Zero-Shot Action Recognition with Error-Correcting Output Codes

Jie Qin[1,2]*, Li Liu[3,4]*, Ling Shao[4], Fumin Shen[5], Bingbing Ni[6], Jiaxin Chen[1,2] and Yunhong Wang[1,2]†
[1]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University
[2]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
[3]Malong Technologies Co., Ltd., [4]University of East Anglia
[5]University of Electronic Science and Technology of China, [6]Shanghai Jiao Tong University
qinjiebuaa@gmail.com, li.liu@malongtech.cn, ling.shao@ieee.org
fumin.shen@gmail.com, nibingbing@sjtu.edu.cn, chenjiaxinX@gmail.com, yhwang@buaa.edu.cn

Abstract

Recently, zero-shot action recognition (ZSAR) has emerged with the explosive growth of action categories. In this paper, we explore ZSAR from a novel perspective by adopting the Error-Correcting Output Codes (dubbed ZSECOC). Our ZSECOC equips the conventional ECOC with the additional capability of ZSAR, by addressing the domain shift problem. In particular, we learn discriminative ZSECOC for seen categories from both category-level semantics and intrinsic data structures. This procedure deals with domain shift implicitly by transferring the well-established correlations among seen categories to unseen ones. Moreover, a simple semantic transfer strategy is developed for explicitly transforming the learned embeddings of seen categories to better fit the underlying structure of unseen categories. As a consequence, our ZSECOC inherits the promising characteristics from ECOC as well as overcomes domain shift, making it more discriminative for ZSAR. We systematically evaluate ZSECOC on three realistic action benchmarks, i.e. Olympic Sports, HMDB51 and UCF101. The experimental results clearly show the superiority of ZSECOC over the state-of-the-art methods.

1. Introduction

During the past decade, human action recognition [1, 27, 55, 52, 54, 6, 44, 7] has been extensively explored. Robust action recognition usually relies on numerous labeled training examples. However, in many realistic scenarios, annotating sufficient examples for ever-growing new categories is exhausting and inapplicable, which inspires us to develop a system that can automatically recognize actions from novel/unseen categories.

* indicates equal contributions.
† indicates corresponding author.

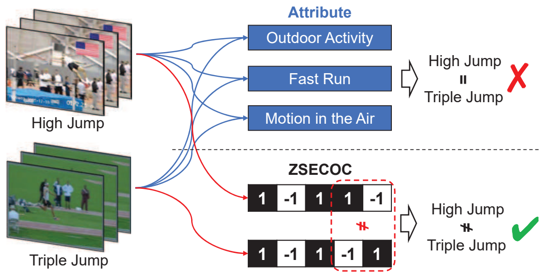


Figure 1. Attributes versus ZSECOC as the label embedding for ZSAR. Semantic attributes are shared across visually similar action categories and fail to distinguish such actions. Our ZSECOC is partially derived from the text corpus that has well-defined class hierarchy relationships, thus 'high jump' and 'triple jump' will have different ZSECOC-based embeddings. Consequently, visually similar actions from different categories can still be separated.

Zero-shot learning (ZSL) [12, 61, 26, 46, 17, 5, 64] has emerged as an effective paradigm for recognizing unseen categories without any labeled examples. Usually, ZSL can be fulfilled with the help of label embeddings (or so-called intermediate representations), among which semantic attributes have been widely utilized. Nevertheless, attributes are often manually-specified and highly subjective, since they are either heuristically defined [12] or provided by domain specialists [25]. Particularly, for zero-shot action recognition (ZSAR), attribute-based methods suffer from several specific drawbacks. First, actions are usually defined by 'verbs', which are lack of well-defined class hierarchy relationships. Second, dynamic actions are more complex than objects, making it very difficult to specify a suitable attribute pool for different actions. The above difficulties significantly limit the capabilities of previous attribute-based ZSL approaches.

To this end, word embeddings have been preferred in recent works [23, 16, 5, 60] for addressing ZSAR. By using word vectors derived from a huge text corpus (e.g.

Wikipedia), we only need category names to construct the label embeddings, instead of time-consuming manually specified attributes. However, the dimensionality m of the embedding space is usually high (typically m >1000), thus word vectors are not scalable for large-scale ZSAR that requires training m visual-semantic mapping functions (i.e. projections from visual features to label embeddings). Moreover, word vectors only take into account the textual distributed representation of category names, without considering the original visual data structures. This will directly lead to poor discriminative capabilities for final ZSAR.

Therefore, it is highly desirable to seek a discriminative and scalable label embedding that can bypass the aforementioned drawbacks. By carefully looking into the essence of ZSAR, we find that our goal intuitively equals designing category-level error-correcting output codes (ECOC). The following superior properties of ECOC [63, 65] motivate us to leverage it for tackling ZSAR:

- Error-correcting abilities. By using some redundant bits, we can tolerate some level of error[1]. This property can be leveraged to enhance the robustness of ZSAR.
- High efficiency. Only a small number of bits are required, and binary code matching is extremely fast, which can make large-scale ZSAR feasible.
- Good diversity. This indicates that the codes are row-wise uncorrelated and column-wise separable. These properties share similar spirits with the principles for designing category-level attributes as in [61].
- Accurate binary classification for each bit. This could lead to reliable visual-semantic mappings.

However, previous ECOC studies mostly addressed multi-class classifications and hardly any efforts have been made to ZSL. This is probably because directly using classifiers trained on seen categories to predict unseen instances will result in poor performance (known as domain shift [23]).

In this paper, we aim to enhance the conventional ECOC with the additional ability of zero-shot recognition (dubbed zero-shot ECOC, ZSECOC). Specifically, we derive the discriminative ZSECOC from category-level semantic correlations which are captured from a large-scale text corpus, i.e. Google News ($\approx$ 100 billion words). The semantic correlations among categories work as tunnels to implicitly transfer crucial knowledge from seen to unseen categories, e.g. the unknown ‘triple jump’ may learn from ‘high jump’ and ‘long jump’. This kind of knowledge transfer can thus address the domain shift problem to some extent. In addition to preserving semantics, the intrinsic local structure of visual data is also considered when designing our discriminative ZSECOC. Furthermore, in contrast to transductive methods [23, 59, 60] that require the access to visual data from unseen categories, a simple semantic transfer strategy *without* using any unseen data is developed to generate effective ZSECOC for unseen categories. This strategy explicitly transforms the learned embeddings of seen categories to better fit the underlying semantic structure of unseen categories. In this way, we can further eliminate the influence of domain shift. As shown in Fig. 1, ZSECOC is more discriminative than attributes. Fig. 2 illustrates the whole learning process of ZSECOC for seen/unseen action categories. Our main contributions are summarized as follows:

1) We address ZSAR by designing discriminative ZSECOC. We equip the conventional ECOC with the capability of ZSAR by discovering the semantic correlations among seen categories, which are quantitatively measured using word vectors of well-defined class hierarchy relationships. The well-established semantic knowledge is further transferred to semantically related unseen categories. As a consequence, the proposed ZSECOC inherits the intrinsic advantages of ECOC as well as overcomes domain shift. **2)** In addition to preserving category-level semantics, our ZSECOC also incorporates instance-level visual data structures. A joint optimization framework is proposed to solve the resultant challenging problem. The high-quality ZSECOC is directly learned via efficient discrete optimization without any relaxations. **3)** The proposed ZSECOC is systematically evaluated on three realistic video action datasets, i.e. Olympic Sports [39], HMDB51 [24] and UCF101 [53]. The state-of-the-art performance in terms of ZSAR clearly demonstrates the superiority of our approach.

2. Related Work

1) Zero-Shot Learning. ZSL aims to recognize unseen categories without any labeled examples. As a common practice, different label embeddings have been employed, e.g. semantic attributes [12, 25, 62, 14, 26] and word vectors [2, 58, 23, 15]. A mapping from visual features to semantic embeddings is learned from seen categories and applied to unseen categories for final recognition. A majority of existing ZSL works focus on object/scene recognition [12, 14, 25, 26, 61, 62, 20, 34, 35] and there are much fewer works on ZSAR [16, 59, 23, 60] due to the challenges previously mentioned. In ZSAR, word vectors have been preferred since only category names are required for constructing the label embeddings.

In addition, previous works also attempted to address the domain shift problem [23] existing in ZSL, since it significantly deteriorates the recognition accuracy. Several domain adaptation methods [13, 14, 23, 60] have been proposed for ZSL, based on transductive learning [23] or data augmentation [60]. However, most of their models were trained using some unseen instances, which violate the fundamental assumption of the standard ZSL setting that no unseen examples could be accessed during training. In this

[1] If the minimum Hamming distance between any pair of codewords is d, the ECOC can correct at least $\frac{d-1}{2}$ single bit errors.

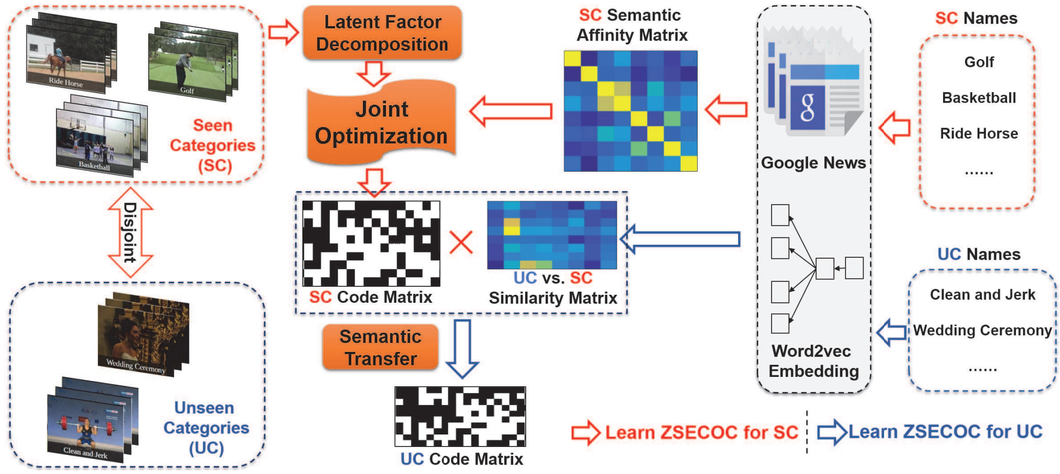


Figure 2. The flow chart of ZSECOC. We design discriminative ZSECOC for seen categories that is both semantics-preserving and data-driven. A joint optimization framework is proposed to solve the resultant challenging problem. We learn the ZSECOC for unseen categories through a simple yet effective semantic transfer strategy. Black and white entries of the code matrix indicate '1' and '-1', respectively.

paper, we develop a simple semantic transfer scheme without any transductive dependency on unseen examples.

2) Error-Correcting Output Codes. ECOC [8, 3, 42, 63, 10] has been explored as an efficient and effective alternative to multi-class classification. Specifically, each class has some pre-specified codeword, e.g. '1010100'. ECOC methods train a set of binary classifiers in terms of each bit of the codes. The final multi-class classification is fulfilled through matching the class-level codewords with the binary code of a test point predicted by binary classifiers.

Lots of efforts have been made for simultaneously optimizing the code matrix and binary classifiers, e.g. random ECOC [3] and discriminant ECOC [42]. However, there are rarely any practices that explore ECOC for ZSL. The most related work to ours is [40], where semantic output codes (SOC) were designed for ZSL. SOC did share some similar spirits with ECOC. Nevertheless, it was directly obtained from semantic knowledge bases, thus lacked the intrinsic characteristics of ECOC (e.g. good diversity). Therefore, to the best of our knowledge, this is the first work that enhances ECOC for the purpose of ZSAR. Next, we will introduce the design of our discriminative ZSECOC in detail.

3. Discriminative ZSECOC

In ZSAR, we aim to recognize any instance $\mathbf{x}^u$ from C^u unseen action categories, given all N instances $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N \in \mathbb{R}^{d\times N}$ from C seen categories, where d is the original feature dimension. In this work, we would like to seek m-bit category-level ZSECOC as the label embedding, by incorporating word vectors as the side information. We denote the ZSECOC of seen and unseen categories as $\mathbf{B} = \{\mathbf{b}_i\}_{i=1}^C \in \{-1,1\}^{m\times C}$ and $\mathbf{B}^u = \{\mathbf{b}_j^u\}_{j=1}^{C^u} \in \{-1,1\}^{m\times C^u}$, respectively. The semantic labels of seen and unseen categories are denoted as $\{y_i\}_{i=1}^C \in \mathcal{Y}$ and $\{y_j^u\}_{j=1}^{C^u} \in \mathcal{Y}^u$ respectively, where $\mathcal{Y} \cap \mathcal{Y}^u = \emptyset$. In the following, we will show the principles for designing ZSECOC for seen categories (i.e. $\mathbf{B}$) from category-level semantics and visual data structures. Subsequently, a joint optimization framework based on alternating iteration is presented for solving the resultant challenging problem.

3.1. Design Principles

1) Preserving Category-Level Semantics. Previous works [3, 42, 18, 8] have shown that ECOC should have good diversity. We find this property is also crucial for ZSAR, thus our code matrix $\mathbf{B}$ is derived from the following properties:

- Column separation: $\max \sum ||\mathbf{b}_i - \mathbf{b}_j||_2^2$,
- Row uncorrelation: $\frac{1}{C}\sum \mathbf{b}_i\mathbf{b}_i^\top = \mathbf{I}$,
- Row-wise balancedness: $\sum \mathbf{b}_i = \mathbf{0}$,

where $\mathbf{I}$ is the identity matrix. Besides, as discussed in [45, 61], preserving semantics is crucial for discrimination. We adopt such a property with the following objective function:

$$\min_{\mathbf{b}_i} \sum s_{ij}||\mathbf{b}_i - \mathbf{b}_j||_2^2, \text{ s.t. } \mathbf{b}_i \in \{-1,1\}^m, \quad (1)$$

where s_{ij} denotes the category-level semantic affinity between the i-th and j-th categories. Specifically, we capture semantic correlations across categories based on the distributed representation of category names. In practice, we employ the skip-gram neural network model [37] trained on the Google News dataset. Each category is thus embedded by a 300-d word vector $\phi(y_i)$, where '$\phi(\cdot)$' is the embedding function. We assign the cosine similarity between $\phi(y_i)$ and $\phi(y_j)$ to s_{ij}, i.e. $s_{ij} = \frac{<\phi(y_i),\phi(y_j)>}{||\phi(y_i)||\cdot||\phi(y_j)||}, i,j = 1,...,C$, where $<,>$ indicates the inner product operation.

The intuition behind formula (1) is that the ZSECOC of similar categories should be close to each other, while different categories should possess distinct codes. Here, we denote this property as *column 'association'*. Similar objectives are also adopted in graph-based hashing methods [57, 33], because of their capabilities of well preserving semantics and achieving high precision in the retrieval task. By combining all the above objectives, we have

$$\min_{\mathbf{b}_i} \sum s_{ij}||\mathbf{b}_i - \mathbf{b}_j||_2^2 - \lambda \sum ||\mathbf{b}_i - \mathbf{b}_j||_2^2,$$
$$\text{s.t. } \frac{1}{C}\sum \mathbf{b}_i\mathbf{b}_i^\top = \mathbf{I}, \sum \mathbf{b}_i = \mathbf{0}, \mathbf{b}_i \in \{-1,1\}^m, \quad (2)$$

where $\lambda > 0$ is the trade-off between columns 'separation' and 'association'. By introducing the matrix form of (2), we have

$$\min_{\mathbf{B}} \mathcal{O}_{\text{sp}} := \text{trace}(\mathbf{B}\mathbf{L}\mathbf{B}^\top)$$
$$\text{s.t. } \mathbf{B}\mathbf{B}^\top = C\mathbf{I},\ \mathbf{B}\mathbf{1} = \mathbf{0}, \mathbf{B} \in \{-1,1\}^{m\times C}, \quad (3)$$

where the subscript 'sp' indicates the semantics-preserving characteristic of our ZSECOC, and $\mathbf{L}$ is the associated Laplacian matrix of the affinity matrix $\mathbf{S}' = \{s'_{ij}\} \in \mathbb{R}^{C\times C}$ where $s'_{ij} = s_{ij} - \lambda, \forall i, j$. Specifically, $\mathbf{L} = \text{diag}(\mathbf{S}'\mathbf{1}) - \mathbf{S}'$.

To tackle this challenging problem, many existing approaches [43, 50, 33, 56, 57] choose to obtain sub-optimal solutions by discarding the binary constraints. As shown in [32, 49, 51], these solutions are of low quality and will lead to less effective classification performance. In this paper, we attempt to address the problem directly without any relaxations and achieve a more accurate solution to $\mathbf{B}$. We will provide the details for optimization in Section 3.2.

2) Capturing Visual Data Structures. According to [63, 65], visual data structures should also be considered for discriminative ECOC. For instance, [63] utilized spectral analysis and [65] employed sum match kernel to acquire useful information from data. We adopt the similar spirit but learn our ZSECOC from data in a different way by using latent factor decomposition (LFD). Specifically, we formulate the problem of learning data-driven ZSECOC as

$$\min_{\mathbf{R},\mathbf{V},\mathbf{B}} \mathcal{O}_{\text{dd}} := ||\mathbf{X} - \mathbf{D}\mathbf{V}||_{\text{F}}^2 + \gamma||\mathbf{V}||_{\text{F}}^2 + \alpha||\mathbf{B}\mathbf{P} - \mathbf{R}\mathbf{V}||_{\text{F}}^2$$
$$\text{s.t. } \mathbf{R}^\top\mathbf{R} = \mathbf{I}, \mathbf{B} \in \{-1,1\}^{m\times C}, \quad (4)$$

where 'dd' denotes the data-driven characteristic of ZSECOC, $||\cdot||_{\text{F}}^2$ denotes the Frobenius norm, $\mathbf{D} \in \mathbb{R}^{d\times m}$ is the pre-computed dictionary (or so-called bases) usually obtained by applying k-means or Gaussian mixture models on seen data $\mathbf{X}$, $\mathbf{V} \in \mathbb{R}^{m\times N}$ is the latent factor matrix, $\mathbf{P} \in \{0,1\}^{C\times N}$ is the category-instance indicator matrix, $\mathbf{R} \in \mathbb{R}^{m\times m}$ is the orthogonal transformation matrix, $\alpha > 0$ is the penalty parameter, and $\gamma > 0$ is the regularization parameter w.r.t. $\mathbf{V}$. In particular, each entry p_{ij} of $\mathbf{P}$ is defined as follows:

$$p_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_j \text{ belongs to the } i\text{−th category}, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

By multiplying the category-level $\mathbf{B}$ with $\mathbf{P}$, we can reconstruct the codes for all seen instances.

The first two terms in (4) correspond to the latent factor decomposition problem. With the dictionary $\mathbf{D}$, a data point $\mathbf{x}_n$ can be reconstructed as $\mathbf{D}\mathbf{v}_n$ by using its latent factor $\mathbf{v}_n$. To approximate the final ZSECOC, we derive the latent factors in terms of the same dimension as with the length of the codes (i.e. m). Furthermore, to fit the codes and the decomposed latent factors, we introduce a penalty term, i.e. the last term in (4). Theoretically, with a sufficiently large α, the resulting ZSECOC can well preserve the intrinsic structure of the visual data. We additionally impose an orthogonal rotation on the factors because such rotation will reduce the quantization loss effectively [19].

Overall Objective Function. By coupling the above two problems, we can learn discriminative ZSECOC from both category-level semantics and visual data structures. The overall objective function is

$$\min_{\mathbf{R},\mathbf{V},\mathbf{B}} \mathcal{O}(\mathbf{R},\mathbf{V},\mathbf{B}) := \mathcal{O}_{\text{dd}} + \beta\mathcal{O}_{\text{sp}}$$
$$\text{s.t. } \mathbf{R}^\top\mathbf{R} = \mathbf{I}, \mathbf{B}\mathbf{B}^\top = C\mathbf{I}, \mathbf{B}\mathbf{1} = \mathbf{0}, \mathbf{B} \in \{-1,1\}^{m\times C}, \quad (6)$$

where $\beta > 0$ weights the importance between the two characteristics, i.e. semantics-preserving and data-driven.

3.2. Alternating Optimization

The above joint problem (6) is generally NP-hard and non-convex due to the discrete constraint on $\mathbf{B}$. Here, we attempt to tackle it by iteratively computing each of the three variables, i.e. $\mathbf{R}$, $\mathbf{V}$ and $\mathbf{B}$. In other words, we find the solution to one variable while fixing the other two. Similar techniques are adopted in [30, 48, 49].

R-Step: With fixed $\mathbf{B}$ and $\mathbf{V}$, the subproblem w.r.t. $\mathbf{R}$ is

$$\min_{\mathbf{R}} ||\mathbf{B}\mathbf{P} - \mathbf{R}\mathbf{V}||_{\text{F}}^2, \text{ s.t. } \mathbf{R}^\top\mathbf{R} = \mathbf{I}. \quad (7)$$

This objective function is equivalent to the classic Orthogonal Procrustes Problem (OPP) [47]. OPP tries to find a rotation to align one point set (i.e. $\mathbf{V}$) with another (i.e. $\mathbf{B}\mathbf{P}$). Specifically, the solution to $\mathbf{R}$ is obtained as follows:

$$\mathbf{U}\mathbf{\Sigma}\hat{\mathbf{U}}^\top = \text{svd}(\mathbf{B}\mathbf{P}\mathbf{V}^\top),\ \mathbf{R} = \mathbf{U}\hat{\mathbf{U}}^\top, \quad (8)$$

where 'svd$(\cdot)$' denotes the singular value decomposition.

V-Step: The subproblem by fixing $\mathbf{B}$ and $\mathbf{P}$ becomes:

$$\min_{\mathbf{V}} ||\mathbf{X} - \mathbf{D}\mathbf{V}||_{\text{F}}^2 + \alpha||\mathbf{B}\mathbf{P} - \mathbf{R}\mathbf{V}||_{\text{F}}^2 + \gamma||\mathbf{V}||_{\text{F}}^2$$
$$\Leftrightarrow \min_{\mathbf{V}} ||\mathbf{X}_{\mathbf{BP}} - \mathbf{D}_{\mathbf{R}}\mathbf{V}||_{\text{F}}^2 + \gamma||\mathbf{V}||_{\text{F}}^2, \quad (9)$$

where

$$\mathbf{X}_{\mathbf{BP}} = \begin{bmatrix} \mathbf{X} \\ \sqrt{\alpha}\mathbf{B}\mathbf{P} \end{bmatrix} \text{ and } \mathbf{D}_{\mathbf{R}} = \begin{bmatrix} \mathbf{D} \\ \sqrt{\alpha}\mathbf{R} \end{bmatrix}.$$

Algorithm 1: Learning ZSECOC for Seen Categories

Input: Seen instances $\mathbf{X} \in \mathbb{R}^{d \times N}$; the Laplacian matrix $\mathbf{L} \in \mathbb{R}^{C \times C}$; the indicator matrix $\mathbf{P} \in \{0,1\}^{C \times N}$; code length m; maximum iteration t; parameters $\alpha, \beta, \gamma, \lambda, \delta, \rho$.

Output: Category-level ZSECOC: $\mathbf{B} \in \{-1,1\}^{m \times C}$.

Generate dictionary $\mathbf{D} \in \mathbb{R}^{d \times m}$ by k-means on $\mathbf{X}$;
Compute $\mathbf{V} = (\mathbf{D}^\top \mathbf{D} + \gamma \mathbf{I})^{-1} \mathbf{D}^\top \mathbf{X}$;
Randomly initialize $\mathbf{B}$, set proximal parameter $\mu = 1$;
Loop until convergence or reach t iterations:
  - R-Step: Compute $\mathbf{R}$ by Eq. (8);
  - V-Step: Compute $\mathbf{V}$ by Eq. (10);
  - B-Step: Compute $\mathbf{B}$ by Eq. (13) and (14).

Problem (9) is equivalent to the regularized least squares problem and thus has a closed-form solution:

$$\mathbf{V} = (\mathbf{D}_{\mathbf{R}}^\top \mathbf{D}_{\mathbf{R}} + \gamma \mathbf{I})^{-1} \mathbf{D}_{\mathbf{R}}^\top \mathbf{X}_{\mathbf{BP}}. \tag{10}$$

B-Step: Given $\mathbf{V}$ and $\mathbf{R}$, the objective function in terms of $\mathbf{B}$ has the following formulation:

$$\begin{aligned} &\min_{\mathbf{B}} \ \alpha ||\mathbf{BP} - \mathbf{RV}||_{\mathrm{F}}^2 + \beta \mathrm{trace}(\mathbf{BLB}^\top) \\ &\text{s.t. } \mathbf{BB}^\top = C\mathbf{I}, \mathbf{B1} = \mathbf{0}, \mathbf{B} \in \{-1,1\}^{m \times C}. \end{aligned} \tag{11}$$

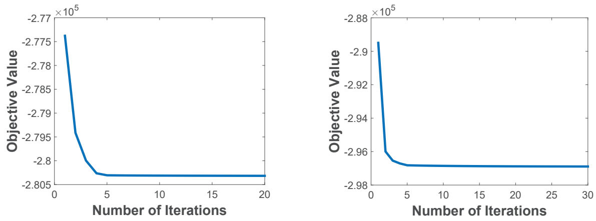
Due to the three constraints on $\mathbf{B}$, this problem is very difficult to solve. To make (11) computationally feasible, we rewrite it by discarding the first two constraints:

$$\begin{aligned} &\min_{\mathbf{B}} \ \mathcal{O}(\mathbf{B}) := \alpha ||\mathbf{BP} - \mathbf{RV}||_{\mathrm{F}}^2 + \beta \mathrm{trace}(\mathbf{BLB}^\top) \\ &+ \frac{\delta}{4} ||\mathbf{BB}^\top||_{\mathrm{F}}^2 + \frac{\rho}{2} ||\mathbf{B1}||_{\mathrm{F}}^2, \text{ s.t. } \mathbf{B} \in \{-1,1\}^{m \times C}. \end{aligned} \tag{12}$$

We can see that with sufficiently large δ and ρ, problems (11) and (12) will be equivalent to each other. However, the above problem is still computationally inapplicable due to the discrete constraint. By carefully looking at (12), we attempt to tackle it by adopting the discrete proximal linearized minimization (DPLM) algorithm recently proposed in [51]. DPLM reformulates the problem into an unconstrained minimization problem, which is then addressed using proximal optimization. In this way, high-quality ZSECOC can be obtained directly without any relaxations. Particularly, we obtain the solution to $\mathbf{B}$ as follows:

$$\mathbf{B}^{(i+1)} = \mathrm{sign}(\mathbf{B}^{(i)} - \frac{1}{\mu} \nabla \mathcal{O}(\mathbf{B}^{(i)})), \tag{13}$$

where $\mathbf{B}^{(i)}$ is the obtained code in the i-th iteration; $\mu > 0$ is the proximal parameter controlling the convergence rate; 'sign$(\cdot)$' returns '1' if the argument is positive and '-1' otherwise. The binary optimization is then performed by updating $\mathbf{B}$ in each iteration with the gradient of $\mathcal{O}(\mathbf{B})$ as



(a) HMDB51 (b) UCF101

Figure 3. Convergence of our optimization procedure.

$$\begin{aligned} \nabla \mathcal{O}(\mathbf{B}) = \ &2\alpha(\mathbf{BPP}^\top - \mathbf{RVP}^\top) + 2\beta \mathbf{BL} \\ &+ \delta \mathbf{BB}^\top \mathbf{B} + \rho \mathbf{B11}^\top. \end{aligned} \tag{14}$$

In practice, we adopt the self-adaptive scheme (SAS) [31] to update the convergence rate for fast convergence of DPLM. Specifically, SAS associates the optimization procedure with an adaptive rate μ. In each iteration, μ is enlarged or reduced according to the changing values of $\mathcal{O}(\mathbf{B})$ between adjacent iterations. We also set an upper bound for the number of iterations for DPLM.

Convergence Analysis. Algorithm 1 illustrates the overall learning process of ZSECOC for seen categories. Theoretically, our optimization procedure can be regarded as the generalized cyclic coordinate descent (CCD) algorithm. Previous studies [4, 19] have guaranteed the convergence of such an optimization scheme. Specifically, in each R/V/B-step, we can obtain either a global or a local optimum. And the overall problem (6) is lower bounded, we can thus ensure the convergence of our method. In our experiments, the overall method can successfully converge within $t = 5 \sim 10$ iterations, as illustrated in Fig. 3.

Until now, the discriminative ZSECOC (i.e. $\mathbf{B}$) for seen categories has been obtained, it is still unclear how to generate the prototype codes of unseen categories (i.e. $\mathbf{B}^u$) for final recognition. We will elaborate on how to acquire $\mathbf{B}^u$ with semantic knowledge transfer in the following section.

4. Semantic Transfer for Unseen Categories

Straightforwardly, we could design $\mathbf{B}^u$ in the same way as with $\mathbf{B}$, by preserving semantics and data structures of unseen categories. However, on the one hand, employing visual data from unseen categories is prohibitive in the training phase of the standard ZSL setting as ours. This makes learning $\mathbf{B}^u$ intractable due to the lack of data, which is also one major reason why previous ECOC has hardly been explored for ZSL. On the other hand, we will encounter domain shift even if we could design $\mathbf{B}^u$ in this way, i.e. visual-semantic mappings trained from seen categories are not suitable for classifying disjoint unseen instances.

In [23, 60, 21], several techniques were proposed to solve domain shift in a transductive way, by incorporating some examples from unseen categories to refine the visual-semantic mapping. However, as we claimed, this is opposed to our standard ZSL setting. Here, we propose a simple semantic knowledge transfer strategy to acquire $\mathbf{B}^u$ without employing any instances from unseen categories. An existing work [61] attempted to utilize a matrix indicating the similarities between seen and unseen categories to fulfill this task. However, the matrix was provided by some volunteers, thus it was highly subjective and unreliable. As we aim to learn ZSECOC automatically without manual intervention, we instead employ a matrix that captures semantic correlations between seen and unseen categories. Particularly, we construct the similarity matrix $\mathbf{S}^u = \{s_{ij}^u\} \in \mathbb{R}^{C \times C^u}$ based on the cosine distances between word vectors of seen and unseen categories:

$$s_{ij}^u = \frac{< \phi(y_i), \phi(y_j^u) >}{||\phi(y_i)|| \cdot ||\phi(y_j^u)||}, i = 1, ..., C, \text{ and } j = 1, ..., C^u.$$

Subsequently, we generate $\mathbf{B}^u = \mathrm{sign}(\mathbf{B}\mathbf{S}^u)$. In this way, $\mathbf{S}^u$ can well transfer the semantic knowledge from correlated seen categories to unseen ones. More importantly, our $\mathbf{S}^u$ is obtained *without* utilizing any unseen instances.

Zero-Shot Recognition. Based on $\mathbf{B}$, i.e. the category-level ZSECOC of seen categories, we can learn the visual-semantic mapping via a set of independent binary classifiers (e.g. linear SVMs). Specifically, we regard each row of $\mathbf{B}$ as the binary labels for seen categories and one binary classifier is trained based on all the seen data and the associated labels. This will result in m independent binary classifiers: $\{f_i\}_{i=1}^m$. Subsequently, for any unseen data point $\mathbf{x}^u$, we can acquire its code through the outputs of these classifiers, i.e. $F(\mathbf{x}^u) = [f_1(\mathbf{x}^u), ..., f_m(\mathbf{x}^u)]^\top$. Finally, we formulate zero-shot recognition as the Hamming decoding process [3] of ECOC. We assign the unseen category label y_{j^*} to a test data point $\mathbf{x}^u$ as follows:

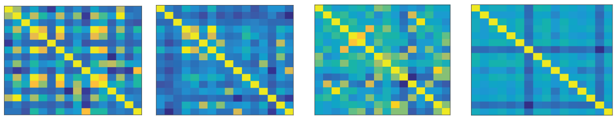
$$j^* = \operatorname*{argmin}_j d_{\mathrm{H}}(F(\mathbf{x}^u), \mathbf{b}_j^u), \tag{15}$$

where d_{H} denotes the Hamming distance and $\mathbf{b}_j^u$ is the prototype code of the j-th unseen category.

5. Experiments

5.1. Experimental Setup

Datasets and Settings. We conduct our experiments on three realistic video action datasets, i.e. Olympic Sports [39], HMDB51 [24] and UCF101 [53], in which there are totally 783, 6766 and 13320 action videos from 16, 51 and 101 categories, respectively. For action representation, we adopt the 50688-d features kindly provided by [60], which are improved dense trajectory (IDT) [55] features encoded



(a) ZSECOC (b) Word Vector (c) Attribute (d) Visual Data

Figure 4. Similarity matrices created using different label embeddings on Olympic Sports. Brighter colors depict larger values.

Table 1. Recognition accuracies with different label embeddings. ZSECOC employs $m = 10log_2(\#\text{category})$ bits of codes.

Embedding	Olympic Sports	HMDB51	UCF101
Word Vector	21.6±0.9 (300-d*)	16.5±3.9 (300-d)	3.2±0.7 (300-d)
Attribute	27.7±4.6 (40-bit)	N/A	13.7±0.5 (115-bit)
ZSECOC	**59.8**±**5.6** (40-bit)	**22.6**±**1.2** (70-bit)	**15.1**±**1.7** (100-bit)

*1-d feature of the double-precision floating-point format equals 64-bit binary code.

by Fisher Vectors (FV) [41]. We adopt the skip-gram neural network model [37] trained on the Google News dataset (≈ 100 billion words) and represent each category name by an L2-normalized 300-d word vector. For any multi-word category name (e.g. 'ride horse'), we generate its vector by accumulating the word vectors of each unique word [38]. For the visual-semantic mapping, we adopt a set of independent linear SVMs [11], as used by the conventional E-COC methods [3]. The lengths of ZSECOC are empirically set to $m = 10log_2(C + C^u)$ as suggested in [3], i.e. 40, 70 and 100 w.r.t. Olympic Sports, HMDB51 and UCF101, respectively. We use cross-validations on seen categories to determine the hyper-parameters for our model.

Evaluation Metric. Following [60], we adopt the class-wise data splits by evenly dividing each dataset into seen/unseen categories, i.e. 8/8, 27/26 and 51/50 splits with regard to Olympic Sports, HMDB51 and UCF101, respectively. We randomly generate 10 splits for each dataset, and the average recognition accuracies and standard deviations are reported. Due to the randomness of initializing $\mathbf{B}$, we report the average results for each split of each dataset based on 5 trials. We conduct the experiments on a PC with an Intel quad-core 3.4GHz CPU and 32GB memory.

In the following, we systematically evaluate our ZSECOC in different aspects. Firstly, ZSECOC is compared to conventional label embebddings. Subsequently, we compare ZSECOC with state-of-the-art ECOC and ZSL methods. Finally, we visualize various qualitative results and present some further analyses as well.

5.2. Experimental Results

Evaluation of Label Embeddings. We first evaluate different strategies for the label embedding, including semantic attributes, word vectors and our ZSECOC. For the real-valued word vectors, we employ linear support vector regression (SVR) instead of SVMs for learning the visual-semantic mapping. In terms of semantic attributes, [29] and [22] provided the 40 and 115 category-level attributes for Olympic Sports and UCF101, respectively. As no semantic

Table 2. Zero-shot action recognition accuracies on the three datasets in terms of different ZSL methods. Feature: Fisher Vectors (**FV**) or Bag of Words (**BoW**); Label embedding - Attribute (**A**) or Word Vector (**WV**); **TD**: Transductive Dependency on unseen data.

Method	Reference	Feature	Label Embedding	TD	Olympic Sports	HMDB51	UCF101
HAA [28]	CVPR 2011	FV	A	×	46.1±12.4	N/A	14.9±0.8
DAP [26]	TPAMI 2014	FV	A	×	45.4±12.8	N/A	14.3±1.9
IAP [26]	TPAMI 2014	FV	A	×	42.3±12.5	N/A	12.8±2.0
ST [59]	ICIP 2015	BoW	WV	×	N/A	13.0±2.7	10.9±1.5
ST [59]	ICIP 2015	BoW	WV	✓	N/A	15.0±3.0	**15.8±2.3**
ESZSL [46]	ICML 2015	FV	WV	×	39.6±9.6	18.5±2.0	15.0±1.3
SJE [2]	CVPR 2015	FV	WV	×	28.6±4.9	13.3±2.4	9.9±1.4
SJE [2]	CVPR 2015	FV	A	×	47.5±14.8	N/A	12.0±1.2
UDA [23]	ICCV 2015	FV	A	✓	N/A	N/A	13.2±1.9
UDA [23]	ICCV 2015	FV	A+WV	×	N/A	N/A	14.0±1.8
MTE [60]	ECCV 2016	FV	WV	×	44.3±8.1	19.7±1.6	**15.8±1.3**
ZSECOC	Ours	FV	ECOC	×	**59.8±5.6** (40-bit)	**22.6±1.2** (70-bit)	15.1±1.7 (100-bit)

Table 3. Recognition accuracies using different ECOC methods.

Method	Olympic Sports	HMDB51	UCF101
RSECOC [3]	18.4±0.6	5.3 ±0.1	3.0±0.8
RDECOC [3]	25.3±1.8	6.2±1.0	2.7±0.2
DECOC [42]	40.1±4.2	6.9±1.0	4.8±0.6
Forest-ECOC [9]	51.0±8.7	9.2±0.7	5.9±0.5
ZSECOC	**59.8±5.6**	**22.6±1.2**	**15.1±1.7**

attributes are available for HMDB51, we omit the attribute-based results on HMDB51. Table 1 shows the ZSAR accuracies on the three datasets. We can have the following observations: (1) Semantic attributes based embeddings can achieve better accuracies than word vectors. (2) Our ZSECOC is the best choice for the label embedding because of its superior characteristics. (3) Shorter codes are required by ZSECOC compared with word vectors and attributes, leading to lower memory load.

We further visually depict the similarity matrices (see Fig. 4) among categories for the three embeddings on Olympic Sports as in [64]. For the binary attributes and ZSECOC, we create their matrices based on the Hamming distance, and the cosine distance is adopted for word vectors. As our ZSECOC is data-driven, we also show the similarity matrix of visual data using the Euclidean distance. We can observe that, in most cases, the colors of the blocks in Fig. 4 (a) are related to the corresponding ones in Fig. 4 (b) and (d). This indicates that ZSECOC can generate a similarity matrix that couples the analogical relationships among both word vectors and visual data. As for the attribute-based matrix, correlations between different categories are stronger than any other embeddings. Thus, the category-level 'separation' is neglected to some extent. This is probably the reason why attributes are not so discriminative as our learned ZSECOC, which owns both category-level 'separation' and 'association'.

Comparison with Other ECOC Methods. As ZSECOC is developed from the conventional ECOC, we also compare ZSECOC with several state-of-the-art ECOC methods. The competitors include data-independent methods: random sparse ECOC (RSECOC) [3] and random dense ECOC (RDECOC) [3]; data-dependent ones: discriminant ECOC (DECOC) [42] and Forest-ECOC [9]. All of the approaches are implemented using the ECOC Library [10]. For fair comparison, we employ the same similarity matrix $\mathbf{S}^u$ to equip data-dependent ECOC with the ZSL ability.

The recognition accuracies of different methods are shown in Table 3. Generally, data-dependent ECOC methods perform better than data-independent ones. This implies the necessity of incorporating visual data structures when learning ECOC. Particularly, our ZSECOC consistently achieves the best accuracies on all the three datasets. The performance gains are especially obvious on large-scale datasets, i.e. HMDB51 and UCF101. This mainly owes to the employment of both category-level semantic correlations and instance-level visual data structures when designing our ZSECOC.

Comparison with State-of-the-Art ZSL Methods. We compare ZSECOC with various contemporary ZSL methods: (1) Direct/Indirect Attribute Prediction (DAP/IAP) [26]: the classic attribute-based ZSL strategy; (2) Human Actions by Attributes (HAA) [28]: we adopt the simplified version provided in [60]; (3) the Self-Training model (ST) [59]: domain shift is solved by a transductive self-training procedure; (4) Embarrassingly Simple Zero-Shot Learning (ESZSL) [46]: the mean square loss is used instead of the regression loss w.r.t. the objective function; (5) Structured Joint Embedding (SJE) [2]: a triplet hinge loss is employed to ensure more related labels correspond to higher mapping values from visual features; (6) Unsupervised Domain Adaptation (UDA) [23]: a target domain specific dictionary is learned by using some unseen data; (7) Multi-Task Embedding (MTE) [60]: multi-task regression is developed to learn the visual-semantic mapping, together with a data augmentation strategy.

We notice that some compared methods (e.g. UDA and ST) require the access to unseen instances and we still compare with their results under this transductive setting. However, [59, 60] developed some data augmentation techniques by using examples from some auxiliary categories. In this setting, categories used for training may be re-used during testing, which seriously violates the fundamental as-

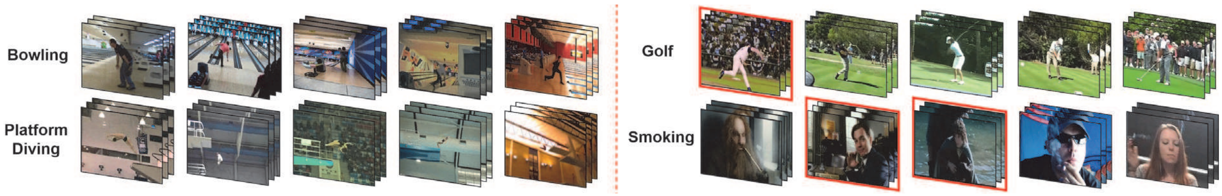
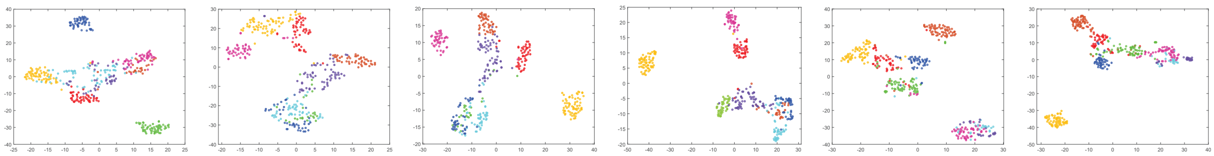


Figure 5. Top-5 returned video examples for unseen categories on Olympic Sports (left) and HMDB51 (right).



(a) ZSECOC (S1) (b) Word Vector (S1) (c) Attribute (S1) (d) ZSECOC (S2) (e) Word Vector (S2) (f) Attribute (S2)

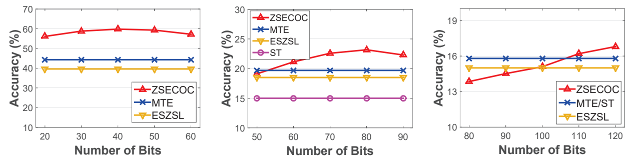
Figure 6. t-SNE [36] visualization between different embeddings of unseen categories w.r.t. two representative (**S**)plits on Olympic Sports.

sumption of ZSL that training/test categories should be mutually excluded. We therefore do not compare with their results in this data augmentation setting.

All the comparison results are illustrated in Table 2. We can conclude that the compared methods can usually achieve better performance by leveraging attributes. This is in accordance with the observations from Table 1. Overall, our ZSECOC consistently outperforms other competitors on Olympic Sports and HMDB51, often by a large margin. It is interesting that ZSECOC also has the advantage over several transductive methods due to its discrimination and semantic transfer ability. On UCF101, ZSECOC is superior to most alternatives and performs slightly worse than ST and MTE. However, ST is based on a transductive setting with the access to unseen data and MTE utilizes the 300-d (1.92×10^4-bit) word vector based embedding which is not so scalable as our 100-bit compact binary embedding.

Qualitative Results. We visualize some ZSAR results in Fig. 5 with the top-5 returned videos corresponding to four unseen categories. The videos within red rectangles depict false-positive examples. Interestingly, we are able to recognize the 'bowling' and 'diving' actions regardless of different view points. As for 'golf', the first returned video is misclassified since 'throwing a baseball' extremely resembles 'playing golf'. Unfortunately, ZSECOC cannot well recognize 'smoking' partially because the IDT features are not so sensitive to the subtle thin smoke. Fig. 6 further visually depict the unseen categories in terms of different embeddings on Olympic Sports to facilitate better understanding of our outstanding performance. As seen in Fig. 6, the ZSECOC-based embeddings appear to be more clustered than those using the original word vector/attribute-based embeddings. This indicates that different unseen categories can be better separated by using our ZSECOC.

Effects of Code Lengths. We show our performance with the increasing numbers of bits. In general, longer codes can achieve better accuracies, especially on larger datasets. Specifically, with 110/120 bits on UCF101, our ZSECOC



(a) Olympic Sports (b) HMDB51 (c) UCF101

Figure 7. Recognition accuracies with increasing code lengths. All the compared methods adopt the fixed-length 300-d word vectors.

can even outperform the best two methods in Table 2, i.e. ST and MTE, which is really encouraging. On the other hand, ZSECOC can obtain the state-of-the-art results with very short codes (e.g. 20 bits w.r.t. Olympic Sports), and this is highly desirable in realistic scenarios.

6. Conclusion

In this paper, we formulated zero-shot learning as designing error-correcting output codes (ECOC). Discriminative ZSECOC was learned in terms of preserving category-level semantics as well as maintaining intrinsic visual data structures. A joint optimization scheme was proposed to iteratively learn the optimal ZSECOC for seen categories. An intuitive semantic transfer strategy was developed to obtain the ZSECOC of unseen categories without any transductive dependency on test data. The extensive experiments in terms of zero-shot action recognition on three public video action datasets demonstrated the state-of-the-art performance of the proposed ZSECOC.

Acknowledgement

This work was partly supported by the National Natural Science Foundation of China (No. 61573045), the Foundation for Innovative Research Groups through the National Natural Science Foundation of China (No. 61421003), the National Natural Science Foundation of China (No. 61502301), and China's Thousand Youth Talents Plan.

References

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *CSUR*, 43(3):16:1–16:43, 2011.

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015.

[3] E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *JMLR*, 1:113–141, Sept. 2001.

[4] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

[5] X. Chang, Y. Yang, G. Long, C. Zhang, and A. Hauptmann. Dynamic concept composition for zero-example event detection. In *AAAI*, 2016.

[6] C. Chen, R. Jafari, and N. Kehtarnavaz. Improving human action recognition using fusion of depth camera and inertial sensors. *IEEE Transactions on Human-Machine Systems*, 45(1):51–61, 2015.

[7] C. Chen, M. Liu, B. Zhang, J. Han, J. Jiang, and H. Liu. 3d action recognition using multi-temporal depth motion maps and fisher vector. In *IJCAI*, 2016.

[8] T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *JAIR*, 2(1):263–286, Jan. 1995.

[9] S. Escalera, O. Pujol, and P. Radeva. Boosted landmarks of contextual descriptors and forest-ecoc: A novel framework to detect and classify objects in cluttered scenes. *PRL*, 28(13):1759 – 1768, 2007.

[10] S. Escalera, O. Pujol, and P. Radeva. Error-correcting output codes library. *JMLR*, 11(Feb):661–664, 2010.

[11] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 9:1871–1874, June 2008.

[12] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

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

[14] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *IEEE TPAMI*, 37(11):2332–2345, Nov 2015.

[15] Z. Fu, T. Xiang, E. Kodirov, and S. Gong. Zero-shot object recognition by semantic manifold distance. In *CVPR*, 2015.

[16] C. Gan, M. Lin, Y. Yang, Y. Zhuang, and A. G. Hauptmann. Exploring semantic inter-class relationships (sir) for zero-shot action recognition. In *AAAI*, 2015.

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

[18] N. Garcia-Pedrajas and C. Fyfe. Evolving output codes for multiclass problems. *IEEE Transactions on Evolutionary Computation*, 12(1):93–106, Feb 2008.

[19] Y. Gong and S. Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *CVPR*, 2011.

[20] D. Jayaraman and K. Grauman. Zero shot recognition with unreliable attributes. In *NIPS*, 2014.

[21] Z. Ji, Y. Yu, Y. Pang, J. Guo, and Z. Zhang. Manifold regularized cross-modal embedding for zero-shot learning. *Information Sciences*, 378:48 – 58, 2017.

[22] Y.-G. Jiang, J. Liu, A. Roshan Zamir, I. Laptev, M. Piccardi, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `/ICCV13-Action-Workshop/`, 2013.

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

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

[25] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[26] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 36(3):453–465, March 2014.

[27] I. Laptev. On space-time interest points. *IJCV*, 64(2):107–123, 2005.

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

[29] J. Liu, B. Kuipers, and S. Savarese. Technical report: Recognizing human actions by attributes, 2011.

[30] L. Liu, Z. Lin, L. Shao, F. Shen, G. Ding, and J. Han. Sequential discrete hashing for scalable cross-modality similarity retrieval. *IEEE TIP*, 2016.

[31] L. Liu, M. Yu, and L. Shao. Projection bank: From high-dimensional data to medium-length binary codes. In *ICCV*, 2015.

[32] W. Liu, C. Mu, S. Kumar, and S.-F. Chang. Discrete graph hashing. In *NIPS*, 2014.

[33] W. Liu, J. Wang, and S.-F. Chang. Hashing with graphs. In *ICML*, 2011.

[34] Y. Long, L. Liu, and L. Shao. Attribute embedding with visual-semantic ambiguity removal for zero-shot learning. In *BMVC*, 2016.

[35] Y. Long, L. Liu, and L. Shao. Towards fine-grained open zero-shot learning: Inferring unseen visual features from attributes. In *WACV*, 2017.

[36] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 9:2579–2605, Nov 2008.

[37] T. Mikolov, I. Sutskever, and K. Chen. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[38] D. Milajevs, D. Kartsaklis, M. Sadrzadeh, and M. Purver. Evaluating neural word representations in tensor-based compositional settings. *EMNLP*, 2014.

[39] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010.

[40] M. Palatucci, D. Pomerleau, G. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

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

[42] O. Pujol, P. Radeva, and J. Vitria. Discriminant ecoc: a heuristic method for application dependent design of error correcting output codes. *IEEE TPAMI*, 28(6):1007–1012, June 2006.

[43] J. Qin, L. Liu, M. Yu, Y. Wang, and L. Shao. Fast action retrieval from videos via feature disaggregation. *CVIU*, 156:104–116, 2017.

[44] J. Qin, L. Liu, Z. Zhang, Y. Wang, and L. Shao. Compressive sequential learning for action similarity labeling. *IEEE TIP*, 25(2):756–769, Feb 2016.

[45] J. Qin, Y. Wang, L. Liu, J. Chen, and L. Shao. Beyond semantic attributes: Discrete latent attributes learning for zero-shot recognition. *IEEE SPL*, 23(11):1667–1671, Nov 2016.

[46] B. Romera-Paredes and P. H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015.

[47] P. H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[48] F. Shen, W. Liu, S. Zhang, Y. Yang, and H. Tao Shen. Learning binary codes for maximum inner product search. In *ICCV*, 2015.

[49] F. Shen, C. Shen, W. Liu, and H. T. Shen. Supervised discrete hashing. In *CVPR*, 2015.

[50] F. Shen, C. Shen, Q. Shi, A. van den Hengel, and Z. Tang. Inductive hashing on manifolds. In *CVPR*, 2013.

[51] F. Shen, X. Zhou, Y. Yang, J. Song, H. T. Shen, and D. Tao. A fast optimization method for general binary code learning. *IEEE TIP*, 25(12):5610–5621, Dec 2016.

[52] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.

[53] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01*, 2012.

[54] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

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

[56] J. Wang, S. Kumar, and S. F. Chang. Semi-supervised hashing for large-scale search. *IEEE TPAMI*, 34(12):2393–2406, 2012.

[57] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, 2009.

[58] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. *arXiv preprint arXiv:1603.08895*, 2016.

[59] X. Xu, T. Hospedales, and S. Gong. Semantic embedding space for zero-shot action recognition. In *ICIP*, pages 63–67, 2015.

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

[61] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S. F. Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013.

[62] X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *ECCV*, 2010.

[63] X. Zhang, L. Liang, and H.-Y. Shum. Spectral error correcting output codes for efficient multiclass recognition. In *ICCV*, 2009.

[64] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, 2016.

[65] B. Zhao and E. P. Xing. Sparse output coding for scalable visual recognition. *IJCV*, 119(1):60–75, 2016.