UEC at TRECVID 2010 Semantic Indexing Task

Yasushi Shimoda, Akitsugu Noguchi and Keiji Yanai
Department of Computer Science, The University of Electro-Communications, JAPAN
{shimoda-y,noguchi-a,yanai}@mm.cs.uec.ac.jp

Abstract

In this paper, we describe our approach and results for the semantics indexing task (SIN) at TRECVID2010. This year, we focused on spatio-temporal features using multi-frame informations, and we used Multiple Kernel Learning as a fusion method to combine all these features in the same way as last year.

Our submitted runs are as follows:

- *(Run1) UEC_MKL*

  fusion of five kinds of the features including color, faces, motion features, Gabor textures and SURF-based spatio-temporal features, with Multiple Kernel Learning (MKL).

- *(Run2) UEC_AVG*

  fusion of five kinds of the same features as Run1 with a standard Support Vector Machine (SVM) and a uniformly-combined kernel.

In all the runs, we used the same five kinds of the features, spatio-temporal features, Gabor texture features, motion histograms, color histograms, and faces. Run1 used Multiple Kernel Learning. Run2 combined the five kernels uniformly each of which corresponds to one of the five kinds of the features, and applied a standard SVM. Since MKL estimates weights to combine kernels, as a result of the full-category SIN task, Run1 yielded the best performance (infAP=0.0478) among our two runs.

1. Introduction

Since TRECVID [8] 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.

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 [2, 11], we implemented a simple version of Adaboost [7] 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 [12] and Multiple Kernel Learning (MKL) [3, 9] both of which achieved the best performance in our preliminary experiments. This year, for the TRECVID2010 Semantic Indexing Task, we used a novel spatio-temporal (ST) feature [6] which is useful for feature-fusion-based action recognition with Multiple Kernel Learning (MKL)

2. Overview

In [6], we proposed new spatio-temporal (ST) features based on SURF and Delaunay triangulation, and used BoFr (Bag-of-Frames) approach with Gabor texture features and motion features. For the TV2010 SIN task, we use these features as main features, and added color histograms and the number of faces as additional features. This year, we do not use sound features and textual ASR features. Then two kinds of fusion methods are applied to model all the features, respectively. One is MKL, and the other one is SVM with a uniformly-weighted combined kernel.

2.1. Features

2.1.1. ST feature

We use a novel spatio-temporal (ST) feature which is based on the SURF (Speeded-Up Robust Feature) features [1] and optical flows detected by the Lucas-Kanade method [4].

For designing a new ST feature, we set the premise that we combine it with holistic appearance features and motion features by Multiple Kernel Learning (MKL). 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 [5]. In the original method, we detect interest points and extract feature vectors employing the SURF method [1], and then we select moving interest points employing the Lucas-Kanade method [4]. 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 [5] is that it is much faster than the other ST features such as cuboid-based features, since it employs SURF [1] and the Lucas-Kanade method [4], both of which are known as very fast detectors. The detail should be referred to [6].

2.1.2. Gabor feature

We use Gabor texture histograms as an appearance feature. A Gabor texture feature represents texture patterns of local regions with several scales and orientations. In this paper, we use 24 Gabor filters with four kinds of scales and six kinds of orientations. Before applying the Gabor filters, we divide a frame image extracted from video shots into 20×20 blocks. We apply the 24 Gabor filters to each block, then average filter responses within the block, and obtain a 24-dim Gabor feature vector for each block. Totally, we extract 400 24-dim Gabor vectors from each frame image.

2.1.3. Motion

As a holistic motion feature, we built motion histograms over a frame image. This feature is expected to have different discriminative power from the proposed ST feature. We extract motion features at grid points with every 8 pixels using the Lucas-Kanade methods [4]. Extracted motion features from each grid are voted to histogram of 7 direction and 8 motion magnitude.

2.1.4. Camera motion detection

If camera motion exists, extracted ST features and motion features do not capture action of people or objects correctly. Therefore, we detect camera motion, and we do not use ST features and motion features when camera motion is detected. To detect camera motion, we calculate motion features based on the Lucas-Kanade method at every 8-pixel grid. If the region where motion is detected is larger than a predefined threshold, we consider camera motion is detected. Camera motion compensation will be our future work.

2.1.5. Vector Quantization of Features: Bag-of-Frames

In most of the existing work on video shot classification, features are extracted only from key frames. However, extracted features depend on selected frames, and it is difficult to select the most informative key frame. Then, we extract features from all frames within each video shot, we vector-quantize all of them and convert them into the bag-of-features (BoF) representation within each shot. While the standard BoF represents the distribution of local features within one image, the BoF employed in this paper represents the distribution of features within one shot which consists of several frame images. We call this BoF regarding one video shot as bag-of-frames (BoFr). ST features are obtained from every N frame images, while motion and appearance features are obtained from one frame image. In both cases, we aggregate Gabor and Motion features within all the frame images extracted from one video shot, and convert them into the BoFr histograms.

2.1.6. Color

Basically, we extract color and faces features from a keyframe of each shot, and we do not used the BoFr method. We use a normal color histogram as the color feature. The axises of RGB, Luv, HSV color space are divided in quarters and a 64-bin histogram is generated. For getting some location information, besides extracting from global scale of the image, we also tried to extract a 768 bins histogram by dividing the image to 4×3 grid segments.

2.1.7. Faces

We perform a face detection by using Haar-like features [10]. The number of faces is expected to help handle with somes concepts related to people.

2.2. Feature Fusion Fusion with Multiple Kernel Learning

Multiple Kernel Learning (MKL) is an extension of a support vector machine (SVM). MKL treats with a combined kernel which is a weighted liner combination of several single kernels, while a normal SVM treats with only a single kernel. MKL can estimates weights for a linear combination of kernels as well as SVM parameters simultaneously in the train step. The training method of a SVM employing MKL is sometimes called as MKL-SVM. MKL-SVM is a relatively new method which was proposed in 2004 in the literature of machine learning [3], and recently MKL is applied to image recognition.

Since by assigning each image feature to one kernel MKL can estimate the weights to combine various kinds of image feature kernels into one combined kernel, we can use MKL as a feature fusion method.

In this paper, we use the multiple kernel learning (MKL) to fuse various kinds of image features. With MKL, we can train a SVM with a adaptively-weighted combined kernel which fuses different kinds of image features. The combined kernel is as follows:

$$K_{comb}(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{K} \beta_j K_j(\mathbf{x},\mathbf{y})$$
$$\text{with } \beta_j \geq 0, \ \sum_{j=1}^{K} \beta_j = 1. \qquad (1)$$

where β_j is weights to combine sub-kernels $K_j(\mathbf{x},\mathbf{y})$. MKL can estimate optimal weights from training data.

3. Experiments

We made two runs as shown in Table 1 and Table 2. The difference among two runs are only fusion methods. The features used in the experiments were the same over all the runs. Our team reached rank 14 (among 30 teams) for the full-category SIN task as shown in Table 1 and Figure 1 and rank 12 (among 37 teams) for the light-category SIN task in TRECVID2010 as shown in Table 2 and Figure 2.

Figure 3 shows the results of Run 1 of the evaluated 30 categories among the submitted 130 categories. For "Dancing" and scenery concepts such as "nighttime" and "mountain", we achieved good results compared with the results of the other teams. For these concepts, the bag-of-frames method worked well. Especially for "dancing" ST features can be regarded as being effective, since it is an action concept. Poor results for some concepts is mainly due to camera motions. Many Web videos used in TV2010 contains various kinds of intentional camera motions such as zoom and pan. In the experiments, in case that camera motion is detected, we regarded ST features and motion features as zero vectors. This led poor performance for some concepts such as "bus", "running" and "cheering" many videos of which are taken in the outdoor.

Figure 4 shows the weight estimated by MKL for Run1. This shows only small differences of the weights among the concepts, and seems to have no prominent tendency.

Table 1. 2 runs for the semantics indexing task (full category) in TRECVID2010.

Runs	Description	infAP
Run1 UEC_MKL	Combine color, face, motion, Gabor and BoSTF model of local pattern features Multiple Kernel Learning (MKL)	0.0478
Run 2 UEC_AVG	Combine color, face, motion, Gabor and BoSTF model of local pattern features SVM with a uniform kernel	0.0117

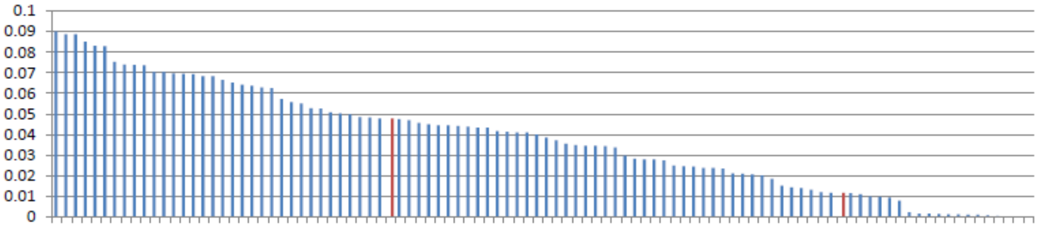

Figure 1. The comparison with results in TRECVID 2010. Red lines show the full-category results of UEC_MKL(Rank 32) and UEC_AVG(Rank 70) among 86 runs.

4. Conclusion

In the semantics indexing task (SIN) of TRECVID2010, we used ST features, Gabor and motion features in addition to the conventional color histogram and the number of faces as features, and used Multiple Kernel Learning to combine them. In the best runs among our submission, we have achieved the 0.0487 average precision (AP).

As future work, we plan to explore feature fusion by MKL and explore features using multi-frame approach furthermore.

References

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

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] P. Viola and M. Jones. Rapid Object Detection Using a Boosted Cascade of Simple Features. In *Proc. of IEEE Computer Vision and Pattern Recognition*, volume 1, 2001.

[11] 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.

[12] 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 2. 2 runs for the semantics indexing task (10 category) in TRECVID2010.

Runs	Description	infAP
Run1 UEC_MKL	Combine color, face, motion, Gabor and BoSTF model of local pattern features Multiple Kernel Learning (MKL)	0.0393
Run 2 UEC_AVG	Combine color, face, motion, Gabor and BoSTF model of local pattern features SVM with a uniform kernel	0.0077

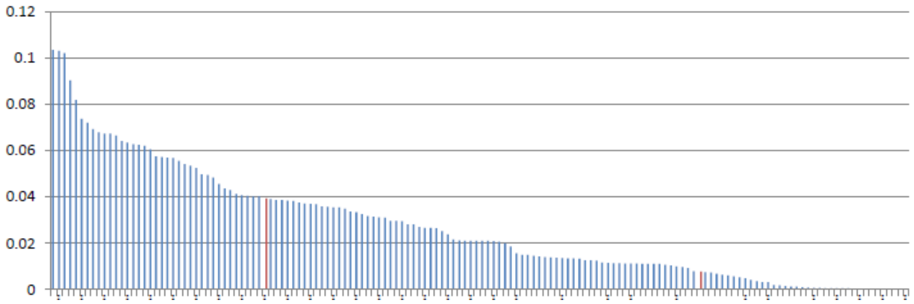

Figure 2. The comparison with light-category results in TRECVID 2010. Red lines show our results of UEC_MKL(Rank 31) and UEC_AVG(Rank 94) among 128 runs.

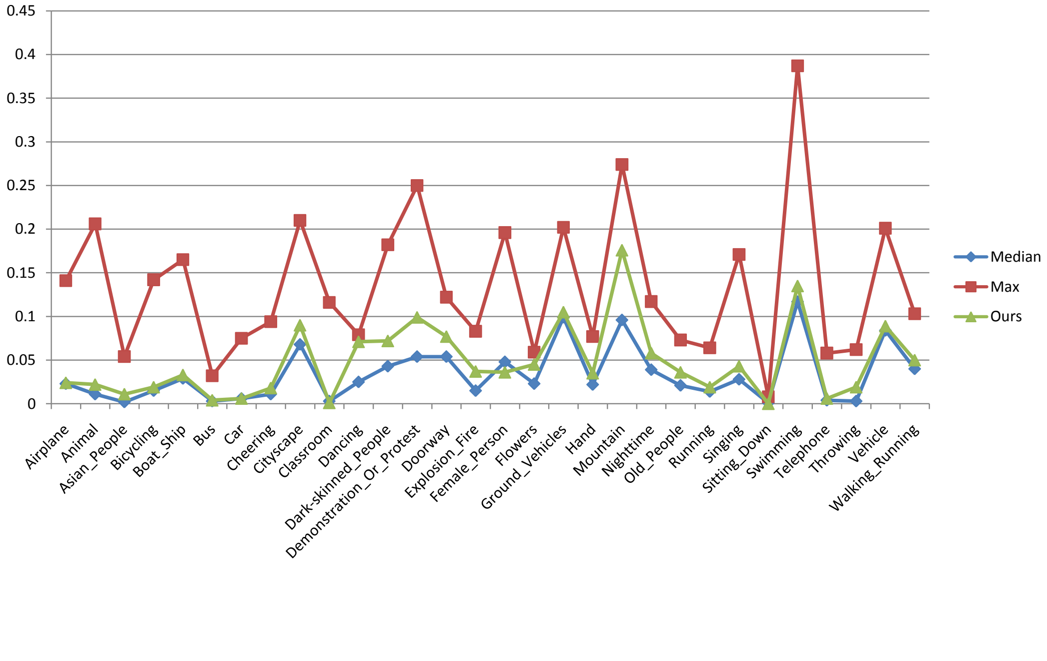

Figure 3. The comparison with median and best results of full category in TRECVID 2010.

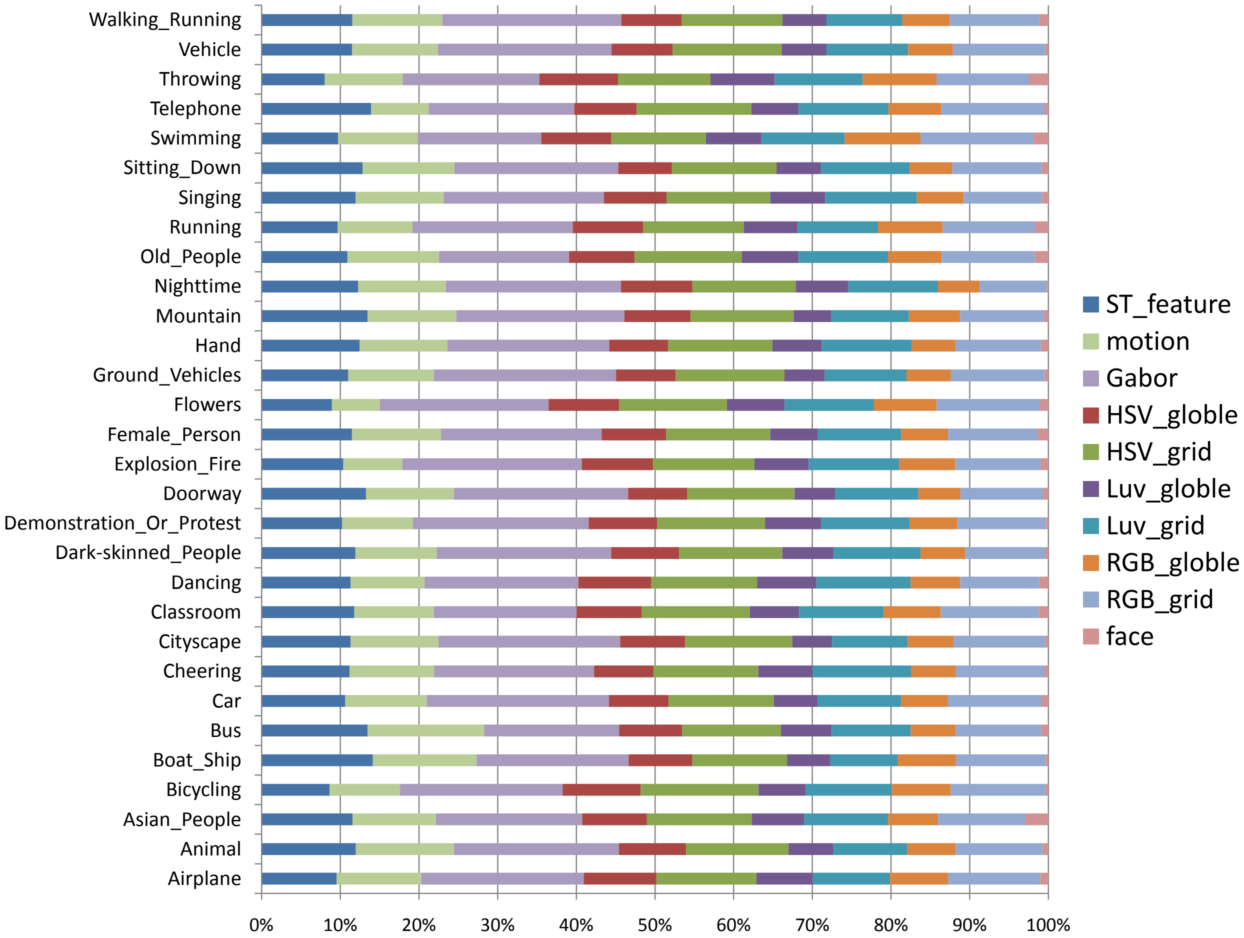

Figure 4. Estimated weights by MKL with full category in TRECVID 2010.