Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Concepts Not Alone: Exploring Pairwise Relationships for Zero-Shot Video Activity Recognition

Chuang Gan[1], **Ming Lin**[3], **Yi Yang**[2], **Gerard de Melo**[1] and **Alexander G. Hauptmann**[4]
[1] IIIS, Tsinghua University, Beijing, China [2] QCIS, University of Technology Sydney, Sydney, Australia
[3] DCM&B, University of Michigan, Ann Arbor, USA [4] SCS, Carnegie Mellon University, Pittsburgh, USA

Abstract

Vast quantities of videos are now being captured at astonishing rates, but the majority of these are not labelled. To cope with such data, we consider the task of content-based activity recognition in videos without any manually labelled examples, also known as zero-shot video recognition. To achieve this, videos are represented in terms of detected visual concepts, which are then scored as relevant or irrelevant according to their similarity with a given textual query. In this paper, we propose a more robust approach for scoring concepts in order to alleviate many of the brittleness and low precision problems of previous work. Not only do we jointly consider semantic relatedness, visual reliability, and discriminative power. To handle noise and non-linearities in the ranking scores of the selected concepts, we propose a novel pairwise order matrix approach for score aggregation. Extensive experiments on the large-scale TRECVID Multimedia Event Detection data show the superiority of our approach.

1 Introduction

Motivation. The increasing ubiquity of devices capable of capturing videos has led to an explosion in the amount of recorded video content. Smartphones, action cameras, as well as surveillance cameras mean that ever-increasing amounts of our daily activities are captured on videos. Due to the torrential volume of this data, the vast majority of videos are never labeled. Moreover, even for those that are shared online, the human-supplied metadata is often vague or unspecific (e.g. "Albufeira, Summer 2015"). Unfortunately, video search engines such as Youtube, Yahoo, and Bing, crucially depend on textual keyword matching. Their approach works well for popular videos but fails hopelessly for long-tail content or personal video collections with insufficient metadata.

Fortunately, encouraging progress has been made on content-based video analysis in recent years. Standard approaches rely on low-level audio/visual input features that are fed into machine learning algorithms such as support vector machines (SVMs) (Chang and Lin 2011) or deep convolutional neural networks (Karpathy et al. 2014; Gan et al. 2015b). These achieve promising results when there are sufficient numbers of labeled training examples for every search query of interest.

However, due to the large number of possible search queries, query-specific training is not always feasible. Zero-shot learning (Lampert, Nickisch, and Harmeling 2009) addresses this problem by providing an alternative paradigm that does not require positive training exemplars for every class of videos. Given a textual query, one aims to retrieve videos that are most relevant to it, exploiting visual attributes of the videos.

While several algorithms (Jiang et al. 2015a; Wu et al. 2014; Dalton, Allan, and Mirajkar 2013; Habibian, Mensink, and Snoek 2014; Liu et al. 2013; Singh et al. 2015) have recently been proposed for such zero-shot video activity recognition, state-of-the-art systems still suffer from brittleness and low precision.

Contributions. In this paper, we show how to make zero-shot learning more robust. Similar to previous work, our system consists of two main components. The first of these aims at a semantic query interpretation, in which the system selects concepts pertaining to the query description from a large pool of potential candidate concepts. The second component produces an aggregation of the individual concept-specific video ranking lists. We propose important strategies for making both more robust:

- We propose a simple yet effective concept selection approach in representing queries. Unlike previous work, concept reliability and discriminative power are considered as critical indicators in order to ensure robust zero-shot activity recognition.
- We devise a novel robust video ranking approach that relies on the recovery of a low-rank order matrix from multiple pairwise order matrices for different concept ranking lists.
- Experimental results on challenging unconstrained video data confirm that the proposed system outperforms the state-of-the-art zero-shot approaches.

2 Related Work

Video analysis has attracted a lot of research interest in the past decade. A recent review can be found in (Jiang et al. 2013). Standard video activity recognition systems, despite their reasonable recognition performance, rely on custom low-level representations, such as improved dense trajectories (Wang and Schmid 2013), and Mel-FrequencyCepstral Coefficients (MFCC) (Rabiner and Schafer 2007). These

suffer from several deal-breaking drawbacks. First, they are incapable of providing a semantic interpretation of a video. Second, because of their high dimensionality, training effective event classifiers on the low-level representation requires a substantial number of training examples per class. When only a few or even no positive examples are available, the power of low-level representations is limited. In our work, we instead draw concept detection to derive higher-level representations.

Semantic video representations describe a video in terms of pre-defined pools of activity concepts and attributes, and have proven both robust for video activity recognition and interpretable by humans (Hauptmann et al. 2007). This form of representation has inspired the development of zero-shot video activity recognition. Most existing zero-shot learning frameworks follow a two-stage classification approach. Given a novel class, first its semantic properties (i.e., attributes or concepts) are identified, then its class label is predicted as a ranking function of those attributes or concepts. To identify relevant concepts, most of the current approaches rely on manually defined concept schemes or blindly adopt inventories based on some outside knowledge sources, such as WordNet (Lin 1998), Wikipedia (Gabrilovich and Markovitch 2007), or label co-occurrence data (Mensink, Gavves, and Snoek). In contrast to our approach, none of the existing approaches take the visual reliability and discriminative power into consideration. As we will see in Section 4, the selection of concepts has a dramatic impact on the retrieval results.

When fusing rankings, most existing approaches to zero-shot activity recognition directly combine the raw ranking scores of related attributes or concepts as the final ranking list. Unlike these methods, we propose a more robust ranking function, by first converting the raw ranking score into scale-invariant pairwise order matrices and then decomposing the multiple pairwise order matrices into a single shared order matrix, which is then consulted to generate the ranking list. Our experiments confirm that this leads to more robust ranking results.

Our work is also related to the sentence generation task (Guadarrama et al. 2013; Sun, Gan, and Nevatia 2015), where the goal is to generate natural language descriptions of a video. Our goal, in contrast, is to address the opposite direction: given a textual query, we seek to retrieve videos that match the query. Matrix factorization is also related to our work, which has been widely used in different tasks (Ye et al. 2012; Fan et al. 2014; Yan et al. 2015).

3 Proposed Method

3.1 Overview

In our approach, we first represent both the video data and the textual description by embedding them in a semantic concept space. The overall framework of our approach is outlined in Figure 1. It consists of two major components.

The first of the two components aims at a semantic query interpretation. Given a query, i.e., a textual description of the target event, we apply text analytics methods to extract salient words that describe this target class. The system selects relevant concepts matching these query words from a large pool of potential candidate concepts.

We are given a textual query q as input. Based on a vocabulary of d visual concepts, we first select relevant concepts according to their semantic similarity with q. A higher similarity score indicates that the corresponding concept is more related to the target video class denoted by q.

We also apply a bank of visual concept detectors to generate semantic representations for the videos. Thus, each video is represented as a vector whose elements are detection scores of different semantic concepts.

Having embedded both the query and the videos as numerical vectors of concepts, we select relevant concepts based on their similarity to the query, but also prune away selected semantic concepts that have insufficient reliability or discriminative power.

The second component produces a ranked list of videos. This is achieved by aggregating the individual concept-specific ranking lists. After the preliminary filtering, the detection scores of selected concepts are still noisy. To alleviate the problems of noise and non-linearity of different ranking lists, we first convert the raw ranking scores of selected concepts into pairwise order matrices, in which each entry characterizes the comparative relationship of two test samples. Our hypothesis is that the relative score relations are consistent within component ranking lists, despite the large variations that may exist in the absolute values of the raw scores. Thus, we take the pairwise order matrices of different semantic concepts as input, and recover a common shared rank-2 order matrix. Even though the original matrix might be inaccurate, the joint relations from multiple matrices may be complementary with respect to each other and hence the shared common matrix may well model the correct order. Finally, we transform the shared order matrix into a ranking list as the final retrieval result. then recovered from these pairwise matrices.

3.2 Query Interpretation

The goal of the query interpretation phase is to create a semantic representation for the query in terms of salient semantic concepts that it expresses. These are taken from a large pool of candidate concepts and should not only be relevant to the textual query but also visually discernable in the videos. It is tedious if not impossible to manually label all related concepts in a large pool of potential candidate concepts. Therefore, most recent work relies on natural language processing technologies to compute the semantic similarity between the query and the candidate concepts (Gan et al. 2015a).

While fully automatic, this approach alone is far from satisfactory, due to the well-known semantic gap between query and visual content. For instance, a given abstract concept such as *meeting* may have vastly different visual representations in different circumstances. It appears unreasonable to choose the relevant visual concepts only based on semantic similarity scores between the concept names and the textual query. In our approach, we posit that the selected relevant concepts should have the following key properties:

1. *Semantic relevance*: the selected concepts should be semantically related with the textual query.

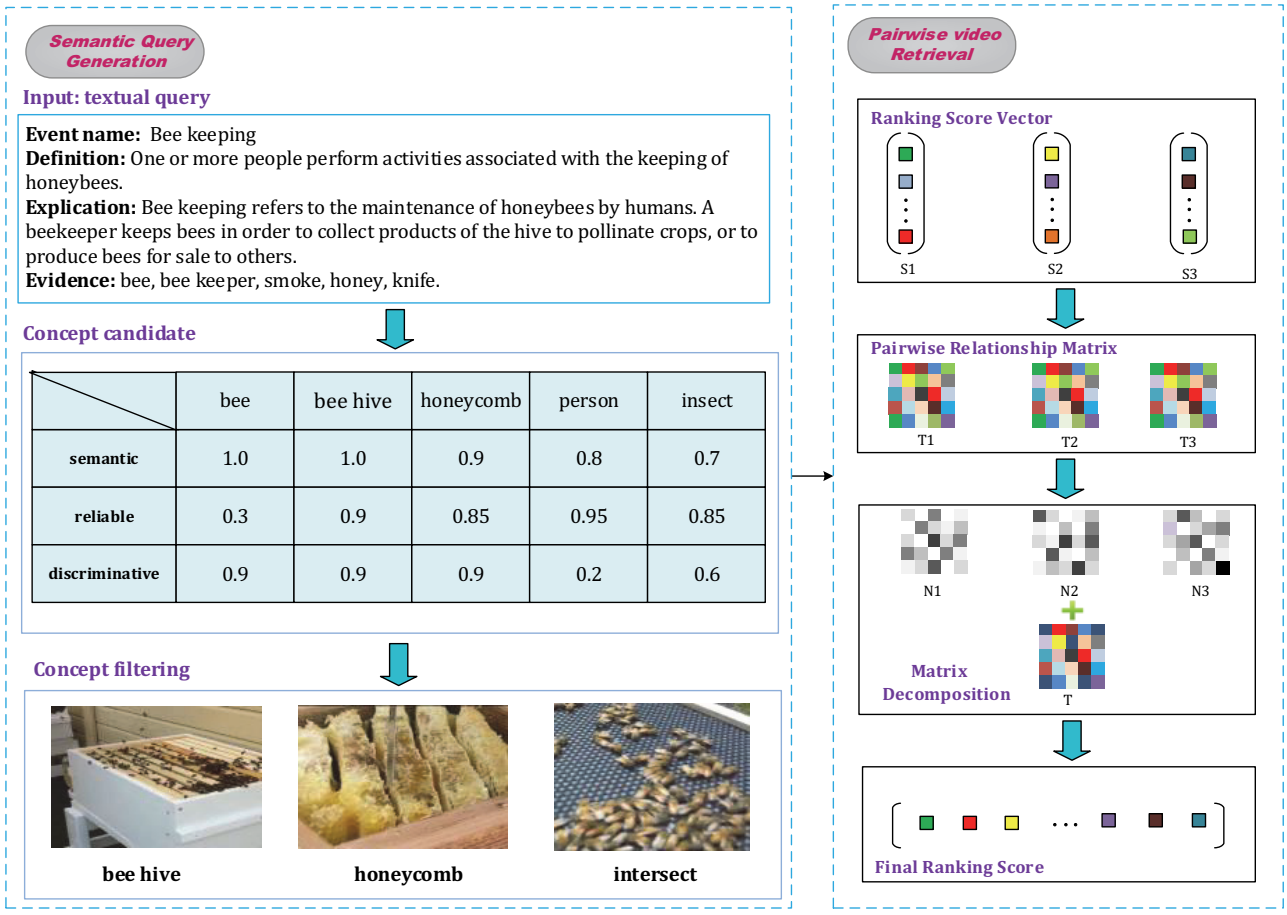

Figure 1: An illustration of our zero-shot video activity recognition framework. We first automatically compute semantic similarity scores between the query and concepts based on the cosine distances of continuous word vectors, and filter out the concepts that are not sufficiently visually reliable or discriminative for events. Then we convert the raw ranking scores of selected concepts to multiple pairwise order matrices, which are taken as input to recover a rank-2 order matrix based on low rank and skew-symmetric constraints. A robust ranking score vector is finally extracted to fit the recovered low-rank order matrix.

2. *Visual reliability*: the selected concepts should be reliable during the detection on different datasets.
3. *Event discriminativeness:* The selected concepts should be discriminative enough in the detection of the video activity.

Query Analysis. To automatically generate query terms from event descriptions, we apply standard natural language processing techniques to clean up its textual description, including removal of common stop words and lemmatization (stemming) to normalize word inflections. Then we compute the TF-IDF score of the remaining terms, and select the top 5 terms as event query terms. Next, we compute the similarity of query terms with concept terms as described below, and average over all query terms.

Semantic Similarity Computation. The semantic similarity computation requires us to have trained a model that can quantify the degree of semantic similarity between two words. This can conveniently be done beforehand in an offline process. We draw on the recent success of the skip-gram with negative sampling neural network model (Mikolov et al. 2013). Given a large text corpus such as Wikipedia, the objective is to produce vectors that represent words such that vector similarity reflects word similarities. The training objective of the skip-gram model achieves this by optimizing for word representations that allow for predicting the surrounding context words in a sentence. More precisely, given a sequence of words $\{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_d\}$, it searches for a vector representation for each word $\mathbf{w}_i$ such that

$$\frac{1}{d}\sum_{t=1}^{d}\sum_{-c\leq j\leq c, j\neq 0}\log(P(\mathbf{w}_{t+j}|\mathbf{w}_t)) \qquad (1)$$

is maximized, where c controls the context window length. The probability of $\mathbf{w}_{t+j}$ given $\mathbf{w}_t$ is defined by the softmax

function

$$P(\mathbf{w}_i|\mathbf{w}_j) = \frac{\exp({\mathbf{w}_i}^T \mathbf{w}_j)}{\sum_{\mathbf{w}} \exp({\mathbf{w}_i}^T \mathbf{w}_j)s} \quad (2)$$

In order to optimize this more efficiently, a binary Huffman tree is used to predict words from the vocabulary, and training is carried out with stochastic gradient ascent, using negative sampling to limit the number of predictions (Mikolov et al. 2013).

Once we have optimized the word vector representations using the corpus, we can measure the similarity between words via the standard cosine measure. The larger the cosine score between two word vectors, the more they are deemed semantically related.

Reliability and Discriminativeness Validation. While concepts selected by the above pipeline are likely semantically related, they need not be visually reliable or discriminative enough for activity recognition. To test the reliability, we use two-fold drop-out cross validation. The averaged precision based on the two-fold cross validation reflects the reliability of the tested concept. We filter out concepts whose precision is below a threshold (set to 80% in all our experiments to reasonably balance precision and recall).

The discriminative power is assessed using detection scores on held-out data. Concepts that have detection scores over $\frac{1}{2}$ for over 50% of all activity classes are deleted. For example, *person* is semantically related to the event *bee keeping* and has a high reliability score. However, it obtains high scores on most of the videos, so the term lacks discriminative power for discerning more specific activities. Therefore, we may prune such concepts from the concept pool.

3.3 Pairwise Order Matrix Construction

The next step is to use semantic representation of the query to retrieve and rank a set of videos. For retrieval, most of existing approaches (Jiang et al. 2015a; Wu et al. 2014; Dalton, Allan, and Mirajkar 2013; Habibian, Mensink, and Snoek 2014; Liu et al. 2013) naively average the detection scores of selected relevant concepts. This leads to a suboptimal solution. First, the scale of different detection scores for different concepts is not comparable even after normalization. Thus, it is unwise to use the same weight when fusing them, but manually fine-tuning weights is also not possible in practice. Second, even within the same concept, the detection scores are not necessarily linear. For example, we may not be able to discern any apparent difference between 0.5 and 0.9, yet perceive a marked difference between 0.1 and 0.15.

Assume that we have d concepts, n videos in the system. We apply d concept detectors to these videos. Our implementation uses a deep Convolutional Neural Network (CNNs) architecture (Krizhevsky, Sutskever, and Hinton 2012). We take the key frames of a given test video as input, run a forward pass through the CNN, and use the softmax score as the concept detection score on the key frame. To arrive at a video-level representation, we rely on simple average pooling.

This process yields a detection score matrix $X \in \mathbb{R}^{d\times n}$. Each column X_i stores the detection scores for the i-th video with respect to all d concepts. For the k^{th} concept, we construct a pairwise order matrix $T^{(k)}$,

$$T^{(k)}_{i,j} = \text{sign}(X_{k,i} - X_{k,j}) \ .$$

The matrix $T^{(k)}$ encodes the pairwise order of every two videos measured under the k-th concept. In particular, $T_{i,j} = 1$ indicates that the i-th sample is more detected as positive with greater confidence than the j-th sample for the k-th concept, while $T_{i,j} = -1$ indicates the opposite comparative relation. Meanwhile, $T_{i,j} = 0$ indicates that the i-th sample has a similar detection confidence value. As this order matrix captures relative assessments between different samples, it is not influenced by the scale or non-linearity of the detection scores within or between concepts.

Assume that there is a ground truth ranking score denoted as $\mathbf{s}$. The corresponding pairwise order matrix $\hat{T}$ is defined by

$$\hat{T} = \mathbf{s}\mathbf{e}^{\mathrm{T}} - \mathbf{e}\mathbf{s}^{\mathrm{T}} \quad (3)$$

where $\mathbf{e} = [1, 1, \cdots]^{\mathrm{T}}$. The next proposition shows that the rank of $\hat{T}$ is exactly 2.

Proposition 1. *The rank of $\hat{T}$ is exactly 2 when the scores in $\mathbf{s}$ are not all equal to a constant value.*

Proof. It is easy to confirm that $\text{rank}(\hat{T}) \leq 2$ from Eq. (3). If there is some $\mathbf{s}$ such that $\hat{T}$ is rank 1, then $\hat{T}_i = c_i\hat{T}_1$ for some constant c_i. Since $\hat{T} = -\hat{T}^{\mathrm{T}}$ and $\hat{T}_{i,i} = 0$, we have $\hat{T} = 0$. □

Proposition 1 entails that $\text{rank}(\hat{T}) = 2$ is a restricted convex relaxation with respect to directly optimizing $\mathbf{s}$ (Yuan, Li, and Zhang 2014). This suggests that we may use rank-2 hard iterative singular value thresholding, as we will explain shortly in more detail.

Based on the above discussion, we now turn to studying how to estimate $\hat{T}$ from a set of pairwise order matrices $T^{(k)}, k = 1, 2, \cdots, d$, with rank-2 and skew-symmetric constraints.

3.4 Recovering the Pairwise Order Matrix

Assume that we have a set of d pairwise order matrices $\hat{T}^{(1)}, \hat{T}^{(2)}, \cdots, \hat{T}^{(k)}$. Because the detection scores are noisy, the order constraints of $\hat{T}^{(k)}$ may contradict each other. We need a robust approach to recover the matrix $\hat{T}$ by adaptively fusing $\hat{T}^{(k)}$. To this end, each time $\hat{T}_{i,j}$ violates the given order $T^{(k)}_{i,j}$, we penalize it with a loss function $\ell(\hat{T}_{i,j}, T^{(k)}_{i,j})$. To maximize the margin, we use the hinge loss in this paper (although theoretically any other loss function is applicable):

$$\ell(\hat{T}, T^{(k)}) \triangleq \sum_{i,j=1}^{n} [1 - T^{(k)}_{i,j}\hat{T}_{i,j}]_+ \quad (4)$$

where $[z]_+ = \max(z, 0)$. Then we have the following optimization problem to recover $\hat{T}$:

Algorithm 1 Hard Iterative Singular Value Thresholding

1: **Input:** $T^{(k)}$, step size η.
2: $\hat{T}_0 = 0$
3: **for** $i = 1, 2, \cdots, L$ **do**
4: $\quad \hat{G}_t = \hat{T}_{t-1} - \eta \nabla_{\hat{T}} \ell(\hat{T})$.
5: $\quad \widetilde{G}_t = \frac{1}{2}(\hat{G}_t - \hat{G}_t^{\mathrm{T}})$
6: $\quad$ SVD: $\widetilde{G}_t = \sum_i \lambda_i U_i V_i^{\mathrm{T}}$
7: $\quad$ rank-2 thresholding: $\hat{T}_t = \lambda_1 U_1 V_1^{\mathrm{T}} + \lambda_2 U_2 V_2^{\mathrm{T}}$
8: **end for**
9: **Output:** $\hat{T} = \hat{T}_L$

$$\min_{\hat{T}} \quad \ell(\hat{T}) \triangleq \sum_{k=1}^{d} \ell(\hat{T}, T^{(k)}) + \lambda \|\hat{T}\|^2 \tag{5}$$

$$\text{s.t.} \quad \operatorname{rank}(\hat{T}) = 2 \tag{6}$$

$$\hat{T} = -\hat{T} \, . \tag{7}$$

The above optimization can be effectively solved with hard iterative singular value thresholding, as depicted in Algorithm 1. It is not difficult to show that Algorithm 1 converges geometrically to the global optimal because the optimization problem is restricted convex.

In Algorithm 1, we first carry out a gradient descent step with respect to $\hat{T}$ with step size η. In line 5, we project the intermediate solution onto skew-symmetric subspace. In lines 6 and 7, we greedily threshold the intermediate solution with the rank-2 constraint. It is important to note that line 5 must be above lines 6 and 7 because a skew-symmetric matrix is still skew-symmetric after rank-2 singular value thresholding but the reverse does not hold.

After getting the optimized matrix $\hat{T}$, we seek to recover the final ranking score vector $\hat{s}$. Based on our rank-2 assumption mentioned above, we expect that $\hat{T}$ is generated from $\hat{s}$ as $\hat{T} = \hat{s}e^T - e\hat{s}^T$. The authors in (Jiang et al. 2011) have shown that using $(1/m)\hat{T}e$ as the recovered s will provide the best least-square approximation, which can be formally described as follows:

$$(1/m)\mathbf{e}\hat{T} = \underset{\hat{\mathbf{s}}}{\operatorname{argmin}} ||\hat{T} - (\hat{\mathbf{s}}\mathbf{e}^{\mathbf{T}} + \mathbf{e}\hat{\mathbf{s}}^{\mathbf{T}})||. \tag{8}$$

Therefore, we can treat $(1/m)\mathbf{e}\hat{T}$ as the recovered $\hat{\mathbf{s}}$ after the retrieval, giving us our final results.

3.5 Out-of-Sample Extension

We can additionally also deal with the case of new out-of-sample test videos. Given a new test video x_{m+1}, we first semantically represent it as an n-dimensional vector. For each dimension, we find its nearest neighbour from the existing test data $X = \{x_1, x_2, ..., x_m\}$. Let x^i denote the nearest example for the i-th semantic concept, and w_i denote the feature similarity based on i-th semantic feature type. Then, the ranking score of x_{m+1} can be computed as $\hat{s}(x_{m+1}) \sum_{i=1}^{n} \frac{w(t^i_{m+1}, x^i)}{\sum_{i=1}^{n} w(t^i_{m+1}, x^i)} \hat{s}(x^i)$, where x^i is the aggregated score for sample x_i.

4 Experiments

4.1 Experimental Setup

Dataset and Metrics. We conduct experiments on the challenging TRECVID Multimedia Event Detection datasets from 2013 (MED13test) and 2014 (MED14test). Each includes 25,000 testing videos (over 960 hours of video) with per-video ground truth annotations for 20 event categories, all officially provided by NIST. Each category has a textual description in the form of event name, definition, explication, and related evidence types. Since we focus on zero-shot event detection, the experiments are conducted without using any examples. To evaluate the results, we apply the official metric: average precision (AP) per event, and mean Average Precision (mAP) by averaging all 20 events.

Image-based Concepts. We obtain 1000 image-based concept detectors using a deep Convolutional Neural Network (CNNs) (Krizhevsky, Sutskever, and Hinton 2012). We use the VGG19 Net (Simonyan and Zisserman 2015) architecture, as implemented in the Caffe (Jia 2013) toolbox. The network is trained on the ImageNet ILSVRC-2014 dataset (Deng et al. 2009), which includes 1.2M training images categorized into 1000 classes.

Video-based Concepts. We also obtain video-based concepts from four publicly available datasets: UCF101 (Soomro, Zamir, and Shah 2012), FCVID (Jiang et al. 2015b), Google Sports1M (Karpathy et al. 2014), and ActivityNet (Heilbron et al.). They contain 101 action categories, 239 action categories, 487 sports categories, and 203 activity categories respectively. We extract the improved dense trajectory (Wang and Schmid 2013) features from videos, and aggregate the local features into video-level feature vectors by Fisher vectors (Oneata et al. 2013). We train linear SVM classifiers and employ 5-fold cross validation to select the parameters.

Held-Out Data. In order to obtain the discriminative power scores, we test on the UCF101 dataset (*crcv.ucf.edu/data/*).

4.2 Experimental Results

Comparison with Previous Work. In Table 1, we compare our approach with other recent state-of-the-art systems, specifically the Bi-Concept approach (Habibian, Mensink, and Snoek 2014), EventNet (Ye et al.), the weak concepts approach (Wu et al. 2014), Selecting (Singh et al. 2015), and SPaR (Jiang et al. 2014). The first three of these only rely on concept aggregation, while Selecting and SPaR combine concept aggregation and re-ranking strategies. We report results on MEDtest13, as this allows us to directly quote the values given in the original papers, for fairness. The results are comparable, as we use the same data split. To better analyse our approach, we also implement a traditional attribute-based retrieval approach ("Basic") (Gan et al. 2015a). However, for a fair comparison, we use our concept features, as these are stronger, as shown later on.

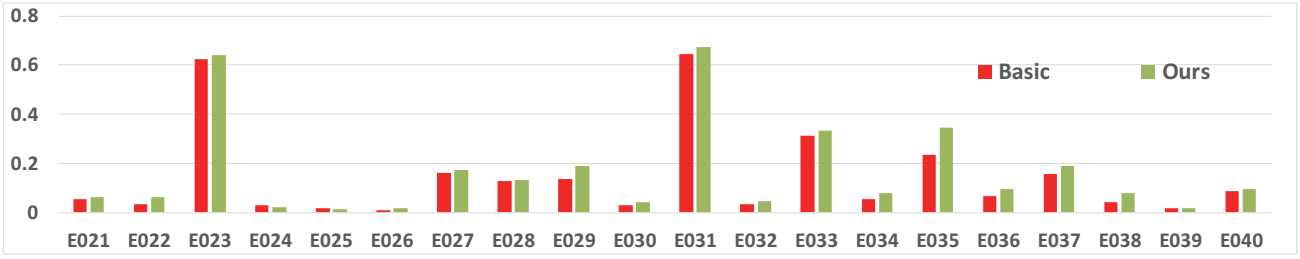

Figure 2: Per-event compared results on TRECVID MED 2014 dataset.

Method	mAP (%)
Bi-concept	6.0
EventNet	8.9
Weak concept	12.7
Selecting	11.8
SPaR	12.9
Basic	14.4
Ours	**16.9**

Table 1: Comparisons with other stat-of-the-art zero-shot event detection systems on MEDtest13.

Dataset	word2vec	Ours
MED13test	14.4	**16.9**
MED14test	13.8	**16.5**

Table 2: Compare mAP (%) scores of query interpretations.

We observe that our proposed algorithm significantly outperforms the previous approaches. Besides our matrix pairwise order ranking approach, we find that our discriminative and reliable concepts filtering is also important factors for obtaining good zero-shot event detection performances. In addition, we also find that the large pre-defined concept pool containing relevant concepts plays a critical role for the sate-of-the-art zero-shot event detection systems. This is why the basic attribute-based approach without concept filtering and rank aggregation also outperforms previous systems.

Detailed Analysis. To isolate the contribution of the different parts, we first evaluate the effectiveness of the proposed semantic query interpretation approach. We compare the experimental results with the "word2vec" approach. We use the top 5 semantically related concepts (without concept selection) to perform the retrieval. The original scores are listed as "word2vec" in Table 2. After removing the concepts that are unreliable or insufficiently discriminative, we find that the retrieval results improve significantly. For additional analysis, we also provide event class-specific results on TRECVID MED 2014 dataset in Figure 2. We observe that for 18 out of 20 classes, our new framework outperforms the attribute-based approach, confirming that our proposed framework makes the traditional zero-shot retrieval system more robust.

Finally, we performed an experiment in which a random noise ranking list was added before the retrieval. The results in Table 3, in comparison with the original scores, reveal that the naive attribute-based approach ("Basic") degrades significantly, from 0.144 to 0.106 on MED13test and from 0.138 to 0.112 on MED14test. In contrast, our pairwise ranking approach was only influenced to a small degree. These results thus further confirm the robustness of the proposed approach.

Dataset	Basic	Ours	Dataset	Basic	Ours
MED13test	10.6	**15.7**	MED14test	11.2	**15.2**

Table 3: mAP (%) scores after adding random noise.

5 Conclusion

In this paper, we have proposed a novel algorithm for zero-shot video activity recognition. We describe a simple but effective concept selection approach, considering semantic relatedness, visual reliability, and discriminative power. Additionally, in order to alleviate the noisiness and non-linearity of the raw ranking scores, we convert the rankings for different concepts into multiple pairwise order matrices, and reconstruct the proper order matrix based on low-rank and skew-symmetric constraints. Finally, we convert that matrix back into a ranking list.

Experimental results show that each of these contributions lead to improvements, outperforming previous work. Hence, these ideas have an important impact on advancing the state-of-the-art work in this area. In the more immediate future, we will consider adding multimodal features (i.e., speech and audio concepts) into our framework.

Acknowledgments. This work was supported in part by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61361136003.

References

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Dalton, J.; Allan, J.; and Mirajkar, P. 2013. Zero-shot video retrieval using content and concepts. In *CIKM*, 1857–1860.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Fan, M.; Zhao, D.; Zhou, Q.; Liu, Z.; Zheng, T. F.; and Chang, E. Y. 2014. Distant supervision for relation extraction with matrix completion. In *ACL*, volume 1, 839–849.

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, 1606–1611.

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

Gan, C.; Wang, N.; Yang, Y.; Yeung, D.-Y.; and Hauptmann, A. G. 2015b. DevNet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*, 2568–2577.

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

Habibian, A.; Mensink, T.; and Snoek, C. G. 2014. Composite concept discovery for zero-shot video event detection. In *ICMR*, 17.

Hauptmann, A.; Yan, R.; Lin, W.-H.; Christel, M.; and Wactlar, H. 2007. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *Multimedia, IEEE Transactions on* 9(5):958–966.

Heilbron, F. C.; Escorcia, V.; Ghanem, B.; and Niebles, J. C. Activitynet: A large-scale video benchmark for human activity understanding.

Jia, Y. 2013. Caffe: An open source convolutional architecture for fast feature embedding. *http://caffe. berkeleyvision. org*.

Jiang, X.; Lim, L.-H.; Yao, Y.; and Ye, Y. 2011. Statistical ranking and combinatorial hodge theory. *Mathematical Programming* 127(1):203–244.

Jiang, Y.-G.; Bhattacharya, S.; Chang, S.-F.; and Shah, M. 2013. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval* 2(2):73–101.

Jiang, L.; Meng, D.; Mitamura, T.; and Hauptmann, A. G. 2014. Easy samples first: Self-paced reranking for zero-example multimedia search. In *Multimedia*, 547–556. ACM.

Jiang, L.; Yu, S.-I.; Meng, D.; Mitamura, T.; and Hauptmann, A. G. 2015a. Bridging the ultimate semantic gap: A semantic search engine for internet videos. *ICMR*.

Jiang, Y.-G.; Wu, Z.; Wang, J.; Xue, X.; and Chang, S.-F. 2015b. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *arXiv preprint arXiv:1502.07209*.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *CVPR)*, 1725–1732.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1097–1105.

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

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

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

Mensink, T.; Gavves, E.; and Snoek, C. G. M. Costa: Co-occurrence statistics for zero-shot classification.

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

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

Rabiner, L. R., and Schafer, R. W. 2007. Introduction to digital speech processing. *Foundations and trends in signal processing* 1(1):1–194.

Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. *ICLR*.

Singh, B.; Han, X.; Wu, Z.; Morariu, V. I.; and Davis, L. S. 2015. Selecting relevant web trained concepts for automated event retrieval. *ICCV*.

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

Sun, C.; Gan, C.; and Nevatia, R. 2015. Automatic concept discovery from parallel text and visual corpora. *ICCV*.

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

Wu, S.; Bondugula, S.; Luisier, F.; Zhuang, X.; and Natarajan, P. 2014. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*, 2665–2672.

Yan, Y.; Tan, M.; Tsang, I.; Yang, Y.; Zhang, C.; and Shi, Q. 2015. Scalable maximum margin matrix factorization by active riemannian subspace search. In *IJCAI*.

Ye, G.; Li, Y.; Xu, H.; Liu, D.; and Chang, S.-F. Eventnet: A large scale structured concept library for complex event detection in video. *Multimedia*.

Ye, G.; Liu, D.; Jhuo, I.-H.; Chang, S.-F.; et al. 2012. Robust late fusion with rank minimization. In *CVPR*, 3021–3028.

Yuan, X.; Li, P.; and Zhang, T. 2014. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *ICML*, 127–135.