Alternative Semantic Representations for Zero-Shot Human Action Recognition

Qian Wang (✉) and Ke Chen

School of Computer Science, The University of Manchester,
Manchester, M13 9PL, UK
{qian.wang,ke.chen}@manchester.ac.uk

Abstract. A proper semantic representation for encoding side information is key to the success of zero-shot learning. In this paper, we explore two alternative semantic representations especially for zero-shot human action recognition: textual descriptions of human actions and deep features extracted from still images relevant to human actions. Such side information are accessible on Web with little cost, which paves a new way in gaining side information for large-scale zero-shot human action recognition. We investigate different encoding methods to generate semantic representations for human actions from such side information. Based on our zero-shot visual recognition method, we conducted experiments on UCF101 and HMDB51 to evaluate two proposed semantic representations . The results suggest that our proposed text- and image-based semantic representations outperform traditional attributes and word vectors considerably for zero-shot human action recognition. In particular, the image-based semantic representations yield the favourable performance even though the representation is extracted from a small number of images per class.

Keywords: Zero-Shot Learning, Semantic Representation, Human Action Recognition, Image Deep Representation, Textual Description Representation, Fisher Vector

1 Introduction

Zero-Shot Learning (ZSL) aims to recognize instances from new classes which are not seen in the training data. It is a promising alternative to the traditional supervised learning which requires labour-intensive annotation work on all the classes involved. As shown in Figure 1, in ZSL, the knowledge learned from training data is transferred to recognise unseen classes through the side information which can usually be acquired with less effort. Although most existing works in ZSL focus on the development of novel recognition models, the side information for knowledge transfer plays an equally important role in the success of ZSL. The most popular side information used in ZSL literature are attributes and word vectors. Although they have been widely used in ZSL [9,12,14,26,27], both of them have obvious drawbacks as well, especially for zero-shot human action recognition in video data.

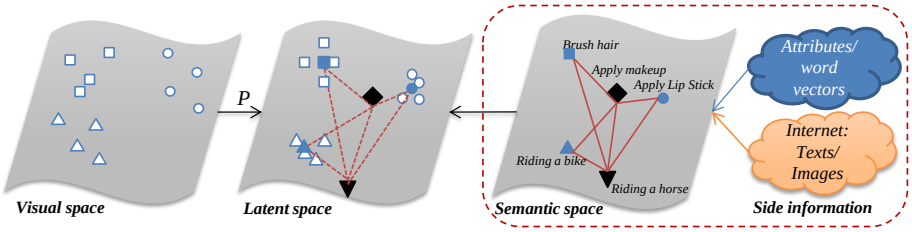

Fig. 1. A schematic diagram of zero-shot learning framework. The work in this paper is highlighted in the dashed box. Human action classes are denoted by coloured markers (blue and black for training and unseen classes respectively) with different shapes. The training data are used to learn the mapping P and training class embedding (blue filled markers in the latent space), then the unseen class embedding (black filled markers in the latent space) is achieved by preserving the semantic distances (red lines). See Section 4.2 for more details of our ZSL method.

The definition and annotation of attributes for human actions (e.g., the attributes defined for UCF101 [10] include "bodyparts-visible: face, fullbody, one-hand", "body-motion: flipping, walking, diving, bending", etc.) are subjective and labour-intensive. When a large number of human actions are involved, more attributes are needed to distinguish one human action from the other. As a result, attributes based semantic representations are inappropriate for large scale zero-shot human action recognition. On the other hand, as stated in [2], using a word vector of the class label to represent a human action is far from adequate to illustrate the rich appearance variations. In addition, the word vectors are learned from textual corpus, thus suffering from the catastrophic semantic gap problem (i.e., the difference of information conveyed by visual media and texts).

To address the limitations of existing semantic representations for ZSL, we attempt to explore alternative side information towards enhanced zero-shot human action recognition. The essentials of side information for ZSL are twofold. Firstly, it should be achievable for a large number of human actions without much effort. More importantly, the side information should be able to capture the visually discriminative semantics thus benefiting the ZSL by easily bridging the semantic gap. To this end, we employ action relevant images as the side information resources to extract the semantics of human actions. With the aid of search engines, it is effortless to collect a set of action relevant images by using the action name as the key words. Although still images lack of temporal information in human actions, they provide abundant visually discriminative information which can be exploited to extract high-level semantic representations for human actions. On the other hand, we aim to enhance the word vectors by collecting and encoding textual descriptions of human actions. We believe that the contextual information in the action relevant texts (e.g., description articles of human actions from the web) will remove the ambiguity of the semantics in the original action word vectors which are based solely on the action labels.

To summarise, the contributions of this paper include:

Table 1. A survey on semantic representations in ZSL

Authors and Year	Semantic Representation
Lampert *et al.* (2009) [12]	Attributes, annotated manually
Sharmanska *et al.* (2012) [21]	Attributes, enhanced by learning from visual data
Liu *et al.* (2011) [14]	Attributes, enhanced by learning from visual data
Qin *et al.* (2016) [18]	Attributes, enhanced by learning from visual data
Fu *et al.* (2014) [8]	Attributes, enhanced by learning from visual data
Inoue *et al.* (2016) [9]	Word vector, enhanced by a weighted combination of related word (from *WordNet*) vectors.
Alexiou *et al.* (2016) [2]	Word vector, enhanced by the synonyms of labels (from Internet dictionaries)
Mukherjee *et al.* (2016) [16]	Word Gaussian distribution
Sandouk *et al.* (2016) [20]	Word vector, enhanced by contexts (from tags)
Elhoseiny *et al.* (2013) [6]	Tf-idf, based on *Wikipedia* articles
Akata *et al.* (2016) [1]	BOW, based on *Wikipedia* articles
Rohrbach *et al.* (2010) [19]	*WordNet* path length, based on *WordNet* ontology Hit-counts, based on web search results
Chuang *et al.* (2015) [5]	*WordNet* path length, based on *WordNet* ontology

- We propose and implement the idea of using textual descriptions to enhance the word vector representations of human actions in ZSL.
- We propose and implement the idea of using action related still images to represent semantics for video based human actions in ZSL.
- Experiments are conducted to evaluate the effectiveness of the proposed semantic representations in zero-shot human action recognition, and significant performance improvement has been achieved.

2 Related Work

The semantic representation is key for the success of ZSL. Recently, attempts have been made to explore more effective semantic representations for objects/actions towards improved ZSL performance. In this section, we will review the prevailing semantic representations used in ZSL (Table 1), including a variety of extensions of attributes and word vectors, as well as many other less popular approaches proposed in literature.

Attributes based semantic representations were firstly proposed for ZSL in [12], thereafter, attributes have been employed for ZSL in many works [3, 26, 27, 28]. A set of binary attributes need to be manually defined to represent the semantic properties of objects. As a result, each object class can be represented by a binary attribute vector in which the value of one and zero indicates the presence and absence of each attribute respectively. Since the attributes are shared by seen and unseen classes, the knowledge transfer is enabled. However, as mentioned above, the definition of attributes require experts with domain knowledge to discriminate different classes, and the attribute annotation for a large number of classes could be subjective and labour-intensive.

Alternatively, attributes can be mined automatically from visual features by discriminative mid-level feature learning [7, 8, 14, 18, 21], but their semantic meanings are unknown, thus inappropriate for direct use in ZSL. To enhance the attributes' discriminative power and semantic meaningfulness, the manually defined attributes and the ones automatically learned from training data are usually combined. However, the data-driven attributes are usually dataset specific and probably fail on a different dataset.

The other kind of prevailing side information used in ZSL is derived from text resources. One of the most popular semantic representations is word vector (e.g., the ones generated by the *word2vec* tool [15]) due to its convenience and effectiveness. A class label can be easily represented with the vector representation of the corresponding word or phrase. However, word vectors are deficient to discriminate different classes from the visual perspective due to the semantic gap, i.e., the gap between visual and semantic information. As a result, word vectors are usually outperformed by attributes in ZSL.

To alleviate the semantic gap problem, some attempts have been made to enhance the word vectors [2, 9, 16, 20]. Inoue *et al.* [9] aim to adapt the original word vectors to make two visually similar concepts close to each other in the adapted word vector space by representing a concept with a weighted sum of its original word vector and its hypernym (based on *WordNet*) word vectors. And the weights are learned from visual resources. Alexiou *et al.* [2] enrich the word vector representation by mining and considering synonyms of the action class labels from multiple *Internet dictionaries*. Mukherjee *et al.* [16] use *Gaussian distribution* instead of a single word vector to model the class labels so that the intra-class variability can be expressed properly in the semantic representations. To address the issue of polysemy, Sandouk *et al.* [20] learn a specific vector representation for a word together with its context. That is to say, the same word could have different vector representations when it is in different contexts. Inspired by these works, our work further investigates the possible side information and enabling techniques to enhance the word vectors for ZSL.

Other than attributes and word vectors, other side information has also been investigated for knowledge transfer in ZSL, only if they are able to model the relationships among different classes and relatively easy to obtain. For example, *WordNet* path length is used to measure the semantic correlations between two concepts in [5, 19]. The Internet together with search engines provides a natural opportunity to get side information to measure between-class semantic relationships based on hit-count on search results [19]. Textual descriptions of a class rather than the single class name are employed to represent a class in [1, 6]. Concept related textual descriptions (e.g., *Wikipedia* page) can be readily obtained from the Internet and then processed with techniques in natural language processing (NLP). Considering our focus on zero-shot human action recognition based on video data, images from the Internet can be alternative side information to texts which have been a typical choice for zero-shot image classification.

3 Method

In this section, we propose our methods of generating semantic representations for zero-shot human action recognition from text and image resources respectively. Firstly, we use search engines to collect action relevant texts and images as the side information. Some typical examples are shown in Fig.2. Once the side information are collected, we use different encoding approaches to generate the semantic representations for human actions.

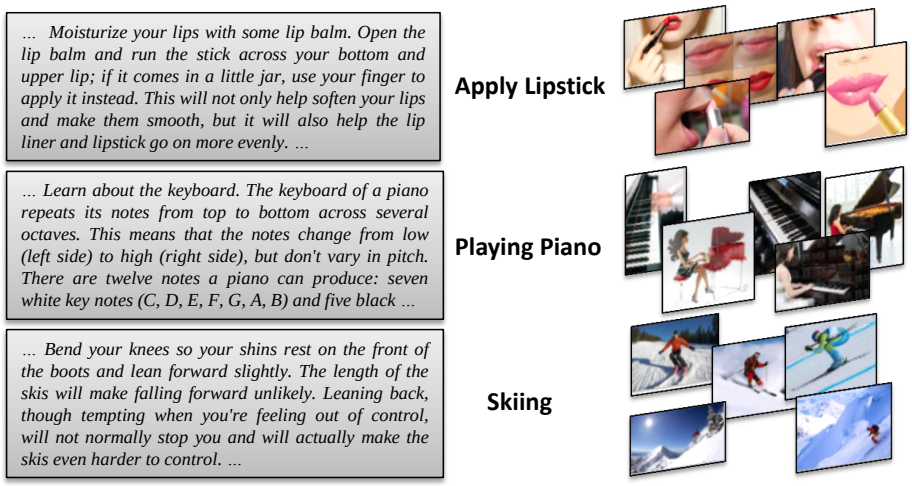

Fig. 2. Examples of collected description texts and images of three human actions from UCF101 (i.e., "Apply Lipstick", "Playing Piano" and "Skiing").

3.1 Text-based Semantic Representation

Texts Collection Motivated by the fact a class label is insufficient to depict the complex concepts in the human action, we try to collect textual descriptions from the web to represent each human action. Textual descriptions of human actions can be derived from *WikiHow*, a website teaching people "how to do anything". Inevitably, the description texts for some actions (e.g., "pick", "sit") are not available from *WikiHow*, for which we turn to alternative sources including *Wikipedia* and *Online dictionary*.

Pre-Processing Once the textual descriptions for all the human actions are collected, we end up with a document for each human action class. We use natural language processing techniques to pre-process the unstructured textual data before encoding them into semantic representations. In the first step, we tokenize the documents to get all the words appearing in the documents. After

removing the stop words (i.e., the words carrying little semantic meanings such as "is", "you", "of"), we have a dictionary containing d words.

Term-Document Matrix (TD) Given the documents and the dictionary containing all the terms/words in the documents, a term-document matrix M is constructed to represent the term frequency in all documents. M_{ij} denotes the frequency of term i in document j, where $i = 1, 2, ..., d$ and $j = 1, 2, ..., C$, C is the number of documents, i.e., the number of human actions in a specific dataset. Thus the column vectors in M can be used to represent the semantic representations of human actions. We denote this approach as **TD** in the following sections.

Average Word Vector (AWV) We aim to enhance the word vectors by incorporating the collected textual information. Taking advantage of the compositional property of word vectors, we can represent a document with the average of all the included word vectors.

$$AWV(j) = \frac{1}{n_j} \sum_{i=1}^{n_j} v_i \tag{1}$$

where n_j is the number of terms in the j-th document, $v_i \in \mathbb{R}^D$ denotes the word vector of the i-th term in the document, and D is the dimensionality of word vectors.

Fisher Word Vector (FWV) In contrast to AWV using the mean of all word vectors to represent a document, FWV aims to model the distribution of word vectors in a document. Fisher Vector represents a document (i.e., a set of words) by the gradient of log likelihood with respect to the parameters of a prelearned probabilistic model (i.e., Gaussian Mixture Model) [17, 25]. A Gaussian Mixture Model (GMM) is used to fit the distribution of the word vectors involved in all documents, where the parameters $\Theta = \{\mu_k, \Sigma_k, \pi_k\}, k = 1, ..., K$. Let $V^j = \{v_1, ..., v_{n_j}\}$ be a set of word vectors from the j-th human action description document. Then the Fisher Vector of j-th document can be denoted by:

$$FWV(j) = [\mathcal{G}^{V^j}_{\mu,1}, ..., \mathcal{G}^{V^j}_{\mu,K}, \mathcal{G}^{V^j}_{\sigma,1}, ..., \mathcal{G}^{V^j}_{\sigma,K}], \tag{2}$$

where

$$\mathcal{G}^{V^j}_{\mu,k} = \frac{1}{\sqrt{\pi_k}} \sum_{v_i \in V^j} \gamma_{k,i} \left(\frac{v_i - \mu_k}{\sigma_k}\right), \tag{3}$$

$$\mathcal{G}^{V^j}_{\sigma,k} = \frac{1}{\sqrt{2\pi_k}} \sum_{v_i \in V^j} \gamma_{k,i} \left(\frac{(v_i - \mu_k)^2}{\sigma_k^2} - 1\right), \tag{4}$$

$$\gamma_{k,i} = \frac{exp[-\frac{1}{2}(v_i - \mu_k)^T \Sigma_k^{-1}(v_i - \mu_k)]}{\Sigma_{t=1}^{K} exp[-\frac{1}{2}(v_i - \mu_t)^T \Sigma_k^{-1}(v_i - \mu_t)]}. \tag{5}$$

The dimension of the Fisher Vector is $2DK$, where D and K are the dimensionality of word vectors and the number of components in the GMM respectively.

3.2 Image-based Semantic Representation

Human actions are difficult to describe with texts due to the complexity and intra-class variations. Although they lack temporal information, still images can provide abundant information for the understanding of human actions. Compared to the video examples, still images are much easier to collect, annotate and store. Thus we hold the view that still images are a proper kind of side information which can benefit modelling human action relationships with little effort.

Image Collection Given a human action, we use the label as the key word and search relevant images with search engines. For most human actions we can get a collection of images each of which gives a view of the action. However, for some action names which could have multiple meanings, the additional explaining key words are needed to get reasonable searching results. For example, we use "salsa spin + dancing" and "playing + hula hoop" for the actions "salsa spin" and "hula hoop" respectively. For each human action, we get different numbers of relevant images after removing the ones of poor quality (e.g., irrelevant ones and the ones smaller than 10Kb) from the returned results. The image collection and filtering can be processed automatically without many human interventions [1].

Feature Extraction We aim to extract useful information from a set of images to represent a human action. Recently, deep convolution neural networks have been used to extract image features carrying high-level conceptual information. By feeding the images into a pre-trained CNN model, the deep image features can be obtained easily. Then each human action is represented with a set of image feature vectors $F^j = \{f_1, ..., f_{n_j}\}$. In the next two sections, we use two approaches to encode the set of image features into the action-level semantic representation.

Average Feature Vector (AFV) Similar to Eq.(1), we can use the average of multiple image features as the human action semantic representation.

$$AFV(j) = \frac{1}{n_j} \sum_{i=1}^{n_j} f_i \tag{6}$$

Fisher Feature Vector (FFV) Similar to the processing applied on word vectors in Section 3.1, we use Fisher Vector to encode a set of image feature vectors relevant to a specific human action.

$$FFV(j) = [\mathcal{G}_{\mu,1}^{F^j}, ..., \mathcal{G}_{\mu,K}^{F^j}, \mathcal{G}_{\sigma,1}^{F^j}, ..., \mathcal{G}_{\sigma,K}^{F^j}], \tag{7}$$

where $\mathcal{G}_{\mu,i}^{F^j}$ and $\mathcal{G}_{\sigma,i}^{F^j}$ can be calculated in the same way as Eq.(3-5).

[1] The image scraper tool is available: http://staff.cs.manchester.ac.uk/ kechen/ASRHAR/

4 Experimental Settings

4.1 Dataset

We use two human action datasets to evaluate the proposed approaches for zero-shot recognition, i.e., UCF101 [22] and HMDB51 [11]. **UCF101** is a human action recognition dataset collected from YouTube. There are 13,320 real action video clips falling into 101 action categories. In our experiments, we use 5 randomly generated 51/50 (seen/unseen) class-wise data splits. **HMDB51** contains 6,766 video clips from 51 human action classes. Similarly, we use 5 randomly generated 26/25 splits in all experiments.

4.2 Zero-Shot Recognition Method

We employ our recently developed ZSL method, bidirectional latent embedding learning (BiDiLEL) [26], as a test bed in our experiments [2]. To make the paper self-contained, we will briefly describe the main idea of BiDiLEL in this section.

The method employs a two-stage latent embedding algorithm to learn a latent space in which the semantic gap is bridged and zero-shot recognition can be done (see Fig.1). In bottom-up stage, we learn a projection matrix P by supervised locality preserving projection (SLPP) [4], such that the examples close to each other in the original visual space will still be close in the latent space. By exploiting the local structures and labelling information in the training data, the learned latent space preserves the data distribution and is more discriminative. The properties are expected to generalise well for test examples from unseen classes.

In the top-down embedding, the latent embedding of each seen class can be calculated by averaging the projections of all the training examples from the class and then serve as landmarks guiding the learning of latent embedding of unseen classes. We use the landmarks based Sammon mapping (LSM) [26] which aims to preserve the inter-class semantic distances (measured in the semantic space). As a result, the semantic distances between seen and unseen classes as well as between any pair of unseen classes will be preserved in the latent space.

Once the latent embedding of both seen and unseen classes are obtained, we can do the zero-shot learning in the latent space using the nearest neighbour method. Specifically, given a test example, we use projection matrix P to map it into the latent space, where its distances to all the class embedding can be calculated, and it will be assigned to the closest class label. For more details, we refer the readers to [26].

4.3 Video Representation

C3D was proposed in [24] for human action recognition. It utilizes 3D ConvNets to learn spatio-temporal features for video streams. According to [26],

[2] Like attributes and word vectors, our proposed semantic representations may be directly deployed in all the existing zero-shot human action recognition methods.

the C3D video representation outperforms its counterparts in zero-shot human action recognition. We use the model pre-trained on Sports-1M dataset and follow the setting in [24, 26] to extract spatio-temporal deep features (i.e., the 4096-dimensional "fc6" activations of the deep neural network) from 16-frame segments. Finally, the visual representation of a video stream is calculated by averaging the features of all the segments from the video.

4.4 Evaluation

In most existing ZSL works, the evaluations are based on the assumption that test examples are only from unseen classes, which is often referred as to conventional zero-short learning (cZSL). In practice, however, the test examples can be from either training classes or unseen classes. To evaluate ZSL methods in a more practical scenario, the problem of generalised ZSL has been formulated and investigated in [3, 27]. In gZSL, given a test example, the label search space consists of both seen and unseen classes. In our experiments, we follow the protocols in [27] and report both conventional and generalised ZSL (cZSL and gZSL) results using per-class accuracy. In the generalised ZSL scenarios, except the examples from test classes, we also reserve 20% examples from each training class for testing and the rest 80% examples from each training class for training.

Concretely, we report the recognition accuracy of test examples from unseen classes by setting the search space in the unseen label set $\mathcal{U}$ for the cZSL; the accuracy is denoted by $A_{\mathcal{U}\to\mathcal{U}}$. For gZSL, we set the search space in the whole label set $\mathcal{T} = \mathcal{S} \cup \mathcal{U}$ and report three types of per-class accuracies, i.e., the recognition accuracy of test examples from unseen classes $A_{\mathcal{U}\to\mathcal{T}}$, the recognition accuracy of test examples from seen classes $A_{\mathcal{S}\to\mathcal{T}}$ and the harmonic mean,

$$H = 2 * A_{\mathcal{U}\to\mathcal{T}} * A_{\mathcal{S}\to\mathcal{T}} / (A_{\mathcal{U}\to\mathcal{T}} + A_{\mathcal{S}\to\mathcal{T}}). \tag{8}$$

The ZSL method employed in our experiments works in the inductive setting (i.e., the test example is processed individually), but can be extended to the transductive setting (i.e., all the test examples are assumed to be available as a collection when doing the recognition) easily by using the structured prediction method [26, 28]. The method of structure prediction uses Kmeans to group all the test examples into clusters (the number of clusters is set to be the number of unseen classes) and find a one-to-one map from the clusters to unseen classes. In our experiments, we will report the results of cZSL in both inductive and transductive settings.

5 Experimental Results

In this section, we present the designed experiments and the results to evaluate the effectiveness of proposed semantic representations [3].

[3] The scripts and data used in our experiments can be available on our project page: http://staff.cs.manchester.ac.uk/ kechen/ASRHAR/

Table 2. Results of different text-based semantic representations (mean±standard error of recognition accuracy %) on UCF101 and HMDB51 datasets. (Sem.Rep.–Semantic Representation, Att–Attributes, WV–Word vector)

Sem. Rep.	UCF101 (51/50)		HMDB51 (26/25)	
	Inductive	**Transductive**	**Inductive**	**Transductive**
Random	2.00	2.00	4.00	4.00
Att	21.54 ± 0.72	$\mathbf{32.00 \pm 2.30}$	-	-
WV	19.42 ± 0.69	22.05 ± 1.74	21.53 ± 1.75	24.14 ± 3.43
TD	19.54 ± 0.75	24.29 ± 0.65	15.26 ± 0.57	15.33 ± 1.72
AWV	$\mathbf{24.38 \pm 1.00}$	30.60 ± 2.67	$\mathbf{21.80 \pm 0.87}$	$\mathbf{26.13 \pm 1.29}$
FWV(K=1)	23.76 ± 0.72	28.54 ± 0.70	19.57 ± 1.21	20.41 ± 1.74
FWV(K=2)	23.61 ± 1.08	28.64 ± 1.45	18.80 ± 1.22	20.01 ± 1.74
FWV(K=3)	22.21 ± 0.96	24.33 ± 2.34	17.35 ± 1.93	21.37 ± 3.16
FWV(K=4)	22.11 ± 0.62	28.76 ± 1.03	17.07 ± 1.41	18.80 ± 2.95
FWV(K=5)	21.50 ± 0.67	27.56 ± 2.43	16.95 ± 1.19	17.20 ± 1.92

5.1 Text-based Representation

We conduct experiments of zero-shot human action recognition by utilising the proposed text-based semantic representations in Section 3.1, i.e., TD, AWV and FWV. We use the 300-dimensional word vectors pre-trained with *word2vec* on Google News dataset (about 100 billion words) [4]. For FWV, we set the value of K in Eq.(2) to be $\{1, 2, 3, 4, 5\}$. The experiments aim to investigate how different text-based semantic representations perform in zero-shot human action recognition. In our experiments, we follow the protocols in [26] using class-wise cross validation to find the optimal values of hyper-parameters. According to the performance on the validation data, cosine distances are employed to calculate the semantic distances for FWV, and Euclidean distances are employed for AWV.

We report the results of conventional ZSL in both inductive and transductive settings in Table 2. With only the textual description sources, the simple encoding method TD can achieve the accuracy of 19.54% and 15.26% respectively on UCF101 and HMDB51, which indicates the textual descriptions collected by search engines are useful for modelling the inter-class relationships. By incorporating the pre-trained word vectors, AWV improves the accuracy to 24.38% and 21.80% respectively on UCF101 and HMDB51. On the other hand, by comparing FWV with different K values, we know that $K = 1$ gives the best results with an accuracy of 23.76% on UCF101 and 19.57% on HMDB51; however, it is still outperformed by AWV on both datasets regardless of inductive or transductive settings. To conclude, AWV performs the best among different text-based semantic representations.

5.2 Image-based Representation

In our experiments, we collect variant numbers of relevant images for different human actions. The average number of relevant images per class is around 200

[4] https://code.google.com/p/word2vec/

Table 3. Results of different image-based semantic representations (mean±standard error of recognition accuracy %) on UCF101 and HMDB51 datasets.

Sem. Rep.	UCF101 (51/50)		HMDB51 (26/25)	
	Inductive	**Transductive**	**Inductive**	**Transductive**
Random	2.00	2.00	4.00	4.00
AFV	37.24 ± 0.89	50.48 ± 1.35	25.55 ± 1.66	30.77 ± 3.23
FFV(K=1)	$\mathbf{40.12 \pm 1.30}$	$\mathbf{50.67 \pm 2.45}$	$\mathbf{25.82 \pm 1.19}$	$\mathbf{31.51 \pm 1.67}$
FFV(K=2)	38.01 ± 1.58	49.60 ± 1.82	25.50 ± 0.95	28.98 ± 1.94
FFV(K=3)	36.52 ± 1.38	45.48 ± 0.73	24.27 ± 1.10	26.95 ± 3.38
FFV(K=4)	35.31 ± 1.17	44.76 ± 2.40	23.22 ± 1.25	25.26 ± 2.32
FFV(K=5)	34.98 ± 0.68	45.08 ± 1.82	23.09 ± 1.12	23.93 ± 2.06

and 100 for UCF101 and HMDB51 respectively. To extract the image features, we use the GoogLeNet [23] model pre-trained on ImageNet dataset [5]. The activations of top fully connected layer of GoogLeNet of 1024 dimensions are used as the deep image features. We evaluate the image-based semantic representations encoded with different approaches described in Section 3.2, i.e., AFV and FFV. Again, we set the values of K in Eq.(7) to be $\{1, 2, 3, 4, 5\}$. We employ the same experiment protocols as those used in the previous experiments (Section 5.1). According to the performance on the validation data, cosine distances are employed to model the semantic distances for FFV, and Euclidean distances are employed for AFV.

The experimental results are shown in Table 3. Apparently, $K = 1$ again gives the best performance of FFV, achieving 40.12% and 25.82% respectively on UCF101 and HMDB51 in the inductive setting, 50.67% and 31.51% respectively on UCF101 and HMDB51 in the transductive setting. Different from the text-based semantic representations, image-based semantic representations FFV encoded by Fisher Vector outperforms the AFV on both datasets.

5.3 Comparison with Other Semantic Representations

In this experiment, we compare the proposed semantic representations with other popular ones. From Table 2 and 3, we know that AWV and FFV(K=1) perform the best among the text- and image-based semantic representations respectively. So we consider AWV and FFV(K=1) as the representatives of the proposed text- and image-based semantic representations. As described in Section 4.4, we conduct the experiments in both conventional and generalised ZSL scenarios in our experiments.

We present the experimental results in Table 4. Clearly, the proposed two semantic representations (i.e., AWV and FFV(K=1)) outperform word vectors and attributes consistently in terms of the conventional ZSL evaluation metric. On UCF101, the use of textual information enhances the word vectors based solely on the action labels by lifting the accuracy from 19.42% to 24.38%, even higher

[5] http://www.vlfeat.org/matconvnet/

Table 4. A comparison of different semantic representations on UCF101 and HMDB51 datasets (mean±standard error)%.

Dataset	Sem. Rep.	cZSL	gZSL		
		$A_{\mathcal{U}\to\mathcal{U}}$	$A_{\mathcal{U}\to\mathcal{T}}$	$A_{\mathcal{S}\to\mathcal{T}}$	H
UCF101	Random	2.00	1.00	1.00	1.00
	WV	19.42 ± 0.69	4.54 ± 0.64	84.79 ± 0.91	8.59 ± 1.17
	Att	21.54 ± 0.72	2.48 ± 0.62	86.39 ± 1.37	4.78 ± 1.18
	AWV	24.38 ± 1.00	5.32 ± 1.53	$\mathbf{86.43 \pm 1.06}$	9.85 ± 2.66
	FFV	$\mathbf{40.12 \pm 1.30}$	$\mathbf{16.55 \pm 1.30}$	82.38 ± 1.17	$\mathbf{27.49 \pm 1.86}$
HMDB51	Random	4.00	2.00	2.00	2.00
	WV	21.53 ± 1.75	2.64 ± 0.33	58.70 ± 1.40	5.05 ± 0.61
	AWV	21.80 ± 0.87	2.99 ± 0.35	$\mathbf{62.00 \pm 2.57}$	5.69 ± 0.64
	FFV	$\mathbf{25.68 \pm 1.07}$	$\mathbf{5.91 \pm 0.90}$	58.57 ± 1.50	$\mathbf{10.65 \pm 1.48}$

than that of labour-intensive attributes (21.54%). The image-based semantic representation FFV encoded with Fisher Vector gives the best accuracy of 40.12%, significantly higher than its counterparts. This is attributed to the narrower semantic gap between video representation space and image-based semantic space. The still images contain abundant visually discriminative information which can be further encoded into high-level semantic representations of human actions. On HMDB51, the same conclusions can be drawn. It is noteworthy that AWV is only slightly better than WV for HMDB51 dataset. The reason might be the existence of actions which are difficult to describe with texts in this dataset, such as, "sit", "talk", "turn", "stand", "pick", "catch", and etc.

Regarding the generalised ZSL scenario, the proposed AWV and FFV perform better on the test examples from unseen classes (with 5.32% and 16.55% respectively on UCF101, 2.99% and 5.91% respectively on HMDB51), outperforming the attributes and word vectors. We also notice that FFV does not perform the best on test examples from seen classes (i.e., $A_{\mathcal{S}\to\mathcal{T}}$), although it is significantly better than others in terms of harmonic mean (H). This is reasonable and practically preferable with the trade-off between recognition accuracy of examples from seen and unseen classes.

5.4 How Many Images Are Enough?

In the previous experiments, we use all the collected images to encode the image-based semantic representations. In this experiment, we investigate how the number of images affects the encoded semantic representations. We use AFV and FFV(K=1) as the encoding methods and generate the semantic representations for each human action with the number of relevant images to be 5, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 respectively (For the case when the total number of collected images for one human action is less than the expected number, we simply use all the collected images of that action in the experiment). The experiments are conducted on two human action datasets in conventional ZSL scenario under both inductive and transductive settings.

The performances of two types of image-based semantic representations with different numbers of images are shown in Fig.3. For a direct comparison, we display the baseline performance of attributes and word vectors in the figure as well. Using more images usually benefits the performance of AFV and FFV on both datasets. In specific, we can see a dramatic performance boost with the number of images increased from 5 to 40 per class for UCF101. A further increase of images does not improve the performance significantly, which is especially true in the inductive setting. For HMDB51 dataset, the similar trend of performance improvement can be observed from Fig.3, and the performance improvement stops until the number of images per class increases to around 80. In addition, the proposed image-based semantic representations using only 5 images per class can achieve better performance on UCF101 than attributes and word vectors, and the number rises to 20 for HMDB51 to beat word vectors. To summarise, we are able to use a small number of relevant images to encode the semantic representations of human actions, yet boosting the zero-shot human action recognition accuracy to a large extent.

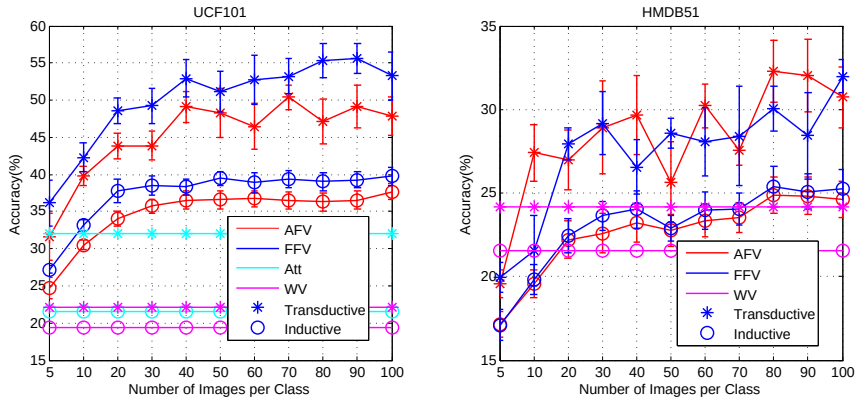

Fig. 3. Effects of number of images on the performance of AFV and FFV(K=1).

6 Conclusions and Future Work

We explore the alternative side information to the existing attributes and word vectors towards improved zero-shot human action recognition. The textual descriptions of human actions from the Internet can be used as side information for knowledge transfer in ZSL. In addition, the combination with pre-trained word vectors can further improve the power of text-based semantic representations, even better than the manually annotated attributes. On the other hand, the image-based semantic representations achieve dramatic performance improvement compared with the ones based on other side information (e.g., texts and

human annotations), due to the narrower semantic gap. Our experiments also show that a small number of images are enough to gain significant performance improvement.

There are quite a few directions we can follow in our future work. Firstly, we only use a very simple encoding method (TD) for text-based semantic representations in this paper, which results in an extremely high dimensionality and sparse vector representation per document. It has been chosen in this work as a proof of concept, but could be optimised by using alternative techniques such as latent Dirichlet allocation (LDA), latent semantic indexing (LSI), etc. Besides, in our methods of text-based representation encoding, only the occurrences of different words in a given document are considered, and the word orders which play an important role in text understanding have been ignored. Thus the meaning of sentences containing "not" and "but" would be destroyed. To overcome this limitation, some potential techniques recently developed in NLP (e.g., *document2vec* [13]) would be investigated. Currently, we extract image features with deep CNN models pre-trained on large scale object classification dataset (i.e., ImageNet). Although the pre-trained models have already shown great generalization and transferability to other visual recognition tasks, better performance can be expected by fine-tuning the models with our specific human action image data. We have done some preliminary experiments on the combination of two different types of semantic representations, but only get results no better than the use of image-based semantic representation alone. We do not want to rush to the conclusion that the image- and text-based semantic representations are not complementary before further studying the combination methods in our future work.

Acknowledgments. The authors would like to thank Ubai Sandouk from MLO group at The University of Manchester for personal communication and the anonymous reviewers for their valuable comments and suggestions.

References

1. Akata, Z., Malinowski, M., Fritz, M., Schiele, B.: Multi-cue zero-shot learning with strong supervision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 59–68 (2016)
2. Alexiou, I., Xiang, T., Gong, S.: Exploring synonyms as context in zero-shot action recognition. In: IEEE International Conference on Image Processing (ICIP). pp. 4190–4194. IEEE (2016)
3. Chao, W.L., Changpinyo, S., Gong, B., Sha, F.: An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: European Conference on Computer Vision (ECCV). pp. 52–68. Springer (2016)
4. Cheng, J., Liu, Q., Lu, H., Chen, Y.W.: Supervised kernel locality preserving projections for face recognition. Neurocomputing 67, 443–449 (2005)
5. Chuang Gan, M.L., Yang, Y., Zhuang, Y., Hauptmann, A.G.: Exploring semantic interclass relationships (sir) for zero-shot action recognition. In: Twenty-Ninth AAAI Conference on Artificial Intelligence. pp. 3769–3775 (2015)

6. Elhoseiny, M., Saleh, B., Elgammal, A.: Write a classifier: Zero-shot learning using purely textual descriptions. In: IEEE International Conference on Computer Vision (ICCV). pp. 2584–2591 (2013)
7. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1778–1785. IEEE (2009)
8. Fu, Y., Hospedales, T.M., Xiang, T., Gong, S.: Learning multimodal latent attributes. IEEE transactions on pattern analysis and machine intelligence 36(2), 303–316 (2014)
9. Inoue, N., Shinoda, K.: Adaptation of word vectors using tree structure for visual semantics. In: ACM on Multimedia Conference. pp. 277–281. ACM (2016)
10. Jiang, Y., Liu, J., Zamir, A.R., Toderici, G., Laptev, I., Shah, M., Sukthankar, R.: Thumos challenge: Action recognition with a large number of classes (2014)
11. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: IEEE International Conference on Computer Vision (ICCV). pp. 2556–2563. IEEE (2011)
12. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 951–958. IEEE (2009)
13. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning (ICML). vol. 14, pp. 1188–1196 (2014)
14. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3337–3344. IEEE (2011)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
16. Mukherjee, T., Hospedales, T.: Gaussian visual-linguistic embedding for zero-shot recognition. Conference on Empirical Methods on Natural Language Processing (EMNLP) (2016)
17. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. pp. 143–156. Springer (2010)
18. Qin, J., Wang, Y., Liu, L., Chen, J., Shao, L.: Beyond semantic attributes: Discrete latent attributes learning for zero-shot recognition. IEEE Signal Processing Letters 23(11), 1667–1671 (2016)
19. Rohrbach, M., Stark, M., Szarvas, G., Gurevych, I., Schiele, B.: What helps where– and why? semantic relatedness for knowledge transfer. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 910–917. IEEE (2010)
20. Sandouk, U., Chen, K.: Multi-label zero-shot learning via concept embedding. arXiv preprint arXiv:1606.00282 (2016)
21. Sharmanska, V., Quadrianto, N., Lampert, C.H.: Augmented attribute representations. In: European Conference on Computer Vision. pp. 242–255. Springer (2012)
22. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
23. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–9 (2015)
24. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: IEEE International Conference on Computer Vision (ICCV). pp. 4489–4497 (2015)

25. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/ (2008)
26. Wang, Q., Chen, K.: Zero-shot visual recognition via bidirectional latent embedding. International Journal of Computer Vision (2017)
27. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning-the good, the bad and the ugly. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
28. Zhang, Z., Saligrama, V.: Zero-shot recognition via structured prediction. In: European Conference on Computer Vision (ECCV). pp. 533–548. Springer (2016)