UEC at TRECVID 2014 SIN task

Keiji Yanai, Hiroyoshi Harada and Do Hang Nga
Department of Informatics, The University of Electro-Communications, JAPAN
{yanai,harada-h,dohang}@mm.cs.uec.ac.jp

Abstract

In this paper, we describe our approach and results for the semantics indexing (SIN) task at TRECVID2014. We submitted four runs for the SIN task of TRECVID 2014 which includes one run submitted last year as a progress run for the 2014 dataset.

In the best run of ours, we used deep convolutional neural network features, Fisher Vector with dense SIFT and Fisher Vector with spatio-temporal local features as features, and combined them in the manner of late fusion. As a result, the best run achieved the mean infAP=0.1537.

1. Introduction

Since TRECVID [14] provides not only a large video date set but also a systematic protocol for evaluating video concept detection performance, it is appreciated by the researchers in the field of video/image recognition. Using this valuable date set, we have been testing our system in these years.

We started participating TRECVID in 2005, and we have been continuously submitting the results to TRECVID for ten years. In the first year, we employed probabilistic region-based image classification for the high-level feature extraction task in TRECVID2005. For the HLF task in TRECVID2006, we extracted some single types of visual features such as color histograms and edge histograms and classified test frames by the support vector machine (SVM). From the results, we realized that a certain feature cannot satisfy all the concepts. For TRECVID2007, we attempted to adopt a kind of fusion to combine some features to get a result that is effective for any kind of concept. What we did is to apply SVM to the extracted features respectively, and then to fuse these SVM classifiers by linear combination with weights selected by cross validation. This method is more effective, however it is intractable to implement when more than 3 kinds of features are extracted. For the TRECVID2008 HLF task, we still used the thought of developing a framework to fuse a number of features to get more effective performance. At that time we added some new features. In addition, inspired by some papers [3, 16], we implemented a simple version of Adaboost [12] algorithm as a method for late fusion. This method can estimate optimal weights automatically no matter how many kinds of features there are. For the TRECVID2009 HLF task, we explore the feature fusion strategy furthermore. In that year, we used the AP-weighted fusion [17] and Multiple Kernel Learning (MKL) [5, 15] both of which achieved the best performance in our preliminary experiments. For the TRECVID2010 Semantic Indexing Task, we used a novel spatio-temporal (ST) feature [9] which is useful for feature-fusion-based action recognition with Multiple Kernel Learning (MKL). For the TRECVID2011 Semantic Indexing task we used six features including ST feature, word histogram and category name detection and use MKL-SVM in all runs. For the TRECVID2012 Semantic Indexing task, we used only three features, SURF, color and spatio-temporal feature. For the TRECVID2013 Semantic Indexing task, the method we used that year was the same as the previous year.

For this year, we introduced the Deep Convolutional Neural Network (DCNN) extracted by DCNN pretrained with ImageNet Challenge 1000 categories in addition to Fisher Vector with dense SIFT and Fisher Vector with spatio-temporal local features.

2. Overview

This year, we use three features, the Deep Convolutional Neural Network (DCNN) extracted by DCNN pre-trained with ImageNet 1000 categories, SIFT with Fisher Vector (FV) coding (SIFT-FV) and spatio-temporal features with FV (ST-FV) except for the progress run we submitted last year. We extracted DCNN features from the keyframe in each shot, SIFT-FV and ST-FV from every five frames. Finally we apply linear SVM and fuse all the SVM output value with late fusion. For Run 1 we used all the three features, for Run 2 we used only DCNN features and for Run 3 we used SIFT-FV and ST-FV. Regarding Run 4 which is a progress run we submitted last year, we used a spatio-temporal (ST) feature [9] with standard bag-of-features coding, SURF with bag-of-features, and color histogram which are the same as the other runs we submitted in 2013.

3. Method

For the runs we submit this year, we used DCNN features, FV-SIFT and FT-ST.

3.1. Deep Convolutional Neural Network (DCNN)

Recently, it has been proved that Deep Convolutional Neural Network (DCNN) is very effective for large-scale object recognition. However, it needs a lot of training images. In fact, one of the reasons why DCNN won the Image Net Large-Scale Visual Recognition Challenge (ILSVRC) 2013 is that the ILSVRC dataset contains one thousand training images per category [4]. This situation does not fit common visual recognition tasks including the TRECVID SIN task. Then, to make the best use of DCNN for common image recognition tasks, Donahue et al. [1] proposed the pre-trained DCNN with the ILSVRC 1000-class dataset was used as a feature extractor.

Following Donahue et al. [1], we extract the network signals from the previous layers of the last output layer in the pre-trained DCNN as a DCNN feature vector. We use the pre-trained deep convolutional neural network in Overfeat [13] with the ILSVRC 2013 dataset. This is slight modification of the network structure proposed by Krizhevsky et al. [4] at the ILSVRC 2012 competition. In the experiments, we extract raw signals from layer-7, where the dimension of the signals are 4096, respectively, and L2-normalize them to use them as DCNN feature vectors.

3.2. Local Feature

As a conventional static appearance feature, we use densely sampled SIFT. We extract random-sampled SIFT features every five frames, and code them with Improved Fisher Vector [11] regarding each shot. We obtained one FV-SIFT vector for each shot

3.3. Spatio-Temporal Feature

We use a spatio-temporal (ST) feature [9, 7] which is based on the SURF (Speeded-Up Robust Feature) features [2] and optical flows detected by the Lucas-Kanade method [6].

For designing a new ST feature, we set the premise that we combine it with holistic appearance features and motion features. Therefore, the important thing is that it has different characteristics from other kinds of holistic features. Following this premise, we extend the method proposed in [8]. In the original method, we detect interest points and extract feature vectors employing the SURF method [2], and then we select moving interest points employing the Lucas-Kanade method [6]. In the original and proposed method, we use only moving interest points where ST features are extracted and discard static interest points, because we expect that it is a local feature which represents how objects in a video are moving. In addition to the original method, we newly introduce Delaunay triangulation to form triples of interest points where both local appearance and motion features are extracted. This extension enables us to extract ST features not from one point but from a triangle surface patch, which makes the feature more robust and informative. The characteristic taken over from the original method [8] is that it is much faster than the other ST features such as cuboid-based features, since it employs SURF [2] and the Lucas-Kanade method [6], both of which are known as very fast detectors. The detail should be referred to [9, 7].

3.4. Vector Quantization of Features: Fisher Vector

After extracting SIFT or ST local features, we apply PCA and code them into Fisher Vectors (FV) with the GMM consisting of 256 Gaussians. After obtaining FV, we normalized them with power normalization and L2-normalization following [11].

3.5. Classification: Linear SVM

For training and classifying, we use standard linear SVM independently for each of the three features, and combine all the SVM output values with averaging. This year we used uniform weights for late fusion, because we had no time to tune weights.

4. Experiments

Table 1 shows the four runs we submitted and the value of infAP. The best runs of ours ranked 48th among 75 runs for the SIN task as shown in Figure 1.

These results shows that DCNN is very effective for SIN tasks in the same way as the other object recognition tasks, because DCNN features extracted from only one keyframe per shot outperformed the combination of FV-SIFT and FV-ST which were extracted from every five frames with large margin. Compared with Run 4 which is a progress run, all the other results were improved by introducing not only DCNN but also Fisher Vector.

5. Conclusions

In the Semantic indexing task of TRECVID2014 [10], we extracted the DCNN, FV-SIFT and FV-ST and combined them using late fusion with uniform weights. We have achieved 0.1537 mean infAP at most. The results indicates that DCNN features is very effective to boost performance.

References

[1] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *Proc. of International Conference on Machine Learning*, 2014.

[2] B. Herbert, E. Andreas, T. Tinne, and G. Luc. Surf: Speeded up robust features. *Computer Vision and Image Understanding*, pages 346–359, 2008.

[3] W. Jiang, S. Chang, and A. Loui. Kernel sharing with joint boosting for multi-class concept detection. In *Proc. of CVPR Workshop on Semantic Learning Applications in Multimedia*, 2007.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[5] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[6] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. of International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[7] D. H. Nga and K. Yanai. A dense surf and triangulation based spatio-temporal feature for action recognition. In *Proc. of International Multimedia Modeling Conference (MMM)*, 2014.

[8] A. Noguchi and K. Yanai. Extracting spatio-temporal local features considering consecutuveness of motions. In *Proc. of Asian Conference on Computer Vision(ACCV)*, 2009.

[9] A. Noguchi and K. Yanai. A surf-based spatio-temporal feature for feature-fusion-based action recognition. In *Proc. of ECCV WS on Human Motion: Understanding, Modeling, Capture and Animation*, 2010.

[10] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quenot. Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2014*, 2014.

[11] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Proc. of European Conference on Computer Vision*, 2010.

[12] R. Schapire, Y. Freund, and R. Schapire. Experiments with a New Boosting Algorithm. In *Proc. of International Conference on Machine Learning*, pages 148–156, 1996.

[13] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proc. of International Conference on Learning Representations*, 2014.

[14] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *Proc. of ACMMM WS on Multimedia Information Retrieval*, pages 321–330, 2006.

[15] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Proc. of IEEE International Conference on Computer Vision*, pages 1150–1157, 2007.

[16] D. Wang, X. Liu, L. Luo, J. Li, and B. Zhang. Video diver: generic video indexing with diverse features. In *Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 61–70, 2007.

[17] M. Wang and X. S. Hua. Study on the combination of video concept detectors. In *Proc. of the 16th ACM international conference on Multimedia*, pages 647–650, 2008.

Table 1. Four kinds of the runs for the SIN tasks of TRECVID 2014.

Runs	Description	infAP
Run1:UEC14_1	DCNN, FV-SIFT and FV-ST	0.1537
Run2:UEC14_2	only DCNN (overfeat)	0.1199
Run3:UEC14_3	FV-SIFT and FV-ST	0.0980
Run4:UEC13_2	BoF-SURF, color histogram, and BoF-ST combined with MKL	0.0663

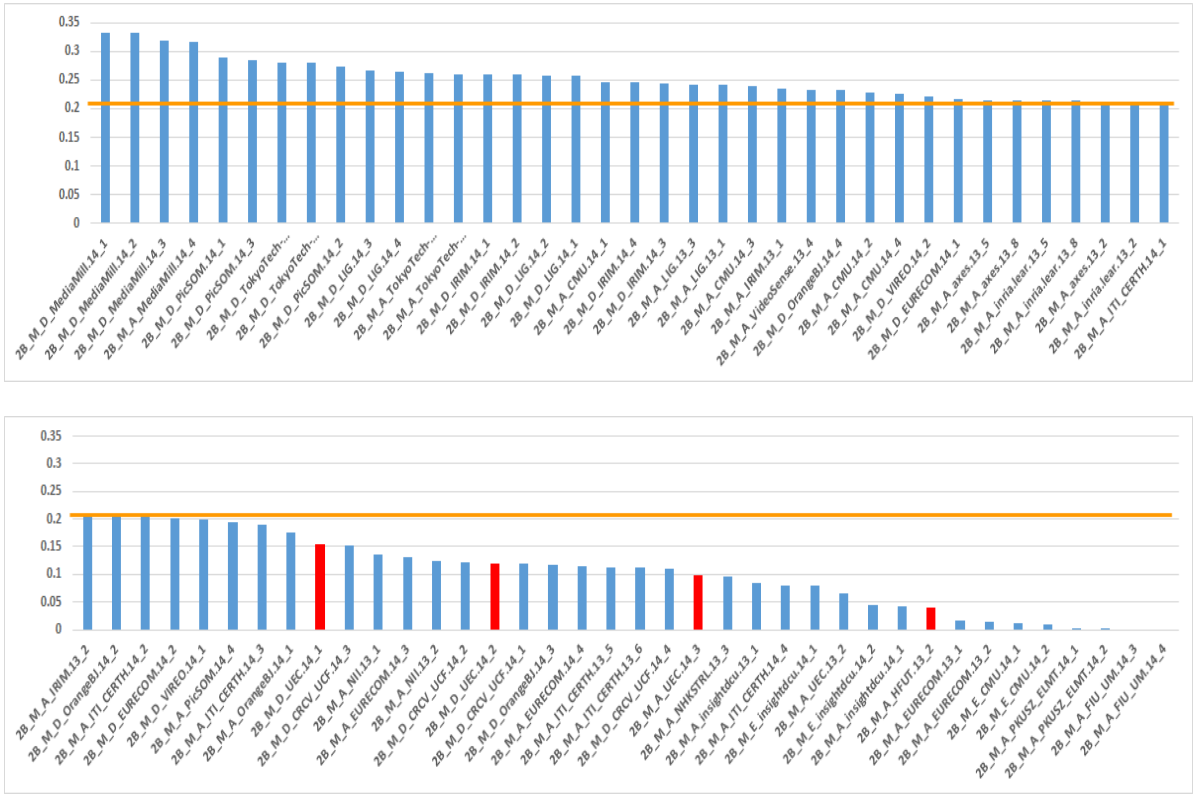

Figure 1. The comparison on the mean average precision (MAP) of all the runs submitted to the TRECVID 2014 SIN task. The red bars show the results of UEC team among 98 runs, and the orange line indicates the median of all the MAPs.