Video2vec Embeddings Recognize Events when Examples are Scarce

Habibian, A.; Mensink, T.; Snoek, C.G.M.

Published in:
IEEE Transactions on Pattern Analysis and Machine Intelligence

DOI:
10.1109/TPAMI.2016.2627563

Citation for published version (APA):
Video2vec Embeddings Recognize Events when Examples are Scarce

Amirhossein Habibian, Thomas Mensink, and Cees G. M. Snoek

Abstract—This paper aims for event recognition when video examples are scarce or even completely absent. The key in such a challenging setting is a semantic video representation. Rather than building the representation from individual attribute detectors and their annotations, we propose to learn the entire representation from freely available web videos and their descriptions using an embedding between video features and term vectors. In our proposed embedding, which we call Video2vec, the correlations between the words are utilized to learn a more effective representation by optimizing a joint objective balancing descriptiveness and predictability. We show how learning the Video2vec using a multimodal predictability loss, including appearance, motion and audio features, results in a better predictable representation. We also propose an event specific variant of Video2vec to learn a more accurate representation for the words, which are indicative of the event, by introducing a term sensitive descriptiveness loss. Our experiments on three challenging collections of web videos from the NIST TRECVID Multimedia Event Detection and Columbia Consumer Videos datasets demonstrate: i) the advantages of Video2vec over representations using attributes or alternative embeddings, ii) the benefit of fusing video modalities by an embedding over common strategies, iii) the complementarity of term sensitive descriptiveness and multimodal predictability for event recognition. By its ability to improve predictability of present day audio-visual video features, while at the same time maximizing their semantic descriptiveness, Video2vec leads to state-of-the-art accuracy for both few- and zero-example recognition of events in video.

1 INTRODUCTION

This paper strives to recognize events such as parking a vehicle, flash mob, and renovating a home in web video content. A problem of increasing importance in a world that is swiftly adapting to video communication. The leading computer vision and multimedia retrieval solutions for this challenging problem, e.g., [35], [53], [68], [74], [81], all learn to recognize events with the aid of many labeled video examples. However, as events become more and more specific, we consider it unrealistic to assume that ample examples to learn from will be commonly available. In practical use only a handful of video examples, an event name and an event definition may be present, such as the ones in Figure 1. We aim for event recognition when video examples are scarce or even completely absent.

Recognizing events from few or zero examples imposes constraints on the video representation. End-to-end deep learning of the video representation [30], [39], [71] demands too many video examples. In [39] for example, Karpathy et al. exploit more than 1 million YouTube videos and their sport category labels to learn the video representation. Alternatively, Xu et al. [81] show how a very deep convolutional neural network (CNN) [63] intended for image classification can be leveraged as representation for event recognition in video. They use responses from intermediate layers of the network to represent frames, which are aggregated over the video using VLAD encoding [29]. When combined with a linear SVM, excellent results on the leading NIST TRECVID event detection benchmarks [57] are reported for scenarios where many and few examples are available. The CNN video representation outperforms more traditional video encodings such as improved dense trajectories [74], [75] and representations combining appearance, motion and audio features [52], [55], [84]. However, both the learned and engineered representations are incapable, nor intended, to recognize events when examples are completely absent. We propose a video representation that can leverage any underlying feature, be it a CNN, improved dense trajectories and/or audio features, while being capable of few- and zero-example event recognition with state-of-the-art accuracy.

The key to few- and zero-example event recognition is to add meaning to the video representation. Inspired by the success in image classification [17], [41], [58], many rely on the predictions made by a set of attribute classifiers [5], [16], [22], [26], [34], [44]–[46], [48], [50], [82]. In [22] Habibian et al. study the properties of 1,346 attribute classifiers trained from ImageNet [61] and TRECVID [57] for representing and recognizing events in web video. Rather than pre-specifying and manually labeling each individual attribute in advance, the attributes can be learned on top of imagery and (weak) labels harvested from the web [10], [12], [79], [83]. To assure visual predictability of the discovered attribute detectors a common tactic is to leverage part of the harvested data for validation [4]. A drawback is that many attribute labels rarely occur. For these infrequent labels only a limited number of positive examples are available, which leads to a biased estimation of their classification reliability. Hence, many of the discovered attributes overfit to their small training set and do not generalize well for new videos. We also discover our semantic representation from the web, but rather than selecting individual, and often unreliable, classifiers per attribute label, we prefer to combine labels automatically into more predictable attributes. By doing so, more training examples are available and a more robust representation is obtained, without losing descriptive ability.

In this paper we present a semantic representation learn-
Event Name: Winning a race without a vehicle
Definition: An individual (or more) succeeds in reaching a pre-determined destination before all other individuals, without vehicle assistance or assistance of a horse or other animal. Racing generally involves accomplishing a task in less time than other competitors. The only type of racing considered relevant for the purposes of this event is the type where the task is traveling to a destination, completed by a person(s) without assistance of a vehicle or animal. Different types of races involve different types of human...

Fig. 1. Video exemplars (top) and the textual definition (bottom) of the event winning a race without a vehicle to illustrate one of the events studied in this paper. Following the NIST TRECVID evaluation guidelines [57], the textual definition is for zero-example recognition, and the ten provided video exemplars are for few-example event recognition.

A preliminary version of this article appeared as [21]. The current version adds i) zero-example event recognition ii) representation learning for multiple video modalities, iii) event specific representation learning, iv) an additional TRECVID video data set, revised experiments, and improved baselines, all using a recent video CNN feature [67], [81], and v) a new related work section, which will be discussed next.

2 RELATED WORK

We focus on three directions of related work we deem most closely connected to ours.

2.1 Representations for Event Recognition

Until recently most event recognition methods exploit video representations based on densely extracted low-level visual features, such as HOG/HOF [42], or MBH [56], [75], often combined with audio features like MFCC features [37], [53], [68]. Currently, most methods extract frame-based deep convolutional neural network (CNN) features [40], [63], [67], using the responses from intermediate layers of the CNN, which is pre-trained on ImageNet images [61], see e.g., [10], [34], [51], [81], [84]. To obtain per video descriptors the local/frame-based descriptors are aggregated by their mean, by using the Fisher vector [62] or VLAD encoding [29]. Despite the fact that these low-level based representations obtain state-of-the-art recognition performance, they suffer from two drawbacks. First, because of their high dimensionality, training effective classifiers on these representations requires a sufficient number of training examples to prevent overfitting. Second, all these representations are incapable of providing a semantic interpretation of the video, which is crucial for zero-example recognition.

Semantic Representations To obtain a semantic representation for videos, inspiration is obtained from describing images with attributes [19], [41], and the video is represented by its attribute predictions. Creating the training data for a set of task-specific attribute classifiers manually involves lots of annotation effort, which is restrictive. Therefore, often public available datasets, such as ImageNet [61] and TRECVID [57], are used to train the attribute classifiers [22], [45], [48]. Although this overcomes the need for (additional) manual annotation, the attributes of these datasets are not necessarily descriptive for event recognition.

To tune the attributes for the task at hand, several works aim to automatically discover the attributes from web images/videos and their textual descriptions [10], [12], [79], [83]. For example Wu et al. [79], start by selecting the most frequent/relevant terms from a set of provided event descriptions as attributes. Then, inspired by Berg et al. [4], they use Internet search engines to gather positive examples. They assure visual predictability of the discovered attribute detectors by cross-validation. Similarly the work of Ye et al. [10] relies on WikiHow to obtain event descriptions, and the visual predictability of the selected terms is ensured by keeping only those terms which are present in existing image classification datasets. We refer to these methods as term attributes since they all discover the attributes from the terms in the descriptions.

Despite their effectiveness in obtaining training data, term attributes suffer from three drawbacks. First, descriptions have a long-tail distribution, therefore many terms rarely occur and these will not have enough positive examples to train reliable visual classifiers. Second, term attributes are selected mostly based on their visual prediction accuracy, while for effective event recognition the selected term attributes should also be descriptive for the target events. Third, contextual information is lost, since term attributes are learned independently by binary classifiers.

In this paper, we also learn the representation from videos and their descriptions. But, rather than selecting individual, and often unreliable, term attributes, we learn the entire representation by automatically combining the terms through embedding. In our embedding, the correlations between the terms are utilized to learn a more effective representation, which is predictable and descriptive.

2.2 Embedding Videos and their Descriptions

To learn correspondences between the visual domain and textual descriptions different embedding methods have been proposed for various purposes i.e., image annotation [6], image classification [2], [78], image captioning [38], [73], video to text translation [15], [80] and cross-modal retrieval [11], [18], [54].

Canonical correlation analysis (CCA) [25] is the classical unsupervised manner to relate different modalities and can be used for cross-modal retrieval [11]. CCA is
the multimodal generalization of PCA, and computed as a generalized eigenvalue problem on the cross-covariance matrix between the visual and textual features. It finds a sequence of uncorrelated projections in which the cross-correlation between the modalities is maximized. This is not necessarily suited to obtain a discriminative representation, as is also validated by our experiments.

Probabilistic topic models, such as the seminal correspondence latent Dirichlet allocation [6] and its extensions [14], [60], learn correspondences by extracting a set of correlated visual and textual topics from imagery and their captions. Despite their effectiveness for captioning images and videos [15], these methods are not designed to learn a representation for the purpose of recognition. Moreover, we note that by design these models are only applicable on discrete features, and therefore they cannot leverage the state-of-the-art video features used in our paper.

Embeddings for multi-class image classification, such as [2], [78] find a low-dimensional subspace in which multi-class classification is performed. The subspace is found jointly with the multi-class classifier by minimizing a classification loss [78], or a zero-shot classification loss [2]. In contrast to this multi-class image classification setting, we are interested in finding descriptions modeled as a multi-label video classification problem. Moreover, we argue that solely optimizing for classification is not sufficient to obtain a predictable and descriptive video embedding.

Recently deep neural network architectures have been proposed to learn multimodal correspondences for retrieval [18], [54] and captioning [38], [73]. Notably the multimodal recurrent neural networks capture the sequential ordering between the words in image captions and generate more accurate image captions, compared to the probabilistic topic models. However, training deep neural networks generally requires large amounts of training examples, while our purpose is to learn from few- and zero-examples.

2.3 Zero-Example Event Recognition

In zero-example event recognition the goal is to recognize an event based on a given textual event definition, without using any video examples. The event definition is usually provided in the form of a title and description, see Figure 1. This zero-example setting is beyond the conventional zero-shot image recognition of objects and scenes [3], [17], [41], [59], where often a training set of related classes is available as well as pre-specified class-to-attribute mappings. This challenging event recognition problem has attracted attention because of its high practical value and the corresponding TRECVID benchmark task initiated by NIST [57]. The common approach is to represent videos and the event queries using a semantic representation, and to rank all video representations based on the cosine similarity to the event query.

To represent the video mostly attributes and term-attributes [10], [12], [32], [33], [79], [83] are exploited. Extensions include combining attributes using logical operators [20], using video-segments to detect attributes [49], and adjustment of attribute scores using an ontology structure [34]. In addition to attributes, automatic speech recognition and optical character recognition have been considered to enrich the semantic video representations [13], [34], [79].

To answer the event query, usually term matching between the semantic video representation and the event definition is performed. The retrieval can be enriched by using contextual information, such as word embeddings and term co-occurrences [8], [83]. For the final ranking, the cosine similarity can be extended by pseudo relevance feedback mechanisms, such as self-paced ranking [32]. This has been used to improve the zero-example event recognition by re-ranking [8], [33], [34]. Our representation learning is orthogonal to these efforts and can be joined with them to further improve the event recognition performance.

3 Video2vec Embeddings

Our goal is to learn a representation function \( f : \mathcal{X} \rightarrow \mathcal{S} \), which maps each low-level video feature \( x_i \in \mathcal{X} \) into the semantic representation \( s_i \in \mathcal{S} \). The representation function is trained on a collection of videos and their semantic descriptions. We represent the descriptions in the form of binary term vectors \( y_i \in \mathcal{Y} \). For event recognition, the \( s_i \) should have two properties: First, it should be descriptive enough to provide a characteristic representation of each video. Second, each dimension of the representation, corresponding to a semantic concept, should be predictable from the low-level video features.

In the attribute approach to learning the representation, the semantic representations are defined to be in the same space as the term vectors \( \mathcal{S} = \mathcal{Y} \). However, in practice, the term vectors are noisy and sparse, which undermines their effectiveness as labels to train \( f \). Therefore, we propose to learn the \( s_i \) on a lower dimensional projection of the term vectors, which are less sparse and less noisy. To remedy that the lower dimensional term vectors may be less descriptive for the video content, we formulate \( f \) as an embedding, coined as Video2vec, which is learned by balancing the descriptiveness and predictability as a joint optimization.

We first introduce the Video2vec embedding and its application for few- and zero-example event recognition (Section 3.1). Then we generalize our embedding to fuse multimedia features (Section 3.2). Finally, we extend the embedding to be learned specifically per event (Section 3.3). We summarize our notation conventions in Table 1.

### Table 1

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>Number of videos</td>
</tr>
<tr>
<td>( M )</td>
<td>Number of unique words in descriptions</td>
</tr>
<tr>
<td>( D )</td>
<td>Dimensionality of low-level feature</td>
</tr>
<tr>
<td>( J )</td>
<td>Number of low-level features to fuse</td>
</tr>
<tr>
<td>( X \in \mathbb{R}^{D \times N} )</td>
<td>Matrix of low-level video features</td>
</tr>
<tr>
<td>( Y \in {0,1}^{M \times N} )</td>
<td>Matrix of binary term vectors</td>
</tr>
<tr>
<td>( W \in \mathbb{R}^{D \times k} )</td>
<td>Video2vec visual projection</td>
</tr>
<tr>
<td>( A \in \mathbb{R}^{M \times k} )</td>
<td>Video2vec textual projection</td>
</tr>
<tr>
<td>( S \in \mathbb{R}^{k \times N} )</td>
<td>Video2vec embedding</td>
</tr>
<tr>
<td>( H \in \mathbb{R}^{M \times M} )</td>
<td>Diagonal matrix with per-term weights</td>
</tr>
<tr>
<td>( \pi_i \in [0,1]^N )</td>
<td>The column representing the i-th video</td>
</tr>
</tbody>
</table>

3.1 Objective Function

We start from a dataset of videos, represented by video features \( X \), and their textual descriptions, represented by
binary term vectors $Y$, indicating which words are present in each video description. Then, our Video2vec representation is learned by minimizing:

$$L_V(A, W) = \min_S L_d(A, S) + L_p(S, W),$$

where $A$ is the textual projection matrix, $W$ is the visual projection matrix, and $S$ is the Video2vec embedding. The loss function $L_d$ corresponds to our first objective for learning a descriptive Video2vec, and the loss function $L_p$ corresponds to our second objective for learning a predictable Video2vec. The Video2vec embedding $S$ connects the two loss functions. **Descriptiveness.** For the $L_d$ function, we use a variant of regularized Latent Semantic Indexing (LSI) [76]. This objective minimizes the quadratic error between the original video descriptions $Y$, and the reconstructed translations obtained from $A$ and $S$:

$$L_d(A, S) = \frac{1}{2} \sum_{i=1}^{N} \|y_i - As_i\|^2 + \lambda_a \Omega(A) + \lambda_s \Psi(S),$$

where $\Psi(\cdot)$ and $\Omega(\cdot)$ denote regularization functions, and $\lambda_a \geq 0$ and $\lambda_s \geq 0$ are regularizer coefficients. We use the squared Frobenius norm for regularization, which is the matrix variant of the $\ell_2$ regularizer, i.e., $\Omega(A) = \frac{1}{2} \|A\|_F^2 = \frac{1}{2} \sum_i \|a_i\|^2 = \frac{1}{2} \sum_{ij} a_{ij}^2$, the sum of the squared matrix elements. Similarly for the Video2vec matrix $\Psi(S) = \frac{1}{2} \|S\|_F^2$.

The main difference with regularized Latent Semantic Indexing [76] is their $\ell_1$ regularizer, $\Omega(A) = \sum_{i} \|a_i\|_1$, which enforces sparsity in the textual projection $A$. However, with our larger representation (typically we use a dimensionality of $k$ between 512 and 2,048 in our experiments compared to only $k = 20$ in [76]) and fewer number of unique words (around 10K, compared to 100K), enforcing sparsity is not necessary for good performance.

Note that other textual embeddings can be formulated similar to Eq. (2), when appropriate regularization functions $\Omega(\cdot)$ and $\Psi(\cdot)$ are used. For example using $\Omega(\cdot) = \|\cdot\|_1$ enforces sparsity [24], [47], [76], or in the extreme case that $A$ is constrained such that each column has a single nonzero value, the objective becomes very close to methods that select the best single term labels [4], and when $A$ is enforced to preserve the taxonomical term relations, the objective resembles taxonomy embedding [72], [77].

**Predictability.** The $L_p$ function measures the occurred loss between the Video2vec $S$ and the embedding of video features using $W$. Since the Video2vec $S$ is real valued, as opposed to a binary or multi-class encoding, we can not rely on standard classification losses such as the hinge-loss used in SVMs. Therefore, we define $L_p$ as a regularized regression, similar to ridge regression:

$$L_p(S, W) = \frac{1}{2} \sum_{i=1}^{N} \|s_i - W^T x_i\|^2 + \lambda_w \Theta(W),$$

where we use (again) the Frobenius norm for regularization of the visual projection matrix $W$, $\Theta(W) = \frac{1}{2} \|W\|_F^2$, and $\lambda_w$ is the regularization coefficient.

**Joint optimization.** To handle large scale datasets and state-of-the-art high-dimensional visual features, e.g., Fisher vectors [62] on video features [75] or deep learned representations [40], we employ a Stochastic Gradient Descent (SGD) [7] optimization, summarized in Algorithm 1. The number of passes over the datasets (epochs) and the stepsize $\eta$ are hyper-parameters of SGD. The Video2vec objective function, as given in Eq. (1), is convex with respect to matrix $A$ and $W$ when the embedding $S$ is fixed. In that case, the joint optimization is decoupled into Eq. (2) and Eq. (3), which are both reduced to a standard ridge regression for a fixed $S$. Moreover, when both $A$ and $W$ are fixed, the objective in Eq. (1) is convex w.r.t. $S$. Therefore we use standard SGD by computing the gradients of a sample w.r.t. the current value of the parameters, and we minimize $S$ jointly with $A$ and $W$.

Let denote a randomly sampled video and description pair at step $t$ by $(x_t, y_t)$, and let $s_t$ denote the current Video2vec embedding of sample $t$. The gradients of Eq. (1) for this sample w.r.t. $A, W, s_t$ are given by:

$$\nabla_{A} L_V = -(y_t - As_t)s_t^T + \lambda_a A,$$

$$\nabla_{W} L_V = -x_t(s_t - W^T x_t)^T + \lambda_w W, \text{ and}$$

$$\nabla_{s_t} L_V = -A^T (y_t - As_t) + (s_t - W^T x_t) + \lambda_s s_t.$$

The effect of joint learning descriptiveness and predictability, becomes clear in Eq. (6), where both the textual projection matrix $A$ and visual projection matrix $W$ contribute to learning the Video2vec embedding $S$. This embedding $S$ is subsequently used to obtain the textual projection $A$ matrix, in Eq. (4), and the visual projection $W$ matrix, in Eq. (5). This leads to the Video2vec embedding, which is both descriptive, by preserving the textual information, and predictable, by minimizing the visual prediction loss.

The parameters $A, S, W$ can be initialized randomly. However, to speed up convergence, we propose to initialize $A$ and $S$ by a rank-reduced singular value decomposition (SVD) of term vectors $Y$. This corresponds to the solution of the unregularized LSI objective of Eq. (2). For this purpose, $A$ and $S$ are initialized by the $k$ first eigenvectors of $YY^T$ and $Y^T Y$, respectively.

After training the visual and textual projection matrices, they are used to predict the Video2vec representation and the term vector of each video. In the case that both a video $x_t$ and description $y_t$ are given, we could obtain the Video2vec representation by returning $s_t$ from Eq. (1), while keeping both $A$ and $W$ fixed. However, in practice most videos are
not provided with a description. Therefore, we use
\[s_i = W^T x_i,\] (7)
to predict our Video2vec representation from the low-level video features \(x_i\). Moreover, using the predicted representation \(s_i\), the term vectors for each unseen video are predicted as follows:
\[\hat{y}_i = As_i = AW^T x_i,\] (8)
where the words with the highest values are most relevant for this video, see the illustration in Figure 2.

**Zero-Example Event Recognition.** To enable zero-example recognition using the Video2vec embedding, we employ the following steps: First, each test video is represented by predicting its term vector \(\hat{y}_i\) using Eq. (8), based on the pre-trained embeddings. Second, we translate the textual event definition into the event query, denoted as \(y^c \in \mathbb{R}^M\), by matching the words in the event definition with the \(M\) unique words in the Video2vec dictionary. Finally, the zero-example ranking is obtained by measuring the similarity between the video representations and the event query based on the cosine similarity:
\[s_c(x_i) = \frac{y^c \cdot \hat{y}_i}{\|y^c\| \|\hat{y}_i\|}.\] (9)

### 3.2 Video2vec Fusion

Videos are inherently multimodal. In general any video contains appearance, motion, and audio cues and sometimes even textual information in the form of subtitles or speech recognition scripts. Fusing the different modalities is typically achieved by early-fusion, i.e., fusion at the level of the representations, and late-fusion, i.e., fusion at the level of prediction scores [64]. Both fusion strategies have been shown to be effective for understanding complex events as well, e.g., [37], [53], [70], [79]. We propose Video2vec\(F\), which extends the Video2vec embedding by learning the semantic representation from multiple modalities.

A straightforward approach to learn the multimodal semantic representation is by fusing multiple Video2vec embeddings, which are independently trained per modality. Our intuition is that the semantic representation is more effective if it is predictable from all the modalities rather than from each individual modality. For Video2vec\(F\) we adjust the predictability loss, of Eq. (3), to incorporate a weighted combination of the predictability from all \(J\) modalities as follows:
\[L^F_p(S, W) = \sum_{j=1}^{J} L_p(S, W^j)\] (10)
where \(S\) is the multimodal Video2vec embedding, and \(W = \{W^j, j = 1, \ldots, J\}\) is a set of projection matrices from the \(J\) modalities. Each feature projection matrix \(W^j \in \mathbb{R}^{D_j \times K}\) projects the low-level feature \(x_i^j \in \mathbb{R}^{1 \times D_j}\) extracted from the video into the Video2vec representation \(s_i\). To balance the impact of feature dimensionality in the multimodal predictability loss, all low-level features from \(x_i^j\) are \(\ell_2\) normalized. Moreover, instead of weighting each modality equally, as in Eq. (10), a term \(\gamma_j \geq 0\) could be added to weight the importance of each modality, if sufficient training examples are available for their cross-validation.

The objective function Eq. (1) is still convex with respect to the parameters \(S, A, \) and \(W^j\) when the other parameters are fixed. However, the gradient with respect to \(s_i\), Eq. (6) becomes:
\[\nabla_s L_V = -A^\top (y_i - As_i) + s_i - \sum_j W^j\top x_i^j + \lambda_s s_i.\] (11)
It can be seen that all the modalities are jointly contributing to learn the multimodal Video2vec representations \(S\).

After training the textual and feature projection matrices, they are used to extract the multimodal Video2vec representation. Each feature projection matrix \(W^j\) predicts the Video2vec representation based on its underlying modality as follows:
\[s_i^j = W^j\top x_i^j.\] (12)
The final Video2vec representation is obtained by aggregating the per-modality representations. We experimentally observe that concatenation performs slightly better than sum or max pooling. The final representation we use is a concatenation of the per-modality embeddings: \(s_i = [s_i^1, \ldots, s_i^J]\).

### 3.3 Event Specific Video2vec

Not all words are equally important for recognizing an event. For each event there are a few terms which are highly informative, while the others are not. For the event “attempting a bike trick” words such as bicycle, jumping, and ramp are highly distinctive while beach, cake, and children are less informative.

In Section 3.1, the descriptiveness loss \(L_d\) is defined as the overall error in reconstructing all the words from the Video2vec representations, see Eq. (2). With this definition, the loss is biased towards the more frequent words, as minimizing their reconstruction error leads to a higher decrease in the overall error. Consequently, the words which
are infrequent in the descriptions might be discarded, which degrades their prediction accuracy from video features. This undermines the effectiveness of our representation learning for event recognition.

We extend the Video2vec embedding to learn a video representation, able to predict the informative words of events. Our extension, which we coin as Video2vec₂, is learned per event. It minimizes the reconstruction error of the terms with respect to their importance for event description, rather than their frequency in the Video2vec train data. We introduce a term sensitive descriptiveness loss:

\[
L_d^t(A, S) = \frac{1}{2} \sum_{i=1}^{N} \| H^e_i (y_i - A s_i) \|_2^2 + \lambda_v \Omega(A) + \lambda_s \Psi(S),
\]

where \( H_e \in \mathbb{R}^{M \times M} \) is a diagonal matrix, denoting the importance of each word for describing an event. By setting a relatively high value for \( h_{jj} \) for term \( j \), its reconstruction error is more penalized compared to the other terms. Hence the word is expected to be more precisely reconstructed.

We determine the term importance matrix \( H_e \) by relying on the presence/absence of terms in the textual event definition. Our assumption is that the words, which are present in event definitions are more important than the absent words. We set each element of the importance matrix \( h_{jj} \) to \( \alpha \), if the word \( j \) is present, and \( 1 - \alpha \) if the word \( j \) is absent in the event definition. \( \alpha \) is a balancing parameter between 0 and 1, which should be higher than 0.5 to assign more importance to the present words. We empirically set this parameter to 0.75 in all our zero-example event recognition experiments.

**Multimodal fusion.** Finally, we can leverage the multimodal features for learning event specific embeddings by combining the multimodal predictability loss and the term sensitive descriptiveness in a joint objective:

\[
L_V(A, W) = \min_S L_d^t(A, S) + L_p^F(S, W).
\]

We coin the learned video representation Video2vec₂, since it is learned event specific on multiple video modalities.

## 4 Experimental Setup

### 4.1 Datasets

We first introduce the dataset used for learning the Video2vec embeddings. Then, we detail the video datasets by which we evaluate the event recognition experiments.

**4.1.1 Video2vec Learning**

In all experiments, we learn the Video2vec embeddings on the VideoStory46K [21] dataset. This collection encompasses 45,826 videos harvested from YouTube, with a total length of 743 hours. Every video comes with a short title caption provided by the user who has uploaded the video, as shown in Figure 3. There are 19,159 unique terms in the captions, most of them occurring infrequently. We filter out the terms occurring only once as they generally are misspelled terms, numbers, or noisy terms. It provides us with 9,828 unique terms, which are used in our experiments.

4.2 Video Features

To cancel out accidental effects of the choice for the underlying features, we consider the same set of appearance,

1. There is also a PROGRESS set with 98K videos, but this partition is for blind testing by NIST only.
motion and audio features for all our experiments, the various baselines, and our Video2vec variants. The appearance features are used for all the experiments. The motion and audio features are added for the experiments involving the multimodal fusion.

**Appearance.** We adopt the video CNN representation [81] as appearance features for event recognition, but found the very deep network of [67] to perform slightly better than [63]. For each video the frames are extracted by uniformly sampling every two seconds. Then, the CNN descriptors are extracted per frame as the 1K dimensional responses from the last fully connected layer (pool3) of the Google Inception network [67]. We train the network on the 15,293 ImageNet categories with more than 200 examples, using the Caffe toolbox [31]. The final video CNN features are obtained by aggregating the frame descriptors over each video by VLAD encoding [29] with a codebook size of 20, resulting in a 20,480 dimensional vector.

**Motion.** We use the MBH descriptors along the motion trajectories [75] as motion features. The extracted 288-dimensional descriptors are reduced to 128 dimensions using PCA and are then aggregated per video using a Fisher vector [62], with 128 Gaussians resulting in a 32,768 dimensional vector. Each Fisher vector is power normalized, with \( \alpha = 0.2 \), as in [27].

**Audio.** We extract MFCCs descriptors [1] over a 10ms window. The descriptors consist of 13 values, 30 coefficients and the log-energy, along with their derivatives and the second derivatives. The MFCC descriptors are aggregated by Fisher vectors using a Gaussian Mixture Model with 256 components, resulting in a 46,080 dimensional vector.

### 4.3 Implementation Details

We learn the Video2vec embeddings by using 75% of the VideoStory46K dataset for training and 25% for validation to set the hyper-parameters of our model \((\lambda_w, \lambda_d, \lambda_a)\) and of SGD (number of epochs, \(\eta\)). As the validation criterion we rely on the objective function value, when using \(S = W^\top X\). The step size \(\eta_t\) is fixed during training.

For few-example event recognition, the event classifiers are trained as binary SVM with RBF kernels, as suggested in [50]. Following [81], [84] we set the SVM regularization and the RBF kernel parameters by a default value of 1, as the train set is not big enough for cross-validation.

As evaluation criteria we follow the standard convention in the literature [36], [57] by relying on the average precision (AP) per event, and we report the mean average precision (mAP) for overall accuracy.

### 5 Experiments

#### 5.1 Video2vec for Few-Example Event Recognition

##### 5.1.1 Effect of Embedding

We first compare Video2vec with term attribute baselines. The baselines learn the representations directly from terms without any embedding. We evaluate all video representations for few-example event recognition using a dimensionality varying from 32 to 8,192.

1. **Term attributes.** This representation is extracted by following the tradition of predicting relevant individual words from the video descriptions e.g., [10], [12], [79], [83]. A linear SVM classifier is trained per term. The classifiers which have the highest prediction accuracy, based on a 2-fold cross-validation [4], are selected as term attributes.

2. **Term attributes-f.** Rather than using cross-validation to select the term attributes it simply selects the words with the highest frequency in the descriptions.

**Results.** The results in Figure 4 demonstrate that Video2vec embeddings outperform the term attribute and term attribute-f baselines on all the three test sets.

Term attributes, which relies on the estimated reliability of individual word classifiers suffers from two drawbacks. First, many of the visual terms are very specific and therefore incapable of characterizing the events of interest i.e., necklace, suitcase, and earring. Although
these words can be accurately predicted from videos, they are incapable of providing a characteristic representation for event recognition. Second, many of the words rarely occur in video descriptions. Hence, only a limited number of positive examples are available to learn the word classifiers, which leads to a biased estimation of their reliability. Consequently, many of the discovered visual terms overfit to their small training set and do not generalize well for new videos.

The drawbacks of term attributes are relaxed by simply relying on the most frequent words. We observe the most frequent terms usually refer to characteristic attributes of events which are frequently used by humans when describing a video, i.e., car, girl, and kid. Because of their large number of positive examples, the trained visual classifiers from term attributes-f are in general more reliable.

The Video2vec embedding represents the words in a reduced-dimensional space, where correlated terms are usually combined together, as visualized in Table 2. Combining correlated words leads to less correlation between dimensions of the learned representation. Moreover, as the positive examples for all correlated terms are combined, it provides more positive video examples to train visual classifiers, often leading to better accuracy.

Besides its effectiveness for few-example event recognition, the Video2vec embedding also improves the representation learning efficiency by training visual embeddings for combination of words rather than individual words.

### 5.1.2 Video2vec vs Other Embeddings

We compare the effectiveness of our proposed Video2vec embedding with two alternative embeddings:

1. **CCA embedding.** This baseline learns the textual and visual projections by CCA [11], which maximizes the cross-correlation between the video features and descriptions. We experimentally observed that the embedding is even more effective when we PCA-reduce the video features to a dimensionality of 1,024 before learning the CCA embedding.

2. **Description embedding.** Similar to the Video2vec, this embedding is learned by minimizing the descriptiveness and predictability losses, but in two disjoint steps: The textual projection is first learned based on the regularized Latent Semantic Indexing [76], as in Eq. (2). Then the visual projection is learned separately, by minimizing the error for predicting the embedded descriptions from the video features based on ridge regression, as in Eq. (3).

### Results

The results in Figure 5 demonstrate that the Video2vec outperforms the CCA and the Description embedding baselines on all the three test sets for a dimensionality larger than 256.

We explain the gain over CCA by the fact that CCA is a symmetric embedding, it learns both the textual and visual projections in the same way. However, the textual and visual features have different distributions and properties, which may require different objective functions to learn their projections. This is achieved by Video2vec as it relies on two separate LSI and ridge regression loss functions to learn the textual and visual projections.

We explain the improvement over the Description embedding by the fact that combining the words based on textual correlation only does not necessarily imply that the corresponding video is visually correlated as well. As it happens, the terms puppy and kid have a high correlation in the descriptions but are visually dissimilar. Combining these two words together, as is done by Description embedding, undermines the accuracy of the classifiers predicting them from videos. In contrast, in a Video2vec the correlated words are combined only if their combination improves their classifier prediction. It prevents the combination of correlated terms which are visually dissimilar.

From now on we fix the Video2vec representation to 2,048 dimensions, which is optimal based on Figure 4.

### 5.1.3 Video2vec vs Other Representations

Next, we compare Video2vec with state-of-the-art video representations for few-example event recognition:

1. **Low-Level.** In this baseline, the event classifiers are trained directly on the low-level video representations, without extracting a semantic representation. We rely on the video CNN features of [81], based on our implementation detailed in Section 4.2.

### Table 3

<table>
<thead>
<tr>
<th></th>
<th>Low-Level</th>
<th>Attributes</th>
<th>Video2vec</th>
</tr>
</thead>
<tbody>
<tr>
<td>MED 2013</td>
<td>28.1</td>
<td>22.5</td>
<td>32.4</td>
</tr>
<tr>
<td>MED 2014</td>
<td>21.8</td>
<td>17.2</td>
<td>24.2</td>
</tr>
<tr>
<td>CCV_{1:10}</td>
<td>50.0</td>
<td>48.8</td>
<td>54.8</td>
</tr>
</tbody>
</table>

---

Fig. 5. Video2vec vs Other Embeddings. Video2vec outperforms the CCA and the Description embedding on all three test sets. The Description embedding is the closest competitor, but it suffers from embedding correlated terms which are visually dissimilar. CCA, uses the same objective function to learn the visual and textual embeddings, which is suboptimal due to intrinsic differences between the visual and textual features.
Table 3 shows that Video2vec outperforms the low-level representation. By increasing training examples the difference becomes more subtle.

2. Attributes. This representation is obtained by adopting the public ImageNet dataset as the source for training attribute classifiers as proposed in [22]. However, instead of training SVM classifiers on bag-of-words encoding of color SIFT descriptors [22], we upgrade the attributes by training a deep CNN with Google Inception architecture on the 15,293 ImageNet categories as detailed in Section 4.2. 

Results. Table 3 shows that Video2vec outperforms the state-of-the-art attributes and low-level video representations on all three test sets.

By comparing the Video2vec and the attribute representation we observe a higher event recognition accuracy of 32.4 vs 22.5 for the MED 2013, 17.2 vs 21.8 for the MED 2014, and 54.8 vs 48.8 for the CCV test set. We explain it by the reliance of the attribute baseline on ImageNet categories. Many of these pre-specified categories are not semantically relevant for the events of interest. For example, many of the categories are devoted to specific animal species. In contrast, the Video2vec embedding is automatically derived from the VideoStory46K dataset, which includes many descriptions relevant to events. Different from previous work, our low-level representation outperforms the attribute representation, indicating the strong low-level features used in this work.

The results further demonstrate that the Video2vec outperforms the low-level representation. As an explanation, we speculate that the low-level representation is prone to overfitting due to its high dimensionality. More specifically, in the low-level baseline, the event classifiers are trained on the 20,480 dimensional low-level features from only 10 positive exemplars. This may lead to overfitting as a result of the curse of dimensionality. In contrast, the Video2vec representation transfers the semantics from descriptions into the video representation to alleviate the overfitting as a sort of regularization.

From 1 to 100 examples. For further investigation, we gradually increase the number of positive examples from 1 to 100, and evaluate the accuracy of the event classifiers trained on both the low-level and Video2vec representation. The positive examples are selected randomly and the results are reported by averaging over 10 repetitions to compensate for the random effect. As Figure 6 shows, when the number of event train examples is less than 10, the Video2vec representation outperforms the low-level representation. By increasing the training examples to 100, the difference becomes more subtle, which confirms our hypothesis.

5.1.4 Video2vec for Zero-Example Event Recognition

We also evaluate the effectiveness of Video2vec for zero-example event recognition. We compare Video2vec with the three semantic representation baselines used before: Attributes, Term attributes, and Term attributes-f.

Results. The results in Table 4 demonstrate that Video2vec outperforms the baselines on both test sets. We again explain the modest performance of the attribute baseline by the fact that this representation relies on ImageNet categories as attributes. Many of which are not semantically relevant to the events of interest. This drawback is relaxed by the term attributes and term-attributes-f baselines, as they are trained on more relevant video descriptions from the VideoStory46K dataset. However, both term attribute baselines rely on individual terms for learning the representation. In contrast, Video2vec learns the representation from term combinations, which leads to a better zero-example event recognition accuracy.

5.2 Video2vec Fusion

We evaluate the effectiveness of the Video2vec fusion (Video2vec^F) proposed in Section 3.2 for few-example event recognition. We perform the experiments using the appearance, motion, and audio modalities, as detailed in Section 4.2.

5.2.1 Effect of fusion

We study the impact of fusing multiple modalities for learning the video representation. We start from using only the appearance features (Appearance) and gradually add the motion (+ Motion) and audio features (+ Audio) to learn the video representation by Video2vec^F.

Results. Table 5 confirms that incorporating more modalities for estimating the predictability loss leads to a more effective representation learning.

We observe that some events are benefiting more from adding the auditory modality, i.e., birthday party and music performance. For these events, there are some distinctive terms which are more effectively predicted from the audio features, i.e., singing, clapping, and piano. Other events improve by including motion features, i.e., parkour and parade, for which some distinctive terms such as jumping, rolling, and running are well predictable from the motion features. It demonstrates that different modalities are complementary for predicting the terms, so they are all required to effectively predict the video descriptions.
Effective than the low-level features in general, as shown on the semantic Video2vec representations, which are more effective than the feature projections directly from low-level video features. This simple baseline is shown to be competitive to more complex multiple kernel learning [70].

2. Late fusion. This baseline fuses the modalities at the level of event classification scores. A separate event classifier is learned by training an SVM on low-level video features per modality. For each test video, the final event detection score is obtained by averaging the detection scores predicted by each classifier, as evaluated in [53].

3. Video2vec early fusion. This baseline learns one Video2vec embedding on the concatenation of various low-level video features, using the standard Video2vec objective of Eq. (1). For each video, a semantic representation is extracted by applying the learned feature projection on the concatenation of the low-level video features. Then, the event classifiers are trained and applied on the semantic video representations.

4. Video2vec late fusion. This baseline learns three separate Video2vec embeddings, one per modality, using the standard Video2vec objective of Eq. (1). For each video, a semantic representation is obtained by concatenating the three Video2vec representations, which are predicted from each modality. Then, the event classifiers are trained and applied on the semantic video representations.

Results. The results are reported in Table 6. The Video2vecF\textsuperscript{2} outperforms the alternative fusions on all three test sets. We explain the better performance of Video2vecF\textsuperscript{2} over early fusion and late fusion by the fact that both baselines train the event classifiers directly from low-level video features. However, in Video2vecF\textsuperscript{2} the event classifiers are trained on the semantic Video2vec representations, which are more effective than the low-level features in general, as shown in Section 5.1.3 for the appearance features.

In the Video2vec late fusion the embeddings are learned separately per modality. In contrast, Video2vecF\textsuperscript{2} relies on all the video modalities jointly to estimate the predictability loss, which in general is more reliable than the per modality predictability estimations. For further investigation, we visualize the learned textual projection matrices A in Figure 7. The Video2vecF\textsuperscript{2} prevents some undesirable combination of terms that happen when the predictability losses are estimated separately per modality. For example, in the Video2vec which is learned only on audio features, the terms laugh, cheer, bark, dog, and woof are all combined as they have similar auditory features (see the right plot). However, the Video2vecF\textsuperscript{2} does not combine laugh and cheer with the bark, dog, and woof terms, as these terms are different in the appearance and motion features. Video2vecF\textsuperscript{2} learns a more reliable combination of terms, leading to a more effective video representation.

Video2vec early fusion learns a single feature projection matrix on the concatenated video features. However, the low-level features from different modalities have a different intrinsic dimensionality, distribution, and meaning, which aggravates the learning from their concatenation. In contrast, the Video2vecF\textsuperscript{2} learns separate feature projection matrices per modality, where each feature projection is optimized based on the features from one modality. It alleviates learning the feature projections, which leads to a more effective video representation.

To conclude, the Video2vecF\textsuperscript{2} effectively fuses the features from various video modalities by embedding them.
into a mutual semantic representation learned jointly over all the modalities.

### 5.3 Event Specific Video2vec

We evaluate the impact of learning event specific Video2vec embeddings, as proposed in Section 3.3, by comparing the Video2vec\(_E\) and its multimodal fusion Video2vec\(_{E,F}\) with Video2vec and Video2vec\(_{E,F}\). Different from Video2vec\(_E\) and Video2vec\(_{E,F}\), which are learned based on the term sensitive descriptiveness loss, in Eq. (13), the others are learned based on LSI, in Eq. (2).

#### Results

The results are reported in Table 7 and Table 8. Both event specific embeddings outperform their generic counterparts on both the MED 2013 and MED 2014 test sets. For zero-example event recognition reported in Table 7, the Video2vec\(_E\) improves the Video2vec from 15.9 to 18.3 on MED 2013, and from 5.2 to 6.8 on MED 2014. Similarly, the Video2vec\(_{E,F}\) improves the Video2vec\(_{E,F}\) from 17.8 to 20.0 on MED 2013 and from 6.6 to 8.0 on MED 2014. A consistent improvement is observed for few-example event recognition also albeit with a smaller margin, as detailed in Table 8. The results confirm the effectiveness of learning event specific Video2vec embeddings based on the term sensitive descriptiveness loss.

We explain the lower performance of the Video2vec and Video2vec\(_{E,F}\) baselines by the fact that the LSI loss treats all the terms equally when measuring the reconstruction error. Hence this loss is biased toward minimizing the reconstruction error for the frequent terms. As a consequence, the Video2vec and Video2vec\(_E\) are less accurate for predicting infrequent terms. This drawback is addressed by the term sensitive descriptiveness loss, which minimizes the reconstruction error for the indicative terms, even if those terms are infrequent in the train data. As a result, the Video2vec\(_E\) and Video2vec\(_{E,F}\) can predict the distinctive terms of an event more accurately, which leads to an improved event recognition accuracy. In Figure 8, we compare the term vectors, which are predicted by Video2vec and Video2vec\(_E\) for two video examples.

We explain the bigger gain for zero-example event recognition by two reasons: First, in few-example event recognition, an SVM event classifier is trained on the provided video exemplars, which makes the recognition somewhat robust to the noise in the video representations. For zero-example recognition the event is recognized by directly matching the event definition with the video representation. Hence, accurate prediction of the terms from videos become crucial. As the second explanation, we highlight that in our
experiments the term importance matrix $H_e$ is determined based on the event definitions. This is reasonable for zero-example event recognition, where the events are recognized using their definition only. However, for few-example event recognition it might be more effective to learn the term importance matrix $H_e$ from the video examples rather than pre-define them from the event definitions.

Finally, the most effective representation is learned by the Video2vec$_E$. It indicates that the term sensitive descriptiveness loss ($L^D_e$) and the multimodal predictability loss ($L^F_E$) are both effective and complementary when learning semantic video representations for event recognition.

### 5.4 Comparison with the state-of-the-art

We evaluate the merit of Video2vec by comparing it with several other recent works on few- and zero-example event recognition. Since in most papers the results are reported only on the MED 2013 test set, we limit our comparisons to this test set. Each approach may vary the data sources used for representation pre-training, yet they all abide to the standard data partitioning for training and testing of event classifiers as prescribed by the NIST evaluation protocol.

The results are reported in Table 9. It shows that our proposed representation learning sets a new state-of-the-art for the both few- and zero-example event recognition. The Video2vec$_E$ more effectively predicts the indicative terms of the event.

### 6 Conclusion

In this paper, we attack the problem of event recognition in video when examples are scarce. We propose the Video2vec embedding that learns a semantic video representation from a set of videos and their textual descriptions by minimizing a joint objective function balancing term descriptiveness and video predictability losses. As a result, the words which are correlated in the descriptions are combined together to improve their video predictability.

In addition, we propose the Video2vec$_E$ embedding with a multimodal predictability loss learned jointly over
video appearance, motion, and audio features. The different modalities are complementary for predicting the words, so they are all required for a richer video description. Moreover, embedding the heterogeneous video features into a mutual semantic space leads to few-example event recognition that is more effective than traditional fusion tactics.

We also propose the Video2vec embedding to learn event specific video representations. This embedding relies on a term sensitive descriptiveness loss to learn a more accurate representation for the words, which are indicative of the event.

Finally, the Video2vec embedding demonstrates that the term sensitive descriptiveness loss and the multimodal predictability loss are both effective and complementary to learn semantic video representations for few- and zero-example event recognition with state-of-the-art accuracy.

We consider Video2vec’s ability to generate human interpretable representations for previously unseen videos most appealing, as it opens up new connections with natural language processing and computational linguistics for describing and querying videos.

REFERENCES


