Zero-Shot Action Recognition with Error-Correcting Output Codes

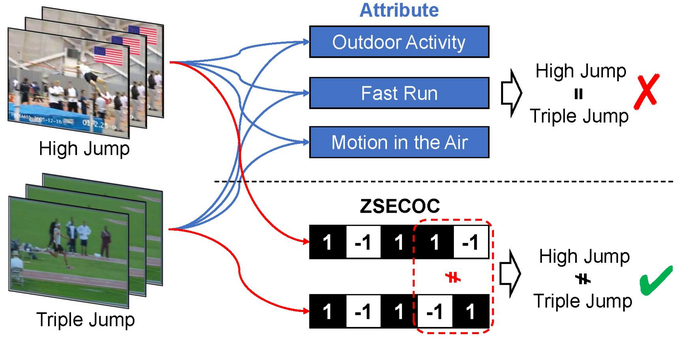
Jie Qin[1,2], Li Liu[3,4], Ling Shao[4], Fumin Shen[5], Bingbing Ni[6], Jiaxin Chen[1] and Yunhong Wang[1]

[1]Beihang University [2]ETH Zurich [3]Malong Technologies Co., Ltd. [4]University of East Anglia
[5]University of Electronic Science and Technology of China [6]Shanghai Jiao Tong University

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 2017 HONOLULU

Motivation:

➢ *Dynamic* actions are more complex than visual objects, making traditional attribute/word vector-based zero-shot learning (ZSL) approaches lack of discrimination and scalability:



- Attributes ⟶ *manually specified* and *highly subjective*
- Word Vectors ⟶ *high-dimensional* and *visual data-irrelevant*

➢ The *superior properties* of error-correcting output codes (ECOC) make it favorable for tackling ZSL problems: *e.g., error-correcting ability, high efficiency, good diversity, and accurate binary classification*

Contributions:

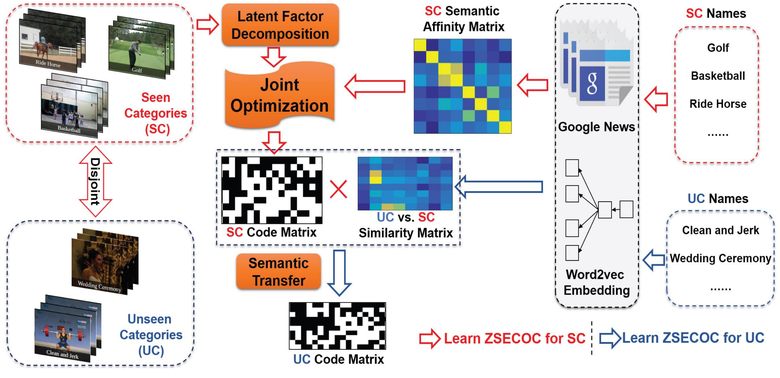
(1) Propose the Zero-Shot ECOC (dubbed ZSECOC): ECOC with the additional capability of zero-shot recognition

(2) ZSECOC inherits the intrinsic advantages of ECOC as well as overcomes the domain shift problem implicitly and explicitly

(3) ZSECOC preserves *instance-level* visual data structures in addition to *category-level* semantics

(4) ZSECOC is systematically evaluated on three realistic video action datasets, i.e., *Olympic Sports, HMDB51* and *UCF101*

Approach:



➢ **ZSECOC for seen categories:**

$$\min_{\mathbf{R},\mathbf{V},\mathbf{B}} \mathcal{O}(\mathbf{R},\mathbf{V},\mathbf{B}) := \underbrace{\mathcal{O}_{\mathrm{dd}}}_{(1)} + \beta \underbrace{\mathcal{O}_{\mathrm{sp}}}_{(2)}$$

$$\text{s.t. } \mathbf{R}^\top\mathbf{R} = \mathbf{I}, \mathbf{B}\mathbf{B}^\top = C\mathbf{I}, \mathbf{B}\mathbf{1} = \mathbf{0}, \mathbf{B} \in \{-1, 1\}^{m \times C}$$

(1) Capturing visual data structures (data-driven):

$$\mathcal{O}_{\mathrm{dd}} := ||\mathbf{X} - \mathbf{D}\mathbf{V}||_{\mathrm{F}}^2 + \gamma||\mathbf{V}||_{\mathrm{F}}^2 + \alpha||\mathbf{B}\mathbf{P} - \mathbf{R}\mathbf{V}||_{\mathrm{F}}^2$$

(2) Preserving category-level semantics (semantics-preserving):

$$\mathcal{O}_{\mathrm{sp}} := \sum s_{ij}||\mathbf{b}_i - \mathbf{b}_j||_2^2 - \lambda \sum ||\mathbf{b}_i - \mathbf{b}_j||_2^2$$

- Optimization: *alternatively* find the solution to each variable

➢ **ZSECOC for unseen categories:**

- Semantic transfer w/o unseen visual data:

$$\mathbf{B}^u = \mathrm{sign}(\mathbf{B}\mathbf{S}^u) \quad s_{ij}^u = \frac{< \phi(y_i), \phi(y_j^u) >}{||\phi(y_i)|| \cdot ||\phi(y_j^u)||}$$

➢ **Zero-shot recognition as Hamming decoding:**

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

Experiments:

- Dataset & Feature: Olympic Sports, HMDB51 and UCF101; IDT + FV
- Semantic embedding: 300-d word vectors via the skip-gram model
- Visual-semantic mapping: a set of independent linear SVMs

➢ **Quantitative results:**

(1) ZSECOC vs. other embeddings

Embedding	Olympic Sports	HMDB51	UCF101
Word Vector	21.6±0.9	16.5±3.9	3.2±0.7
Attribute	27.7±4.6	N/A	13.7±0.5
ZSECOC	**59.8±5.6**	**22.6±1.2**	**15.1±1.7**

(2) ZSECOC vs. other ECOC

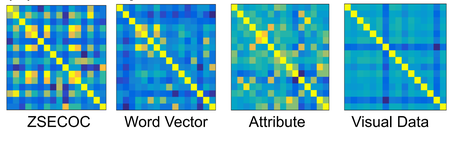
Method	Olympic Sports	HMDB51	UCF101
RSECOC [3]	18.4±0.6	5.3 ±0.1	3.0±0.8
RDECOC [3]	25.3±1.8	6.2±1.0	2.7±0.2
DECOC [42]	40.1±4.2	6.9±1.0	4.8±0.6
Forest-ECOC [9]	51.0±8.7	9.2±0.7	5.9±0.5
ZSECOC	**59.8±5.6**	**22.6±1.2**	**15.1±1.7**

(3) ZSECOC vs. state-of-the-art

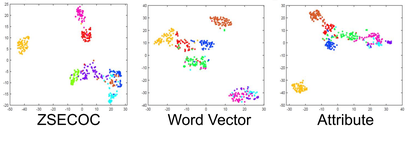
Method	Reference	Feature	Label Embedding	TD	Olympic Sports	HMDB51	UCF101
HAA [28]	CVPR 2011	FV	A	×	46.1±12.4	N/A	14.9±0.8
DAP [26]	TPAMI 2014	FV	A	×	45.4±12.8	N/A	14.3±1.9
IAP [26]	TPAMI 2014	FV	A	×	42.3±12.5	N/A	12.8±2.0
ST [59]	ICIP 2015	BoW	WV	×	N/A	13.0±2.7	10.9±1.5
ST [59]	ICIP 2015	BoW	WV	✓	N/A	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)

➢ **Qualitative results:**

(1) Similarity matrices



(2) t-SNE of unseen categories



(3) Top-5 returned video examples for unseen categories