Instance Recognition

CS 783: Visual Recognition

Course Outline

- Introduction
- Exact instance retrieval
- Classification
- Detection
- Segmentation
- Weak Supervision
- Active Learning
- Domain Adaptation
- Unsupervised Representation learning
- Vision and Language

Traditional learning based

Tentative set of advanced topics

Course Outline

- Introduction
- **Exact instance retrieval**
- Classification
- Detection
- Segmentation

Traditional learning based

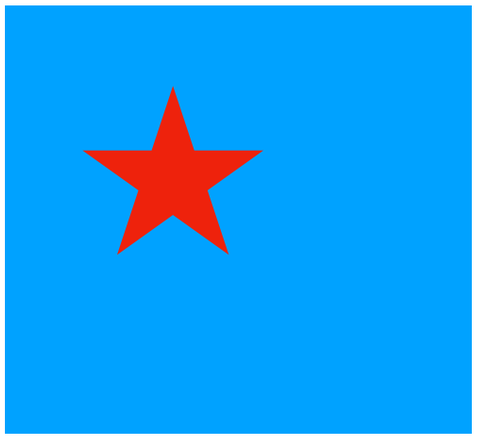
- Weak Supervision
- Active Learning
- Domain Adaptation
- Unsupervised Representation learning
- Vision and Language

Tentative set of advanced topics

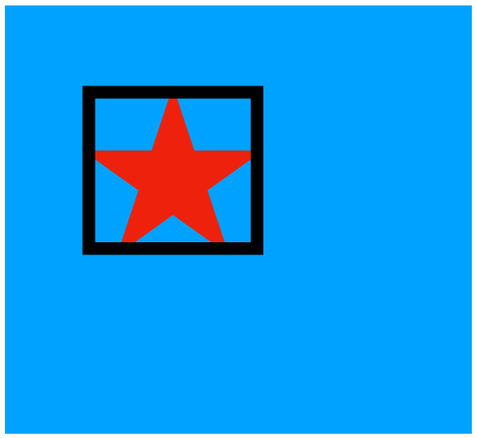
Problem

- We are given a database of images that contain various objects at different locations.
- We are given a query image with a specific bounding box B (Left, Top, Right, Bottom)
- Retrieve all instances of images that contain the bounding box B.
- Assumption: The bounding box can be found without much variation in the database

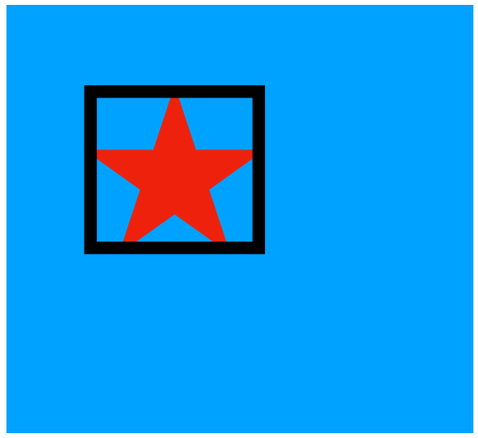
Problem

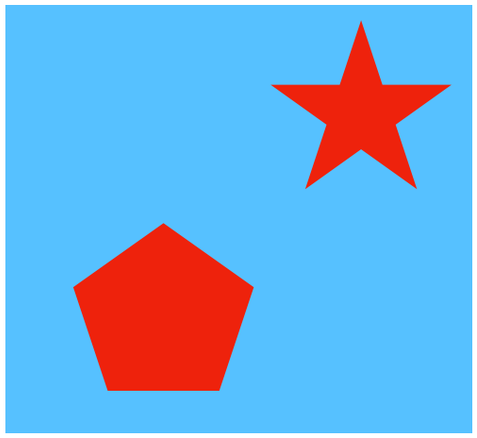


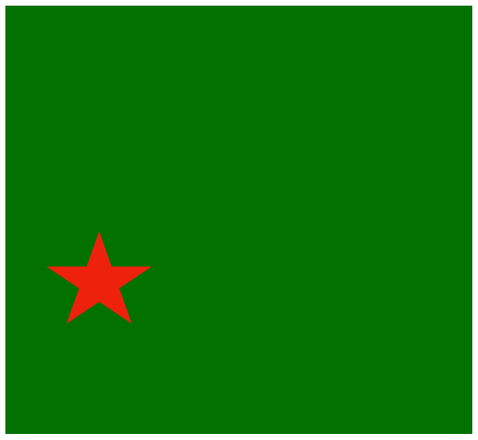
Problem

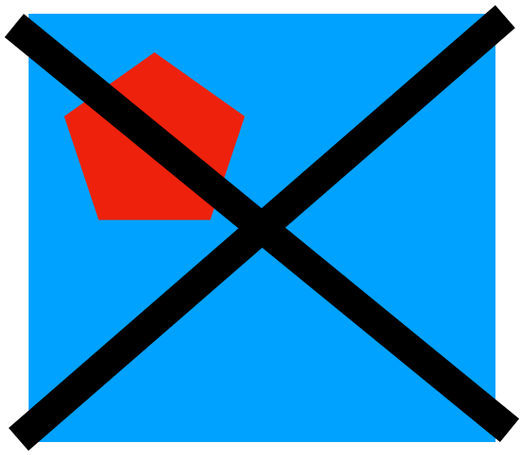


Problem





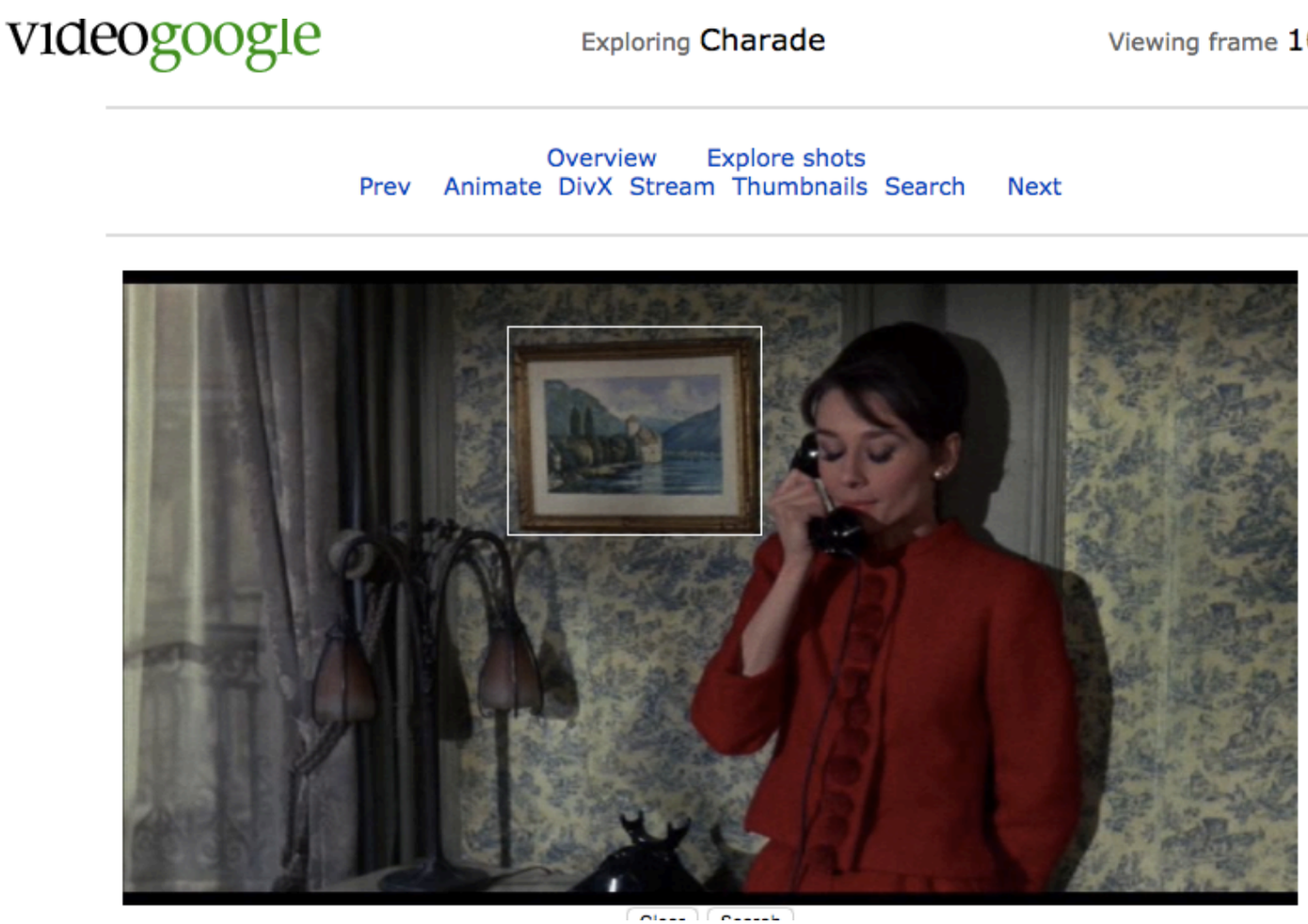




Problem

- Problem is different from image retrieval as done on web search engines as the query in this case is part of an image
- It is termed content based image retrieval
- In usual content based image retrieval, full images are used. In this case part of an image is considered

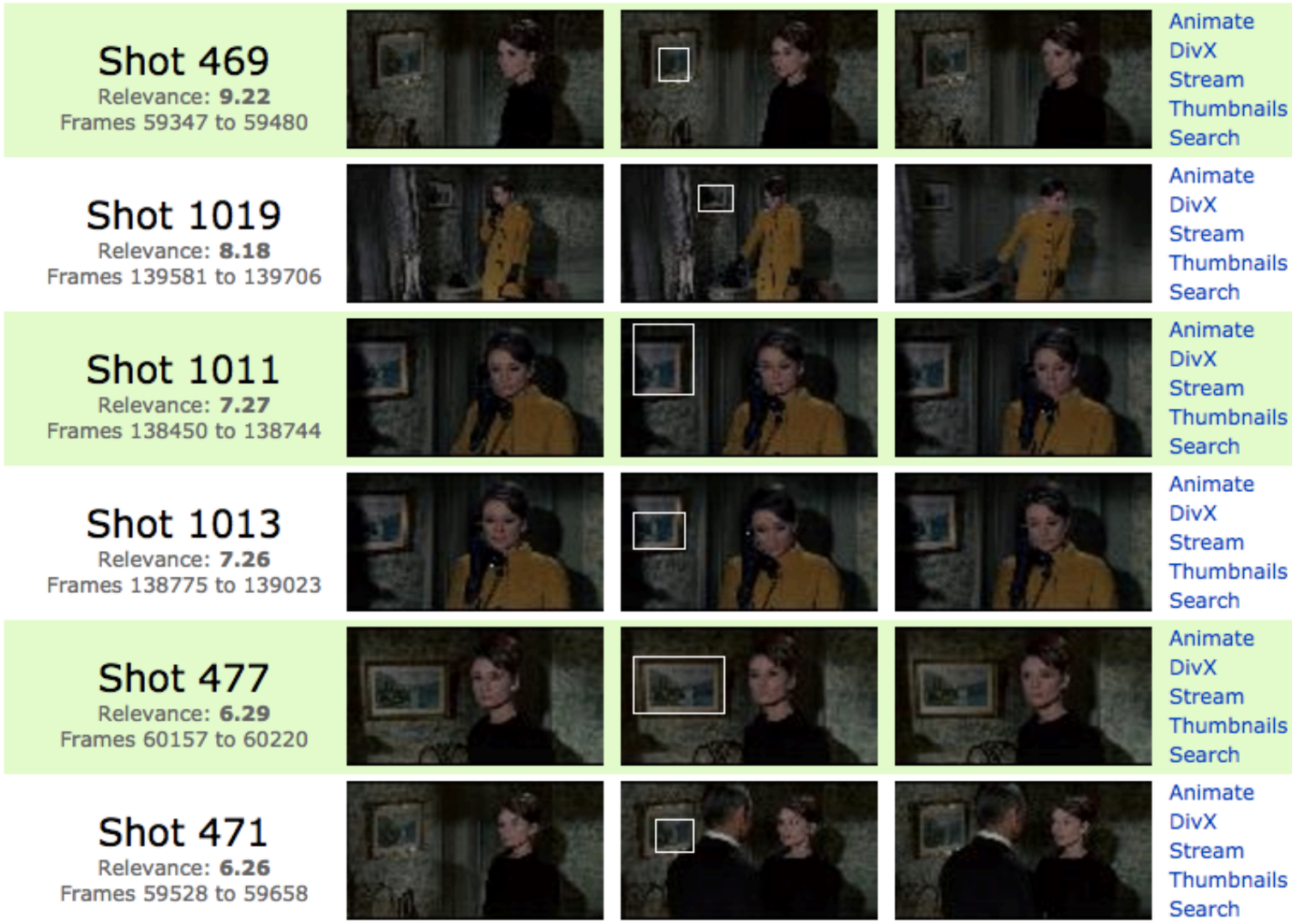
Example



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



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

Naive Approach

12.34
234.45
22.45
11.55
10.95
9.56
2.34
1.34
5.67
8.76
8.88

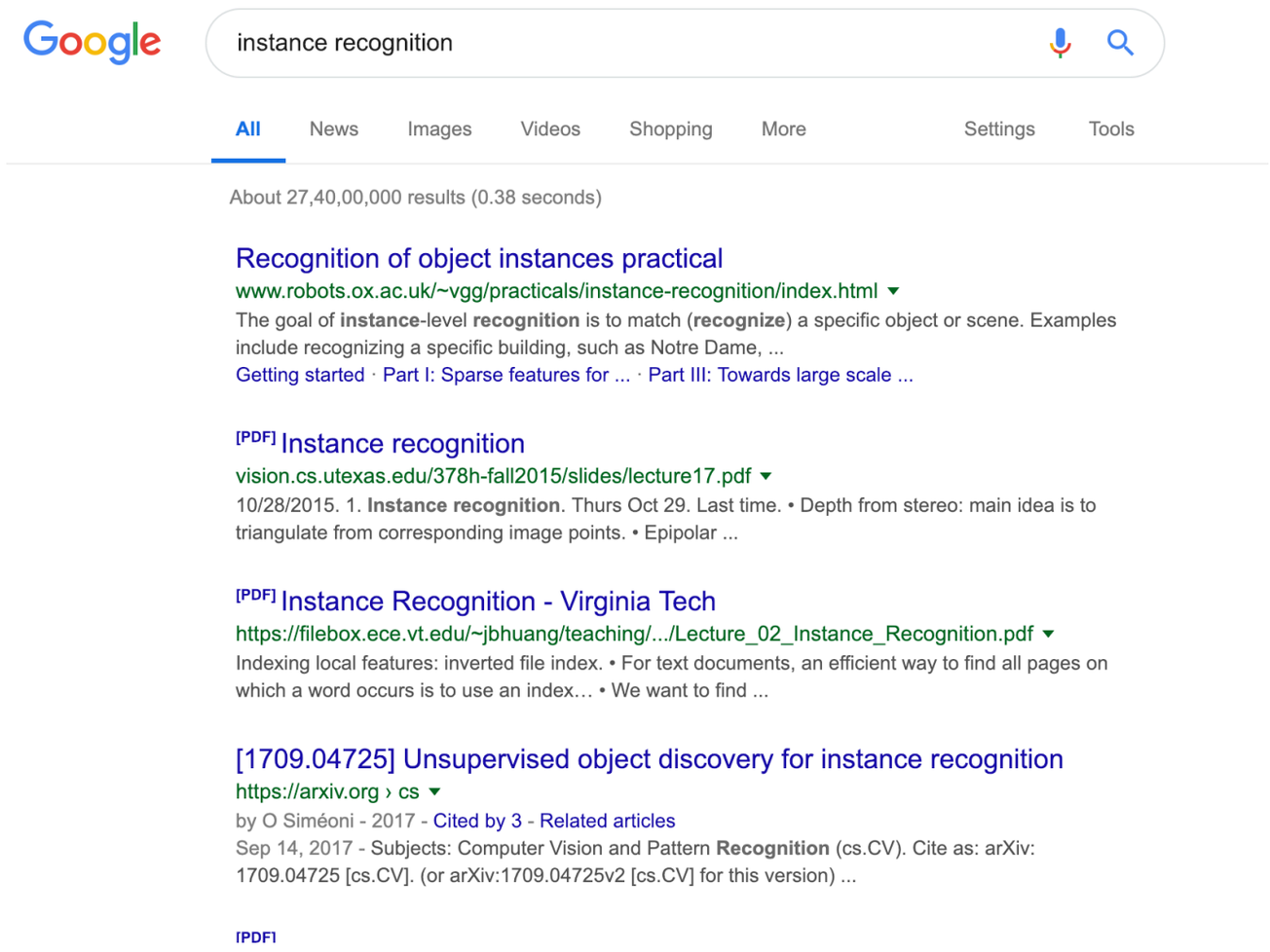
A

10.55
23.2
11.11
11.11
23.45
22.45
11.55
10.95
9.56
2.34
1.34
5.67
8.76
8.88

DA

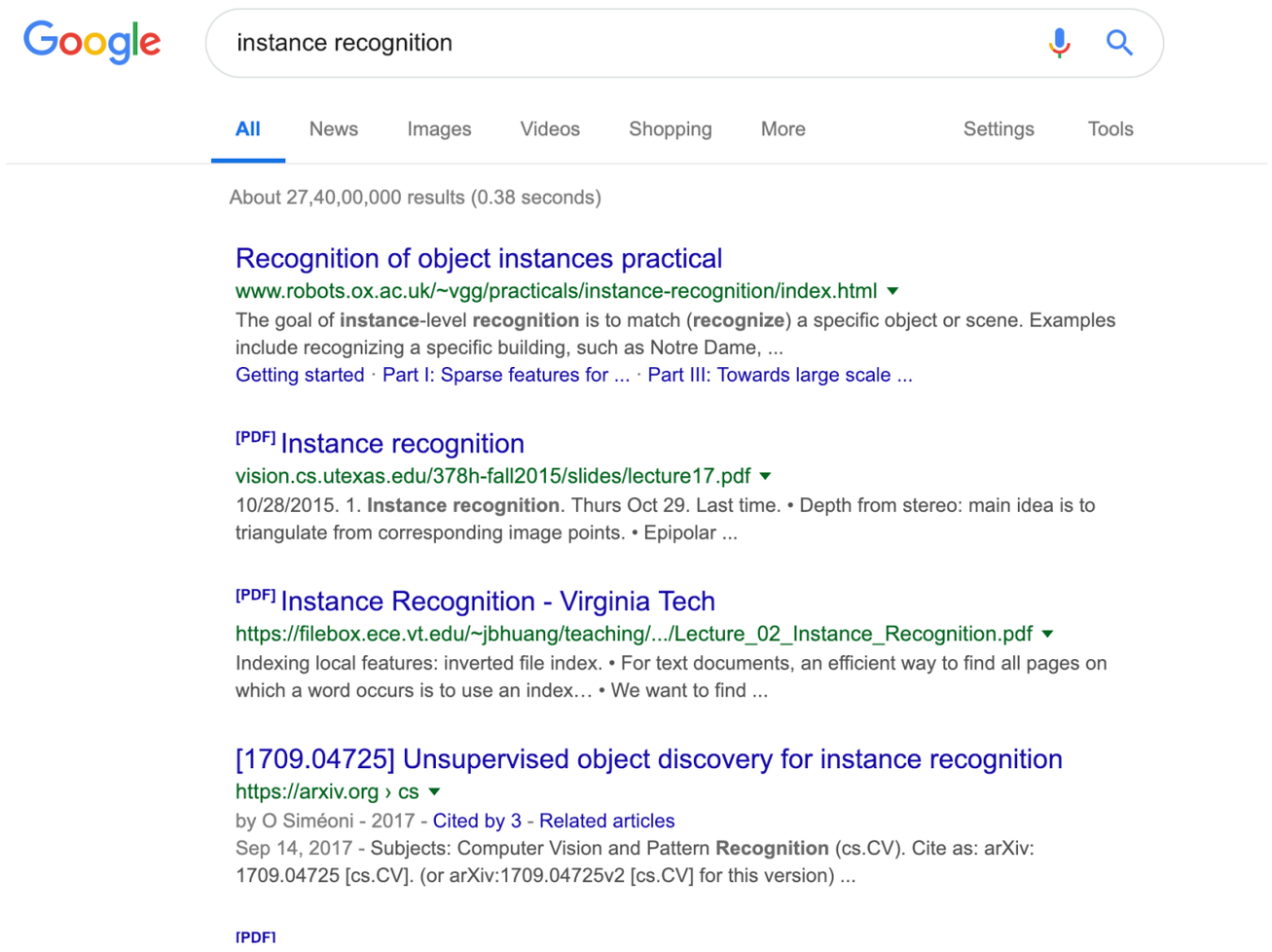
Naive approach based on matching pixel values will not work…, Why?

Another approach - Text Retrieval



In text retrieval we have examples of searching documents based on terms

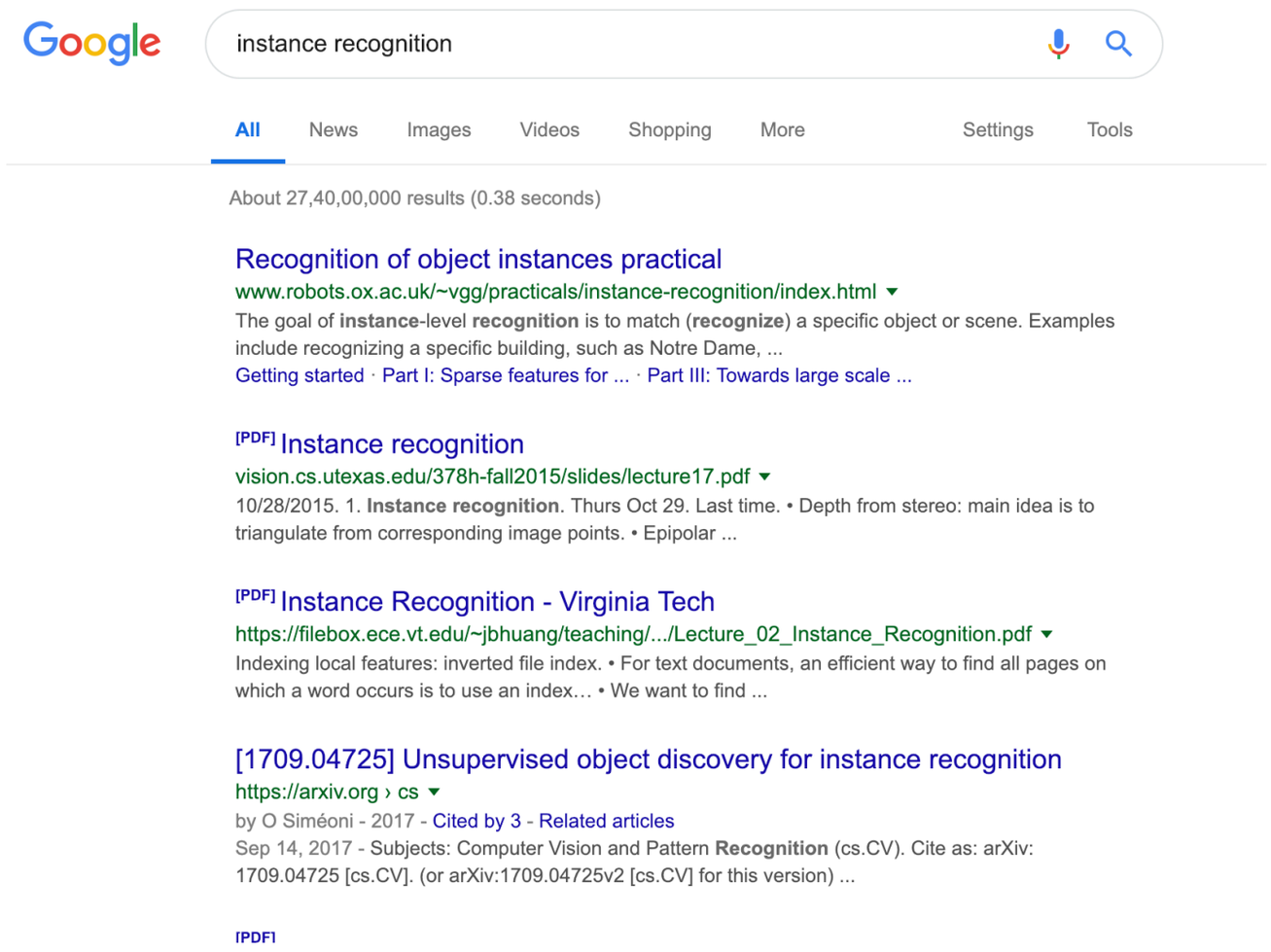
Another approach - Text Retrieval



In text retrieval we have examples of searching documents based on terms

Can we approach the problem in that manner?

Another approach - Text Retrieval



In text retrieval we have examples of searching documents based on terms

Can we approach the problem in that manner?

- Analogy: Given a term or set of terms, look up and retrieve pages that are most relevant for the term

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

An excerpt from Fortune magazine on Deep learning http://fortune.com/ai-artificial-intelligence-deep-machine-learning/

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

An excerpt from Fortune magazine on Deep learning http://fortune.com/ai-artificial-intelligence-deep-machine-learning/

Example

advance image recognition Medical startup

rapid accurate radiologists diagnose cancer

pharmaceutical

improve robotics autonomy drone self-driving car

An excerpt from Fortune magazine on Deep learning http://fortune.com/ai-artificial-intelligence-deep-machine-learning/

Words - Visual Words?

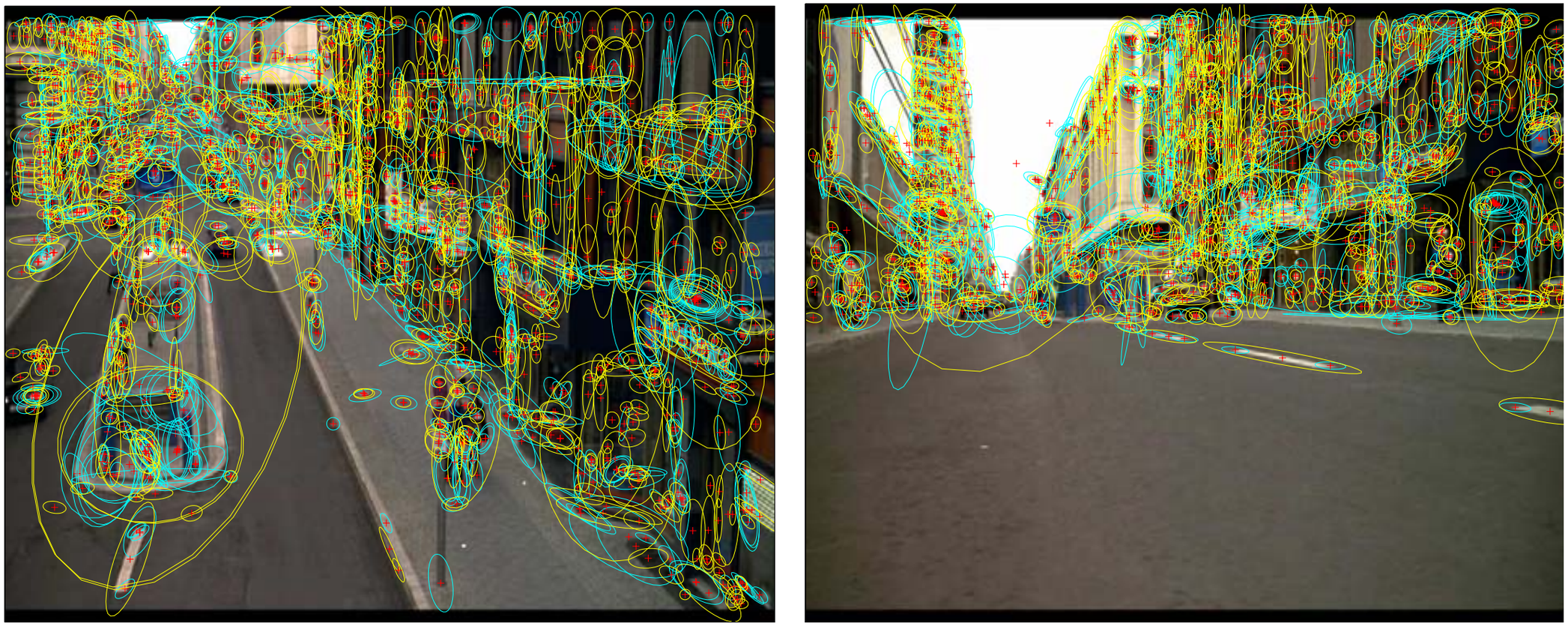
One option - Segmentation?

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 later

Words - Visual Words?

One method for obtaining visual words is based on “Scale Invariant Feature Transform” (SIFT)

Stemming and Lemmatization

am, are, is -> be

car, cars, car's, cars' -> car

the boy's cars are different colors

the boy car be differ color

Visual Word Stemming?

- Centroids found by clustering…

K-Means method

- Top-down method
- Start with random initialisation of *k* cluster centres
- Assign points to the closest cluster centre
- Recompute mean of the cluster centres till convergence

K-Means Method

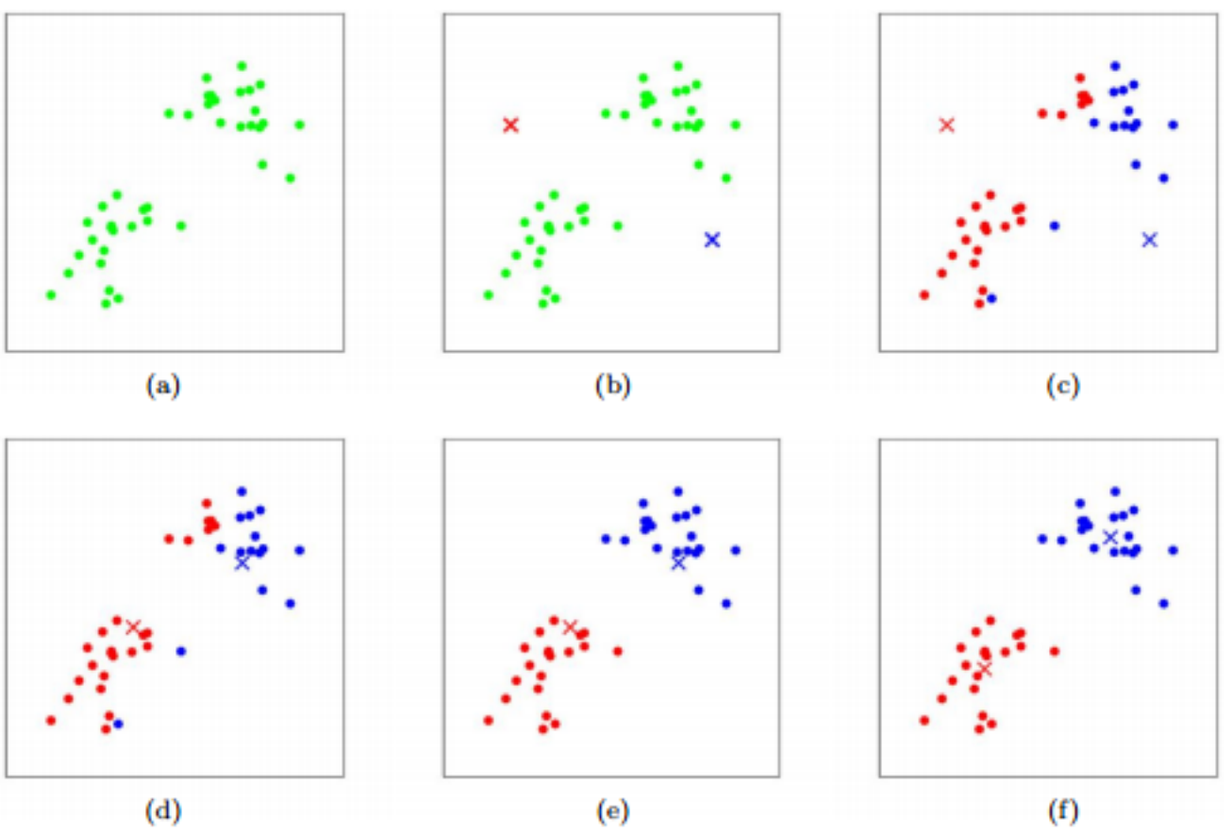


Fig: courtesy Chris Piech

Algorithm

Algorithm 1 Algorithm to obtain visual word vocabulary using K-Means

Result: Clustering using K-Means for obtaining Visual Word Vocabulary
Input: A set of visual words V_i over all images D
Output: A set of visual word centroids C_k that is the set of k visual vocabulary words
Initialization: Randomly initialize C_j^0 to a set of visual words that are randomly chosen
while $C_j^{t+1} - C_j^t > \epsilon$ **do**
    $\forall\ V_i$, obtain distance d_{ij} from each centroid C_j
    Assign V_i to the nearest cluster centroid C_j^t to which distance d_{ij} is minimum;
    For each cluster j obtain the new centroid by obtaining **mean**($V_i \in$ cluster j) ;
    obtain new centroid $C_j^{t+1} =$ **mean**$_j$;
end

Visual Word-> Visual Word Vocabulary

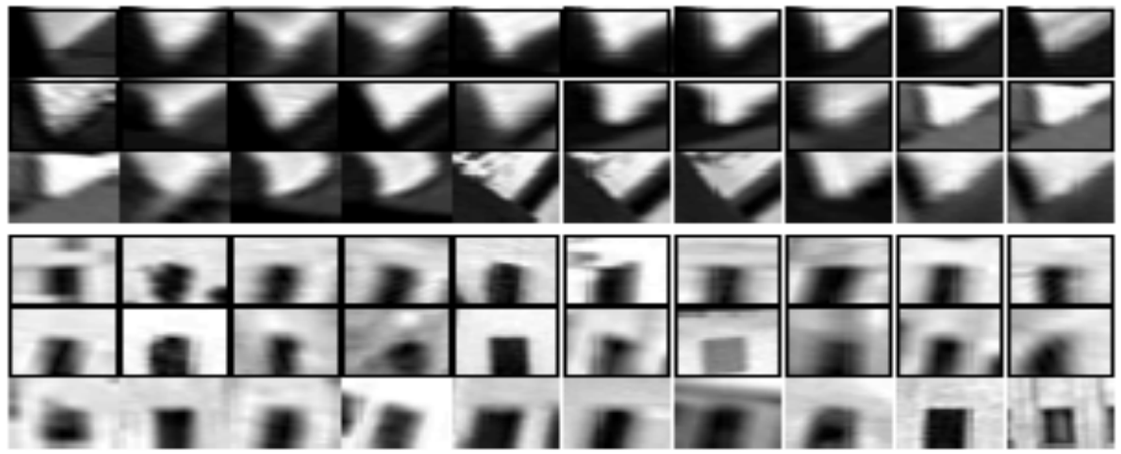
- The number of clusters can be thought of as the dictionary size

- Each visual word extracted is matched against all the centroids and assigned to the centroid it most closely matches

$$\mathbf{dist}(V_j, C_i) > \mathbf{dist}(V_j, C_m) \forall m \neq i.$$

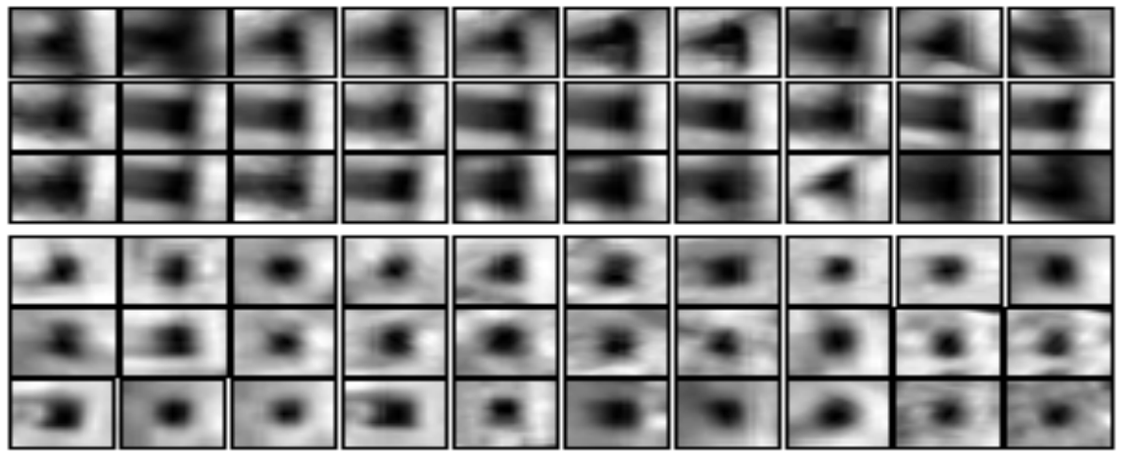
Examples of visual word-> centroid

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.

Visual Indexing

- After extracting visual words, they are matched with the visual word vocabulary
- The information is indexed in an inverted list that matches each visual vocabulary word with the matched visual words in each document and the location of each word

Scoring of words

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

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

- 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.

Inverted List example for text

ID	Text
1	Baseball is played during summer months.
2	Summer is the time for picnics here.
3	Months later we found out why.
4	Why is summer so hot here

↑ Sample document data

Dictionary and posting lists →

Term	Freq	Document ids
baseball	1	[1]
during	1	[1]
found	1	[3]
here	2	[2], [4]
hot	1	[4]
is	3	[1], [2], [4]
months	2	[1], [3]
summer	3	[1], [2], [4]
the	1	[2]
why	2	[3], [4]

Procedure for Database Images

- Given a database of images, initially visual word descriptors are extracted from all images.
- These descriptors are clustered using k-means algorithm to obtain a set of k visual word vocabulary (cluster centroids) C_k
- For each image, each of the set of visual words are associated with the nearest cluster centroid and the corresponding TF-IDF weighted vector is obtained for each document V_{kd} where each document d is represented in terms of k centroids.
- These are stored in an inverted list.

Procedure for Querying

- A bounding box in an image is specified as a query. The local feature extraction method is followed to extract visual words in the query bounding box region.
- The visual words extracted are then stemmed using the visual word vocabulary centroids C_k by finding for each word the nearest centroid. In this way a query vector Q_k is obtained.

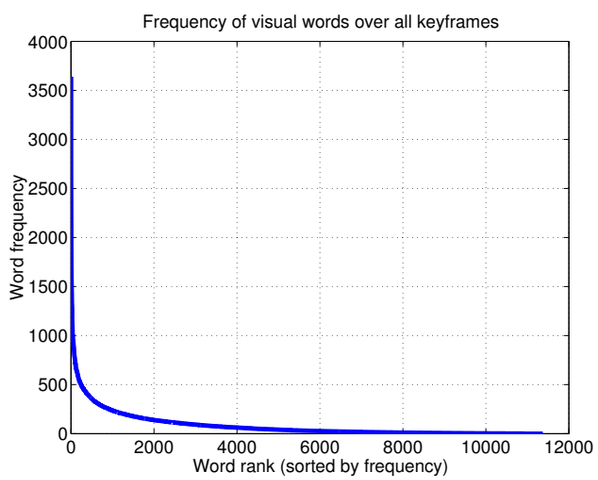
Procedure for Querying

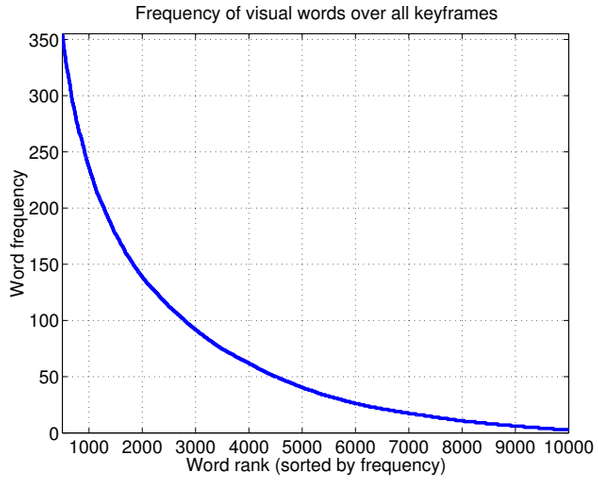
- Corresponding to each term t_i occurring in Q_k we obtain through the inverted list a set of documents that have the term t_i appearing in them along with the weight of the term t_i in that document.
- Based on the distance we compute the distance between query vector Q_k and the ith database document vector V_{ki} where each term has its corresponding weight for both query and database.

Stop words

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

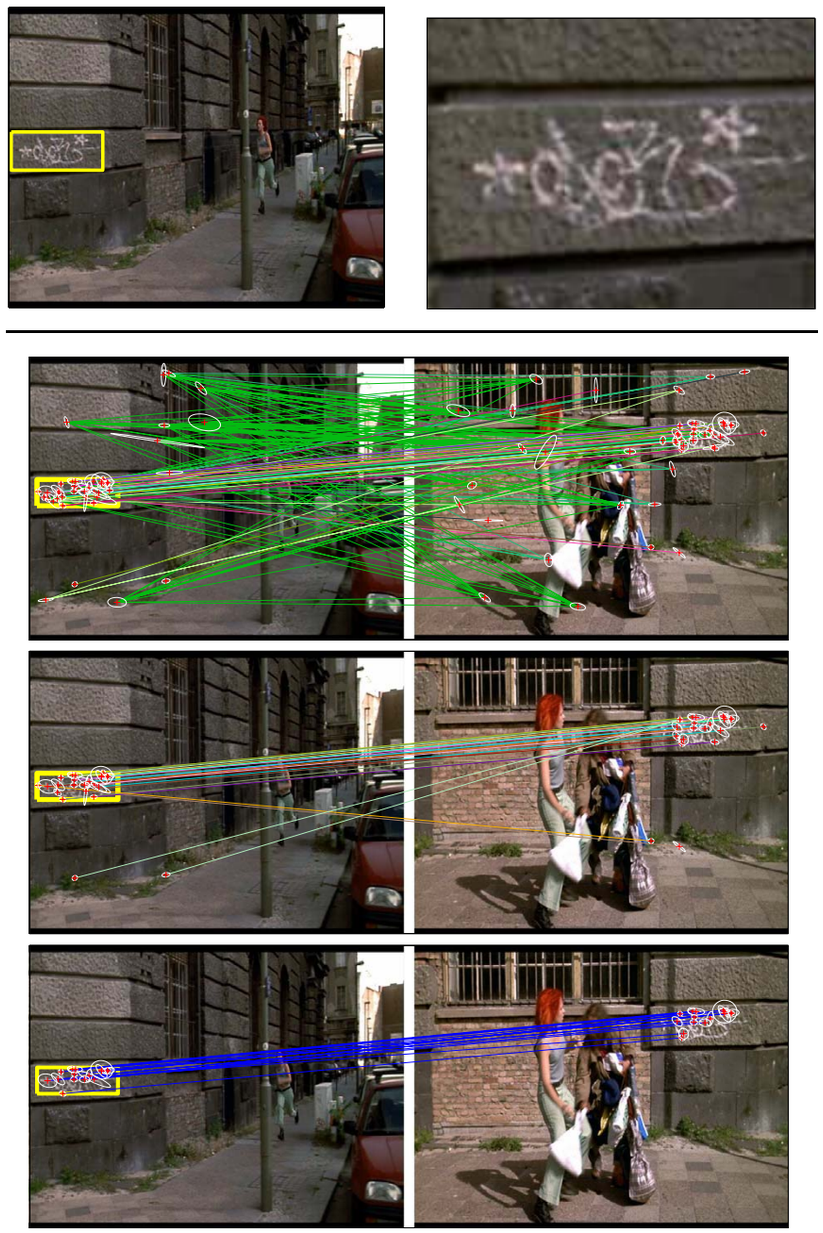
- Similar strategy is used in VideoGoogle





Spatial consistency

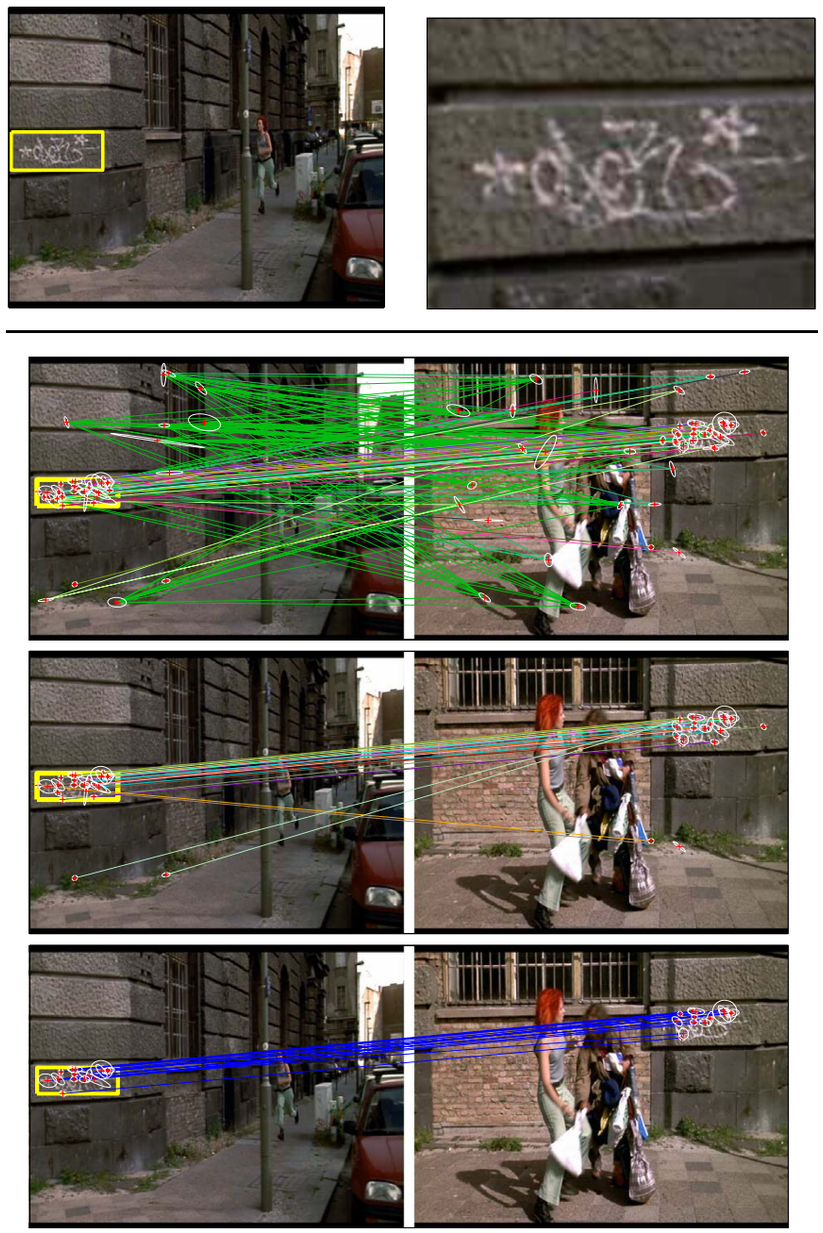
- The matches are scored for being spatially consistent



Spatial consistency

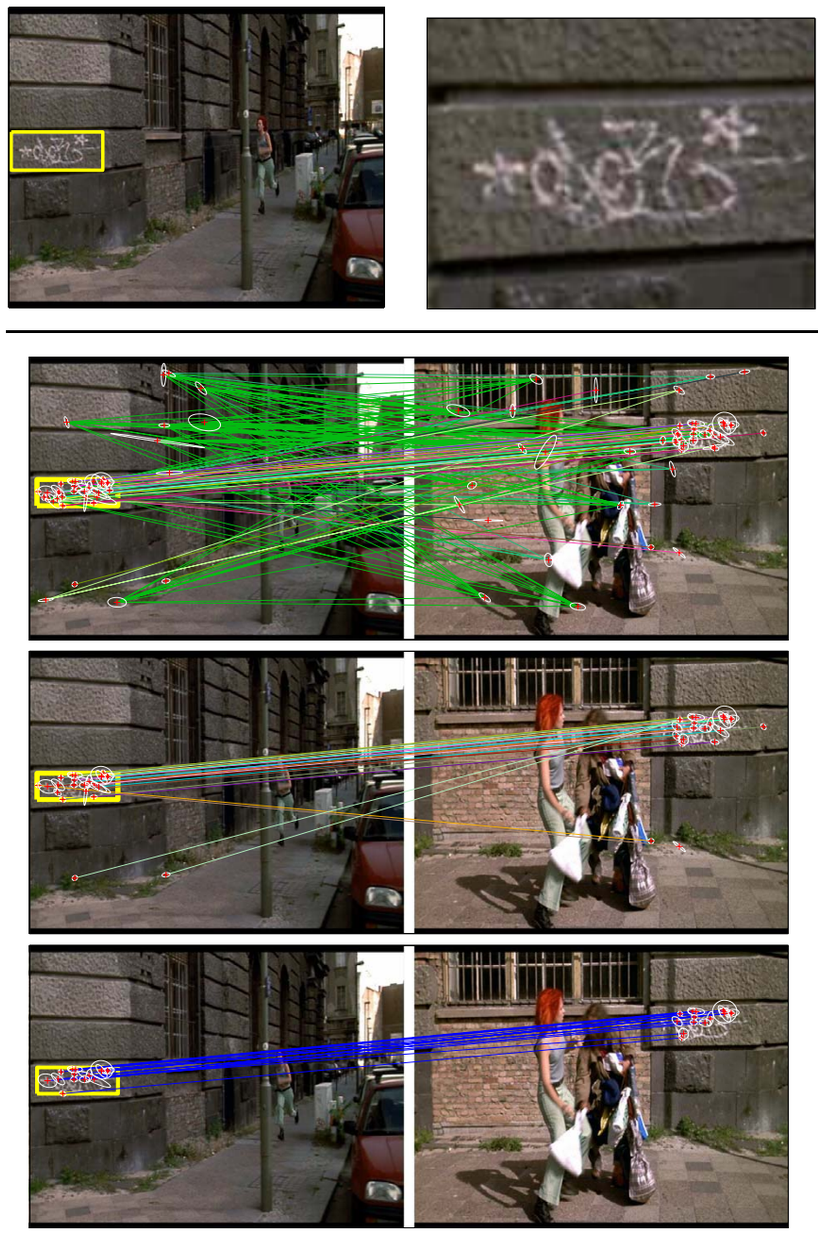
- The matches are scored for being spatially consistent

- More matches in a neighbourhood indicate high likelihood of a correct match

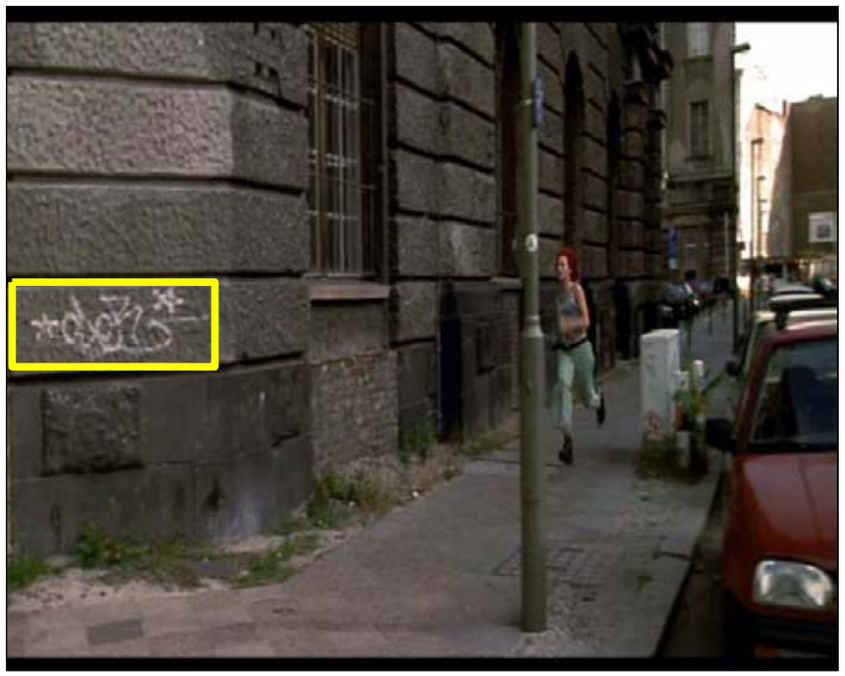


Spatial consistency

- 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

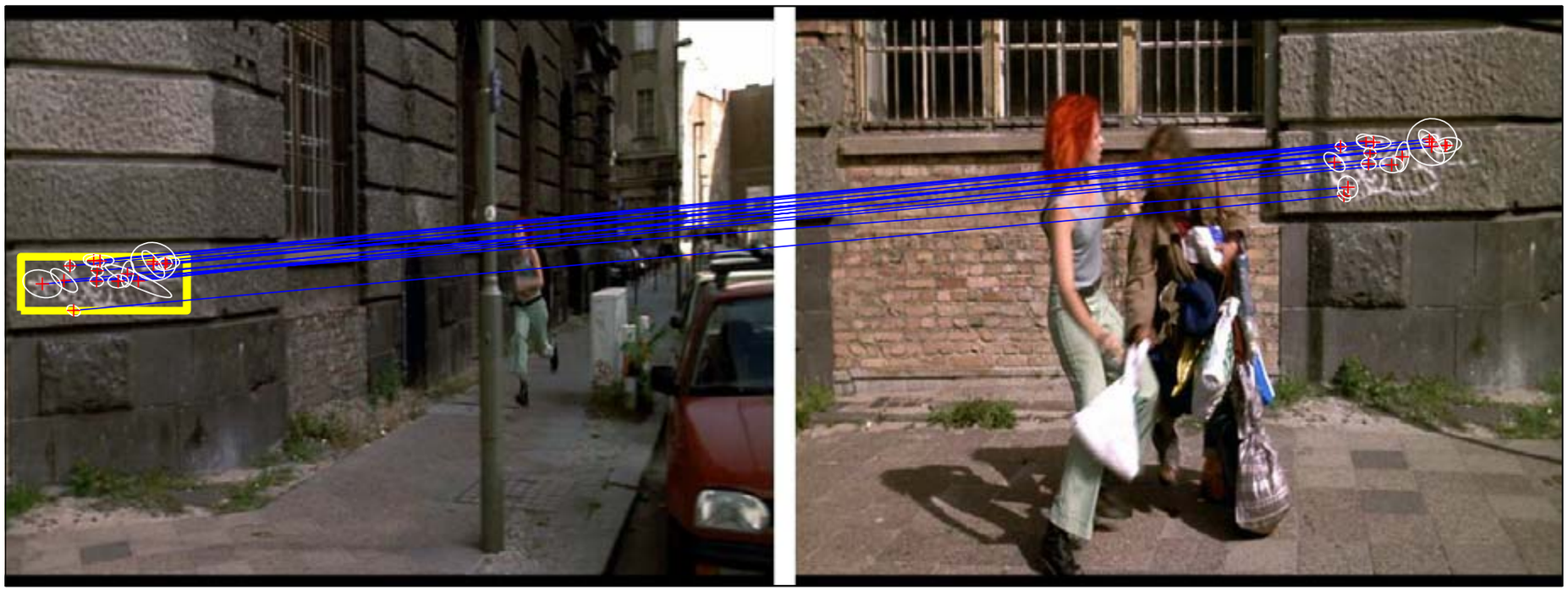


Example

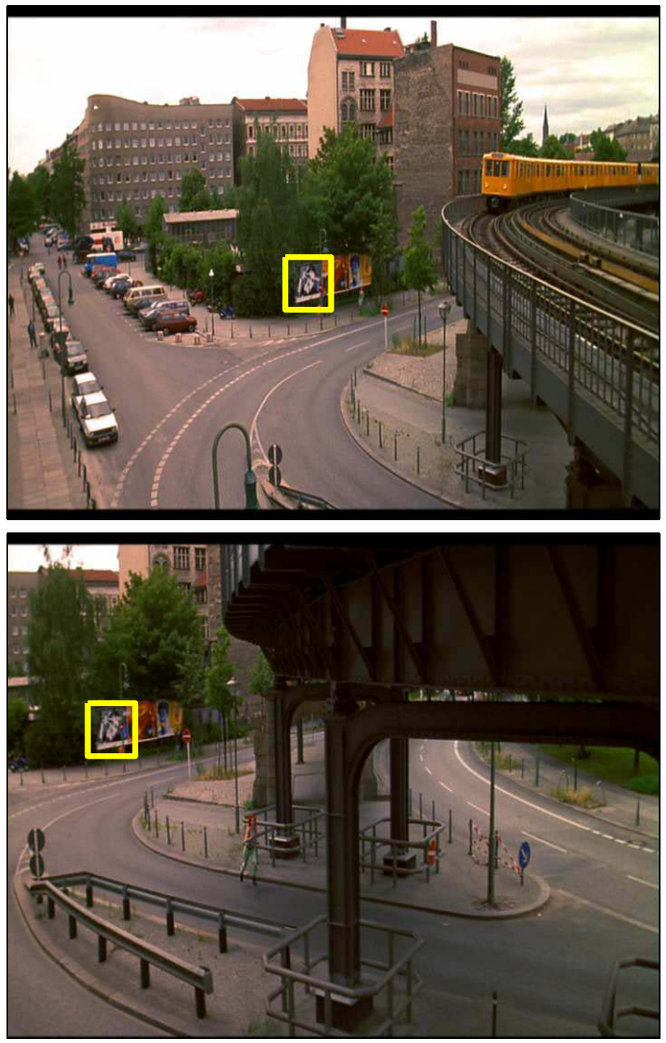
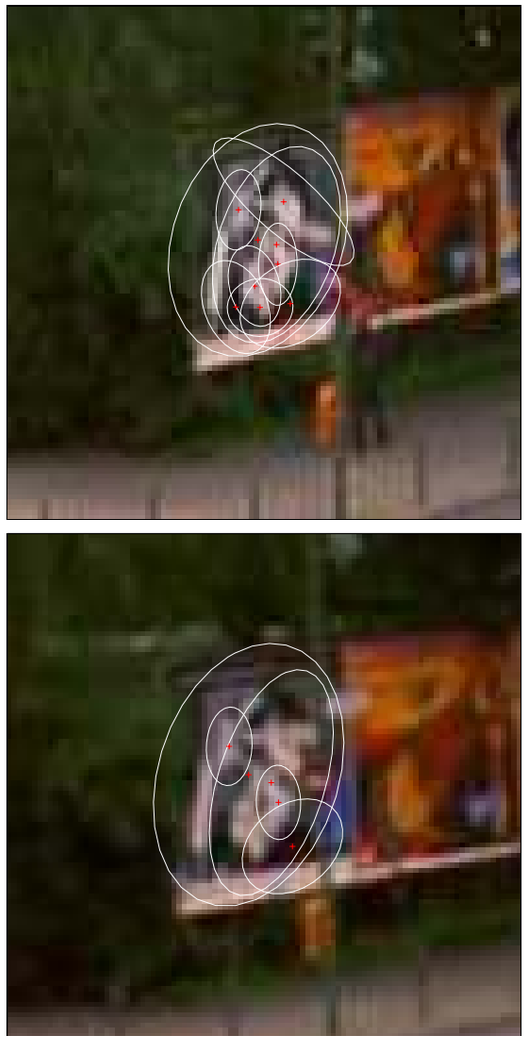
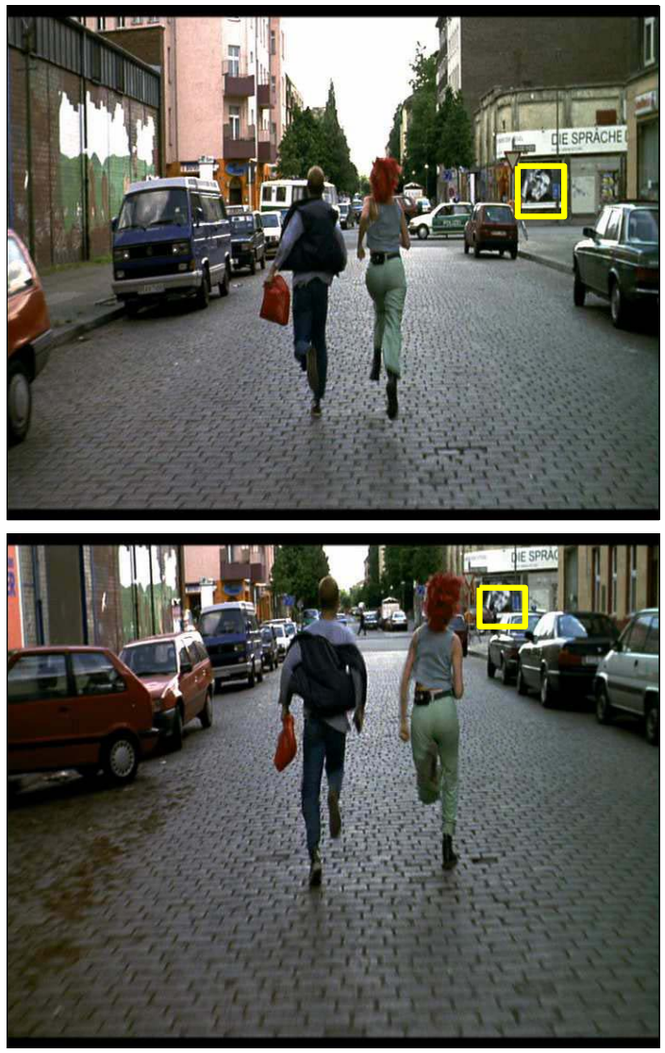


Example

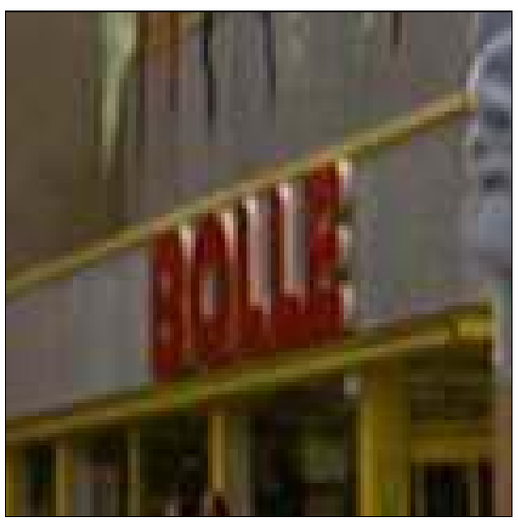
Example

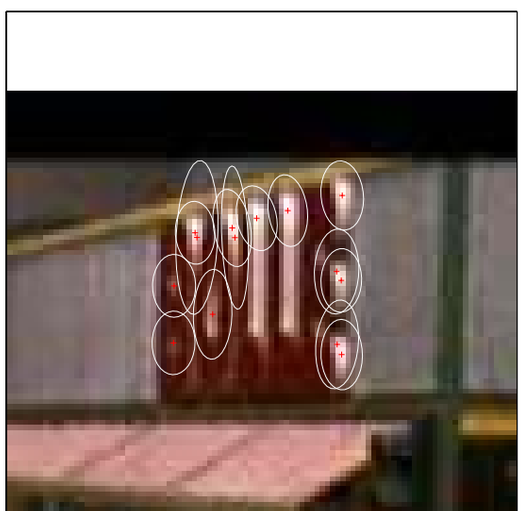


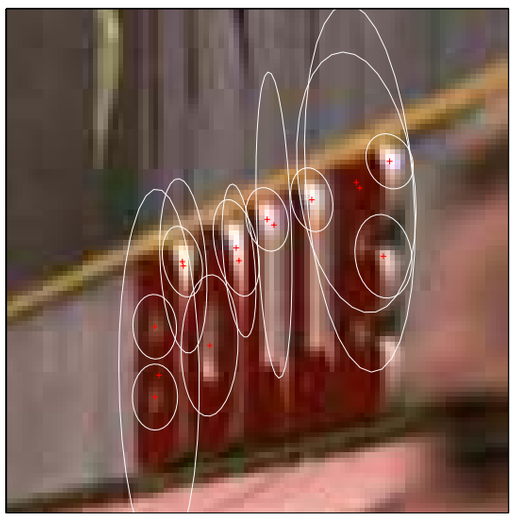
Example

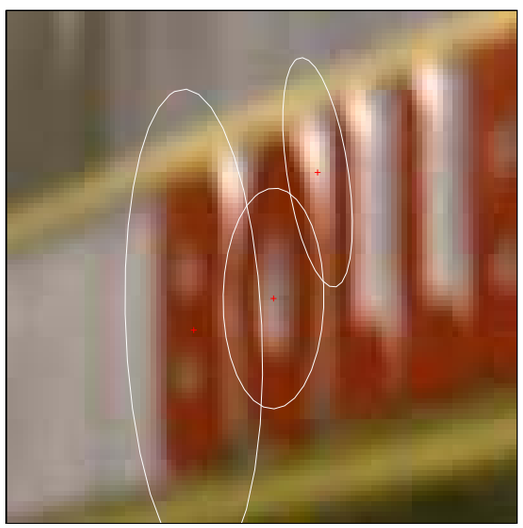





Example







How to improve over Video Google?

How to improve over Video Google?

- Dependency on the visual word vocabulary.
- If we use higher vocabulary size scalability of method may be effected
- Crucial dependency on geometric layout of words

Scalable Vocabulary Tree Approach

- The scalable vocabulary tree approach by Nister and Strewenius in CVPR 2006 addresses these limitations

- We present the main idea using the authors illustrations

Scalable Recognition with a Vocabulary Tree

David Nistér, Henrik Stewénius

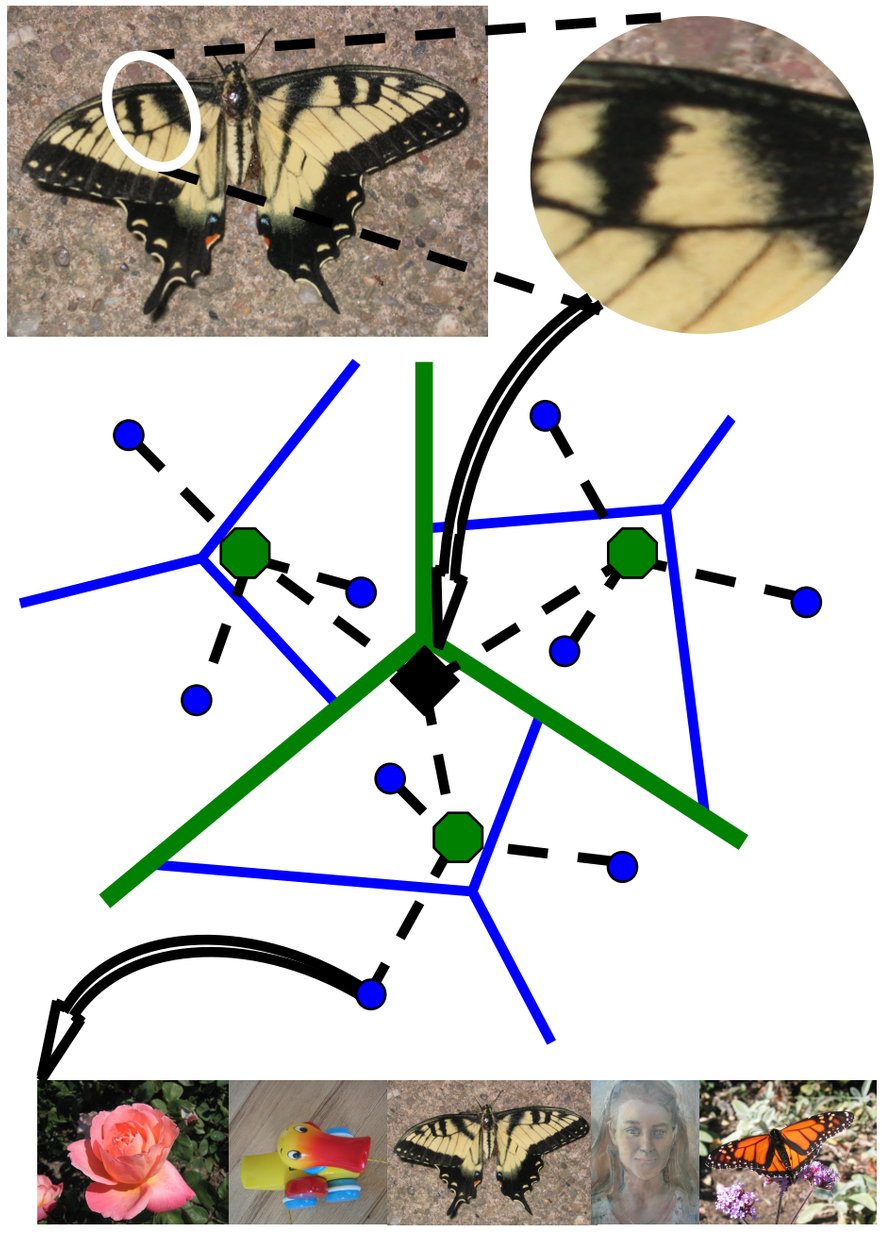
UK

Center for Visualization & Virtual Environments

New York City CVPR 2006

Scalable Recognition with a Vocabulary Tree

David Nistér, Henrik Stewénius

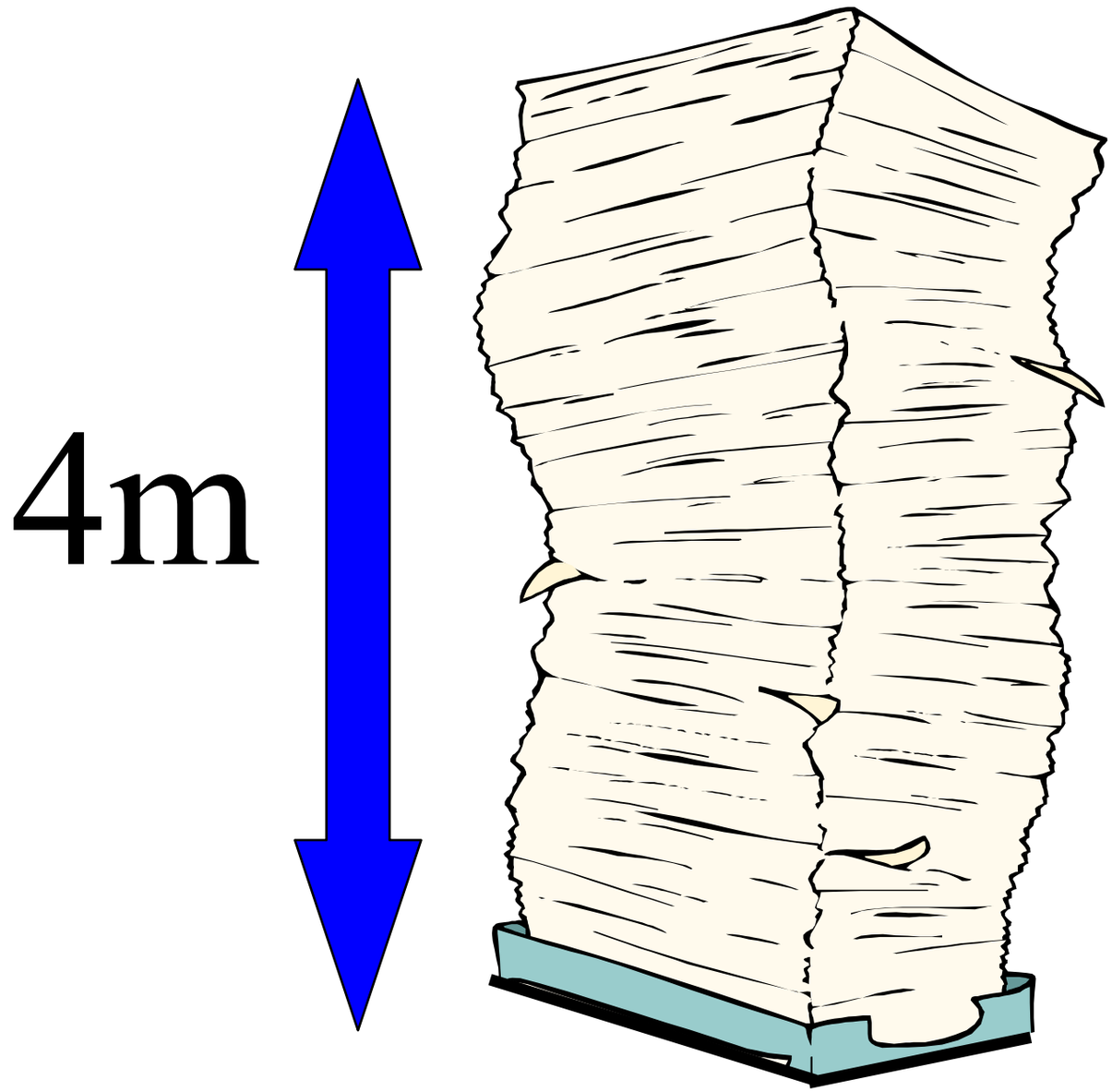


UK

Center for Visualization & Virtual Environments

New York City CVPR 2006

50 Thousand Images



110,000,000
Images in
5.8 Seconds

UK

Center for
Visualization & Virtual
Environments


MetLife
Kodak
EASYSHARE Digital Camera and Dock
Shoot.
Touch.
Share.
Ask Merrill
MUSIC
TIMES SQUARE GALLERY
New York Sightseeing
GRAY LINE
www.grayline newyork.com
PLANET HOLLYWOOD

Center for Visualization & Virtual Environments

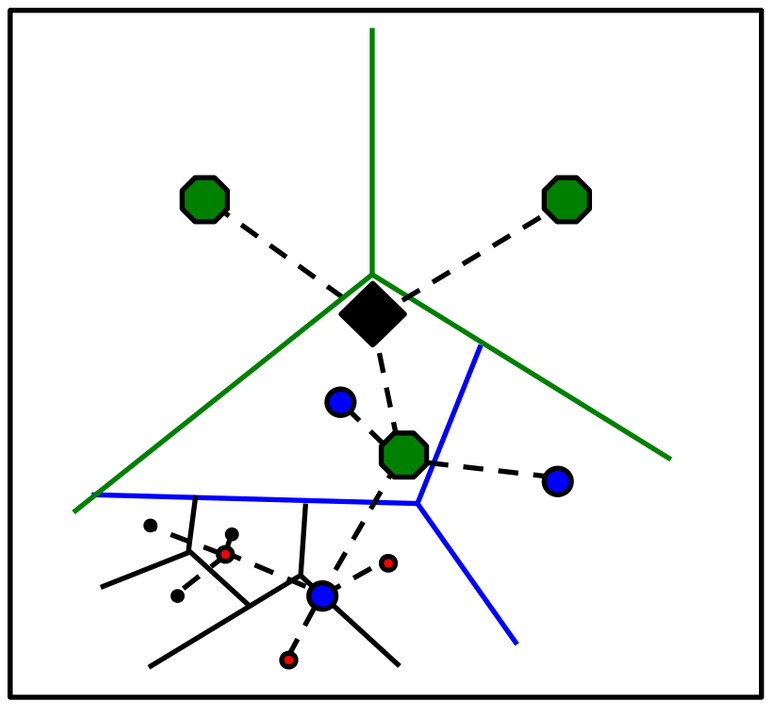
UK

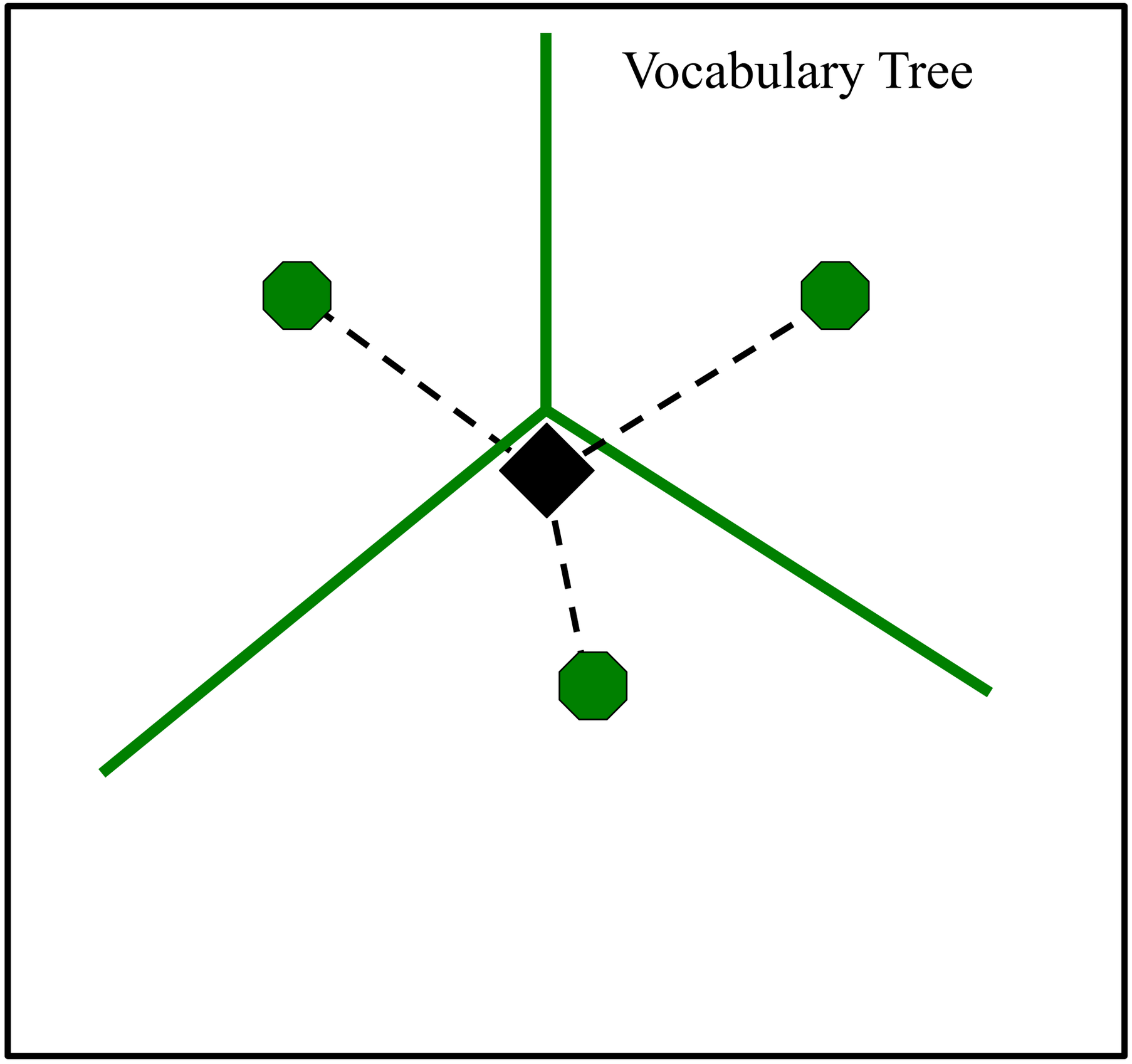
Center for Visualization & Virtual Environments

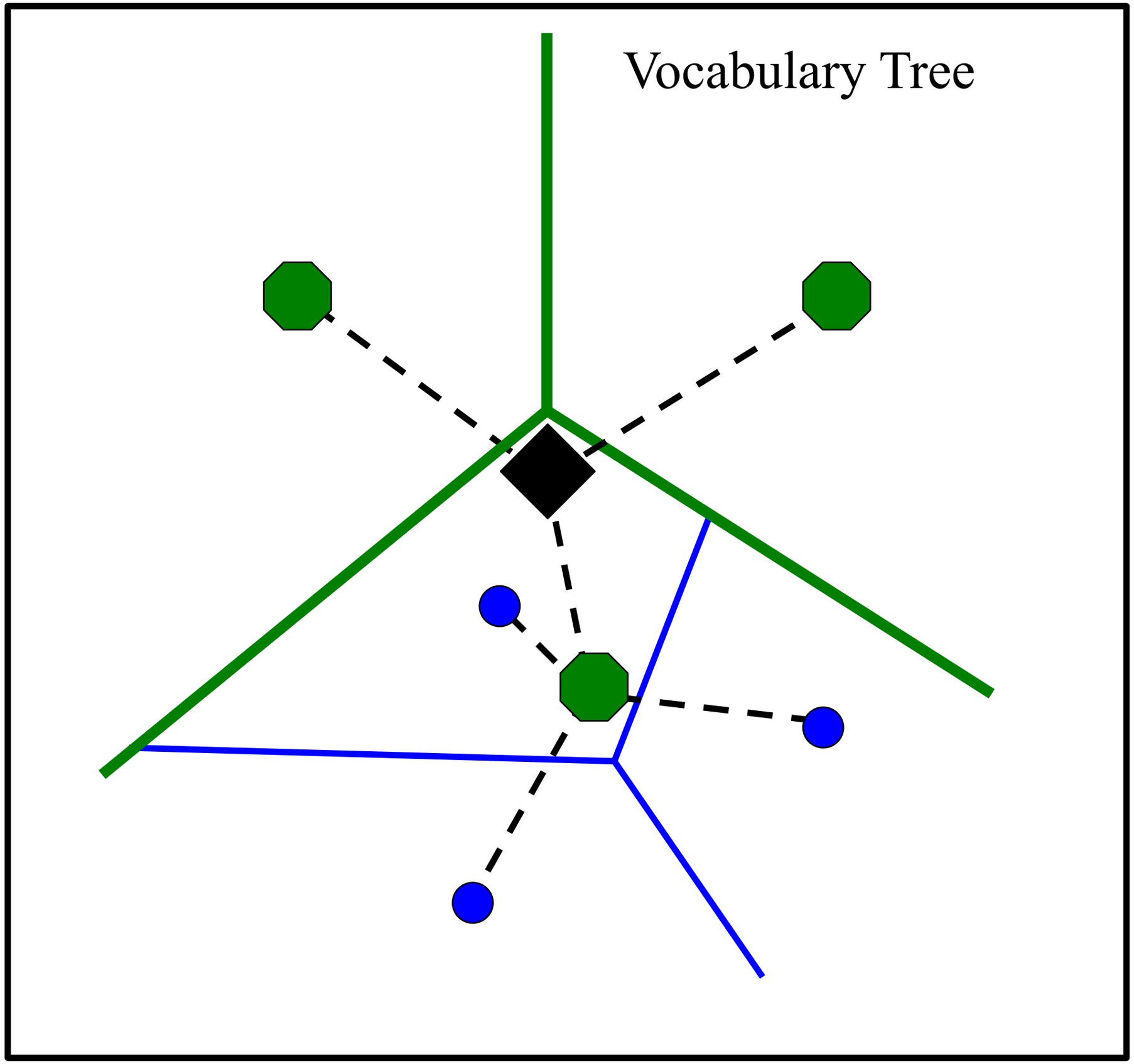
Take-Home Message

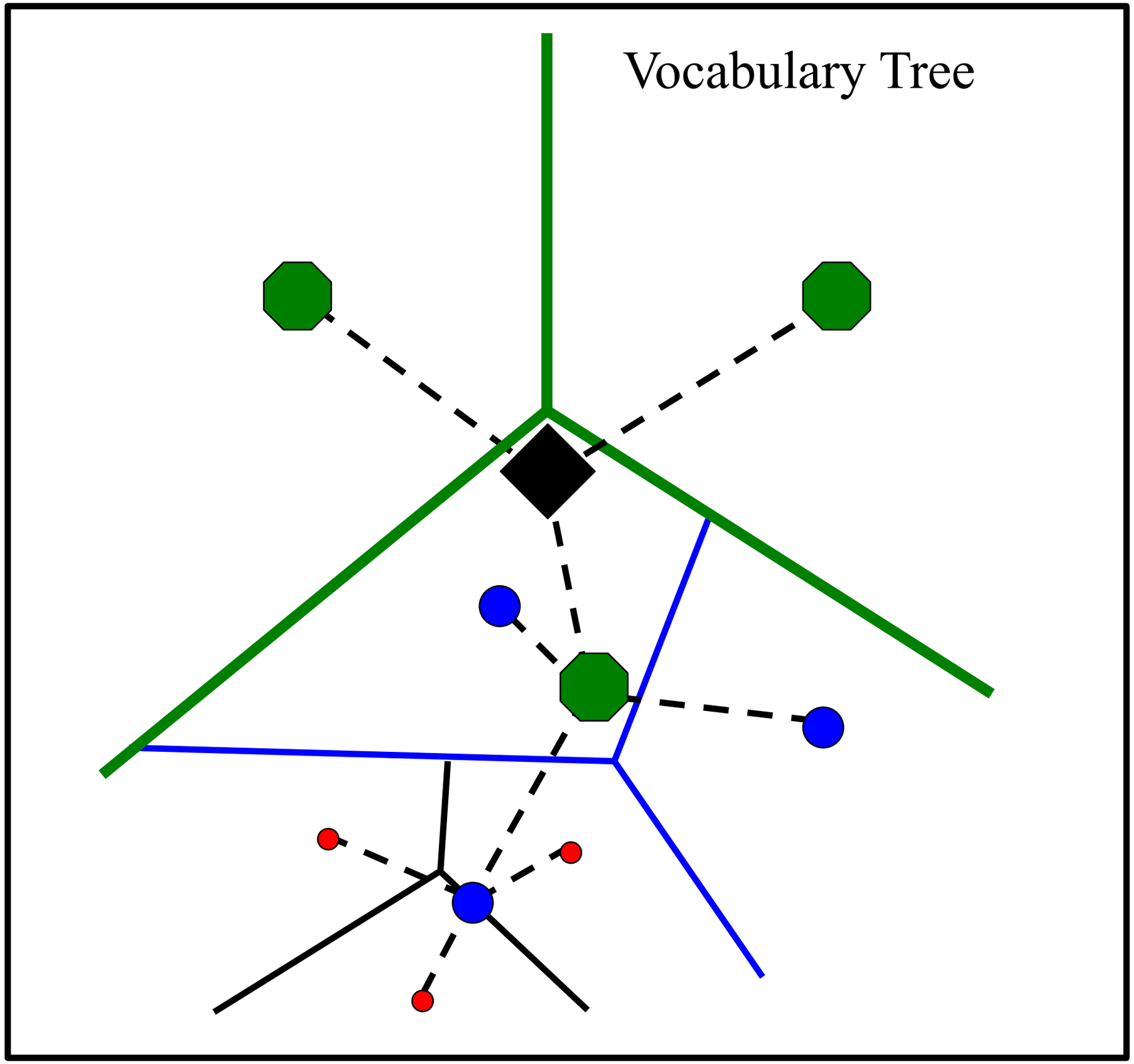
If we can get repeatable, discriminative features,

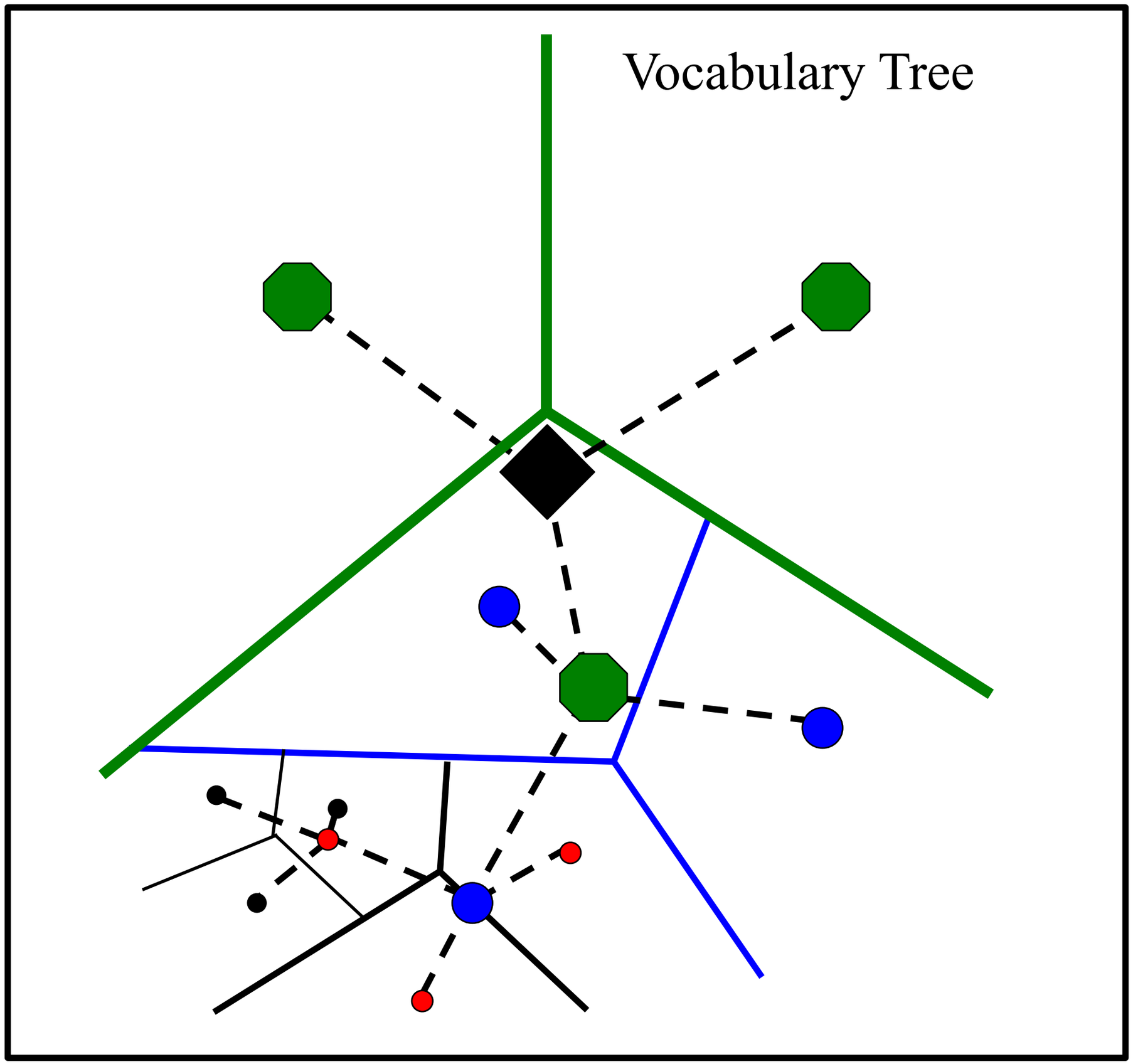
then recognition can scale to very large databases
using the vocabulary tree and indexing approach
described in Nistér & Stewénius CVPR 2006.




Vocabulary Tree


Vocabulary Tree


Vocabulary Tree


Vocabulary Tree

Adding, Querying and Removing Images at full speed

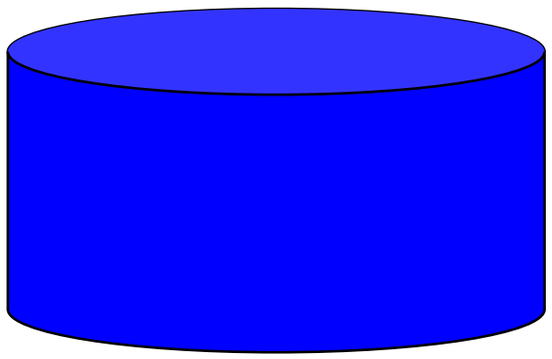
Query

Add

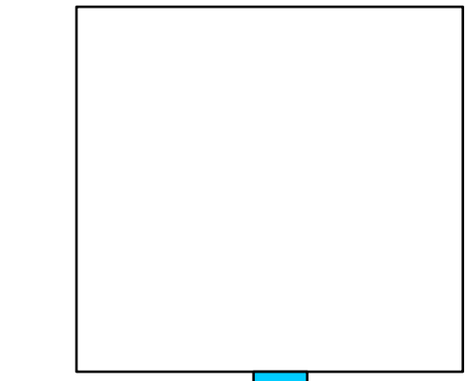
Remove

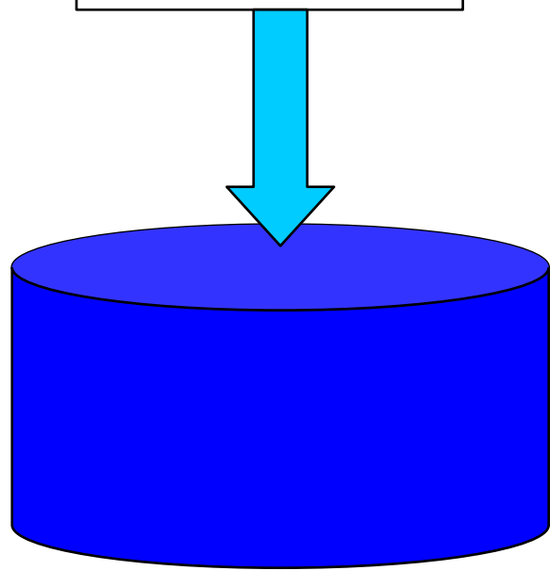
Training and Addition are Separate

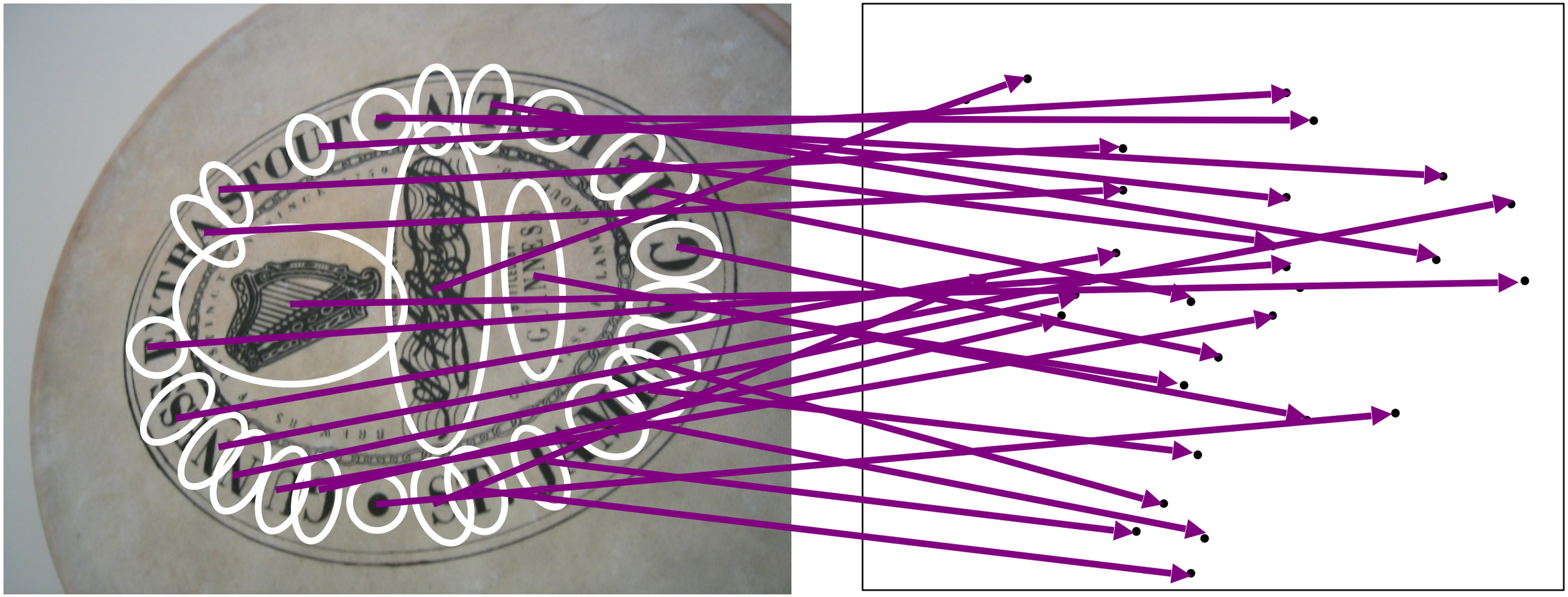
Common Approach

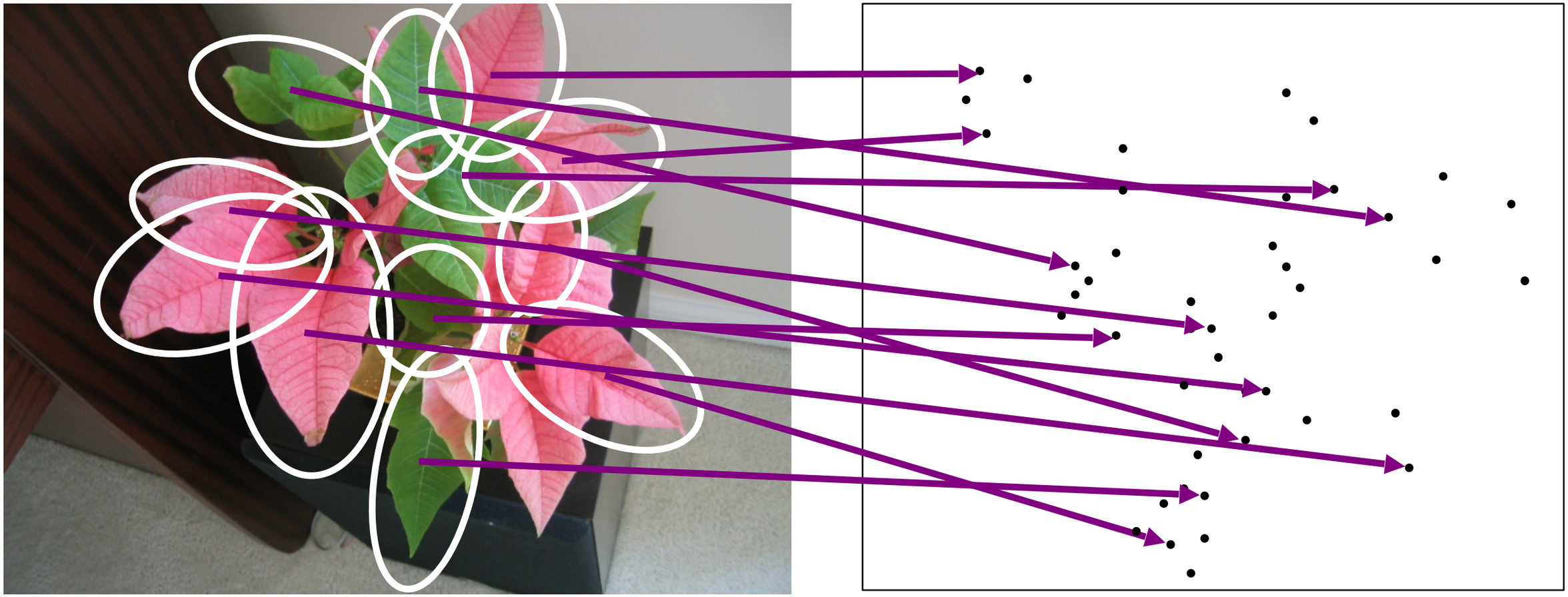


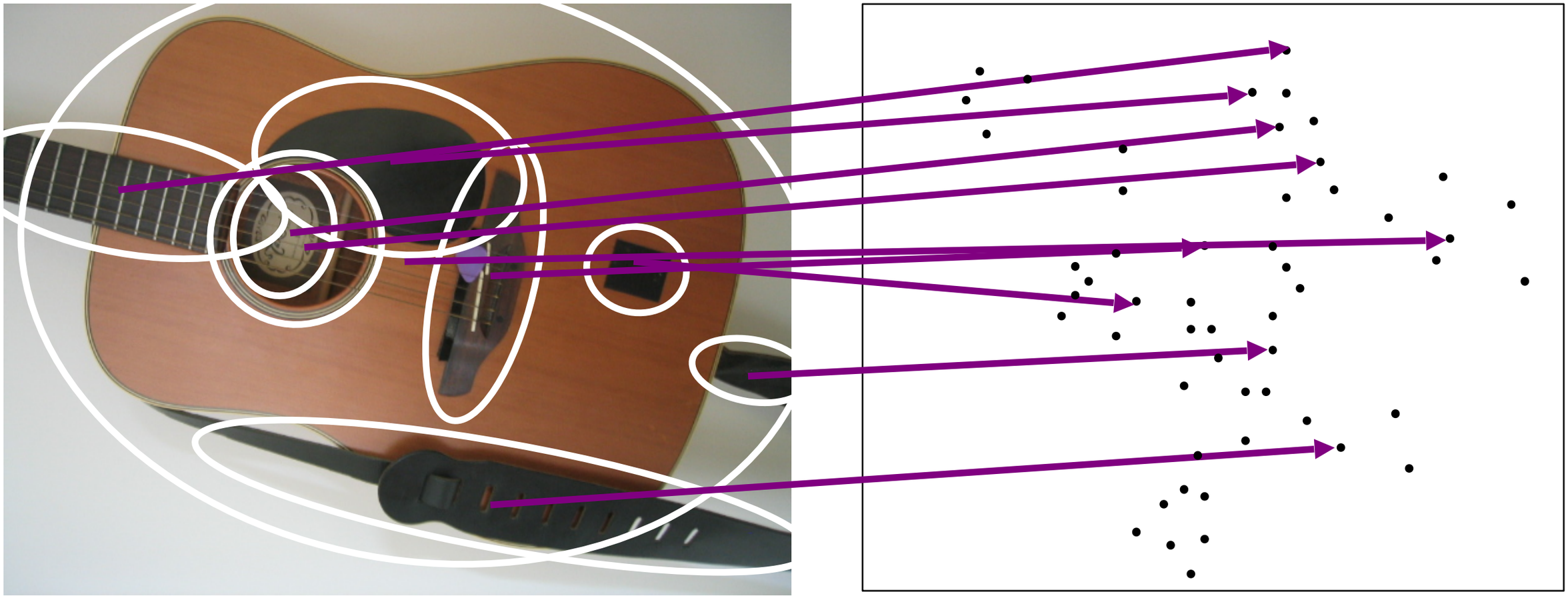
Our approach

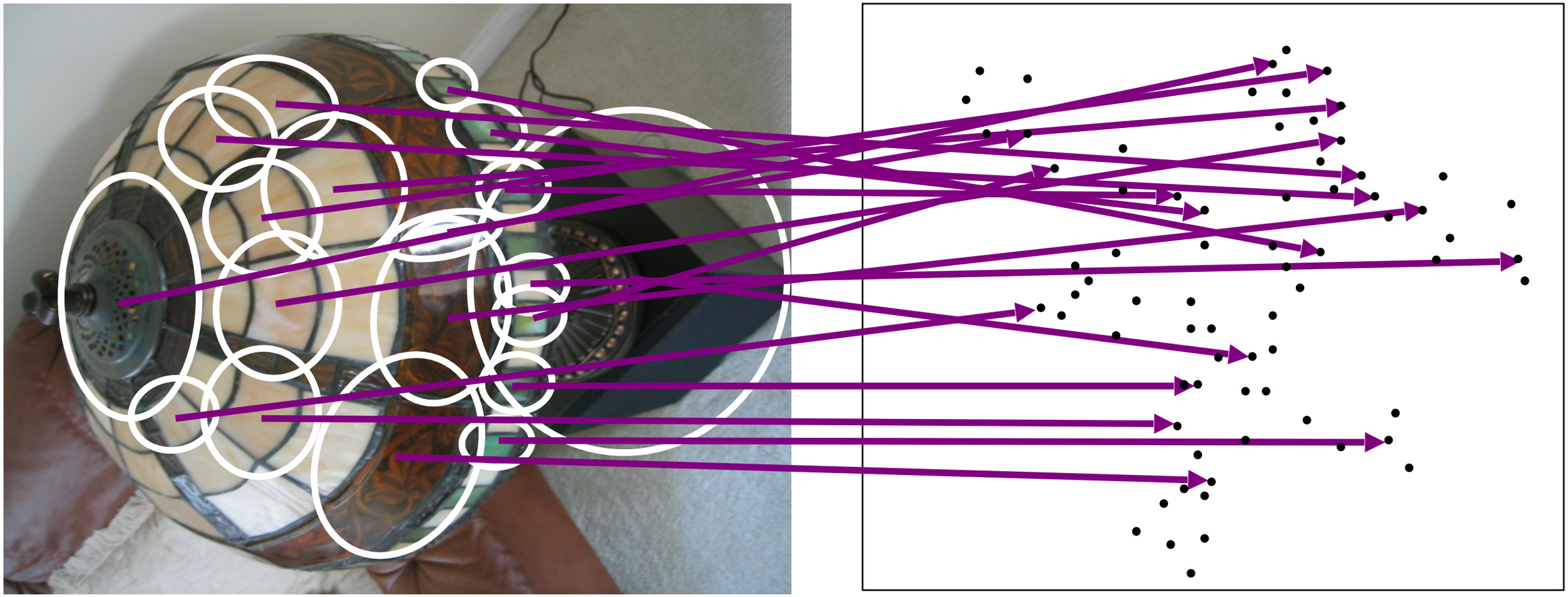


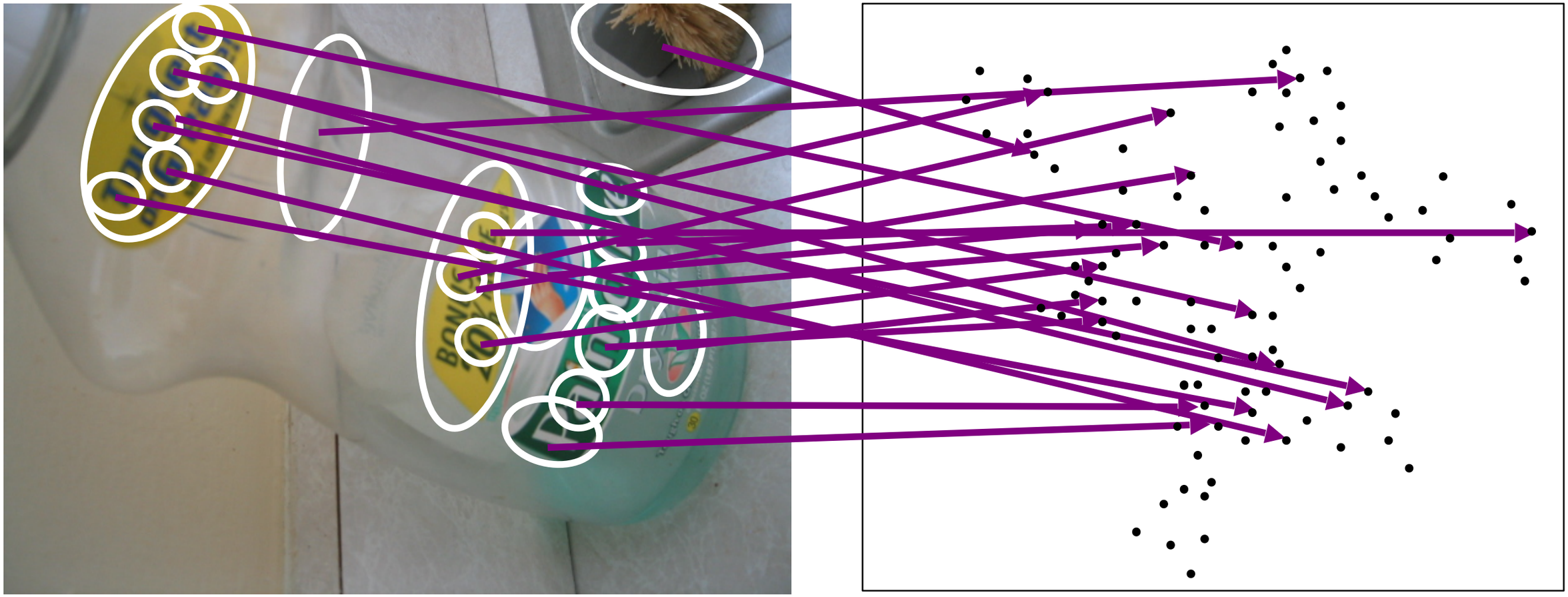


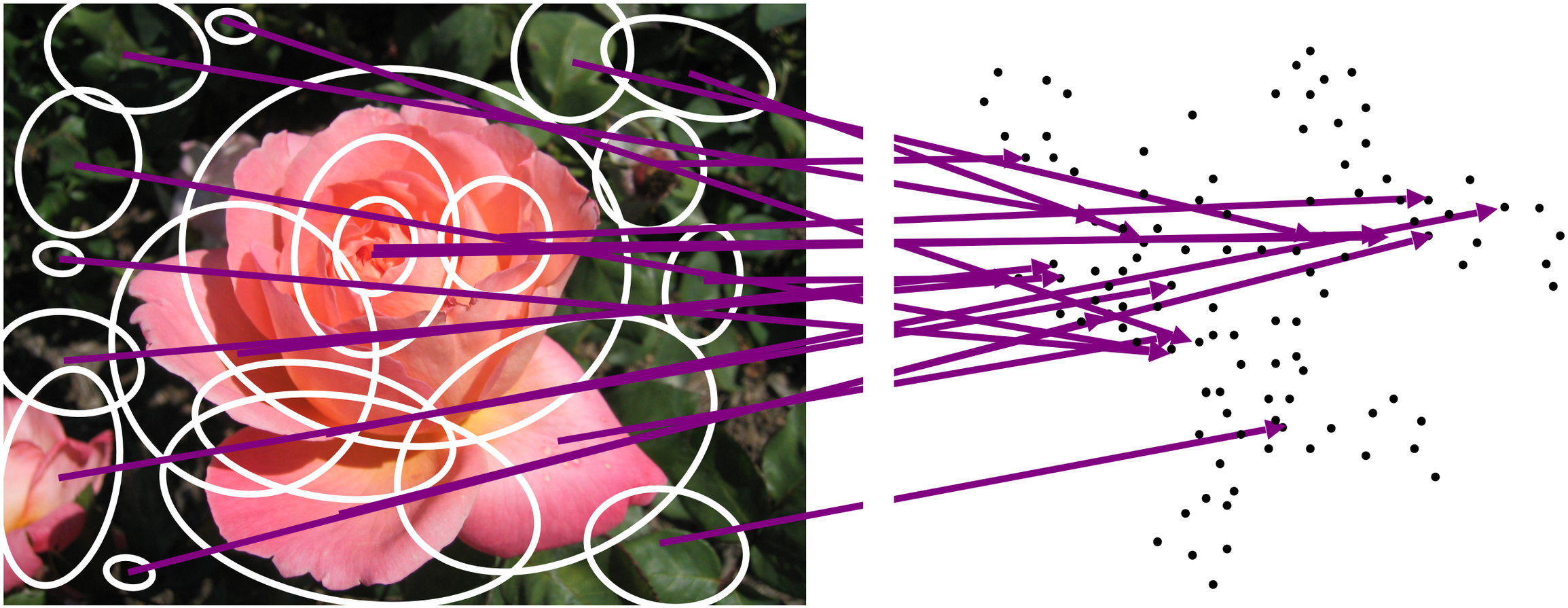


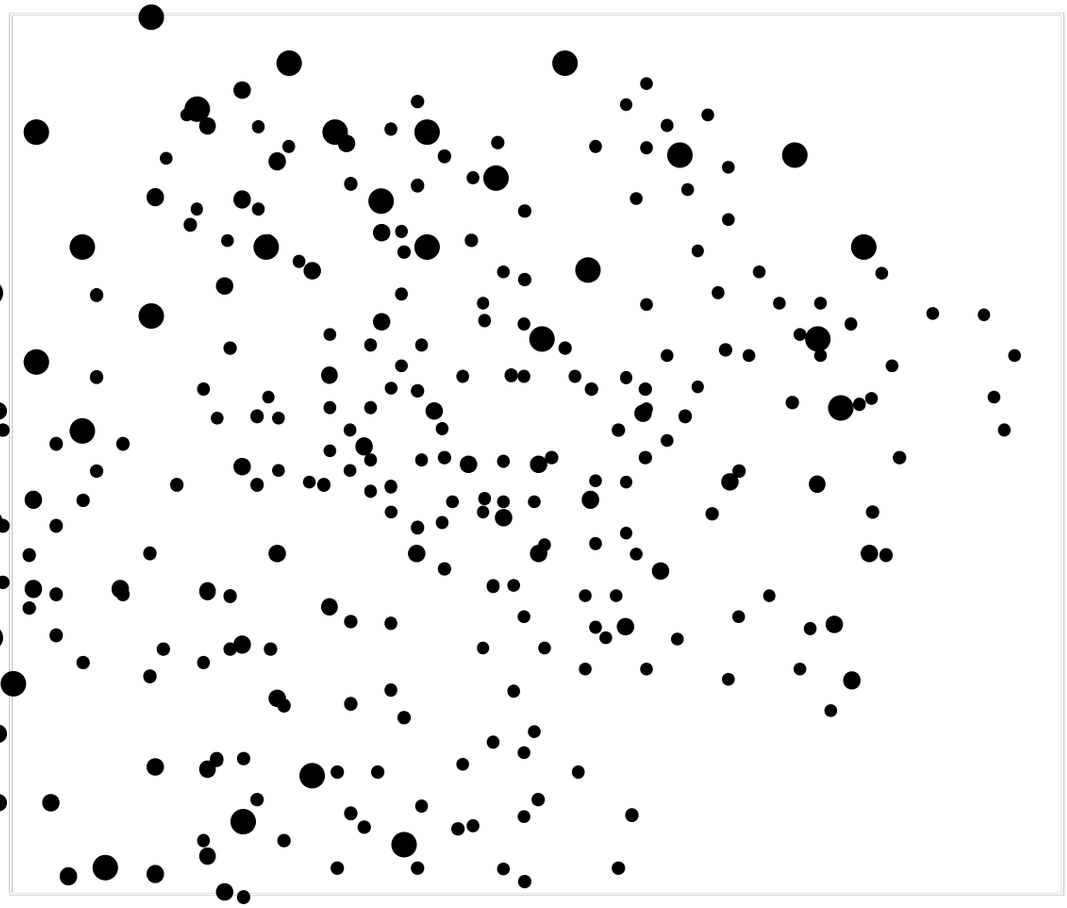






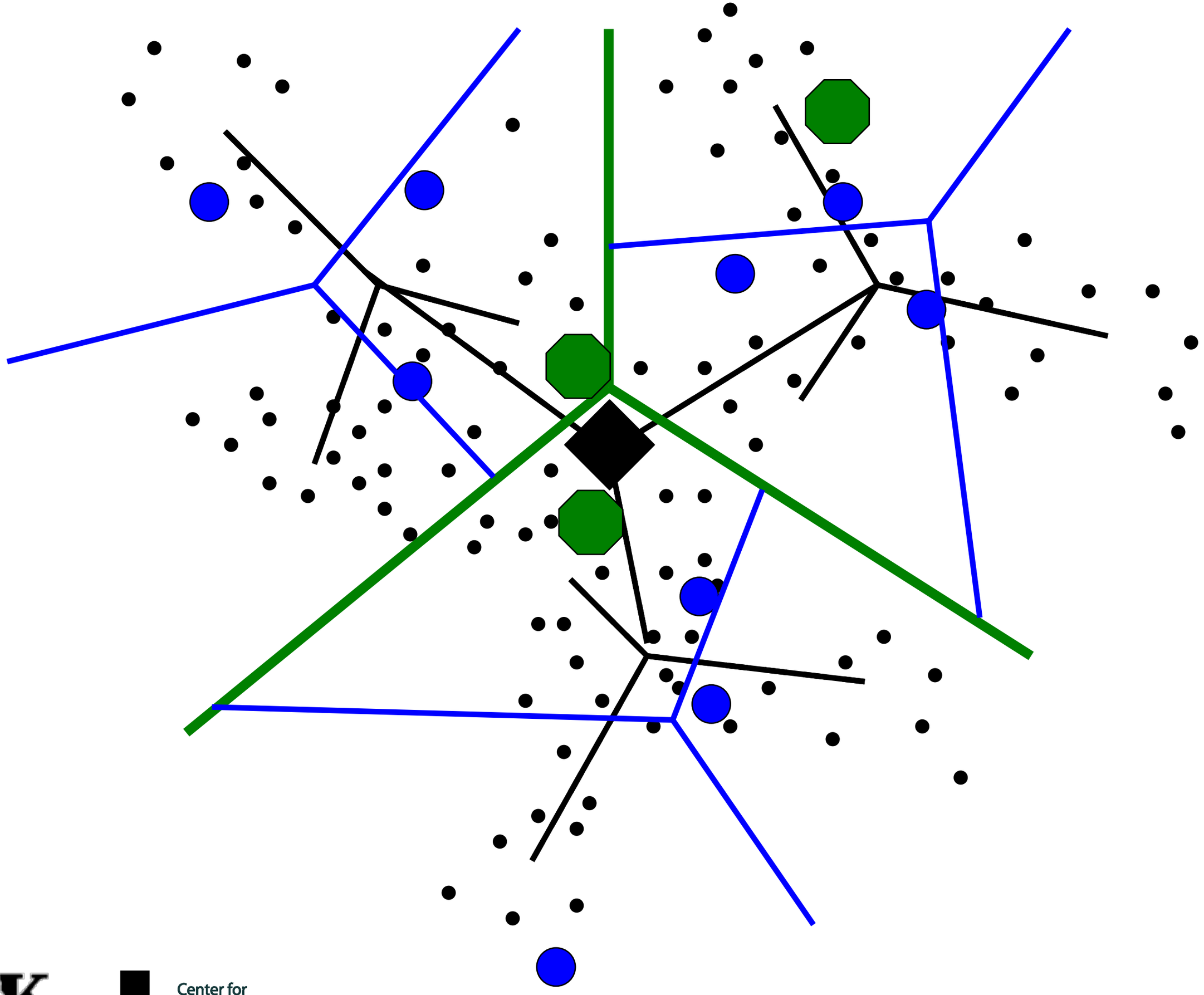


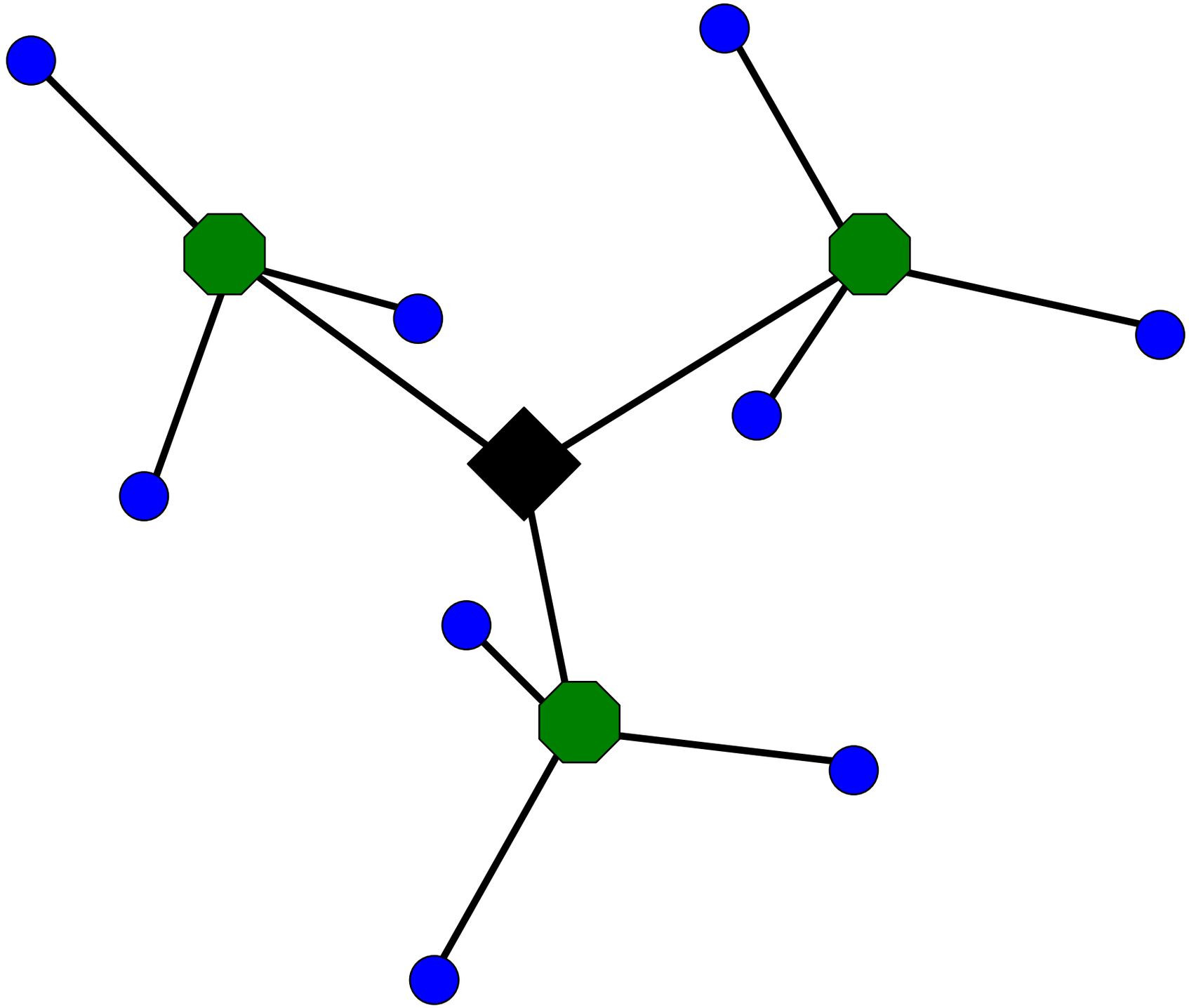


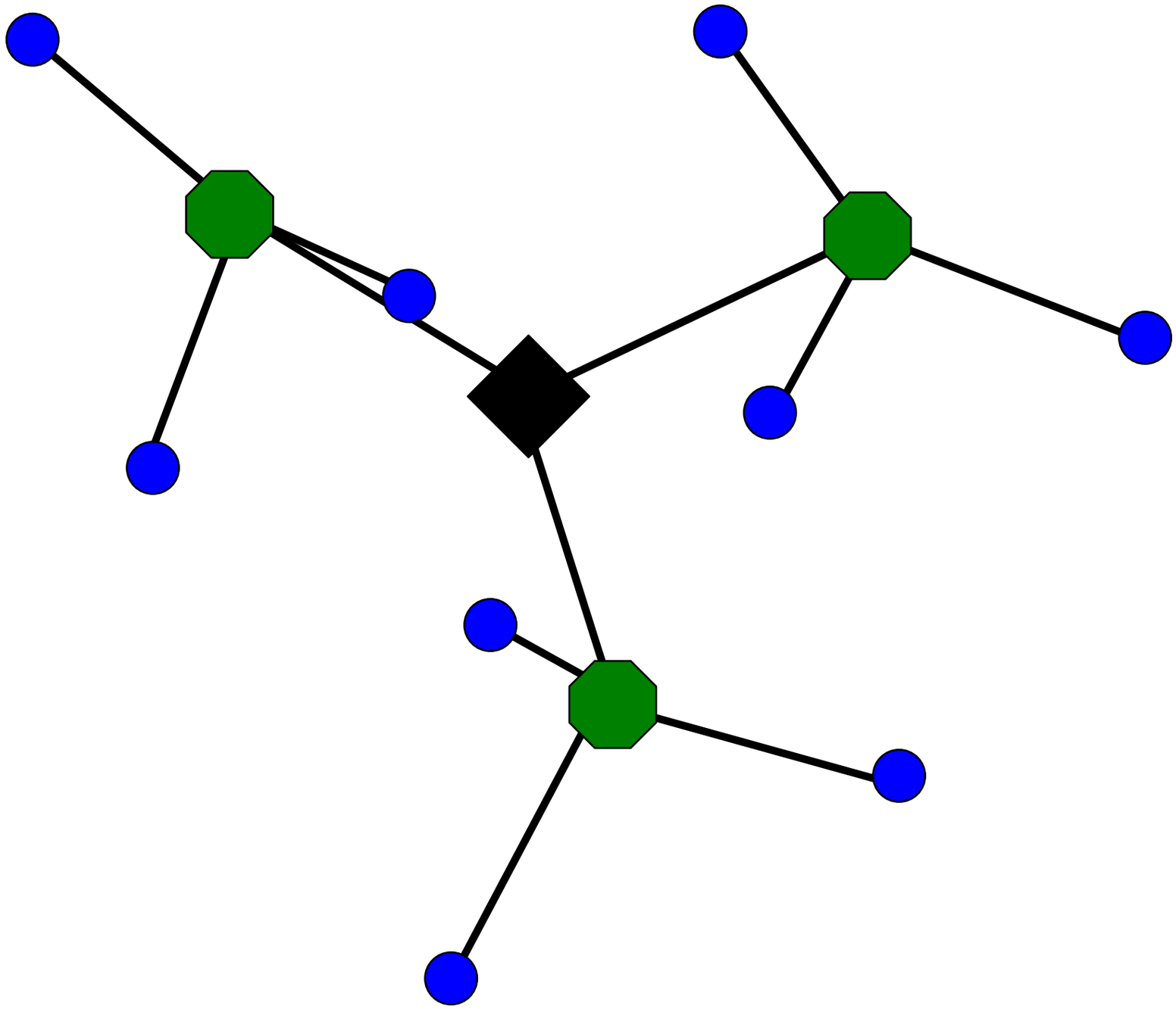


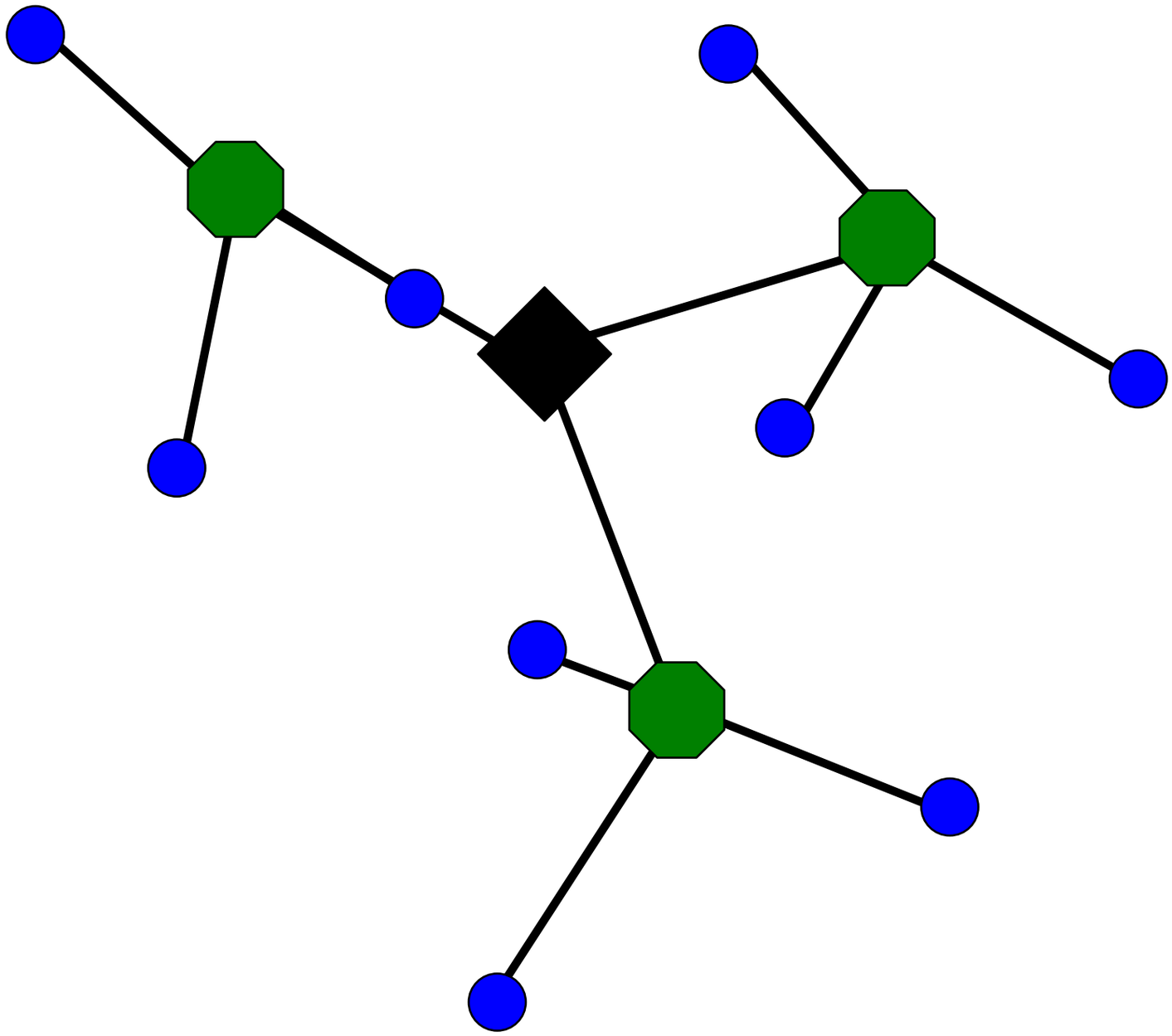
UK

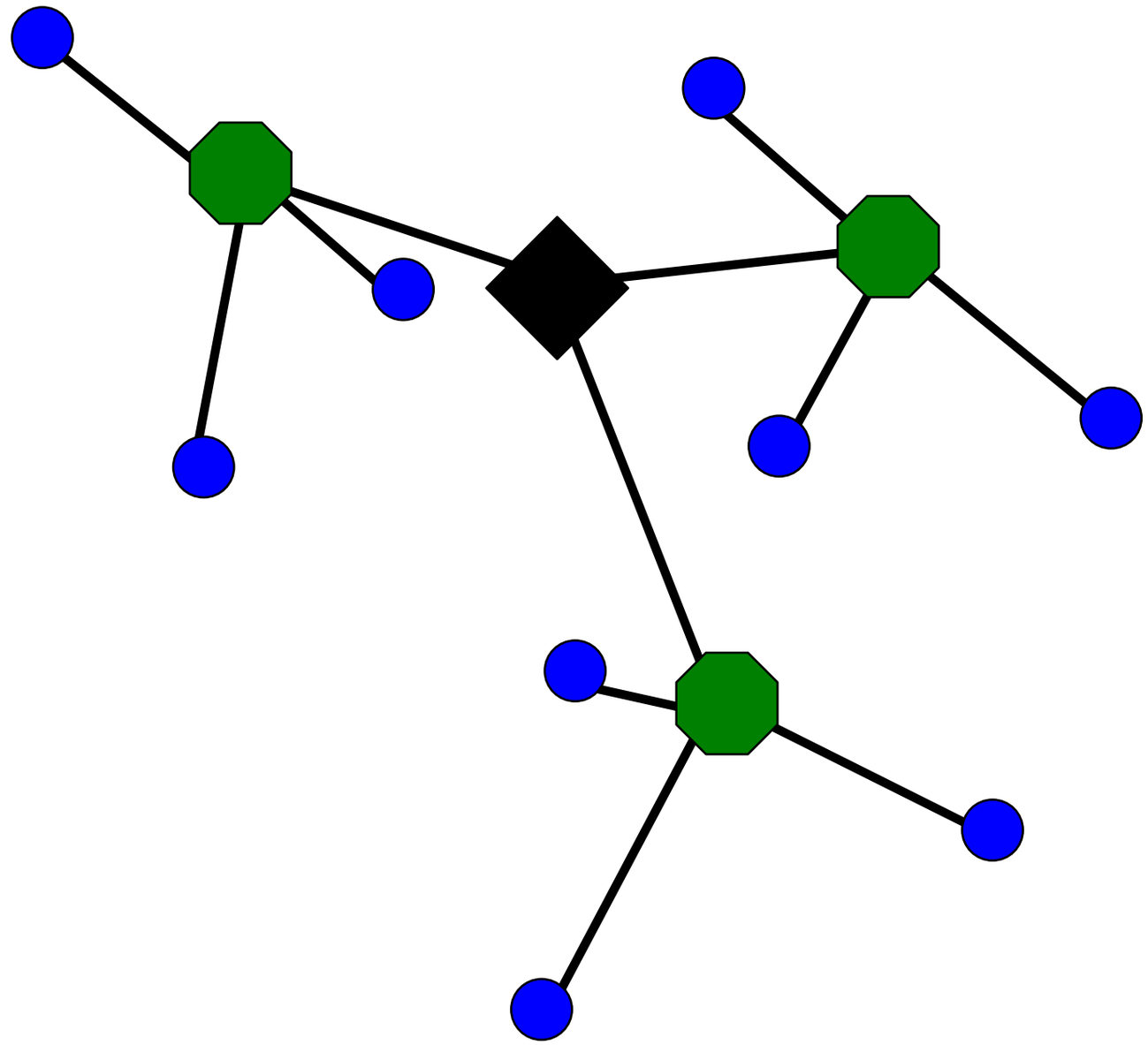
Center for
Visualization & Virtual
Environments

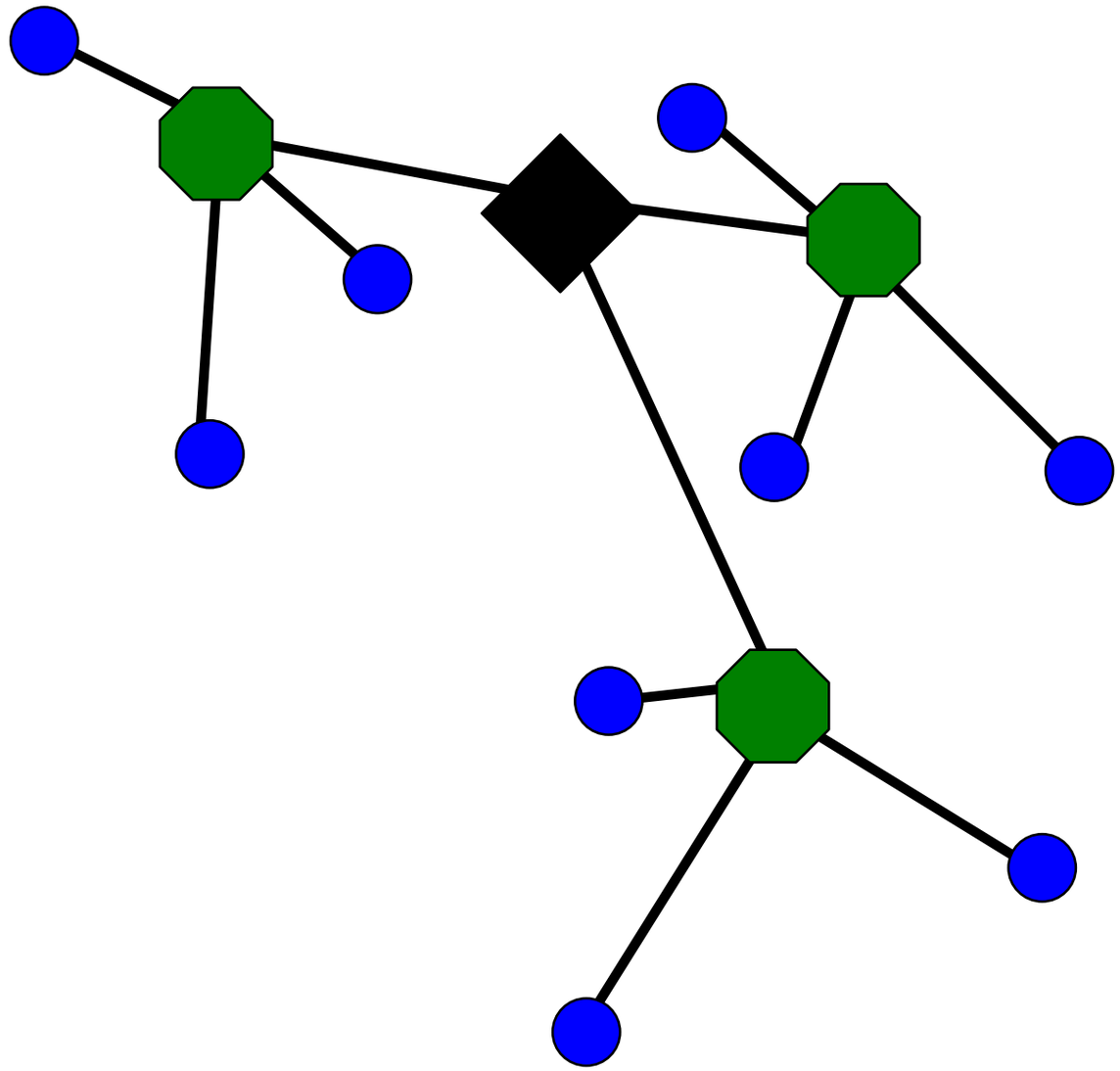


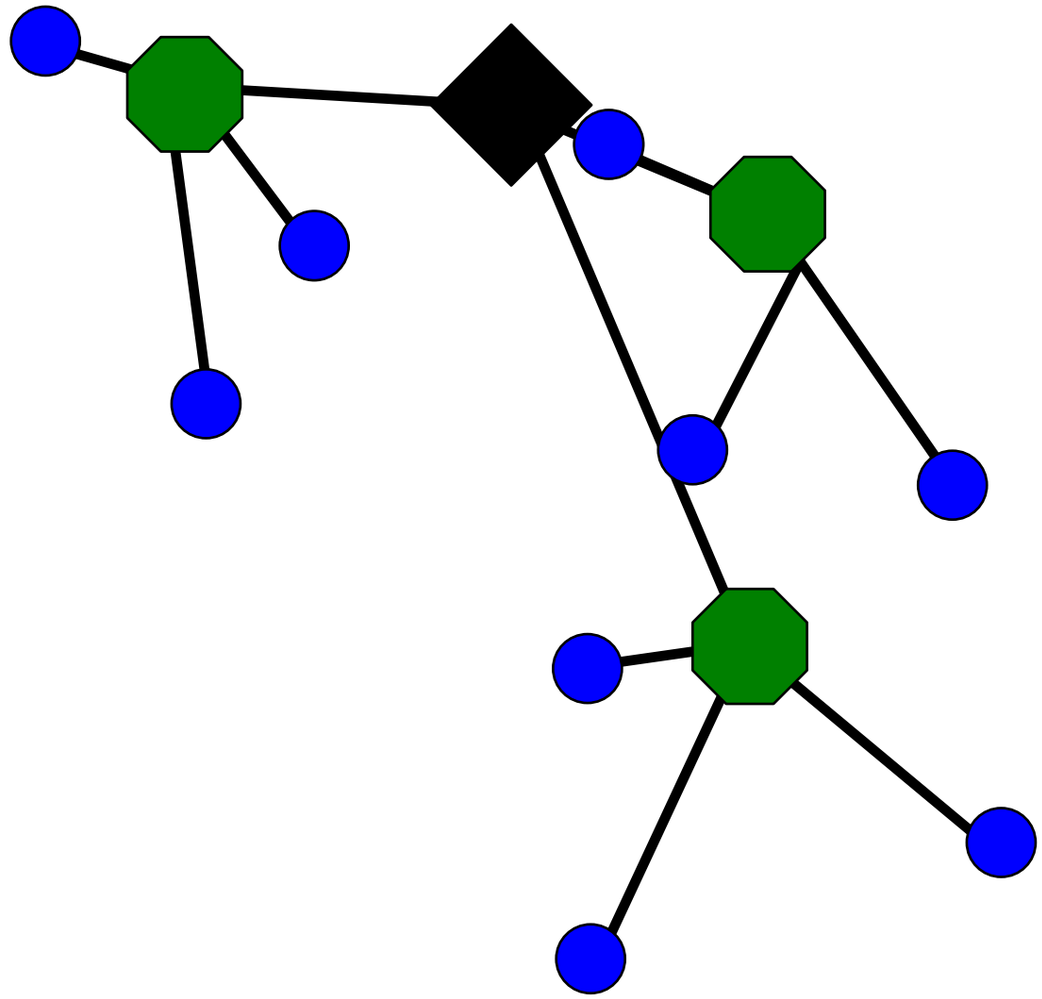


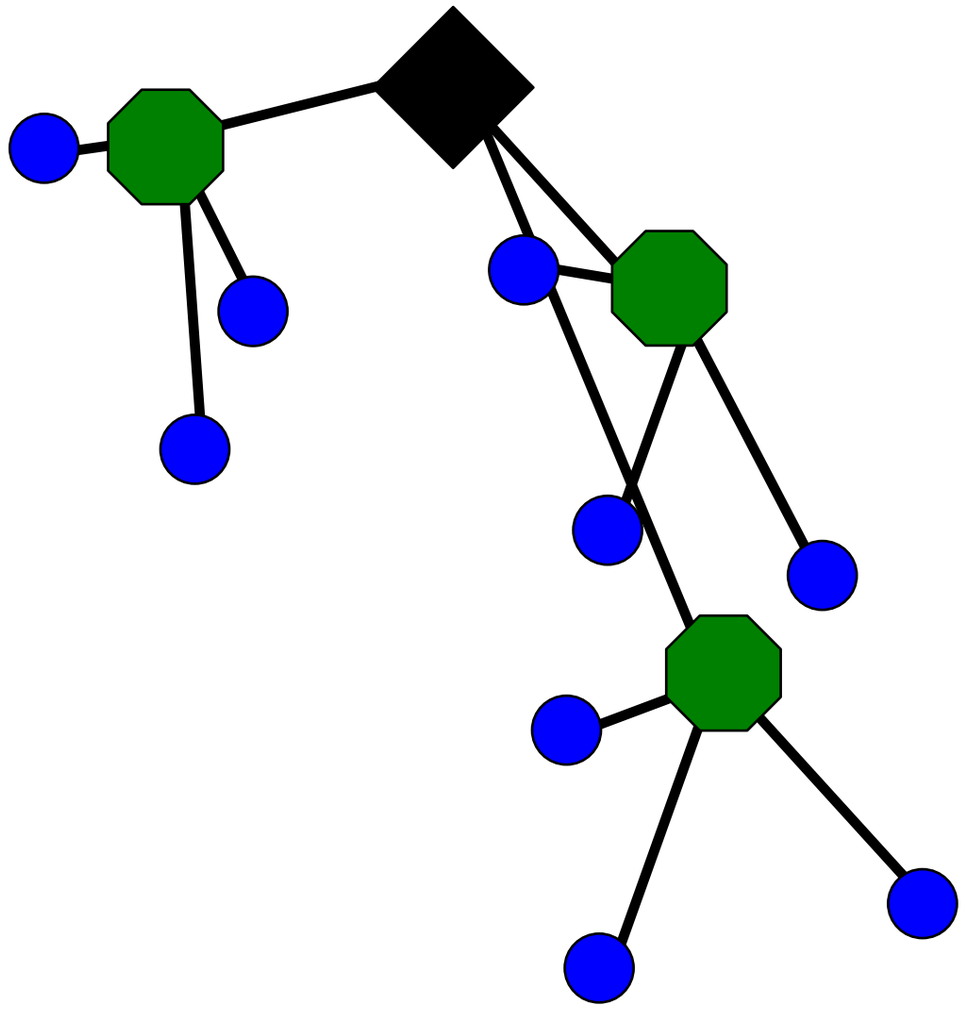


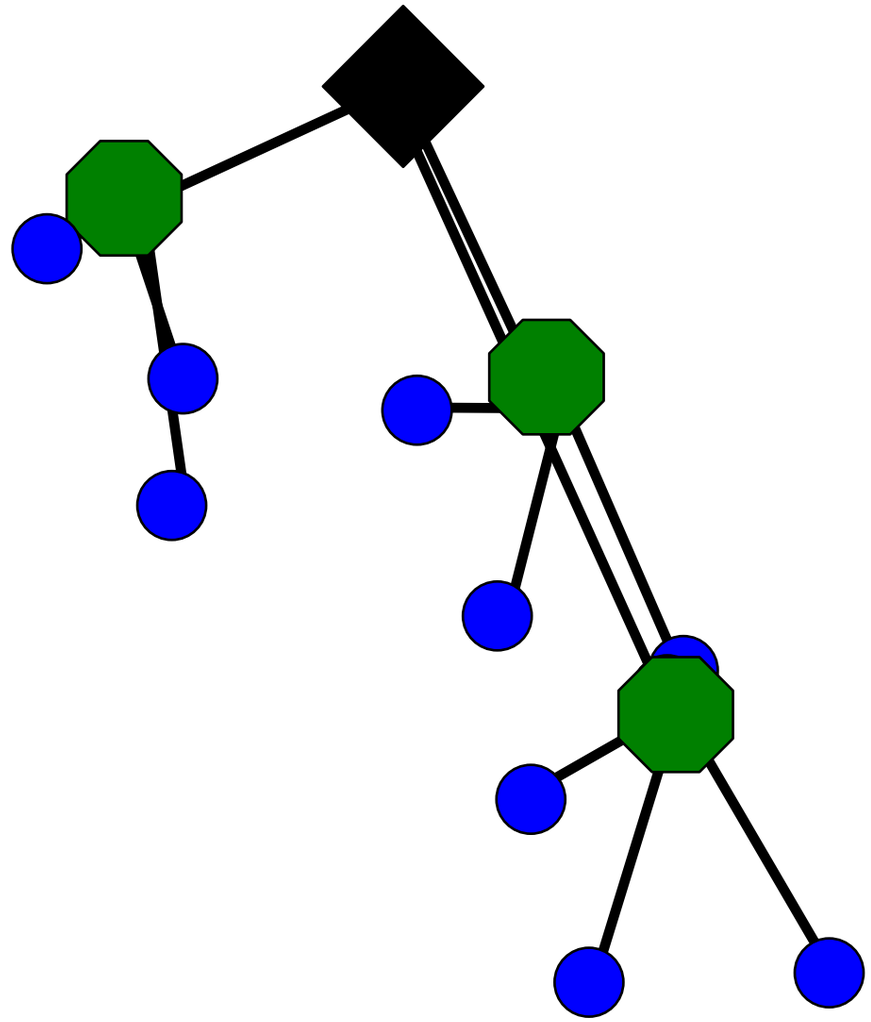


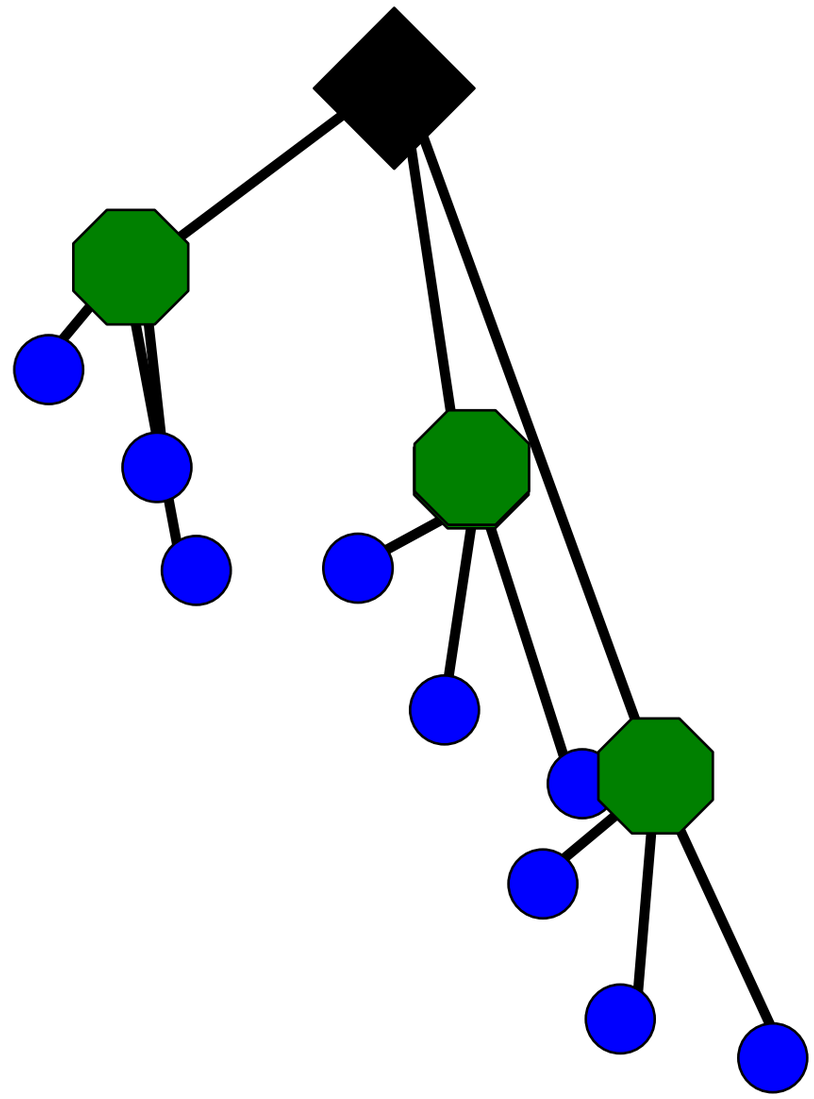


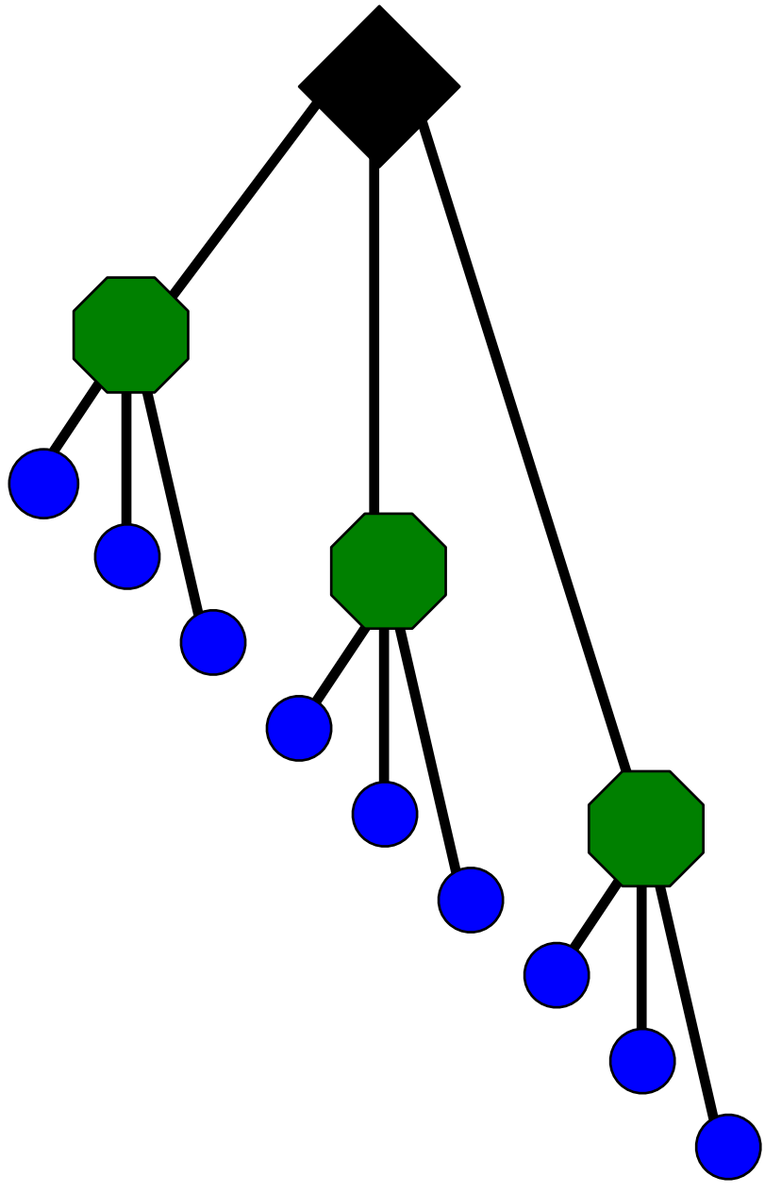


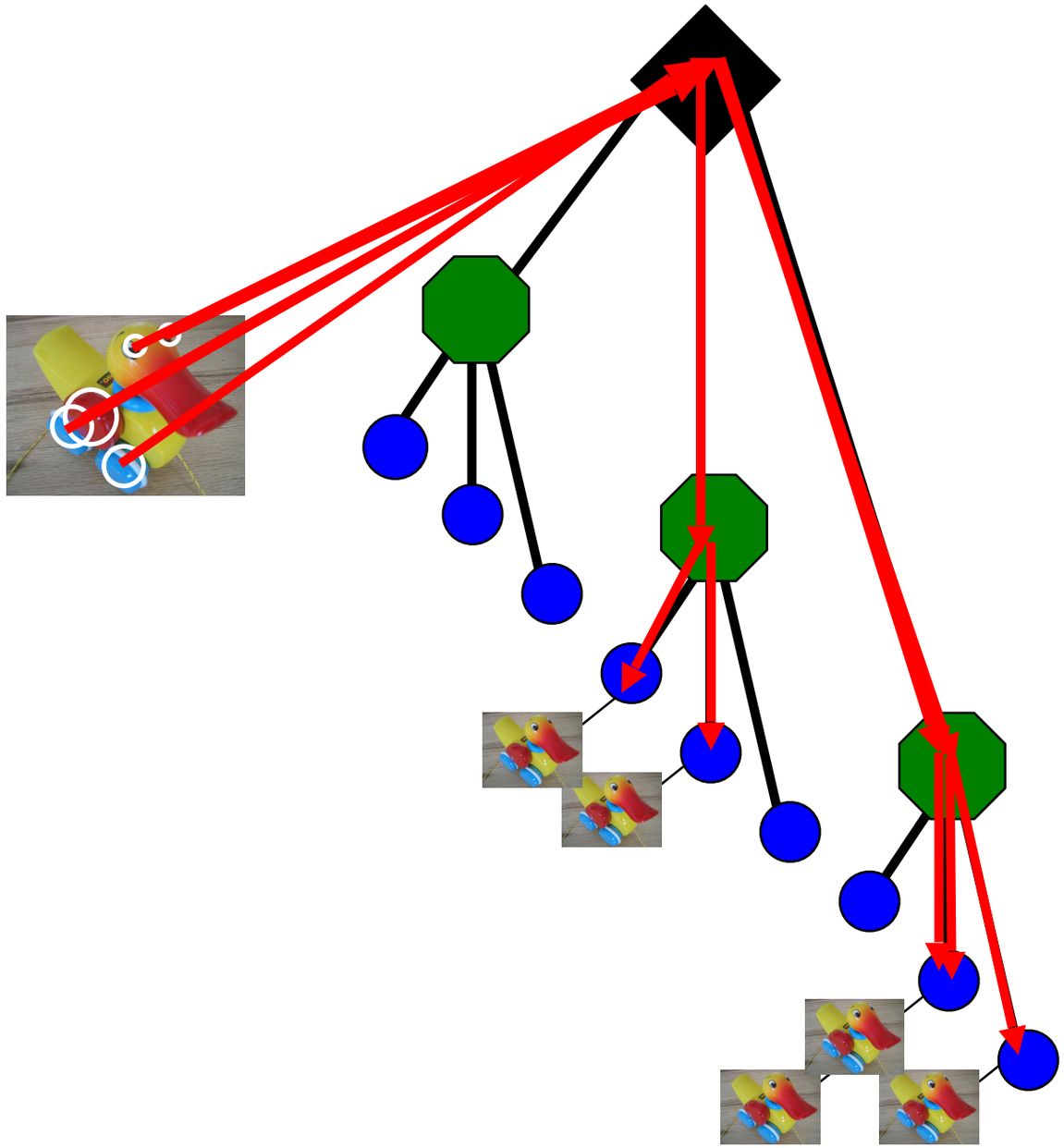


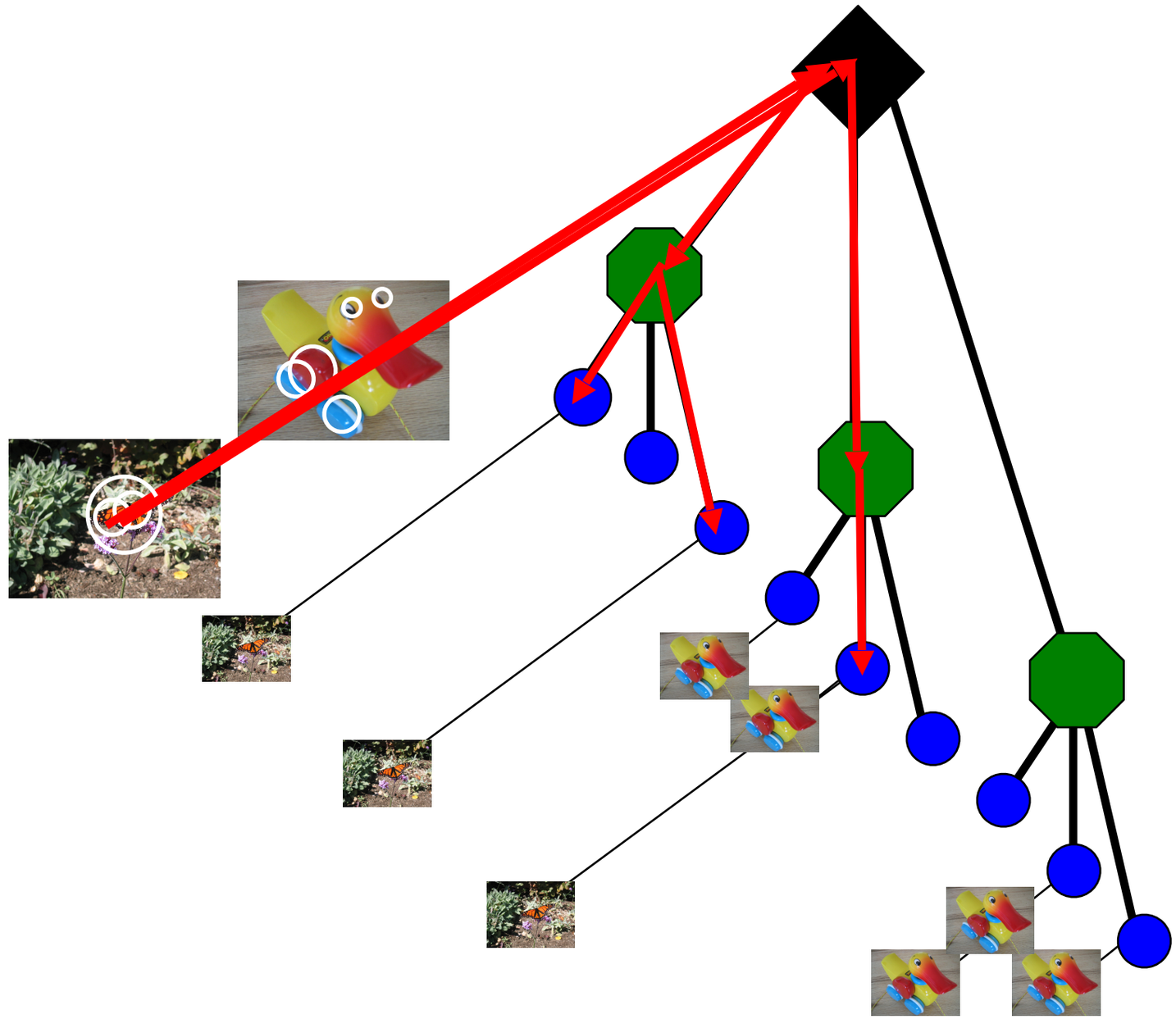


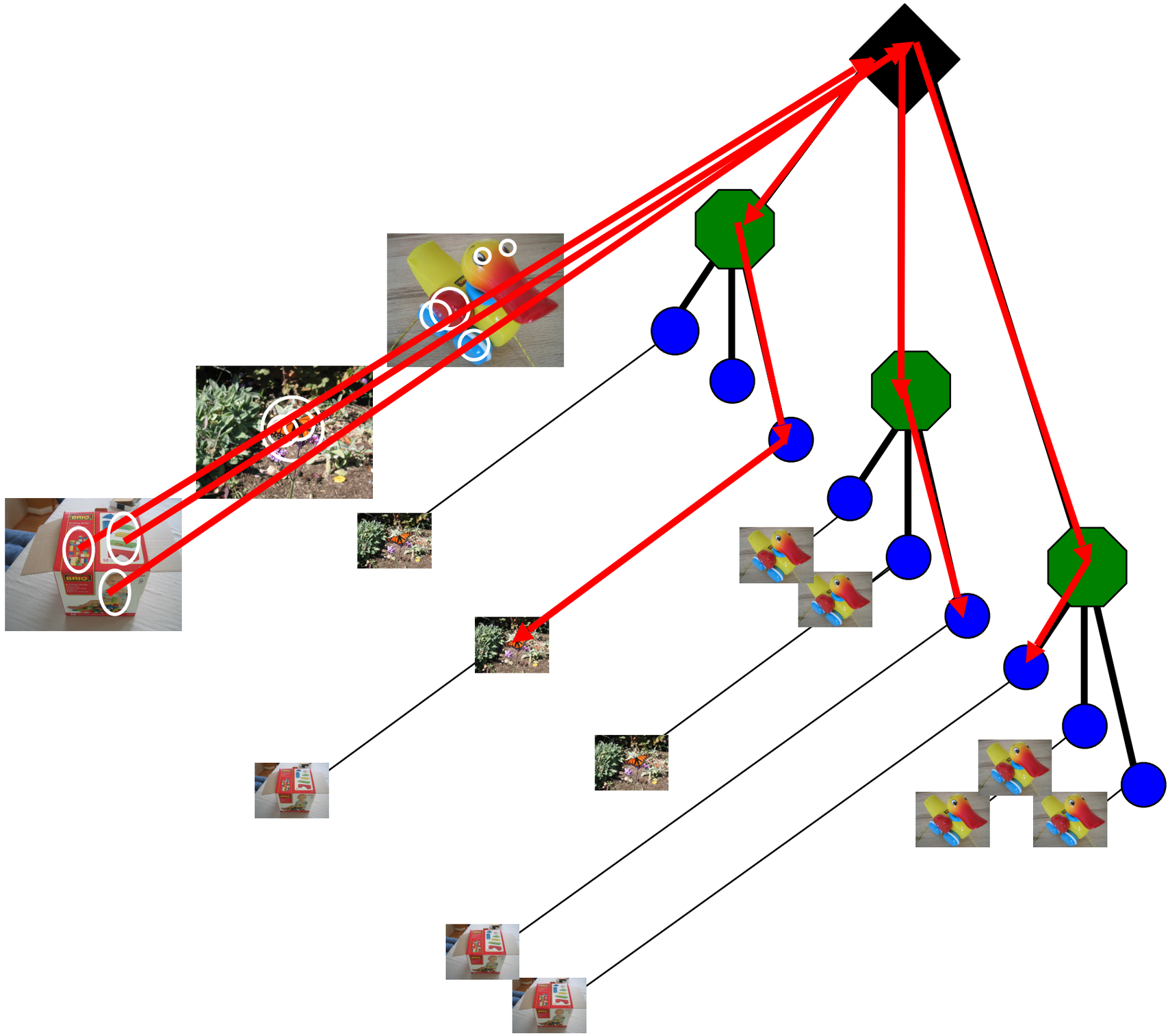


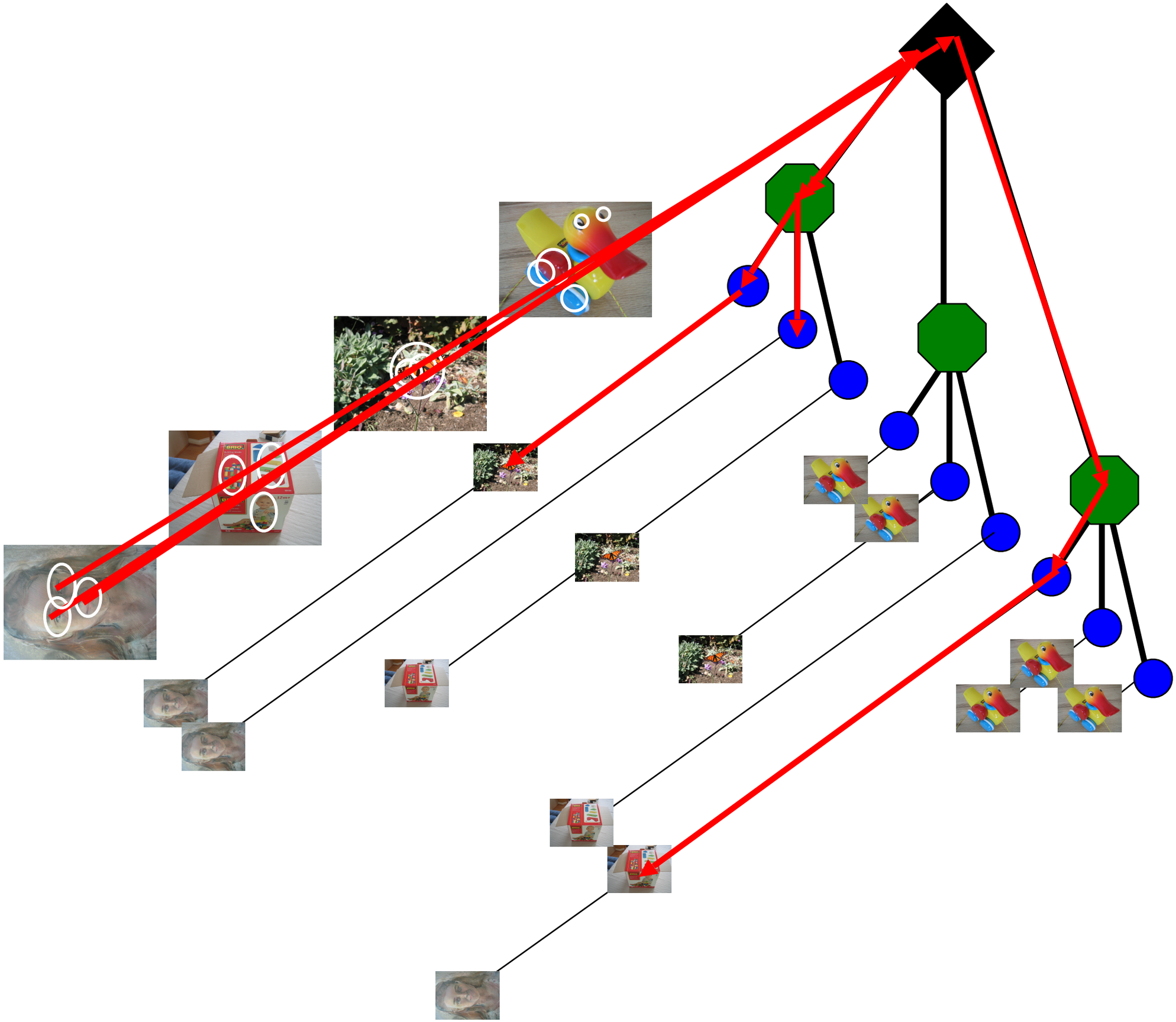


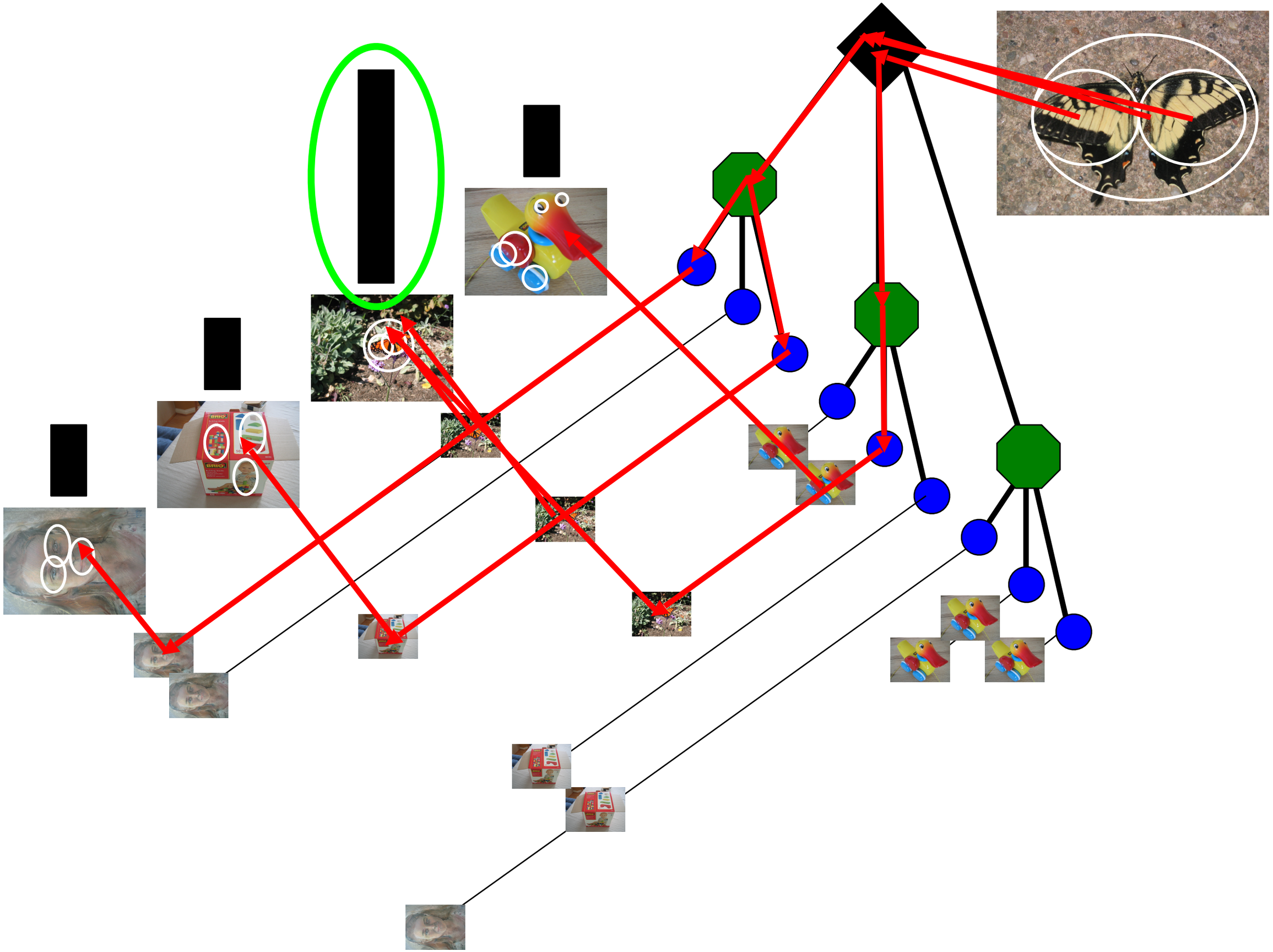




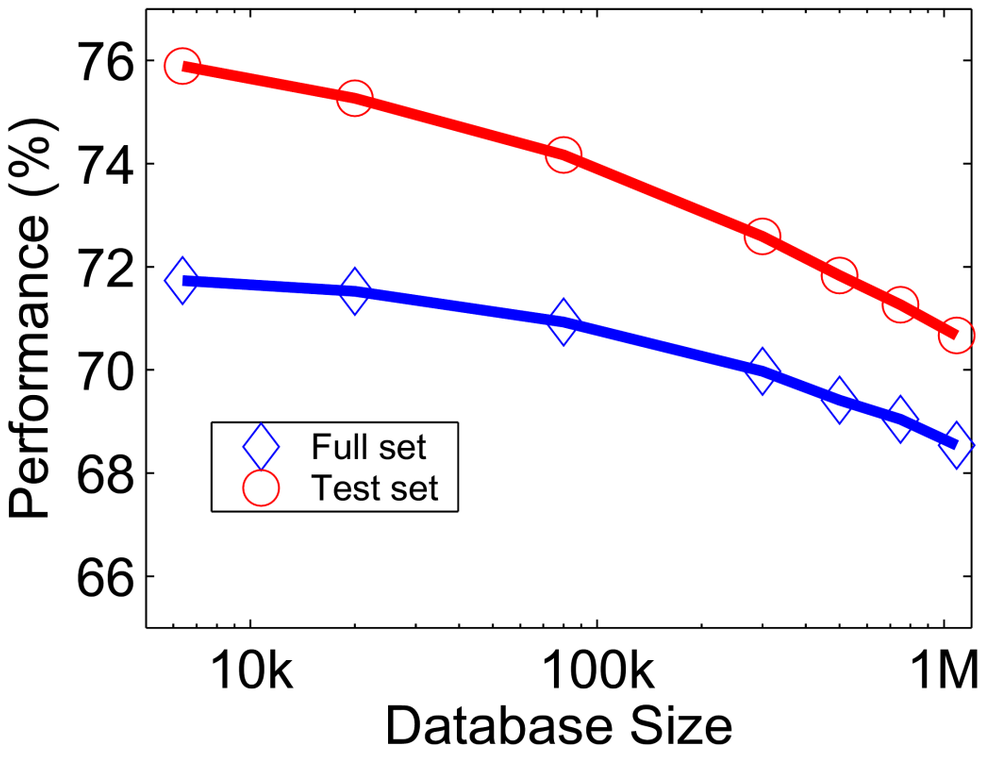


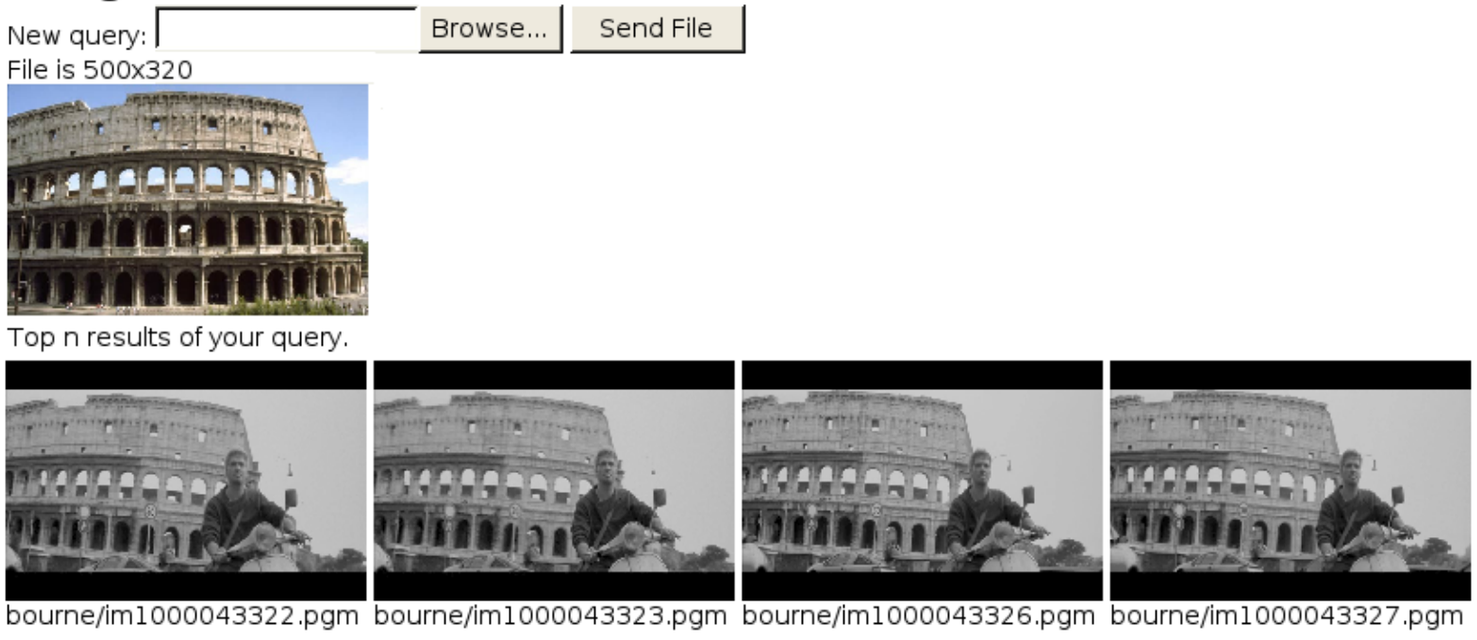






Performance





UK

Center for Visualization & Virtual Environments

Center for Visualization & Virtual Environments

Recognition Benchmark Images

Henrik Stewénius and David Nistér

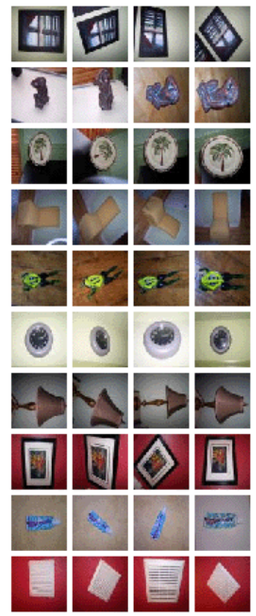
The set consists of 2604 groups of 4 images each for a total of 10416 images. All the images are 640x480.

If you use the dataset, please refer to:
- D. Nistér and H. Stewénius, Scalable Recognition with a Vocabulary Tree, CVPR 2006. PDF

Subsets

For users of subsets of the database please note that the difficulty is dependent on the chosen subset. Important factors are:

1. Difficulty of the objects themselves. CD-covers are much easier than flowers. See performance curve below.
2. Sharpness of the images. Many of the indoor images are somewhat blurry and this can affect some algorithms.
3. Similar or identical objects. All the pictures where taken by CS students/faculty/staff and thus keyboards and computer equipment are popular motives. So is computer vision literature.



Download

Please note BEFORE starting your download that the file is almost **2GB**. Please save a local copy in order to save bandwidth at our server.
- Zipped File.

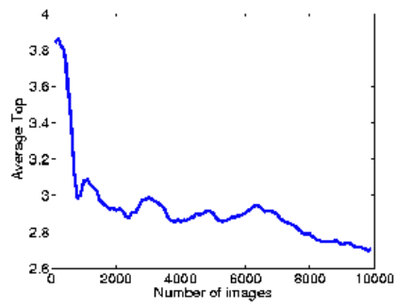
Performance

In the paper we give results either for a subset of 6376 images (all we had at that time) or a smaller subset of 1400 images. The smaller set was used when we did not have an efficient enough implementation in order to handle the larger set.

Performance Measures

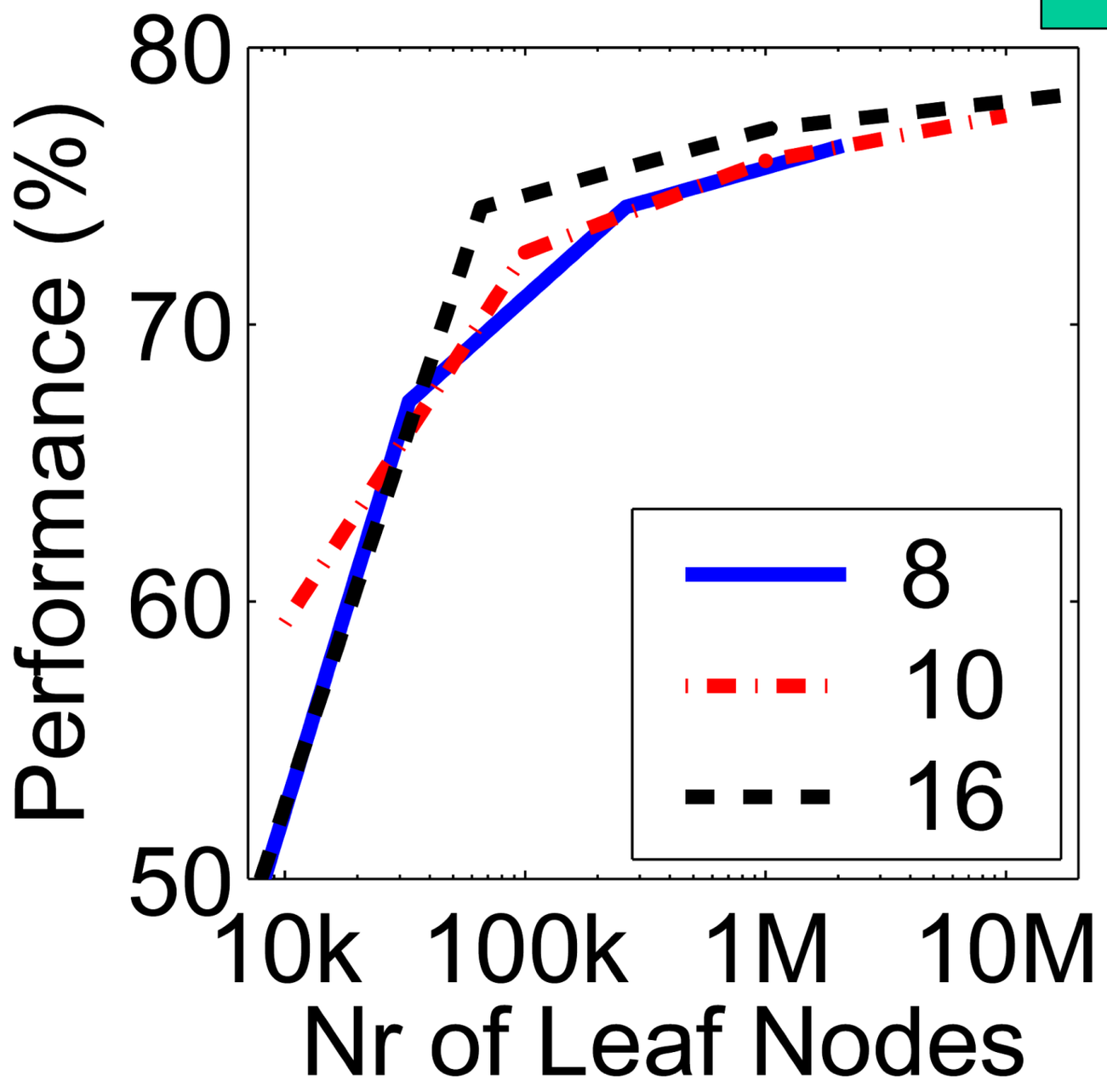
- Our simplest measure of performance is to count how many of the 4 images which are top-4 when using a query image from that set of four images.

A matlab implementation which computes this measure: Download.



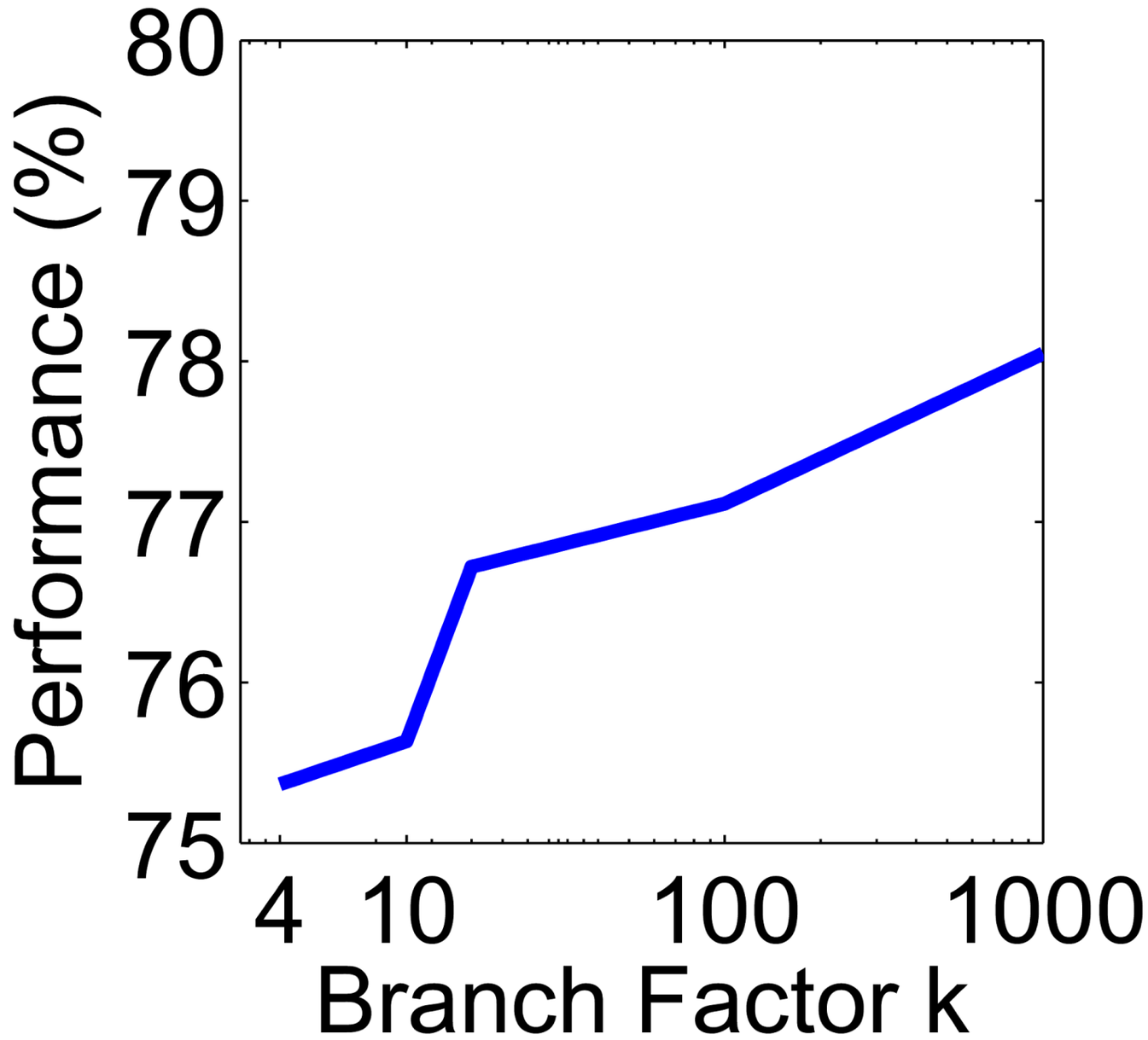
How our performance varies when taking subsets 0:n from the set. These results were run with settings optimized for speed.

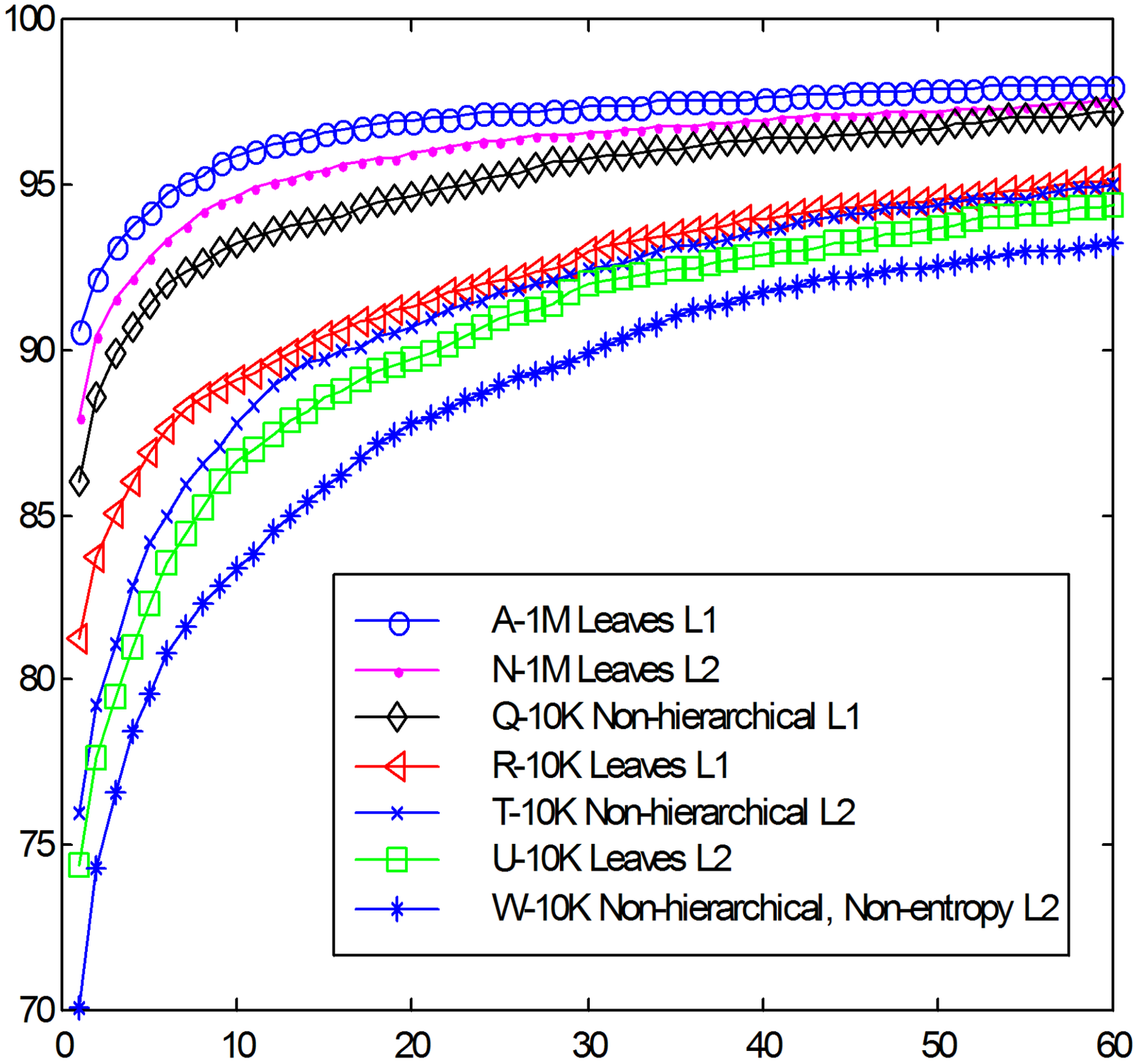
Size Matters



Improves Retrieval

Improves Speed


80
79
78
77
76
75
Performance (%)
4
10
100
1000
Branch Factor k



UK

Center for Visualization & Virtual Environments

Robust to Clutter and Occlusion

- Local Regions
- Like Web-search