PicSOM Experiments in TRECVID 2013

Workshop draft – Revision: 1.29

Satoru Ishikawa, Markus Koskela, Mats Sjöberg, Jorma Laaksonen, Erkki Oja
Department of Information and Computer Science
Aalto University School of Science
P.O. Box 15400, FI-00076 Aalto, Finland

Ehsan Amid, Kalle Palomäki, Annamaria Mesaros, Mikko Kurimo
Department of Signal Processing and Acoustics
Aalto University School of Electrical Engineering
P.O. Box 13000, FI-00076 Aalto, Finland

firstname.lastname@aalto.fi

Abstract

Our experiments in TRECVID 2013 include participation in the Semantic Indexing (SIN), Multimedia Event Detection (MED), and Multimedia Event Recounting (MER) tasks.

In Semantic Indexing, we participated in the main and paired tasks. We implemented linear and non-linear SVM-based classifiers on six visual features extracted from the main keyframes and also additional frames from longer shots. We used homogeneous kernel map approximations for the linear classifiers, which narrow the performance gap to the non-linear SVMs. We submitted the following four runs to the main task:

- `PicSOM_M_1`: 4 *SIFT* features with exponential χ^2 kernels + *Centrist* and *ScalableColor* RBF SVMs fused over features with arithmetic mean and over frames with the maximum operator
- `PicSOM_M_2`: like submission `PicSOM_M_1`, but with additional 4 SIFT homogeneous kernel map detectors
- `PicSOM_M_3`: 4 *SIFT* features with homogeneous kernel map + *Centrist* and *ScalableColor* RBF SVMs fused over features with arithmetic mean and over frames with the maximum operator
- `PicSOM_M_4`: like submission `PicSOM_M_3` but the detectors were trained with training data from last year's evaluation

The run `PicSOM_M_1` obtained the highest MXIAP score of 0.2055.

We submitted to the paired task the following two runs:

- `PicSOM_P_5`: Baseline run with direct multiplication ("AND") of shot-level scores of each member concept
- `PicSOM_P_6`: Another run with 80% multiplication ("AND") + 20% sum ("OR") of shot-level scores of each member concept

The run `PicSOM_P_6` obtained the highest MXIAP score of 0.1126.

In the Multimedia Event Detection task, we participated in the PROGAll search task. We submitted VisualSys as our FullSys for the pre-specified event kit, and VisualSys, AudioSys, and real FullSys (weighted fusion of visual system and audio results) for the ad-hoc event kit. We used 10Ex and 100Ex conditions of both pre-specified and ad-hoc event kits as the sole training data for the detectors in the runs:

- `PicSOM_FullSys_PROGAll_PS_100Ex_2`: full/visual system (pre-specified event kit 100Ex)
- `PicSOM_FullSys_PROGAll_PS_10Ex_2`: full/visual system (pre-specified event kit 10Ex)
- `PicSOM_VisualSys_PROGAll_AH_100Ex_1`: visual system (ad-hoc event kit 100Ex)
- `PicSOM_AudioSys_PROGAll_AH_100Ex_1`: audio system (ad-hoc event kit 100Ex)
- `PicSOM_FullSys_PROGAll_AH_100Ex_1`: weighted fusion of visual system and audio system (ad-hoc event kit 100Ex)
- `PicSOM_VisualSys_PROGAll_AH_10Ex_1`: visual system (ad-hoc event kit 10Ex)
- `PicSOM_AudioSys_PROGAll_AH_10Ex_1`: audio system (ad-hoc event kit 10Ex)
- `PicSOM_FullSys_PROGAll_AH_10Ex_1`: weighted fusion of visual system and audio system (ad-hoc event kit 10Ex)

The mean average precision (MAP) score of `PicSOM_FullSys_PROGAll_PS_100Ex_2` was 6.4 and that of `PicSOM_FullSys_PROGAll_PS_10Ex_2` was 3.2. The run `PicSOM_VisualSys_PROGAll_AH_100Ex_1` failed due to a programming error, also leading to the failure of `PicSOM_FullSys_PROGAll_AH_100Ex_1`. Consequently the obtained results were not on the level we expected based on the results of the pre-specified event kit runs.

In Multimedia Event Recounting we participated with the run:

- `PicSOM_FullSys_PROGAll_PS_100Ex_2`: visual system (pre-specified event kit 100Ex)

The recountings were solely based on the visual detections where those parts of the test videos were included in the recountings where the detection score exceeded the detection threshold value used in the MED task. Our initial MER evaluation scores were surprisingly good, being 36.39 % in recounting review time, 64.34 % in accuracy and 1.96 in precision of the observation text.

I. Introduction

In this notebook paper, we describe our experiments for the TRECVID 2013 evaluation [1]. We participated in the Semantic Indexing (SIN, Section II), Multimedia Event Detection (MED, Section III) and Multimedia Event Recounting (MER, Section IV) tasks. Overall conclusions are presented in Section V.

II. Semantic indexing

Our system for the Semantic Indexing (SIN) task is based on fusing several supervised detectors trained for each concept, based on different shot-level image and video features. The basic system architecture is the same as we have used in previous editions of TRECVID [2], [3], [4]. The accuracy of linear classifiers is improved by employing explicit kernel maps. As the concept-wise ground-truth for the supervised detectors we used the annotations gathered by the organized collaborative annotation effort [5]. All our runs were submitted to the full task and are of type A. This year we also participated in the paired concepts task.

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 and used all I-frames provided in the test data set. We extracted six image features from all extracted frames, four of them BoV-type (*SIFT*, *ColorSIFT*, *SIFTds*, and *ColorSIFTds*) and two others (*Centrist* and *ScalableColor*). See [6], [3] for details.

For the non-linear SVM classifiers we use the exponential χ^2 kernel for the BoV features and the RBF kernel for *Centrist* and *ScalableColor*. For the linear BoV classifiers, we utilize homogeneous kernel maps [7] of order $d = 3$ 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 as

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

where N is the number of used features, n_i is the number of frames in shot i and $r_{i,j,k}$ is the detection score for feature k in frame j of shot i.

C. Concept pairs

We participated in the concept pair task for the first time. Based on naïve assumptions, we used as our baseline result the product of the detection scores of the single-concept detectors as the detection score for the paired concept. As a slight generalization of this "logical AND" operation, we made another run, where we used 80% of "logical AND" and 20% of "logical OR".

These entities can be formalized as

$$r(a \wedge b) = r(a)\, r(b) \qquad (2)$$

$$r(a \vee b) = r(a) + r(b) - r(a)\, r(b) \qquad (3)$$

$$r_\alpha(a, b) = \alpha\, r(a \wedge b) + (1 - \alpha)\, r(a \vee b) \qquad (4)$$

where detections scores r are assumed to be $0 \leq r \leq 1$ and α is the weighting coefficient whose value was $\alpha = 1$ in our first run and $\alpha = 0.8$ in the second.

D. Submitted runs

This section describes our submitted main and paired Semantic Indexing runs. Table I shows an overview of the main task submissions, where the three columns in the middle refer to the used classifiers: non-linear SVMs for BoV features, linear classifiers for BoV features, and the RBF-kernel SVMs for the non-BoV features. The score values for the shots were obtained in the same manner for each run as the maximum over the frame-wise scores resulting from the arithmetic mean over all features. The rightmost column lists the corresponding mean extended inferred average precision (MXIAP) [8] values. Figure 1 illustrates the concept-wise XIAP results of the runs.

Comparison of the MXIAP values for runs `PicSOM_M_1`, and `PicSOM_M_3` shows the expected result: the non-linear χ^2 kernel SVM performs better than the linear homogeneous kernel map approximating the intersection kernel. `PicSOM_M_2` additionally hints that fusing the non-linear and linear detection scores from the same features does not bring any further improvement over the non-linear results.

The difference between the `PicSOM_M_3` and `PicSOM_M_4` runs was that in the latter we used concept detectors that had been trained with TRECVID 2012 SIN training data. The better performance of `PicSOM_M_3` shows that the increased amount of annotated training data has been beneficial in the training of the detectors.

Table II shows an overview of the paired task submissions and Figure 2 illustrates the concept-pair-wise XIAP results of these runs. For both runs, we used the results of our best-performing single-concept run `PicSOM_M_1` as the individual detection values.

As can be seen, `PicSOM_P_6` that combined a small portion of "logical OR" to the "logical AND" gave a slightly better MXIAP value than the baseline, but the difference is not significant. We expect to obtain better pair-wise concept detection performance in our ongoing experiments by implementing the pair fusion at least partially already on the frame level as now the fusion was solely done on the shot level.

TABLE I
An overview of the submitted runs in the main Semantic Indexing task. See text for details.

#	run id	classifiers			MXIAP
		non-linear	linear	non-BoV	
1	`PicSOM_M_1`	•		•	0.2055
2	`PicSOM_M_2`	•	•	•	0.2012
3	`PicSOM_M_3`		•	•	0.1829
4	`PicSOM_M_4`		•	•	0.1524

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

#	run id	baseline	α in (4)	MXIAP
1	PicSOM_P_5	•	1.0	0.1111
2	PicSOM_P_6		0.8	0.1126

III. MULTIMEDIA EVENT DETECTION

In Multimedia Event Detection (MED) task, we participated in the PROGAll category and used both pre-specified and ad-hoc event kits with the 100Ex and 10Ex conditions. Our detection system consisted of two sub-systems: the visual system and the audio system, but unfortunately the latter one was not yet ready at the time when the pre-specified event kit detections were submitted.

The detection thresholds required for all submissions were selected based on the naïve heuristic assumption that 1/100 of the videos were relevant. The threshold was thus selected so that one percent of the top-scoring test set videos had the detection score over the selected threshold. (We are not fully sure if this is acceptable use of the test set data as the thresholds are dependent on the detection values on the test set.)

A. VisualSys

Our visual system for the MED task uses only visual detectors trained with positive event kit videos as positive examples and background training videos as negative examples of the event. We extracted one keyframe per second from the event kit videos and approximately six keyframes per ten seconds from the PROGAll video set. From these frames, we calculated the same four *SIFT*, *Centrist* and *ScalableColor* features as in the SIN task and described in [6], [3].

As detectors for the visual content we trained homogeneous kernel map approximations for the linear classifiers of the *SIFT* features and RBF kernel SVMs for *Centrist* and *ScalableColor*. Feature-wise detections were first fused with arithmetic mean over all features and then fused with maximum value over the frames to get video-level detection scores. The setup is thus exactly the same as for the PicSOM_M_3 run in the SIN task shown as run #3 in Table I.

B. AudioSys

Our audio system includes an initial step of music detection to exclude all the video files which contain music, since in that case, the audio content does not provide any important clues to recognize the underlying event. We performed this stage by detecting the stable peaks of the short-time spectrogram of the signal [9]. We first modeled the signal in a short-time (30 ms) overlapping window by a high-order (600) autoregressive filter obtained using linear prediction with sampling frequency of 44100 Hz and then, calculated the spectrogram using the impulse response of the filter. This resulted in a smoother version of the spectrogram. Then, we summed up the stable spectral peak intensities over 0.5 s time frames and compared it with a predefined threshold. All the frames with a total peak intensity value above the threshold was marked as music. Finally, all the files containing a music ratio above a certain value were discarded.

Next, we extracted an extensive set of well-known features e.g. MFCC, zero crossing, etc. over short-time (30 ms) overlapping windows and then, calculated their mean and variance over a 0.5 s frame and considered it as the corresponding feature vector. Then, we trained a stacked denoising autoencoder (SDAE) [10] to obtain a higher-order presentation of the initial features as well as reducing the dimensionality. The initial feature extraction step helps the system focus on the desired aspects of the signal. Applying the SDAE, we obtained the final feature representation of the signal over every 0.5 s frame.

Since the videos are of different lengths, the corresponding sets of feature vectors have different cardinalities. In order to attain a comparable feature representation for the videos, we quantized the feature vectors using k-means++ algorithm [11] and then, formed the histogram of the quantization cluster indices for each video. Additionally, we applied a TF-IDF [12] scaling on the frequency values to reduce the effect of redundant bins and considered the normalized weighted histograms as the final presentation of the videos.

For the classification step, we first defined a weighted Jensen-Shannon [13] distance measure over the histograms as the kernel function. The weights were obtained based on the mean and variance of the histograms of a particular class and the background videos. Then, we trained a set of 100 support vector regression (SVR) models for each class and averaged over the outputs to obtain the probability values of the events for each video.

C. FullSys

The VisualSys and AudioSys results were fused together with weighted arithmetic average inverse rank. The weight of VisualSys was heuristically set to $w_v = 0.95$ and that of AudioSys to $w_a = 0.05$ to reflect their expected relative performances. This expectation followed from experiments with the event kits available in the MED Test collection. (We are not fully sure if this is acceptable use of the MED Test collection.)

The weighted inverse rank r_i used as the final score value for video i can be obtained as

$$r_i = \frac{w_v}{v_i} + \frac{w_a}{a_i} = \frac{0.95}{v_i} + \frac{0.05}{a_i} \,, \tag{5}$$

where $v_i \in \{1, 2, \ldots\}$ is the rank of video i in their ordering based on the VisualSys detections and $a_i \in \{1, 2, \ldots\}$ is that based the AudioSys detections.

D. Submitted runs

Our submitted runs in the Multimedia Event Detection task are summarized in Table III together with their mean average precision (MAP) scores. Submissions #1 and #2 were actually made twice, the second time as VisualSys because we then did not yet have the AudioSys available.

TABLE III
AN OVERVIEW OF THE SUBMITTED RUNS IN THE MULTIMEDIA EVENT DETECTION TASK. SEE TEXT FOR DETAILS.

#	run id	MAP
1	PicSOM_FullSys_PROGAll_PS_100Ex_2	6.4
2	PicSOM_FullSys_PROGAll_PS_10Ex_2	3.2
3	PicSOM_VisualSys_PROGAll_AH_100Ex_1	0.6
4	PicSOM_AudioSys_PROGAll_AH_100Ex_1	0.1
5	PicSOM_FullSys_PROGAll_AH_100Ex_1	0.6
6	PicSOM_VisualSys_PROGAll_AH_10Ex_1	2.2
7	PicSOM_AudioSys_PROGAll_AH_10Ex_1	0.2
8	PicSOM_FullSys_PROGAll_AH_10Ex_1	2.1

The ad-hoc VisualSys 100Ex experiment resulting in the submission shown in Table III as run #3 unfortunately failed due to a programming error, and consequently also the corresponding FullSys run #5 went wrong.

The ad-hoc VisualSys result with the 10Ex training condition is somewhat worse than the corresponding pre-specified result. This does not seem to have happened with other groups' submissions, so we might have had some technical problems with that experiment as well, unless it is purely coincidental. The AudioSys was still performing quite poorly and requires additional effort for the forthcoming years' evaluations.

Overall, our pre-specified event kit MAP result with the 100Ex training condition is clearly worse than what we expected based on the corresponding experiments with the MED Test collection. We will study the reason for this behavior in more detail.

IV. MULTIMEDIA EVENT RECOUNTING

Our MER results are based on the FullSys 100Ex MED submission where only the VisualSys detections were used. The initial MER evaluation measures of our run are shown in Table IV.

Those parts of the test videos where the detection score exceeded the detection threshold value used in the MED task were included in the recountings as positive evidence of the existence of the event. These evidence parts were selected on the frame level and then expanded in time two seconds in both directions to always get at least four seconds long video snippets. Overlapping and adjacent snippets were finally concatenated to reduce the total number of snippets extracted.

We always used the full video frame as the bounding box, importance value 1.0 and description *"visual content matches examples"*. The confidence value was obtained from the maximum frame-wise score value inside the snippet and the snippets were ordered based on the confidence value so that the most confident snippets would be presented first.

TABLE IV
AN OVERVIEW OF RESULTS IN MER SUBMISSION. SEE TEXT FOR DETAILS.

characteristic	value	rank
Percent Recounting Review Time	36.39 %	1st
Accuracy	64.34 %	3rd
Precision of the observation text	1.96 = 'Fair'	2nd

We calculated the inverse of our MER submission's average inverse ranks for the three individual characteristics shown in Table IV. This figure equals to 1.64 and is higher than that of any other submission. For comparison, this combined score was 1.89 for Sesame, 2.26 for SRIAURORA and 2.77 for BBNVISER. We consider this outcome more as an artefact or an indication of the immaturity of the evaluation measures than as an evidence of a good performance of our rather primitive MER system.

V. CONCLUSIONS

Concerning the SIN task results, it seems that our position in the result table of all participating groups is slightly worse than in the previous years. This could be expected as we did not have any methodological improvements in our system compared to that of the last year.

The first time participation in the MED task was quite demanding and we were not able to implement all the methods we had planned. Both the visual and audio systems have room for improvement. Also an unfortunate programming error hindered our ad-hoc event kit results. In the forthcoming years we plan to participate also in the 0Ex training condition, and implement at least the OCRSys subsystem if not also the ASRSys subsystem.

Our relatively good performance in the MER task was a delightful surprise when compared to our corresponding less than mediocre MED submission result. We assume that the definitions of the performance measures used in the MER task still require some effort.

ACKNOWLEDGMENTS

This work has been funded by the grants 255745, 251170 and 136209 of the Academy of Finland, TFMC 11863 of EIT ICT Labs, and *Next Media* and *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, Greg Sanders, Barbara Shaw, Wessel Kraaij, Alan F. Smeaton, and Georges Queenot. Trecvid 2012 – an overview of the goals, tasks, data, evaluation mechani sms and metrics. In *Proceedings of TRECVID 2012*. NIST, USA, 2012.

[2] Mats Sjöberg, Markus Koskela, Milen Chechev, and Jorma Laaksonen. PicSOM experiments in TRECVID 2010. In *Proceedings of the TRECVID 2010 Workshop*, Gaithersburg, MD, USA, November 2010.

[3] Mats Sjöberg, Satoru Ishikawa, Markus Koskela, Jorma Laaksonen, and Erkki Oja. PicSOM experiments in TRECVID 2011. In *Proceedings of the TRECVID 2011 Workshop*, Gaithersburg, MD, USA, December 2011.

[4] Mats Sjöberg, Satoru Ishikawa, Markus Koskela, Jorma Laaksonen, and Erkki Oja. PicSOM experiments in TRECVID 2012. In *Proceedings of the TRECVID 2012 Workshop*, Gaithersburg, MD, USA, November 2012.

[5] 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.

[6] 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.

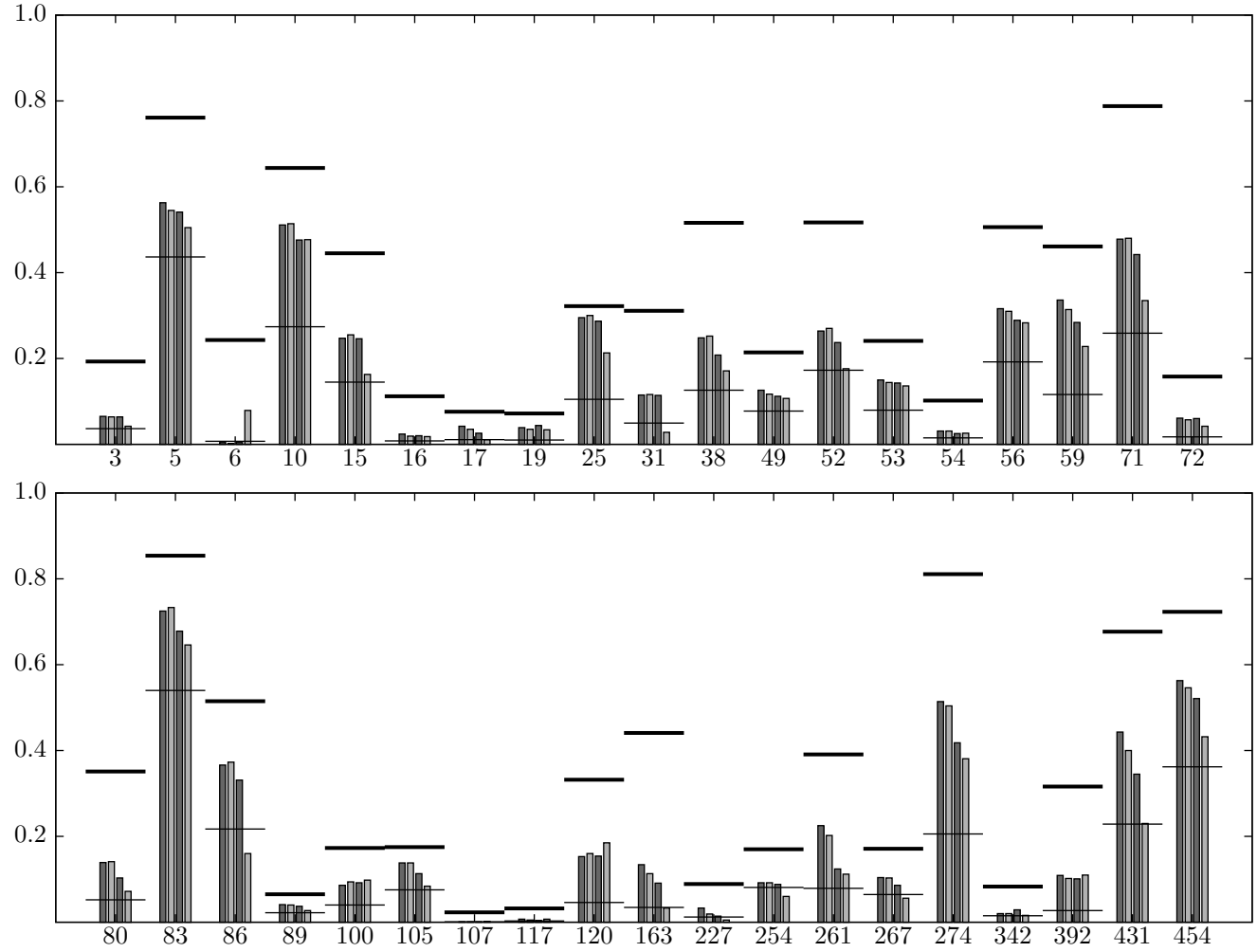

Fig. 1. The concept-wise XIAP results of our submitted runs for each evaluated concept in the main Semantic Indexing task. The order of the runs is as in Table I, i.e. `PicSOM_M_1`, ..., `PicSOM_M_4`. The median and maximum values over all submissions are illustrated as horizontal lines.

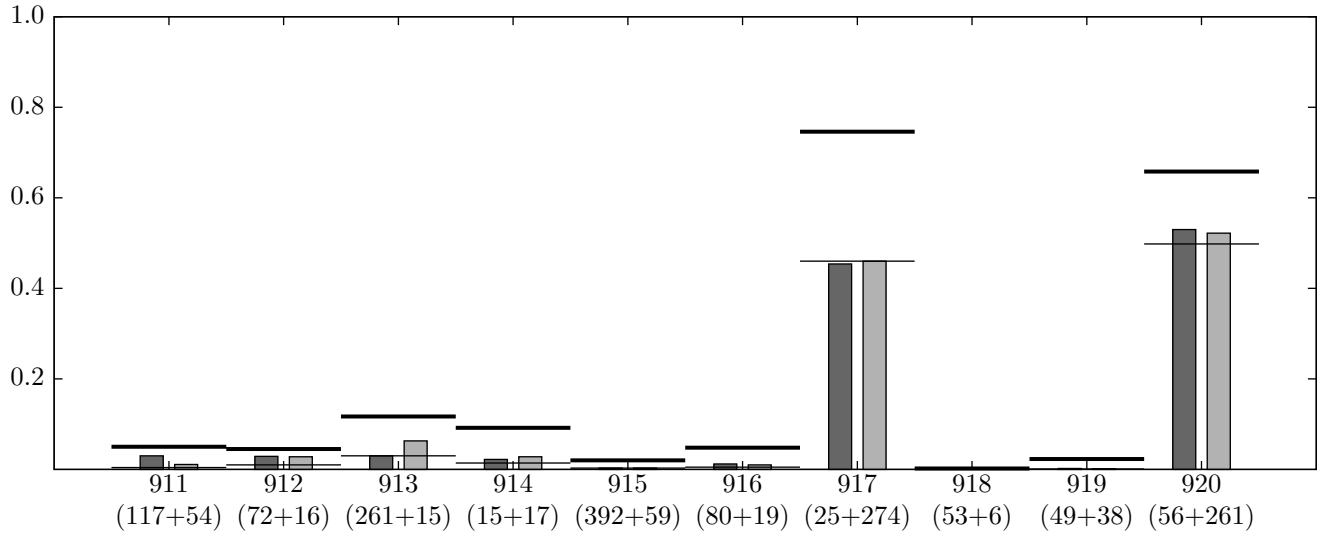

Fig. 2. The concept-wise XIAP results of our submitted runs for each evaluated concept in the concept pairs task of Semantic Indexing (left: `PicSOM_P_5`, right: `PicSOM_P_6`). The median and maximum values over all submissions are illustrated as horizontal lines.

[7] 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.

[8] 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.

[9] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura. Video handling with music and speech detection. *MultiMedia, IEEE*, 5(3):17–25, 1998.

[10] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, December 2010.

[11] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[12] Thomas Roelleke and Jun Wang. Tf-idf uncovered: a study of theories and probabilities. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 435–442, New York, NY, USA, 2008. ACM.

[13] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86, 1951.