NTTFudan Team at TRECVID 2016: Multimedia Event Detection

Yonqqing Sun[1], Rui-Wei Zhao[2], Minjun Li[2], Chuan Lu[2], Hiroyuki Arai[1], Tetsuya Kinebuchi[1], and Yu-Gang Jiang[2]

[1]NTT Media Intelligence Labs, Japan
[2]School of Computer Science, Fudan University, China
{yongqing.sun,arai.hiroyuki,kinebuchi.t}@lab.ntt.co.jp
{rwzhao14,minjunli13,chuanlu13,ygj}@fudan.edu.cn

December 12, 2016

Abstract

The TRECVID 2016 Multimedia Event Detection (MED) challenge evaluates the detection performances of high level complex events in Internet videos with limited number of positive training examples [1]. In this notebook paper, we present an overview of our system, highlighting on the selection and fusion of multiple classification models from a wide range of feature representations to improve the performance. Our MED submissions include 5 system runs for the Pre-Specified (PS) sub-task under 010Ex and the 100Ex condition, and 2 runs for the Ad-Hoc (AH) sub-task under the 10Ex condition. We verify the effectiveness of our developed system by very competitive results obtained. Especially, our primary run ranks first in the PS 100Ex EvalFull task.

1 Introduction

Event detection for complex high-level video is a very difficult task. Some recent survey papers have reviewed the existing approaches used in video event detection systems, covering general frameworks, key modules as well as deep learning based methods used to solve this problem [3, 17]. In this paper, we summarize the methods used in our TRECVID 2016 MED submissions and the results obtained by our system. Figure 1 gives an overview of our developed system for the task. The system first extracts the video features by concept detectors, deep CNN feature, along with some other traditional visual and audio features. These extracted features are fed into trained supervised classifiers to predict the video events. In addition, we also incorporate some zero-shot based methods to rank the test videos according to the estimated relevancies to the target class. In the end, the selection and fusion module tries to select the most important models before fusing them together for the final prediction.

2 System Components

In this section, we briefly introduce each of our used components in the submitted system.

2.1 Concept Detectors

Concept detectors are proved to be very successful in video classification and widely used in the previous MED systems [8]. In this work, we try to incorporate

Table 1: A summary of our submissions.

	Run	Features	Fusion
PS @ 10Ex	primary-1	concept, global spatial, object, motion, audio, zero-shot	weighted fusion by threshold selection
	contrastive-2	concept, global spatial, object, motion, audio, zero-shot	linear regression fusion
	contrastive-3	concept, global spatial, object, motion, audio, zero-shot	logistic regression fusion
	contrastive-4	concept, global spatial, object, motion, audio, zero-shot	linear regression fusion with 0-1 normed scores
	contrastive-5	concept, global spatial, object, motion, audio, zero-shot	logistic regression fusion with 0-1 normed scores
PS @ 100Ex	primary-1	concept, global spatial, object, motion, audio, zero-shot	weighted fusion by threshold selection
	contrastive-2	concept, global spatial, object, motion, audio, zero-shot	linear regression fusion
	contrastive-3	concept, global spatial, object, motion, audio, zero-shot	logistic regression fusion
	contrastive-4	concept, global spatial, object, motion, audio, zero-shot	linear regression fusion with threshold selection
	contrastive-5	concept, global spatial, object, motion, audio, zero-shot	logistic regression fusion with threshold selection
AH @ 10Ex	primary-1	concept, global spatial, object, motion, audio, partial zero-shot	weighted fusion by threshold selection
	contrastive-2	concept, global spatial, object, motion, audio, partial zero-shot	weighted fusion by threshold selection with 0-1 normed scores

the concept information from a wide range of available image and video datasets to help improve our classification system performance.

ImageNet-1000: The ImageNet-1000 concepts comes from the ImageNet contest dataset [10]. The ImageNet dataset contains around 1.26 million training images from 1,000 categories. In our system, we use the pre-trained ResNet152 model to generate the class prediction probabilities on each frames in the video by the last softmax outputs [2]. In this way, the 1000-d outputs are supposed to indicate the likelihood of presence of the corresponding classes in each frame. To get the video level concept scores, we simply average pool all the frame features in the videos.

ImageNet-20574: Similar to our last year's system, we adopt the concept detectors on the ImageNet dataset with 20,574 categories [13]. We retrain the VGG19 model as in [11] on this dataset with extended class labels and take the last softmax outputs as high-level semantic concepts. Similarly, average pooling over all extracted video frames are used.

Places-205: We extract the scene information from the video by using the concept detectors trained on the MIT places dataset [20]. The network we

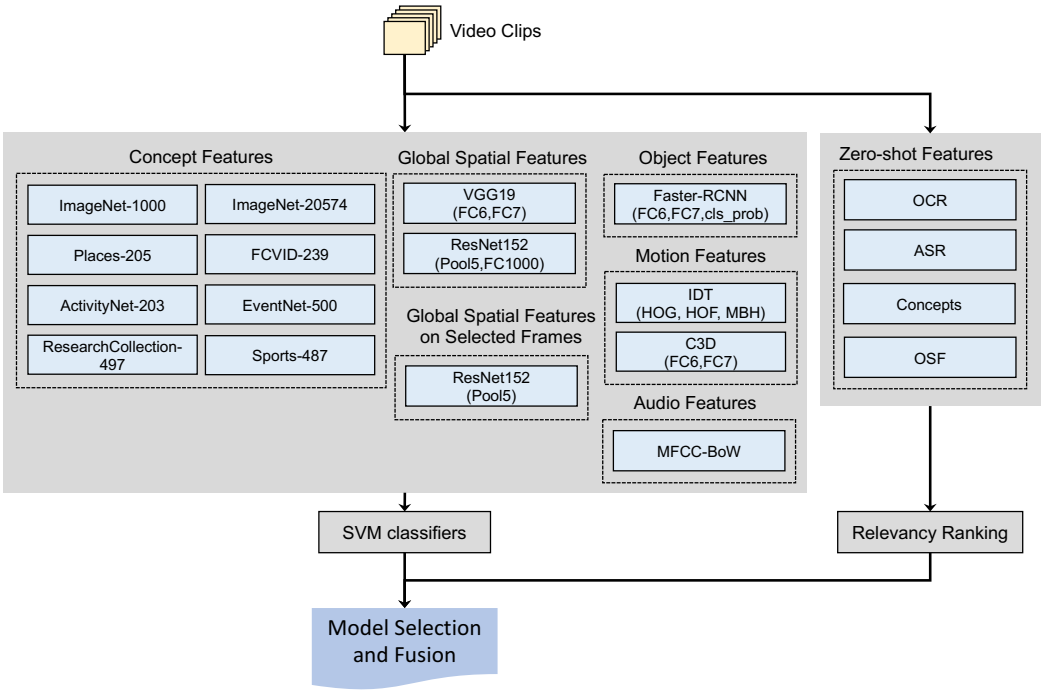

Figure 1: An overview of the key components in our system.

use to extract the scene concepts is the fine-tuned AlexNet model. Also, we simply take the final softmax outputs as concept scores on each frame and average pool all the scores in the same video.

ResearchCollection-497: Our system also includes the concept detectors trained on ResearchCollection-497 dataset provided in [19]. This dataset is derived from the original MED'14 Research Collection dataset. It contains 497 manually annotated concepts based on video frame level. We train 497 concept detectors based on the annotated frame data. For each test video, we apply the trained concept detectors on each video frames. The detection scores are average pooled over all frames in each video to generate the video concept scores.

FCVID-239: We first use the 239 video categories from the recently released Fudan-Columbia Video Dataset (FCVID) [4] to train the video class detectors with CNN model, like in our last year's system [13]. FCVID contains about 90K videos annotated into 239 classes, covering a wide range of topics like social events, procedural events, objects, scenes, etc. We take spatial frames as inputs to fine-tune the VGG19 model for classifying the 239 video categories. Similarly, the frame level responses are averaged to obtain the 233-d video level concept scores for all the MED videos.

ActivityNet-203: We also use the 203 video categories from the ActivityNet video dataset to train the concept detectors with the CNN model. The ActivityNet dataset provides a large scale video benchmark for human activity understanding. It aims at covering a wide range of complex human activities that are of interest to people in their daily living, including eating, drinking, sports, socializing, household activities, etc. We use the VGG19 model pre-trained on ImageNet dataset to extract the FC6 layer features on the ActivityNet video frames. The frame-level responses are averaged to obtain the 4096-d video level

features for all the videos. Then SVM classifiers are trained on the 203 video events as concept detectors.

EventNet-500: The EventNet dataset is another large scale structured video concept library consisting of 500 events [18]. It includes automatic detection models for the video events and constituent concepts using deep learning with around 95K training videos from YouTube. Like in the ActivityNet, we use pre-trained VGG19 model to generate video features and train SVM classifiers as concept detectors.

Sports-487: The 487 sports related concepts are defined by the Sports-1M dataset [6]. The Sports-1M dataset consists of 1 million YouTube videos belonging to 487 classes, mainly providing action and motion concepts annotated at video level. In our experiments, we take the pre-trained C3D model on Sports-1M dataset to generate the concept scores for each video [14].

2.2 Global Spatial Features

Deep CNN models have shown their advantages on various visual recognition tasks, including image classification, object detection, etc. In our developed system, we use two deep CNN models to extract global spatial features on the videos, as described in the following.

VGG19: The VGG19 model used to extract mid-level global spatial features is the same as the one we use in ImageNet-20574 concept detectors [10, 13]. Instead of taking the last softmax outputs, here we keep the outputs of both the FC6 and FC7 layers as mid-level frame features. All the extracted features are average pooled across the video frames to form the video level feature representation.

ResNet152: The Residual Network is the winner model for the ILSVRC 2015 competitions in the ImageNet classification task [2]. We use the 152 layers version of the network, or ResNet152, pre-trained on ImageNet dataset to extract frame features of the MED videos. Here we keep both the Pool5 and FC1000 layers features extracted by the model. Like in the case in the VGG19 model, we simply average pool all the frame features of the video to generate the video level feature representation.

2.3 Global Spatial Features on Selected Frames

As the MED data mainly contains user generated videos, it is possible that some of the video frames in a video clip are not directly relevant to the class label. In our developed system, we utilize a simple video frame selection strategy based on concept detectors in order to alleviate this problem. The main idea is to select a subset of the concept detectors as discussed in Section 2.1 and keep only the frames with at least one high concept score detected. The concept subset is built by keeping all the relevant concepts related to the video classes. The relevances are estimated by the similarity scores between the Word2Vec representation of the words in concept names and video class names. Those frames with all low relevant concept scores are discarded and the extracted features on these frames will not join the average pooling process in the video level feature generation step. In our system, we only conduct selected frame pooling with the extracted ResNet152-Pool5 features.

2.4 Object Features

In addition to the global spatial features, we extract the object level features to describe the video content. The objects in the video frames, like bikes and dogs, are highly related some of the video events like bike tricks and dog shows. To achieve this aim, we use the Faster RCNN model pre-trained on the MS COCO dataset to extract the deep CNN object features [7]. The Faster RCNN make the embedded Region Proposal Network (RPN) to detect objects and extract object features with ROI pooling in an end-to-end manner [9]. For the video data, we first max pool all the object features in each video frame to produce the frame level features. Then the frame level features are average pooled across all the video frames to produce the video level feature representation.

We also use the techniques described in [12] to exploit the object information in the video data. In specific, we first treat the prediction scores of the CNN model trained on ImageNet-20574 data as the object detection scores on each video frames. Then a two layer LSTM model is applied to generate the

video level feature representation.

2.5 Motion Features

To capture the motion information, we extract the improved dense trajectories (IDT) features according to [15]. Briefly, densely sampled local frame patches are first tracked over time and three descriptors are then computed for each trajectory: a 96-d histogram of oriented gradients (HOG) descriptor, a 108-d histogram of optical flow (HOF) descriptor, and a 108-d motion boundary histogram (MBH) descriptor. We first reduce the dimension of these descriptors by a factor of 2 using Principle Component Analysis (PCA) then we encode these features into Fisher Vectors with a codebook of 256 words.

In addition to the traditional IDT features, we also extracted C3D features from the video frame sequences to capture both appearance and motion information [14]. The C3D feature is proved to be very successful in the MED tasks [19]. In our implementation, both the FC6 and FC7 layer features are kept in the C3D model.

2.6 Audio Feature

The audio sound tracks contain useful clues for identifying some video semantics which are usually complementary to the visual features. We utilize the popular MFCCs (Mel-Frequency Cepstral Coefficients), which are computed for every 32ms timewindow with 50% overlap and then quantized into a 4000-d soft weighted bag-of-words representation [5].

2.7 Classification

Similar to our last year's implementation, we simply adopt SVMs as the classifiers. Linear kernel SVM is applied to the IDT features, since it is found working well with the high-dimensional Fisher vector based representations. We adopt χ^2-kernel for all the other features, including concept detector scores, deep CNN features and MFCC features. As the χ^2-kernel SVM is much expensive to compute than the linear kernel one, we use the GPU parallel computing to aid the calculation of the χ^2 kernels among large number of training and test MED data.

2.8 Zero-shot Learning

In this year, We further try to use the zero-shot learning methods to help the video classification in the MED tasks. This is proved to be effective especially when training examples are limited. In specific, the following approaches are used.

OCR: We use the OCR method by tesseract toolkit to detect characters on the video frames. The extracted words are pre-processed as in [19] and converted to feature representation by Word2Vec model pre-trained on the Wikipedia corpus. Relevance between each word extracted in the video and video class names are compared by calculating the cosine distance between their word vectors. Videos with larger accumulated relevant words to the video class name are ranked higher.

ASR: Besides the OCR approach, we further use the ASR methods to extract more text data from the videos. In our ASR system, we use the Kaldi toolkit to extract the speaking English scripts in the videos. The way to deal with the extracted texts are the same as in the OCR system.

Concepts: As described above, we have already trained various concept detectors in the supervised classification systems. These concept detectors and also be used in the zero-shot component. In specific, the relevance between each concept names and video class names can also be directly estimated by calculating the cosine distance between their word vectors. In this way, videos with large accumulated relevant concept scores to the video class name are ranked higher.

OSF: In our developed system, we further use the Object-Scene semantic Fusion (OSF) based zero-shot learning method proposed in [16]. One of the key assumptions OSF makes for zero-shot recognition is that object-scene semantic space is a good proxy for measuring semantic distance of video content, which means that video samples containing similar objects and scenes are likely to belong to the same video class. It compares the object-scene vector representation to video class prototypes represented in the same object-

scene semantic space. OSF models the prototype of test classes by some defined OSR matrix. More details are available in [16].

2.9 Fusion

Given the prediction scores of multiple models, we can capture the video characteristics from different aspects.

It is critical to effectively fuse the multiple scores to generate the final predictions. For the MED video data, some classes are strongly related with particular objects that could be effectively recognized by the CNN features, while others may contain dramatic movements so motion features such as IDT can contribute more significantly. On the other side, some extracted features may not be suitable at predicting certain video events. For example, we find it very helpful to use object features to predict events like parking cars and tailgating, possibly thanks to its successful detection of cars in the video frames. While the object features work poorly in events like metal crafts, possibly due to its incapability of good quality feature extraction related to the metal craft event.

To explicitly select a subset of good important models for the video classification, we first evaluate the classification performance of every models on a split validation set from the training set. Then we rank each models by the evaluated mAP in decreasing order. The subset is selected by setting an MAP threshold and keep only the top ranked models. To fuse the classification scores, we heuristically sum the prediction scores weighted by the corresponding MAP scores evaluated on the validation set.

In addition to the heuristic weighting strategy, we also try to use linear regression and logistic regression to learn the fusing weights of each type of features by the validation set.

3 Result and Discussion

In TRECVID MED 2016, we submitted 5 runs for the PS sub-task under both 10Ex and 100Ex conditions. We also submitted 2 runs for AH sub-task under the 10Ex condition. The submission details are listed in Table 1. In the AH sub-task, the ASR and OCR models are not used. For the fusion methods with 0-1 normed scores described in the table, we simply normalize the all SVM outputs to the range of 0-1 using the scale coefficients learnt on training data.

The evaluation results of MED 2016 are displayed in Figure 2 to 6. Figure 2 and 3 show the performance of the primary runs for PS 10Ex sub-tasks evaluated on the full and sub test set. Figure 4 and 5 show the performance of the primary runs for PS 100Ex sub-tasks evaluated on the full and sub test set. Figure 6 shows the performance of the primary runs for AH 10Ex sub-task evaluated on the full test set. Note that the results of PS tasks are reported by MAP and the ones in AH tasks are reported by MInfAP200%. We can see from these results that our submitted runs (named "nttfudan" with highlighted bars in the figures) achieves very competitive performance in all sub-tasks. Especially in the PS 100Ex EvalFull sub-task, our system achieves the best performance. In the other sub-tasks, our results all rank at the 3rd place and the MAP scores are close to the best reported performances. All these results indicate that our system components and the selection and fusion strategy are effective.

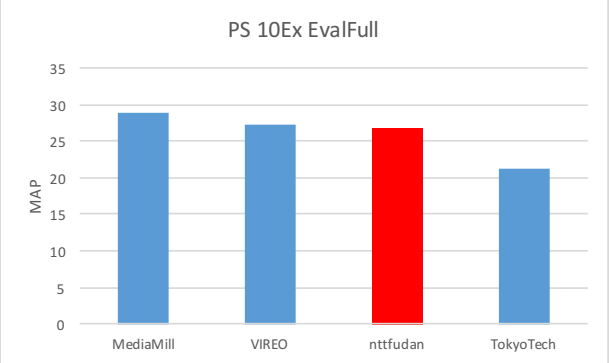

Figure 2: Performance of primary runs for PS 10Ex EvalFull sub-task.

In our contrastive runs, we mainly compare different fusion strategies of the system components. The evaluation results of our runs on the PS tasks are

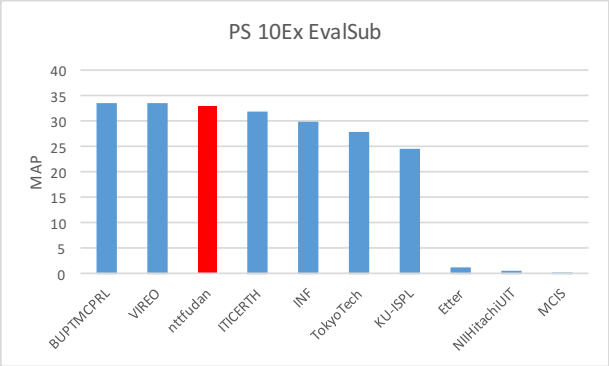

Figure 3: Performance of primary runs for PS 10Ex EvalSub sub-task.

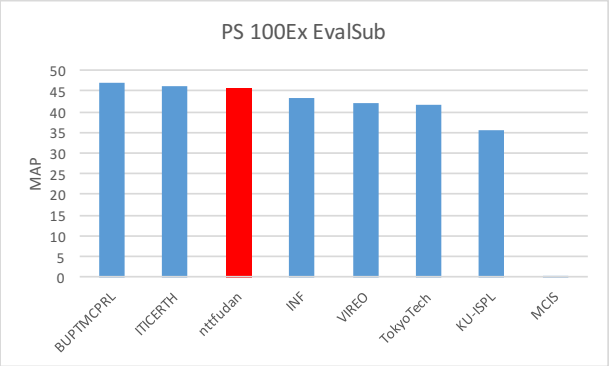

Figure 5: Performance of primary runs for PS 100Ex EvalSub sub-task.

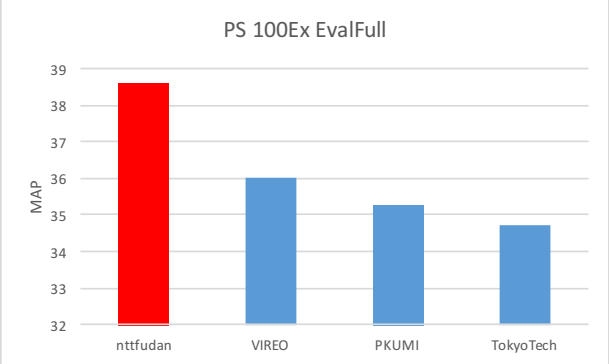

Figure 4: Performance of primary runs for PS 100Ex EvalFull sub-task.

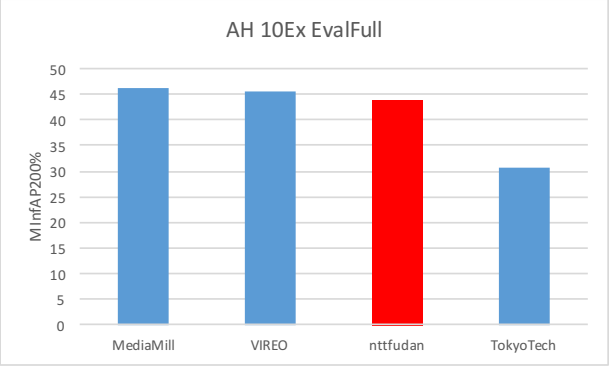

Figure 6: Performance of primary runs for AH 10Ex EvalFull sub-task.

listed in Table 2. Column P1 in the table refers to our primary runs and the columns C2-C5 refer to the four contrastive runs. We can see from the table that our primary runs with heuristically weighted sum of selected models is able to achieve stable and high performances in all cases. The linear regression methods are more reliable than the logistic regression ones. And linear regression methods (C2) on fewer positive examples conditions are slightly better than the primary runs. According to the results in C4 and C5, we find that either norm the individual model scores or applying regression methods based on selected models can improve the performances. This further verifies that model selection is necessary when multiple models are employed in the system.

In addition to the MED'16 test data, we also evaluate our developed system on the MED'14 test dataset to explore the effectiveness of our system components. Under the 10Ex condition, we find that by using the deep spatial features only, we can reach the MAP score of 26.7. If we further add object features in the system, the MAP can reach to 28.6. Then incorporating the motion and audio features can help improve the MAP to 31.6 and 32.3. By further introducing the concept detectors into the system, the performance rises to 34.3. When fusing with the global features on selected frames, the MAP can rise by a margin of 0.5. At last, fusing with zero-shot methods

Table 2: Results (MAP) of our primary and contrastive runs on PS tasks.

Task	P1	C2	C3	C4	C5
PS 10Ex EvalFull	26.7	27.0	17.2	23.2	19.0
PS 10Ex EvalSub	32.8	35.1	23.6	30.6	26.0
PS 100Ex EvalFull	38.6	21.8	5.9	38.1	28.9
PS 100Ex EvalSub	45.7	27.8	9.7	47.6	35.8

can further boost the MAP by 1.3.

References

[1] G. Awad, J. Fiscus, M. Michel, D. Joy, W. Kraaij, A. F. Smeaton, G. Quénot, M. Eskevich, R. Aly, and R. Ordelman. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *Proceedings of TRECVID 2016*. NIST, USA, 2016.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *arXiv.org*, Dec. 2015.

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

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

[5] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID Workshop*, 2010.

[6] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[7] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *Lecture Notes in Computer Science*, pages 740–755. Springer International Publishing, Cham, 2014.

[8] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, G. Quéenot, and R. Ordelman. Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *NIST TRECVID Workshop*, 2015.

[9] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, pages 91–99. Curran Associates, Inc., 2015.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, Apr. 2015.

[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.

[12] Y. Sun, Z. Wu, X. Wang, H. Arai, T. Kinebuchi, and Y.-G. Jiang. Exploiting Objects with LSTMs for Video Categorization. In *ACM Multimedia*, pages 142–146, Amsterdam, The Netherlands, 2016.

[13] Y. Sun, Z. Wu, X. Wang, K. Sudo, Y. Taniguchi, T. Kinebuchi, and Y.-G. Jiang. Ntt-fudan team trecvid 2015: Multimedia event detection. In *NIST TRECVID Workshop*, 2015.

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *ICCV*, pages 4489–4497, Santiago, Chile, Dec. 2015. IEEE.

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

[16] Z. Wu, Y. Fu, Y.-G. Jiang, and L. Sigal. Harnessing object and scene semantics for large-scale video understanding. In *CVPR*, 2016.

[17] Z. Wu, T. Yao, Y. Fu, and Y.-G. Jiang. Deep learning for video classification and captioning. *arXiv preprint arXiv:1609.06782*, 2016.

[18] G. Ye, Y. Li, H. Xu, D. Liu, and S.-F. Chang. Eventnet: A large scale structured concept library for complex event detection in video. In *ACM MM*, MM '15, pages 471–480, New York, NY, USA, 2015. ACM.

[19] H. Zhang, Y. Lu, M. de Boer, Z. Qiu, L. Pang, and C. Ngo. Vireo-tno @ trecvid 2015: Multimedia event detection and video hyperlinking. 2015.

[20] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *NIPS*, pages 487–495. Curran Associates, Inc., 2014.