Article

A Continuous Semantic Embedding Method for Video Compact Representation

Tingting Han ^{1,*} , Yuankai Qi ² and Suguo Zhu ¹¹ Computer Science Department, Hangzhou Dianzi University, Hangzhou 310018, China; zsg2016@hdu.edu.cn² Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China; yk.qi@hit.edu.cn

* Correspondence: ttinghan@hdu.edu.cn

Abstract: Video compact representation aims to obtain a representation that could reflect the kernel mode of video content and concisely describe the video. As most information in complex videos is either noisy or redundant, some researchers have instead focused on long-term video semantics. Recent video compact representation methods heavily rely on the segmentation accuracy of video semantics. In this paper, we propose a novel framework to address these challenges. Specifically, we designed a novel continuous video semantic embedding model to learn the actual distribution of video words. First, an embedding model based on the continuous bag of words method is proposed to learn the video embeddings, integrated with a well-designed discriminative negative sampling approach, which helps emphasize the convincing clips in the embedding while weakening the influence of the confusing ones. Second, an aggregated distribution pooling method is proposed to capture the semantic distribution of kernel modes in videos. Finally, our well-trained model can generate compact video representations by direct inference, which provides our model with a better generalization ability compared with those of previous methods. We performed extensive experiments on event detection and the mining of representative event parts. Experiments on TRECVID MED11 and CCV datasets demonstrated the effectiveness of our method. Our method could capture the semantic distribution of kernel modes in videos and shows powerful potential to discover and better describe complex video patterns.

Keywords: discriminative sampling; video compact representation; video embedding

Citation: Han, T.; Qi, Y.; Zhu, S. A Continuous Semantic Embedding Method for Video Compact Representation. *Electronics* **2021**, *10*, 3106. <https://doi.org/10.3390/electronics10243106>

Academic Editor: Yung-Lyul Lee

Received: 13 October 2021

Accepted: 10 December 2021

Published: 14 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Video representation is a classical topic in computer vision. Generally, to obtain video representations, the primary task is to extract the useful features from the videos. From the earlier hand-crafted spatial-temporal features to recently learned deep video features, tremendous efforts have been made to obtain better understandings of videos. Hand-crafted features such as the STIP [1] and improved dense trajectories [2] model video motion from both the spatial and temporal dimensions along the trajectories. However, in addition to the heavy computational cost over all the video frames, the time and storage costs of these frame-level features are huge. On the other hand, the learned deep video features such as the two-stream ConvNet [3], CNN video representation [4], C3D ConvNet [5], VGAN [6], mGRU [7], ActionVLAD [8], the I3D network [9], and the Slowfast network [10] have recently achieved significantly better performance in action detection or event detection tasks. Recurrent neural networks (RNNs), particularly long short-term memory (LSTM) [7,11–13] and modified hierarchical recurrent neural encoder (HRNE) [14] are also used to model temporal information and represent videos. These frame-level or segment-level features are learned from well-trained deep models and could embed the visual semantic information of video content and temporal information together.

After all the frame-level features or segment-level features of a video are obtained, there are various approaches to obtaining the final video representation. Given all the

features of each video, some researchers generate the video representation by simple strategies such as average pooling or max pooling. Others try to encode the features in specific rules. Xu et al. [4] utilized VLAD encoding [15] to generate video representation from CNN descriptors—their method is promising. Vondrick et al. [6] proposed an adversarial network for video with a spatio-temporal convolutional architecture. Their generative model learns a useful internal representation for video and helps with video understanding and simulation. Girdhar et al. [8] proposed an end-to-end model for spatio-temporal video feature aggregation. The trainable ActionVLAD representation outperforms most prior approaches based on the two-stream.

However, when it comes to complex videos, which consist of complex videos content with tremendous intra-class variations, traditional video representation methods for these videos seem to be ineffective with limited performances. It becomes urgent and challenging to propose a rational compact representation for complex videos. Some researchers [16–19] try to take the video as a combination of manually defined or pre-trained concepts and obtain mid-level video representations based on event-specific concepts. Though some progress has been achieved, these methods heavily rely on the enormous human-defined concepts and their corresponding classifiers. Jiang et al. [20] tried to find the minimum number of visual frames needed to obtain a comparable event recognition performance. Mazloom et al. [21] and Phan et al. [22] learned the discriminative concept detectors and key evidence for complex videos. In actor identification works, representative actions [23] are mined from the entire movie to help with the identification. Some researchers exploited the latent concepts [24] of complex videos with hierarchical models to capture the semantic cues. Xie et al. [25] observed the fact that only some specific video parts or information are shared by the same type of video, and they propose a multi-mode representation method to reflect the kernel modes of video content and concisely describe the videos. As videos are divided into a fixed number of segments to mine latent patterns, the performance of their method heavily relies on the segmentation accuracy of video semantics. To address these challenges, in this work, we propose a novel method to learn the actual distribution of video words. We hope that our method could not only reflect the long-term semantics but also better accommodate the kernel modes of videos, as most information in complex videos is either noisy or redundant.

Intuitively, to generate a rational compact video representation, we try to answer two questions in this paper. First, how does one obtain the actual distribution of video words and utilize them for video analysis tasks such as event detection? Second, how does one capture the semantic distribution of kernel mode in videos?

To answer the first question, we claim that the continuous long-term semantics within a video are coherent. Inspired by continuous bag-of-words (CBOW) model in natural language processing, we designed a continuous semantic embedding model to learn the distribution of video words and hoped that the learned embedding could encode the context coherence of video semantics. We also deemed that semantics among different videos show different effects for representing the videos. Figure 1 shows illustrations of the confusing and convincing video semantics in videos. Given different videos, a variety of long-term video semantics can be found in the videos. “Diving”, “Soccer”, and “Parade” videos and their long-term semantics are shown in the figure. Watching the video semantics together, we obtain some interesting findings. It seems that the “crowd” semantics in the red boxes are confusing. They appear in all the videos and are little help in terms of distinguishing the videos. However, the video semantics in the blue boxes can convincingly help us distinguish the videos. For example, the “penalty kick” clip in the second row almost implies that this is a “soccer” video. Based on the above, in this paper, we pay more attention to discriminating the confusing and convincing video semantics and utilize them for generating our compact video representation. To achieve this goal, we propose a novel discriminative negative sampling approach when training the continuous video embedding model to ensure that the learned semantic embedding of video words encode both the context coherence and the discriminative degree of video semantics.

Figure 1. Given different videos, a variety of long-term video semantics can be found in the videos. This figure shows video semantics in the “Diving”, “Soccer”, and “Parade” videos. Watching these video semantics together, we obtain some interesting findings. It seems that the “crowd” semantics in the red boxes are confusing. They appear in all the videos and are little help in terms of distinguishing the videos. However, the video semantics in the blue boxes can convincingly help us distinguish the videos. For example, the “penalty kick” clip in the second row almost implies that this is a “soccer” video. Based on the above, in this paper, we pay more attention to discriminating the confusing and convincing video semantics and utilize them for generating our compact video representation. To achieve this goal, we propose a novel continuous video semantic embedding model integrated with a well-designed discriminative negative sampling approach.

To answer the second question, we propose an aggregated distribution pooling method to capture the semantic distribution of kernel mode in videos. Our basic assumption is that the kernel mode of video semantics will occur more frequently and densely in the video. Taking the “Diving” video in Figure 1 as an example, the kernel mode “air manoeuvre” and “take into water” may occur much more frequently and densely in the video than other semantics such as “scenery”. Based on this assumption, the proposed aggregated distribution pooling method aims to ensure that the kernel modes of each video are closely aggregated after all the semantic embedding of video clips are obtained.

The main contributions of our work are three-fold:

- We propose a continuous video semantic embedding model to learn the actual distribution of video words. Our embedding could encode the context coherence of video semantics.
- We propose a well-designed discriminative negative sampling approach. We pay more attention to discriminating the confusing and convincing video semantics and utilize them for generating the compact video representation. Incorporating with the continuous semantic embedding model, our semantic embedding could encode both the context coherence and the discriminative degree of video semantics. Experimental results demonstrate that discriminative negative sampling helps emphasize the convincing clips in the embedding while weakening the influence of the confusing ones—which also helps to learn more appropriate video semantic embedding.
- We propose an aggregated distribution pooling method to capture the semantic distribution of kernel modes in videos and generate the final video representation. We validated our method on event detection and the mining of representative event parts.

The remainder of this paper is organized as follows. Section 2 reviews related work. We describe our Video2Word framework in Section 3. Section 3.1 presents the continuous semantic embedding model integrated with a well-designed discriminative negative sampling approach. Section 3.2 describes the proposed aggregated distribution pooling method

to generate the final video method. Finally, the experimental results and conclusions are presented in Sections 4 and 5, respectively.

2. Related Work

2.1. Action Recognition

Action recognition is vital for the comprehension of video content and other higher-level video analysis tasks. In the beginning, researchers design spatial-temporal features, such as spatio-temporal interesting points (STIPs) [1] and improved dense trajectories (IDT) [2] to capture the motion information in the videos. Recently, with the development of deep learning methods, researchers have utilized the learned deep features to represent video information; among these deep features, the two-stream convolution network [3], the C3D network [5], the I3D network [9], and the Slowfast network [10] have achieved significantly better performances. However, in untrimmed videos, actions may take place in different time intervals and the other video parts might be noisy for recognition, which means a better action recognition method must consider the convincing video parts that help to distinguish the videos.

2.2. Temporal Embedding

Word embedding [26] methods are usually utilized in natural language processing (NLP) to model the relationship of a word and its context. Inspired by word embedding, Ramanathan et al. [27] treated videos as sequences of frames and learned temporal embedding for the video frames to model the relationship of the video frames and their neighborhoods. Their method outperforms the previous frame-level representation methods on video analysis tasks. In this paper, our inspiration for continuous semantic embedding also came from the continuous bag-of-words (CBOW) model in NLP. We trained our embedding model with a well-designed discriminative sampling approach, hoping that the learned semantic embeddings of video words encode both the context coherence and the discriminative degree of video semantics. It needs to be noticed that our work was the first job in segment-level video embedding. Our method is more comprehensive than the frame-level temporal embedding method [27].

2.3. Event Understanding

In human activity recognition tasks, videos are treated as a sequence of short-term action clips called action-words. Wu et al. [28] thought that an activity is about a set of action-topics and proposed a probabilistic model relating the action-words and the action topics. Their method is effective in action segmentation and recognition, which inspired us to train and learn the real distribution of video words by an end-to-end continuous semantic embedding model.

In activity prediction and actor identification tasks, Xu et al. [29] predicted human activities from only partially observed videos, which revealed that partial video information might be sufficient for effective classification. Xie et al. [23,25] mined the representative actions of actors which distinguish them from others in entire movies to support identification. Inspired by these works, we sought to capture the semantic distribution of kernel mode in videos for compact video representation.

2.4. Complex Event Detection

Complex event detection aims to detect the event in complex videos. Compared to simple videos in which events are usually well defined and describable by short frame sequences, complex events contain more people, objects, human actions, multiple scenes, and have significant intra-class variations. Usually, these events are presented in much longer and more video clips. Despite this difficulty, steady efforts have been made in terms of MED task [19,25,30,31].

Ramanathan et al. [27] focused on learning a frame representation that can capture the semantic and temporal context of a video. Li et al. [24] analyzed the latent concepts from

segment-level activities of events with an approach similar to ours. They assumed that events are composed of activity concepts, and these concepts are organized in a hierarchical structure which can be used to uncover video semantics by capturing segment-level activity concepts. In contrast to latent concepts models in which relations between instances are exploited, the model in [32] assumes that the instances in a video are mutually independent. Though the proposed model shows improvement, the concepts of complex events do not always evenly distribute among videos. This means that the number of latent concepts varies in different scenes, making the above assumption invalid. Xie et al. [25] encoded video clips by a deep-visual-word-based method and mined the visual topics of each video. They claimed that the obtained visual topics are not only representative but discriminative for event detection. However, as videos are divided into a fixed number of segments, the performances of their method always suffer from the poor segmentation accuracy of video semantics. Li et al. [33] presented a framework with hierarchical attention for video event recognition, and claimed that their model has advantages in long-term and short-term motion information acquisition. In this work, we treated a video as the combination of continuous small video clips, and proposed learning continuous semantic embedding for them. We conducted discriminative negative sampling incorporated with our model to ensure that the learned embedding (the distribution of video words) encoded both the context coherence and the discriminative degree of video semantics. The proposed method can achieve a much better video representation performance with no restrictions on video segmentation.

3. Our Approach

The motivation for our paper was to obtain the optimal compact video representation for long and untrimmed complex videos and we then utilized these video representations for better event detection and video understanding. Given an input video, our method will generate semantic embedding for all the video clips and utilize them to generate the final compact video representations. In our embedding model, we chose to give high embedding values video clips that contained convincing semantics, while for video clips that consisted of confusing semantics, we gave low values. Furthermore, our aggregated distribution pooling approach was designed to capture the semantic distribution of kernel mode in videos. In this section, we describe the details of the proposed continuous video embedding model to learn the actual distribution of video words.

3.1. Continuous Video Semantic Embedding Model

Figure 2 demonstrates the framework of our method. Given a set of raw videos, we first extracted the features of video clips by the pre-trained C3D architecture. The continuous semantic embedding model aims to learn the distribution of video words or semantics. In the figure, S_2 in the blue color is the anchor clip, S_1 and S_3 are the contexts or positive clips, and S_i in red is the negative clips. Integrated with a novel discriminative negative sampling approach, the obtained video embedding encodes both the long-term semantic and the discriminative information of video clips. We then utilized an aggregated distribution pooling approach that captures the distribution of the kernel mode of video semantics to obtain the compact video representations.

Our continuous video semantic embedding model is inspired by the classical continuous bag-of-words approach [34] in natural language processing for word embedding learning. Figure 3 demonstrates the framework of the CBOW. In the model, context is represented by multiple words for a given target word. For example, we could use “cat” and “tree” as context words for “climbed” as the target word. With the above configuration to specify context words, each word coded using 1-out-of- V representation means that the hidden layer output is the average of word vectors corresponding to context words at the input.

Input. Given a long, untrimmed video. We first divided it into small video clips. Similar to the CBOW method, we treated each video clip as the target input and tried to

model its correlation with the context clips, video clips near the target clips are treated as context inputs. We encoded the input pairs by the pre-trained C3D architecture to obtain the input features.

Figure 2. The proposed Video2Word framework for compact video representation. Given a set of raw videos, we first extracted features of video clips by pre-trained models. Inspired by continuous bag-of-words (CBOW) model, we then conducted a continuous video semantic embedding model to learn the distribution of video words. Integrated with a novel discriminative negative sampling approach, the obtained video embedding encodes both the long-term semantic and the discriminative information of video clips. Finally, we proposed an aggregated distribution pooling approach that captures the distribution of kernel mode of video semantics to obtain the compact video representations.

Figure 3. The framework of the classical CBOW model.

Video Features Extraction. To generate action proposals for the input videos, we first needed to obtain the visual content representation of the video clips. Action recognition models can be used for extracting frame or snippet-level visual features in long and untrimmed videos; among them, the C3D model has been utilized by most temporal action

proposal generation methods. In this work, for a better comparison with state-of-the-art methods, we followed related works such as SST [35], multi-mode [25], and BMN [36] using a pre-trained C3D model for video clips' representation. The C3D will encode every 16 frames into a 4096-dimension vector. We treated the 16 frames as a single video clip. Given a new video without any annotation, we extracted C3D features for every 16-frame clip and utilized a video word embedding method to encode it into a video word. Since the C3D was well trained, we just needed to use the C3D model to inference each video once. After feature extracting, videos are represented by non-overlapping clips. To this point, the basic input of our model was the features of video clip pairs.

Video Embedding. Given the features of the video clips, we propose to learn video embedding for the clip-level video semantics to obtain intuitive and effective video embeddings.

3.1.1. Continuous Video Semantic Embedding

Our video embedding first needs to encode the context coherence of video semantics. Based on the observation that video content remains consistent within a short temporal interval, we put the target clips and its context as input pairs of the model hoping that the obtained embeddings of the target clip and its context remain close. Figure 4 presents the framework of our continuous video semantic embedding methods. We utilized the continuous bag-of-video-words methods to encode the context information into the embedding. From this figure, the model takes the features of the context clips and the target clip as input, and we also provide hard negative samples for embedding by a well-designed discriminative negative sampling approach to ensure that the learned embedding not only encodes semantic coherence but also the discriminative degree of video semantics.

Figure 4. The frameworks of our continuous video semantic embedding methods. We utilized the continuous bag-of-video-words methods to encode context information into the embedding. Our methods try to model the correlation between the input pairs. It takes features of the context clips and the target clip as input, and we also provided hard negative samples by a well-designed discriminative negative sampling approach to ensure that the learned embedding not only encodes semantic coherence but also the discriminative degree of video semantics.

In this paper, we presented a fast but effective implementation scheme for the continuous bag-of-video-words method. In contrast to traditional CBOW which takes the context words as inputs and tries to learn the prior network parameters to fit the target word, we simply applied the dot production of the two embedded format input pairs to calculate their correlation scores. In detail, we enumerated all the video clips. For each clip, we treated it as the target clip t_i when modeling the relationship between context clips $C_i = \{c_1, c_2, \dots, c_n\}$ and the target clip, where n is the context window size, in this paper, we chose two clips before and after the target clips as the context clips. As the CBOW method claims that the contexts are closely related with each other, our network should give a high correlation score of the input pairs, that is:

$$F(C_i, t_i) = 1, \quad (1)$$

where F indicates the training network. While the input pairs that contain negative samples h_i , the output of our network should be 0, namely:

$$F(C_i, h_i) = 0. \quad (2)$$

As shown in Figure 4, we average pool the two branches' layer and use the sigmoid activation after the dot production to fit the correlation between the input pairs. The cross-entropy loss was chosen in the training step:

$$Loss = - \sum_{i=1}^N y_i \log p_i + (1 - y_i) \log(1 - p_i), \quad (3)$$

where N is the number of input pairs, y_i is the ground-truth correlation of input pair i , p_i is the predicted correlation. From Equations (1) and (2), given the input pairs of the target clip and its context, y_i should be 1. Otherwise, y_i is set to 0. With our simplified CBOW embedding strategy, our model can obtain video embeddings containing abundant semantic information within a short time.

3.1.2. Discriminative Negative Sampling

Correctly sampling context frames and negatives is quite important in order to learn high-quality embeddings. In this paper, we followed the strategy of [27], using hard negatives from the same video. The negatives were chosen from outside the range of the context window within the video.

Our video embedding was also designed to encode the discriminative degree of video semantics. As shown in Figure 1, video semantics show different effects for representing the videos. Some of them can convincingly distinguish the video, while others may be confusing. Our embedding model needs to have the ability to discriminate whether a video clip is confusing or convincing in terms of differentiating the video, and encode this difference into the final results. We propose a novel discriminative negative sampling approach to accomplish the above requirements. Specifically, we utilized a simple approach to evaluate the confidence of discriminative attribute, i.e., the convincing or confusing degree, and chose the hard negative sample for each target clip independently according to the video clips' confidence ranking. For each video clip, we determined its k (typically equal to 20) nearest neighbors in the training partition, namely all the videos. We scored each clip based on how many nearest neighbors were within the class as opposed to the number of those out of the class. For example, if there were 19 neighbors of clip x_i within the class, and only one neighbor was out of the class, the discriminative confidence score of x_i should be $19 = 19/1$, which means clip x_i helps distinguish the video a lot. For another clip x_j , if there are five neighbors of clip x_j are within the class, while 15 neighbors are out of the class, the discriminative confidence score of x_j should be $1/3 = 5/15$, which means clip x_j is confusing in terms of classifying the video. Our algorithm will choose a low confidence clip for it as its hard negative sample. In reverse, we will choose a high confidence clip for the target clip as its negative sample. To demonstrate the effectiveness of the proposed discriminative sampling approach, we conducted experiments on the performances in the event detection task between the continuous video semantic embedding method and the model integrated with discriminative negative sampling. The detailed results are shown in our later experimental part. The experimental results demonstrate that the discriminative negative sampling helps emphasize the convincing clips in the embedding while weakening the influence of the confusing ones, which also helps to learn a more appropriate video semantic embedding for event detection.

3.2. Aggregated Distribution Pooling

From Figure 2, after the continuous video semantic embedding model, we can obtain the video embeddings of all the video clips in a video. We then proposed an aggregated distribution pooling approach that captures the distribution of kernel mode of videos semantics to obtain the compact video representations by direct model inference with no restriction of video segmentation. Xie et al. [25] demonstrated that the meaningful high-level visual semantics or kernel modes of event frequently occur and are the abstract of the long-term video content. We followed them and considered that the kernel mode of each video more frequently occurs which could discriminate the videos. Taking the “Diving” video in Figure 1 as an example, the kernel mode “air maneuver” and “take into the water” occur much more frequently and densely in the video than other semantics such as “scenery”. The proposed aggregated distribution pooling was designed to ensure that the kernel modes of each video are closely aggregated. All the embeddings of a video share the same vocabulary size D (which was set to 1024 in our later experiments); in other words, we obtained a set of D -dimensional embedding vectors to generate the final video representation. For each video word or each dimension in the vocabulary, the histogram was introduced to calculate the embedding value distribution. We simply evenly divided all the values of each dimension into 10 bins from the minimum to the maximum value and then chose the highest bin for aggregation. For all the D video words in the vocabulary, we obtained their most frequent value interval, which in other words is the densest value distribution interval of the video words. Thus, the kernel mode of the video semantic distribution was obtained. When aggregating for each dimension of the embedding, we only chose numbers within the densest value interval for aggregation and then averaged them. As such, the final video representation was generated by our aggregated distribution pooling approach.

4. Experiments

We first present the results of our mining of representative video parts using our method for video understanding. Then, we test our compact video representation on event detection tasks and compare our experimental results with other state-of-art methods.

4.1. Datasets and Settings

We utilized the TRECVID MED11 dataset and the Columbia Consumer Video (CCV) dataset for evaluating our method on video understanding and event detection tasks.

4.1.1. TRECVID MED11 Dataset

MED11 (<https://www-nlpir.nist.gov/projects/tv2011/index.html#med>, accessed on 1 October 2021) was released by NIST for the purpose of TRECVID evaluation (<http://trecvid.nist.gov>, accessed on 1 October 2021) which is available to all participants and has become a standard benchmarking dataset for the research community. There are 2047 Internet videos for 15 complex events such as “Birthday party”, “Parade”, and “Landing a fish”. Each kind of video has approximately 100 exemplars for training and 40 exemplars for testing. The total duration of videos in each event collection is approximately 11 h. We followed [24,25], and used 70% of randomly selected videos from MED11 EventKit for training and 30% for testing.

4.1.2. CCV Dataset

CCV [37] contains 9317 videos from YouTube for 20 semantic categories, covering events such as “baseball”, “ice skating”, and “swimming”. The database typically has very little textual annotation and has become a standard benchmark for event detection and understanding. The total duration of videos is approximately 100 h.

To compare our model with other event detection methods [4,24,27,32,38], we followed the same protocol as in [24,25]. The event names and train/test splits can be found in the original paper [37].

4.1.3. Implementation Details

We trained the continuous video semantic embedding with context video size $n = 4$ and we chose 3 hard negative samples in the corresponding video for data augmentation. We optimized our model using *ADAM* step updates, with an initial learning rate of 5×10^{-4} and weight decay of 10^{-3} , separately. After video embedding, each video clip was embedded into a 1024-dimension vector. We then utilized the obtained video embeddings for aggregated distribution pooling to obtain the final compact video representation. Our models were implemented by TensorFlow, with training executed on GeForce TITAN X GPUs.

4.2. Representative Video Parts Mining

Representative video parts are the parts of a video that are most helpful in differentiating the video [39]. In our work, we aimed to propose a compact video representation that captures the semantic distribution of kernel mode in videos. We considered that the representative video parts should be not only representative but also discriminative. The discriminative attribute refers to the degree that the video part helps to distinguish the video from other types of video. The representative attribute refers to the degree that the video part can represent other items of the same category. In other words, the video clip more frequently occurs in the videos.

We visualized the results of our mining of the representative video parts with our method on the MED11 Dataset in Figure 5. We present representative video parts examples of the “Diving” videos. As shown in the figure, our approach could capture the video clips that contained semantics such as “air maneuver” and “take into water”, which are the kernel modes of diving videos. From the figure, we conclude that our method can capture the semantic distribution of kernel mode in videos.

Figure 5. Results of our mining of representative video parts with our method on the MED11 Dataset. The illustrations are of representative parts of diving videos. Our approach can capture the video clips that contain semantics such as “air manoeuvre” and “take into water”, which are the kernel modes of diving videos.

4.3. Event Detection

We also tested the ability of our method in terms of compact video representation. Video compact representation aims to discover the main concepts of video semantics that correspond to the human cognition of the video category. These key concepts are then used to compact represent the whole video regardless of other trivial video semantics.

Compared with the state-of-the-art methods, we followed [4,25] and conducted experiments on each video type with 100 positive exemplars. Mean average precision (mAP) was used as the performance metric following the NIST standard.

Tables 1 and 2 present the performance comparisons of the event detection performances between the continuous video semantic embedding method and the model integrated with discriminative negative sampling on the MED11 dataset and CCV dataset, separately. From the tables, we noticed that both of our methods have good performances on the datasets. Our embedding model, when integrated with the proposed discriminative negative sampling, has a significant performance improvement as compared to the initial

model. We concluded that the discriminative negative sampling helps to emphasize the convincing clips in the embedding while weakening the influence of confusing ones. Thus, the integrated model could learn the best video semantic embedding on both the MED11 and the CCV datasets. In the remainder of our paper, we utilize the integrated model as our continuous video embedding model.

Table 1. Event detection performances of our continuous video semantic embedding model compared with the model integrated with discriminative negative sampling (DNS) on the MED11 dataset.

Algorithms	mAP (%)
CNN [4]	74.7
Multi-Mode [25]	83.0
Video Embedding	74.4
Video Embedding + DNS	87.2

Table 2. Event detection performances of our continuous video semantic embedding model compared with the model integrated with discriminative negative sampling (DNS) on the CCV dataset.

Algorithms	mAP (%)
CNN [4]	61.1
Multi-Mode [25]	68.9
STDRN-HA [33]	74.2
Video Embedding	61.8
Video Embedding + DNS	72.4

We also present the performance comparison between our method and the other state-of-the-art event detection methods on the MED11 dataset in Table 3. Our method significantly improves the mean average precision (mAP). Compared with the previous state-of-the-art multi-mode event representation method, the proposed method achieves a performance improvement of 4.2% (from 83.0% to 87.2%).

Table 3. Performance comparisons of the proposed method with other event detection methods on the MED11 dataset.

Methods	mAP (%)
Joint+LL [40]	66.1
topic SR [41]	66.4
K-means-State [24]	68.7
HRNE [14]	68.8
fc7 [27]	69.1
InstanceInfer [32]	70.3
HMMFV [42]	70.8
Temporal Embedding [27]	71.1
Semantic-Visual [18]	73.3
CNN [4]	74.7
Latent Concepts [24]	76.8
Multi-Mode [25]	83.0
Ours	87.2

Table 4 shows the performance comparisons of our method with the other event detection methods on the CCV dataset. The proposed method also significantly improves the mean average precision (mAP). Compared with the previous state-of-the-art multi-mode event representation method, the proposed method with TDN proposals achieved

a performance improvement of 3.5% (from 68.9% to 72.4%). By capturing the semantic distribution of kernel mode of video discriminative semantics, our method also showed competitive results compared with the STDRN-HA method [33] but without intensive frame attention extractions. From Tables 3 and 4, it can be concluded that our method is effective for event detection on complex videos and shows great potential in terms of video compact representation. Again, we noticed that the proposed model did not achieve as much performance improvement on the CCV dataset as compared to the MED11. As we previously analyzed in the last section, videos in the CCV dataset are shorter and less complex than the MED11. Compared with all state-of-the-art methods on two benchmark datasets, the performances of our method were superior which demonstrates the effectiveness of the proposed method. Our method could capture the semantic distribution of kernel mode in videos and showed powerful capabilities to discover and better describe complex video patterns.

Table 4. Performance comparisons of the proposed method with other event detection methods on the CCV dataset.

Methods	mAP (%)
K-means-State [24]	56.7
HRNE [14]	57.0
InstanceInfer [32]	58.3
Temporal Embedding [27]	61.7
CNN [4]	61.1
Latent Concepts [24]	64.7
Multi-Mode Event Representation [25]	68.9
Ours	72.4

5. Conclusions

In this paper, to obtain the compact video representation that could reflect the kernel mode of video contents and concisely describe the video, we proposed a novel continuous video semantic embedding model to learn the actual distribution of video words. Integrated with a well-designed discriminative negative sampling approach, our model could encode both the context coherence and the discriminative degree of video semantics to video embeddings. We proposed an aggregated distribution pooling method to capture the semantic distribution of kernel mode in videos and generated the final video representation. Experimental results of the TRECVID MED11 dataset and the CCV dataset demonstrated the effectiveness of our embedding model integrated with discriminative negative sampling and the promising potential of our framework for video understanding and event detection. In our future work, we would like to explore how we can make better use of applying our compact video representation on other video analysis tasks such as unsupervised video segmentation and captioning.

Author Contributions: Conceptualization, T.H. and S.Z.; methodology, T.H.; software, T.H.; validation, T.H., S.Z. and Y.Q.; formal analysis, Y.Q.; investigation, S.Z.; resources, T.H.; data curation, T.H.; writing—original draft preparation, T.H.; writing—review and editing, Y.Q.; visualization, S.Z.; supervision, Y.Q.; project administration, S.Z.; funding acquisition, T.H. and Y.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partly supported by the National Natural Science Foundation of China under Project No. 62002091, 61902092 and the Zhejiang Province Natural Science Foundation of under Project No. LQ21F020014.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Laptev, I. On space-time interest points. *Int. J. Comput. Vis.* **2005**, *64*, 107–123. [\[CrossRef\]](#)
2. Wang, H.; Klaser, A.; Schmid, C.; Liu, C.L. Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *Int. J. Comput. Vis.* **2013**, *103*, 60–79. [\[CrossRef\]](#)
3. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 5–10 December 2014; pp. 568–576.
4. Xu, Z.; Yang, Y.; Hauptmann, A.G. A discriminative CNN video representation for event detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1798–1807.
5. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
6. Vondrick, C.; Pirsaviash, H.; Torralba, A. Generating videos with scene dynamics. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 613–621.
7. Zhu, L.; Xu, Z.; Yang, Y. Bidirectional Multirate Reconstruction for Temporal Modeling in Videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2653–2662.
8. Girdhar, R.; Ramanan, D.; Gupta, A.; Sivic, J.; Russell, B. ActionVLAD: Learning spatio-temporal aggregation for action classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
9. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
10. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6202–6211.
11. Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
12. Gao, L.; Guo, Z.; Zhang, H.; Xu, X.; Shen, H.T. Video captioning with attention-based lstm and semantic consistency. *IEEE Trans. Multimed.* **2017**, *19*, 2045–2055. [\[CrossRef\]](#)
13. Škrlj, B.; Kralj, J.; Lavrač, N.; Pollak, S. Towards robust text classification with semantics-aware recurrent neural architecture. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 575–589. [\[CrossRef\]](#)
14. Pan, P.; Xu, Z.; Yang, Y.; Wu, F.; Zhuang, Y. Hierarchical recurrent neural encoder for video representation with application to captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1029–1038.
15. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
16. Yang, X.; Zhang, T.; Xu, C.; Hossain, M.S. Automatic visual concept learning for social event understanding. *IEEE Trans. Multimed.* **2015**, *17*, 346–358. [\[CrossRef\]](#)
17. Ye, G.; Li, Y.; Xu, H.; Liu, D.; Chang, S.F. Eventnet: A large scale structured concept library for complex event detection in video. In Proceedings of the ACM Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 471–480.
18. Zhang, X.; Yang, Y.; Zhang, Y.; Luan, H.; Li, J.; Zhang, H.; Chua, T.S. Enhancing video event recognition using automatically constructed semantic-visual knowledge base. *IEEE Trans. Multimed.* **2015**, *17*, 1562–1575. [\[CrossRef\]](#)
19. Mazloom, M.; Li, X.; Snoek, C.G. Tagbook: A semantic video representation without supervision for event detection. *IEEE Trans. Multimed.* **2016**, *18*, 1378–1388. [\[CrossRef\]](#)
20. Jiang, Y.G.; Dai, Q.; Mei, T.; Rui, Y.; Chang, S.F. Super fast event recognition in internet videos. *IEEE Trans. Multimed.* **2015**, *17*, 1174–1186. [\[CrossRef\]](#)
21. Mazloom, M.; Gavves, E.; Snoek, C.G. Conceptlets: Selective semantics for classifying video events. *IEEE Trans. Multimed.* **2014**, *16*, 2214–2228. [\[CrossRef\]](#)
22. Phan, S.; Le, D.D.; Satoh, S. Multimedia event detection using event-driven multiple instance learning. In Proceedings of the ACM Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 1255–1258.
23. Xie, W.; Yao, H.; Sun, X.; Zhao, S.; Han, T.; Pang, C. Mining representative actions for actor identification. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 20–25 March 2016; pp. 1253–1257.
24. Li, C.; Huang, Z.; Yang, Y.; Cao, J.; Sun, X.; Shen, H.T. Hierarchical Latent Concept Discovery for Video Event Detection. *IEEE Trans. Image Process.* **2017**, *26*, 2149–2162. [\[CrossRef\]](#)
25. Xie, W.; Yao, H.; Sun, X.; Han, T.; Zhao, S.; Chua, T.S. Discovering Latent Discriminative Patterns for Multi-Mode Event Representation. *IEEE Trans. Multimed.* **2018**, *21*, 1425–1436. [\[CrossRef\]](#)
26. Morin, F.; Bengio, Y. Hierarchical probabilistic neural network language model. In Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, Bridgetown, Barbados, 6–8 January 2005; pp. 246–252.
27. Ramanathan, V.; Tang, K.; Mori, G.; Fei-Fei, L. Learning temporal embeddings for complex video analysis. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4471–4479.
28. Wu, C.; Zhang, J.; Savarese, S.; Saxena, A. Watch-n-patch: Unsupervised understanding of actions and relations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4362–4370.

29. Xu, Z.; Qing, L.; Miao, J. Activity Auto-Completion: Predicting Human Activities From Partial Videos. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3191–3199.
30. Li, C.; Cao, J.; Huang, Z.; Zhu, L.; Shen, H.T. Leveraging Weak Semantic Relevance for Complex Video Event Classification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3647–3656.
31. Zhao, S.; Yao, H.; Gao, Y.; Ding, G.; Chua, T.S. Predicting Personalized Image Emotion Perceptions in Social Networks. *IEEE Trans. Affect. Comput.* **2017**, 526–540. [[CrossRef](#)]
32. Lai, K.T.; Yu, F.X.; Chen, M.S.; Chang, S.F. Video event detection by inferring temporal instance labels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2243–2250.
33. Li, Y.; Liu, C.; Ji, Y.; Gong, S.; Xu, H. Spatio-temporal deep residual network with hierarchical attentions for video event recognition. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2020**, 16, 1–21. [[CrossRef](#)]
34. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
35. Buch, S.; Escorcia, V.; Shen, C.; Ghanem, B.; Carlos Niebles, J. SST: Single-stream temporal action proposals. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2911–2920.
36. Lin, T.; Liu, X.; Li, X.; Ding, E.; Wen, S. Bmn: Boundary-matching network for temporal action proposal generation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 3889–3898.
37. Jiang, Y.G.; Ye, G.; Chang, S.F.; Ellis, D.P.W.; Loui, A.C. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In Proceedings of the ACM International Conference on Multimedia Retrieval, Trento, Italy, 17–20 April 2011.
38. Huang, T.; Zhao, R.; Bi, L.; Zhang, D.; Lu, C. Neural Embedding Singular Value Decomposition for Collaborative Filtering. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–9.
39. Xie, W.; Yao, H.; Zhao, S.; Sun, X.; Han, T. Event patches: Mining effective parts for event detection and understanding. *Signal Process.* **2018**, 149, 82–87. [[CrossRef](#)]
40. Izadinia, H.; Shah, M. Recognizing complex events using large margin joint low-level event model. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 430–444.
41. Ramanathan, V.; Liang, P.; Fei-Fei, L. Video event understanding using natural language descriptions. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 2–8 December 2013; pp. 905–912.
42. Sun, C.; Nevatia, R. Active: Activity concept transitions in video event classification. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 2–8 December 2013; pp. 913–920.