PicSOM Experiments in TRECVID 2014 Semantic Indexing Task

Jorma Laaksonen

Aalto University School of Science
Department of Information and Computer Science
Espoo, Finland

10 Nov 2014

Contents

overview

related works

training and detection

details

conclusions

demo

The team

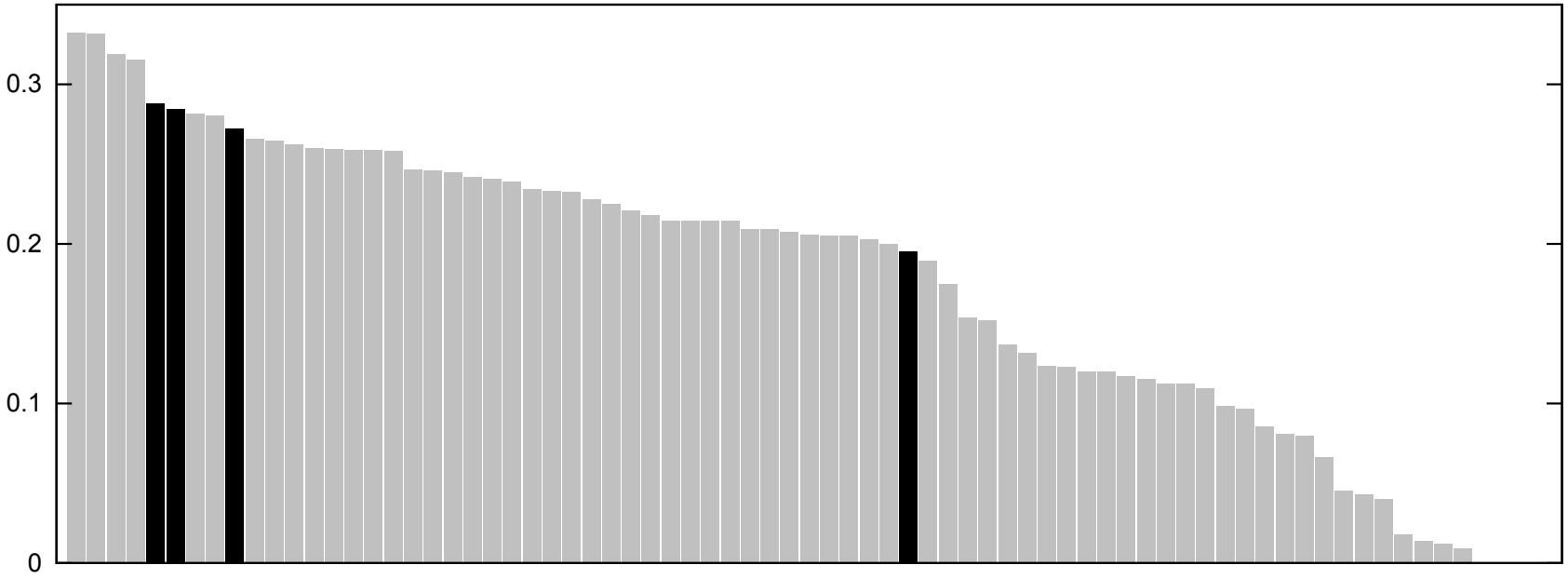
@ Aalto University School of Science, Espoo, Finland

- Satoru Ishikawa, doctoral student
- Markus Koskela, post.doc., left the group in summer 2014
- Mats Sjöberg, PhD to be, left the group in summer 2014
- Rao Muhammad Anwer, post.doc, started in winter 2014
- Jorma Laaksonen, teaching research scientist
- Erkki Oja, professor retiring in winter 2015

Overview

the big picture

- Four submissions in SIN Main task:
  - PicSOM_4 Muminpappan A 0.2000 (0.1951)
  - PicSOM_3 Hattifnattar D 0.2900 (0.2843)
  - PicSOM_2 Snusmumriken D 0.2777 (0.2722)
  - PicSOM_1 Mårran D 0.2936 (0.2880)



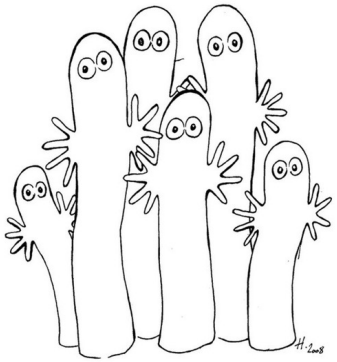
Some characters from Moomin Valley

Naming of our runs

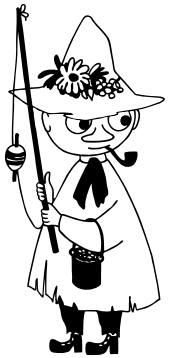
- Tove Jansson
  - Finnish Swede novelist, painter and comic strip author
  - creator of the Moomins
  - 9 Aug 1914 – 27 Jun 2001

Tove

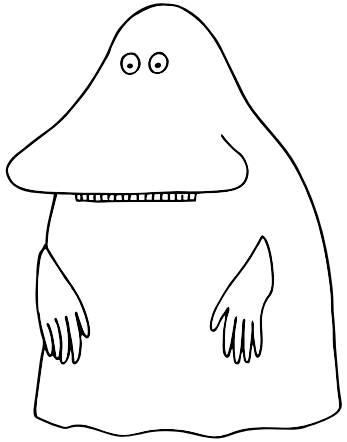
Muminpappan



Hattifnattar



Snusmumriken



Mårran

Contents

overview

related works

training and detection

details

conclusions

demo

Linear Homogeneous Kernel Map SVM classifiers

old works

- Mats Sjöberg, Markus Koskela, Satoru Ishikawa, and Jorma Laaksonen. *Real-time large-scale visual concept detection with linear classifiers.* In Proceedings of 21st International Conference on Pattern Recognition, Tsukuba, Japan, November 2012.
- Mats Sjöberg, Markus Koskela, Satoru Ishikawa, and Jorma Laaksonen. *Large-scale visual concept detection with explicit kernel maps and power mean SVM.* In Proceedings of ACM International Conference on Multimedia Retrieval (ICMR2013), pages 239–246, Dallas, Texas, USA, April 2013. ACM.

Fusion of CNN activation features

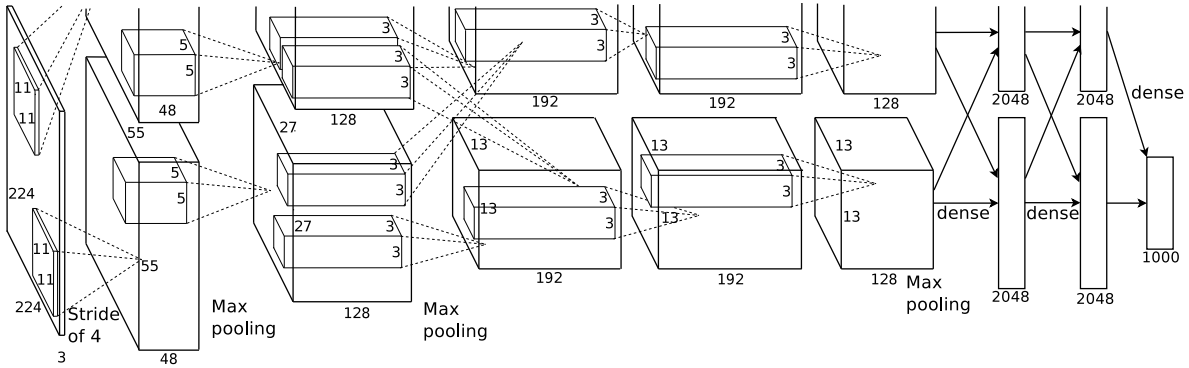
recent work

- Markus Koskela and Jorma Laaksonen. *Convolutional network features for scene recognition.* In Proceedings of the 22nd International Conference on Multimedia, Orlando, Florida, November 2014:
  - state-of-the-art results in scene recognition with four benchmarks:
    - scenes-15 0.921
    - uiuc-sports 0.948
    - indoor-67 0.701
    - sun397 0.547
  - four different CNN features as combinations of
    - 2 different training sets: ILSVRC 2010 and 2012
    - 2 different CNN architectures: Krizhevsky and Zeiler
  - full image features vs. spatial pyramid features
  - late geometric mean fusion

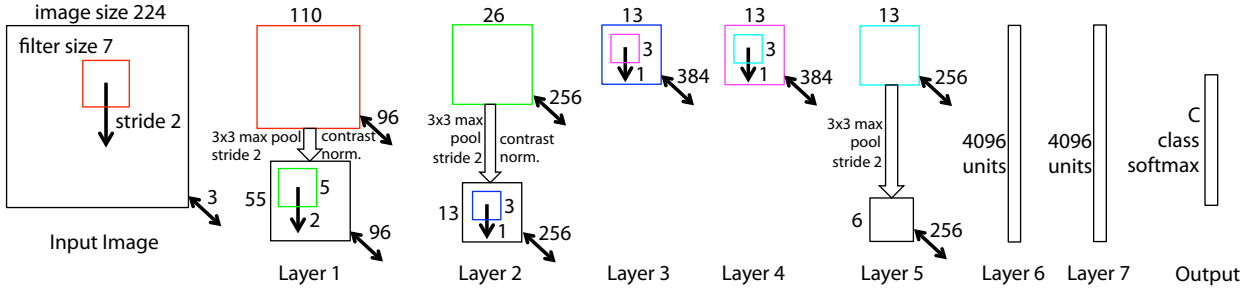
Fusion of CNN activation features

CNN network models

- Caffe library implementations of:
  - Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. *ImageNet classification with deep convolutional neural networks.* In NIPS, 2012:



  - Matthew Zeiler and Rob Fergus. *Visualizing and understanding convolutional networks.* arXiv:1311.2901, November 2013:



Contents

overview

related works

training and detection

details

conclusions

demo

Training procedure

same as before

- 6 old features: used old detectors trained in 2013
  - libsvm
  - rbf / exp χ^2
- 30 new features: trained detectors using same images
  - liblinear
  - homogeneous kernel map, order 0 / 1 / 2
  - histogram intersection
  - hard negative mining

purpose	*dataset*	*videos*	*shots*	*images*	*comment*
development	IACC.1.*	28003	546530	546530	keyframes
validation	IACC.2.A	2418	112677	1679245	i-frames
evaluation	IACC.2.B	2373	106913	1573832	i-frames

Detection procedure

same as before

- detections scores calculated for each i-frame
- feature-wise scores fused in each i-frame
  - arithmetic mean
  - no concept-dependent feature selection
  - no concept- or feature-dependent weighting
- i-frame-wise scores fused in each shot
  - maximum value with no within-shot weighting
  - no between-shot / within-video processing
  - no between-concept processing

Contents

overview

related works

training and detection

details

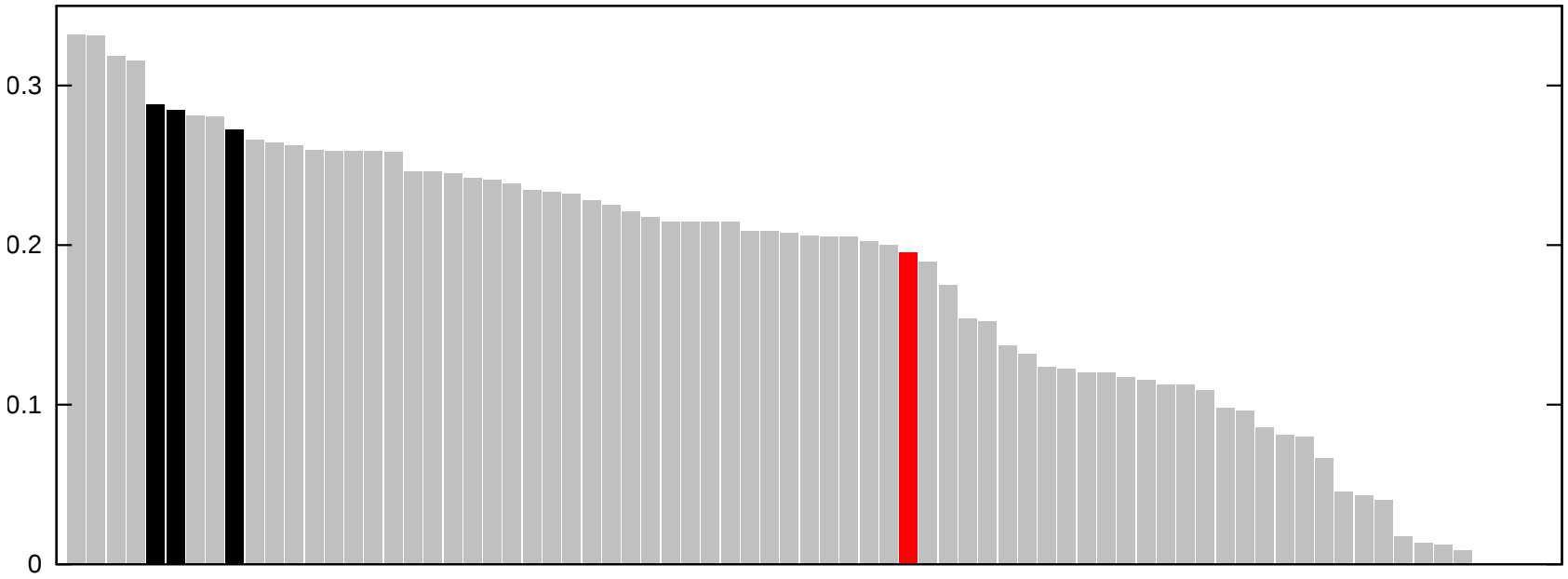
conclusions

demo

Run 4 Muminpappan, MXIAP = 0.2000 (0.1951)

our best TRECVID 2013 result

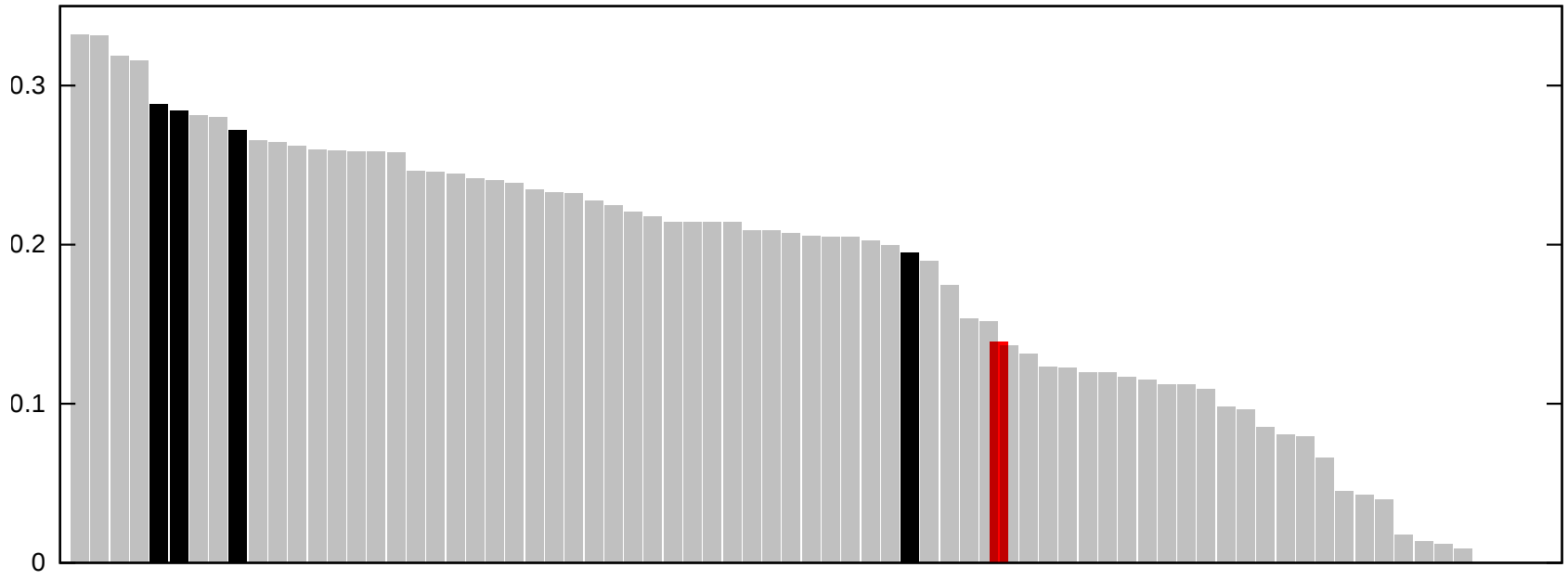
feature	*dim.*	*classifier*	*MXIAP*
ColorSIFTds-1x1-2x2	5000	SVM exp χ^2	0.1609
SIFTds-1x1-2x2	5000	SVM exp χ^2	0.1537
SIFT-1x1-2x2	5000	SVM exp χ^2	0.1368
ColorSIFT-1x1-2x2	5000	SVM exp χ^2	0.1330
OCVCentrist	1302	SVM RBF	0.1173
scalablecolor	256	SVM RBF	0.0437
fusion			0.2000



Fisher vector, VLAD, LBP and SIFT features

experimented with in fall 2013

feature	*dim.*	*classifier*	*MXIAP*
ColorSIFTds-1x1-2x2-1x3	8000	lin hkm1 int	0.1259
ColorSIFT-1x1-2x2-1x3	8000	lin hkm1 int	0.0989
OCVMlhmsLbp-10-1234	10240	lin hkm1 int	0.0915
OCVMlhmsLbp-10-12	5120	lin hkm1 int	0.0762
vlfeat-dsift-128-gmm-128-FV	32768	lin int	0.1251
vlfeat-dsift-128-kmeans-512-VLAD	65536	lin int	0.1392



CNN activation features

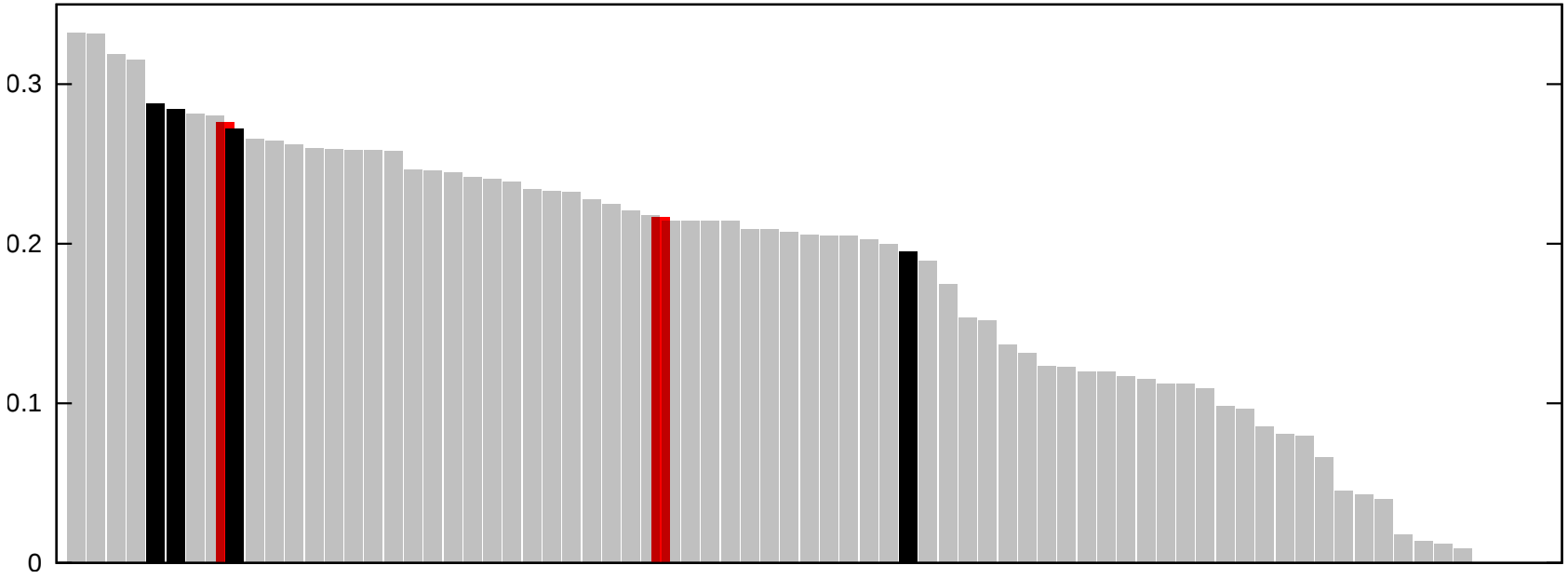
extraction and detector training

- 4 different CNN Caffe networks trained:
  - two training sets: ILSVRC 2010 and 2012
  - two network architectures: Krizhevsky (2012) and Zeiler & Fergus (2013)
  - two image scalings: aspect ratio preserving (Zeiler) and distorting (Krizhevsky)
- 24 different CNN Layer 6 activation features
  - four networks above
  - three feature-level fusions: center only, average, maximum
  - full image features or two-level spatial pyramid
- liblinear + HKM order 2 + histogram intersection

CNN activation features

increasing their number

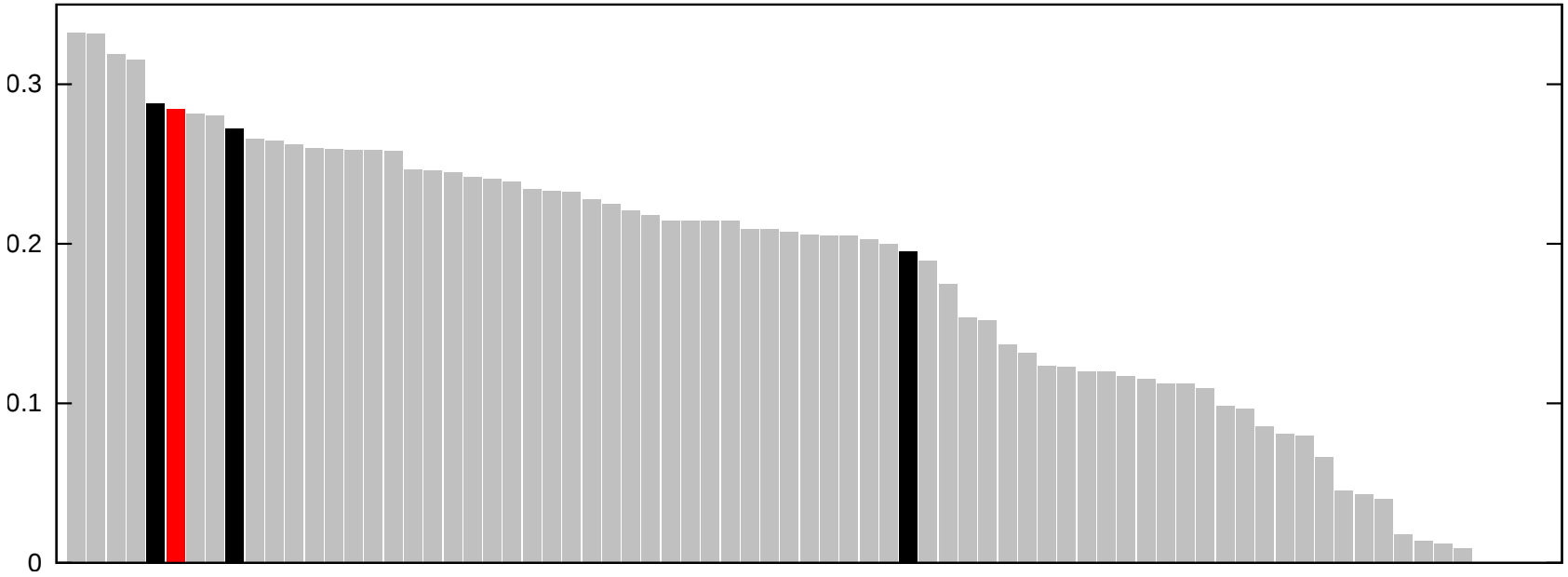
feature	*dim.*	*classifier*	*MXIAP*
worst individual full	4096	lin hkm2 int	0.1550
best individual full	4096	lin hkm2 int	0.1979
worst individual pyram.	8192	lin hkm2 int	0.2118
best individual pyram.	8192	lin hkm2 int	0.2164
fusion 12 full			0.2637
fusion 12 full + 12 pyram.			0.2759



Run 3 Hattifnattar, MXIAP = 0.2900 (0.2843)

applying hard negative mining

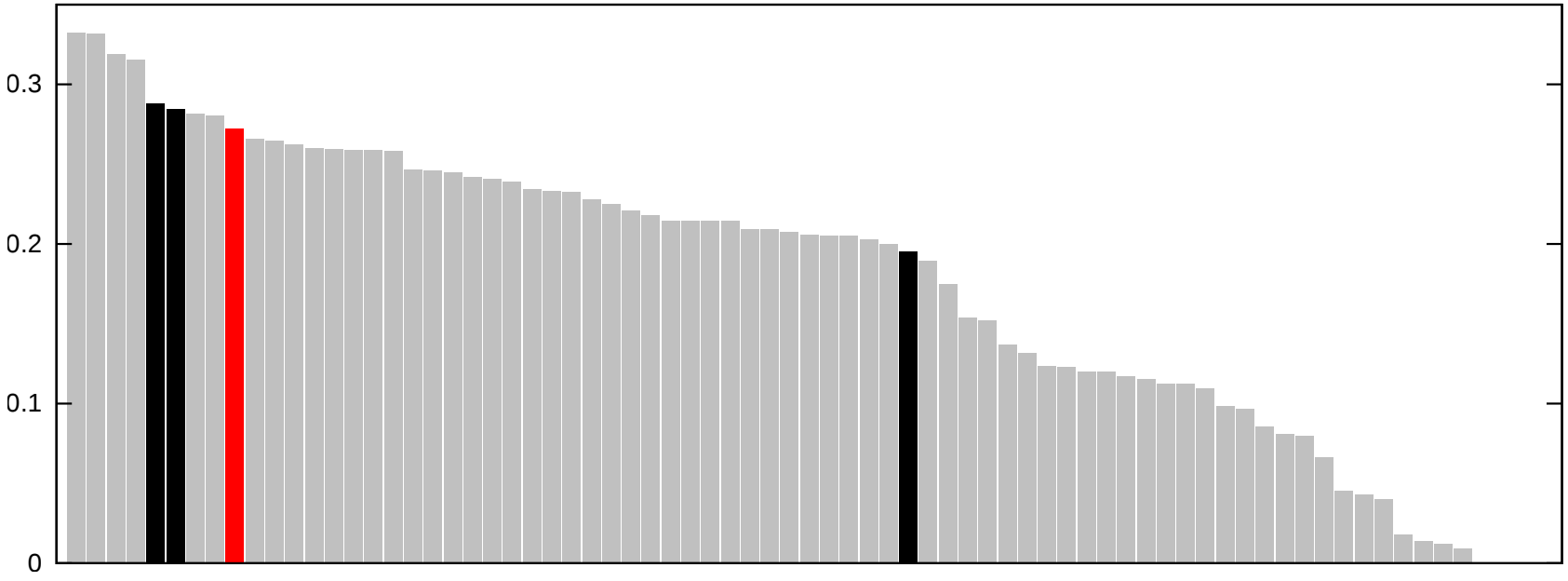
id	*setup*	*hard neg.m.*	*MXIAP*
0	*12 full*	no	0.2637
1	*12 full*	1 round	0.2504
2	*12 full*	2 rounds	0.2585
	fusion of 0+1		0.2742
	fusion of 0+1+2		0.2737
	24 full	no	0.2759
	24 full, fusion 0+1		0.2900



Run 2 Snusmumriken, MXIAP = 0.2777 (0.2722)

combining most of the detectors

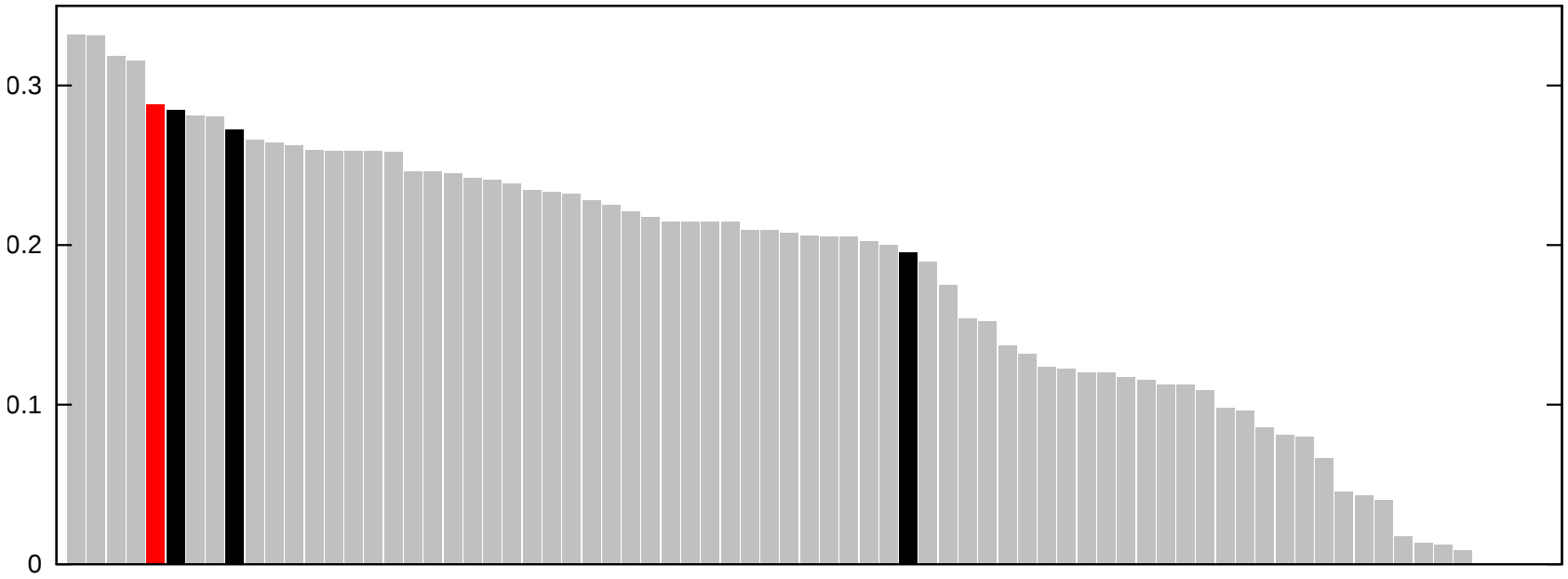
- 4 old SIFT/ColorSIFT BoV features
- old centrist feature
- old scalablecolor feature
- 2 new ColorSIFT 3-level pyramid features
- new Fisher vector feature
- new VLAD feature
- 24 new CNN activation features



Run 1 Mårran, MXIAP = 0.2936 (0.2880)

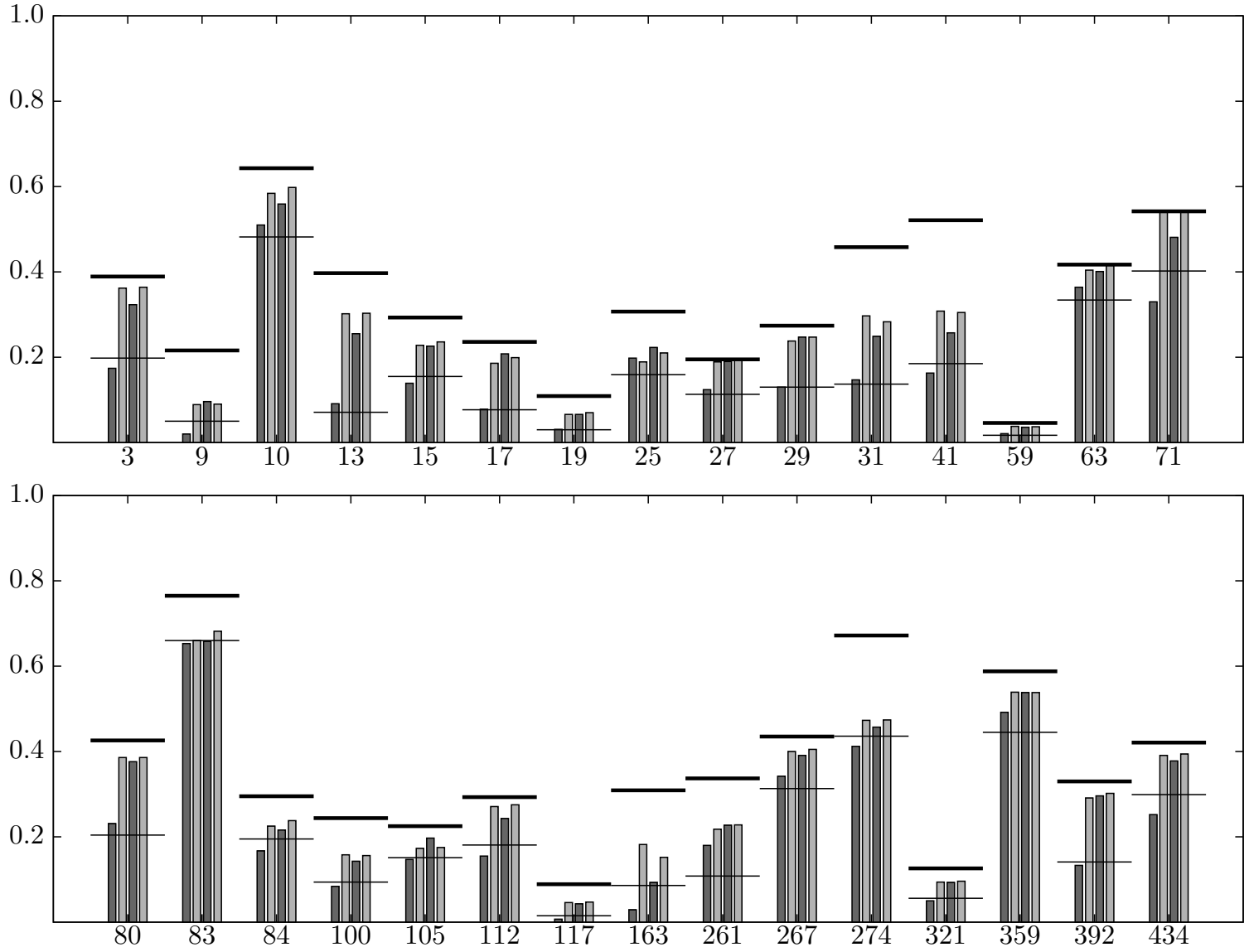
everything put together

- like Hattifnattar and Snusmumriken combined
  - one round of hard negative mining with CNN features
  - all features



Run 1 Mårran, Concept-wise results

top results for concepts 27 and 71



Contents

overview

related works

training and detection

details

conclusions

demo

Conclusions

- CNN activation features have a great promise as universal image representation:
  - fast to extract ($\approx$ 100ms CPU)
  - moderate feature dimensionalities
  - superior accuracy
  - suitable for use with linear classifiers ($\approx$ 1ms CPU)
  - variations can be generated
  - fusion provides additional accuracy
- hard negative mining is useful, but not many rounds are needed

Contents

overview

related works

training and detection

details

conclusions

demo

Demo with a documentary film

breaking the ice

Demo with a documentary film

entering the room