Aalto University
School of Science

Scaling up semantic indexing

Mats Sjöberg
Satoru Ishikawa, Markus Koskela, Jorma Laaksonen, Erkki Oja

CBIR research group (PicSOM)
http://research.ics.tkk.fi/cbir/

Department of Information and Computer Science
Aalto University, School of Science
mats.sjoberg@aalto.fi

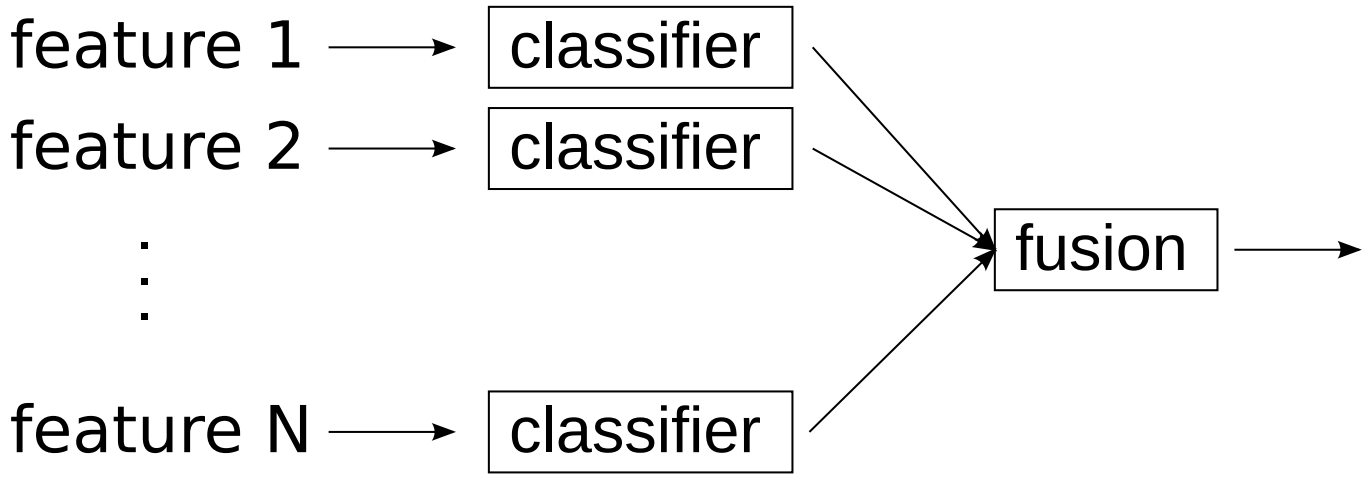
About us

- The PicSOM group from Aalto University has taken part in TRECVID since 2005.
- Before 2010 the university was called Helsinki University of Technology (Aalto = HUT + HSE + UIAH).
- In this year we participated in the semantic indexing (SIN) and known-item search (KIS) tasks.

Motivation

- We are currently working with the Finnish Broadcasting Company (YLE) and the National Audiovisual Archive (KAVA) on content-based analysis on the live TV signal.
- This includes doing fast online semantic indexing on streaming video
  $\Rightarrow$ increased emphasis on scalability and speed.
- Also, improving the speed of offline training of detectors.
- In TRECVID 2011 we focused on radically improving the speed of both the online and the offline components of the semantic indexing pipeline.

Semantic indexing pipeline



- (Color)SIFT + SVM (χ^2) + (weighted) geom. mean fusion.
- *Similarity Cluster* weighting (Wilkins et al, 2007).
- Offline: extract features from training data, train classifiers (parameter selection most time consuming).
- Online: extract features from new image(s), predict with trained detectors.

Feature extraction

- Bag-of-visual-words features (BoV) very successful.
- Best results for PicSOM group in TRECVID: ColorSIFT with dense sampling, 1x1-2x2 pyramid, soft assignment,
- However, computationally very expensive: about 1 image per second.
- Consider: (online) 25 frames per second video (!), or (offline) 3 million image database: 35 days.

Feature extraction, cont.

- We have looked at other non-BoV features.
- Local Binary Patterns (LBP)[1], simple and efficient texture operator, useful e.g. for face description.
- A promising choice: CENsus TRansform hISTogram (Centrist)[2].
- Basically an LBP histogram reduced in dimensionality (40) with PCA, plus mean and stddev.
- This done in a 2 level spatial pyramid, giving a dimensionality of $(40 + 2) \times (25 + 5 + 1) = 1302$.

[1] Pietikäinen, Hadid, Zhao, Ahonen:, Computer Vision Using Local Binary Patterns, Springer, 2011

[2] Wu, Rehg: CENTRIST: A Visual Descriptor for Scene Categorization, PAMI, 2011.

SIFT vs Centrist

Example: extract features for 2268 images

- ColorSIFT: 43 minutes, about 1 image per second
- Centrist: 49 seconds, about 50 images per second

Centrist is roughly 50 times faster.

Now live video starts to look feasible!

Training classifiers

- Kernel SVM's state-of-the-art, but computationally expensive.
- Linear classifiers fast, but less accurate.
- Offline, but constrains database size, concept vocabulary, less room for experimentation.

Parameter selection most time consuming phase:

- C-SVM has two parameters (C, γ) (LIBSVM[1]),
- linear classifier (L^2 regularised logistic regression solver from LIBLINEAR) has only one parameter (C).

[1] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, ACM TIST, 2011.

Training classifiers, cont.

- Parameter selection times in TRECVID 2011, with a somewhat naive line search followed by grid search.
- SVM: on average 3 days!
- linear: on average a bit more than 1 hour!
- (A strong bias towards SVM since our cluster has a maximum run-time of 7 days!)

hours	SVM	linear	×
min	0.6	0.2	3.5
max	168.0	4.2	40.3
median	33.9	1.2	27.2
average	79.1	1.3	61.1

Prediction with trained classifier

- Critical in online scenario: detect concepts in new images.
- Prediction with LIBSVM takes around 100–500 milliseconds per image with ColorSIFT features
- Consider: with 300 concepts (e.g. TRECVID) this is in the order of 100 seconds per image.
- LIBLINEAR takes 1–3 milliseconds per image.
- In the order of 1 second per image or less for 300 concepts
- Real-time video is typically 25 images per second or more, of course not all frames need to be classified

Experiments

classifier	feature	MXIAP
SVM	ColorSIFT	0.1233
	SIFT	0.1139
	Centrist	0.0939
linear	ColorSIFT	0.0329
	SIFT	0.0292
	Centrist	0.0289
	EdgeFourier	0.0101
	ScalableColor	0.0182

- Centrist not quite as good as BoV features, but quite good considering 50-fold speedup.
- LIBLINEAR for single features much worse than LIBSVM.

Time estimates

classifier + features	MXIAP	offline (days)	online (secs)
SVM ColorSIFT	0.1233	77.0	45.6
SVM Centrist	0.0939	5.5	45.0
SVM 3 best fusion	0.1363	123.3	136.0
linear ColorSIFT	0.0329	73.7	1.1
linear 3 best fusion	0.0827	113.5	2.3
linear 12 fusion	0.0986	189.2	7.0
linear 14 fusion	0.1145	591.2	11.4
SVM Centrist + linear 10	0.1116	81.2	50.2
SVM 3 + linear 14	0.1398	601.1	146.4

- Rough estimate of offline and online processing times.
- Scenario: 1M images, detecting 300 concepts online.

Time estimates, cont.

classifier + features	MXIAP	offline (days)	online (secs)
SVM ColorSIFT	**0.1233**	**77.0**	45.6
SVM Centrist	**0.0939**	**5.5**	45.0
SVM 3 best fusion	0.1363	123.3	136.0
linear ColorSIFT	0.0329	73.7	1.1
linear 3 best fusion	0.0827	113.5	2.3
linear 12 fusion	0.0986	189.2	7.0
linear 14 fusion	0.1145	591.2	11.4
SVM Centrist + linear 10	**0.1116**	**81.2**	50.2
SVM 3 + linear 14	0.1398	601.1	146.4

- Centrist result is in the same order of magnitude as ColorSIFT, but much faster to calculate.

Time estimates, cont.

classifier + features	MXIAP	offline (days)	online (secs)
SVM ColorSIFT	0.1233	77.0	45.6
SVM Centrist	0.0939	5.5	45.0
SVM 3 best fusion	0.1363	123.3	**136.0**
linear ColorSIFT	**0.0329**	73.7	1.1
linear 3 best fusion	**0.0827**	113.5	2.3
linear 12 fusion	**0.0986**	189.2	7.0
linear 14 fusion	**0.1145**	591.2	**11.4**
SVM Centrist + linear 10	0.1116	81.2	50.2
SVM 3 + linear 14	0.1398	601.1	146.4

- Linear results improve strongly by adding features.
- Even with five times more features, 10-fold speed increase compared to SVM.

Time estimates, cont.

classifier + features	MXIAP	offline (days)	online (secs)
SVM ColorSIFT	0.1233	77.0	45.6
SVM Centrist	0.0939	5.5	45.0
SVM 3 best fusion	0.1363	123.3	136.0
linear ColorSIFT	0.0329	73.7	1.1
linear 3 best fusion	0.0827	113.5	**2.3**
linear 12 fusion	0.0986	189.2	**7.0**
linear 14 fusion	0.1145	591.2	**11.4**
SVM Centrist + linear 10	0.1116	81.2	50.2
SVM 3 + linear 14	0.1398	601.1	146.4

- Linear prediction is fast even with many features.

Conclusions

- For offline speed, fast feature calculation is most critical.
- Centrist is 50 times faster than best BoV feature.
- For online speed, prediction time of classifier is most critical.
- Linear classifier is $50 - 100$ times faster than kernel SVM.
- With many features, linear classifier can achieve same order of magnitude MXIAP as single best SVM.