What do 15,000 object categories tell us about classifying and localizing actions?

Mihir Jain[†] Jan C. van Gemert[†] Cees G. M. Snoek[†⋆]

[†]University of Amsterdam [⋆]Qualcomm Research Netherlands

Abstract

This paper contributes to automatic classification and localization of human actions in video. Whereas motion is the key ingredient in modern approaches, we assess the benefits of having objects in the video representation. Rather than considering a handful of carefully selected and localized objects, we conduct an empirical study on the benefit of encoding 15,000 object categories for action using 6 datasets totaling more than 200 hours of video and covering 180 action classes. Our key contributions are i) *the first in-depth study of encoding objects for actions,* ii) *we show that objects matter for actions, and are often semantically relevant as well.* iii) *We establish that actions have object preferences. Rather than using all objects, selection is advantageous for action recognition.* iv) *We reveal that object-action relations are generic, which allows to transferring these relationships from the one domain to the other. And,* v) *objects, when combined with motion, improve the state-of-the-art for both action classification and localization.*

1. Introduction

This paper contributes to automatically classifying and localizing human actions like *phoning*, *horse-riding*, and *sumo wrestling* in video content. Different from the leading techniques in these two challenging problems, *e.g.* [32, 33, 51] and [15, 22, 44], which all emphasize on encoding motion for action, we study the benefits of having objects in the video representation. For action classification the relationship between objects and actions has been considered earlier [10, 11, 46], but only for a handful of carefully selected and localized objects. By contrast, we are the first to encode the *presence* of thousands of object categories for action classification and localization in a large, diverse, and comprehensive evaluation.

Fueled by large-scale image collections such as ImageNet [6], containing more than 14M labeled images of 22K objects, an extensive evaluation of objects in action is feasible. Moreover, inspired by the renaissance of convolutional neural networks, classifiers for objects are now more efficient to train, faster to apply and better in accuracy than ever before [3, 20, 59]. These deep nets learn the invariant representation and object classification result simultaneously by back-propagating information, through stacked convolution and pooling layers, with the aid of a large number of labeled examples. In this paper, we train 15,000 object classifiers with a deep convolutional neural network [20] and use their pooled responses as a video representation for action classification and localization. Figure 1 demonstrates the interest of encoding objects for actions.

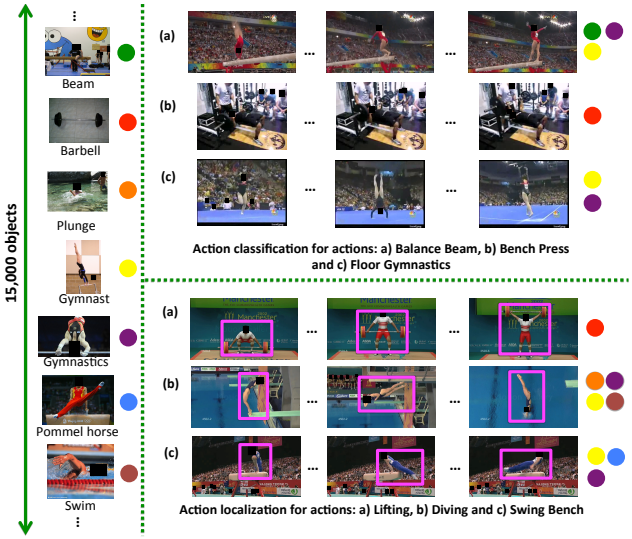


Figure 1. In this paper we ask ourselves the question: *"What do 15,000 object categories tell us about classifying and localizing actions?"* and conduct an empirical study on the benefit of encoding object categories for action. We show that: objects matter for actions, actions have object preference, object-action relations are generic, and adding object encodings improves the state-of-the-art in action classification and localization.

Being the first to consider such a large set of object category responses for action classification and localization, we conduct an empirical study on the benefit of encoding objects for action. Our study is inspired by the example set by Deng *et al.* [5] for image categorization. Our key contributions are *i)* the first in-depth study of encoding objects for actions, *ii*) we show that objects matter for actions, and are often semantically relevant as well, especially when

the actions interact with objects. What is more, the object representation is complimentary to modern motion encodings. *iii)* We establish that actions have object preferences. Rather than using all objects, selection is advantageous. Surprisingly, adding a video representation consisting of just 75 selected object categories to a motion encoding leads to an absolute improvement of 9.8% on the validation set of the THUMOS challenge 2014 [18]. *iv)* We reveal that object-action relations are generic, which allows the transfer of these (learned) relationships from the one dataset to the other. In fact, when tested on HMDB51 [21], the object-action relations learned on UCF101 [41] are stronger predictors than the ones learned on HMDB51 itself. And *v)* objects improve the state-of-the-art on five datasets for action classification and localization. Before detailing our empirical study and findings, we start with a review of related work on video representations for action classification and localization.

2. Related work

Advancements in action classification have resulted in a mature repertoire of elegant and reliable techniques yielding good accuracy. Such techniques include sampling at interest points [7, 23], densely [40, 54] or along dense trajectories [9, 17, 28, 50]. Such samples are represented by powerful local descriptors [4, 19, 24] which are robust to modest appearance and motion changes. Invariance to camera motion is either directly modeled from the video [14, 45, 51] or built into the local MBH descriptor [4, 50]. After aggregating local descriptors in a global video representation such as versions of VLAD [14, 32] or Fisher [30, 33, 51] they are input to a classifier such as SVM. The success of these techniques are due to their excellent performance as well as their practical simplicity which facilitates easy adoption by others. We follow the action classification tradition and study how it can profit from inclusion of object categories.

Though most of the emphasis in the action literature has been on classification, progress has also been made in action localization [2, 15, 22, 29, 44]. In addition to the class, localization requires specifying the location of the action in the video. Instead of localizing cuboids [2], recently action location is more precisely defined as a sequence of bounding boxes [15, 22]. To handle the huge search space that comes with such precise localization, methods to efficiently sample action proposals [15, 29] are combined with the motion encodings used in action classification. In this paper, we study whether object categories can also play a role in action localization.

Actions have been shown to correlate with their object and scene surroundings. The type of scene puts a prior on the possible actions [27, 49, 56] and vice-versa, an action may disambiguate similar looking objects [10, 42]. The most important object to detect is arguably the person performing the action. The pose of the person allows modeling human-object interactions [13, 35, 57]. In addition to the person detector, the number of detected objects has steadily increased from 2 objects [11], 4 objects [46] to 27 objects [58]. We follow these methods to exploit object-action relations, although we only detect the presence of an object, not its location. This allows us to scale up the number of objects with several orders of magnitude to a set of 15,000 object categories.

As an alternative to object categories, attributes for action classification have traditionally focused on the motion or on the action. A video can be embedded in a representation of holistic actions [36, 38], or, attributes are defined on atomic parts of the action [53]. Such attributes allow zero shot recognition [26], or modeling semantic relations over time [25, 43]. Instead of action attributes, we consider object categories in addition to the powerful motion features as common in action classification. By using object categories we do not intend to replace the action, rather, we study how objects can augment the action representation.

3. Empirical study

3.1. Object responses

We compute the likelihood of the presence of object categories in each frame of the considered videos. We use an in-house implementation of a Krizhevsky style cuda-convnet with dropout [20]. The convolution network has eight layers with weights and is trained using error back propagation. In our set of objects to learn, we include all ImageNet object categories that have more than 100 examples available, leading to 15k objects in total. At test time we obtain the likelihood of the presence of each category in a provided video frame. The N ($\sim$15k) dimensional vector of object attribute scores ($S(i)$; $i = 1...N$) is computed for each frame and to obtain the video representation these vectors are simply averaged across the frames: $\psi_x = \frac{1}{F}\sum S_{x_f}$, where F is the number of frames in video x, S_{x_f} is the object vector representation per frame f.

3.2. Motion

We capture motion information by several local descriptors (HOG, HOF and MBH) computed along the improved trajectories [51]. Improved trajectories is one of the recently proposed approaches that takes into account camera motion compensation, which is shown to be critical in action recognition [14, 51]. To encode the local descriptors, we use Fisher vectors. We first apply PCA on these local descriptors and reduce the dimensionality by a factor of two. Then 256,000 descriptors are selected at random from the training set to estimate GMM with K (=256) Gaussians. Each video is then represented by $2DK$ dimensional Fisher vector, where D is the dimension of descriptors after PCA.

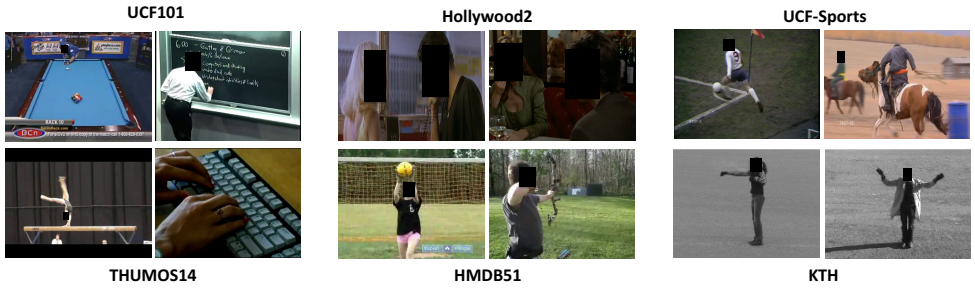


Figure 2. Action examples from the UCF101 [41], THUMOS14 [18], Hollywood2 [27], HMDB51 [21], UCF Sports [37] and KTH [39] datasets that we consider in our empirical study to reveal what 15,000 object categories tell us about classifying and localizing actions.

Finally, we apply power and ℓ_2 normalization to the Fisher vector as suggested in [34].

3.3. Datasets and evaluation criteria

Below we present the databases used in this paper. Some example frames and action classes from these datasets are shown in Figure 2.

UCF101. The UCF101 dataset [41] is a large action recognition dataset containing 13,320 videos and includes 101 action classes. It has large variations (camera motion, appearance, scale, etc) and exhibits a lot of diversity in terms of actions. It also contains many action classes that involve objects. We perform evaluation on three train/test splits and report the mean average accuracy over all classes.

THUMOS14. THUMOS14 [18] is the largest action dataset proposed to date in terms of number of classes, length and number of videos. It includes UCF101 as its train set and also has a background, validation and test sets. The validation and test set contain 1010 and 1584 temporally untrimmed long videos respectively. When testing on validation set, UCF101 is used for training. And for test set both UCF101 and validation set are used for training. Mean average precision (mAP) is the measure for evaluation.

Hollywood2. The Hollywood2 [27] dataset contains 1,707 video clips from 69 movies representing 12 action classes. The dataset is divided into a train set and test set of 823 and 884 samples respectively. Following the standard evaluation protocol of this benchmark, we use mAP as the evaluation measure.

HMDB51. The HMDB51 [21] dataset contains 6,766 video clips extracted from various sources, ranging from movies to YouTube. It consists of 51 action classes, each having at least 101 samples. We follow the evaluation protocol of [21] and use three train/test splits. The average classification accuracy is computed over all classes. We use the original set, which is not stabilized for camera motion.

UCF Sports. This dataset consists of sports broadcasts with realistic actions captured in dynamic and cluttered environments [37]. It has 10 classes and 150 videos (103 for train and 47 for testing). We use it in this work for action localization experiments as it has groundtruth for action location as a sequence of bounding-boxes. The area under the ROC curve (AUC) is the standard evaluation measure used, and we follow this convention.

KTH. This dataset consists of 6 action classes [39]. Each action is performed several times by 25 subjects. The background is homogeneous and static in most sequences and no objects are involved. We use it here to gauge the impact of objects in such an artificial scenario. The average classification accuracy is used for evaluation measure.

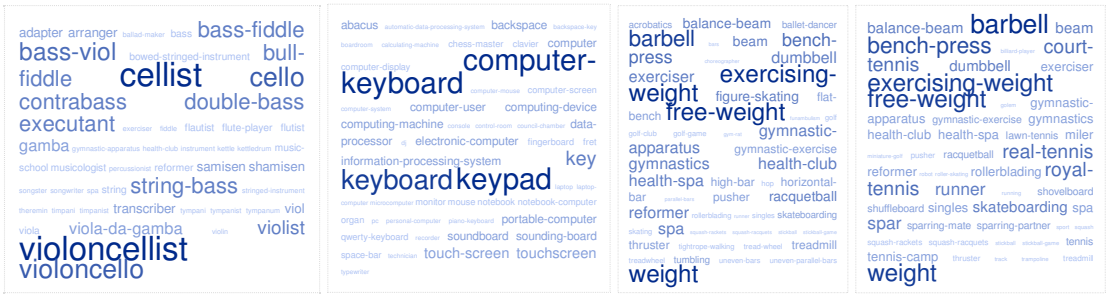
3.4. Classification

For the classification, we use a linear SVM. When combining different descriptors, we simply add the kernel matrices, as done in [46]. The multi-class classification problem that we consider is addressed by applying a *one-against-rest* approach.

4. Objects matter for actions

Qualitative experiment: We first perform a qualitative experiment that visualizes the contribution per object category on the action classes of UCF101. To compute the contribution, we first ℓ_1 normalize the ψ_x vectors for all the videos in the UCF101 training set to make them comparable. Then, to compute the contribution per object for the j^{th} action class, c_j, we sum the normalized response vectors for the videos belonging to c_j, *i.e.*, $\beta_j = \sum_{x \in c_j} \frac{\psi_x}{\sum_i \psi_x(i)}$. The response of the i^{th} object category for the j^{th} action class is then simply $\beta_j(i)$ and can be used to visualize the proportional contribution per object. We show qualitative results for four action classes in Figure 3.

In many cases the object responses seem semantically related to the action classes. For *playing cello*, the objects cellist and cello are obviously characteristic and indeed they result in a high response as well. The same holds for the response of keyboard and keypad for *typing*. In case an action involves very small objects the relationship is more



(a) *Playing cello* (b) *Typing* (c) *Body weight squats* (d) *Lunges*

Figure 3. Qualitative experiment visualizing the contribution per object category on four action classes of UCF101. For *playing cello* and *typing*, the prominent object categories in terms of response are also semantically relevant. For *body weight squats* and *lunges* the dominant objects are not essential for the actions, but do appear frequently in their background context. Depending on whether actions share the same background, these objects may or may not lead to a discriminative representation for actions.

difficult to capture with the current implementation of object classifiers, we observe this for *juggling balls* (data not shown). Naturally the co-occurrence between actions and objects can also capture objects that are not necessarily part of the action, for *body weight squats* and *lunges*, for example, we find objects like barbell and weight that commonly appear in the gym.

To get further insight into the object and action interdependence, a heat-map between action classes of UCF101 and the set of most responsive object categories are shown in Figure 4. For clarity we only show 34 action classes (every third class) and the union of the most responsive object for each action class. For each action class, we find that the object categories that have highest responses are semantically related to the action class. *Trampoline jumping* has high response from trampoline, *Rafting* has objects raft and kayak with high responses. Thus, the object responses seem to make sense for the action classes from a semantic point of view. Next, we will assess if they also contribute to the visual recognition.

Quantitative classification experiment: To assess the quantitative value of objects for action we perform an action classification experiment on the UCF101 and THUMOS14 validation datasets, which both have a large variety of action classes. On purpose, we also include the KTH dataset, which is a dataset with six action classes that are devoid of objects. We compare action classification using a representation of object responses with one using motion, as detailed in Section 3. For KTH we encode motion with HOF only, because the number is close to 100% already. We summarize the results in Table 1.

A representation using object responses performs reasonably on UCF101 and THUMOS14, but 78.7% on KTH is rather poor on this relatively simple dataset. The results confirm that objects allow for action classification, but only if the actions interact with objects. As expected, a representation based on motion is better on all three datasets.

Method	UCF101	THUMOS14 val	KTH
Objects	65.6%	49.7%	78.7%
Motion	84.2%	56.9%	94.9%
Objects + Motion	88.1%	66.8%	95.4%

Table 1. Quantitative action classification experiment comparing objects, motion and their combination. For the realistic actions of UCF101 and THUMOS14 val, objects added to motion, lead to a considerable improvement in the performance. The gain in case of KTH is minimal as it does not involve objects. When objects are combined with motion it always improves performance further. We conclude that objects matter for actions.

However, when objects are combined with motion it always improves performance further, even on KTH. For UCF101 the absolute improvement over motion only is 3.9% and for THUMOS14 is as much as 9.9%. This is a considerable increase and demonstrates that objects matter for actions.

Quantitative localization experiment: Another interesting aspect is to see where the informative object responses are located with respect to the action. Are the object responses from the area encompassing the action more discriminative than those in the background? Or the ones from everywhere in the video, together are more discriminative? This is an important aspect because if the responses in the proximity of actions are informative, then the objects can also play a role in action localization. We conduct an experiment on the UCF Sports dataset, which comes with localization groundtruth. In absence of object location groundtruth, we compute the 15,000 object responses for the full video, inside and outside the groundtruth tube. Action classification is done using these three representations and the average precisions are reported in Table 2.

We obtain the best results when the objects are encoded inside the groundtruth tube, leading to an mAP of 74.4%, where encoding the entire video scores 60.7% and outside the groundtruth 53.5%. Except for *lifting*, all action classes improve. Note that there is no motion encoding used here.

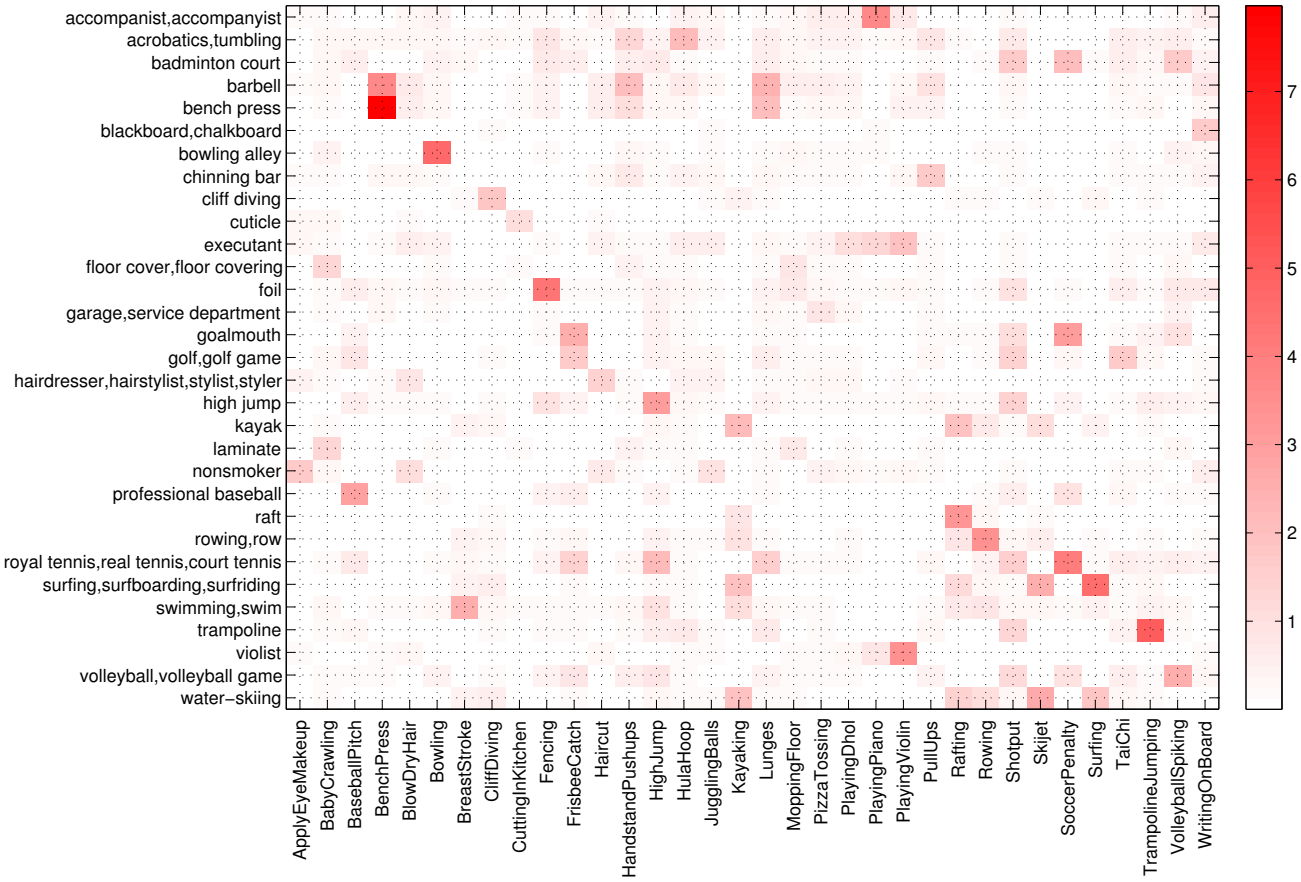


Figure 4. Qualitative experiment visualizing the heat-map between the 31 most responsive objects (y-axis) and 34 action classes (x-axis) of the UCF101 dataset (every3^{rd} action class is chosen for clarity). The high responses for semantically related action-object pairs is apparent. Note the high responses for the objects trampoline, raft, blackboard, kayak and their associated actions.

Although it can depend on the action class and the videos, in general the results suggest that object responses close to and involved in the actions matter most. We show the advantage of using objects for action localization in Section 7.2.

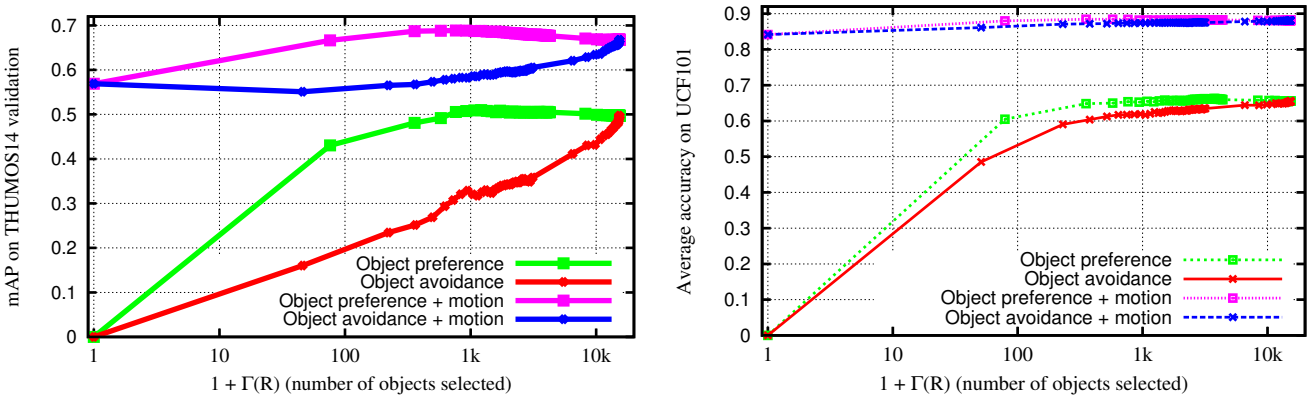
5. Actions have object preference

For a given video dataset of n action classes, we might not need all N object categories to obtain an optimal representation. Only the categories relevant to the action classes, ideally corresponding to those leading to a discriminative representation, are required. So the objective is to find a subset of m object categories from the N object categories, such that the discriminative power of the representation is maximized for a given set of action classes. We refer to these objects as *preferred* objects. To each action class j we assign a set of the top R most responsive object categories, $top_R(c_j) = R\text{-}\arg\max_i \psi_x(i)$. The union of these sets of object categories, for $j = 1..n$, gives us a set of preferred objects for the given set of action classes, $\Gamma(R) = \bigcup_j top_R(c_j)$. While the absence of an object category in an action class is also informative, it would be less discriminating as it may be absent for many other action classes as well.

Classes	Video	Outside tube	Inside tube
Diving	100.0%	100.0%	100.0%
Kicking	66.7%	16.7%	66.7%
Lifting	100.0%	100.0%	50.0%
Riding-horse	100.0%	100.0%	100.0%
Running	50.0%	50.0%	75.0%
Skateboarding	0	0	25.0%
Swinging	66.7%	16.7%	100.0%
Swinging-bar	0	75.0%	75.0%
Swinging-golf	66.7%	33.3%	66.7%
Walking	57.1%	42.9%	86.7%
Mean	60.7%	53.5%	74.4%

Table 2. Average precisions for action classes of UCF Sports dataset using object responses from: the whole video, only the background of the action, and only in the vicinity of the action. Evidently, object responses in the vicinity of the action matter most.

We evaluate the impact of object preference on action classification by varying the value of R in light of a representation consisting of (a) objects only and (b) objects



(a) THUMOS14 (b) UCF101

Figure 5. Accuracies of (1) objects only and (2) objects+motion, as a function of R for THUMOS14 and UCF101. On x-axis number of selected objects are shown for a different values of R from 0 to 15,000. *Preference* plots reveal that only a few hundred most responsive categories are needed to achieve the maximum gain. *Avoidance* plots keep on improving till the last and the most preferred category is added, again showing that actions have object preference.

with motion on both the THUMOS14 and UCF101 datasets. The accuracies are computed for different values of R (one-against-rest SVMs are re-trained for each value of R), *i.e.* starting with no objects, then progressively adding the most responsive object categories for the given dataset, till all the object categories are used. We refer to this plot as the one with object preference. We also do the reverse, *i.e.* start with no objects, then the least responsive objects and progressively add more responsive ones until the most responsive object is also added at the end. This plot is referred to as the one with object avoidance. We present results in Figure 5.

The results on both THUMOS14 and UCF101 show that some objects are more important than others for action classification. When comparing object preference (green line) with object avoidance (red line) we observe that adding the most preferred objects first, results in a better mAP and mean accuracy than adding the least preferred object. More importantly, the accuracies for the object avoidance plots keep on increasing till the last and most preferred object is added. Again, the combination of objects and motion results in a big increase in performance. Interestingly, when just one most preferred object category per action class is added to the motion encoding it results in a big jump (pink line). On THUMOS14 this leads to an improvement of 9.8% with an object encoding of only 75 dimensions, where on UCF101 the results improve from 84.2% to 88.0%. For UCF101 the pink plot peaks at around R=6 and then stabilizes. In case of THUMOS14 it dips a bit as the videos are not temporally constrained and extra object categories can cause confusion because of the objects from the parts of the videos not containing actions of interest. The selection is therefore more important for THUMOS14 and hence the larger gain. For both datasets, plots consistently peak at around R=6 or 11 and then stabilize, so from here on for all the datasets we use R=11 as the optimal choice of R ($R^\star$). We further note that using $R^\star$ results in higher performance than using all objects for both THUMOS14 (68.8% vs 66.8%) and UCF101 (88.5% vs 88.1%). These results show that actions have object preference.

6. Object-action relations are generic

We achieved a significant boost in the numbers with just a few object categories selected for a given set of actions. Now, we evaluate if this knowledge of characteristic objects learned from one dataset can be transferred to another dataset. For this we conduct experiments on HMDB51 and UCF101 as they have 12 action classes in common. We learn the preferred set of objects from the training sets of HMDB51 and another preferred set from UCF101 (using the same simple procedure as before). Then we use these sets for the object representation of videos in the test set of HMDB51. We compare the impact of the representations for these two transfers in Table 3.

We first consider the motion baseline for the 12 action categories on HMDB51 which scores a mean classification accuracy of 83.6%. It obtains the best overall result for *dive*, as this action in HMDB51 includes bungee jumping, base jumping, cliff diving etc, there is no representative object only the act of jumping is common. For all other actions, adding objects in the video representation, as learned from training data, is better. Including only the top object per action, as learned from the HMDB51 train set, increases the mean accuracy to 85.2%, while adding the best set of objects increases the mean accuracy to 87.5% and obtains the best overall results for *climb pullup*, *punch*, *shoot ball* and *shoot bow*. The most prominent objects for these classes are: climber, barbell, sparring, volleyball, and archery.

Classes	HMDB51	HMDB51		UCF101	
	Motion	+ Objects			
		R=1	$R^\star$	R=1	$R^\star$
Brush hair	96.7%	95.6%	96.7%	96.7%	98.9%
Climb	87.8%	92.2%	92.2%	90.0%	92.2%
Dive	87.8%	81.1%	84.4%	84.4%	85.6%
Golf	98.9%	98.9%	98.9%	98.9%	98.9%
Handstand	90.0%	85.6%	90.0%	90.0%	88.9%
Pullup	91.1%	86.7%	92.2%	88.9%	92.2%
Punch	85.6%	87.8%	88.9%	84.4%	87.8%
Pushup	72.2%	77.8%	88.9%	90.0%	88.9%
Ride bike	76.7%	95.6%	91.1%	95.6%	93.3%
Shoot ball	86.7%	91.1%	93.3%	87.8%	92.2%
Shoot bow	92.2%	92.2%	94.4%	93.3%	94.4%
Throw	37.8%	37.8%	36.7%	36.7%	43.3%
Mean	83.6%	85.2%	87.5%	86.4%	88.1%

Table 3. The characteristic object categories learned from the training sets of HMDB51 or UCF101 transfer to the test set of HMDB51, and always improve over the motion-only baseline. Interestingly, learning the characteristic objects on UCF101 results in the biggest gain in mean accuracy. We conclude that object-action relations are generic.

Interestingly, learning the characteristic set of objects on UCF101, leads to even better mean accuracy on HMDB51. Including only the most preferred object per action results in an accuracy of 86.4%, whereas including the best set of objects obtains 88.1%, a considerable improvement over using motion only. The improvement for *pushup* is significant with the most preferred objects only. The key object here is benchpress, which has same type of posture as in pushups. For *ride bike*, both the sets lead to large improvement. The key object here from both the sets is safety bicycle. The better performance with UCF101 is probably because of a richer learning set due to a larger number of videos in it. The ability to learn a discriminative set of objects from action classes that transfers to unknown videos with the same action classes opens up future possibilities for action recognition such as zero-shot classification. For the moment we conclude that object-action relations are generic.

7. Objects improve the state-of-the-art

In this section, we conduct experiments for action classification and action localization, and compare with state-of-the-art methods for both these tasks.

7.1. Action classification

For action classification, we experiment on four datasets, namely UCF101, THUMOS14 validation/test, Hollywood2 and HMDB51. The bottom half of Table 4 lists the numbers for objects, just motion, objects and motion (for all objects, $R = 1$ and R^*). For UCF101, just by adding one top object category for each action class improves the average accuracy from 84.2% to 88.0%. The same tactic leads to an improvement of 9.8% on THUMOS14 validation and 5.3% on THUMOS14 test set with just a 75 (=$\Gamma(1)$) dimensional representation. Corresponding improvements for HMDB51 and Hollywood2 are 3.1% and 2.0% respectively. It is also interesting to see the action classification performance with just objects is competitive on UCF101 and THUMOS14 datasets as they involve more actions based on objects.

In the top half of Table 4, we compare our results with the best methods in the literature on these four datasets. Three of the best performers on UCF101 [1, 32, 55] propose improved encoding methods. Another one is the approach of Wang *et al*. [51] with spatial pyramid. When objects are combined with the motion representation we achieve the best average accuracy. The most competitive method is the recent work of Peng *et al*. [32] that uses higher order VLAD and learns codebooks in a supervised manner, which does better than Fisher vectors. On the recent THUMOS14 dataset, we report for the top 3 performing approaches in the competition (*i.e*. with best mAPs on test set). In the column for the validation set we include [47] as they have the same train/test setting. Our approach achieves the state-of-the-art mAPs of 71.6% and 68.8% on test and validation sets respectively. Due to object selection this is slightly better than our winning approach at THUMOS14 [16].

Robust motion descriptors along the dense trajectories with higher order encodings have also done well [14, 30, 32, 51] on the Hollywood2 and HMDB51 datasets. Our motion baseline is similar to these methods and by adding objects we achieve considerable improvements on both the datasets. The most recent methods of Peng *et al*. [33], Hoai *et al*. [12] and concurrent work of Fernando *et al*. [8], have further raised the bar. Temporal ordering in video as motion or as evolution of appearance are exploited in [8, 12] for action classification. We expect these methods to improve further by adding our object representation. In [33], Fisher vectors are combined with stacked Fisher vectors (2-layers of Fisher vectors). The authors provided us with their stacked Fisher vectors for HMDB51, which on combining with our Fisher vectors achieves an average accuracy of 69.9%. After adding our object representation we obtain 71.3%. Objects combined with any of the above representation boosts the performance and leads to state-of-the-art action classification results on the UCF101, THUMOS14 (validation and test) and HMDB51 datasets.

7.2. Action localization

The objective of action localization is to detect when and where an action of interest occurs. We follow the convention in the literature to localize an action as a sequence of bounding boxes [15, 22, 44]. We conduct this experiment on UCF Sports. We obtained the tubelet action proposals using independent motion evidence from Jain *et al*. [15]. As motion representation, we use the motion boundary histogram as local descriptor and aggregate as bag of features.

UCF101		THUMOS14 val		THUMOS14 test		Hollywood2		HMDB51	
Soomro *et al.* [41]	43.9%	Varol *et al.* [47]	62.3%	Varol *et al.* [47]	63.2%	Zhu *et al.* [60]	61.4%	Zhu *et al.* [60]	54.0%
Cai *et al.* [1]	83.5%	Jain *et al.* [16]	66.8%	Oneata *et al.* [31]	67.2%	Vig *et al.* [48]	61.9%	Oneata *et al.* [30]	54.8%
Wu *et al.* [55]	84.2%			Jain *et al.* [16]	71.0%	Jain *et al.* [14]	62.5%	Wang *et al.* [51]	57.2%
Wang *et al.* [52]	85.9%					Oneata *et al.* [30]	63.3%	Peng *et al.* [32]	59.8%
Peng *et al.* [32]	87.7%					Wang *et al.* [51]	64.3%	Hoai *et al.* [12]	60.8%
						Hoai *et al.* [12]	73.6%	Fernando *et al.* [8]	63.7%
						Fernando *et al.* [8]	**73.7%**	Peng *et al.* [33]	66.8%
Objects	65.6%		49.7%		44.7%		38.4%		38.9%
Motion	84.2%		56.9%		63.1%		64.6%		57.9%
Objects + Motion	88.1%		66.8%		70.8%		66.2%		61.1%
Objects (R=1) + Motion	88.0%		66.7%		68.4%		66.6%		61.0%
Objects ($R^\star$) + Motion	**88.5%**		**68.8%**		**71.6%**		66.4%		61.4%
Objects + Peng *et al.* [33]	–		–		–		–		**71.3%**

Table 4. Comparison of our approach using objects and motion with the state-of-the-art on the UCF101, THUMOS14 validation and test set, Hollywood2 and HMDB51 datasets. Adding object categories in the video representation improves the state-of-the-art in action classification for these datasets.

	Overlap threshold					
	0.1	0.2	0.3	0.4	0.5	0.6
Objects	55.5	55.0	48.2	38.1	30.3	19.7
Motion	51.6	49.2	44.2	30.9	20.6	14.0
Objects + Motion	56.1	54.1	51.2	42.1	**34.3**	**27.3**
Motion (+2x2)	57.4	56.6	51.8	41.8	31.0	23.3
Objects + Motion (+2x2)	**58.1**	**57.8**	**52.2**	**42.2**	33.7	24.0

Table 5. Impact of encoding tubelet proposals from [15] using object category responses, motion and a spatial 2x2 grid for action localization on UCF Sports.

The codebook size is set to $k = 500$ and we report with and without a spatial grid of 2×2. We extract 15,000 object responses from each bounding box of the tubelet and average them. Power normalization is followed by l_2 normalization on the averaged vector for the final tubelet representation. Linear SVM is used for the classification. The top five detections are considered for each video after non-maximum suppression.

The area under the ROC (AUC) is reported in Table 5. Objects alone attains excellent AUCs, even greater than motion without spatial grid, up to 10% absolute improvement for an overlap threshold of 0.5. When object and motion are combined the results improve further, and adding the spatial grid improves the AUC even more. We compare the results of this best representation with the state-of-the-art methods in Figure 6. Our numbers are better than the current best approach of Jain *et al.* [15], while we use only a small subset of their tubelet proposals. We obtain an absolute improvement of 6%-7% for difficult thresholds of above 0.3 overlap with the groundtruth action sequence of bounding boxes. We conclude that objects improve the state-of-the-art in action localization.

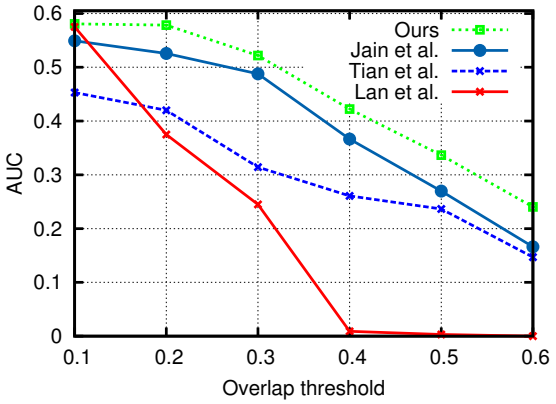


Figure 6. Comparison of our approach, using tubelet proposals from [15] encoded with objects and motion, with Lan *et al.* [22], Tian *et al.* [44] and Jain *et al.* [15] on the UCF Sports dataset. The Area under the ROC curve is shown for overlap thresholds from 0.1 to 0.6. Objects improve the state-of-the-art in action localization.

8. Conclusions

In this paper we ask ourselves the question: *"What do 15,000 object categories tell us about classifying and localizing actions?"* and conduct an empirical study on the benefit of encoding 15,000 object categories for actions. Our experiments show that objects matter for actions, and are often semantically relevant as well, especially when the actions interact with objects. What is more, the object representation is complementary to modern motion encodings. We establish that actions have object preferences. Rather than using all objects, selection is advantageous both in terms of the compact video representation as well as in terms of its discriminative action classification ability. The learned object-action relationships are generic, and transfer from one dataset to another. When our object representations are combined with modern motion encodings and spatio-temporal action proposals it leads to a new state-of-the-art on five datasets for action classification and localization.

We believe that learning generic object-action relationships opens up new directions. For example, selecting the relevant subset of objects to include in the action representation or the possibility for zero-shot action recognition. Another appealing next step is to localize objects and encode their locations for classifying and localizing actions.

Acknowledgments This research is supported by the STW STORY project, the Dutch national program COMMIT, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20067. 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/NBC, or the U.S. Government.

References

[1] Z. Cai, L. Wang, X. Peng, and Y. Qiao. Multi-view super vector for action recognition. In *CVPR*, 2014. 7, 8

[2] L. Cao, Z. Liu, and T. S. Huang. Cross-dataset action detection. In *CVPR*, 2010. 2

[3] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014. 1

[4] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006. 2

[5] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010. 1

[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. 1

[7] I. Everts, J. C. van Gemert, and T. Gevers. Evaluation of color spatio-temporal interest points for human action recognition. *IEEE TIP*, 23(4):1569–1580, 2014. 2

[8] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars. Modeling video evolution for action recognition. In *CVPR*, 2015. 7, 8

[9] A. Gaidon, Z. Harchaoui, and C. Schmid. Recognizing activities with cluster-trees of tracklets. In *BMVC*, 2012. 2

[10] A. Gupta and L. S. Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, 2007. 1, 2

[11] D. Han, L. Bo, and C. Sminchisescu. Selection and context for action recognition. In *ICCV*, 2009. 1, 2

[12] M. Hoai and A. Zisserman. Improving human action recognition using score distribution and ranking. In *ACCV*, 2014. 7, 8

[13] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *ECCV*, 2010. 2

[14] M. Jain, H. Jégou, and P. Bouthemy. Better exploiting motion for better action recognition. In *CVPR*, 2013. 2, 7, 8

[15] M. Jain, J. C. van Gemert, H. Jégou, P. Bouthemy, and C. G. M. Snoek. Action localization with tubelets from motion. In *CVPR*, 2014. 1, 2, 7, 8

[16] M. Jain, J. C. van Gemert, and C. G. M. Snoek. University of Amsterdam at THUMOS Challenge 2014. In *ECCV workshop*, 2014. 7, 8

[17] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo. Trajectory-based modeling of human actions with motion reference points. In *ECCV*, 2012. 2

[18] 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. 2, 3

[19] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008. 2

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 2

[21] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011. 2, 3

[22] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *ICCV*, 2011. 1, 2, 7, 8

[23] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003. 2

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

[25] W. Li and N. Vasconcelos. Recognizing activities by attribute dynamics. In *NIPS*, 2012. 2

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

[27] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 2, 3

[28] P. Matikainen, M. Hebert, and R. Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. In *ICCV Workshop*, 2009. 2

[29] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid. Spatio-temporal object detection proposals. In *ECCV*, 2014. 2

[30] D. Oneata, J. Verbeek, and C. Schmid. Action and Event Recognition with Fisher Vectors on a Compact Feature Set. In *ICCV*, 2013. 2, 7, 8

[31] D. Oneata, J. Verbeek, and C. Schmid. The LEAR submission at Thumos 2014. In *ECCV workshop*, 2014. 8

[32] X. Peng, L. Wang, Y. Qiao, and Q. Peng. Boosting vlad with supervised dictionary learning and high-order statistics. In *ECCV*, 2014. 1, 2, 7, 8

[33] X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked fisher vectors. In *ECCV*, 2014. 1, 2, 7, 8

[34] F. Perronnin, J.Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010. 3

[35] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *IEEE TPAMI*, 34(3):601–614, 2012. 2

[36] D. Ramanan and D. A. Forsyth. Automatic annotation of everyday movements. In *NIPS*, 2003. 2

[37] M. D. Rodriguez, J. Ahmed, and M. Shah. Action MACH: a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 3

[38] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012. 2

[39] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004. 3

[40] F. Shi, E. Petriu, and R. Laganiere. Sampling strategies for real-time action recognition. In *CVPR*, 2013. 2

[41] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, 2012. 2, 3, 8

[42] A. Srikantha and J. Gall. Discovering object classes from activities. In *ECCV*, 2014. 2

[43] C. Sun and R. Nevatia. Active: Activity concept transitions in video event classification. In *ICCV*, 2013. 2

[44] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *CVPR*, 2013. 1, 2, 7, 8

[45] H. Uemura, S. Ishikawa, and K. Mikolajczyk. Feature tracking and motion compensation for action recognition. In *BMVC*, 2008. 2

[46] M. M. Ullah, S. N. Parizi, and I. Laptev. Improving bag-of-features action recognition with non-local cues. In *BMVC*, 2010. 1, 2, 3

[47] G. Varol and A. A. Salah. Extreme Learning Machine for Large-Scale Action Recognition. In *ECCV workshop*, 2014. 7, 8

[48] E. Vig, M. Dorr, and D. Cox. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *ECCV*, 2012. 8

[49] T. Vu, C. Olsson, I. Laptev, A. Oliva, and J. Sivic. Predicting actions from static scenes. In *ECCV*, 2014. 2

[50] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 2

[51] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In *ICCV*, 2013. 1, 2, 7, 8

[52] H. Wang and C. Schmid. LEAR-INRIA submission for the THUMOS workshop. In *ICCV workshop*, 2013. 8

[53] L. Wang, Y. Qiao, and X. Tang. Mining motion atoms and phrases for complex action recognition. In *ICCV*, 2013. 2

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

[55] J. Wu, Y. Zhang, and W. Lin. Towards good practices for action video encoding. In *CVPR*, 2014. 7, 8

[56] X. Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. In *CVPR*, 2011. 2

[57] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 2

[58] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011. 2

[59] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 1

[60] J. Zhu, B. Wang, X. Yang, W. Zhang, and Z. Tu. Action Recognition with Actons. In *ICCV*, 2013. 8