Imperial College London

Video Retrieval using Search and Browsing with Key Frames

Daniel Heesch, Marcus Pickering, Stefan Rüger, Alexei Yavlinsky

Multimedia Knowledge Management, Department of Computing, South Kensington Campus, Imperial College London, SW7 2AZ, UK
http://km.doc.ic.ac.uk/
{daniel.heesch,m.pickering,srueger,agy02} @ imperial.ac.uk

Detection Tasks

Shot boundary detection

Overview

- Colour histograms used to characterise frames.
- Frame divided into 9 blocks, histogram taken for each of R,G,B components for each block.
- Look at differences between histograms up to 16 frames either side of current frame.
- Distance measure calculated at each frame:

$$d_n(t) = \frac{1}{n}\sum_{i=0}^{n-1} D(t+i, t-n+i)$$

where $D(i,j)$ represents the median block distance between the histograms of frames i and j.

- Declaration of shot boundaries is based on characteristics of peaks in the distance measure.

Results

- Our best run achieved the following results:

	Recall	Precision
All	0.85	0.87
Cuts	0.91	0.89
Graduals	0.70	0.81

- Global comparison shows our system to be the third best amongst all groups.
- Performance was particularly high relative to other systems in the challenging task of detection of gradual transitions.

Feature Detection

- Attempted “vegetation” feature only using colour-based classifier trained for grass.
- Images segmented into regions, regions characterised by centroids of pixel clusters in RGB space.
- Training set created from several hundred positive and negative examples.
- Test region classified by finding its 25 nearest neigbours in the training set, where “nearest” is defined using the earth-mover’s distance.
- Relevance score for a shot was square of highest region score.
- Average precision for our two runs was between 0.08 and 0.09 (compared to median 0.15). Hit count for our best run was 360 (compared to median 367).

Search Task

Features

HSV Global Colour Histograms

- Quantised distribution in 3D colour space.
- Feature vector is list of proportions of pixels that fall into respective 3D histogram bins.

HSV Focus Colour Histograms

- As above, but only central 25% of image considered.
- Close similarities between images that differ primarily with respect to background.

Colour Structure Descriptor (MPEG-7)

- 8x8 structuring window slid over image, each bin contains number of window positions for which there is at least one pixel falling into that bin.

Marginal RGB Colour Moments

- Histograms formed for each colour channel and the mean and 2nd, 3rd and 4th central moments computed for each marginal colour distribution.

Thumbnail

- Grey values from scaled down original image.
- Suited to detection of near-identical image copies.

Convolution Filters

- Feature vector generated through application of convolution filters to the three colour channels.
- Three stage process captures arrangements of features in the image.

Variance

- Grey value standard deviations in a 5x5 sliding window for each of 9 tiles.

Smoothness

- Smoothness measure for each of 64 image tiles.

Uniformity

- Uniformity measure for each of 64 image tiles.

Bag of Words

- Stemmed words from associated transcript accompanied by corresponding tf-idf weights.

Text

- Test data is from LIMSI speech recogniser
- Query taken from XML topic definition and relevance of each test shot determined by the Managing Gigabytes search engine.

Retrieval using *k*-nearest neighbours

- Distance for descriptor d from test image T_i to each of k nearest (Manhattan distance Dis_d between feature vectors) positive or negative examples

$$\mathrm{Dis}_d(Q,T_i) = \frac{\sum_{n\in N}\left(\mathrm{dist}(T_i,n)+\varepsilon\right)^{-1}}{\sum_{q\in \hat{Q}}\left(\mathrm{dist}(T_i,q)+\varepsilon\right)^{-1}+\varepsilon}$$

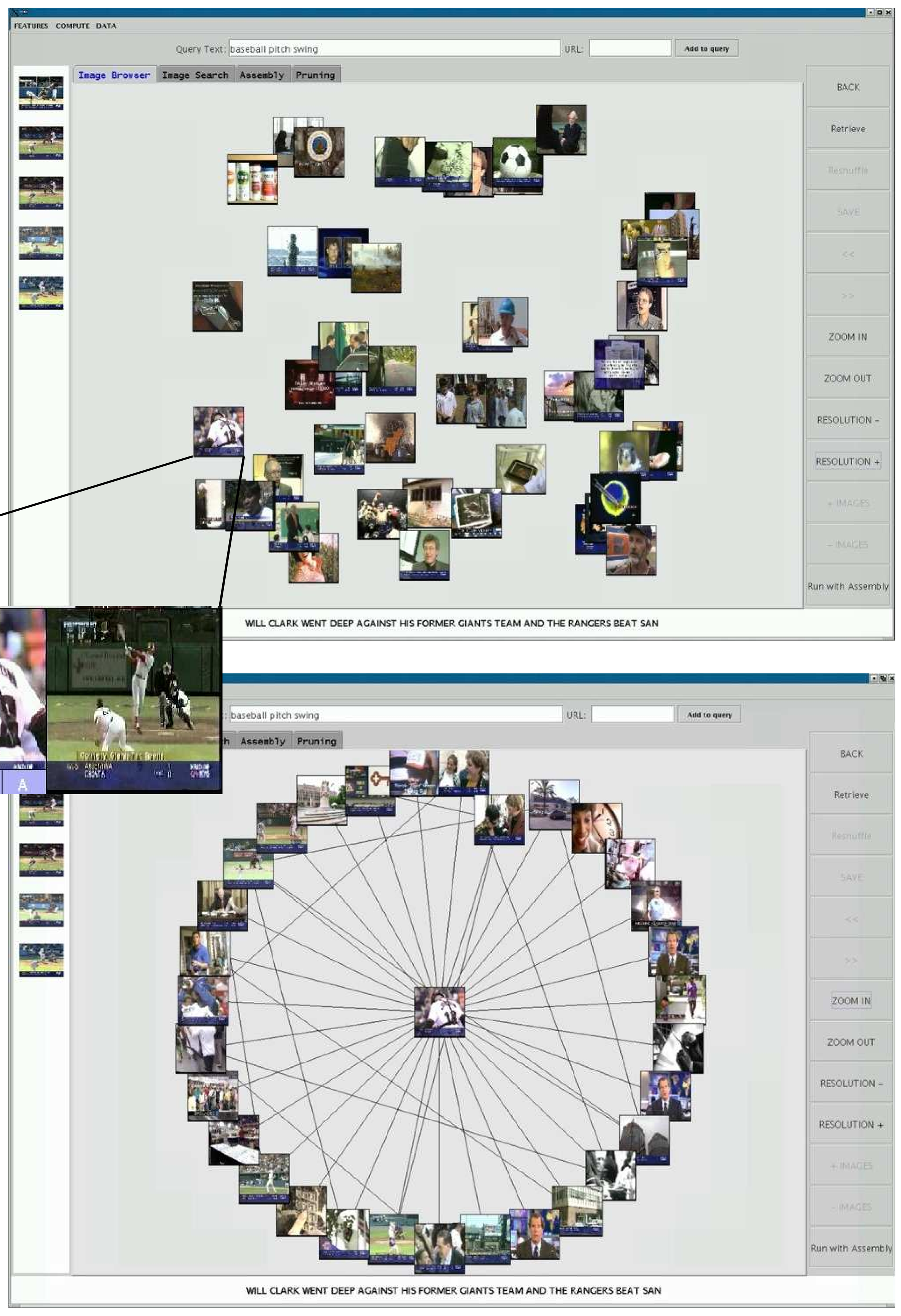
where Q and N are the sets of positive and negative examples amongst the k nearest neighbours, such that $|Q|+|N|=k$

Relevance feedback

- Distance from centre is proportional to dissimilarity from query.
- The sum of squared errors:

$$SSE(w) = \sum_{i=1}^{N}\left[D_s^w(Q,T_i) - D_u(Q,T_i)\right]^2$$

is minimised with respect to w



Browsing

Temporal browsing

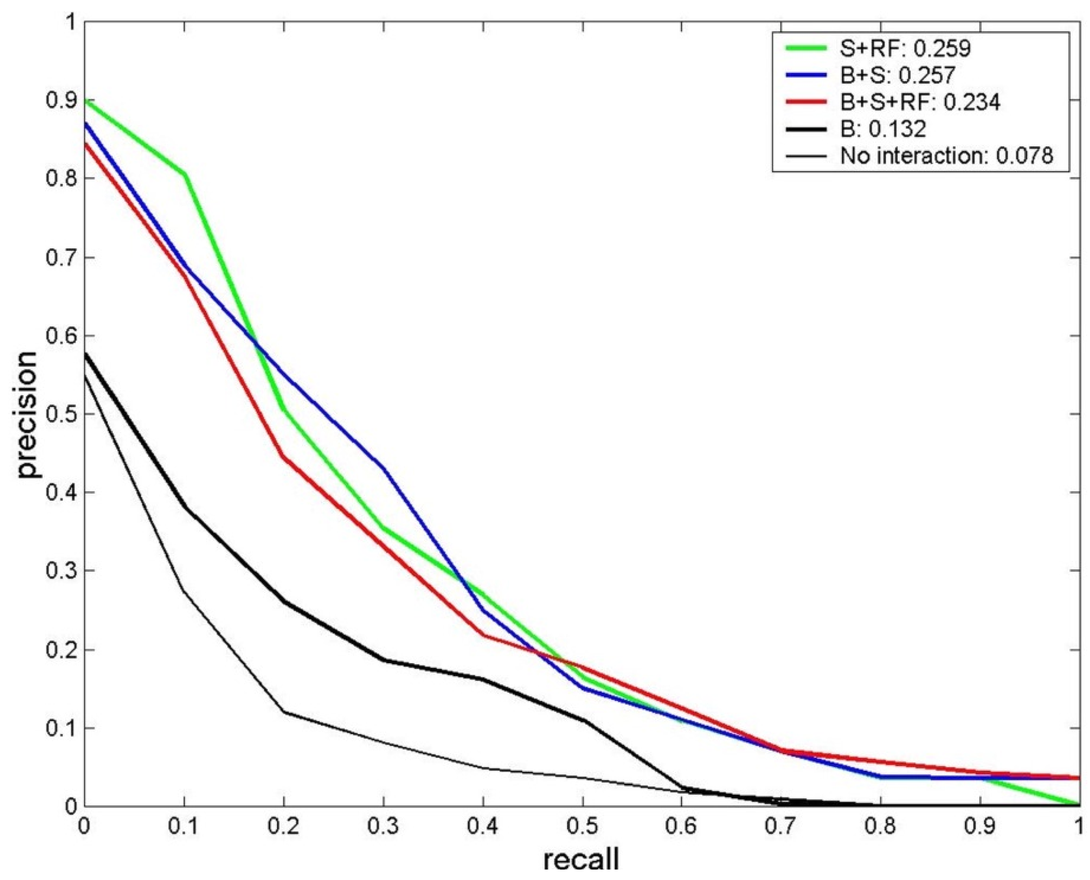
- Sliding window consisting of an image and its left and right shot neighbours.
- Window can be slid in either direction along the broadcast.

Lateral browsing

- Each image is connected to ‘similar’ images in a large pre-computed directed graph.
- Clicking an image displays its graph neighbours. Two are connected if there exists at least one feature combination for which one image is ranked top when querying with the other.

Results

- We entered four interactive runs:
  - S+RF - Search and relevance feedback.
  - B+S - Browsing and search.
  - B+S+RF - Browsing, search and relevance feedback.
  - B - Browsing only.



- 3 out of 4 interactive runs amongst top 8.
- All of the top 3 runs have significantly better performance than the browsing-only run, which itself has performance significantly better than the manual run (both at alpha = 0.05).
- “Browsing only” run better than 25% of all interactive runs.

Conclusions

Detection tasks

Shot boundary detection

- Highly accurate despite minimal training in news domain, returning some of the best results amongst all systems.

Feature detection

- Promising approach which will fare better when properly trained.

Search

- Browsing improves significantly over manual search and provides a viable alternative to interactive search by example.
- Temporal browsing was a useful tool since relevant shots were often located near each other in the broadcast.
- Although adding lateral browsing did not statistically significantly change the overall interactive performance, it did subjectively add to user satisfaction.