Multi-Frame, Multi-Modal, and Multi-Kernel Concept Detection in Video

Cees G.M. Snoek¹, Koen E.A. van de Sande¹, Jasper R.R. Uijlings¹, Miguel Bugalho², Isabel Trancoso², Fei Yan³, Muhammed A. Tahir³, Krystian Mikolajczyk³, Josef Kittler³, Theo Gevers¹, Dennis C. Koelma¹, Arnold W.M. Smeulders¹

¹ University of Amsterdam ² INESC-ID ³ University of Surrey

Conclusions TRECVID 2008

• Good settings for Bag-of-Words
  - SIFT + colorSIFT improves ~8%
  - Soft codebook assignment improves ~7%
  - Multi-frame analysis improves ~20%

http://www.MediaMill.nl
Myth: TRECVID incremental only

>100% improvement in just 3 years

Snoek et al, TRECVID 2008
Van de Sande et al, PAMI 2010
Van Gemert et al, PAMI 2010

State-of-the-Art

Spatio-temporal sampling → Visual feature extraction → Codebook transform → Kernel-based learning
State-of-the-Art

Software available for download at http://color descriptors.com

Our TRECVID 2009 focus

Snoek et al, TRECVID 2008
Van de Sande et al, PAMI 2010
Van Gemert et al, PAMI 2010
Our TRECVID 2009 focus

Roadmap

- Spatio-temporal sampling
- Visual feature extraction
- Codebook transform
- Audio concept detection
- Kernel-based learning
1,000,000 frames analyzed

- Multi-frame biggest improvement in 2008
  - Extend further by analyzing up to 10 extra i-frames/shot
  - Yields 1M frames to analyze for the test set collection
- Need to speed-up by being “smart and strong”
  - Speed-up feature extraction
  - Speed-up quantization
  - Speed-up kernel-based learning
  - Speed-up by computing

Roadmap

- Spatio-temporal sampling
- Visual feature extraction
- Codebook transform
- Kernel-based learning
- Audio concept detection
Fast dense descriptors

\[
\begin{align*}
A &= \begin{pmatrix}
1 & 0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 0 & 1
\end{pmatrix} \\
R &= \begin{pmatrix}
\cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\end{pmatrix}
\end{align*}
\]

\[
x = \begin{pmatrix}
1 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & \cdots \\
0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & \cdots \\
\end{pmatrix}
\]

Linear Interpolation

Image Patch

Pixel-wise Responses \( R \)

Final Descriptor

\[
\begin{pmatrix}
2x & \text{speed-up} \\
16x & \text{speed-up}
\end{pmatrix}
\]

Roadmap

- Spatio-temporal sampling
- Visual feature extraction
- Codebook transform
- Audio concept detection
- Kernel-based learning
Fast quantization

- Random forests
  - Randomized process makes it very fast to build
  - Tree structure allows fast vector quantization
  - Logarithmic rather than linear projection time

- Real-time BoW
  - When used with fast dense sampling
  - SURF 2x2 descriptor instead of 4x4
  - RBF kernel

GPU-empowered quantization

- Achieve data-parallelism by writing Euclidean distance in vector form
Roadmap

- Spatio-temporal sampling
- Visual feature extraction
- Codebook transform
- Kernel-based learning
- Audio concept detection

SVM pre-computed kernel trick

- Use distance between feature vectors
  - Feature length easily > 100,000
    \[ k(F, F') = e^{-\frac{1}{\lambda} \text{dist}(F, F')} \]
- Increase efficiency significantly
  - Pre-compute the SVM kernel matrix
  - Long vectors possible as we only need 2 in memory
  - Parameter optimization re-uses pre-computed matrix
GPU-empowered pre-computed kernel

1. Compute average distances per $N^2$ kernel sub-block
2. Compute kernel function values

Computing

- 2009 system much more efficient than 2008 system
  - 6x more visual data analyzed using less compute power

- Some best estimates:
  - Visual feature extraction: 8400 Processor-Node-Hours
  - Training concept detectors: 4000 PNH
  - Applying concept detectors: ~1 week GPU
Roadmap

- Spatio-temporal sampling
- Visual feature extraction
- Codebook transform
- Kernel-based learning
- Audio concept detection

Audio concept detection

- External sound corpus: ~100 concepts
- siren, water, ... speech, female voice ...
- monologue, dialogue ...
- music events
- low frequency

- Early fusion of features:
  - MFCCs (+ deltas), PLPs (+ deltas), Brightness, Bandwidth, ZCR, Pitch, Harmonicity, Shifted delta cepstra, Audio spectrum envelope and flatness
  - 0.50s window length, with 0.25s spacing

Bugalho et al, InterSpeech 2009
Trancoso et al, ICME 2009
Roadmap

- Spatio-temporal sampling
- Visual feature extraction
- Codebook transform
- Kernel-based learning
- Audio concept detection

Multi-kernel learning

- **Kernel Discriminant Analysis** combined with spectral regression [Tahir09]
  - We use SR-KDA with 6 visual kernels
  - Weighted output combined using SUM rule
- **Multi-Kernel Fisher Discriminant Analysis**
  - We use non-sparse L2 MK-FDA [Yan09]
  - Fusion of 1 audio and 6 visual kernels
    - 20 audio concept detector scores used as input for RBF kernel

Tahir et al, ICCV-Subspace 2009
Yan et al, ICDM 2009
Experiments (all type A)

- **Baseline**: single-frame SFS on all visual kernels
- **Experiment 1**: multi-modal & multi-kernel
  - SR-KDA (visual only)
  - MK-FDA (audiovisual fusion)
- **Experiment 2**: multi-frame
  - Visual fusion: 5 extra i-frames + MAX fusion [donated]
  - Best-of: 1 to 10 extra i-frames + MAX/AVG fusion
  - SFS: all multi-frame visual kernel combinations

Results: experiment 1

- Multi-kernel improves upon baseline: ~9%
- Multi-modal kernel outperforms uni-modal kernel only slightly: ~2%
  - ...but for specific (audiovisual) concepts more impressive improvement, up to 50%

![Graph showing TRECVID 2009 High-level Feature Task Benchmark Comparison]
Results: experiment 2

- Multi-frame is true performance booster, improvement over baseline: ~30%
- Best to select optimal number of extra frames, per kernel, per concept,
  - On average 6 additional i-frames with MAX or AVG fusion is a solid choice

Visualizing multi-frame impact
Conclusions TRECVID 2009

- Multi-modal using multi-kernel seems promising
  - More experiments needed to be conclusive
- Multi-frame is true performance booster
  - 30% improvement over single-frame baseline
  - Time for the community to move on to video analysis

References

References II


Contact info

- Cees Snoek
  http://staff.science.uva.nl/~cgmsnoek