Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Exploring Semantic Inter-Class Relationships (SIR) for Zero-Shot Action Recognition

Chuang Gan[1], **Ming Lin**[3], **Yi Yang**[2], **Yueting Zhuang**[4] and **Alexander G. Hauptmann**[3]
[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[2] Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, Sydney, Australia
[3] School of Computer Science, Carnegie Mellon University, Pittsburgh, USA
[4] College of Computer Science, Zhejiang University, Zhejiang, China
ganchuang1990@gmail.com, linming04@gmail.com,
yiyang@cs.cmu.edu, yzhuang@zju.edu.cn, alex@cs.cmu.edu

Abstract

Automatically recognizing a large number of action categories from videos is of significant importance for video understanding. Most existing works focused on the design of more discriminative feature representation, and have achieved promising results when the positive samples are enough. However, very limited efforts were spent on recognizing a novel action without any positive exemplars, which is often the case in the real settings due to the large amount of action classes and the users' queries dramatic variations. To address this issue, we propose to perform action recognition when no positive exemplars of that class are provided, which is often known as the zero-shot learning. Different from other zero-shot learning approaches, which exploit attributes as the intermediate layer for the knowledge transfer, our main contribution is SIR, which directly leverages the semantic inter-class relationships between the known and unknown actions followed by label transfer learning. The inter-class semantic relationships are automatically measured by continuous word vectors, which learned by the skip-gram model using the large-scale text corpus. Extensive experiments on the UCF101 dataset validate the superiority of our method over fully-supervised approaches using few positive exemplars.

Introduction

Recent studies in computer vision and multimedia have explored the action recognition in the real world videos and made significant progress over the last decade. In literature, reliable low-level features such as STIP (Laptev et al. 2008), Mosift (Chen and Hauptmann 2009), dense trajectory (Wang et al. 2011) and improved dense trajectory (Wang, Schmid, and others 2013), combined with a modern learning algorithm such as support vector machines (SVM), have achieved promising recognition results.

To obtain good performances in action recognition, existing approaches require sufficient positive exemplars to train a series of action classifiers, i.e. one for each action. However, due to the large number of action classes, a main challenge is to gather adequate positive exemplars that exactly match the target action. Zero-shot learning addresses this problem by providing an alternative approach that does not require any positive exemplars. Instead, a user may provide other forms of side information, such as visual class hierarchy (Mensink et al. 2013; Wang et al. 2009) or attributes (Farhadi et al. 2009; Lampert, Nickisch, and Harmeling 2009). Then a transfer function is applied at test time for the unseen class, e.g., its position in the class hierarchy or an attributes-to-class mapping. As discussed in (Liu, Kuipers, and Savarese 2011), obtaining class hierarchy relationships between verbs/actions is much more difficult than discovering relationships between nouns/objects, due to the fact that verbs do not have the same well-built ontological relationships as nouns, such as WordNet. Thus, Liu *et al.* (Liu, Kuipers, and Savarese 2011) proposed to recognize actions by a piece of well-structured attribute lists, which is probably the first attempt to recognize actions only using texts. The attributes are consist of human readable properties that are shared across different classes. As discussed in object categorization (Farhadi et al. 2009; Lampert, Nickisch, and Harmeling 2009), action recognition (Liu, Kuipers, and Savarese 2011) and multimedia event recognition (Liu et al. 2013), the ability of characterizing objects and actions by attributes is not only helpful for recognizing available objects, actions and events, but also powerful for recognizing classes that have never been seen before.

Although promising for the above applications, attribute-based representation suffers from several drawbacks. Firstly, generating semantic attributes is tedious and unreasonable for many visual concepts, especially for actions. For example, to recognize the action *walking*, they need to define attributes such as *arm pendulum-like motion* and *translation motion* are positive, but the attributes like *torso up-down motion* and *torso twist* are negative. However, describing these scenes to a computer would require a lengthy textual description but still might not capture the full nuance. In addition, the description template designed in (Liu, Kuipers, and Savarese 2011) is static, making it difficult to scale up to a variety of ad-hoc actions. Therefore, in their experiments, action classes are restricted to a few simple ones such as *walk* and *jump forward*, in the clean and lab-generated video datasets such as the KTH dataset (Laptev et al. 2008) and the Weizmann dataset (Blank et al. 2005). It remains unclear

how to recognize complex actions such as *soccer penalty*, with a limited number of pre-specified attributes.

Secondly, it is difficult to obtain reliable attribute-based representations. Previous research (Ma et al. 2013) has revealed the inherent uncertainty in terms of the accuracy and reliability of the attribute representation. As action recognition directly relies on the attribute representation, the performances will degrade if attribute classifiers are inaccurate. Moreover, it remains unclear how many attributes will be sufficient and which attributes will be particularly helpful for an unknown action. Thus, it is extremely difficult and even impossible to design a static attribute pool for different actions, because actions are dynamic and diverse.

After carefully analyzing different classes of actions, we find that a series of action classes may share some elements if they are semantically similar to each other. Instead of explicitly modeling the shared information by using attribute representation, we propose to exploit the action-action inter-class relationships to transfer knowledge. Figure 1 illustrates an example. An action class *front crawl* can be well described by highly like *breaststroke*, like *crawl*, but unlike *basketball* and *volleyball*. Based on this observation, we propose a novel approach for zero-shot learning. Our method is based on semantic label transfer learning. We first train a concept detector for each known classes in the video collections. Then we conduct the semantic correlation computation from a large-scale text corpus, i.e. Wikipedia. Finally, we estimate a classifier for the testing action class as a weighted combination of related classes, using the semantic correlations to define weight. Our method avoids the high-level attributes-to-class mappings, which are challenge to be defined, tedious to be annotated, and unreliable to be used. We summarize our contributions as follows:

- We propose a simple but effective framework for zero-shot action recognition without positive exemplars and attributes.
- We demonstrate the action-action inter-class relationships can be obtained from an external ontology, which allows for effortless zero-shot action recognition, compared to the attribute-based representation.
- We conduct extensive experiments to demonstrate the effectiveness of our approach for unseen action recognition, and achieve better results than the fully-supervised approaches using few positive exemplars.

The rest of the paper is organized as follows. In Section 2, we review the related work of zero-shot learning and action recognition. In Section 3, we describe our approach in detail. Experimental settings and evaluation results are presented in Section 4. Section 5 concludes the paper.

Related Work

Our framework involves two research directions: zero-shot learning and action recognition, which will be presented in this section, respectively.

Zero-shot Learning

The task of zero-shot learning is to recognize classes that have never been seen before. Namely, there are no positive exemplars available. Attribute-based representation is introduced as an interpretable level of indirections between classes, which can be shared and reused among different classes. Most recent methods harvest the attributes by manual labelling (Farhadi et al. 2009; Lampert, Nickisch, and Harmeling 2009; Yu et al. 2013; Parikh and Grauman 2011; Larochelle, Erhan, and Bengio 2008), mining knowledge from other domains (Rohrbach et al. 2010), or extracting the features themselves (Sharmanska, Quadrianto, and Lampert 2012; Yu et al. 2013; Liu, Kuipers, and Savarese 2011; Feng et al. 2013). After obtaining attributes, the effectiveness of knowledge transferring always depends on the performances of trained classifiers independently (Lampert, Nickisch, and Harmeling 2009) or the mapping function between low-level features and attribute labels (Akata et al. 2013). However, those approaches take attributes as intermediate information, so they are indirect ways to solve zero-shot learning problems. Training attribute classifiers always costs huge human efforts and increases computational load. Therefore, they are not applicable for large-scale settings and high-speed requirements, as discussed in (Rohrbach, Stark, and Schiele 2011; Kankuekul et al. 2012). Besides object recognition, there is also some concern about zero-shot video search using visual concept as attribute (Lin et al. ; Guadarrama et al. 2013; Jiang et al. 2014).

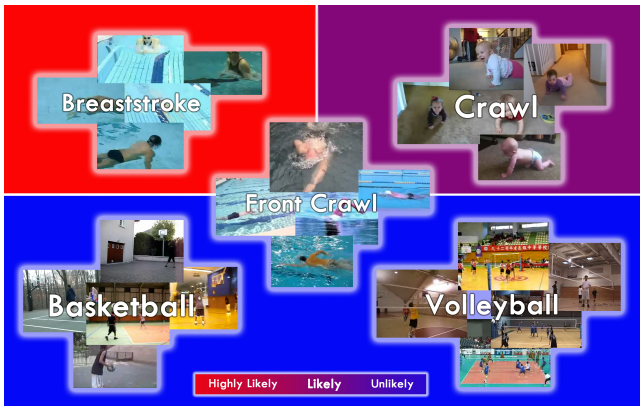


Figure 1: An example of recognizing the unknown action *front crawl*. It is highly like *breaststroke*, like *crawl*, but unlike *basketball* and *volleyball*.

Exploiting inter-class relationships for zero-shot learning has been discussed in the object recognition. In (Mensink et al. 2013), it proposed to put the new class position in the class hierarchy. In (Mensink, Gavves, and Snoek 2014), they proposed to utilize the hit counts in Yahoo search engine to measure the inter-class relationships. However, most of these works focus on the object zero-shot learning. However, it also remains unclear how these methods can be applied in action recognition. Because verbs don't have the well-defined class hierarchy relationships. And the queries of video search also change more dramatically than the image search, thus the inter-class relationships can hardly be measured through the web statistic.

Action Recognition

The problem of video analysis has been widely explored in the community of computer vision and multimedia (Gan et al. 2013). Action recognition is one of the major concern. Recently, researches focus on realistic datasets collected from movies (Laptev et al. 2008) and web videos (Reddy and Shah 2013). UCF101 (Soomro, Zamir, and Shah 2012) is the most challenging large-scale action dataset and has driven more difficult action recognition. Most successful approaches are based on some local space-time forms of features that are then encoded in fisher vectors (FV). The fisher vectors are finally fed into a SVM classifier to train specific action recognition models. In addition, several mid-level representations also draw attentions in action recognition. Action bank (Sadanand and Corso 2012) has been proposed as a new mid-level feature based on atomic action. Besides, Liu *et al.* (Liu, Kuipers, and Savarese 2011) proposed attributes as the mid-level representation and also applied it to zero-shot action recognition. However, it relied on manually-defined and data-driven attributes, so it is not applicable for the large-scale setting.

Proposed method

In this section, we define our zero-shot action recognition framework by leveraging the semantic correlations between the known and unknown action classes. In the first subsection, we introduce how we off-line get the classifiers of the known action classes. Then we describe the various linguistic knowledge bases and semantic correlations approaches we exploited to obtain the inter-class relationships between actions. Finally, we present how we use the classifiers of the known action classes and their semantic correlations to estimate the classifier of the unknown action class.

Learning Concept detectors

We define a concept collection $C = \{C_1, C_2, ..., C_N\}$, where N is the number of concepts. We then employ well-established techniques to build a classifier for each known action class. In particular, improved dense trajectory features (Wang, Schmid, and others 2013) and fisher vector (Oneata et al. 2013) representations defined over gaussian mixture model (GMM) codes of these video are used to represent each video.

For each known action class k, we use all the videos belonging to that class as positive data, and 5000 null videos from the development set as the negative data to train a concept detector by least square regression (LR) as

$$\arg\min_{w_k} \sum_{n} \left({w_k}^T x_i - y_i\right)^2 + \lambda ||w_k||^2. \quad (1)$$

where $x_i \in R^d$ is the low level feature for video i, and $y_i \in \{0, 1\}$ is the associated binary label. Then for each action class k, we will get a weight vector $w_k \in R^d$. We can also call this weight vector as the action concept detector.

Semantic correlation

We exploit the semantic correlations between action names to automatically measure the inter-class relationships. Many existing Natural Language Processing techniques can be used to measure semantic correlations (SC) between concepts. The most widely used resources in Natural Language Processing (NLP) to calculate SC of concepts are WordNet (as the largest machine readable expert-created language ontology), and Wikipedia (as the largest online encyclopedia).

WordNet path length-based SC measures. WordNet(Miller 1995) is a large-scale lexical database of the English language, originally intended as modeling the human lexical memory. English words are organized into concepts according to synonym sets or synsets. Due to its impressive size (over 100,000 concepts) and richness in encoded semantic relationships, WordNet became an important expert-created source of language information. SC measures on WordNet mostly use its graph structure (i.e., the encoded relations) to determine the path length between concepts or the shared information content of concepts. We use the similarity measure proposed by Lin (Lin 1998) to define the correlations of two concepts c_1 and c_2 as

$$\text{sim}_{Lin}(c_1, c_2) = \frac{2 * IC(lcs)}{IC(c_1) + IC(c_2)}. \quad (2)$$

where lcs denotes the lowest common subsumer of the two concepts in the WordNet hierarchy (i.e., the lowest common hypernym) and *IC* denotes the information content of a concept. *IC* is computed as $IC(c) = logp(c)$ where $p(c)$ is the probability of encountering an instance of *c* in a corpus. The probability *p(c)* can be estimated from the relative corpus frequency of *c* and the probabilities of all concepts that *c* subsumes.

Wikipedia vector based SC measures. Wikipedia is the largest online collaboratively built encyclopedia, with more than 3 million articles for the English version. It contains pages for concepts and each page provides a detailed and human edited description of the corresponding concept. Inspired by the recent success of the skip-gram language model (Miller 1995; Mikolov et al. 2013), an efficient and effective method for learning high-quality vector representation of words from large amounts of unstructured text data, we adopt it into our problem. We take a text corpus, i.e. Wikipedia as the input and produces the word vectors as the output. It first constructs a vocabulary from the training text data and then learns vector representation of words. The output word vectors can be used to measure the correlations via cosine distances. The larger cosine score between two word vectors means that they are more semantically related.

The word representation using neural networks is very promising, because the learned vectors can explicitly encode the linguistic regularities and patterns. The training objective of skip-gram model is to find a word representation that is useful for predicting the surrounding words in a sentence. More generally, given a sequence of words $\{e_1, e_2, ..., e_T\}$, it searches for a vector representation for each word e_i, denoted by v_{e_i}, such that

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{-c \le j \le c, j \ne 0} \log(P(e_{t+j}|e_t)) \quad (3)$$

is maximized. c controls the training context size, and the probability of e_{t+j} given e_t is defined by the softmax function

$$P(e_i|e_j) = \frac{\exp(v_{e_i}^T v_{e_j})}{\sum_e \exp(v_{e_i}^T v_{e_j})} \quad (4)$$

This objective function attempts to make the vector representation of semantically close words behave similarity in predicting their contexts. In practice, a hierarchical softmax function (Mikolov et al. 2013) is used to make the training process computationally feasible. After trained on large-text corpus, the cosine distances between vectors of semantically similar words are larger.

Compared with the rule-based Wordnet approach, continuous word vector-based representation is both data-driven and flexible. Once word vectors are trained from independent corpus, one can measure the semantic correlations for an arbitrary set of words. In the experiment, we find that Wikipedia vector based representation can better measure the inter-class relationships between actions. This will be discussed in the experiment part.

Label transfer

Inspired by the probabilistic formulation of attribute-based DAP (Lampert, Nickisch, and Harmeling 2009) approach, our method can be defined as a modification of the attribute-based model that represents the unknown action class as a linear combination of a set of K semantically related known classes z_k. We formulate it as

$$p(y_u|x) = \sum_{k=1}^{K} p(y_u|y_k)p(y_k|x), \quad (5)$$

where $p(y_u|x)$ is what we want to estimate, the probability of the unseen action class u given the low-level feature x. $p(y_k|x)$ models the probability of the related action class k given low-level feature. $p(y_u|y_k)$ is the transition probability from the related known class k to the unseen class u.

For each related action class k, we obtain the probability $p(y_k|x)$ by applying the action detectors and taking the response value as $p(y_k|x) = {w_k}^T x$. Then each test video is represented by a vector of action detector responses $S = [s_1, s_2, ...s_N]$. Each dimension corresponds to a type of known class action detector. The conventional fully-supervised concept bank approach (Sadanand and Corso 2012) obtains the transition probability $p(y_u|y_k)$ by learning a weight vector from positive data and negative data, to recognize the target action.

Since we do not have positive data for the unseen action class, we can not apply the traditional concept bank approach to learn the weight of different concepts. Instead we utilize the semantic correlations computed from the external ontology as the weight to estimate a classifier for an unseen action class. Then our approach can be considered as estimating a weight vector w_u to recognize the novel action classes u, as

$$w_u = \sum_K \mathrm{w}_k s_{uk}. \quad (6)$$

where s_{uk} represents the semantic correlation between the know action action class k and the unseen action class u. $w_k \in R^d$ is the classifier of the known action class k, which can be off-line obtained. Then the estimated classifier w_u can be directly used for recognizing the unseen action class u.

Experiment

We present the dataset, experimental settings, evaluation criteria and the experimental results in this section.

Low-level Feature Representation

Trajectory features have been proved to be the most reliable features for action recognition and multimedia event recognition, which consist of five different descriptors (trajectory, HOG, HOF, MBHX and MBHY) to capture the shape and temporal motion information of videos. We adopt the improved trajectories proposed by (Wang, Schmid, and others 2013) to extract local features for each video in the UCF101 dataset. We use the default parameters, which results in 426 dimensions in total. Then the PCA operations are performed separately on each of the 5 descriptor types to keep half of the dimensions. After PCA, the local features reduce to 213 dimensions. Finally, each video is encoded in a Fisher Vector (Oneata et al. 2013) based on a GMM of 256 Gaussians, producing a 109056 dimensional vector.

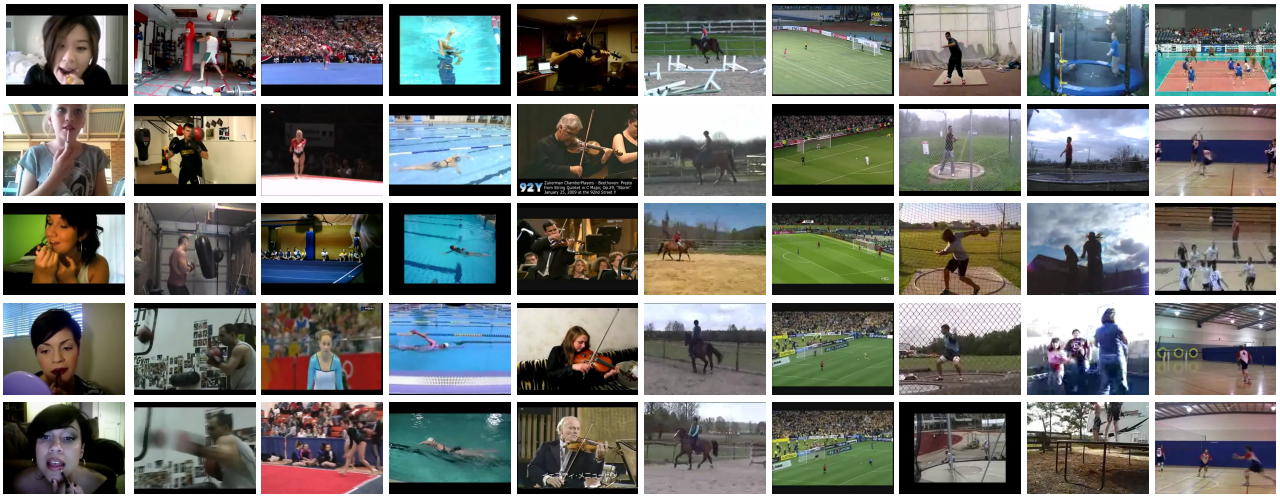
Dataset

To illustrate the effectiveness of the proposed approach, we do experiments on the largest action recognition dataset UCF101 (Soomro, Zamir, and Shah 2012). It consists of 101 action classes, over 13K clips and 27 hours of video data, which makes it much more diverse than other datasets for action recognition. The videos in UCF101 were downloaded from YouTube, containing poor lighting, cluttered background, and severe camera motion. Frames of example videos are shown in Figure 2. These videos have also been divided into five types: *human-object interaction, body motion only, human-human interaction, playing musical instruments* and *sports*. The reasons that we choose UCF101 as experimental dataset are as follows:

- As it is collected for YouTube, it contains real actions and poses significant challenges on action recognition.
- It contains nearly complete action classes in other action recognition datasets.
- It can be divided into different action types, which is suitable for our large-scale zero-shot learning task.

Implementation

For all the concept training, where least square regression used, we employ 5-fold cross validations for the value of λ. The search ranges of this parameter are $\lambda \in \{0.01, 0.1, 1, 10, 100\}$.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Figure 2: Figures (a) -(j) represent the highest ranking results for testing class *apply lipstick*, *boxing punching bag*, *floor gymnastics*, *front crawl*, *playing violin*, *horse riding*, *soccer penalty*, *throw discus*, *trampoline jumping* and *volleyball spiking* in the UCF101 dataset. Uniquely characterized classes are well identified, e.g. *apply lipstick* and *front crawl*. Confusions occur between visually similar classes, e.g. *floor gymnastics* and *trampoline jump*.

Action name	0 shot (our method)	1 shot (SVM)	1 shot (LR)	3 shots (SVM)	3 shots (LR)	5 shots (SVM)	5 shots (LR)
apply lipstick	**92.75%**	54.14%	66.67%	78.34%	84.52%	86.09%	89.22%
boxing punching bag	81.04%	69.45%	74.14%	82.08%	83.82%	87.93%	**93.21%**
floor gymnastics	**91.64%**	47.23%	53.19%	65.44%	71.78%	84.35%	88.76%
front crawl	**97.77%**	45.37%	53.51%	86.63%	95.36%	94.57%	96.71%
horse ride	**65.44%**	42.38%	45.89%	53.40%	57.42%	62.07%	63.35%
play violin	**74.55%**	43.46%	39.29%	55.09 %	43.32 %	58.65%	52.87%
soccer penalty	86.22%	54.78%	64.73%	79.88 %	83.92%	86.36%	**91.41%**
throw discus	**80.84%**	41.51%	38.51%	68.24%	58.24%	72.58%	72.77%
trampoline jump	88.74%	56.73%	65.42%	70.98 %	86.78%	87.46%	**91.24%**
volleyball spike	**58.69%**	38.42%	32.55%	44.18 %	47.82%	54.71%	56.77%
mAP	**81.77%**	49.35%	53.39%	68.23%	71.30%	77.48%	79.63%

Table 1: The mean Average Precision (mAP) comparisons with using few positive exemplars.

Experiment Setup

Since our goal is action recognition that no training exemplars are available, we split the labels of the UCF101 dataset into two disjoint sets: known classes and unknown classes. One set contains 91 action classes as known classes for training. The other set contains 10 action classes as unknown classes for testing. We apply Average Precision (AP) and mean Average Precision (mAP) as evaluation criteria.

For the consistent evaluation of zero-shot action recognition, we have selected 10 testing classes for public comparison: *apply lipstick, boxing punching bag, floor gymnastics, front crawl, horse riding, playing violin, soccer penalty, throw discus, trampoline jumping* and *volleyball spiking*. Thus our testing set consists of 1400 videos of those class actions, while the 12000 videos of the remaining 91 classes can be used for training. Additionally, we also encourage the use of the dataset for the regular complex large-scale zero-shot action recognition setting. In particular, we expect the splits of the UCF101 dataset to be suitable to test the performances of zero-shot action recognition, because the choice of testing classes covers different action types and some classes also look visual similar, which makes the zero-shot action recognition be difficult.

Furthermore, to make our results in a larger perspective, we also perform five trial experiments by randomly splitting the training classes and testing classes. The comparison results have also been reported in Table 2.

Baseline

Leveraging related source videos for action recognition without positive exemplars and attributes is so far an unexplored area. To the best of our knowledge, there is no directly

Approach	trial 1	trial 2	trial 3	trial 4	trial 5
0 shot (Our method)	0.7204	0.7124	0.8464	0.7724	0.8246
1 shot (SVM)	0.4518	0.4215	0.4962	0.4652	0.4974
1 shot (LR)	0.4752	0.4321	0.5218	0.5287	0.5296
2 shots (SVM)	0.5013	0.4895	0.6353	0.5657	0.5417
2 shots (LR)	0.5278	0.5496	0.6620	0.5958	0.5770
3 shots (SVM)	0.5398	0.5772	0.6932	0.6367	0.6459
3 shots (LR)	0.5557	0.6034	0.7296	0.6576	0.6776
4 shots (SVM)	0.5876	0.6259	0.7466	0.6952	0.7072
4 shots (LR)	0.6245	0.6464	0.7690	0.7014	0.7265
5 shots (SVM)	0.6478	0.6649	0.8065	0.7347	0.7865
5 shots (LR)	0.6942	0.6842	0.8186	0.7537	0.8143

Table 2: The mean Average Precision (mAP) of action recognition for 5 trials.

Evaluation Metric	WordNet	Wikipedia
mAP	63.41%	**81.77%**

Table 3: The mean Average Precision (mAP) of zero-shot action recognition using different SC approaches.

related algorithm to compare with. Typical attribute-based methods are not feasible in our large-scale action recognition setting. Thus, we compare the fully-supervised, i.e. support vector machines (SVM) and least square regression (LR) approaches using few positive exemplars (n-shot), where n means the number of positive exemplars available. For the compared algorithms, we randomly select 1, 2, 3, 4 and 5 videos as as positive data, and 5000 null videos (collected from Youtube, don't belong to any action class in UCF101) from development set as negative data to train a binary classifier. To be noted, the 5 positive data used in the 5-shot experiment are excluded from the testing set in all experiments, so the testing data are the same in each experiment. And we repeat experiments on 10 groups of randomly-generated training and testing set. The average mAP scores are then reported in Table 1 and Table 2.

Experiment Result

We first show the compared results for each action class between our proposed 0-shot and traditional n-shot (only using n positive exemplars and 5000 null videos) in the public setting. It can be found from Table 1 that the proposed method achieves the highest accuracies for 7 testing classes among the whole 10 classes when the number of the positive exemplars increases to be 5.

Then we also show the five trial experiment results by randomly splitting the training classes and testing classes in Table 2. The mAP scores of our method are between 0.8464 and 0.7124, and all beat the 5-shot settings. The experiment results validate the proposed approach can be an effective way to solve the ad-hoc action recognition problem, where few or even no positive samples available. And for n-shot experiments, we find least square regression (LR) can achieve better results than support vector machines (SVM).

Comparisons of different SC approaches

In this section, we compared the results between WordNet path length-based SC measures and Wikipedia vector based SC measures in Table 3. It can be seen that the Wikipedia vector based approach can achieve better results. The reasons lie in that the verbs don't have the well-defined hierarchical relationships, which make the traditional image-based automatic knowledge transfer approaches could hardly be applied in the action recognition setting. The proposed Wikipedia vector based semantic correlation approach improved the results significantly.

Conclusion

In this paper, we have proposed a novel approach by using semantic inter-class relationships for zero-shot action recognition. To the best of our knowledge, we are the first to perform zero-shot action recognition without positive exemplars and attributes. We leverage the pre-trained classifiers of known action classes and their semantic correlations with the novel action class for the label transfer. The proposed method is fully-automatic, which not only saves tedious human efforts, but also achieves promising performances for action recognition on UCF101 dataset. We consider the findings in this paper as a starting point for future research.

Acknowledgments

This work was supported by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the Natural Science Foundation of China Grant 61033001, 61361136003. This work was also supported by the open project program of the state key lab of CAD&CG (Grant No. A1402).

References

Akata, Z.; Perronnin, F.; Harchaoui, Z.; Schmid, C.; et al. 2013. Label-embedding for attribute-based classification. In *CVPR*, 819–826.

Blank, M.; Gorelick, L.; Shechtman, E.; Irani, M.; and Basri, R. 2005. Actions as space-time shapes. In *ICCV*, volume 2, 1395–1402.

Chen, M.-y., and Hauptmann, A. 2009. Mosift: Recognizing human actions in surveillance videos.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. In *CVPR*, 1778–1785.

Feng, J.; Jegelka, S.; Yan, S.; and Darrell, T. 2013. Learning scalable discriminative dictionary with sample relatedness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1645–1652.

Gan, C.; Qin, Z.; Xu, J.; and Wan, T. 2013. Salient object detection in image sequences via spatial-temporal cue. In *Visual Communications and Image Processing (VCIP), 2013*, 1–6. IEEE.

Guadarrama, S.; Krishnamoorthy, N.; Malkarnenkar, G.; Venugopalan, S.; Mooney, R.; Darrell, T.; and Saenko, K. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2712–2719. IEEE.

Jiang, L.; Mitamura, T.; Yu, S.-I.; and Hauptmann, A. G. 2014. Zero-example event search using multimodal pseudo relevance feedback. In *International Conference on Multimedia Retrieval*, 297. ACM.

Kankuekul, P.; Kawewong, A.; Tangruamsub, S.; and Hasegawa, O. 2012. Online incremental attribute-based zero-shot learning. In *CVPR*, 3657–3664.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 951–958.

Laptev, I.; Marszalek, M.; Schmid, C.; and Rozenfeld, B. 2008. Learning realistic human actions from movies. In *CVPR*, 1–8.

Larochelle, H.; Erhan, D.; and Bengio, Y. 2008. Zero-data learning of new tasks. In *AAAI*, 646–651.

Lin, D.; Fidler, S.; Kong, C.; and Urtasun, R. Visual semantic search: Retrieving videos via complex textual queries.

Lin, D. 1998. An information-theoretic definition of similarity. In *ICML*, volume 98, 296–304.

Liu, J.; Yu, Q.; Javed, O.; Ali, S.; Tamrakar, A.; Divakaran, A.; Cheng, H.; and Sawhney, H. S. 2013. Video event recognition using concept attributes. In *WACV*, 339–346.

Liu, J.; Kuipers, B.; and Savarese, S. 2011. Recognizing human actions by attributes. In *CVPR*, 3337–3344.

Ma, Z.; Yang, Y.; Sebe, N.; Zheng, K.; and Hauptmann, A. G. 2013. Multimedia event detection using a classifier-specific intermediate representation. *IEEE Transactions on Multimedia* 15(7):1628–1637.

Mensink, T.; Verbeek, J.; Perronnin, F.; and Csurka, G. 2013. Distance-based image classification: Generalizing to new classes at near zero cost.

Mensink, T.; Gavves, E.; and Snoek, C. G. M. 2014. Costa: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 3111–3119.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Oneata, D.; Verbeek, J.; Schmid, C.; et al. 2013. Action and event recognition with fisher vectors on a compact feature set. In *ICCV*.

Parikh, D., and Grauman, K. 2011. Relative attributes. In *ICCV*, 503–510.

Reddy, K. K., and Shah, M. 2013. Recognizing 50 human action categories of web videos. *Machine Vision and Applications* 24(5):971–981.

Rohrbach, M.; Stark, M.; Szarvas, G.; Gurevych, I.; and Schiele, B. 2010. What helps where–and why? semantic relatedness for knowledge transfer. In *CVPR*, 910–917.

Rohrbach, M.; Stark, M.; and Schiele, B. 2011. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, 1641–1648.

Sadanand, S., and Corso, J. J. 2012. Action bank: A high-level representation of activity in video. In *CVPR*, 1234–1241.

Sharmanska, V.; Quadrianto, N.; and Lampert, C. H. 2012. Augmented attribute representations. In *ECCV*. 242–255.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Wang, H.; Ullah, M. M.; Klaser, A.; Laptev, I.; Schmid, C.; et al. 2009. Evaluation of local spatio-temporal features for action recognition. In *BMVC*.

Wang, H.; Klaser, A.; Schmid, C.; and Liu, C.-L. 2011. Action recognition by dense trajectories. In *CVPR*, 3169–3176.

Wang, H.; Schmid, C.; et al. 2013. Action recognition with improved trajectories. In *ICCV*.

Yu, F. X.; Cao, L.; Feris, R. S.; Smith, J. R.; and Chang, S.-F. 2013. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 771–778.