Learning to Detect, Associate, and Recognize Human Actions and Surrounding Scenes in Untrimmed Videos*

Jungin Park
Yonsei University
Seoul, Korea
newrun@yonsei.ac.kr

Sangryul Jeon
Yonsei University
Seoul, Korea
cheonjsr@yonsei.ac.kr

Seungryong Kim
Yonsei University
Seoul, Korea
srkim89@yonsei.ac.kr

Jiyoung Lee
Yonsei University
Seoul, Korea
easy00@yonsei.ac.kr

Sunok Kim
Yonsei University
Seoul, Korea
kso428@yonsei.ac.kr

Kwanghoon Sohn†
Yonsei University
Seoul, Korea
khsohn@yonsei.ac.kr

ABSTRACT

While recognizing human actions and surrounding scenes addresses different aspect of video understanding, they have strong correlations that can be used to complement the singular information of each other. In this paper, we propose an approach for joint action and scene recognition that is formulated in end-to-end learning framework based on temporal attention techniques and the fusion of them. By applying temporal attention modules to the generic feature network, action and scene features are extracted efficiently, and then they are composed to a single feature vector through the proposed fusion module. Our experiments on the CoVieW18 dataset show that our model is able to detect temporal attention with only weak supervision, and remarkably improves multi-task action and scene classification accuracies.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**;

KEYWORDS

Video Classification; Action Classification; Scene Classification; Semantic Feature Fusion

ACM Reference Format:
Jungin Park, Sangryul Jeon, Seungryong Kim, Jiyoung Lee, Sunok Kim, and Kwanghoon Sohn. 2018. Learning to Detect, Associate, and Recognize Human Actions and Surrounding Scenes in Untrimmed Videos. In *The 1st Workshop and Challenge on Comprehensive Video Understanding in the Wild(CoVieW'18), October 22, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3265987.3265989

*Produces the permission block, and copyright information
†Corresponding author

CoVieW'18, October 22, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5976-4/18/10…$15.00
https://doi.org/10.1145/3265987.3265989

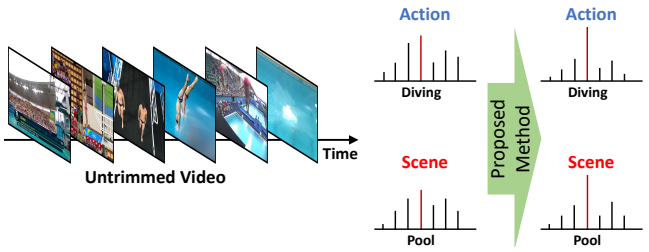

Figure 1: Basic concept of our approach. Video classification performance can be improved by considering semantic correlations between actions and scenes.

1 INTRODUCTION

Comprehensive video understanding has recently received increasing attention from the research communities and industries with the arising of large scale video data and the efficient machines learning techniques that can learn and understand like humans [2, 7, 18, 24].

In computer vision, video understanding is often addressed in the form of recognition or localization of human-centric events [3, 8, 16, 22, 23]. However, video understanding *in the wild* is an extremely challenging problem due to the irrelevance, which means that most videos in the real world contain large numbers of irrelevant frames pertaining to target tasks, which hinders to extract salient information.

Under such challenges, we argue that recognizing human activities and surrounding scenes of untrimmed videos can be used to complement the singular information of each other. The intuition behind this is straightforward: the presence (or absence) of particular scene properties can often be used to infer the possible subset of actions that can take place and vice versa. For example, if there is a 'pool' within the scene, then 'diving' becomes a possible action (See fig.1). On the contrary, if there is no 'pool', but a basketball court, then the probability of the 'diving' action decreases.

Motivated by this intuition, we introduce a novel cascade network that sequentially detects, associates, and recognizes both human activities and surrounding scenes in large-scale untrimmed videos. Our approach first starts to determine when the semantic information of interests occur in untrimmed videos by learning

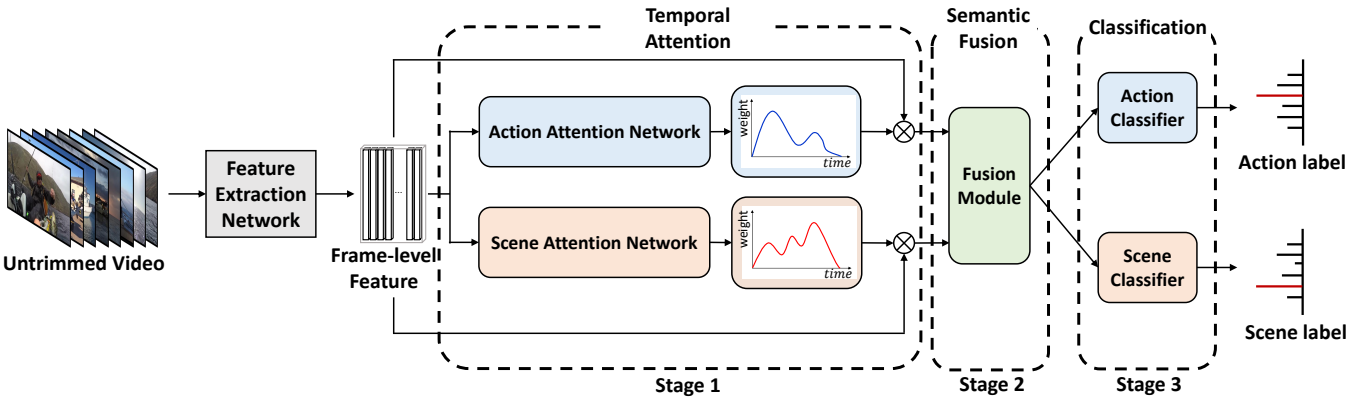

Figure 2: An overview of the proposed model for detecting temporal attention, fusing the action and scene feature, and performing the classification. First, a temporal attention networks detect an action and a scene temporal attention in the form of the probability distribution. Later, a fusion module combines semantic features about the action and scene with the temporal attentions. Finally, the fused feature is further analyzed by an action classifier and a scene classifier

temporal attention module. Without temporal annotations of instances, the attention module directly takes an untrimmed video as input and learns to predict the frame-wise probabilities of its video-level label pertaining to target tasks. Then the detected semantic entities (actions and scenes) are deeply fused to represent a long untrimmed video into a single feature vector. A number of ways of differentiable fusion are investigated, such as concatenation, sum, and multiplication, which facilitates joint training of all components in an end-to-end manner. Finally, the classification module predicts the probabilities of actions and scenes through each classifier. We test our approach over the large-scale challenging CoVieW18 dataset where both categories of action and scene are annotated in each video. The results demonstrate that all components of the proposed network are effectively engineered, composing semantic cues for action and scene recognition.

2 RELATED WORKS

In video categorization, some works in the literature use semantic representations. Liu et al. [11] focus on attribute-based event recognition. Jain et al. [14] consider the relationships between object and action, and Ikizler et al. [16] consider the combination of object and scene to improve action classification. While these works mainly focused on improving action classification performance, we aim to comprehensively understand untrimmed videos by leveraging the interactions between semantic entities. Gao and Ma [10] proposed a panoramic frame and representative feature patches as middle-level features for movie scene recognition approach. They used the informative correlations between video scenes to enhance the recognition performance of individual video scene. Ding et al. [20] proposed a Multi-view Multi-instance learning model (MMIL) which considers both context and independent instance in a bag simultaneously. They extracted the same context cues from different view and effectively integrated them into a unified learning framework based on joint sparse coding. However, these works experiments on the movie clips, thus, we do not know how they perform in real-world videos.

Several approaches have combined context cues to improve action recognition performance in controlled scenarios [4, 9, 12, 16, 24]. Marszaek et al. [15] show the relation of the co-occurrence between actions and scenes to develop useful visual representations for extracting short actions in movie clips. To extract the robust visual features for action understanding, [1, 4, 14, 16] show that implicit and explicit modeling of the relationships between objects in the video allows to discriminate action occurring in videos, especially by reducing the confusion between actions with similar motions such as drinking and smoking. More recently, Wu et al. [24] use robust discriminative networks to learn object, scene, and action relationships that tend to improve activity classification performance. In the same context as this work, Heilbron et al. [9] expand this idea further by exploiting semantic information acquired from action-object and action-scene relationships to address action detection.

3 PROPOSED METHOD

Given an untrimmed video V, our goal is to estimate the posterior probabilities of human activity and surrounding scene such that

$$p(z^a|V) \quad \text{and} \quad p(z^s|V), \tag{1}$$

where each of z^a and z^s is the annotated action and scene class label of video V.

Inspired by the attention-based video understanding model [22] and the semantic context fusion model [9, 16, 24], we propose a novel deep architecture that boosts action and scene classification performance. Our model employs the *soft attention mechanism* to obtain more discriminative features, and further effectively uses interactions which exist between action and scene features. Our method first estimates the temporal attentions using two sequential convolutional activations to obtain the independent generic features

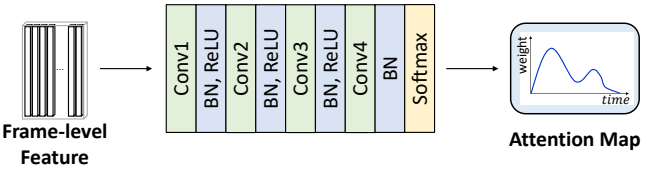

Figure 3: The details of the attention network. Frame-level feature fed into the sequential convolution, batch normalization, and ReLU layers. After passing through the last softmax layer, output the attention map in the form of probability distribution.

related to the action and the scene, and then performs fusion of two generated features for the action and the scene classification.

Fig.2 shows an overview of our approach which consists of three stages: (i) temporal attention module, (ii) semantic feature fusion module and (iii) classification module. At first, the temporal attention module detects the frames that are likely to contain the action or scene corresponding to the annotated labels by leveraging soft attention mechanism. For example, the probabilities are close to zero for the frames that does not have any related activity and scene to the annotated label, and vice versa. The attention networks are trained in weakly-supervised manner without the ground-truth temporal annotaions following [13]. Second, extracted features from the attention module are deeply combined through the semantic fusion module to complement the singular insformation of each other. We consider three different fusion strategies such as concatenation, sum, and multiplication, where a global average pooling layer is additionally applied to the end for the aggregation of temporal informations. Finally, we are able to produce the prediction score by passing fused video-level features through the classification module which consists of two fully-connected classifiers for each action and scene recognition task.

3.1 Stage 1: Temporal Attention Module

At first, a set of video frames are fed into deep networks for feature extraction. These feature representations are utilized for providing visual content to the temporal attention module. Formally, given video frames $\mathbf{F}$ and the feature extraction network parameters w_h, we extract the representations as

$$\mathbf{H} = \mathcal{F}(\mathbf{F}; W_h). \tag{2}$$

The temporal attention module learns to rank these frame-level features according to the importance weight through sequential 1-dim convolutional layers followed by batch normalization and ReLU. The estimated attentions are further normalized along temporal dimension with softmax operator, such that

$$\mathbf{W}'_i = \frac{exp(\mathcal{F}(\mathbf{H}_i; W_t))}{\sum_{l=1}^{L} exp(\mathcal{F}(\mathbf{H}_l; W_t))} \tag{3}$$

where W_t is the temporal attention network parameters and L is an arbitrary length of video. This is illustrated in Fig. 3. The estimated attentions are then multiplied with the frame-level features to highlight the discriminative frames and suppress the irrelevant ones as follows:

$$\mathbf{X} = \mathbf{W}'\mathbf{H}^T. \tag{4}$$

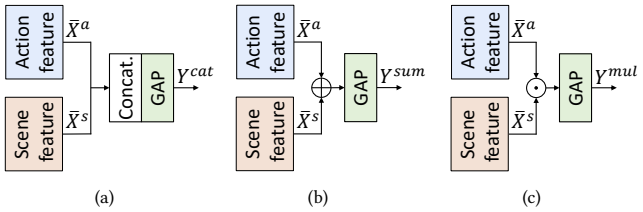

Figure 4: Three different methods for the semantic feature fusion: (a) concatenate the outputs of stage 1 in first layer of the fusion module, (b) element-wise summation between two outputs, and (c) element-wise multiplication between two outputs.

Note that we implemented temporal attention networks for each action and scene recognition task thus $\{\mathbf{W}^a, \mathbf{W}^s\}$ and $\{\mathbf{X}^a, \mathbf{X}^s\}$ are obtained where each a and s represents action and scene.

3.2 Stage 2: Semantic Fusion Module

Motivated from that semantic contexts of untrimmed video have strong correlations between them and can help interpreting each other, we consider three different methods to fuse the attended features extracted from the temporal attention module as illustrated in Fig. 4.

Our intention is to fuse the two networks such that channel responses at the same temporal position are put in correspondence. To motivate this, consider the previous example of the action, diving and the scene, pool. It is highly likely that the pool will come to the same scene at the moment of diving and thus, it is reasonable to consider temporal correspondence to maximize the correlation of the action and the scene features. This temporal correspondence is easily achieved as we employ the same architecture in two attention networks. A fusion function $f : \mathbf{X}^a, \mathbf{X}^s \rightarrow \mathbf{Y}$ fuses two feature maps $\mathbf{X}^a, \mathbf{X}^s \in \mathbb{R}^{L \times D}$ to produce an output map $\mathbf{Y}_i \in \mathbb{R}^{L \times D'}$, where $1 \leq i \leq D'$ and D' are the number of channels of the output feature maps. The fused feature is fed into a Global Average Pooling (GAP) layer, and so that the output of the semantic feature fusion module such that:

$$y_i = \frac{1}{L} \sum_{j=1}^{L} \mathrm{Y}_{i,j} \tag{5}$$

where $1 \leq i \leq D'$ and L is the length of the frames.

3.2.1 Concatenation fusion. $\mathbf{Y}^{cat} = f^{cat}(\mathbf{X}^a, \mathbf{X}^s)$ stacks the two feature maps across the feature channels D:

$$\mathbf{Y}^{cat} = \mathbf{X}^a || \mathbf{X}^s \tag{6}$$

where $\mathbf{Y}^{cat} \in \mathbb{R}^{W \times 2D}$ and $||$ represents the concatenation operator. Concatenation is the most common strategy keeping the correspondence of two features by simple stacking. The drawback of this feature is that it uses twice as many parameters in stage 3 compared to other fusion methods.

3.2.2 Sum fusion. $\mathbf{Y}^{sum} = f^{sum}(\mathbf{X}^a, \mathbf{X}^s)$ computes the sum of the two feature maps at the same temporal locations t and feature channel d:

$$\mathbf{Y}^{sum}_{t,d} = \mathbf{X}^a_{t,d} \oplus \mathbf{X}^s_{t,d} \tag{7}$$

where $1 \leq t \leq L$ and $1 \leq d \leq D$ and $\oplus$ represents element-wise summation operator. Since the outputs of stage 2 have the same temporal resolution, sum fusion can consider more detailed correspondence using element-wise process.

3.2.3 Multiplication fusion. $\mathbf{Y}^{mul} = f^{mul}(\mathbf{X}^a, \mathbf{X}^s)$ performs element-wise multiplications of $\mathbf{X}^a$ and $\mathbf{X}^s$:

$$\mathbf{Y}^{mul}_{t,d} = \mathbf{X}^a_{t,d} \odot \mathbf{X}^s_{t,d} \tag{8}$$

where $1 \leq t \leq L, 1 \leq d \leq D$ and $\odot$ represents the element-wise multiplication operator. The resulting feature $Y^{mul} \in \mathbb{R}^{W \times D}$ captures multiplicative interactions at corresponding features. In this fusion, the features with high weights in both attention maps are enhanced and with low weights in both attention maps are further attenuated to obtain a more discriminative feature.

3.3 Stage 3: Classification Module

In the classification module, we aim to classify each video into the action and scene categories by using two classifiers, based on the output features of the fusion module $\bar{\mathbf{y}}$. Suppose we have n_a action classes and n_s scene classes, we learn a linear mapping $W^a_c \in \mathbb{R}^{n_a \times D}$ and $W^s_c \in \mathbb{R}^{n_s \times D}$ to transform the feature representation $\bar{\mathbf{y}}$ into a n_a and n_s-dimensional score vector $\bar{\mathbf{z}}^a$ and $\bar{\mathbf{z}}^s$, i.e., $\bar{\mathbf{z}}^k = W^k_{cls}\bar{\mathbf{y}}$, where $k \in \{a, s\}$ and W^k_{cls} are the classification network parameters. This score vector can be also passed through a softmax layer as follows:

$$\hat{z}^k_i = softmax(\bar{z}^k_i), \tag{9}$$

where $\bar{z}^k_i$ denotes the i^{th} dimension of $\bar{\mathbf{z}}^k$. For clarity, we use the notation $\bar{z}^k$ to denote the original classification score of video and $\hat{z}^k$ to represent the softmax classification score.

3.4 Training

We first train our proposed model without semantic fusion module so that the temporal attention module is previously learned in a weakly-supervised manner. Only with the video-level labels, the temporal attention networks are trained to automatically reason the temporal weights of possible instances that may lead to superior classification performances. Specifically, our loss function is defined as a summation of cross-entropys from each action and scene label,

$$\mathcal{L} = \mathcal{L}_{action} + \mathcal{L}_{scene}. \tag{10}$$

where

$$\begin{aligned} \mathcal{L}_{action} &= -\sum_{i=1}^{M} y_i^{action} log(\hat{z}_i^a) \\ \mathcal{L}_{scene} &= -\sum_{i=1}^{N} y_i^{scene} log(\hat{z}_i^s), \end{aligned} \tag{11}$$

and y is the annotated labels for action and scene, M and N are the number of each action or scene classes. Note that the parameters of temporal attention network are initialized before training to estimate a uniform probability distribution over the video frames in a manner similar to [21].

We further fine-tune the whole networks initialized with learned parameters from temporal attention networks. The components of our model are all differentiable, facilitating an end-to-end training of our model with the same cross-entropy loss in (10).

3.5 Implementation details

To make our findings reproducible, we describe here the implementation details of our model. We use PyTorch [17] to implement our models. Since each video sequence from CoVieW'18 has arbitrary number of frame, we applied zero padding to the extracted features so that the length of all the features fed into the attention module is fixed to 300, i.e., $\mathbf{H} \in \mathbb{R}^{300 \times 1024}$. We trained our model with the adaptive momentum estimation algorithm [6], where the batch size is set to 70. The initial learning rate is set to 0.001 and decreases every 1,200 iterations by a factor of 10, and it stops training at 6,000 iterations. Four 1-dim convolutional layers are used in temporal attention network with kernel size of 3. Each of them has 512, 128, 32, and 1 channels. The output dimension of the fusion layer depends on the fusion method, 2048 for concatenation fusion and 1024 for sum and multiplication fusion. Each of action and scene classifiers consists of two fully-connected convolutional layers where the first one has 1,024 channels and the last one has 285 and 29 channels for each action and scene classification.

4 EXPERIMENTS

In this section, we describe the experimental results of our method. We introduce the evaluation datasets and show the effectiveness of the temporal attention and feature fusion compared to direct classification without attention or fusion.

4.1 Dataset

We have used CoVieW18 dataset in our experiment. The CoVieW18 dataset is released for multi-task action and scene recognition in untrimmed video that sampled from the Youtube-8M dataset [18] with annotated action and scene class labels for each video. Each video in dataset is composed of pre-extracted frame-level feature using Inception network trained on ImageNet [19] and about Youtube video URLs. Although our model can be trained *end-to-end* from the feature extraction network using back-propagation, we train from the attention module since the Coveiw 2018 dataset is pre-extracted features. Our model can be improved by end-to-end learning as shown in [5].

The dataset is consist of 90,853 videos, 285 actions and 29 scenes and each video has one action and one scene associated with it. We take 84, 853, 3, 000 and 3, 000 videos for training, validation, and testing, respectively. All the videos in the dataset are decoded at 1 frame-per-second up to the first 300 seconds (5 minutes) and the shortest video has 120 frames (2 minutes).

4.2 Results

In this section, we show quantitative evaluation of the temporal attention networks. And then, we compare the results according to the fusion method: (1) Concatenation; (2) Sum; (3) Multiplication. Finally, to validate the effectiveness of the temporal attention and the feature fusion, we compare with the following baselines: (1) Without attention and fusion; (2) Use attention without fusion; (3) Use fusion without attention.

Figure 5: An example of temporal attention for a video sequence of *"Surfing on the Beach"*. Our method is able to generates weights of how much each frame is associated with the classification tasks. The region denoted by a bold color is the frame which have a large value of the weight.

Table 1: Performance comparison of different fusion strategies on CoVieW dataset. Accuracies are measured using the mAP metric at top 1 and top 5.

Methods	Action@1	Scene@1	Action@5	Scene@5
Concatenation	83.5	94.63	97.17	99.1
Sum	84.23	95.07	97.2	99.2
Multiplication	**84.8**	**96.07**	**97.3**	**99.27**

4.2.1 Temporal attention module. We first evaluate the temporal attention module to investigate the significance of leveraging the temporal clues for video classification. In fig 4, we show the example of the extracted temporal attention weights and actual frames of the video sequence. The temporal attention is illustrated by heatmaps, in which frames related with the action "surfing" or the scene "beach" have high scores, and the frames which are less discriminative for the semantic representation have low scores.

4.2.2 Semantic fusion module. We compare different fusion strategies in Table 1. Results are evaluated using the mean Average Precision (mAP) metric at top 1 and top 5. We observe that Concatenation fusion perform considerably lower than Sum and Multiplication fusion. Multiplication fusion performs best and is slightly better than Sum fusion. What stand out is that Concatenation fusion shows lowest performance even though it uses nearly twice as many parameters in classifier. And also this is interesting, since this, as well as the high result of Sum-fusion and Multiplication-fusion, suggest that simply summing or multiplication the feature maps is already a good fusion technique and learning a randomly initialized combination does not lead to significantly different/better results.

4.2.3 Ablation studies. To evaluate the contribution of each module in our model, we break down our network with different combinations. The results are shown in Table 2. We adopt a 3-layer classifier (one pooling layer and two fully-connected layers) for the given frame-level dataset; and a variant network which considers either of *Attention* or *Fusion*. We apply element-wise multiplication method to fusion networks, which shows the best performance compared with other fusion methods. We can see that *Attention* has more contribution to the improved performance than *Fusion*. It indicates that though correlation between the action and the scene information is important, the attention feature has significant discriminative information which is very useful for classification. All results achieves significantly better results on the scene classification since the number of classes much more in the action. Considering the attention and combining two informations offers better performance.

Table 2: Comparison of action and scene classification accuracy(mAP) on the CoVieW dataset. Accuracies are measured with using only classifier, attention without fusion, fusion without attention and attention and fusion both.

Attention	Fusion	Action(mAP)	Scene(mAP)
-	-	78.8	93.4
✓	-	82.3	94.57
-	✓	81.78	94.23
✓	✓	**84.8**	**96.07**

5 CONCLUSION

We proposed a framework that recognize both human actions and scene using the temporal attention and feature fusion of untrimmed video based on deep neural networks. The experimental results on the CoVieW18 dataset demonstrated the effectiveness of the proposed temporal attention module and feature fusion module.

ACKNOWLEDGEMENT

This research was supported by the International Research & Development Program of the National Research Foundation of Korea (NRF) funded by Ministry of Science and ICT (NRF-2017K1A3A1A16066838).

REFERENCES

[1] A.Gupta and L.S.Davis. 2007. Object in Action: An Approach for Combining Action Understanding and Object Perception. *CVPR* (2007).

[2] A.Karpathy, G.Toderici, S.Shetty, T.Leung, R.Sukthankar, and L.Fei-Fei. 2014. Large-scale Video Calssification with Convolutional Neural Networks. *CVPR* (2014).

[3] A.Montes, A.Salvador, S.Pascual, and X.Giro i Nieto. 2017. Temporal Activity Detection in Untrimmed Videos with Recurrent Neural Networks. *1st NIPS Workshop on Large Scale Computer Vision Systems* (2017).

[4] A.Prest, V.Ferrari, and C.Schmid. 2013. Explicit modeling of human-object interactions in realistic videos. *IEEE transactions on pattern analysis and machine intelligence* (2013).

[5] B.Fernando and S.Gould. 2016. Learning End-to-end Video Classification with Rank-Pooling. *ICML* (2016).

[6] D.Kingma and J.Ba. 2014. Adam: A method for stochastic optimization. *arXiv* (2014).

[7] F.C.Heilborn, V.Escorcia, B.Ghanem, and J.C.Niebles. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. *CVPR* (2015).

[8] F.C.Heilbron, JC.Niebles, and B.Ghanem. 2016. Fast Temporal Activity Proposals for Efficient Detection of Human Actions in Untrimmed Videos. *CVPR* (2016).

[9] F.C.Heilbron, W.Barrios, V.Escorcia, and B.Ghanem. 2017. SCC: Semantic Context Cascade for Efficient Action Detection. *CVPR* (2017).

[10] GY.Gao and HD.Ma. 2014. Movie Scene Recognition using Panoramic Frame and Representative Feature Patches. *JCST* (2014).

[11] J.Liu, Q.Yu, O.Javed, S.Ali, A.Tamrakar, A.Divakaran, H.Cheng, and H.Sawkney. [n. d.]. Video event recognition using concept attributes. *IEEE Workshop on WACV* ([n. d.]).

[12] L.-J.Li and L.Fei-Fei. 2007. What, where and who? classifying events by scene and object recognition. *ICCV* (2007).

[13] L.Wang, Y.Xiong, D.Lin, and L.V.Gool. 2017. UntrimmedNets for Weakly Supervised Action Recognition and Detection. *CVPR* (2017).

[14] M.Jain, J.C.v.Gemert, T.Mensink, and C.G.Snoek. 2015. Object2action: Classifying and Localizing Actions without any Video Example. *ICCV* (2015).

[15] M.Marszalek, I.Laptev, and C.Schmid. 2009. Actions in context. *CVPR* (2009).

[16] N.Ikizler-Cinbis and S.Sclaroff. 2010. Object, Scene and Actions: Combining Multiple Features for Human Action Recognition. *ECCV* (2010).

[17] Paszke, Adam, Gross, Sam, Chintala, Soumith, Chanan, Gregory, Yang, Edward, DeVito, Zachary, Lin, Zeming, Desmaison, Alban, Antiga, Luca, Lerer, and Adam. 2017. Automatic differentiation in PyTorch. *NIPS 2017 Workshop* (2017).

[18] S.Abu-El-Haija, N.Kothari, J.Lee, P.Natsev, G.Toderici, B.Varadarajan, and S.Vijayanarasimhan. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. *arXiv* (2016).

[19] S.Ioffe and C.Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariance Shift. *ICML* (2015).

[20] X.Ding, B.Li, W.Hu, W.Xiong, and Z.Wang. 2012. Horror Video Scene Recognition Based on Multi-view Multi-instance Learning. *ACCV* (2012).

[21] X.Glorot and Y.Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. *AISTATS* (2010).

[22] Y.Peng, Y.Zhao, and J.Zhang. 2017. Two-stream Collaborative Learning with Spatial-temporal Attention for Video Classification. *IEEE Transactions on Circuits and Systems for Video Technology* (2017).

[23] Z.Wu, X.Wang, Y.Jiang, H.Ye, and X.Xue. 2015. Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification. *ACMMM* (2015).

[24] Z.Wu, Y.Fu, Y.Jiang, and L.Sigal. 2016. Harnessing Object and Scene Semantics for Large-Scale Video Understanding. *CVPR* (2016).