Video Google:
A Text Retrieval Approach to Object Matching in Videos

Joseph Bonath
22 April, 2014
ECG 782

Overview

- Background
- Descriptors
  - Shape Adapted
  - Maximally Stable
- Implementation
  - Visual Indexing
  - Object Retrieval
- Summary

Background
Text Retrieval ("Google")

- Process
  - (1) Documents are parsed into words
  - (2) Words are represented by stems (root words)
  - (3) Stop list is created to reject common words

- Each document is represented by a vector
  - Components are given by the frequency of occurrence of the words contain in the document

Background
Text Retrieval ("Google")

- Set of all document vectors are organized into an "inverted file"
  - Composed of an entry for each word followed by a list of documents and the position in which the word occurs
  - Allows for efficient retrieval
- Text Retrieval
  - Compute vector of word frequencies
  - Return documents with closest vectors (measured by angle)

Descriptors

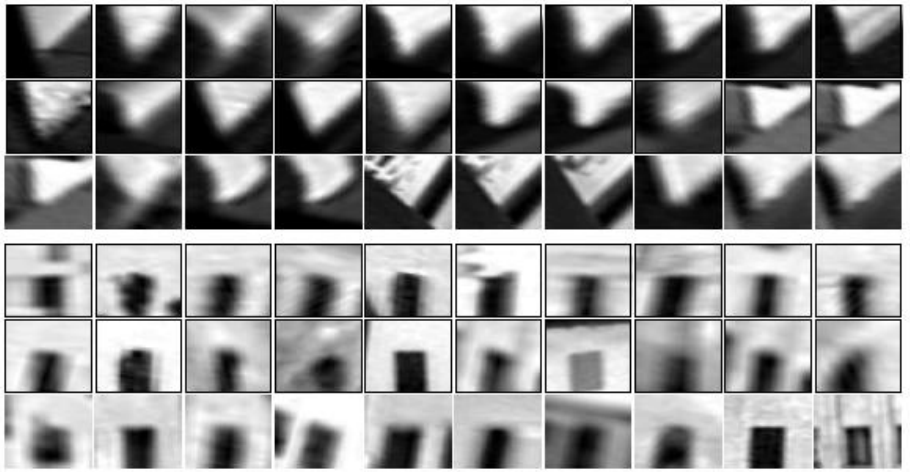
Shape Adapted (SA)

- Iteratively determine ellipse shape, scale, and center about an interest point
- Scale – local exremum of a Laplacian
- Shape – maximize intensity gradient isotropy over elliptical region
- Tend to center on corner-like features

Maximally Stable (MS)

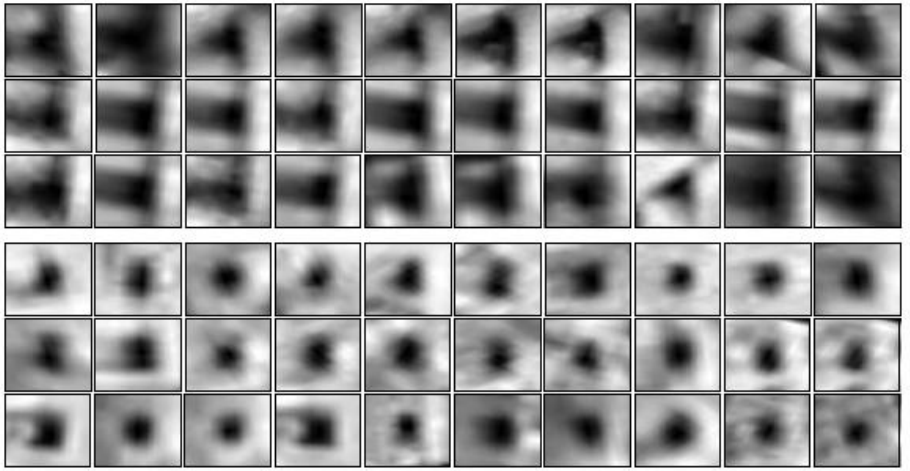
- Select areas from an intensity watershed image segmentation
- Regions chosen are those approximately stable over a varying threshold
- Correspond to blobs of high contrast to the surroundings

SA



(a)

MS



(b)

Samples from the clusters corresponding to a single visual word. (a) Two examples of clusters of Shape Adapted regions. (b) Two examples of clusters of Maximally Stable regions.

Descriptors

- SA / MS regions are clustered separately
  - Cover independent regions of a scene
  - Different “vocabularies”
- Each region represented by a 128-dimensional vector
  - Uses SIFT descriptor – emphasizes orientation of gradient
  - Invariant to small translations in region position
- Combine information across sequence of frames
  - Detected regions are tracked using constant velocity dynamical model and correlation
  - Regions not surviving minimum 3 frames are rejected
  - Descriptor is averaged over the track
  - Reduces noise in descriptor and rejects unstable regions

Implementation

- Vector quantize descriptors into clusters
  - Visual “words” for text retrieval
  - K-means clustering
- Regions are tracked through contiguous frames
  - Mean vector descriptor $\bar{x}_i$ computed for each of the i regions
- Reject unstable regions
  - Regions with 10% largest diagonal covariance
- Only use a subset of full film
  - Fraction of each shot to minimize computation

Visual Indexing

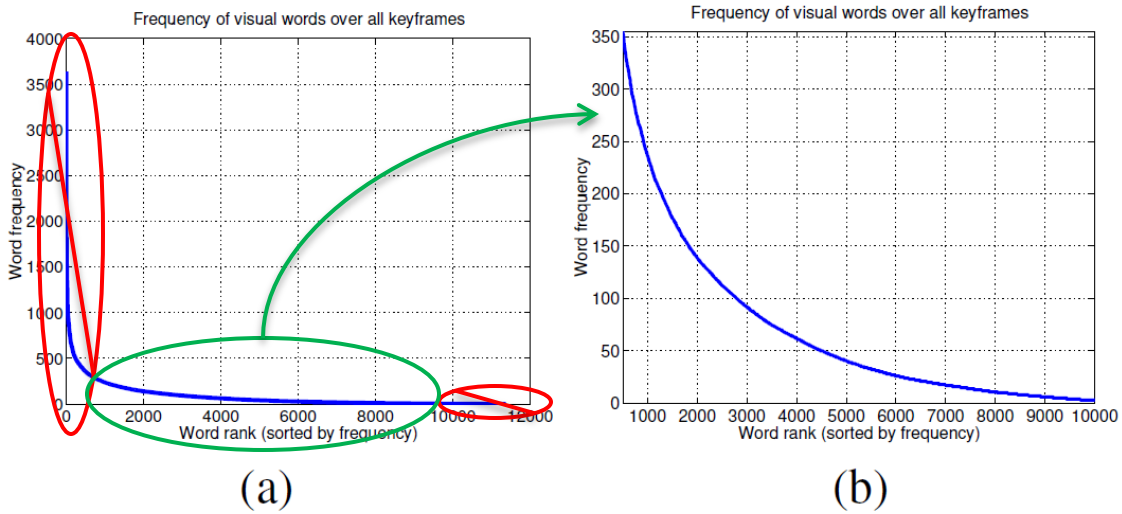
- Weighting method
  - "term frequency – inverse document frequency" (tf-idf)
  - Each document is represented by term $V_d = (t_1, \ldots, t_i, \ldots, t_k)^T$ where

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i}$$

  - Word frequency weights words occurring often in a particular document thus describing it well
  - Inverse document frequency down weights words appeaing often throughout the database

Object Retrieval

- Stop List
  - Most frequent visual words appearing in almost all images are suppressed
  - Analogous to removing common words from a text search (a, the, it, etc.)

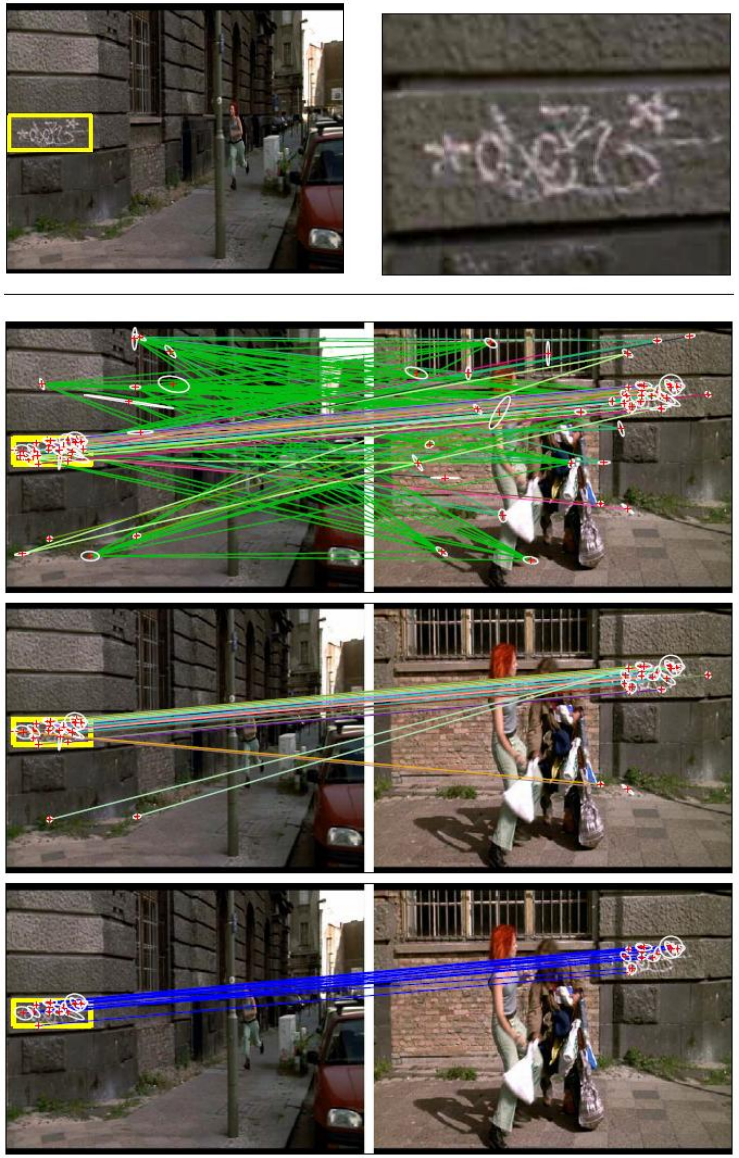


Frequency of MS visual words among all 3768 keyframes of Run Lola Run (a) before, and (b) after, application of a stoplist.

Object Retrieval

- Spatial Consistency
  - After retrieval using weighted frequency vector, re-rank frames based on spatial consistency
  - Loosely – neighboring matches can simply lie somewhere in the surrounding area
  - Strictly – neighboring matches must have the same spatial layout
    - Matched regions can provide affine transformation
  - 15 nearest neighbors are used, each one voting for the frame
    - Number of votes determines rank

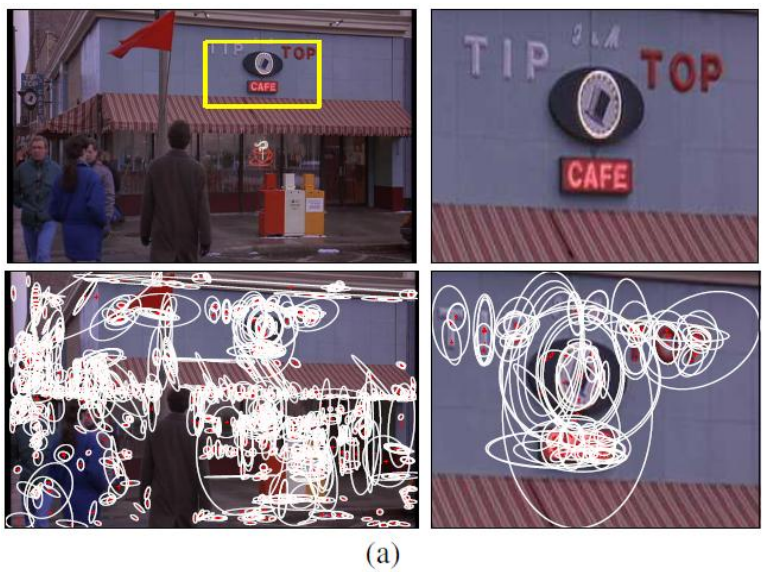
Matching stages. Top row: (left) Query region and (right) its close-up. Second row: Original word matches. Third row: matches after using stop-list, Last row: Final set of matches after filtering on spatial consistency.



Summary

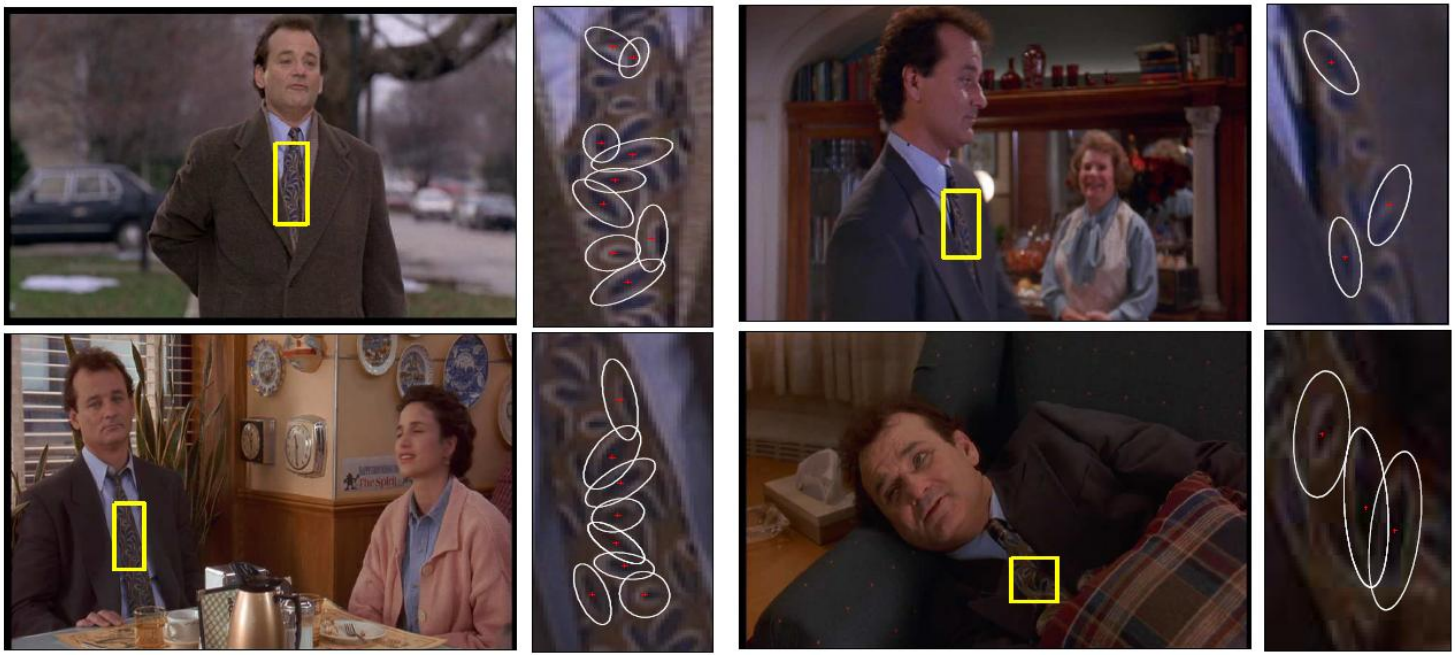
- Analogy of text retrieval
  - Immediate run-time object retrieval throughout a movie database
- Invariance to affine transformation
- Building a visual vocabulary
- Future Improvements
  - Current low rankings due to lack of visual descriptors for some scenes
  - “Upgrade” vocabulary for different scene types

Example object query

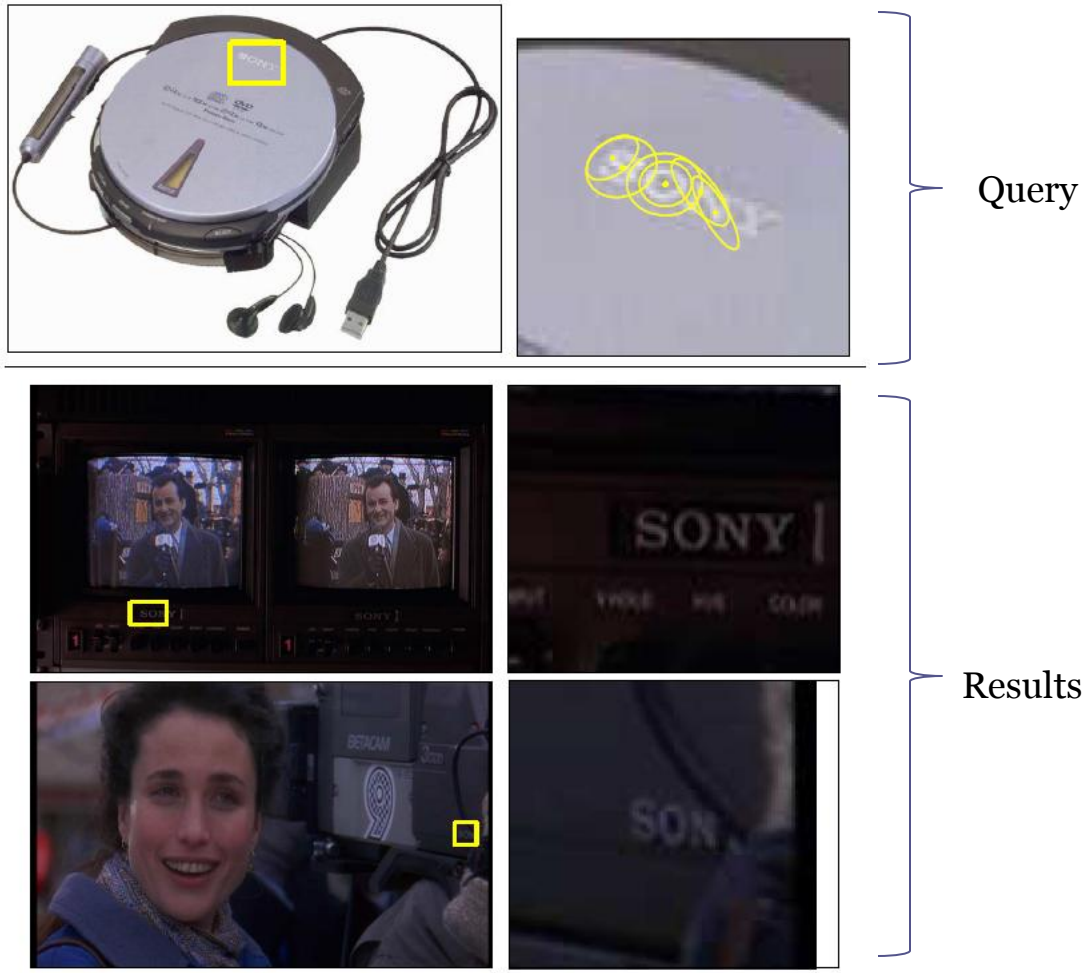


Query

Results



Example object query using external source



Automatic association of multiple aspects of a 3D object

Multiple Aspects of 3D Object taken from one shot

(a)

Query Region

Results

(b)