Video Google: Text Retrieval Approach to Object Matching in Videos

Authors: Josef Sivic and Andrew Zisserman
University of Oxford
ICCV 2003

Motivation

- Retrieve key frames and shots of video containing particular object with ease, speed and accuracy with which Google retrieves web pages containing particular words
- Investigate whether text retrieval approach is applicable to object recognition
- Visual analogy of word: vector quantizing descriptor vectors

Benefits

- Matches are pre-computed so at run time frames and shots containing particular object can be retrieved with no delay

- Any object (or conjunction of objects) occurring in a video can be retrieved even though there was no explicit interest in the object when the descriptors were built

Text Retrieval Approach

- Documents are parsed into words
- Words represented by stems
- Stop list to reject common words
- Remaining words assigned unique identifier
- Document represented by vector of weighted frequency of words
- Vectors organized in inverted files
- Retrieval returns documents with closest (angle) vector to query

Data Structure for Rapid Document Retrieval

color	doc. 1, doc. 2, doc 27, …
edge finding	doc. 12, doc. 24, doc 37, …
interest operators	doc. 1, doc. 2, doc 27, …
line finders	doc. 17, doc. 24, doc 37, …
motion	doc. 1, doc. 28, doc 27, …
object recognition	doc. 1, doc. 2, doc 43, …
relational matching	doc. 15, doc. 43, doc 92, …
segmentation	doc. 13, doc. 25, doc 27, …
texture	doc. 12, doc. 25, doc 47, …

inverted index

Viewpoint invariant description

- Two types of viewpoint covariant regions computed for each frame
  - Shape Adapted (SA) Mikolajczyk & Schmid
  - Maximally Stable (MS) Matas *et al.*
- Detect different image areas
- Provide complimentary representations of frame
- Computed at twice originally detected region size to be more discriminating

Shape Adapted Regions: the Harris-Affine Operator

- Elliptical shape adaptation about interest point
- Iteratively determine ellipse center, scale and shape
- Scale determined by local extremum (across scale) of Laplacian
- Shape determined by maximizing intensity gradient isotropy over elliptical region
- Centered on corner-like features

Examples of Harris-Affine Operator

140 K. Mikolajczyk and C. Schmid

(a) b)

Fig. 6. (a) Example of a 3D scene observed from significantly different viewpoints. There are 14 inliers to a robustly estimated fundamental matrix, all of them correct. (b) An image pairs for which our method fails. There exist, however, corresponding points which we have selected manually.

Maximally Stable Regions

- Use intensity watershed image segmentation
- Select areas that are approximately stationary as intensity threshold is varied
- Correspond to blobs of high contrast with respect to surroundings

Defs. From the Matas Paper

Idea: think of a stack of binary images corresponding to the 256 possible thresholds of a gray tone image.

Region $\mathcal{Q}$ is a contiguous subset of $\mathcal{D}$, i.e. for each $p, q \in \mathcal{Q}$ there is a sequence $p, a_1, a_2, \ldots, a_n, q$ and $pAa_1, a_iAa_{i+1}, a_nAq$.

(Outer) Region Boundary $\partial\mathcal{Q} = \{q \in \mathcal{D} \setminus \mathcal{Q} : \exists p \in \mathcal{Q} : qAp\}$, i.e. the boundary $\partial\mathcal{Q}$ of $\mathcal{Q}$ is the set of pixels being adjacent to at least one pixel of $\mathcal{Q}$ but not belonging to $\mathcal{Q}$.

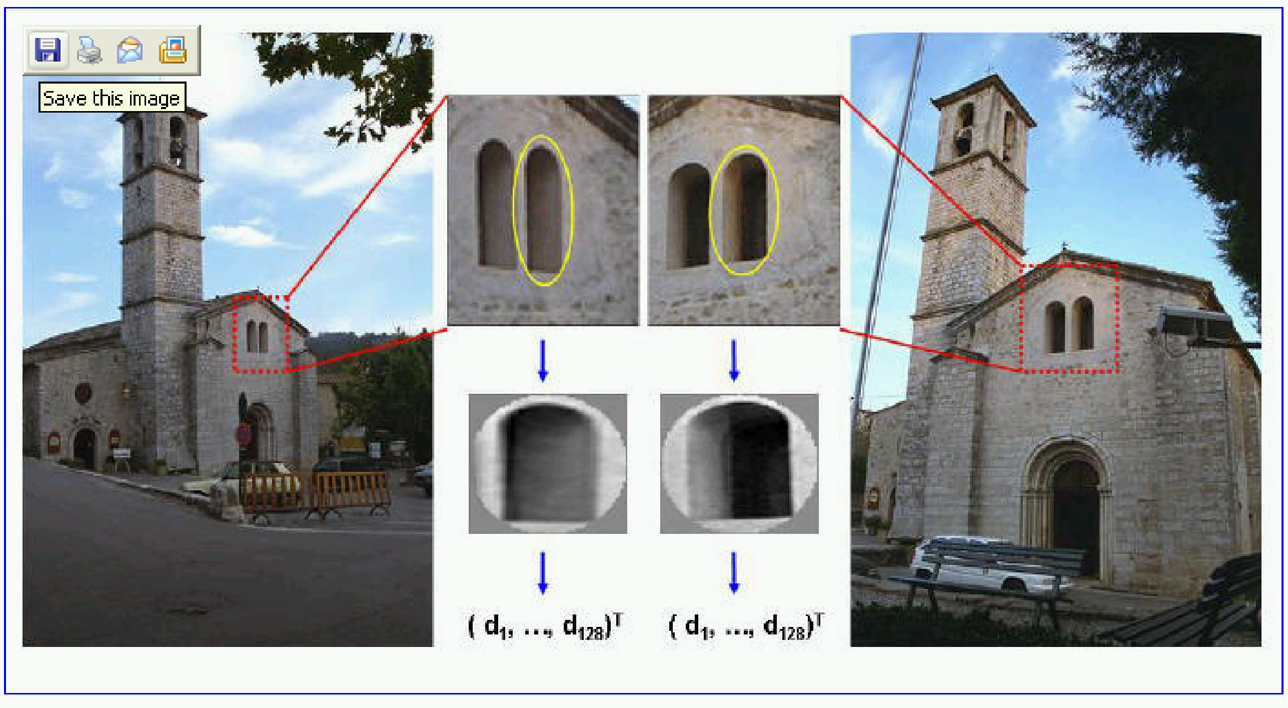
Extremal Region $\mathcal{Q} \subset D$ is a region such that for all $p \in \mathcal{Q}, q \in \partial\mathcal{Q} : I(p) > I(q)$ (maximum intensity region) or $I(p) < I(q)$ (minimum intensity region).

Maximally Stable Extremal Region (MSER). Let $\mathcal{Q}_1, \ldots, \mathcal{Q}_{i-1}, \mathcal{Q}_i, \ldots$ be a sequence of nested extremal regions, i.e. $\mathcal{Q}_i \subset \mathcal{Q}_{i+1}$. Extremal region $\mathcal{Q}_{i*}$ is maximally stable iff $q(i) = |\mathcal{Q}_{i+\Delta} \setminus \mathcal{Q}_{i-\Delta}|/|\mathcal{Q}_i|$ has a local minimum at i^* ($|.|$ denotes cardinality). $\Delta \in \mathcal{S}$ is a parameter of the method.

Examples of Maximally Stable Regions

Feature Descriptor

- Each elliptical affine invariant region represented by 128 dimensional vector using SIFT descriptor



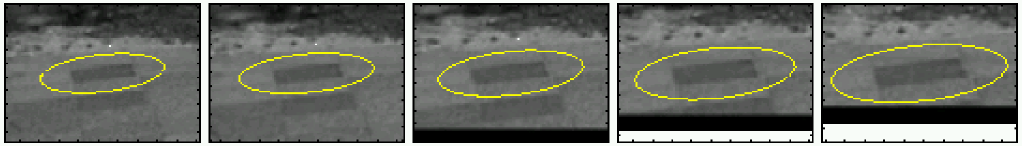
Noise Removal

- Information aggregated over sequence of frames
- Regions detected in each frame tracked using simple constant velocity dynamical model and correlation
- Region not surviving more than 3 frames are rejected
- Estimate descriptor for region computed by averaging descriptors throughout track

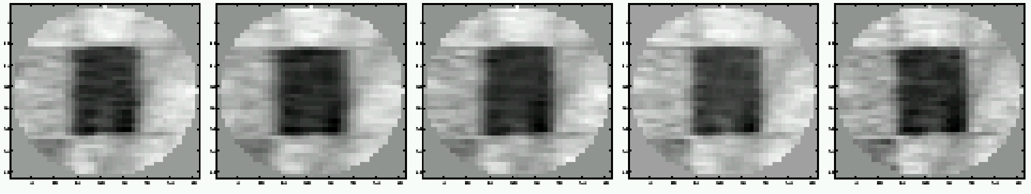
Noise Removal

- **Tracking region over 70 frames**

First (left) and last (right) frame of the track.



Close-up of the 1st, 20th, 40th, 55th, 70th frame.



Visual Vocabulary

- Goal: vector quantize descriptors into clusters (visual words)
- When a new frame is observed, the descriptor of the new frame is assigned to the nearest cluster, generating matches for all frames

Visual Vocabulary

- Implementation: K-Means clustering
- Regions tracked through contiguous frames and average description computed
- 10% of tracks with highest variance eliminated, leaving about 1000 regions per frame
- Subset of 48 shots (~10%) selected for clustering
- Distance function: Mahalanobis
- 6000 SA clusters and 10000 MS clusters

Visual Vocabulary

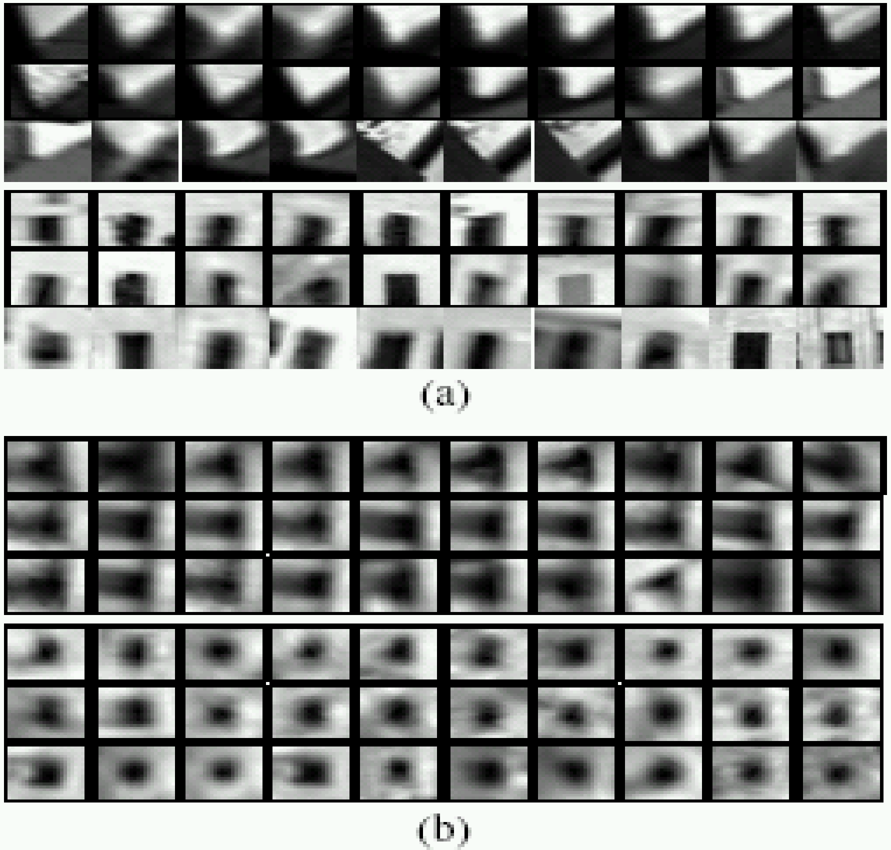


Figure 2: Samples from the clusters corresponding to a single visual word. (a) Two examples of clusters of Shape Adapted regions. (b) Two examples of clusters of Maximally Stable regions.

Visual Indexing

- Apply weighting to vector components
- Weighting: term frequency-inverse document frequency (tf-idf)
- Vocabulary k words, each doc represented by k-vector $V_d = (t_1,\ldots,t_i,\ldots,t_k)^T$ where

$$t_i = \underbrace{\frac{n_{id}}{n_d}}_{\text{term frequency}} \underbrace{\log \frac{N}{n_i}}_{\text{inverse document frequency}}$$

n_{id} = # of occurences of word i in doc d

n_d = total # of words in doc d

n_i = # of occurences of word i in db

N = # of doc in db

Experiments - Setup

- Goal: match scene locations within closed world of shots
- Data:164 frames from 48 shots taken at 19 different 3D locations; 4-9 frames from each location

Experiments - Retrieval

- Entire frame is query
- Each of 164 frames as query region in turn
- Correct retrieval: other frames which show same location
- Retrieval performance: average normalized rank of relevant images

$$\widetilde{Rank} = \frac{1}{NN_{rel}} \left(\sum_{i=1}^{N_{rel}} R_i - \frac{N_{rel}(N_{rel}+1)}{2} \right)$$

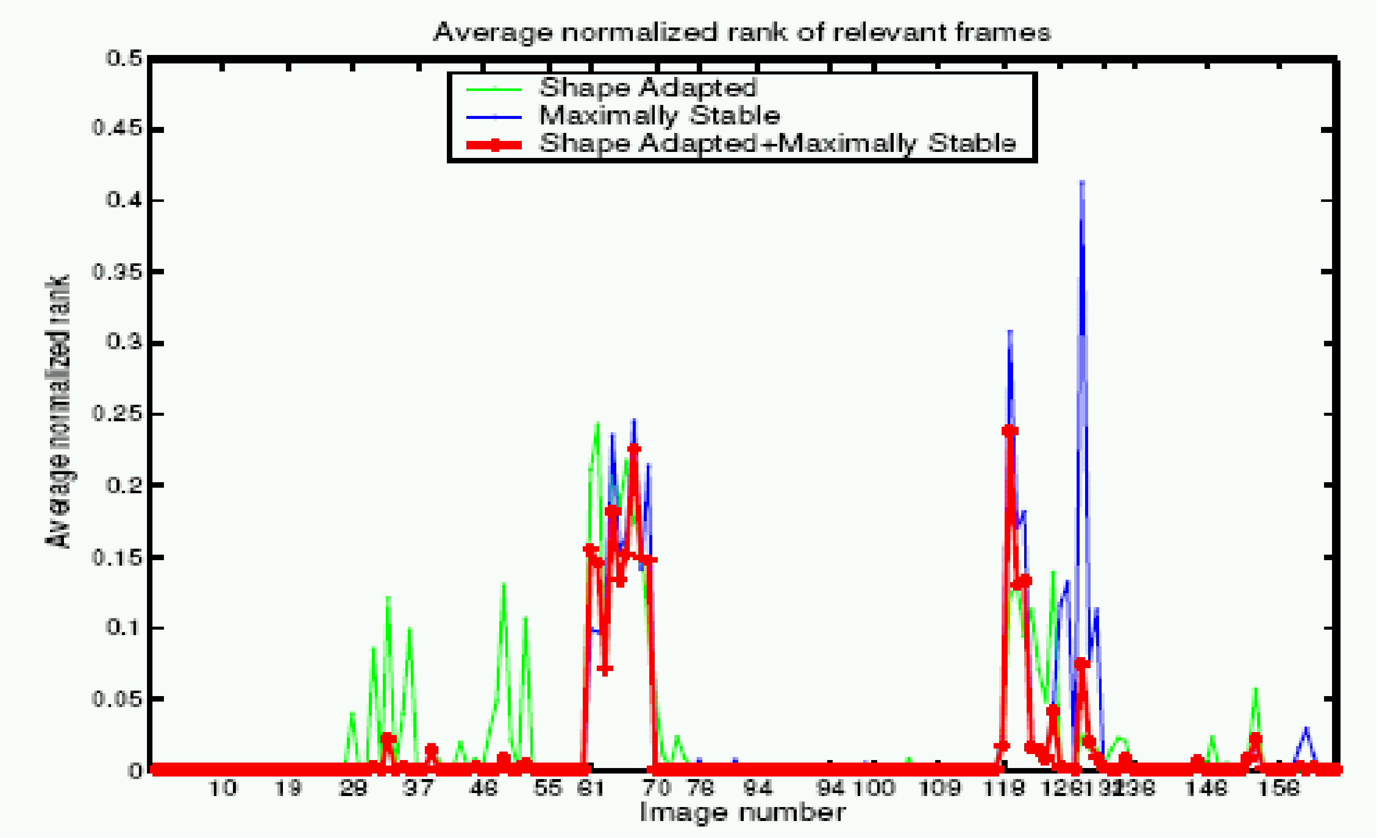
N_{rel} = # of relevant images for query image

N = size of image set

R_i = rank of ith relevant image

Rank lies between 0 and 1.
Intuitively, it will be 0 if all relevant images are returned ahead of any others.
It will be .5 for random retrievals.

Experiment - Results



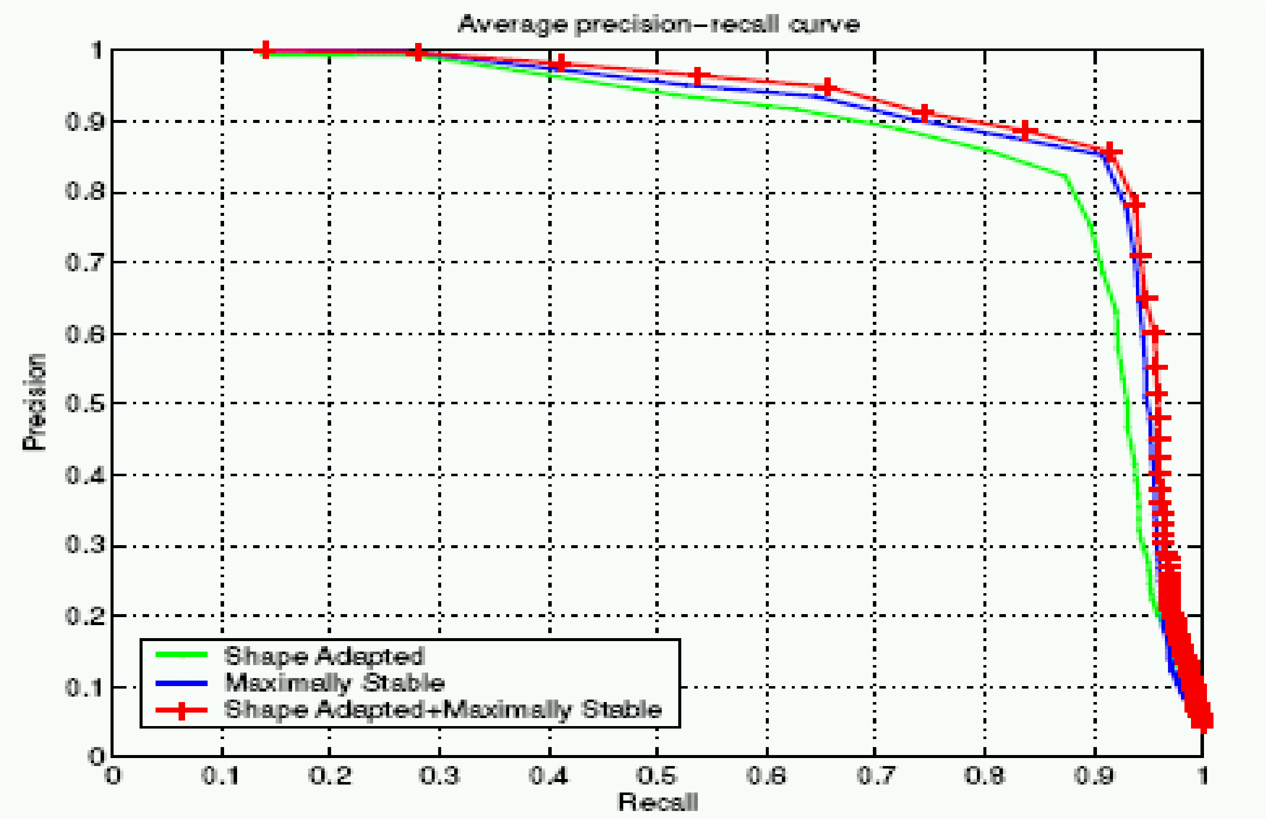
Zero is good!

Experiments - Results

	binary	*tf*	*tf-idf*
SA	0.0265	0.0275	0.0209
MS	0.0237	0.0208	0.0196
SA+MS	0.0165	0.0153	0.0132

Table 1: The mean of the $\widetilde{Rank}$ measure computed from all 164 images of the ground truth set for different term weighting methods.

Experiments - Results



Precision = # relevant images/total # of frames retrieved

Recall = # correctly retrieved frames/ # relevant frames

Stop List

- Top 5% and bottom 10% of frequent words are stopped

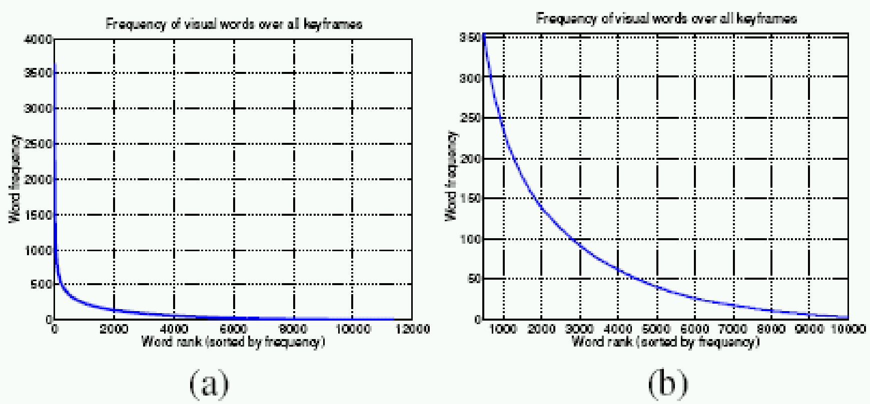


Figure 5: Frequency of MS visual words among all 3768 keyframes of Run Lola Run (a) before, and (b) after, application of a stoplist.

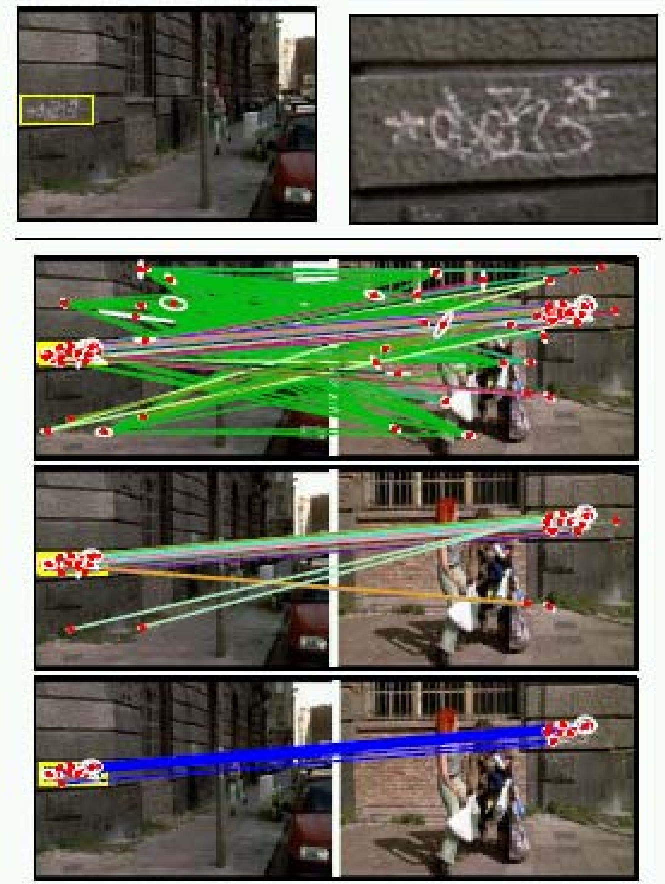


Figure 6: Matching stages. Top row: (left) Query region and (right) its close-up. Second row: Original word matches. Third row: matches after using stop-list, Last row: Final set of matches after filtering on spatial consistency.

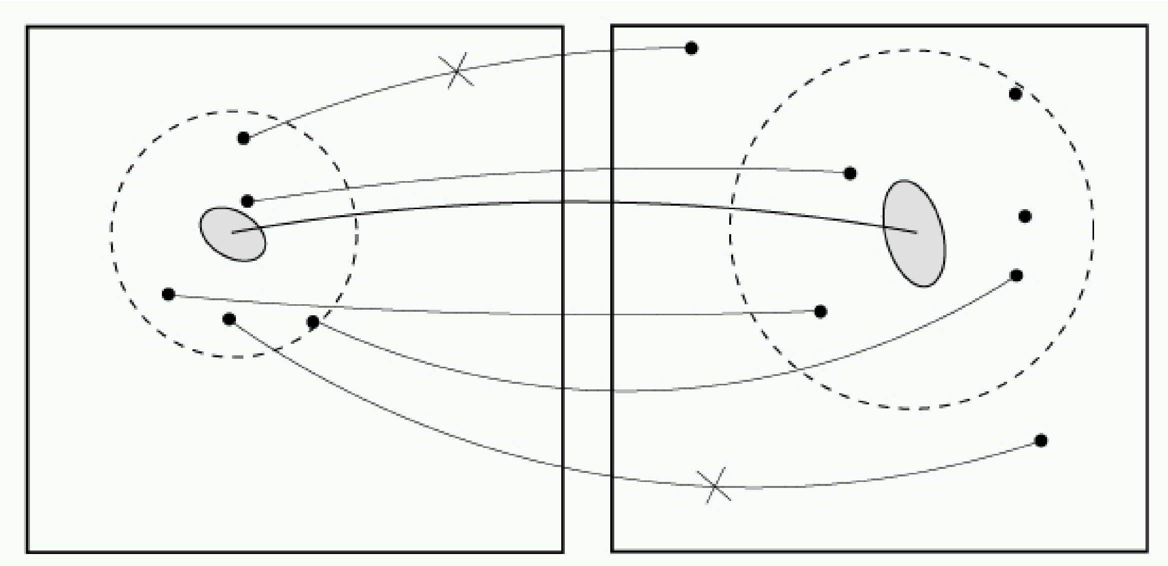
Spatial Consistency

- Matched region in retrieved frames have similar spatial arrangement to outlined region in query

- Retrieve frames using weighted frequency vector and re-rank based on spatial consistency

Spatial Consistency

- Search area of 15 nearest neighbors of each match cast a vote for the frame
- Matches with no support are rejected
- Total number of votes determine rank

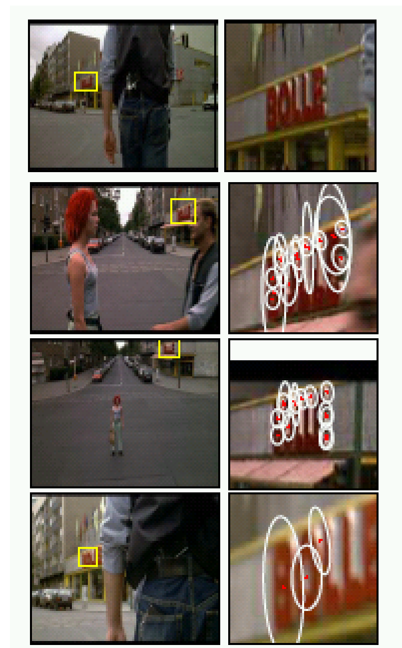


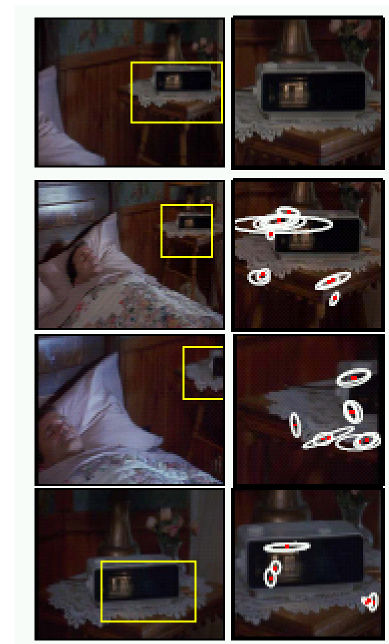
circular areas are defined by the fifth nearest neighbour and the number of votes cast by the match is three.

Inverted File

- Entry for each visual word
- Store all matches : occurences of same word in all frames

More Results





Demo

- http://www.robots.ox.ac.uk/~vgg/research/vgoogle/how/method/method_a.html
- http://www.robots.ox.ac.uk/~vgg/research/vgoogle/index.html