

Instance Recognition

CS 6980: Visual Recognition

Course Outline

- Introduction

- Exact instance retrieval
- Classification
- Detection
- Segmentation

Deep or traditional
learning
based

- Weak Supervision
- Active Learning
- Domain Adaptation
- Unsupervised Representation learning
- Dynamic Temporal Aspects

Tentative set
of
advanced topics

Course Outline

- Introduction

- Exact instance retrieval
- Classification
- Detection
- Segmentation

Traditional
learning
based

- Weak Supervision
- Active Learning
- Domain Adaptation
- Unsupervised Representation learning
- Dynamic Temporal Aspects

Tentative set
of
advanced topics

Course Outline

- Introduction

- **Exact instance retrieval**

- Classification

- Detection

- Segmentation

Traditional
learning
based

- Weak Supervision

- Active Learning

- Domain Adaptation

- Unsupervised Representation learning

- Dynamic Temporal Aspects

Tentative set
of
advanced topics

Problem

- Given a bounding box in an image, extract similar regions in the image for a database of images
- Analogy: Given a term or set of terms, look up and retrieve pages that are most relevant for the term
- Assumption: The bounding box can be found without much variation in the database

Example

videogoogle

Exploring **Charade**

Viewing frame 106725

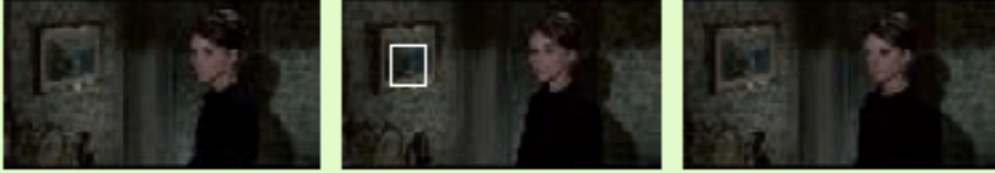

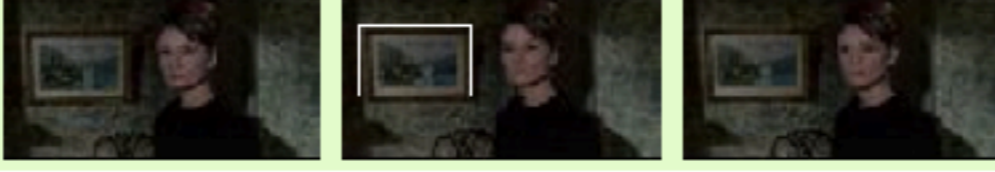
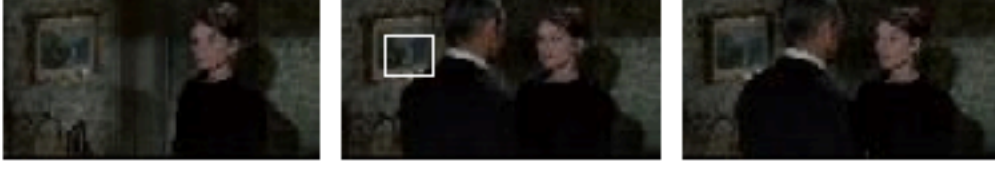
Overview Explore shots
Prev Animate DivX Stream Thumbnails Search Next



Video Google: A Text Retrieval Approach to Object Matching in Videos
Josef Sivic and Andrew Zisserman
ICCV 2003

link to demo: <http://www.robots.ox.ac.uk/~vgg/research/vgoogle/index.html>

Example

Shot 469 Relevance: 9.22 Frames 59347 to 59480		Animate DivX Stream Thumbnails Search
Shot 1019 Relevance: 8.18 Frames 139581 to 139706		Animate DivX Stream Thumbnails Search
Shot 1011 Relevance: 7.27 Frames 138450 to 138744		Animate DivX Stream Thumbnails Search
Shot 1013 Relevance: 7.26 Frames 138775 to 139023		Animate DivX Stream Thumbnails Search
Shot 477 Relevance: 6.29 Frames 60157 to 60220		Animate DivX Stream Thumbnails Search
Shot 471 Relevance: 6.26 Frames 59528 to 59658		Animate DivX Stream Thumbnails Search

Video Google: A Text Retrieval Approach to Object Matching in Videos
Josef Sivic and Andrew Zisserman
ICCV 2003

Approach

- Text retrieval systems
- Documents are parsed into words
- Words are stemmed
- Stored in an inverted file index
- Documents are matched using TF-IDF score

Example

- The **advances** in **image recognition** extend far beyond cool social apps. **Medical startups** claim they'll soon be able to use computers to read X-rays, MRIs, and CT scans more **rapidly** and **accurately** than **radiologists**, to **diagnose cancer** earlier and less invasively, and to accelerate the search for life-saving **pharmaceuticals**. Better image recognition is crucial to unleashing **improvements** in **robotics**, **autonomous drones**, and, of course, **self-driving cars**—a development so momentous that we made it a cover story in June

Example

- The **advance** in **image recognition** extend far beyond cool social apps. **Medical startup** claim they'll soon be able to use computers to read X-rays, MRIs, and CT scans more **rapid** and **accurate** than **radiologists**, to **diagnose cancer** earlier and less invasively, and to accelerate the search for life-saving **pharmaceutical**. Better image recognition is crucial to unleashing **improve** in **robotics**, **autonomy drone**, and, of course, **self-driving car**—a development so momentous that we made it a cover story in June

Example

advance image recognition

Medical startup

rapid accurate radiologists diagnose cancer

pharmaceutical

improve robotics autonomy drone self-driving car

Words - Visual Words?



One option - Segmentation?



One option - Segmentation?

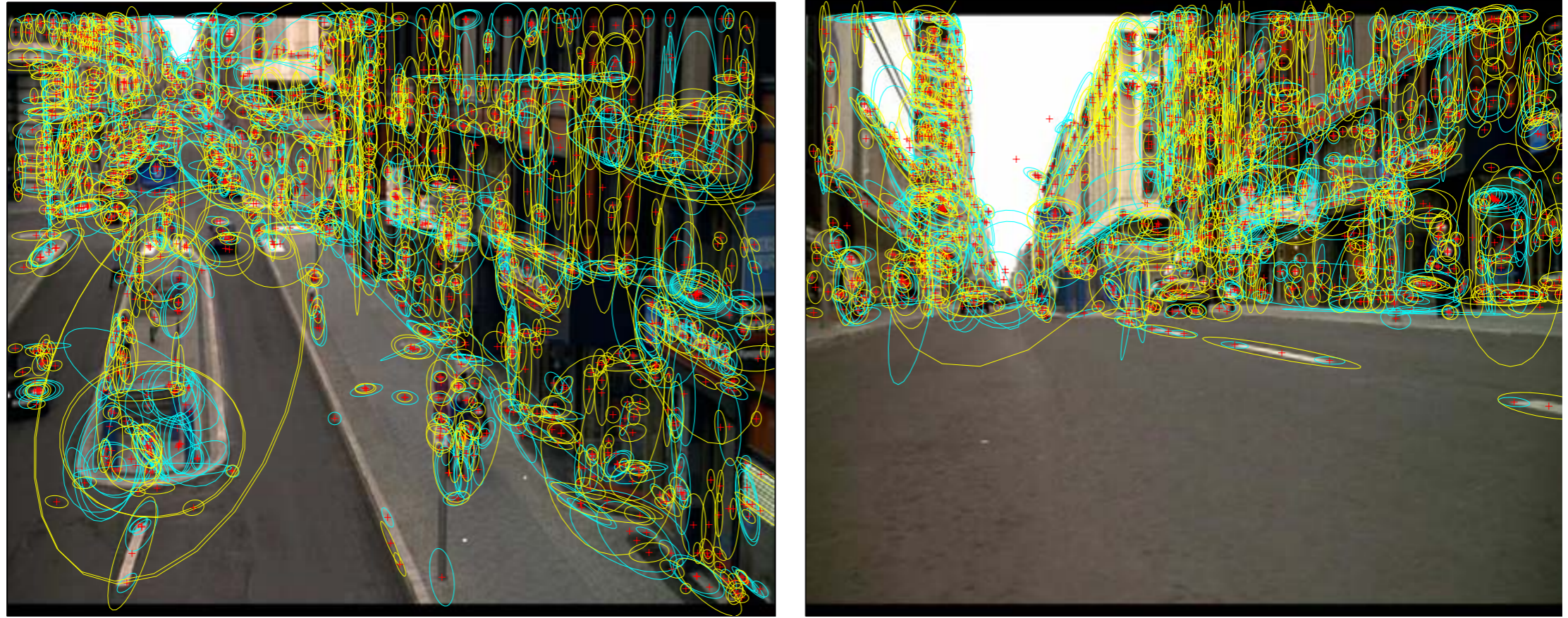


One option - Segmentation?



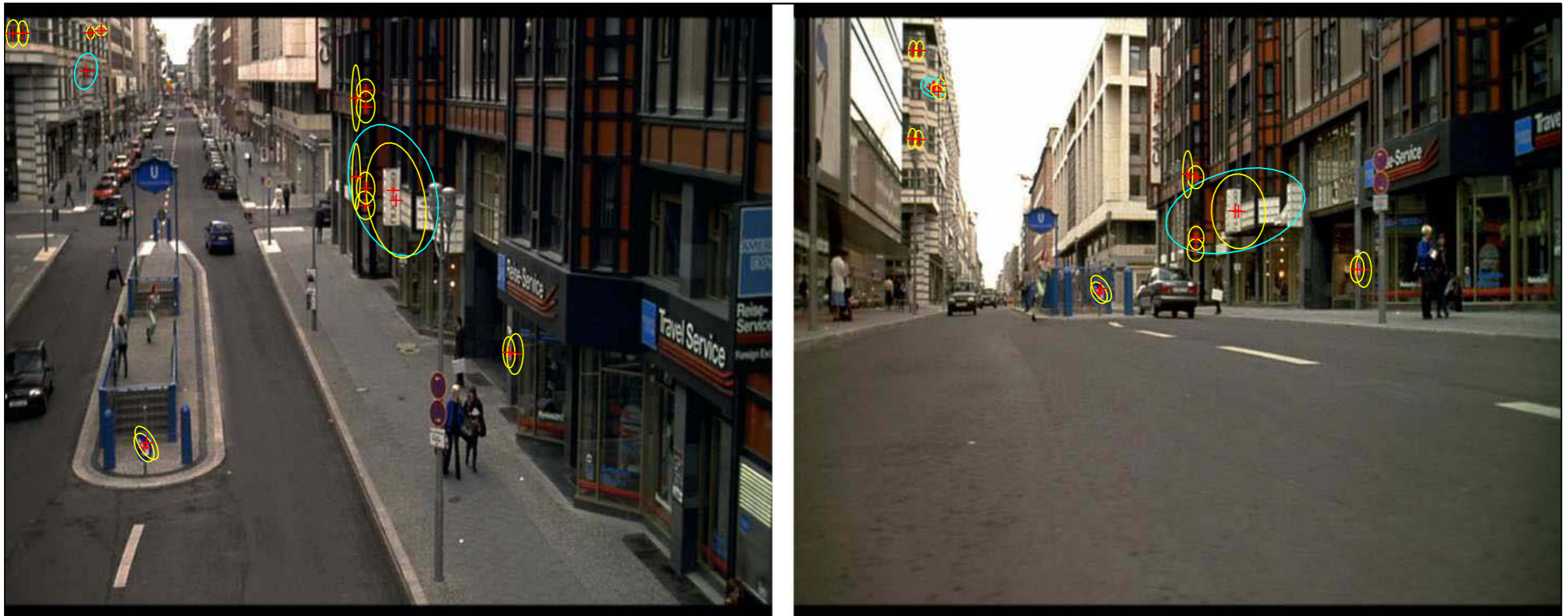
Will not create repeatable segments that can be matched

Words - Visual Words?



Solution proposed: Very local patches that can be well represented. Will see more about these in next class

Words - Visual Words?



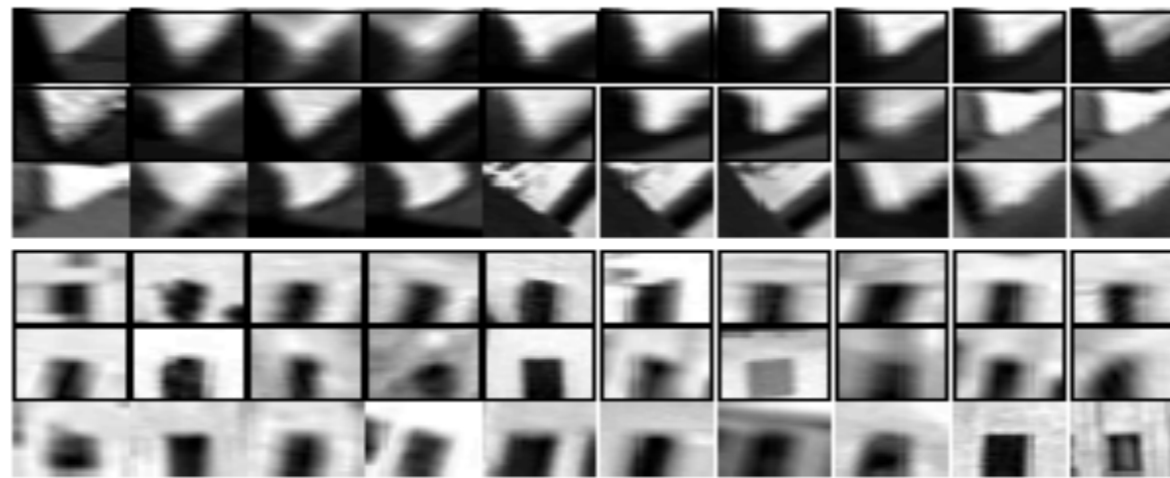
Visual words obtained by local interest operators (MSER and SA) that are described using SIFT

Stemming?

- Centroids - obtained by clustering visual words and using centroid for representation

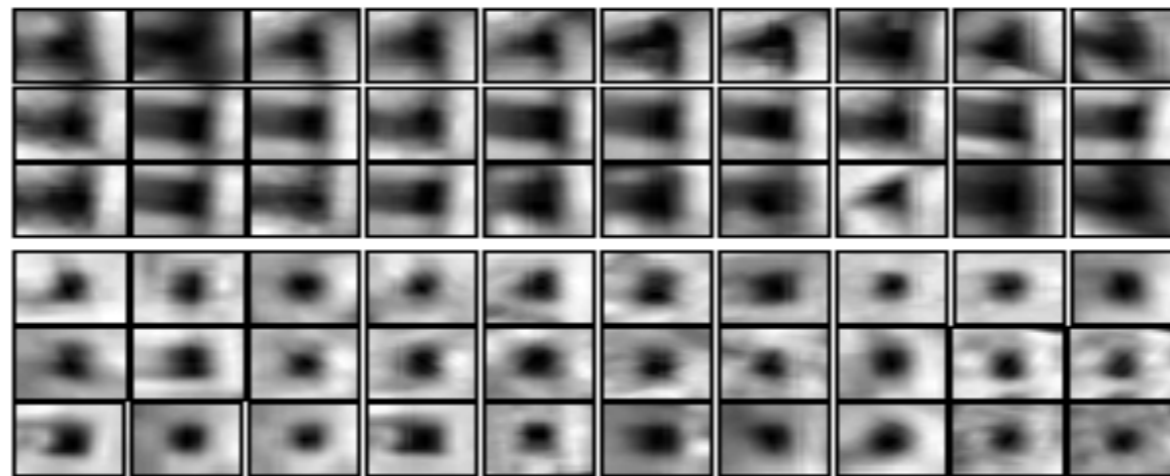
MS and SA “Visual Words”

SA



(a)

MS

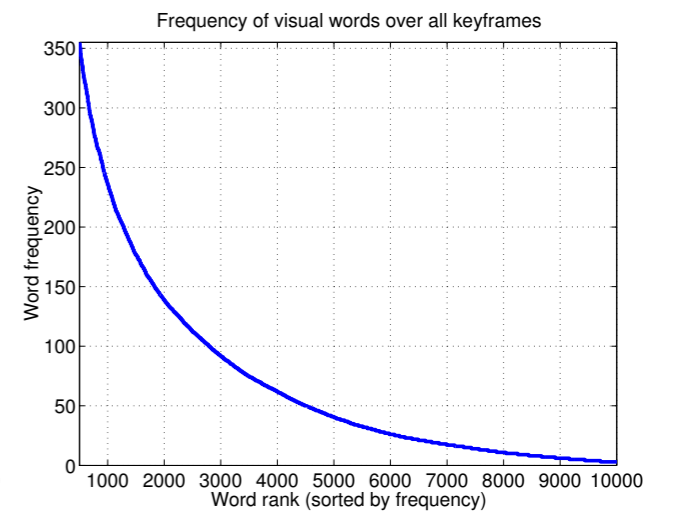
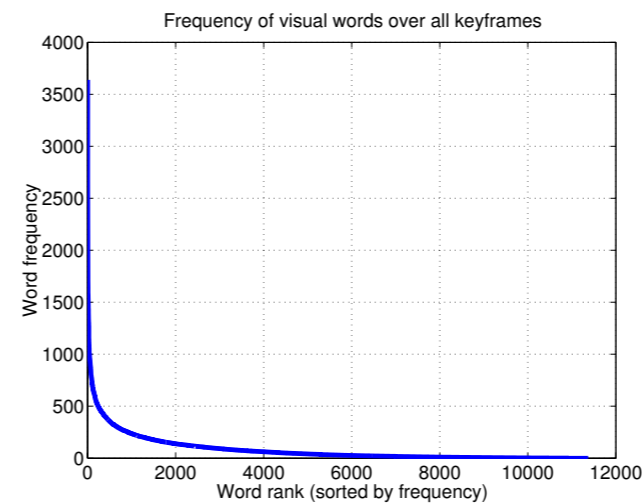


(b)

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.

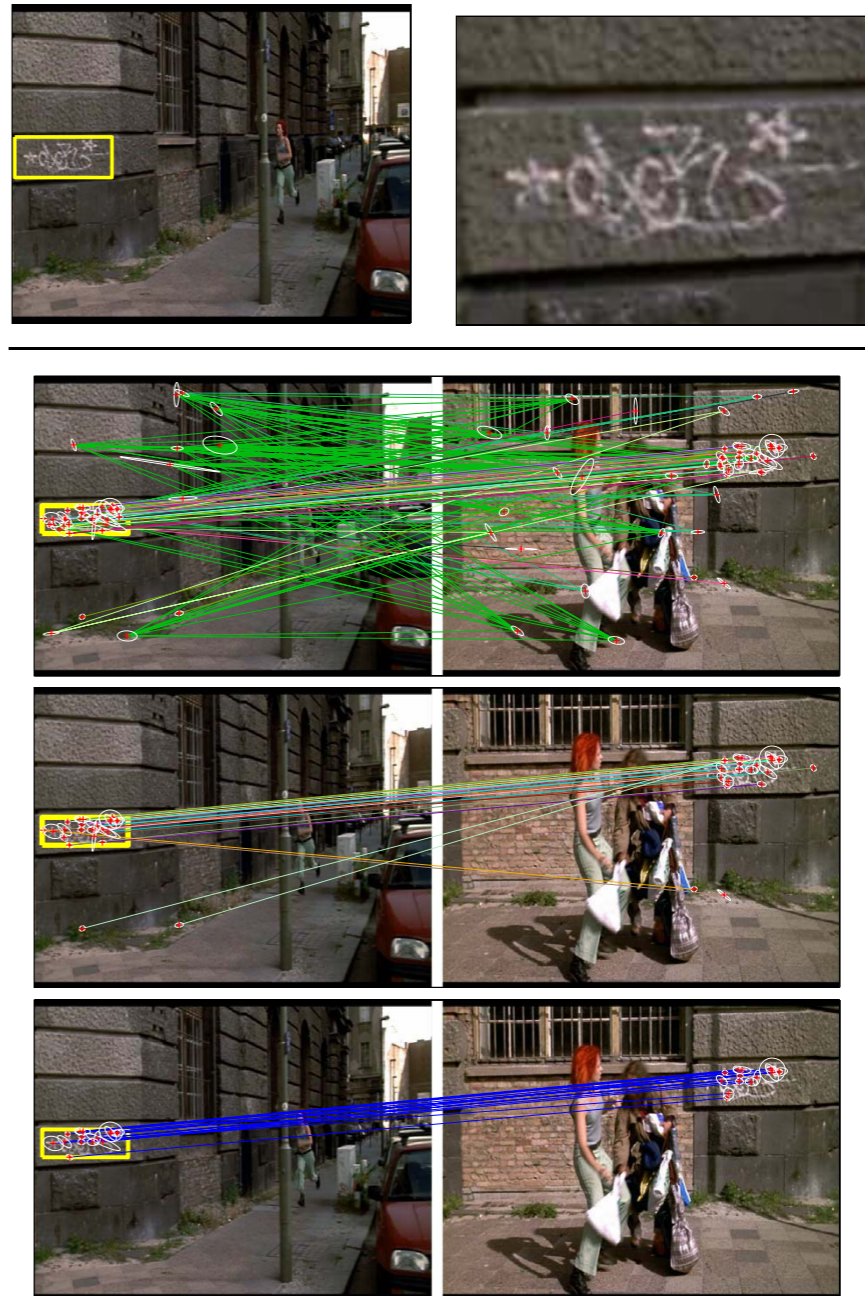
Stop words

- In text retrieval, in order stop words (most common words such as and, or etc) are removed
- Similar strategy is used in VideoGoogle



Spatial consistency

- The matches are scored for being spatially consistent
- More matches in a neighbourhood indicate high likelihood of a correct match
- Obtained by considering a region from 15 nearest neighbour matches, more matches in the region increase the support for the match



Scoring of words

- Words are scored using TF-IDF
- Each document is represented by a k terms $t_1 \dots t_k$

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

- where n_{id} is the number of occurrences of word i in document d
- and n_d is the number of words in document d .
- n_i is the number of term i in the whole database
- and N is the number of documents in the database.

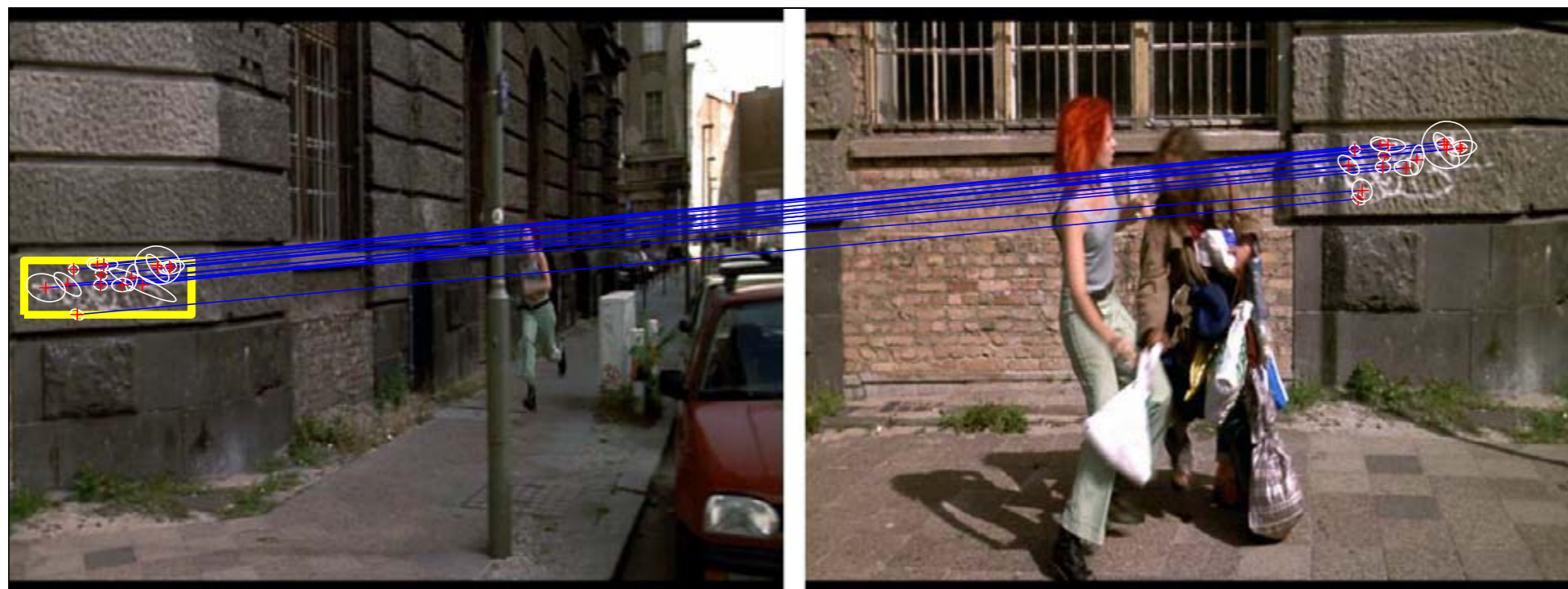
Example



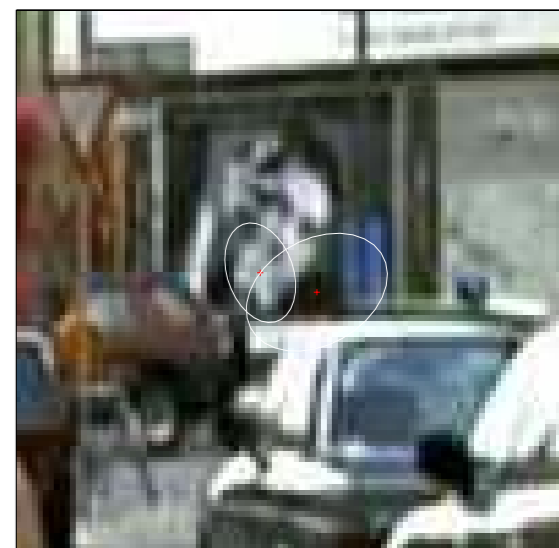
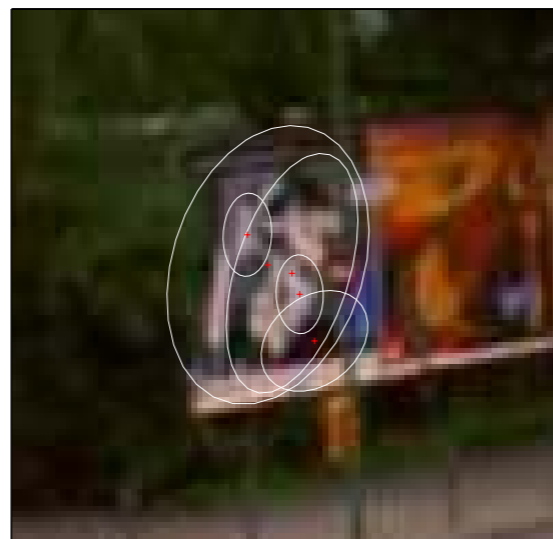
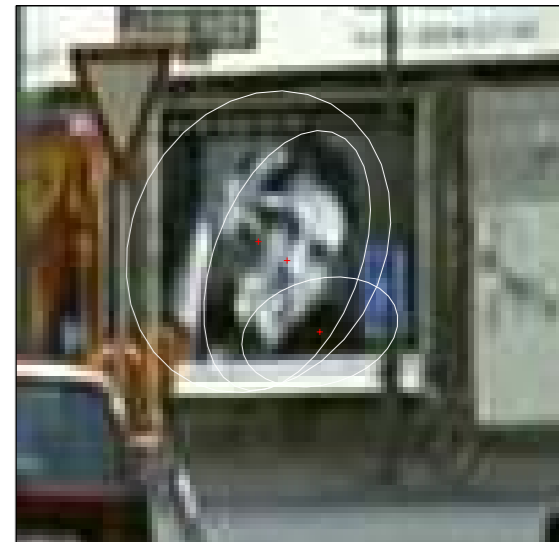
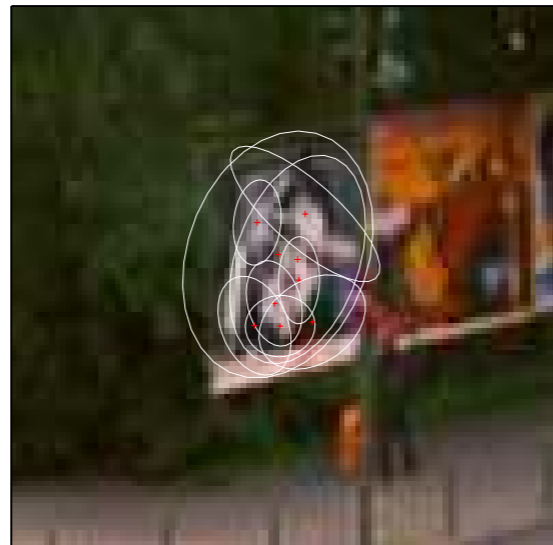
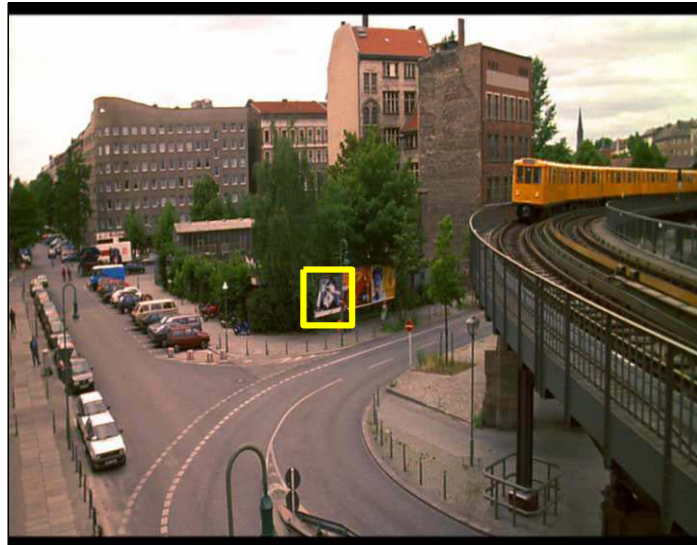
Example



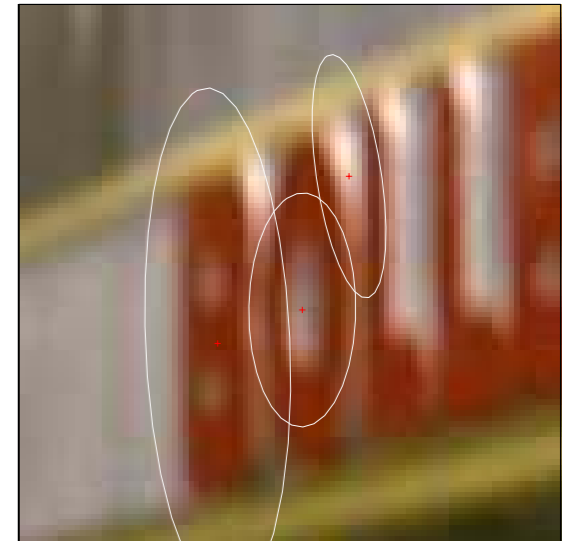
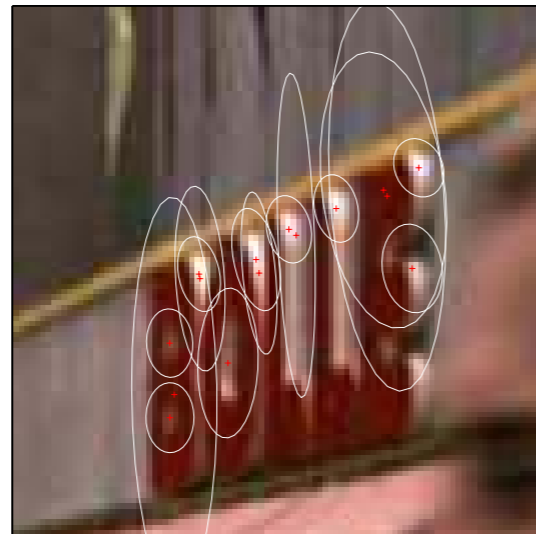
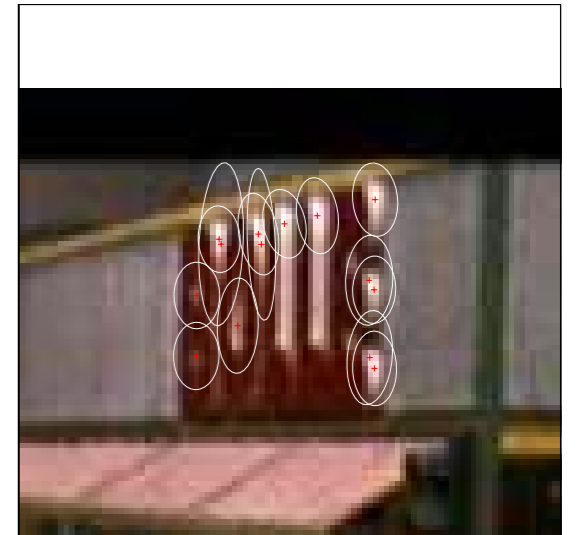
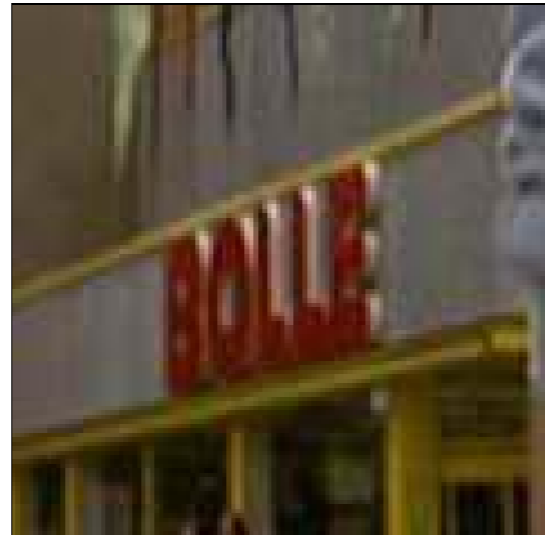
Example



Example



Example



Conclusion

- Local features enable us to obtain word analogues for representing images
- Exact instance recognition can be obtained by using a text retrieval approach to match sets of local features
- Limitations: this particular approach (using only features) cannot generalise to match categories unless

Next class

- A brief overview of local feature representation