50,000 Tiny Videos: A Large Dataset for Non-Parametric Content-Based Retrieval and Recognition

Alexandre Karpenko (University of Toronto)

Main Contribution

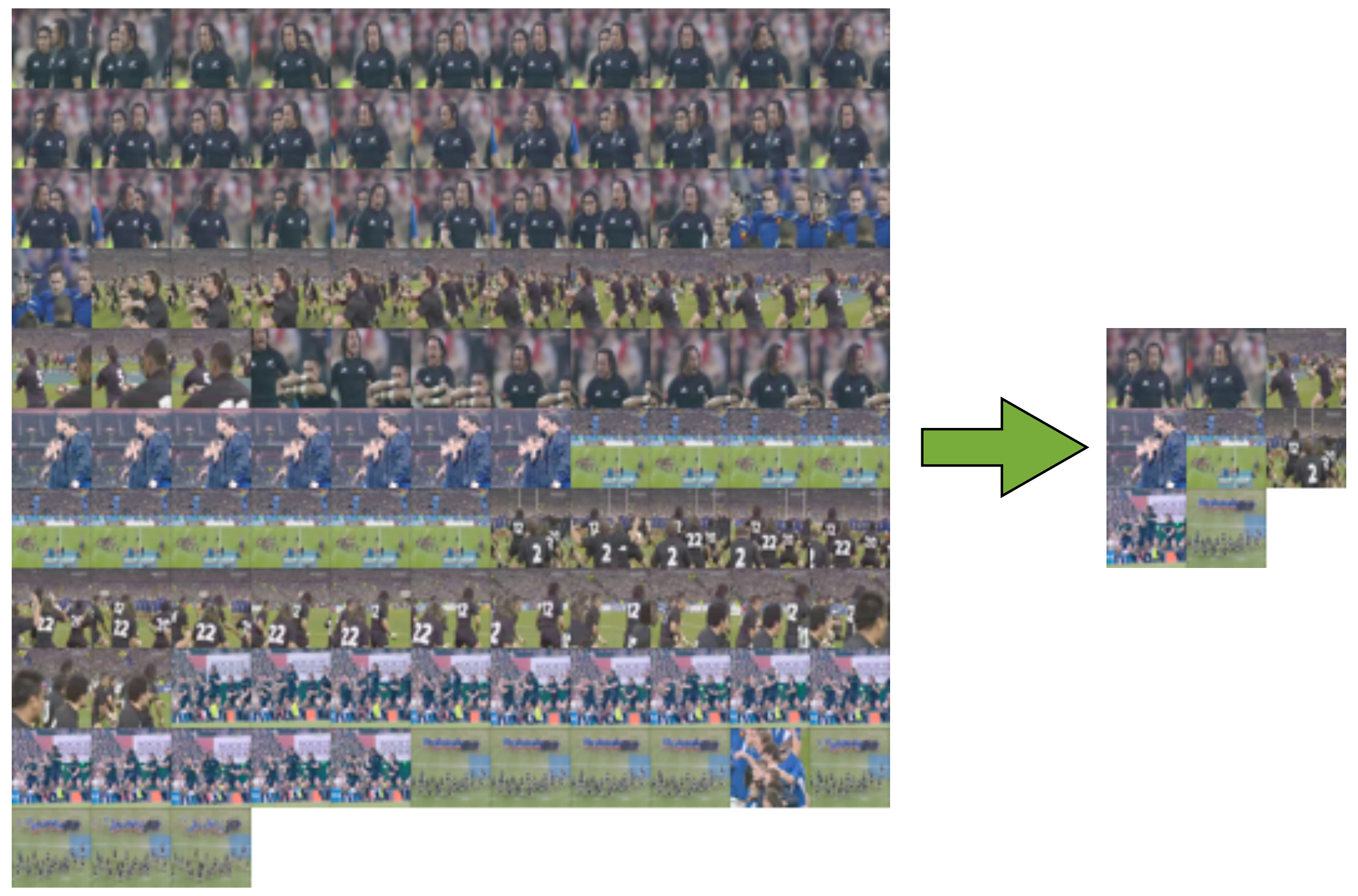
- A data-mining approach to content-based retrieval and recognition using a large collection of videos
- **Applications:**
  - Content-based copy detection
  - Related video retrieval
  - Content-based search
  - Classification

Prior Work

- Tiny Images by Torralba et al.
  - A large dataset of 80 million tiny (32x32) images
  - Dataset has many applications:
    - Person detection and localization
    - Scene recognition
    - Image colorization and orientation detection

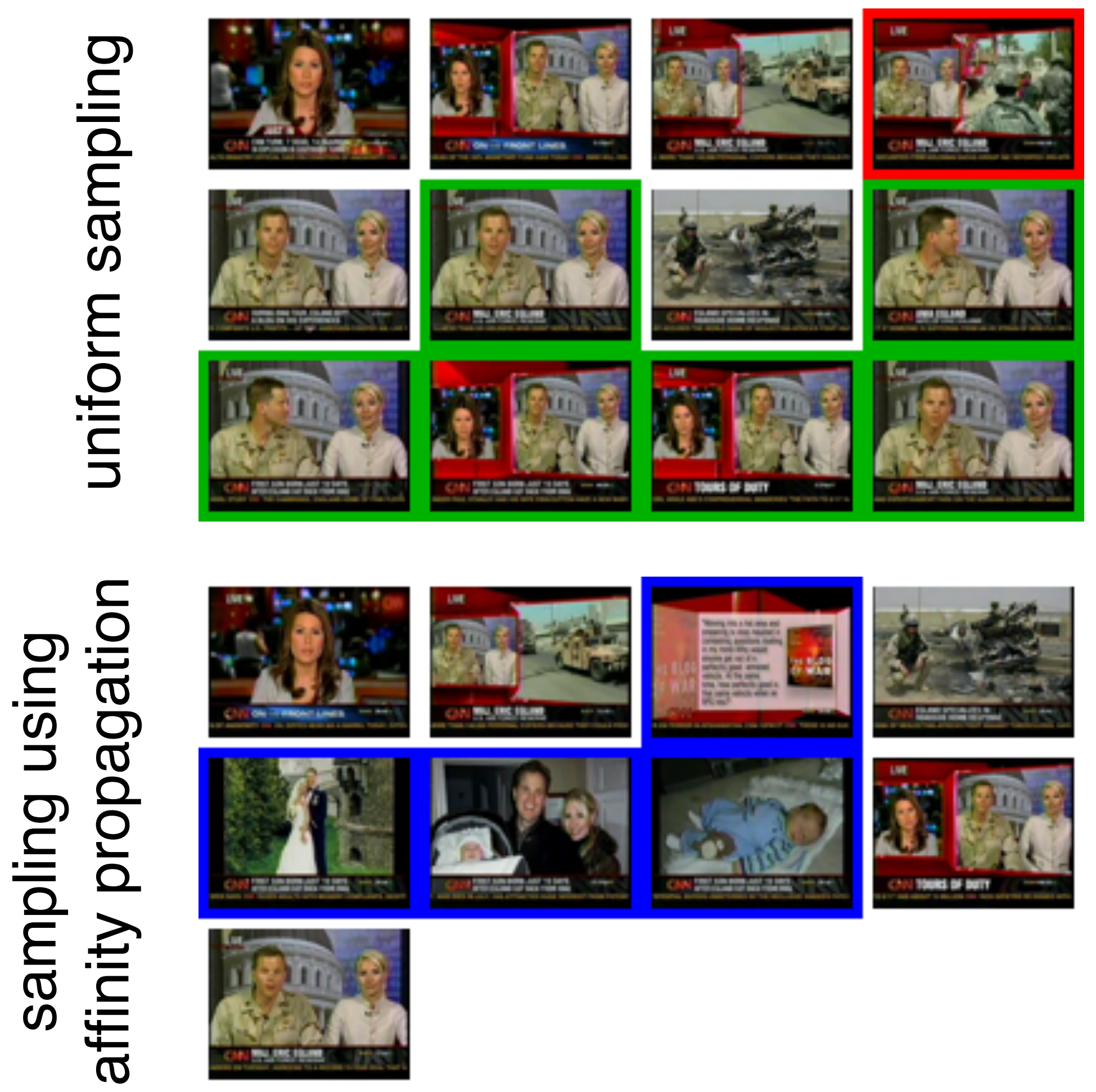
Our Video Database

- Over ***50,000 Videos*** were collected from YouTube over 4 months, in the categories:
  - News, Sports, People, Travel, and Technology
- Occupies ***500GB*** of disk space
- Totals 170 days of continuous playback
- Metadata includes:
  - View count, rating, title, description, category, user assigned tags (lables), etc
- Largest labelled research database to date
- Small compared to YouTube (>100M videos)



Tiny Video Representation

- Uses same descriptor for frames as tiny images:
  - Frames are resized to 32x32 pixels and normalized to zero mean and unit magnitude
- Videos are represented with only a few ***keyframes*** using an exemplar-based clustering algorithm called ***Affinity Propagation***



Similarity Metrics

- Between two ***frames*** I_a and I_b
  - sum of squared differences:

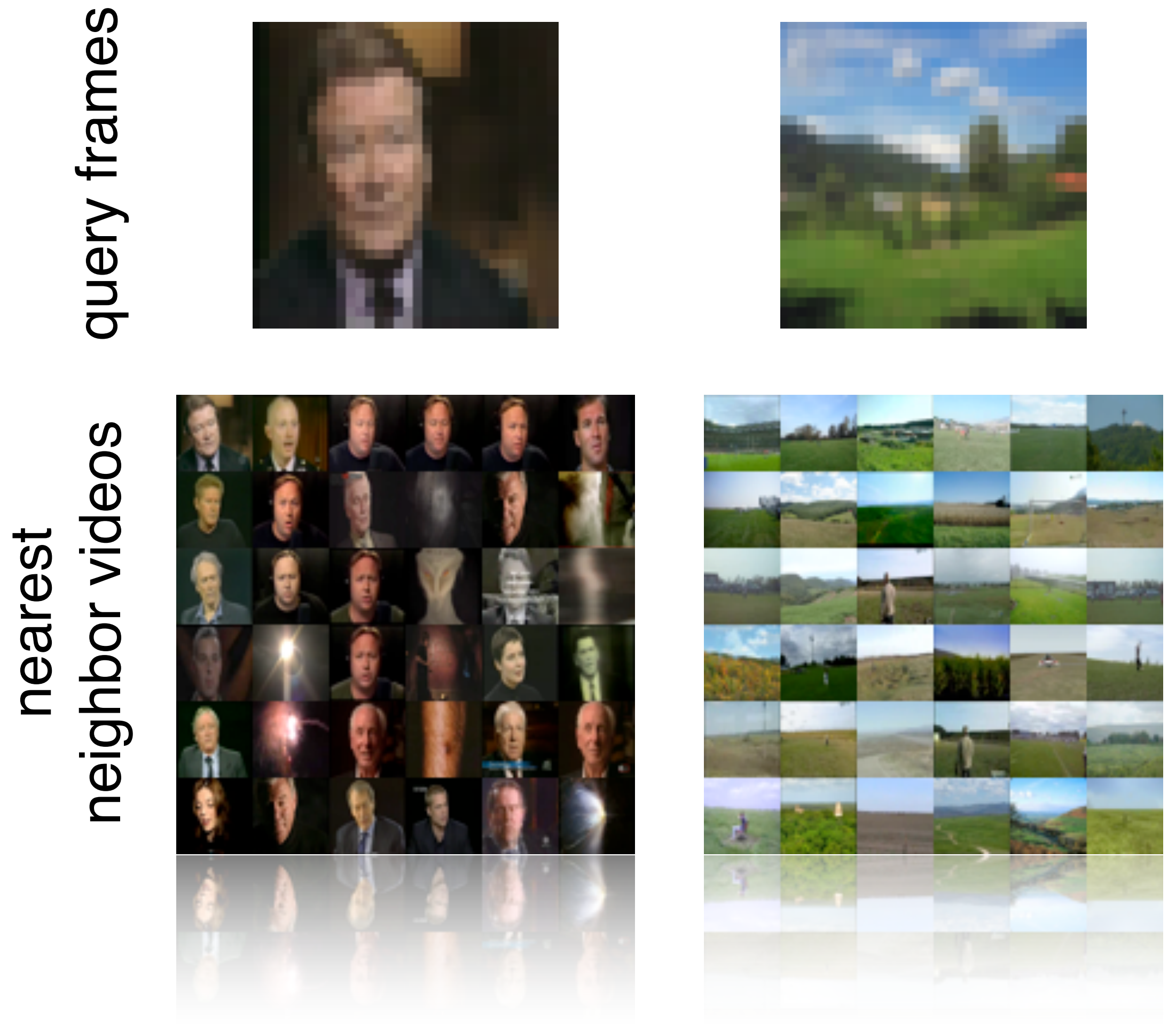
$$D^2_{ssd}(a,b) = \sum_{x,y,c} (I^*_a(x,y,c) - I^*_b(x,y,c))^2$$

  - Improved metric allows pixels to shift slightly:

$$D^2_{shift}(a,b) = \sum_{x,y,c} \min_{|D_{x,y}| \leq w} (I^*_a(x,y,c) - \hat{I}^*_b(x+D_x, y+D_y, c))^2$$

- Between two ***videos*** V_α and V_β

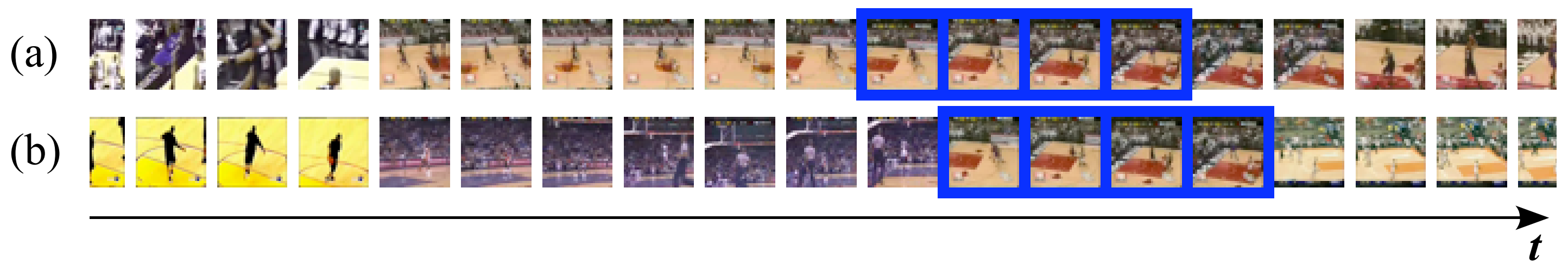
$$\hat{D}^2_{ssd/shift}(\alpha,\beta) = \min_{I_a \in V_\alpha, I_b \in V_\beta} (D^2_{ssd/shift}(a,b))$$

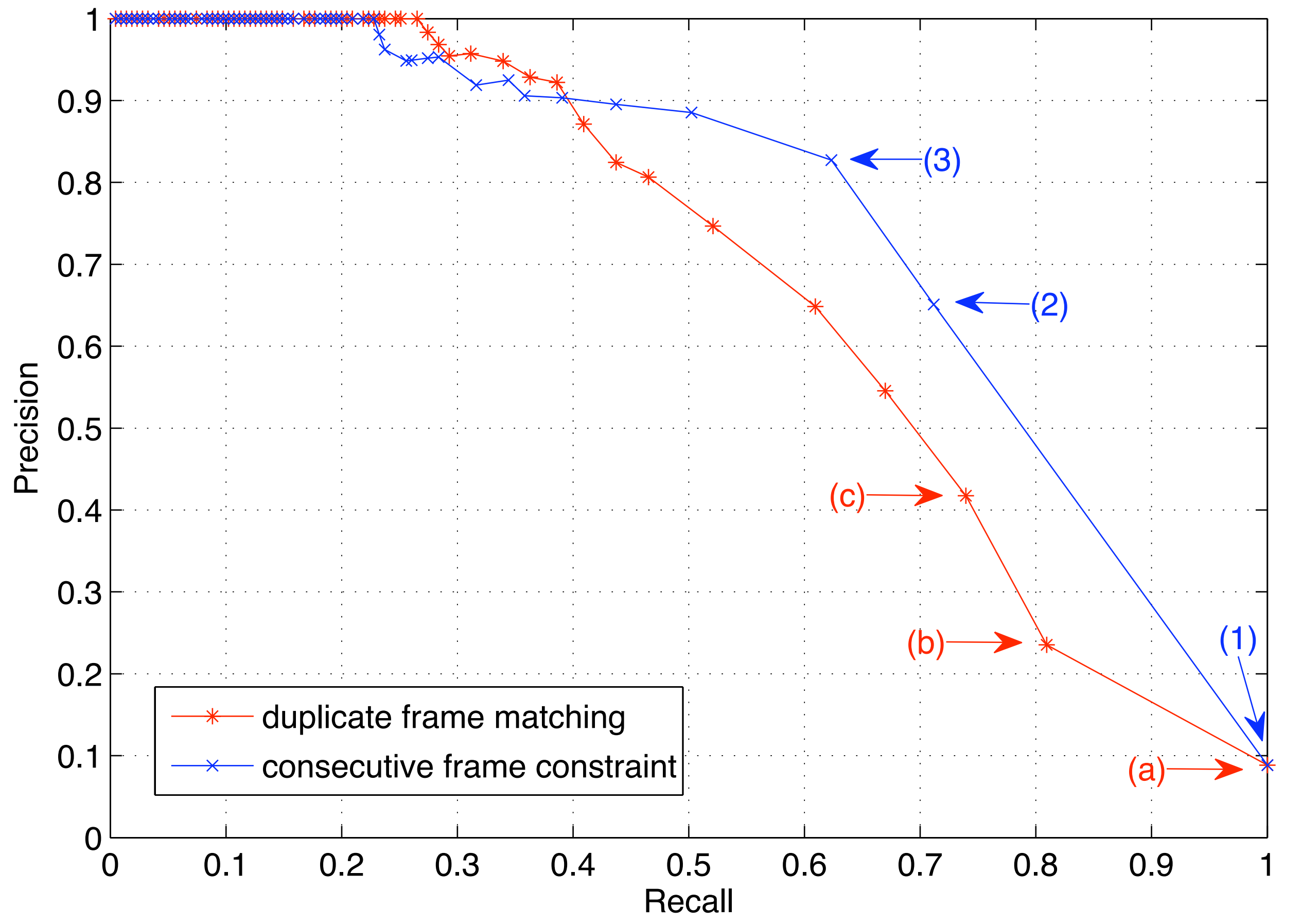


Applications

1. Related Video Retrieval

- Find duplicate videos by detecting frames with high correlation
- 16% of videos in our YouTube dataset have at least one duplicate shot

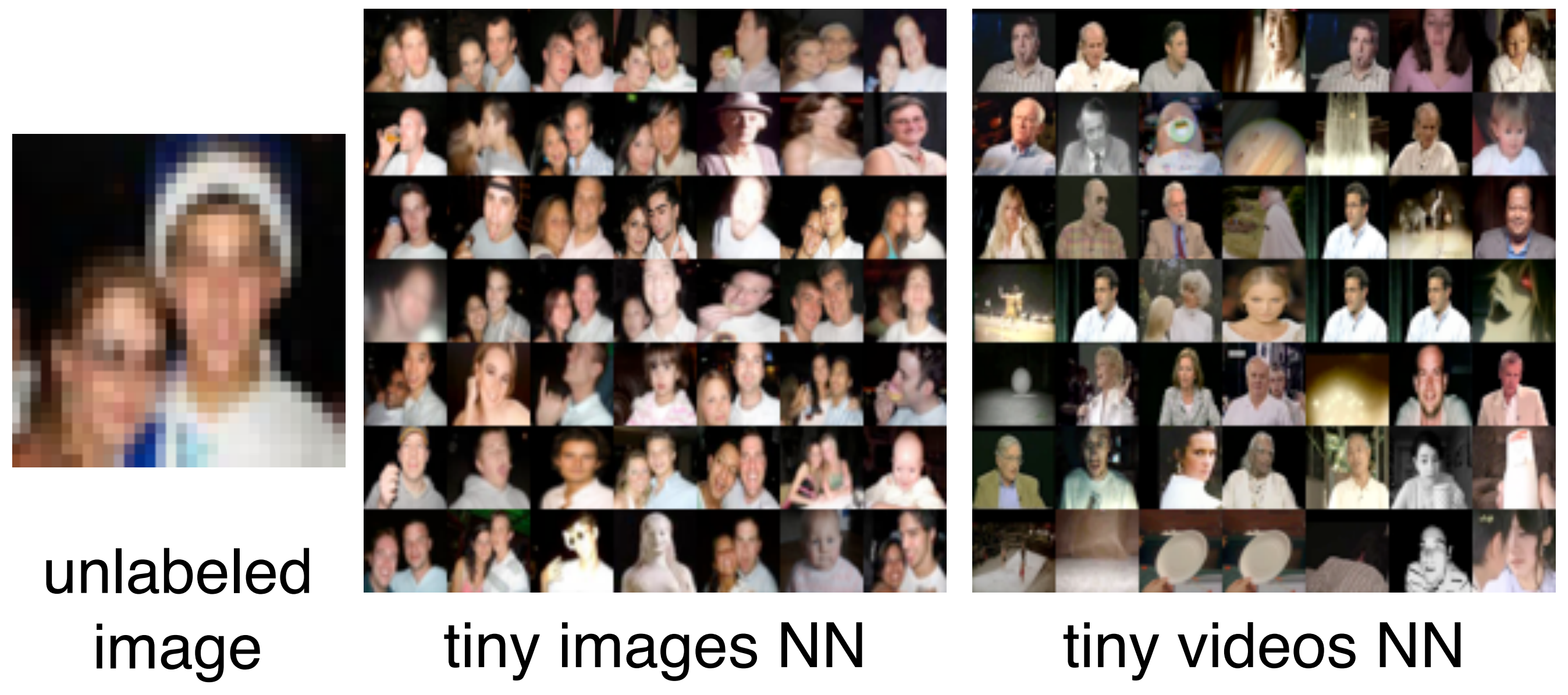




- We detect duplicate videos with high precision and recall
- Achieve a perfect score on the MUSCLE-VCD-2007 content-based copy detection evaluation corpus

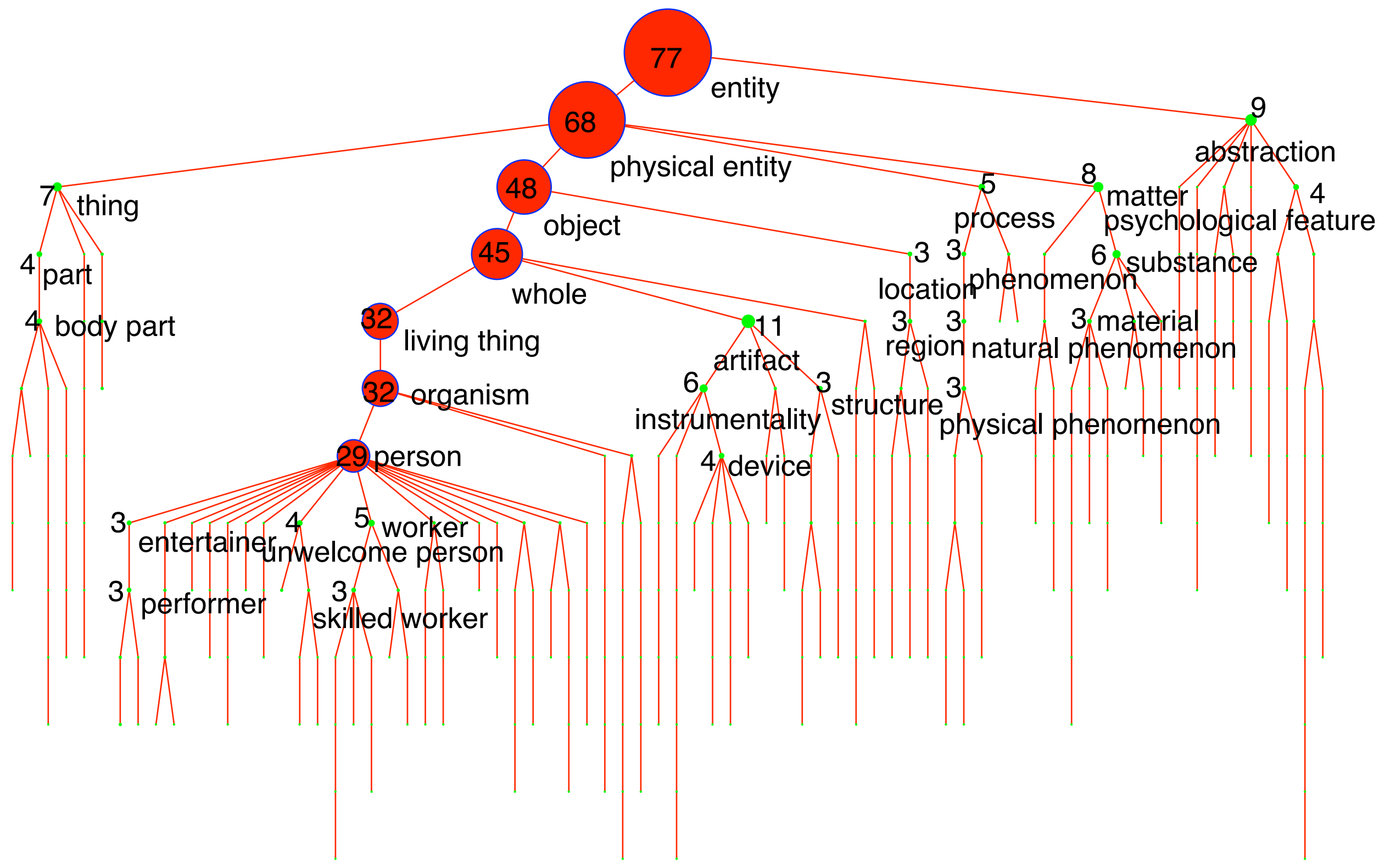
2. Classification with WordNet

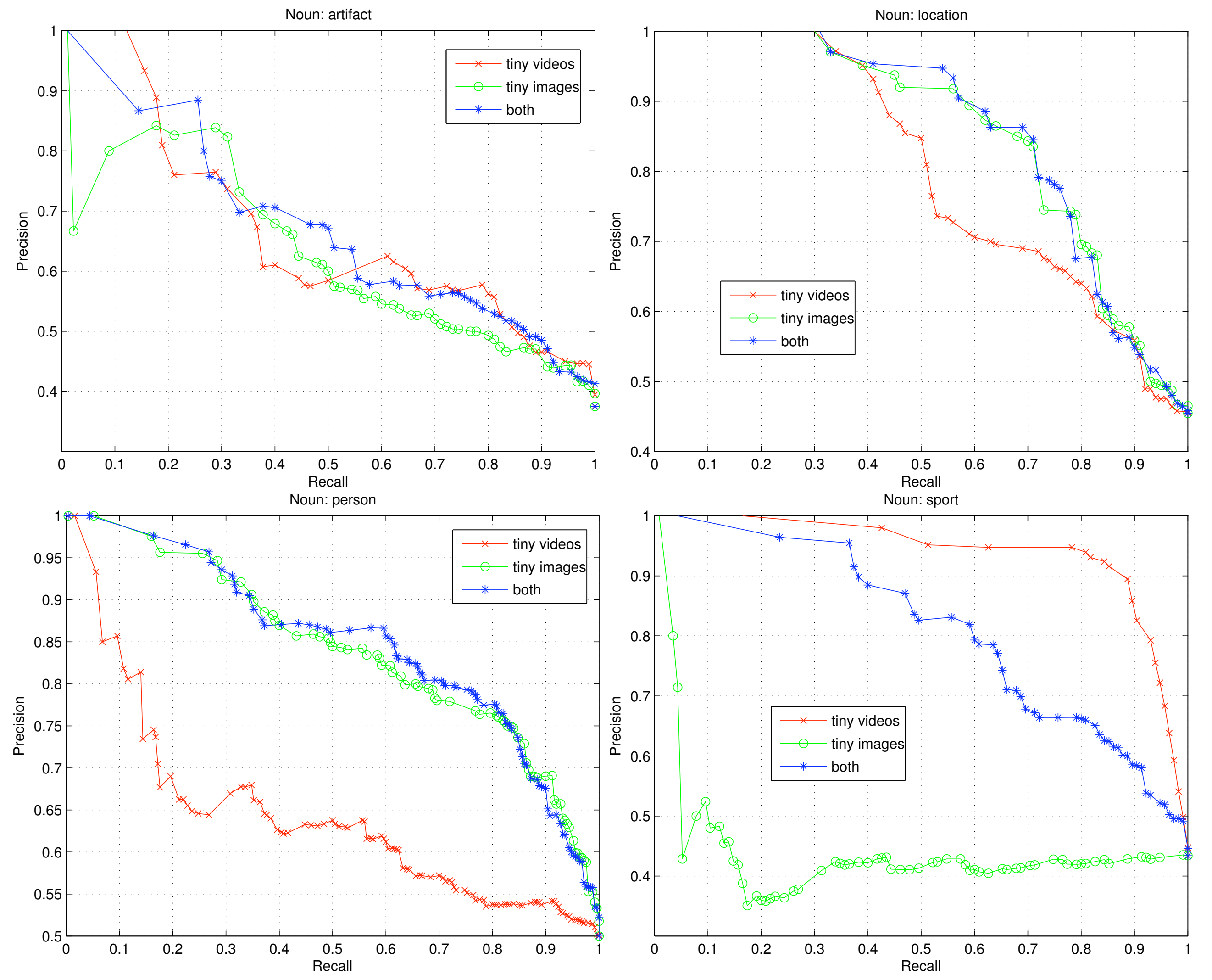
- Find k-NN for an unlabeled input image
- Use labels of NN to vote for a category by accumulating votes at multiple semantic levels (in order to reduce labeling noise)



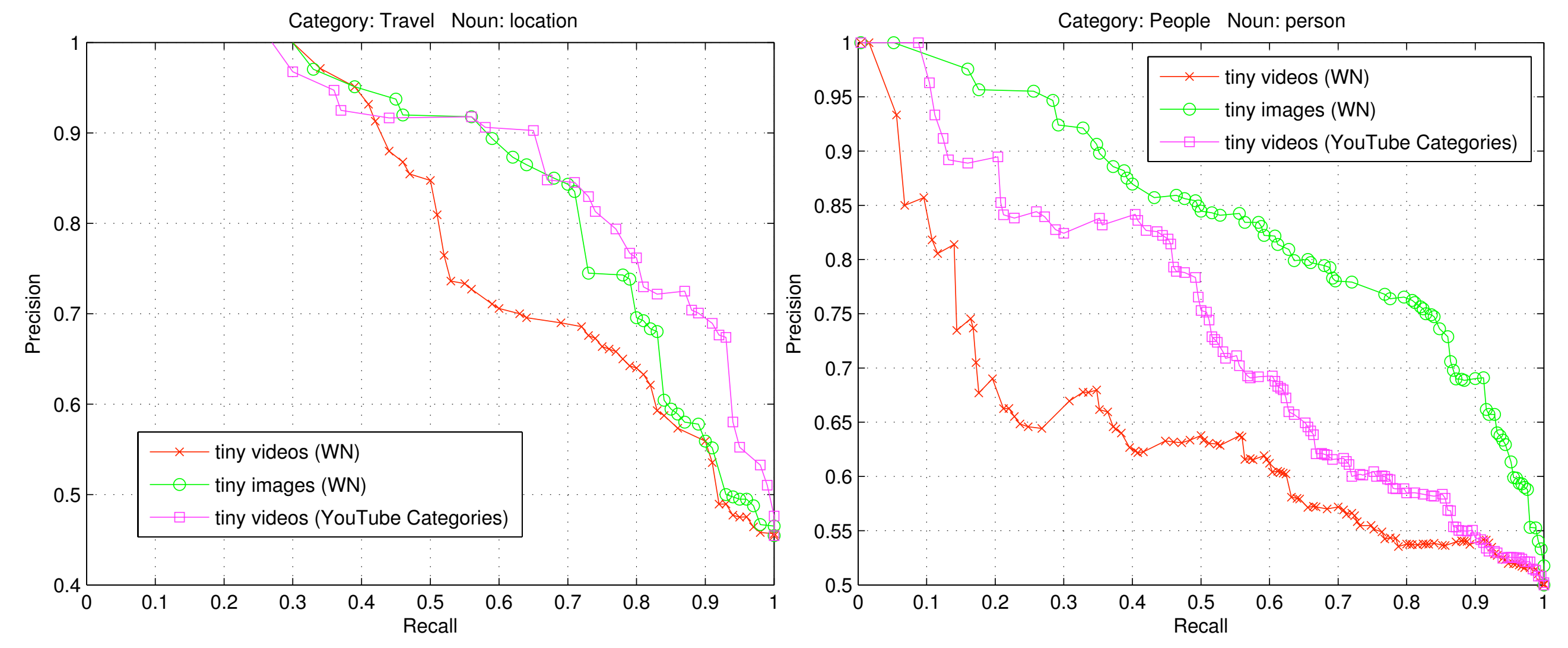
unlabeled image tiny images NN tiny videos NN

- Videos generally focus on activities, while images on objects
- Both datasets can be combined to improve classification results in a wider range of cases





- Can use additional metadata for tiny videos to further improve classification for some categories



Conclusion & References

- A large amount of data can aid a variety of computer vision tasks. Other applications include:
  - optical flow prediction for single images
  - semantic video segmentation

Tiny Videos: Non-Parametric Content-based Video Retrieval and Recognition
A. Karpenko and P. Aarabi

80 million tiny images: a large dataset for non-parametric object and scene recognition
A. Torralba, R. Fergus, W. T. Freeman