ_{SAMV}.

5 Conclusion

In this paper we present multi-label verb-only representations for visual actions, for both action recognition and action retrieval. We collect annotations for three action datasets which we use to create the multi-verb representations. We show that such representations embrace class overlaps, dealing with the inherent ambiguity of single-verb labels, and encode both the action's *manner* and its *result*. The experiments, on three datasets, highlight how a multi-verb approach with hard assignment is best suited for recognition tasks, and soft-assignment for retrieval tasks, including cross-dataset retrieval of similar visual actions.

We plan to investigate other uses of these representations for few-shot and zero-shot learning as well as further investigate the relationships between result and manner verb types, including for imitation learning.

Annotations The annotations for all three datasets can be found at
https://github.com/mwray/Multi-Verb-Labels.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[2] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. Convolutional image captioning. In *CVPR*, 2018.

[3] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016.

[4] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015.

[5] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio Mayol-Cuevas. You-do, I-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, 2014.

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018.

[7] Fernando De La Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) database. *Robotics Institute*, 2008.

[8] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. Language models for image captioning: The quirks and what works. In *ACL*, 2015.

[9] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, and Xun Wang. Dual dense encoding for zero-example video retrieval. *CoRR abs/1809.06181*, 2018.

[10] Alireza Fathi, Yin Li, and James M. Rehg. Learning to recognize daily actions using gaze. In *ECCV*, 2012.

[11] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016.

[12] Spandana Gella and Frank Kella. Analysis of action recognition tasks and datasets for still images. In *ACL*, 2017.

[13] Spandana Gella, Mirella Lapata, and Frank Keller. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *NAACL*, 2016.

[14] Albert Gordo and Diane Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *CVPR*, 2017.

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *ICCV*, 2017.

[16] Jess Gropen, Steven Pinker, Michelle Hollander, and Richard Goldberg. Affectedness and direct objects: The role of lexical semantics in the acquisition of verb argument structure. *Cognition*, 1991.

[17] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018.

[18] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013.

[19] Meera Hahn, Andrew Silva, and James M. Rehg. Action2vec: A crossmodal embedding approach to action learning. In *BMVC*, 2018.

[20] Malka Hovav-Rappaport and Beth Levin. Building verb meanings. *The projection of Arguments*, 1998.

[21] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *CoRR abs/1705.05950*, 2017.

[22] Sameh Khamis and Larry S Davis. Walking and talking: A bilinear approach to multi-label action recognition. In *CVPRW*, 2015.

[23] Hilde Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011.

[24] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *ECCV*, 2018.

[25] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *ICLR*, 2015.

[26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR abs/1301.3781*, 2013.

[27] George A. Miller. Wordnet: a lexical database for english. *CACM*, 1995.

[28] Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. Large-scale multi-label text classification-revisiting neural networks. In *ECML PKDD*, 2014.

[29] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[30] Filip Radenović, Giorgos Tolias, and Ondřej Chum. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *ECCV*, 2016.

[31] Piyush Rai, Abhishek Kumar, and Hal Daume. Simultaneously leveraging output and task structures for multiple-output regression. In *NIPS*, 2012.

[32] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *GCPR*, 2014.

[33] Matteo Ruggero Ronchi and Pietro Perona. Describing common human visual actions in images. In *BMVC*, 2015.

[34] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016.

[35] Gunnar A Sigurdsson, Olga Russakovsky, and Abhinav Gupta. What actions are needed for understanding human actions in videos? In *ICCV*, 2017.

[36] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *CVPR*, 2018.

[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *Technical Report CRCV*, 2012.

[38] Ekaterina Taralova, Fernando De La Torre, and Martial Hebert. Source constrained clustering. In *ICCV*, 2011.

[39] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. In *NAACL HLT*, 2015.

[40] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *CVPR*, 2018.

[41] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. Video captioning via hierarchical reinforcement learning. In *CVPR*, 2018.

[42] Michael Wray, Davide Moltisanti, Walterio Mayol-Cuevas, and Dima Damen. Sembed: Semantic embedding of egocentric action videos. In *ECCVW*, 2016.

[43] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*, 2015.

[44] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *ICCV*, 2015.

[45] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *CVPR*, 2016.

[46] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, 2016.

[47] Rowan Zellers and Yejin Choi. Zero-shot activity recognition with verb attribute induction. In *EMNLP*, 2017.

[48] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, 2016.

[49] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, 2018.