shot boundary detection combining similarity analysis and classification

Matthew Cooper\textsuperscript{1}, Ting Liu\textsuperscript{2}, and Eleanor Rieffel\textsuperscript{1}

\textsuperscript{1}FX Palo Alto Laboratory
http://www.fxpal.com

\textsuperscript{2}Dept. of Computer Science
Carnegie Melon University
http://www.autonlab.org
traditional video segmentation

what’s working and what’s not?
- features are YUV histograms (block and global)
- replace ad hoc peak detection with supervised classification as in [Qi, et al., 2003]

reformulating segmentation

Low-level Feature Extraction → Local Novelty Analysis → Boundary / Non-boundary Classification

VIDEO

Segment

Low-level Features

Pairwise similarity comparison(s) → Linear kernel correlation → Novelty Features

FEATUREREFERENCE

LOCAL

SEGMENTS
inter-frame similarity analysis

- concatenate YUV histogram features
  \[ f_i \rightarrow x_i \ (x_i \in \mathbb{R}^p) \]
- construct L1 similarity matrix:
  \[ S(i, j) = \sum_{p=1}^{P} |X_i(p) - X_j(p)| \]
novelty via kernel correlation

- scale-space kernel linearly combines adjacent frame comparisons

- more generally:

\[ \nu(n) = \sum_{l=-L}^{L-1} \sum_{m=-L}^{L-1} K(l, m) S(n + l, n + m) \]
related work: dissimilarity kernels

- **scale-space (SS) kernel** weights only adjacent inter-frame similarities [e.g. Witkin, 1984]

- **diagonal cross-similarity (DCS)** kernel weights inter-frame similarity of pairs $L$ frames apart [Pye et al., 1998; Pickering et al., TRECVIDs]

- **row (ROW)** kernel compares current frame to each frame in local neighborhood [Qi, et al., 2003]
dissimilarity kernels

- cross similarity (CS) kernel is a matched filter for ideal dissimilarity boundary

- full similarity (FS) kernel penalizes within-segment dissimilarity
  [Cooper and Foote, ICIP 2001]
input features for classification

\[ \nu(n) = \sum_{l=-L}^{L-1} \sum_{m=-L}^{L-1} K(l, m) S(n + l, n + m) \]

- **kernel-based features**: concatenate frame-indexed kernel correlations \( \nu_L(n) \) for \( L=2,3,4,5 \), for both global histogram similarity and block histogram similarity

- **raw similarity features**: concatenate all raw similarity comparisons that contribute to kernel correlation for \( L=5 \) (without linearly combining them)
experimental setup

- efficient exact kNN classifier provided by T. Liu and A. Moore at CMU (http://www.autonlab.org)
  - ball-tree implementation \~ 10 times speedups over naïve kNN
  - for details, see [Liu, Moore, Gray, NIPS 2003]

- TRECVID 2002 test set for \textbf{cut} boundary detection
  - almost 6 hours of broadcast news data
  - manual ground truth, 1466 cut boundaries
  - medians from TV02: recall = 0.86, precision = 0.84
  - hold-one-out cross validation, $k = 11$
comparative results

- FS similarity features provide most information and achieve best overall performance
setup for SB04

- to extend to cut and gradual detection, we follow two-step binary classification approach in [Qi, et al., 2003]

- unlike prior work no smoothing of classifier outputs, no motion, flash, etc.
- efficient exact kNN classifier $k = 11$
- 8 CNN and ABC videos from SB03 test set
- hold-one-out cross validation
training – varying the similarity measure

- FS pairwise similarity features used
- 8 ABC and CNN videos in SB03 test set used for training
- testing similarity measures

\[
S(i, j) = \sum_{p=1}^{P} |X_i(p) - X_j(p)|
\]

\[
S(i, j) = \sqrt{\sum_{p=1}^{P} (X_i(p) - X_j(p))^2}
\]

\[
S(i, j) = \sum_{p=1}^{P} \frac{(X_i(p) - E_{ij}(p))^2}{(X_i(p) + E_{ij}(p))}
\]

\[
S(i, j) = \sum_{p=1}^{P} \frac{(X_i(p) - E_{ij}(p))^2}{(X_i(p) + E_{ij}(p))^2}
\]

- testing different lag L=5, 10
- random projection for dimension reduction for L=10
comparing similarity measures
training – varying $L$

- $L=10$ implies **FS** feature dimensionality is $d=380$
- problem of fast kNN
  - significant speed-up when $d$ is small: $O(1) \sim O(dN\log N)$
  - little speed-up when $d$ is large: $O(dN^2)$
- random projection

THM (Johnson-Lindenstrauss lemma) For any $0 < c < 1$ and any integer $N$, let $d'$ be a positive integer such that

$$d' \geq 4\left(\frac{c^2/2 - c^3/3}{c^2/2 - c^3/3}\right)^{-1} \ln N$$

Then for any set $V$ of $N$ points in $\mathbb{R}^d$, there is a map $f: \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ such that for all $u, v \in V$,

$$(1-c)||u - v||^2 \leq ||f(u) - f(v)||^2 \leq (1+c)||u - v||^2.$$
varying $L$ for fixed featured dimensionality
SB04 systems

- training data consists of 8 ABC, CNN videos from SB03 set
- 90% of non-boundary frames discarded
- $k = 11$
- sensitivity determined by $0 \leq \kappa \leq k$
- post-processing to avoid spurious boundaries in local temporal neighborhood
Cut Results

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>P</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg</td>
<td>0.831</td>
<td>0.762</td>
<td>0.776</td>
</tr>
<tr>
<td>Best</td>
<td>0.920</td>
<td>0.951</td>
<td>0.935</td>
</tr>
<tr>
<td>&lt;FXPAL&gt;</td>
<td>0.903</td>
<td>0.940</td>
<td>0.921</td>
</tr>
</tbody>
</table>

Cut Detection Performance: SB04

- L=5
- L=10 (RP)
- Competition
gradual results

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>P</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg</td>
<td>0.503</td>
<td>0.578</td>
<td>0.565</td>
</tr>
<tr>
<td>Best</td>
<td>0.846</td>
<td>0.775</td>
<td>0.8089</td>
</tr>
<tr>
<td>&lt;FXPAL&gt;</td>
<td>0.756</td>
<td>0.789</td>
<td>0.769</td>
</tr>
</tbody>
</table>
mean results

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>P</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg</td>
<td>0.7255</td>
<td>0.727</td>
<td>0.709</td>
</tr>
<tr>
<td>Best</td>
<td>0.884</td>
<td>0.896</td>
<td>0.890</td>
</tr>
<tr>
<td>&lt;FXPAL&gt;</td>
<td>0.856</td>
<td>0.891</td>
<td>0.872</td>
</tr>
</tbody>
</table>
time complexity

- 1 decode run includes histogram extraction (code never optimized) for all SysIDs
- 2 classification runs correspond to 10 SysIDs
- all times for all 12 videos
conclusions

- many segmentation approaches can be formulated within the framework of inter-frame similarity analysis and linear kernel correlation
- non-parametric supervised classification is effective for media segmentation
- very general framework

- thanks to Andrew Moore at CMU
- for more information: cooper@fxpal.com