Spatial-Aware Object Embeddings for Zero-Shot Localization and Classification of Actions

Pascal Mettes and Cees G. M. Snoek
University of Amsterdam

Abstract

We aim for zero-shot localization and classification of human actions in video. Where traditional approaches rely on global attribute or object classification scores for their zero-shot knowledge transfer, our main contribution is a spatial-aware object embedding. To arrive at spatial awareness, we build our embedding on top of freely available actor and object detectors. Relevance of objects is determined in a word embedding space and further enforced with estimated spatial preferences. Besides local object awareness, we also embed global object awareness into our embedding to maximize actor and object interaction. Finally, we exploit the object positions and sizes in the spatial-aware embedding to demonstrate a new spatio-temporal action retrieval scenario with composite queries. Action localization and classification experiments on four contemporary action video datasets support our proposal. Apart from state-of-the-art results in the zero-shot localization and classification settings, our spatial-aware embedding is even competitive with recent supervised action localization alternatives.

1. Introduction

We strive for the localization and classification of human actions like *Walking a dog* and *Skateboarding* without the need for any video training examples. The common approach in this challenging zero-shot setting is to transfer action knowledge via a semantic embedding build from attributes [23, 30, 57] or objects [2, 19, 55]. As the semantic embeddings are defined by image or video classifiers, they are unable, nor intended, to capture the spatial interactions an actor has with its environment. Hence, it is hard to distinguish who is *Throwing a baseball* and who is *Hitting a baseball* when both actions occur within the same video. We propose a spatial-aware object embedding for localization and classification of human actions in video, see Figure 1.

We draw inspiration from the *supervised* action classification literature, where the spatial connection between actors and objects has been well recognized, *e.g.* [14, 34, 37, 53]. Early work focused on capturing actors and objects implicitly in a low-level descriptor [4, 60], while more recently the benefit of explicitly representing detected objects [8], their scores, and spatial properties was proven effective [16, 49, 61, 62]. Both [7] and [40] demonstrate the benefit of temporal actor and object interaction, by linking detected bounding boxes over time via trackers. By doing so, they are also capable of (supervised) action localization. We also detect actors and objects, and link them over time to capture spatio-temporal interactions. Different from all of the above works, we do no rely on any action class and/or action video supervision to get to our recognition. Instead, we introduce an embedding built upon actor and object detectors that allows for zero-shot action classification and localization in video.

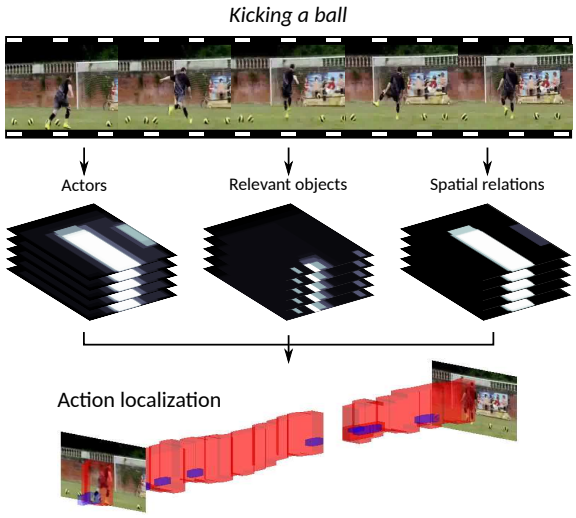


Figure 1: **Spatial-aware object embedding.** Actions are localized and classified by information about actors, relevant objects, and their spatial relations.

Our main contribution is a spatial-aware object embedding for zero-shot action localization and classification. The spatial-aware embedding incorporates word embeddings,

box locations for actors and objects, as well as their spatial relations, to generate action tubes. This enables us to both classify videos and to precisely localize where actions occur. Our spatial-aware embedding is naturally extended with contextual awareness from global objects. We furthermore show how our embedding generalizes to any query involving objects, spatial relations, and their sizes in a new spatio-temporal action retrieval scenario. Action localization and classification experiments on four contemporary action video datasets support our proposal.

2. Related work

2.1. Supervised action localization and classification

A wide range of works have proposed representations to classify actions given video examples. Such representations include local spatio-temporal interest points and features [26, 31, 52] and local trajectories [1, 50], typically aggregated into VLAD or Fisher vector representations [38, 39]. Recent works focus on learning global representations from deep networks, pre-trained on optical flow [43] or large-scale object annotations [20, 22, 59]. We also rely on deep representations for our global objects, but we emphasize on local objects and we aim to classify and localize actions without the need for any video example.

For spatio-temporal action localization, a popular approach is to split videos into action proposals; spatio-temporal tubes in videos likely to contain an action. Annotated tubes from example videos are required to train a model to select the best action proposals at test time. Action proposal methods include merging supervoxels [18, 44], merging trajectories [3, 35], and detecting actors [63]. The current state-of-the-art action localizers employ Faster R-CNN [41] trained on bounding box annotations of actions in video frames [13, 51]. We are inspired by the effectiveness of actor detections and Faster R-CNN for localization, but we prefer commonly available detectors trained on images. We employ these detectors as input to our spatial-aware embedding for localization in video in a zero-shot setting.

2.2. Zero-shot action localization and classification

Inspired by zero-shot image classification [24], several works have performed zero-shot action classification by learning a mapping of actions to attributes [11, 30, 64]. Models are trained for the attributes from training videos of other actions and used to compare test videos to unseen actions. Attribute-based classification has been extended *e.g.* using transductive learning [9, 58] and domain adaptation [23, 57]. Due to the necessity to manually map each action to global attributes a priori, these approaches do not generalize to arbitrary zero-shot queries and are unable to localize actions, which is why we do not employ attributes in our work.

Rather than mapping actions to attributes, test actions can also be mapped directly to actions used for training. Li *et al.* [27] map visual video features to a semantic space shared by training and test actions. Gan *et al.* [12] train a classifier for an unseen action by relating the action to training actions at several levels of relatedness. Although the need for attributes is relieved with such mappings, this approach still requires videos of other actions for training and is only able to classify actions. We localize and classify actions without using any videos of actions during training.

A number of works have proposed zero-shot classification by exploiting large amounts of image and object labels [6]. Given deep networks trained on image data, these approaches map object scores in videos to actions *e.g.* using word vectors [2, 17, 19, 55] or auxiliary textual descriptions [10, 15, 54]. Objects as the basis for actions results in effective zero-shot classification and generalizes to arbitrary actions. However, these approaches are holistic; object scores are computed over whole videos. In this work, we take the object-based perspective to a local level, which allows us to model the spatial interaction between actors and objects for action localization, classification, and retrieval.

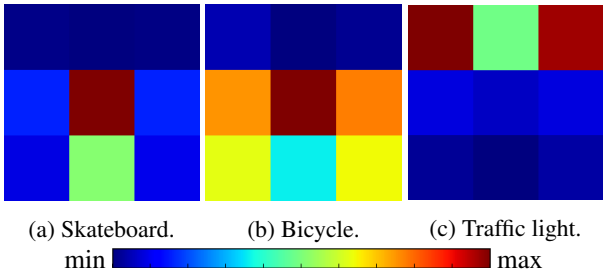
The work of Jain *et al.* [19] has previously performed zero-shot action localization. Their approach first generates action proposals. In a second pass, each proposal is represented with object classification scores. The proposals best matching the action name in word2vec space are selected. Their approach does not use any object detectors, nor is there any explicit notion of spatial-awarenes inside each action proposal. Finally, spatial relations between actors and objects are ignored. As we will show in the experiments, inclusion of our spatial-awareness solves these limitations and leads to a better zero-shot action localization.

3. Spatial-aware object embeddings

In our zero-shot formulation, we are given a set of test videos $\mathcal{V}$ and a set of action class names $\mathcal{Z}$. We aim to classify each video to its correct class and to discover the spatio-temporal tubes encapsulating each action in all videos. To that end, we propose a spatial-aware embedding; scored action tubes from interactions between actors and local objects. We present our embeddings in three steps: (i) gathering prior knowledge on actions, actors, objects, and their interactions, (ii) computing spatial-aware embedding scores for bounding boxes, and (iii) linking boxes into action tubes.

3.1. Prior knowledge

Local object detectors. We first gather a set of local detectors pre-trained on images. Let $\mathcal{O} = \{O_D, O_N\}$ denote the objects with detectors O_D and names O_N. Furthermore, let $\mathcal{A} = \{A_D, \texttt{actor}\}$ denote the actor detector. Each detector outputs a set of bounding boxes with corresponding object probability scores per video frame.



(a) Skateboard. (b) Bicycle. (c) Traffic light.

Figure 2: **Examples of preferred spatial relations of objects relative to actors.** In line with our intuition, skateboards are typically on or below the actor, while bicycles are typically to the left or right of actors and traffic lights are above the actors.

Textual embedding. Given an action class name $Z \in \mathcal{Z}$, we aim to select a sparse subset of objects $\mathcal{O}_Z \subset \mathcal{O}$ relevant for the action. For the selection, we rely on semantic textual representations as provided by word2vec [36]. The similarity between object o and the action class name is given as:

$$w(o, Z) = \cos(e(o_N), e(Z)), \tag{1}$$

where $e(\cdot)$ states the word2vec representation of the name. We select the objects with maximum similarity to the action.

Actor-object relations. We exploit that actors interact with objects in preferred spatial relations. To do so, we explore where objects tend to occur relative to the actor. Since we can not learn precise spatial relations between actors and objects from examples, we aim to use common spatial relations between actors and objects, as can be mined from large-scale image data sets. We discretize the spatial relations into nine relative positions, representing the preposition *in front of* and the eight basic prepositions around the actor, *i.e. left of, right of, above, below*, and the four corners (*e.g. above left*). For each object, we obtain a nine-dimensional distribution specifying its expected location relative to the actor, as detailed in Figure 2.

3.2. Scoring actor boxes with object interaction

We exploit our sources of prior knowledge to compute a score for the detected bounding boxes in all frames of each test video $V \in \mathcal{V}$. Given a bounding box b in frame F of video V, we define a score function that incorporates the presence of (i) actors, (ii) relevant local objects, and (iii) the preferred spatial relation between actors and objects. A visual overview of the three components is shown in Figure 3. More formally, we define a score function for box b given an action class Z as:

$$s(b, F, Z) = p(A_D|b) + \sum_{o \in \mathcal{O}_Z} r(o, b, F, Z), \tag{2}$$

where $p(A_D|b)$ is the probability of an actor being present in bounding box b as specified by the detector A_D. The function r expresses the object presence and relation to the actor, it is defined as:

$$r(o, b, F, Z) = w(o, Z) \cdot \left(\max_{f \in F_n} p(o_D|f) \cdot m(o, \mathcal{A}, b, f) \right), \tag{3}$$

where $w(o, Z)$ states the semantic relation score between object o and action Z and F_n states all bounding boxes within the neighourhood of box b in frame F. The second part of Equation 3 states that we are looking for a box f around b that maximizes the joint probability of the presence of object o (the function $p(o_D|f)$), the match between the spatial relations of (b, f) and the prior relations of the actor and object o (the function m). We define the spatial relation match as:

$$m(o, \mathcal{A}, b, f) = 1 - JSD_2(d(\mathcal{A}, o)||d(b, f)), \tag{4}$$

where $JSD_2(\cdot||\cdot) \in [0, 1]$ denotes the Jensen-Shannon Divergence with base 2 logarithm [28]. Intuitively, the Jensen-Shannon Divergence, a symmetrized and bounded variant of the Kullback-Leibler divergence, determines to what extent the two 9-dimensional distributions match. The more similar the distributions, the lower the divergence, hence the need for the inversion as we aim for maximization.

3.3. Linking spatial-aware boxes

The score function of Equation 2 provides a spatial-aware embedding score for each bounding box in each frame of a video. We apply the score function to the boxes of all actor detections in each frame. We form tubes from the individual box scores by linking them over time [13]. We link those boxes over time that by themselves have a high score from our spatial-aware embedding and have a high overlap amongst each other. This maximization problem is solved using dynamic programming with the Viterbi algorithm. Once we have a tube from the optimization, we remove all boxes from that tube and compute the next tube from the remaining boxes.

Let T denote a discovered action tube in a video. The corresponding score is given as:

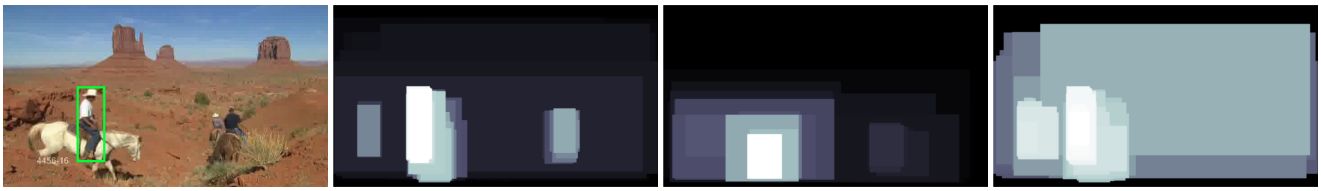
$$t_{\text{emb}}(T, Z) = \frac{1}{|T|} \sum_{t \in T} s(t_b, t_F, Z), \tag{5}$$

where t_b and t_F denote a bounding box and the corresponding frame in tube T.

In summary, we propose spatial-aware object embeddings for actions; tubes through videos by linking boxes based on the zero-shot likelihood from the presence of actors, the presence of relevant objects around the actors, and the expected spatial relations between objects and actors.

4. Local and global object interaction

To distinguish tubes from different videos in a collection, contextual awareness in the form of relevant global object



(a) Video frame. (b) Actor detection. (c) Object detection (horse). (d) Spatial relation match.

Figure 3: **Example of our spatial-aware embedding.** The actor sitting on the left horse (green box) is most relevant for the action *Riding horse* based on the actor detection, horse detection, and spatial relations between actors and horses.

classifiers is also a viable source of information. Here, we first outline how to obtain video-level scores based on object classifiers. Then, we show how to compute spatial- and global-aware embeddings for action localization, classification, and retrieval.

4.1. Scoring videos with global objects

Let $\mathcal{G} = \{G_C, G_N\}$ denote the set of global objects with corresponding classifiers and names. Different from the local objects $\mathcal{O}$, these objects provide classifier scores over a whole video. Given an action class name Z, we again select the top relevant objects $\mathcal{G}_Z \subset \mathcal{G}$ using the textual embedding. The score of a video V is then computed as a linear combination of the word2vec similarity and classifier probabilities over the top relevant objects:

$$t_{\text{global}}(V, Z) = \sum_{g \in G_Z} w(g, Z) \cdot p(g|V), \tag{6}$$

where $p(g|V)$ denotes the probability of global object g of being in video V.

4.2. Spatial- and global-aware embedding

The information from local and global objects is combined into a spatial- and global-aware embedding. Here, we show how this embedding is employed for spatial-aware action localization, classification, and retrieval.

Action localization. For localization, we combine the tube score from our spatial-aware embedding with the video score from the global objects into a score for each individual tube T as:

$$t(T, V, Z) = t_{\text{emb}}(T, Z) + t_{\text{global}}(V, Z). \tag{7}$$

We note that incorporating scores from global objects does not distinguish tubes from the same video. The global scores are however discriminative for distinguishing tubes from different videos in a collection $\mathcal{V}$. We compute the final score for all tubes of all videos in $\mathcal{V}$ using Equation 7. We then select the top scoring tubes per video, and rank the tubes over all videos based on their scores for localization.

Action classification. For classification purposes, we are no longer concerned about the precise location of the tubes from the spatial-aware embeddings. Therefore, we compute the score of a video V given an action class name Z using a max-pooling operation over the scores from all tubes T_V in the video. The max-pooled score is then combined with the video score from the global objects. The predicted class for video V is determined as the class with the highest combined score:

$$c^*_V = \arg\max_{Z \in \mathcal{Z}} \left(\max_{T \in T_V} t_{\text{emb}}(T, Z) + t_{\text{global}}(V, Z) \right). \tag{8}$$

Spatial-aware action retrieval. Spatial-aware action retrieval from user queries resembles action localization, *i.e.* rank the most relevant tubes the highest. However, different from localization, we now have the opportunity to specify actor and object relations via the search query. Given the effectiveness of size in actor-object interactions [7], we can also allow users to specify a relative object size r. By altering the size of queries objects, different localizations can be retrieved of the same action. To facilitate spatial-aware action retrieval, we alter the spatial relation match of Equation 4 with a match for a specified relative object size:

$$q(o, \mathcal{A}, b, f, r) = m(o, \mathcal{A}, b, f) + \left(1 - |\frac{s(b)}{s(f)} - r|\right), \tag{9}$$

where $s(\cdot)$ denotes the size of a bounding box. Substituting the spatial relation match with Equation 9, we again rank top scoring tubes, but now by maximizing a match to user-specified objects, spatial relations, and relative size.

5. Experimental setup

5.1. Datasets

UCF Sports consists of 150 videos from 10 sport action categories, such as *Skateboarding*, *Horse riding*, and *Walking* [42]. We employ the test split as suggested in [25].

UCF 101 consists of 13,320 videos from 101 action categories, such as *Skiing*, *Basketball dunk*, and *Surfing* [45]. We use this dataset for classification and use the test splits as provided in [45], unless stated otherwise.

J-HMDB consists of 928 videos from 21 actions, such as *Sitting*, *Laughing*, and *Dribbling* [21]. We use the

	Localization (mAP @ 0.5)				**Classification** (mean accuracy)			
	# local objects				# local objects			
	0	1	2	5	0	1	2	5
Embedding I: ***Actor-only***	0.083	-	-	-	0.100	-	-	-
Embedding II: ***Actors and objects***	-	0.175	0.182	0.193	-	0.205	0.117	0.139
Embedding III: ***Spatial-aware***	-	**0.221**	0.209	0.199	-	0.180	0.196	**0.255**

Table 1: **Influence of spatial awareness.** On UCF Sports we compare our spatial-aware object embedding to two other embeddings; using only the actors and using actors with objects, while ignoring their spatial relations. Our spatial-aware embedding is preferred for both localization (one object per action) and classification (five objects per action).

bounding box around the binary action masks as the spatio-temporal annotations for localization. We use the test split as suggested in [21].

Hollywood2Tubes consists of 1,707 videos from the Hollywood2 dataset [32], supplemented with spatio-temporal annotations for localization [35]. Actions include *Fighting with a person*, *Eating*, and *Getting out of a car*. We use the test split as suggested in [32].

5.2. Implementation details

Textual embedding. To map the semantics of actions to objects, we employ the skip-gram network of word2vec [36] trained on the metadata of the images and videos from the YFCC100M dataset [47]. This model outputs a 500-dimensional representation for each word. If an action or object consists of multiple words, we average the representations of the individual words [19].

Actor and object detection. For the detection of both the actors and the local objects, we use Faster R-CNN [41], pre-trained on the MS-COCO dataset [29]. This network consists of the actor class and 79 other objects, such as *snowboard*, *horse*, and *toaster*. After non-maximum suppression, we obtain roughly 50 detections for each object per frame. We apply the network to each frame (UCF Sports, J-HMDB), or each 5^{th} frame (UCF 101, Hollywood2Tubes) followed by linear interpolation.

Spatial relations. The spatial relations between actors and objects are also estimated from the MS-COCO dataset. For each object instance, we examine the spatial relations with the closest actor (if any actor is close to the object). We average the relations over all instances for each object.

Object classification. For the global objects, we employ a GoogLeNet network [46], pre-trained on a 12,988-category shuffle [33] of ImageNet [6]. This network is applied to each 5^{th} frame of each video. For each frame, we obtain the object probabilities at the softmax layer and average the probabilities over the entire video. Following [19], we select the top 100 most relevant objects per action.

Evaluation. For localization, we compute the spatio-temporal intersection-over-union between top ranked actor tubes and ground truth tubes. We report results using both the (mean) Average Precision and AUC metrics. For classification, we evaluate with mean class accuracy.

6. Experimental results

6.1. Spatial-aware embedding properties

In the first experiment, we focus on the properties of our spatial-aware embedding, namely the number of local objects to select and the influence of the spatial relations. We also evaluate qualitatively the effect of selecting relevant objects per action. We evaluate these properties on the UCF Sports dataset for both localization and classification.

Influence of local objects. We evaluate the performance using three settings of our embeddings. The first setting is using solely the actor detections for scoring bounding boxes. The second setting uses both the actor and the top relevant objects(s), but ignores the spatial relations between actors and objects. The third setting is our spatial-aware embedding, which combines the information from actors, objects, and their spatial relations.

In Table 1, we provide both the localization and classification results. For localization using only the actor results in tubes that might overlap well with the action of interest, but there is no direct means to separate tubes containing different actions. This results in low Average Precision scores. For classification, using only the actor results in weak accuracy scores. This is because there is again no mechanism to discriminate videos containing different actions.

The second row of Table 1 shows the result when incorporating local object detections. For both localization and classification, there is a considerable increase in performance, indicating the importance of detections of relevant objects for zero-shot action localization and classification.

In the third row of Table 1, we show the performance of our spatial-aware embedding. The embedding outperforms the other settings for both localization and classification. This result shows that gathering and capturing information about the relative spatial locations of objects and actors provides valuable information about actions in videos.

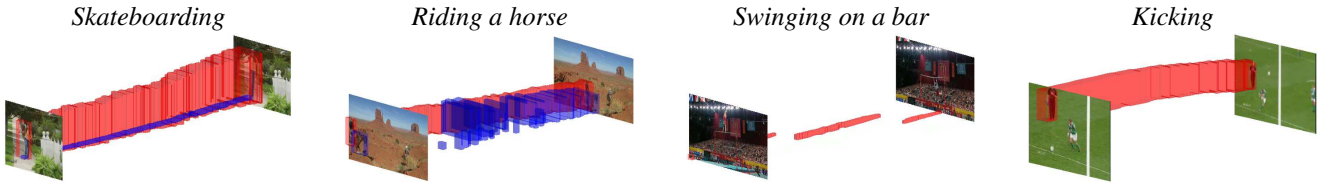


Figure 4: **Qualitative action localization results.** For *Skateboarding* and *Riding a horse*, relevant objects (blue) aid our localization (red). For *Swinging on a bar* and *Kicking*, incorrectly selected objects result in incorrect localizations. We expect that including more object detectors into our embedding will further improve results.

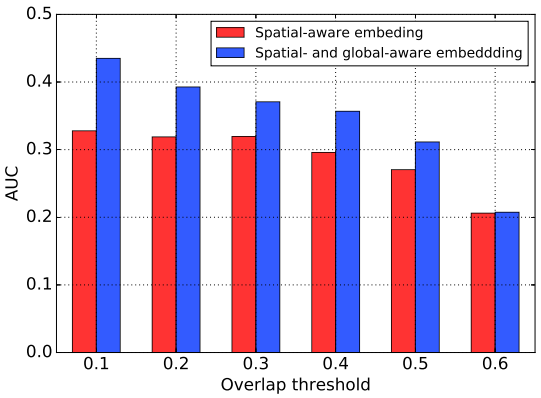


Figure 5: **Local and global object interaction effect on localization.** Adding global object awarereness further improves our spatial-aware object embedding on UCF Sports, especially at low overlap thresholds.

The spatial-aware embedding is most beneficial for the action *Riding a horse* (from 0.03 to 0.75 mAP), due to the consistent co-occurrence of actors and horses. Contrarily, the performance for *Running* remains unaltered, which is because no object relevant to the action is amongst the available detectors.

We have additionally performed an experiment with finer grid sizes on UCF Sports. For localization with the top-5 objects, we reach an mAP of 0.170 (4x4 grid) and 0.171 (5x5 grid), compared to a score of 0.199 with the 3x3 grid. Overall, the scores descrease slightly with finer grid sizes, indicating that coarse spatial relations are preferred over fine spatial relations.

How many local objects? In Table 1 we also consider how many relevant local objects to maintain per action. For localization, we observe a peak in performance using the top-1 local object per action, with a mean Average Precision (mAP) of 0.221 at an overlap threshold of 0.5; a sharp increase in performance over the 0.083 mAP using only the actor. When more objects are used, the performance of our embeddings degrades slightly, indicating that actors are more likely to interact with a single object than multiple objects on a local level. At least for the UCF Sports dataset.

For classification, we observe a reverse correlation; the more local objects in our embedding, the higher the classification accuracy. This result indicates that for classification, we want to aggregate more information about object presence in videos, rather than exploit the single most relevant object per action. This is because a precise overlap with the action in each video is no longer required for classification. We exploit this relaxation with the max-pooling operation in the video-level scoring of Equation 8.

Selecting relevant objects. In our zero-shot formulation, a correct action recognition depends on detecting objects relevant to the action. We highlight the effect of detecting relevant objects in Figure 4. For successful actions such as *Skateboarding* and *Riding a horse*, the detection of respectively skateboards and horses help to generate a desirable action localization. For the actions *Swinging on a bar* and *Kicking*, the top selected objects are however incorrect, either because no relevant object is available or because of ambiguity in the word2vec representations.

Conclusions. We conclude from this experiment that our spatial-aware embedding is preferred over only using the actor and using actors and objects without spatial relations. Throughout the rest of the experiments, we will employ the spatial-aware embedding, using the top-1 object for localization and the top-5 for classification.

6.2. Local and global object interaction

In the second experiment, we focus on the localization and classification performance when incorporating contextual awareness from global object scores into the spatial-aware embedding. We perform the evaluation on the UCF Sports dataset.

Effect on localization. In Figure 5, we show the AUC scores across several overlap thresholds. We show the results using our spatial-aware embedding and the combined spatial- and global-aware embedding.

We observe that across all overlap thresholds, adding global object classifier scores to our spatial-aware embedding improves the localization performance. This result in-

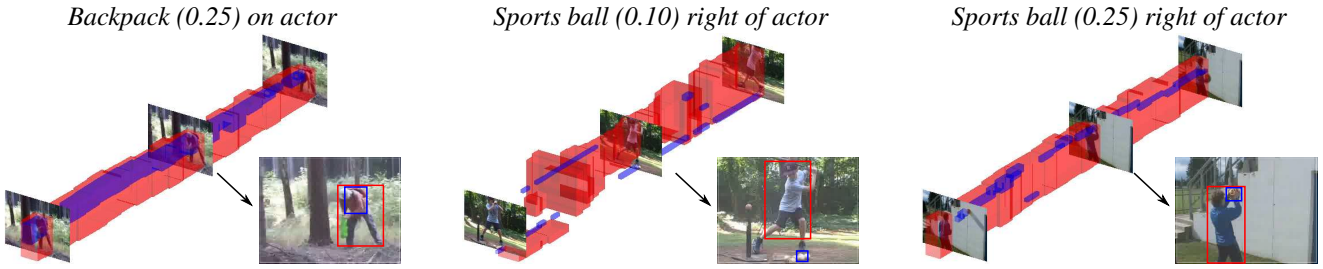


Figure 6: **Spatial-aware action retrieval.** Top retrieved results on J-HMDB given specified queries. Our retrieved localizations (red) reflect the prescribed object (blue), spatial relation, and object size.

	Accuracy
Random	0.100
Jain *et al*. [19]	0.264
Spatial-aware embedding	0.255
Spatial- and global-aware embedding	0.645

Table 2: **Local and global object interaction for classification.** Adding global object awarereness improves our spatial-aware embedding considerably on UCF Sports.

dicates the importance of global object classification scores for discriminating tubes from different videos. The increase in performance is most notable at lower overlap thresholds, which we attribute to the fact that no localization information is provided by the global objects. The higher the overlap threshold, the more important selecting the right tube in each video becomes, and consequently, the less important the global object scores become.

Effect on classification. In Table 2, we show the classification accuracies on the UCF Sports dataset. We first observe that our spatial-aware embedding yields results competitive to the global object approach of Jain *et al*. [19], who also report zero-shot classification on UCF Sports. We also observe a big leap in performance when using our spatial- and global-aware embedding, with an accuracy of 0.645. We note that the big improvement is partially due to our deep network for the global object classifiers, namely a GoogleNet trained on 13k objects [33]. We have therefore also performed an experiment with our spatial- and global-aware embedding using the network of [19]. We achieved a classification accuracy of 0.374, still a considerable improvement over the accuracy of 0.264 reported in [19].

Conclusion. We conclude from this experiment that including global object classification scores into our spatial-aware embedding improves both the zero-shot localization and classification performance. We will use this embedding for our comparison to related zero-shot action works.

6.3. Spatial-aware action retrieval

For the third experiment, we show qualitatively that our spatial-aware embedding is not restricted to specific action queries and spatial relations. We show that any object, any spatial relation, and any object size can be specified as a query for spatial-aware action retrieval. For this experiment, we rely on the test videos from J-HMDB. In Figure 6, we show three example queries and their top retrieved actions.

The example on the left shows how we can search for a specific combination of actor, object, and spatial relation. The examples in the middle and right show that specifying different sizes for the query object leads to a different retrieval. The examples show an interaction with a baseball (middle) and a soccer ball (right), which matches with the desired object sizes in the queries.

We conclude from this experiment that our embedding can provide spatio-temporal action retrieval results for arbitrarily specified objects, spatial relations, and object sizes.

6.4. Comparison to state-of-the-art

For the fourth experiment, we perform a comparative evaluation of our approach to the state-of-the-art in zero-shot action classification and localization. For localization, we also compare our results to supervised approaches, to highlight the effectiveness of our approach.

Action classification. In Table 3, we provide the zero-shot classification results on the UCF-101 dataset, which provides the most comparisons to related zero-shot approaches. Many different data splits and evaluation setups have been proposed, making a direct comparison difficult. We have therefore applied our approach to the three most common types of zero-shot setups, namely using the standard supervised test splits, using 50 randomly selected actions for testing, and using 20 actions randomly for testing.

In Table 3, we first compare our approach to Jain *et al*. [19], who like us do not require training videos. With an accuracy of 0.328 we ouperform their approach (0.303). We also compare to approaches that require training videos for their zero-shot transfer, using author suggested splits. For the (random) 51/50 splits for training and testing, we

	Train	Test	Splits	Accuracy
Jain *et al*. [19]	–	101	3	0.303 ± 0.00
Ours	–	101	3	**0.328** ± 0.00
Kodirov *et al*. [23]	51	50	10	0.140 ± 0.02
Liu *et al*. [30]	51	50	5	0.149 ± 0.01
Xu *et al*. [56]	51	50	30	0.186 ± 0.02
Xu *et al*. [58]	51	50	50	0.222 ± 0.03
Xu *et al*. [57]	51	50	50	0.229 ± 0.03
Li *et al*. [27]	51	50	30	0.268 ± 0.04
Ours	–	50	10	**0.404** ± 0.01
Kodirov *et al*. [23]	81	20	10	0.225 ± 0.04
Gan *et al*. [11]	81	20	10	0.311 ± 0.01
Ours	–	20	10	**0.512** ± 0.05

Table 3: **Comparison to state-of-the-art** for zero-shot action classification on UCF101. For all protocols and test splits we outperform the state-of-the-art, even without us needing any training videos for action transfer.

obtain an accuracy of 0.404. Outperform the next best zero-shot approach (0.268) considerably. We like to stress that all other approaches in this regime use the videos from the training split to guide their zero-shot transfer, while we ignore these videos. When using 20 actions for testing, the difference to other zero-shot approaches increases from 0.255 [23] and 0.311 [11] to 0.512. The lower the number of actions compared to the number of objects in our embedding, the more beneficial for our approach.

Action localization. In Table 4, we provide the localization results on the UCF Sports, Hollywood2Tubes, and J-HMDB datasets. We first compare our result to Jain *et al*. [19] on UCF Sports in Table 4a, which is the only zero-shot action localization work in the literature we are aware of. Across all overlap thresholds, we clearly outperform their approach. At the challenging overlap threshold of 0.5, we obtain an AUC score of 0.311, compared to 0.071 for Jain *et al*. [19]; a considerable improvement.

Given the lack of comparison for zero-shot localization, we also compare our approach to several *supervised* localization approaches on UCF Sports (Table 4a) and Hollywood2Tubes (Table 4b). We observe that we can achieve results competitive to supervised approaches [5, 18, 48], especially at high overlaps. Naturally, the state-of-the-art *supervised* approach [13] performs better, but requires thousands of hard to obtain video tube annotations for training. Our achieved performance indicates the effectiveness of our approach, even though no training examples of action videos or bounding boxes are required. Finally, to highlight our performance across multiple datasets, we provide the first zero-shot localization results on J-HMDB in Table 4c.

Conclusion. For classification, we outperform other zero-shot approaches across all common evaluation setups.

	UCF Sports				
	0.1	0.2	0.3	0.4	0.5
Supervised					
Gkioxari *et al*. [13]	0.560	0.560	0.560	0.520	0.495
Jain *et al*. [18]	0.550	0.525	0.490	0.370	0.270
Tian *et al*. [48]	0.455	0.425	0.315	0.265	0.240
Cinbis *et al*. [5]	0.292	0.169	0.128	0.102	0.049
Zero-shot					
Jain *et al*. [19]	0.288	0.232	0.162	0.099	0.072
Ours	**0.435**	**0.393**	**0.371**	**0.357**	**0.311**

(a)

	Hollywood2Tubes				
	0.1	0.2	0.3	0.4	0.5
Supervised					
Mettes *et al*. [35]	0.345	0.240	0.154	0.092	0.048
Cinbis *et al*. [5]	0.121	0.051	0.020	0.007	0.001
Zero-shot					
Ours	0.210	0.138	0.086	0.047	0.020

(b)

	J-HMDB				
	0.1	0.2	0.3	0.4	0.5
Zero-shot					
Ours	0.346	0.333	0.305	0.268	0.230

(c)

Table 4: **Comparison to state-of-the-art** for zero-shot action localization on (a) UCF Sports, (b) Hollywood2Tubes, and (c) J-HMDB. The only other zero-shot action localization approach is [19], which we outperform considerably. We also compare with several *supervised* alternatives. We are competitive, especially at high overlaps thresholds.

For localization, we outperform the zero-shot localization of [19], while even being competitive to several supervised action localization alternatives.

7. Conclusions

We introduce a spatial-aware embedding for localizing and classifying actions without using any action video during training. The embedding captures information from actors, relevant local objects, and their spatial relations. The embedding further profits from contextual awareness by global objects. Experiments show the benefit of our embeddings, resulting in state-of-the-art zero-shot action localization and classification. Finally, we demonstrate our embedding in a new spatio-temporal action retrieval scenario with queries containing object positions and sizes.

Acknowledgements

This research is supported by the STW STORY project.

References

[1] B. B. Amor, J. Su, and A. Srivastava. Action recognition using rate-invariant analysis of skeletal shape trajectories. *TPAMI*, 38(1):1–13, 2016. 2

[2] S. Cappallo, T. Mensink, and C. Snoek. Video stream retrieval of unseen queries using semantic memory. *BMVC*, 2016. 1, 2

[3] W. Chen and J. J. Corso. Action detection by implicit intentional motion clustering. In *ICCV*, 2015. 2

[4] W. Choi, K. Shahid, and S. Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *ICCV wshop*, 2011. 1

[5] R. G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, 2014. 8

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 5

[7] V. Escorcia and J. C. Niebles. Spatio-temporal human-object interactions for action recognition in videos. In *ICCV wshop*, 2013. 1, 4

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 2010. 1

[9] Y. Fu, T. Hospedales, T. Xiang, Z. Fu, and S. Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *ECCV*, 2014. 2

[10] C. Gan, M. Lin, Y. Yang, G. de Melo, and A. G. Hauptmann. Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition. In *AAAI*, 2016. 2

[11] C. Gan, T. Yang, and B. Gong. Learning attributes equals multi-source domain generalization. In *CVPR*, 2016. 2, 8

[12] C. Gan, Y. Yang, L. Zhu, D. Zhao, and Y. Zhuang. Recognizing an action using its name: A knowledge-based approach. *IJCV*, 2016. 2

[13] G. Gkioxari and J. Malik. Finding action tubes. In *CVPR*, 2015. 2, 3, 8

[14] A. Gupta and L. S. Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, 2007. 1

[15] A. Habibian, T. Mensink, and C. Snoek. Video2vec embeddings recognize events when examples are scarce. *TPAMI*, 2017. 2

[16] D. Han, L. Bo, and C. Sminchisescu. Selection and context for action recognition. In *ICCV*, 2009. 1

[17] N. Inoue and K. Shinoda. Adaptation of word vectors using tree structure for visual semantics. In *MM*, 2016. 2

[18] M. Jain, J. an Gemert, H. Jégou, P. Bouthemy, and C. Snoek. Action localization with tubelets from motion. In *CVPR*, 2014. 2, 8

[19] M. Jain, J. van Gemert, T. Mensink, and C. Snoek. Objects2action: Classifying and localizing actions without any video example. In *ICCV*, 2015. 1, 2, 5, 7, 8

[20] M. Jain, J. C. van Gemert, and C. Snoek. What do 15,000 object categories tell us about classifying and localizing actions? In *CVPR*, 2015. 2

[21] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, 2013. 4, 5

[22] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 2

[23] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, 2015. 1, 2, 8

[24] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 2014. 2

[25] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *ICCV*, 2011. 4

[26] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 2

[27] Y. Li, S.-H. Hu, and B. Li. Recognizing unseen actions in a domain-adapted embedding space. In *ICIP*, 2016. 2, 8

[28] J. Lin. Divergence measures based on the shannon entropy. *Trans. on Inf. Theory*, 1991. 3

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[30] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011. 1, 2, 8

[31] L. Liu, L. Shao, X. Li, and K. Lu. Learning spatio-temporal representations for action recognition: A genetic programming approach. *Trans. Cybernetics*, 46(1):158–170, 2016. 2

[32] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 5

[33] P. Mettes, D. C. Koelma, and C. Snoek. The imagenet shuffle: Reorganized pre-training for video event detection. In *ICMR*, 2016. 5, 7

[34] P. Mettes, C. Snoek, and S.-F. Chang. Localizing actions from video labels and pseudo-annotations. In *BMVC*, 2017. 1

[35] P. Mettes, J. van Gemert, and C. Snoek. Spot on: Action localization from pointly-supervised proposals. In *ECCV*, 2016. 2, 5, 8

[36] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 3, 5

[37] D. J. Moore, I. A. Essay, and M. H. Hayes III. Exploiting human actions and object context for recognition tasks. In *ICCV*, 1999. 1

[38] D. Oneata, J. Verbeek, and C. Schmid. Action and event recognition with fisher vectors on a compact feature set. In *ICCV*, 2013. 2

[39] X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked fisher vectors. In *ECCV*, 2014. 2

[40] A. Prest, V. Ferrari, and C. Schmid. Explicit modeling of human-object interactions in realistic videos. *TPAMI*, 2013. 1

[41] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2, 5

[42] M. D. Rodriguez, J. Ahmed, and M. Shah. Action MACH: a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 4

[43] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 2

[44] K. Soomro, H. Idrees, and M. Shah. Action localization in videos through context walk. In *ICCV*, 2015. 2

[45] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012. 4

[46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 5

[47] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *CACM*, 2016. 5

[48] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *CVPR*, 2013. 8

[49] M. M. Ullah, S. N. Parizi, and I. Laptev. Improving bag-of-features action recognition with non-local cues. In *BMVC*, 2010. 1

[50] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. 2

[51] P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Learning to track for spatio-temporal action localization. In *ICCV*, 2015. 2

[52] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, 2008. 2

[53] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg. A scalable approach to activity recognition based on object use. In *CVPR*, 2007. 1

[54] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*, 2014. 2

[55] Z. Wu, Y. Fu, Y.-G. Jiang, and L. Sigal. Harnessing object and scene semantics for large-scale video understanding. In *CVPR*, 2016. 1, 2

[56] X. Xu, T. Hospedales, and S. Gong. Semantic embedding space for zero-shot action recognition. In *ICIP*, 2015. 8

[57] X. Xu, T. Hospedales, and S. Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *ECCV*, 2016. 1, 2, 8

[58] X. Xu, T. Hospedales, and S. Gong. Transductive zero-shot action recognition by word-vector embedding. *IJCV*, 2017. 2, 8

[59] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative cnn video representation for event detection. In *CVPR*, 2015. 2

[60] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010. 1

[61] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 1

[62] B. Yao, A. Khosla, and L. Fei-Fei. Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses. In *ICML*, 2011. 1

[63] G. Yu and J. Yuan. Fast action proposals for human action detection and search. In *CVPR*, 2015. 2

[64] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu. Robust relative attributes for human action recognition. *PAA*, 2015. 2