PicSOM Experiments in TRECVID 2015

Workshop notebook paper – Revision: 1.16

Satoru Ishikawa[+], Rao Muhammad Anwer[+], Markus Koskela[*], Jorma Laaksonen[+]
[+]Department of Computer Science
Aalto University School of Science
P.O. Box 15400, FI-00076 Aalto, Finland
firstname.lastname@aalto.fi

[*] Helsinki Institute for Information Technology HIIT
Department of Computer Science, University of Helsinki
P.O. Box 68, FI-00014 University of Helsinki, Finland
firstname.lastname@helsinki.fi

Abstract

The PicSOM Group's participation in TRECVID 2015 includes successful submissions in the semantic indexing (SIN) and localization (LOC) tasks. We also registered for the multimedia event detection (MED) and linking (LNK) tasks, but didn't run any experiments nor submit any results.

In semantic indexing (SIN), we participated in the main task only. We extended our last year's set of features with five new convolutional neural network (CNN) activation features based on three new different network architectures. We also implemented two simple feature weighting schemes to improve the precision of our detections. We submitted the following four SIN runs:

- 4 HEMULEN: Baseline run matching the best PicSOM SIN submission in TRECVID 2014
- 3 SNIFF: Similar to the 4 HEMULEN baseline run with five new convolutional neural network (CNN) activation features
- 2 MUMINMAMMAN: Similar to 3 SNIFF, but the weights of the individual detectors based on the different features were optimized jointly for all concepts
- 1 LILLAMY: Similar to 2 MUMINMAMMAN, but the optimization was class-specific

The run 1 LILLAMY obtained the highest MXIAP score of 0.2794. The run 4 HEMULEN which matched our best submission in TRECVID 2014 produced MXIAP score of 0.2445 which is considerably lower than the corresponding result 0.2880 last year. We thus assume that the evaluated SIN concepts this year were more difficult than those of last year.

We submitted four runs in the localization (LOC) task:

- 4: Baseline run with I-frame level temporal localizations from our SIN system and spatial detections from class-specific average locations from the training data
- 3: Deformable Part-Based Model (DPM) result with HOG features for both temporal and spatial localization
- 2: Temporal localizations from our SIN system and spatial localizations from DPM-based results
- 1: Temporal localizations combined from our SIN results and DPM results, spatial localizations from DPM

This was our first participation in the LOC task. Our results were not good, but still valuable for our own knowledge. It seems in the light of the F-scores, that our SIN method was better for I-frame level temporal localization, but when pixel-level spatial localization was considered, the SIN detection results didn't bring any advantage to the DPM-based results.

I. INTRODUCTION

In this notebook paper, we describe our experiments for the TRECVID 2015 evaluation [1]. We participated in two tasks, the semantic indexing (SIN, Section II) and localization (LOC, Section III). In addition, we registered and planned to participate in the multimedia event detection (MED) and linking (LNK) tasks, but didn't finally have the necessary resources. Overall conclusions are presented in Section IV.

II. SEMANTIC INDEXING

Our submissions to the semantic indexing (SIN) task are based on fusing several supervised detectors trained for each concept, based on different shot-level image features. The basic system architecture is the same as we have used in previous editions of TRECVID [2]. As the concept-wise ground-truth for the supervised detectors we used the annotations gathered by the organized collaborative annotation effort [3]. All our runs were submitted to the *main* task and are of *training type D*. We did not participate in the *no annotation* condition.

A. Features and classifiers

In addition to the main keyframes provided in the master shot reference, we extracted additional frames from training data shots longer than two seconds. In 2013 and 2014 we had used all I-frames provided in the test data set, but this time we reduced the number of I-frames used with a logarithmic curve which resulted in using a total of approximately 500,000 I-frames, which is roughly 30% of the available I-frames.

1) Old global, BoW, FV and VLAD features: We used the six image features from our previous TRECVID submissions: two global features (*Centrist* and *ScalableColor*)

and four BoV-type features (*SIFT*, *ColorSIFT*, *SIFTds*, and *ColorSIFTds*). Non-linear SVM classifiers were used with the exponential χ^2 kernel for the BoV features and the RBF kernel for the global features. See [4], [2] for details.

Similarly to TRECVID 2014 SIN, we extracted dense SIFT descriptors and encoded them using both Fisher vectors [5] and VLAD [6]. The codebooks were constructed using a 128-component GMM and k-means with 512 clusters, respectively. The corresponding classifiers were trained using linear SVMs.

2) CNN features: For the feature extraction in the keyframes we use CNNs pre-trained on the ImageNet database for object classification [7]. We used all the same CNN features as in TRECVID 2014, see [8]. In this year, we additionally used three new different CNN architectures, namely, 16-layer and 19-layer VGG [9] nets, and GoogLeNet [10]. In the case of VGG nets, we extract the activations of the network on the first fully connected (*fc6*, 4096-dimensional) layer with the given input images as the features. For GoogLeNet, we used similarly the output of the 5th Inception module, having the dimensionality of 1000.

Both a single center region or ten regions as suggested in [11] were extracted from all images. In the case of ten regions, both average and maximum pooling of the region-wise features were used. Furthermore, we augment these features with the reverse spatial pyramid pooling proposed in [12] with two scale levels. The second level consists of a 3×3 grid with overlaps and horizontal flipping, resulting in a total of 26 regions, on the scale of two. The activations of the regions are then pooled using average and maximum pooling. Finally, the activations of the different scales are concatenated. The resulting spatial pyramid features are therefore 8192- and 2000-dimensional for the VGG nets and GoogLeNet, respectively. See [13] for more details.

As classifiers for the CNN features, we utilized linear SVMs with homogeneous kernel maps [14] of order $d = 2$ to approximate the intersection kernel.

B. Classifier fusion

Classifier outcomes were in the first stage fused over the features for each frame with arithmetic mean. In the second fusion stage over the frames of each shot we used the maximum value. This can be written for concept class c as

$$r_{i,c} = \max_{j=1,\ldots,n_i} \frac{1}{N} \sum_{k=1}^{N} w_{k,c}\, r_{i,j,k,c} \;, \qquad (1)$$

where N is the number of used features, n_i is the number of frames in shot i and $r_{i,j,k,c}$ is the detection score for class c with feature k in frame j of shot i. The weighting term $w_{k,c}$ is an additional factor we hadn't used in our earlier experiments. Index c refers to the concept class and k to the feature in question. We studied both concept-specific and concept-independent selection of w. In the latter case the weight term simplifies to w_k.

Otherwise, the score values for the shots were obtained in the same manner as in TRECVID 2014 for each run as the maximum over the frame-wise scores resulting from the now weighted arithmetic mean over all features.

TABLE I
AN OVERVIEW OF THE SUBMITTED RUNS IN THE SEMANTIC INDEXING TASK. SEE TEXT FOR DETAILS.

	features		weight in (1)		
run id	TV14	5×CNN	w_k	$w_{k,c}$	MXIAP
4 HEMULEN	•				0.2445
3 SNIFF	•	•			0.2646
2 MUMINMAMMAN	•	•	•		0.2765
1 LILLAMY	•	•		•	0.2794

C. Mining hard negatives

A concept-wise, two-class classifier generally produced false positives on negative examples that were similar to the positive examples according to the used feature space. Therefore, to acquire more relevant negative examples, we performed n rounds of hard negative mining [15] and sampled 10 000 negative examples on each round. The final classifier for a given feature was obtained by fusing the classifier trained with the original, randomly sampled negatives and the n classifiers using mined relevant negatives.

In preliminary experiments, we observed that a single round of mining hard negatives already brought the greatest improvement. We therefore used the value $n = 1$ in the following experiments. This procedure is equal to that we used in TRECVID 2014 for the first time.

D. Submitted SIN runs

This section describes our submitted semantic indexing runs. Table I shows an overview, where the first two columns in the middle refer to the used features: the 38 features of our best-performing TRECVID 2014 [8] submission (TV14) and the five new CNN activation features. We used hard negative mining for all CNN features in both two sets. The next two columns indicate whether class-independent (w_k) or class-dependent ($w_{k,c}$) feature weighting scheme was used. The rightmost column lists the corresponding mean extended inferred average precision (MXIAP) [16] values.

Figure 1 shows how our runs were positioned in the MXIAP scores among all the 86 submitted runs. Figure 2 further illustrates the concept-wise XIAP results of our runs together with the maximum and median results of all the submissions. All our submissions were of training type D.

The run 4 HEMULEN is intended to match the best PicSOM submission in TRECVID 2014, denoted then as `1 Mårran`, i.e. to use the same features, classifiers, and method of fusion [8]. The run 3 SNIFF differs from the baseline only by its use of the five additional CNN-based features.

In the run 2 MUMINMAMMAN, the class-independent feature weighting was used. The feature-specific values were optimized by a gradient search where the tests sets of TRECVID 2013 and 2014 were used as a validation set. The run 1 LILLAMY is equal to the previous one, but now the optimization was run independently for all the 60 semantic concepts.

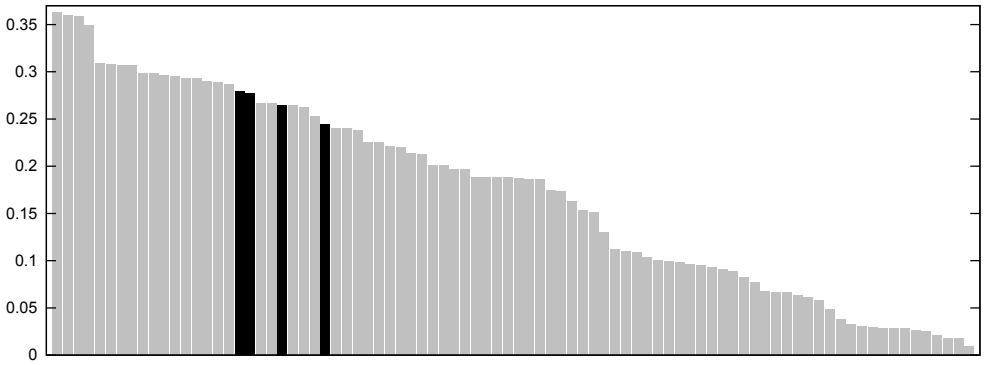

Fig. 1. Overview of the MXIAP all SIN runs submitted with our submissions in bolded bars.

According to our MXIAP results, the most notable increase of performance compared to our last year's submission comes from the new additional features. Both types of feature weight optimization further improve the result, while the class-dependent weighting scheme is slightly better than the class-independent one. As can be seen in Figure 2, the superiority of these two approaches is dependent on the concept in question.

III. Localization

The Deformable Part-Based Method (DPM) of Felzenszwalb et al. [17] has shown to provide excellent performance in human and generic object detection. An object is modeled as a collection of parts in the DPM model. The parts are constructed with a root model that can be seen as analogous to the standard HOG-based representation of Dalal and Triggs [18]. The DPM model employs latent SVM formulation for learning. The root filter, the part filters and the deformation cost of the configuration of all parts are concatenated to obtain a detection score for a window. The standard DPM framework employs HOG features computed over a dense grid of 8×8 non-overlapping cells. Several variants for DPM framework [19], [20], employing color and texture features have been proposed in literature. In our submitted runs, we used the standard DPM model with HOG features [17]. In the future we will experiment with a variant of HOG with color name features [19].

A. Submitted LOC runs

Table II summarizes or four submitted LOC runs. For run 4, we used just the I-frame-wise detections from our SIN subsystem described in the previous section. The spatial localizations were based on input-independent class-specific means of the object bounding boxes in the training data. For those classes that didn't have any training data we used the averages of the bounding boxes of the other classes. The required threshold parameter was optimized for each concept separately based on the distributions of the SIN scores for the positive and negative samples in the training set. These run 4 localizations were also used as a backup result for concepts *5 Anchorperson* and *31 Computers*, for which there were no training data available, in all other submitted runs.

TABLE II
Our LOC runs and their I-frame Fscores (IF), Recalls (IR) and Precisions (IP) and corresponding mean pixel measures Fscores (PF), Recalls (PR) and Precisions (PP)

runId	IF	IR	IP	PF	PR	PP
1	0.6321	**0.6275**	0.7741	0.3875	0.4530	0.4340
2	**0.6643**	0.6173	**0.8034**	0.3868	**0.4719**	0.4289
3	0.5232	0.5351	0.7445	**0.3944**	0.4503	**0.4344**
4	**0.6643**	0.6173	**0.8034**	0.2670	0.2934	0.3771

Run 3 is based on the use of the Deformable Part-Based Model (DPM) score with HOG features for both temporal and spatial localization. A common threshold value was used for all concepts.

Run 2 was formed so that SIN scores were first used to select the I-frames that were the most likely to contain the concept in question. Then the spatial region proposal by the DPM model was used to provide the bounding box.

In run 1, the selection of the I-frames was implemented by summing our SIN and DPM scores for the temporal part of the localization and then the spatial part was carried out by using the bounding box proposed by the DPM model.

Our LOC results are overall quite bad, but still valuable for our own knowledge. The DPM model seems to be working because the mean pixel measures are clearly better with it than with the average proposals. However, on the I-frame level our SIN result is better than the DPM result, but their combination does not seem to bring any improvement.

IV. Conclusions

Concerning the SIN task results, it seems that other groups have improved their methods since TRECVID 2014 more than what we have been able to do. Looking at our own SIN results only, we are satisfied with the progress we have made. Both the additional CNN features and the two feature weight optimization schemes gave clear improvements in our MXIAP scores.

Concerning the LOC task, this was our first time participation and we didn't expect much of the outcome yet. Still, there

is some promise in the results, even though the combination of the image-level detection value and the spatial region proposal did not work quite as well as we expected. In the forthcoming years we plan to investigate more efforts in this task if it will be continued.

ACKNOWLEDGMENTS

This work has been funded by the grants 255745 and 251170 of the Academy of Finland and *Data to Intelligence (D2I)* SHOK projects. The calculations were performed using computer resources within the Aalto University School of Science "Science-IT" project.

REFERENCES

[1] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Wessel Kraaij, Alan F. Smeaton, Georges Quénot, and Roeland Ordelman. Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2015*. NIST, USA, 2015.

[2] Satoru Ishikawa, Markus Koskela, Mats Sjöberg, Jorma Laaksonen, Erkki Oja, Ehsan Amid, Kalle Palomäki, Annamaria Mesaros, and Mikko Kurimo. PicSOM experiments in TRECVID 2013. In *Proceedings of the TRECVID 2013 Workshop*, Gaithersburg, MD, USA, November 2013.

[3] Stéphane Ayache and Georges Quénot. Video corpus annotation using active learning. In *Proceedings of 30th European Conference on Information Retrieval (ECIR'08)*, pages 187–198, Glasgow, UK, March-April 2008.

[4] Markus Koskela, Mats Sjöberg, Ville Viitaniemi, Jorma Laaksonen, and Philip Prentis. PicSOM experiments in TRECVID 2007. In *Proceedings of the TRECVID 2007 Workshop*, Gaithersburg, MD, USA, November 2007.

[5] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8, June 2007.

[6] H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[8] Satoru Ishikawa, Markus Koskela, Mats Sjöberg, Rao Muhammad Anwer, Jorma Laaksonen, and Erkki Oja. PicSOM experiments in TRECVID 2014. In *Proceedings of the TRECVID 2014 Workshop*, Orlando, FL, USA, November 2014.

[9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, abs/1409.1556, 2014.

[10] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv*, abs/1409.4842, 2014.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[12] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. arXiv.org:1403.1840, March 2014.

[13] Markus Koskela and Jorma Laaksonen. Convolutional network features for scene recognition. In *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, Florida, November 2014.

[14] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2010)*, 2010.

[15] Xirong Li, Cees G. M. Snoek, Marcel Worring, Dennis C. Koelma, and Arnold W. M. Smeulders. Bootstrapping visual categorization with relevant negatives. *IEEE Transactions on Multimedia*, 15(4):933–945, June 2013.

[16] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08)*, pages 603–610, 2008.

[17] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[18] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *In* Proceedings of Computer Vision and Pattern Recognition, 2005.

[19] Fahad Shahbaz Khan, Rao Muhammad Anwer, Joost van de Weijer, Andrw D. Bagdanov, and Maria Vanrell. Color attributess for object detection. In *In Proceedings* of Computer Vision and Pattern Recognition, 2012.

[20] J. Zhang, K.Huang, Y. Yu, and T. Tan. Boosted local structured hog-lbp for object localization. In *In* Proceedings of Computer Vision and Pattern Recognition, 2010.

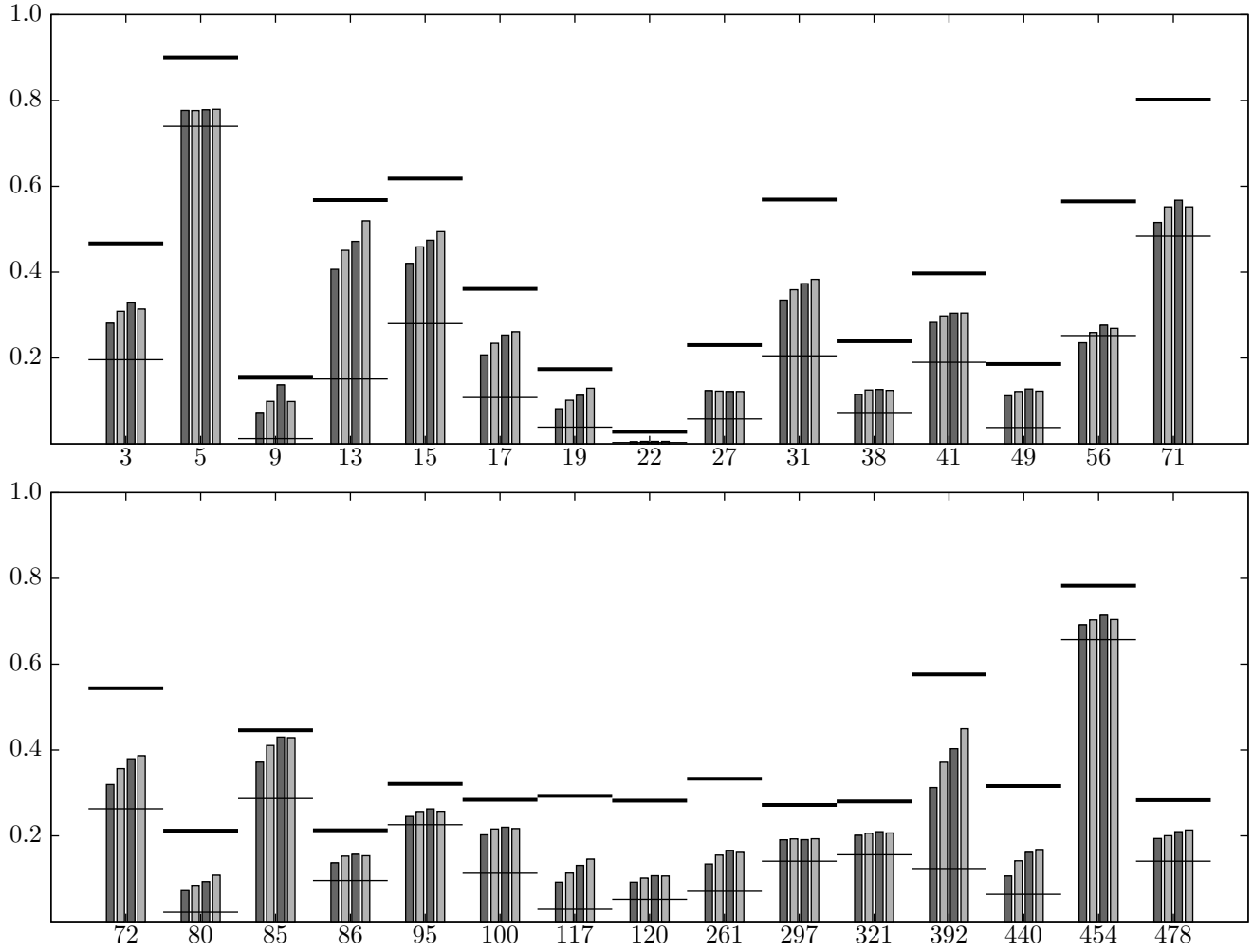

Fig. 2. The concept-wise XIAP results of our submitted runs for each evaluated concept in the semantic indexing task. The order of the runs is as in Table I, i.e. 4 HEMULEN, ..., 1 LILLAMY. The median and maximum values over all submissions are illustrated as horizontal lines.