UEC at TRECVID 2012 SIN and MED task

Kazuya Hizume and Keiji Yanai
Department of Computer Science, The University of Electro-Communications, JAPAN
{hizume-k,yanai}@mm.cs.uec.ac.jp

Abstract

In this paper, we describe our approach and results for the semantics indexing (SIN) task and Multimedia event detection (MED) task at TRECVID2012.

In our run of SIN task, we used three features, spatio-temporal (ST) features, SURF and color features. This year, we use all frame to extract features. This run used Multiple Kernel Learning as a fusion method to combine all these features in the same way as last year. Our submitted run is F_A_UEC1_1. As a result of the full-category SIN task, run reached a performance infAP=0.116.

In MED task, we divide videos to shots which are 3000 frames at most and extract SURF, ST features from shots. Then, we select positive shots with VisualRank method from. We get the average of the top three shot scores as the original video score.

1. Introduction

Since TRECVID [11] 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, 14], we implemented a simple version of Adaboost [10] 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 [15] and Multiple Kernel Learning (MKL) [4, 13] 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 use six features including ST feature, word histogram and category name detection and use MKL-SVM in all runs. For the TRECVID2012 Semantic Indexing task, this year, we use only three features, SURF, color and spatio-temporal feature. We also participate in Multimedia event detection task, and use the system of SIN task that splitting videos into shots and VisualRank method[3] are added.

2. Overview

Semantic Indexing

This year, we use three features, SURF, color and spatio-temporal (ST) feature[9]. SURF and color features are extracted from all frames. We quantize these features by Bag-of-Features representation, and apply MKL-SVM to model all features.

Multimedia event detection

We use two features, SURF and ST feature. Extraction methods of these features are the same as SIN task. We divide each video into shots which consists of 3000 frames at most, then extract features from each shot and create an unsupervised shot ranking with VisualRank method[3]. We use the top 500 shots of the ranking as positive shots. Scores of each shot regarding the given events are calculated with MKL-SVM, and the final score is the average among the top three shot scores for each video.

3. Semantic Indexing

3.1. Feature extraction

3.1.1. ST feature

We use a spatio-temporal (ST) feature [9] which is based on the SURF (Speeded-Up Robust Feature) features [1] and optical flows detected by the Lucas-Kanade method [7].

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 [8]. 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 [7]. 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 [1] and the Lucas-Kanade method [7], both of which are known as very fast detectors. The detail should be referred to [9].

3.1.2. Vector Quantization of Features: Bag-of-Features

In most of existing works on video shot classification, features are extracted only from key frames. However, the extracted features depend on selected frames, and it is difficult to select the most informative key frame.

This year, we extract the features from all frames. The extracted features is vector-quantized and converted into the bag-of-features (BoF) representation within each shot.

Then we use spatial pyramid matching technique[5] to BoF representation. We divide the frames to 2 × 2 regions, and generate BoF vectors within each region. We applied this technique to SURF and color features, because these features are extracted from one frame.

Also, we use soft assignment[12] to BoF representation. When allocating to the code word of BoF, make assignments to a plurality of code words. It is to be noted that the value to be assigned is the L1 normalized inverse of the distance between codewords.

3.1.3. Local pattern

We use SURF [6] as a local pattern feature. The local patches are sampled randomly, and they are vector-quantized to convert them into BoF vectors. The codebook is built by performing the k-means clustering with features extracted from one key frame of all the shots in the training videos. We set the size of the codebook as 1000. Since we use a spatial pyramid with 1 × 1 and 2 × 2 regions, totally we generate a 5000 dimensional feature vector.

3.1.4. Color

We extract RGB color histogram features from all pixels of selected frames of each shot. In the same way as SURF, we generate a 5000 dimensional BoF vector.

3.1.5. 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 [4], 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 an adaptively-weighted combined kernel which fuses different kinds of image features. The combined kernel is as follows:

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

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

3.2. Experiments

Table 1 shows runs we submitted and the value of infAP. Figure 1 shows the result of our run of the evaluated 50 categories among the submitted 346 categories and compare with median and best of all team. Our run reached rank 37 (among 51 runs) for the full-category SIN task as shown in Figure 2 and rank 60 (among 89 runs) for the light-category.

This year there is also time constraints, we use only the amount of relatively basic feature. Some categories of high recognition rate for the entire are relatively high in our result, so there is some degree of effectiveness of the current feature amount. But accuracy is much lower than in the category that the image feature is more important than spatio-temporal feature, for example Airplane, Baby, Kitchen and Glasses. It is necessary to add a valid image features.

4. Multimedia event detection

We apply the system used in the SIN task to the MED task. Our system is intended to recognize per shot, not to recognize per video, so we need to divide the videos of the dataset in the MED task into shots.

By shot segmentation, there are many shots that do not include the event. If we use these 'not include' shots as a positive data, adverse effect on learning. So, we select a shot by unsupervised ranking with VisualRank method[3], and treat a shot of the top ranking as a positive shot.

4.1. Dividing videos into shots

Shot segmentation is performed to calculate the color histogram difference between the frame images. Each video is divided into shots which consists 3000 frames at most, then the features are extracted from each shot.

4.2. Feature extraction

For the MED task, we use two features, SURF and ST feature. The feature extraction methods are the same as the SIN task.

4.3. Selecting shots with VisualRank

As a method on visual-feature-based shot ranking, we employ the VisualRank method[3], which is an image ranking method based on the widely known Web page raking method, PageRank. PageRank calculates ranking of Web pages using hyper-link structure of the Web. In VisualRank, index value of the image is estimated by iterative calculation using the image similarity matrix instead of hyper-link structure. To compute the similarity of shot, we use the histogram intersection of BoF vector of ST feature in our system. An equation to compute VisualRank is as follows:

$$r = \alpha S r + (1-\alpha) p \quad (0 \leqq \alpha \leqq 1) \quad (2)$$

Table 1. run for the semantics indexing task in TRECVID2012.

Runs	Description	full	light
Run1:UEC1_1	Combine SURF, color, features and Multiple Kernel Learning (MKL)	0.116	0.144

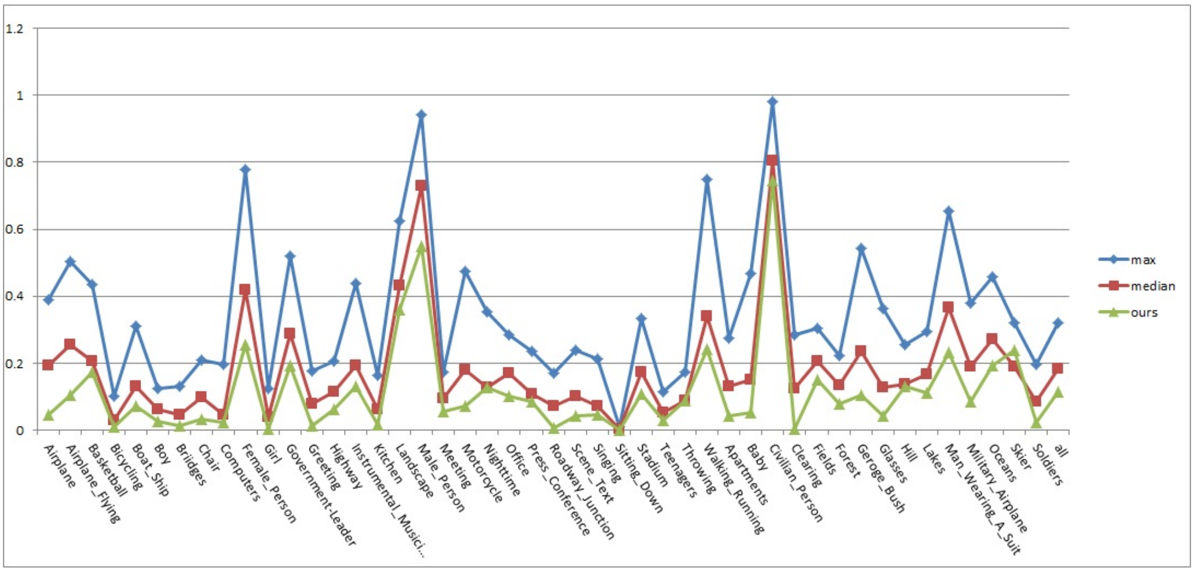

Figure 1. The comparison with median, best and Our run of full category in TRECVID 2012.

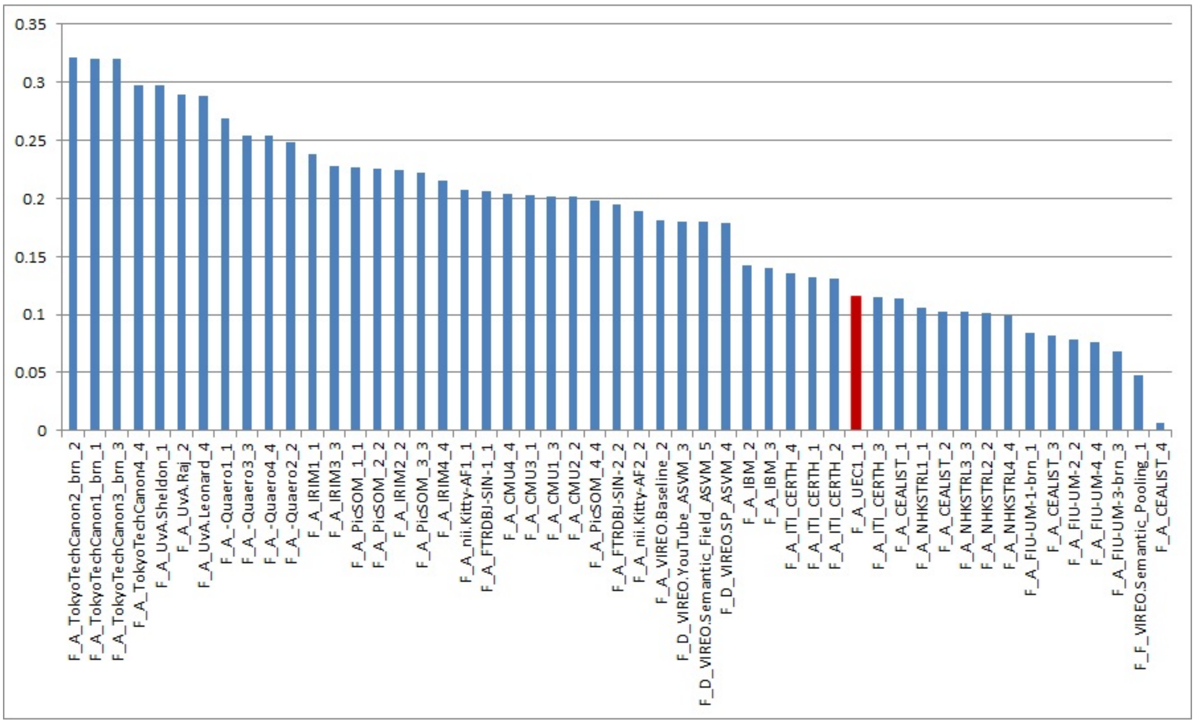

Figure 2. The comparison with results in TRECVID 2012. Red lines show the full-category results of UEC team among 51 runs.

where S is the column-normalized similarity matrix of images, p is a damping vector, and r is the ranking vector each element of which represents ranking score of each image. α plays a role to control the extent of effect of p. Commonly, α is set as 0.8 or more and we set 0.85. Moreover, we set a uniform damping vector.

After the ranking vector is obtained, we treat top 500 shot of the ranking vector as a positive shots and use for the training data.

4.4. Score and threshold decision

The score is calculated per shot by MKL-SVM. The average of the top three scores of video shots is used for the original video score.

Threshold is calculated from 2-fold cross validation scores of learning videos classified by MKL-SVM. We use the average of the original video scores for threshold.

4.5. Experiments

Figure 3 and 4 shows the result score of NDC, P_{MD} and P_{FA}. Our team is the result of lower unfortunately. This is due to that the value of P_{MD} can be very large when compared with other teams. Although not seen in so that there is a big difference with other teams for P_{FA}, the difference between the NDC had spread by the difference of the P_{MD} because $Cost_{MD}$ is much larger than $Cost_{FA}$. This has affected the accuracy of the system used in the SIN task, it is also necessary to reconsider a method to determine the score of the original video from the score of the shot. Our system uses the values of the top three shots to determine the original video score, however, it is required how to determine the score with an awareness of the original video, such as the number and time of each shot of the original video.

Figure5 shows a precision of selected shots when applying VisualRank method for 20 categories. We calculated the precision rate of a) 100 shots randomly chosen as the baseline, b) 100 shots randomly chosen from the top 500 shots applied VisualRank and c) top 100 shots applied VisualRank. In the category of almost all we obtained a precision of b) higher than baseline, but of c) lower than baseline. This is because the shot in the middle of a large and complex behavior has gone into the top. Features cannot be obtained sufficiently from such a shot because there is little movement in the shot and the time of shot is very short. In addition, because there are a relatively large number of intermediate shots, they were ranked in the top. Therefore, it is necessary to apply the method to eliminate as much as possible such intermediate shot.

5. Conclusion

In the Semantic indexing task of TRECVID2012, we extracted SURF, color and ST features from all frame and used Multiple Kernel Learning to combine them. We have achieved 0.116 average precision. In the Multimedia event detection task, we divided videos to shots which are 3000 frames at most, then extracted SURF and ST features from each shots. We select a shot by unsupervised ranking with VisualRank method, and treat a shot of the top ranking as a positive shot. The original video score was the average of the top three shot scores.

As future work, it is necessary to reconsider the effective image features to be added as a whole system. We are considering the introduction of the feature that has been successful in the field of object recognition such as the Fisher vector. In the MED task, since we split shot as additional processing, we explore how to determine the final score of the video taking into consideration the percentage of the original video of each shot and the algorithms to select the positive shot.

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] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(11):1877–1890, 2008.

[4] 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.

[5] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc.of IEEE Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.

[6] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

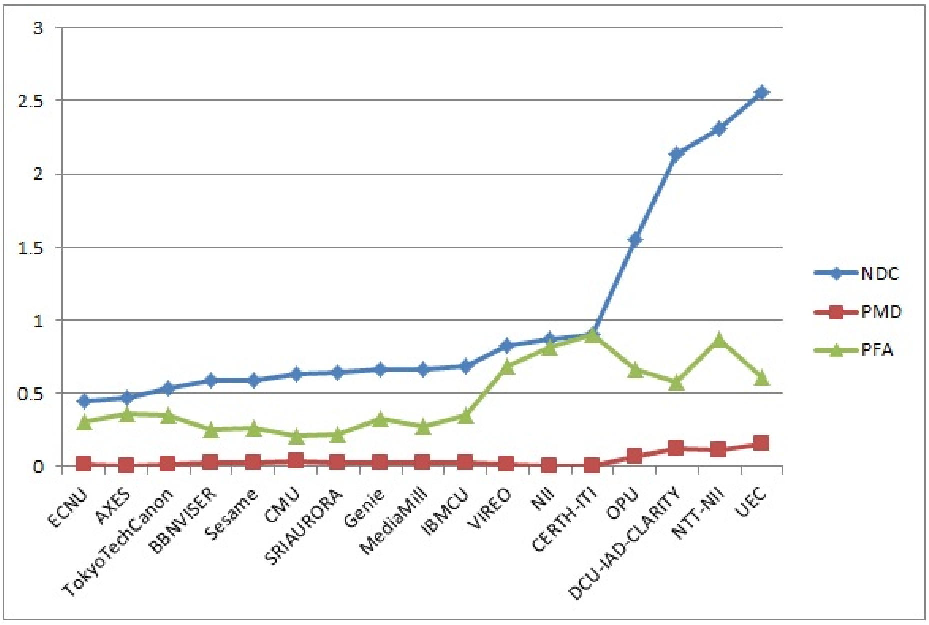

Figure 3. result PreSpec events in TRECVID 2012.

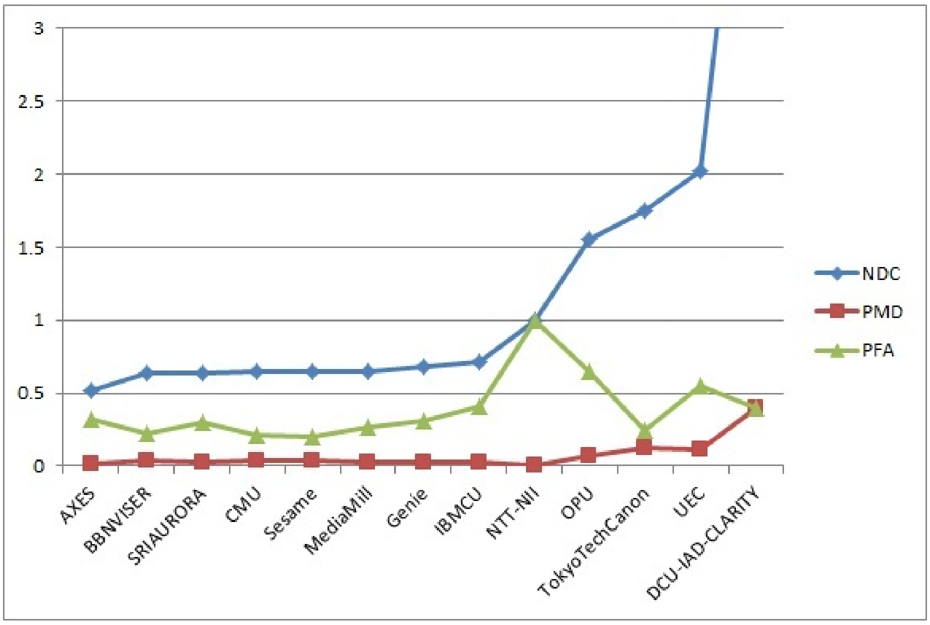

Figure 4. result AdHoc events task in TRECVID 2012.

[7] 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.

[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] 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.

[11] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *Proc. of ACMMM WS on Multimedia Infor-*

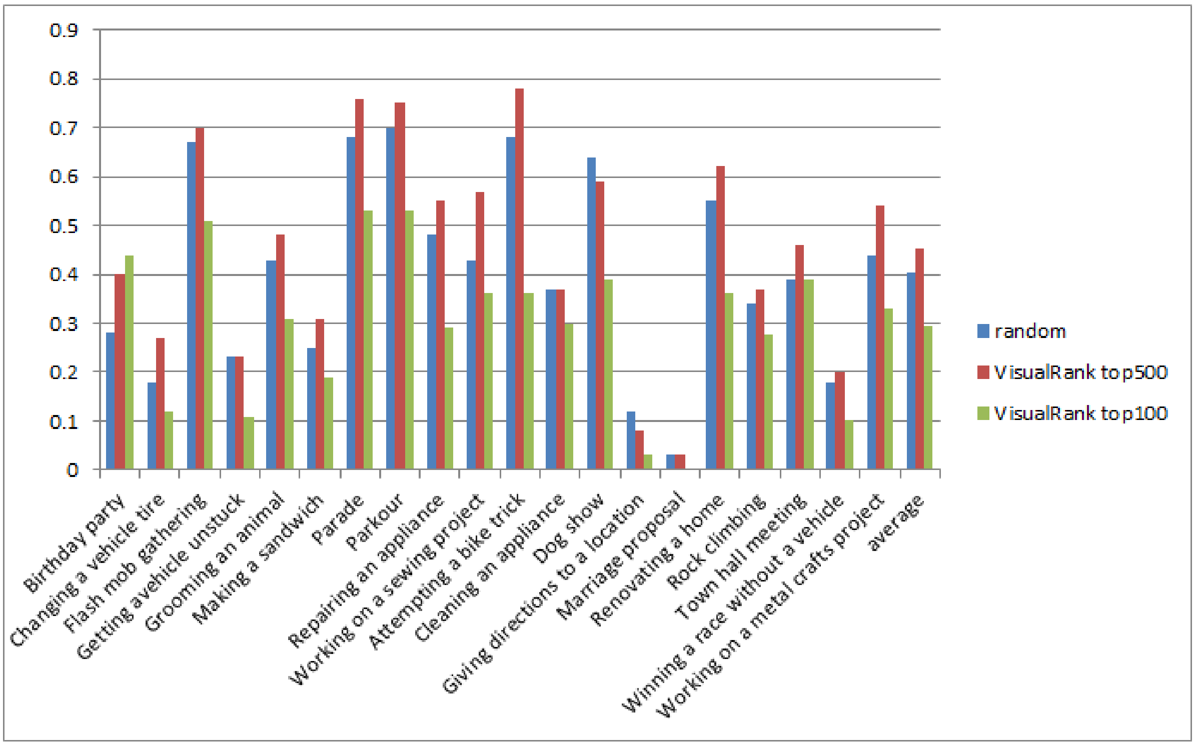

Figure 5. Verification of VisualRank. Random is baseline. VisualRank top500 is 100 shots at random from the top 500 ranking. VisualRank top100 is top 100 shots ranking.

mation Retrieval, pages 321–330, 2006.

[12] J. van Gemert, J. Geusebroek, C. Veenman, and A. Smeulders. Kernel codebooks for scene categorization. *Computer Vision–ECCV 2008*, pages 696–709, 2008.

[13] 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.

[14] 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.

[15] 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.