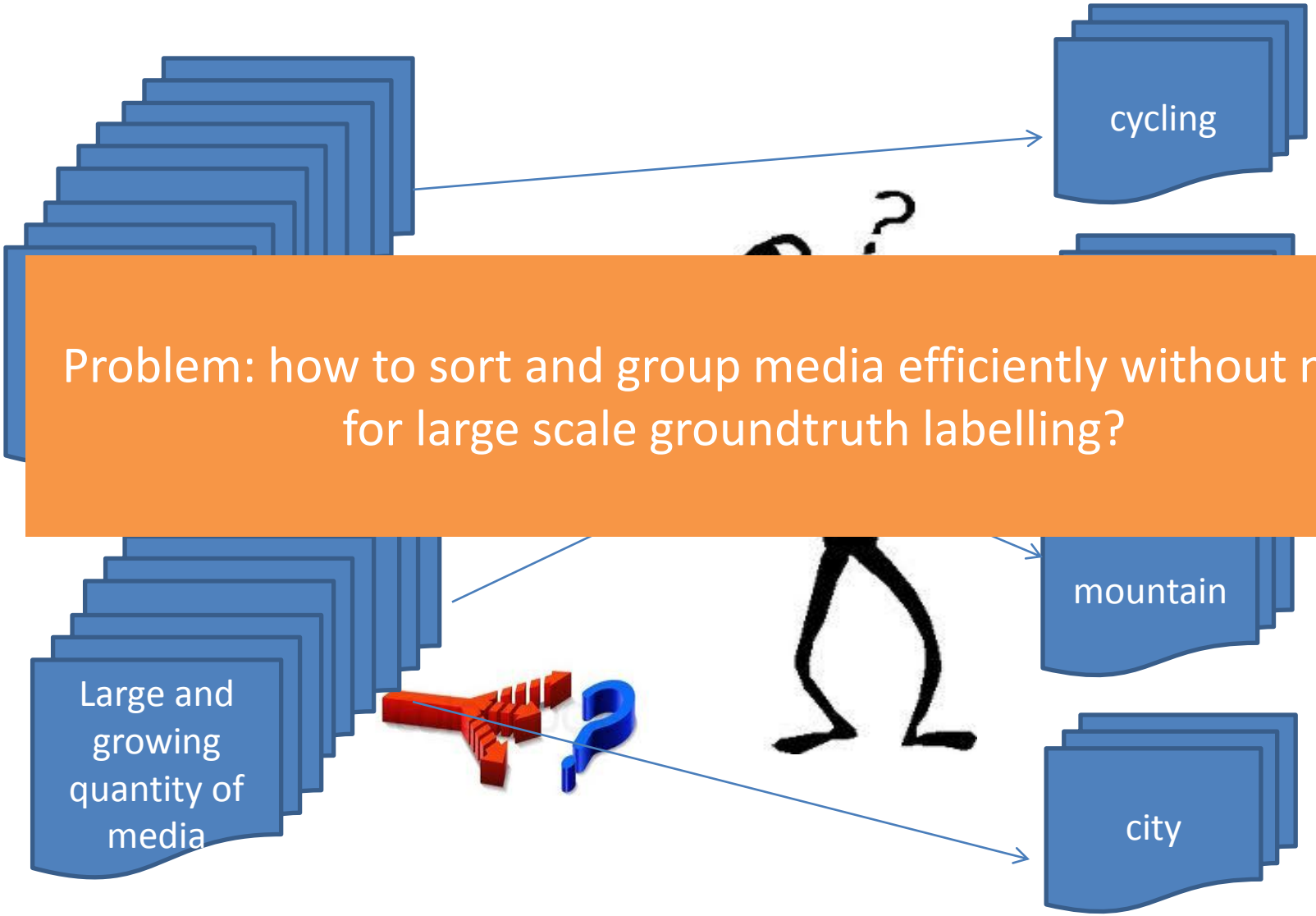
Paper

iGroup: Weakly supervised image and video grouping, A. Gilbert and R. Bowden, ICCV 2011.

Ahmet BUĞDAY

- Introduction
- Related Work
- Input Data Signature
- Similarity of signatures
  - The Min-Hash Algorithm
  - Histogram Weighting Approximation
- Expanding Signatures Through Co-occuring Discriminatory Features
- Iterative Signature Learning
- Results
  - Image
  - Video
  - Computational Costs
- Conclusion

Introduction

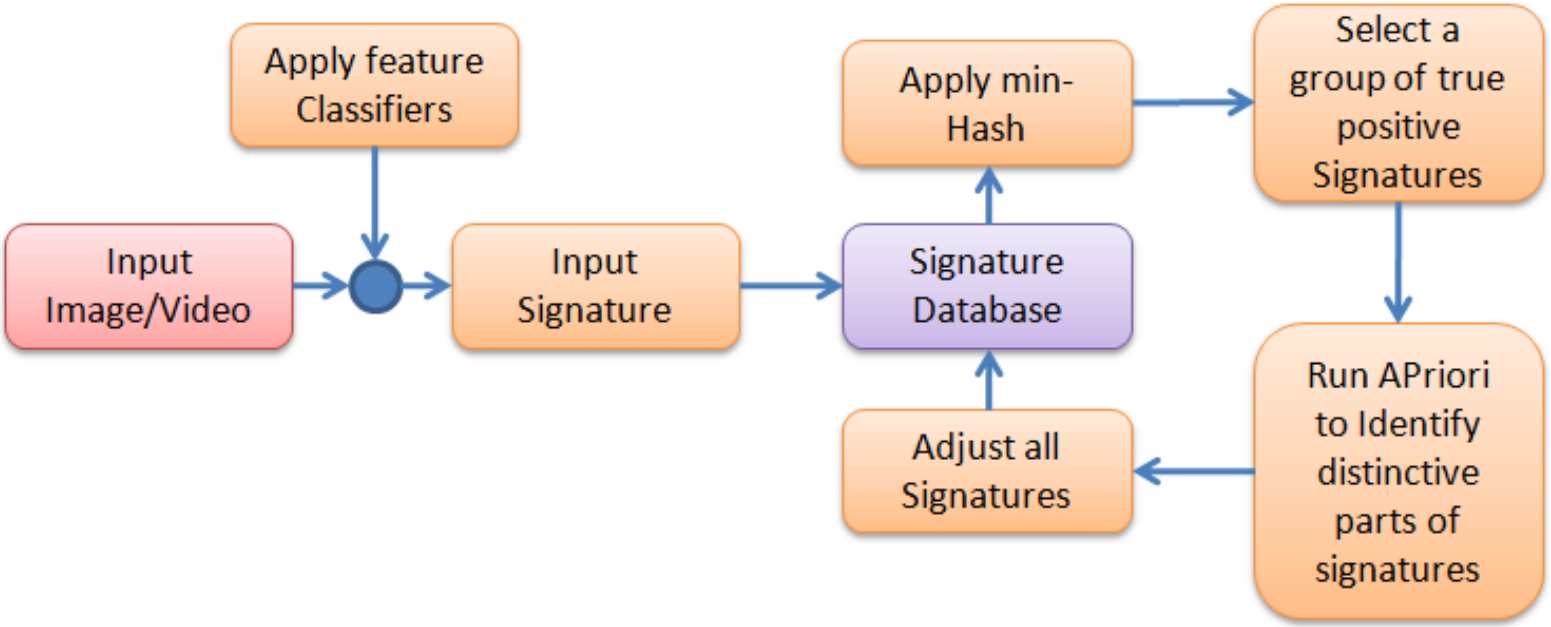


Problem: how to sort and group media efficiently without need for large scale groundtruth labelling?

Introduction

- Generic, efficient and iterative algorithm for clustering classes of images and videos
- Weakly supervised
  - Users pick a few images or videos that belongs the same class or group
- Works with different datasets not like single example or one shot learning approaches
  - Sensitive to training examples ability to generalise
  - Applied to simple staged datasets
- Data mining tools originally developed for text analysis used and extended
  - Min-Hash
  - Apriori
- Inspired by the Bag-of-Words (BoW), image signature is introduced as a simple descriptor

Introduction



An overview of the proposed approach

Introduction

1. Use feature classifier to form initial signature
2. Signature is converted from a weighted histogram into the min-Hash representation
   - facilitate high speed similarity measures between the input samples
   - User can utilise to select M true positive and a single false positive signature
3. Selected signatures are then mined to identify the distinctive combinations of words
4. The rules of mining are then converted into new compound visual words and appended to all the signatures
   - Effect of pulling the signatures from the positive examples closer together
5. This process is then iterated, allowing a user to iteratively cluster data

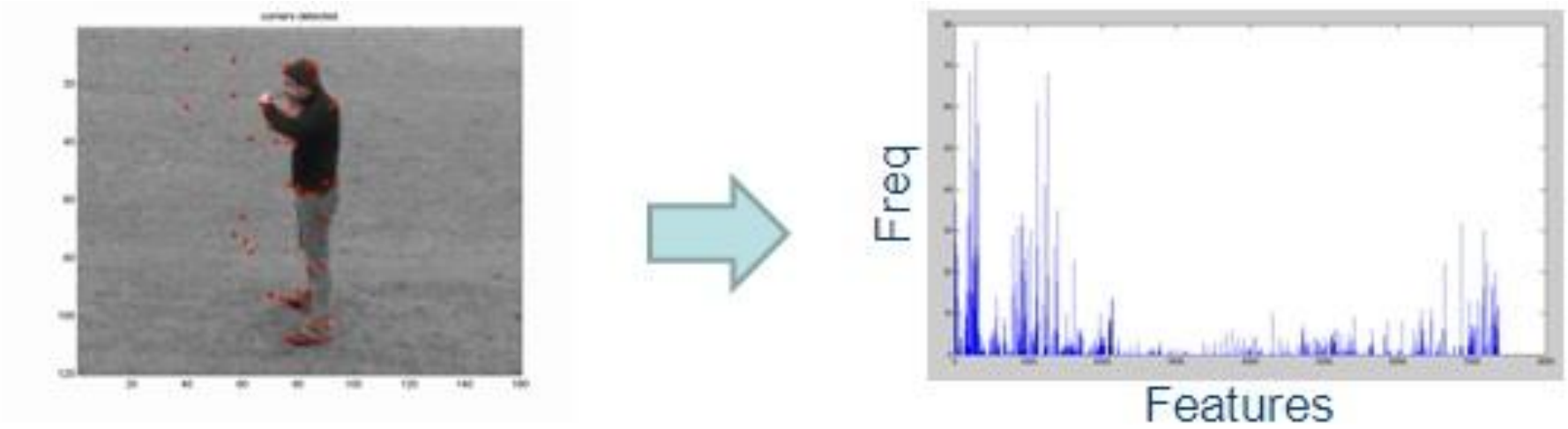
Related Work

- There has been a number of approaches that utilise data mining's ability to work efficiently with large amounts of images or video
- Within image domain
  - Quack et al. applied association rule by mining spatially grouped SIFT features.
  - Chum proposes a min-Hash based approach to detect rare co-occurences in a similar fashion to APriori data mining

Related Work

- Within temporal domain
  - Gilbert et al. take an over complete set of Harris corners [13], group them spatially and temporally and mine out the optimal feature combinations, to be used to classify the video sequences.
  - Chum et al.[6] demonstrated the ability of min-Hash to efficiently identify near duplicate images within datasets
  - Chum et al. [7] propesed an efficient fast method to approximate the histogram intersection of images to improve near duplicate image detection.

Input Data Signature



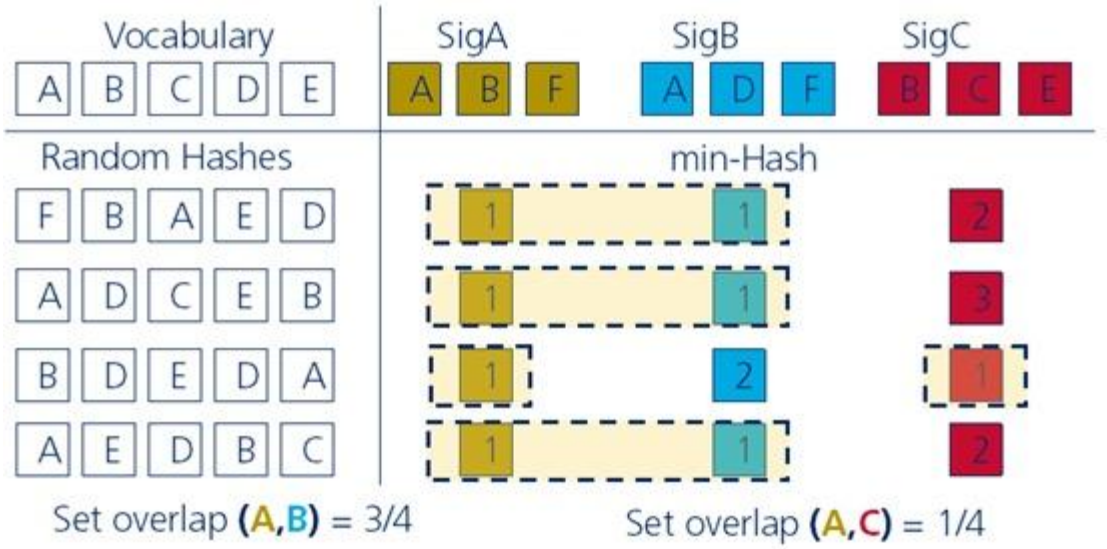
- An image signature is a frequency histogram of a set of discrete symbols similar in nature to the popular Bag of Words (BoW) model
- Two key differences from BoW:
  - The signature is increased in size as a result of mining rule
  - The signature can be based on any feature
    - BoW representations
    - other classifier responses

Similarity of Signatures

- Min-Hash is used as similarity measure
- The computation is proportional only to the number of input samples rather than the size of the vocabulary
  - It is ideally suited for image signatures which can be of high and increasing dimensionality
- $sim(S_1, S_2) = \left|\frac{S_1 \cap S_2}{S_1 \cup S_2}\right|$
  - exhaustive naive element by element comparision
- $\sum(min\pi(S_1) = min\pi(S_2))/N$, min-Hash similarity
  - $min\pi(S_1)$, min-Hash; π, permutation; S_1, S_2 sets

Similarity of Signatures

- Min-Hash Algorithm
  - The minimum element position of the permutation that is present in the set S.
- Table 1. Example of the 4 random permutations and resultant min-Hashes



Similarity of Signatures

- The false positive rate can be further reduced by grouping the min-Hash results into "sketches"
  - Consisting n hashes
- The grouping process will only work for input sample pairs that have at least m (set as 1) identical sketches
- Sketches for Table 1 with sketch size n=2
  - ((1,1),(1,1)); ((1,1),(2,1)); ((2,3),(1,2))

Similarity of Signatures

- Histogram Weighting Approximation
  - The original min-Hash algorithm [6] is designed for a set of uniformly weighted symbols
  - In order to convert the frequency based image signature into a min-Hash set of uniform symbols, the symbols are duplicated based upon their frequency
  - e.g
    - X={A,B,C} vocabulary containing visual words
    - t_i frequency response of the feature
      - t1 = {3,0,2} and t2 = {2,1,0}
    - t1={A1,A2,A3,C1,C2} and t2={A1,A2,B1}

Expanding signatures through co-occurring discriminatory features

- It is likely that min-Hash will produce falsely matching signatures because of noise
- A novel approach to “pull” positive sets together is proposed
- Association Rule data mining APriori[1] will be used to identify the compound visual words
- It searches databases and identify the set elements that co-occur most frequently within the positive sets with respect to the negative sets

Expanding signatures through co-occurring discriminatory features

- The frequency of a set element is related to the support and confidence of an association rule $A \Rightarrow B$
- $\sup(A \Rightarrow B) = \frac{|\{T|T \epsilon D, (A \cup B) \subseteq T|}{|D|}$
- $conf(A \Rightarrow B) = \frac{\sup(A \cup B)}{\sup(A)}$

$$= \frac{|\{T|T \epsilon D, (A \cup B) \subseteq T|}{|\{T|T \epsilon D, A \subseteq T|}$$

Expanding signatures through co-occurring discriminatory features

- In addition to the set elements being frequent, they must also be discriminative with respect to the negative set
- The transaction vectors of all examples are appended with a label, α, that identifies if the set is a positive or negative example
- The results of data mining include rules of the form $(A, B) \Rightarrow \alpha$

Expanding signatures through co-occurring discriminatory features

- The support threshold is the number of positive transactions over the total number of transactions
- The confidence threshold is 100, to ensure an association rule is only found if the elements are within all the positive sets and none of the negative sets.

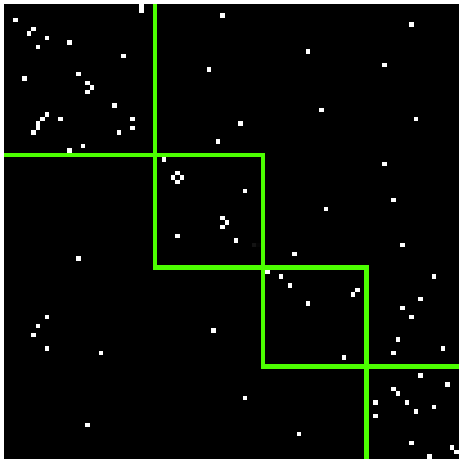
Iterative Signature Learning

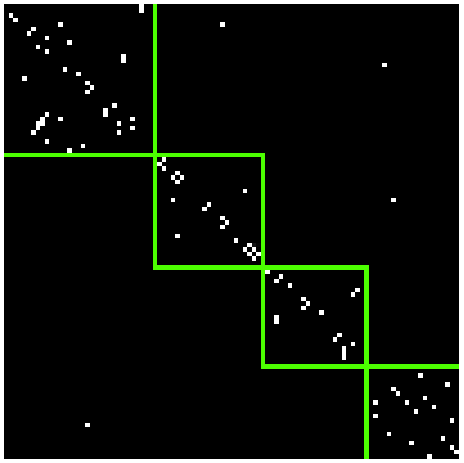
- This will accentuate common and distinctive elements of the positive rules from the mining and increase their overall similarity
  - “pull” the positive signatures together
- Ex
  - t1={A1,A2,A3,C1,C2}, t2={A1,A2,B1}
  - If mining rules return Ax; A4 will be added to t1 and A3 to t2
  - If mining rules return (A2,B1), no addition to t1 and t2 will be {A1,A2,B1, AB1}

Results

- Images
  - E. Ong and R. Bowden[21] dataset
    - 100 images 4 classes (city, jungle, mountain, winter)
    - 11 bin color and 42 bin edge histogram concatenated (all images used)
    - Image signature db has v=432, N=1500 min-hash permutation, sketch size n=3

Results





(a) Initial confusion between query image and dataset,
(b) Confusion after 7 iterations usingM = 1 (Green lines indicate class boundaries

Results

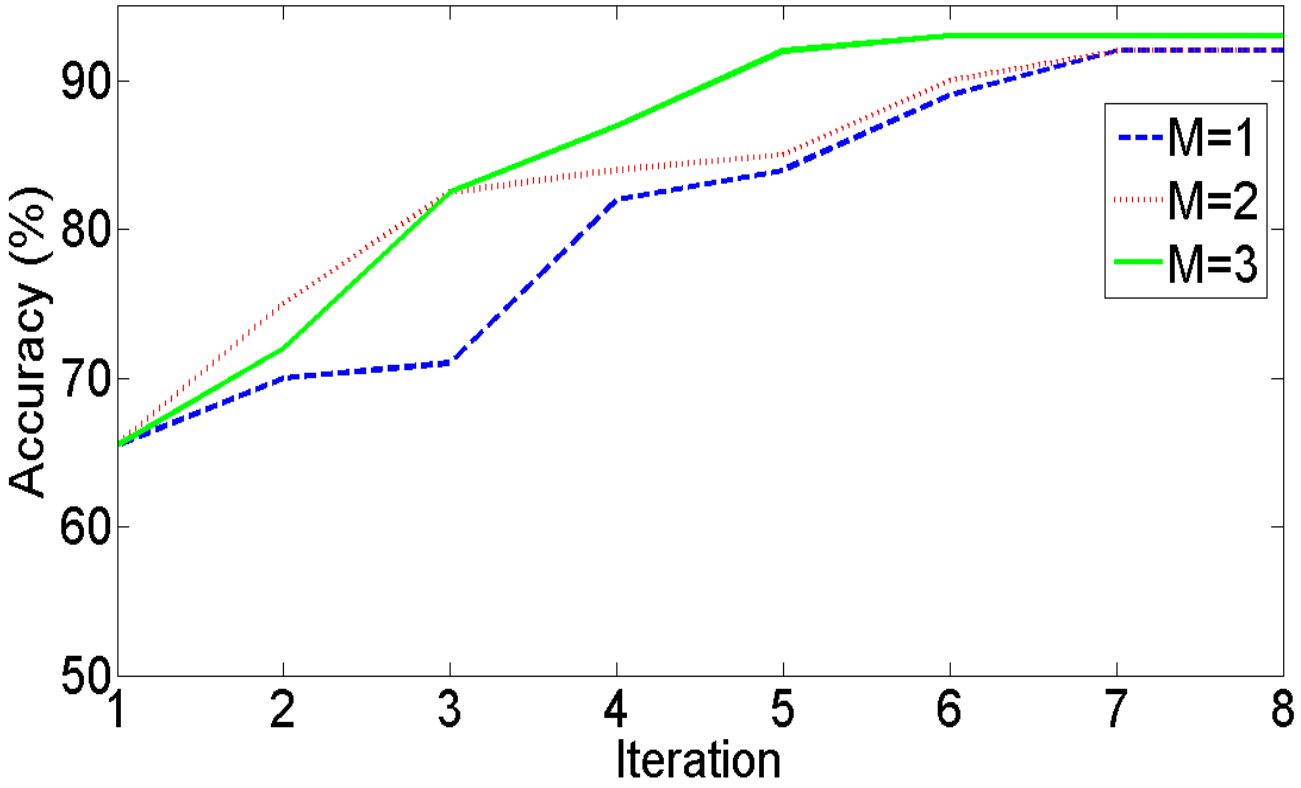


Image Dataset Accuracy with respect to iteration level, and varying M

Results

- Caltech101[9] dataset
  - 101 object categories 31 to 800 images per category
  - 15 training examples randomly selected from each class
  - BoW histograms of standard SIFT descriptors with the dimension reduced to 30 as employed in [4] used as initial signature
  - these image signatures were tested on another 50 unseen test images from each class
  - This process repeated 10 times

Results

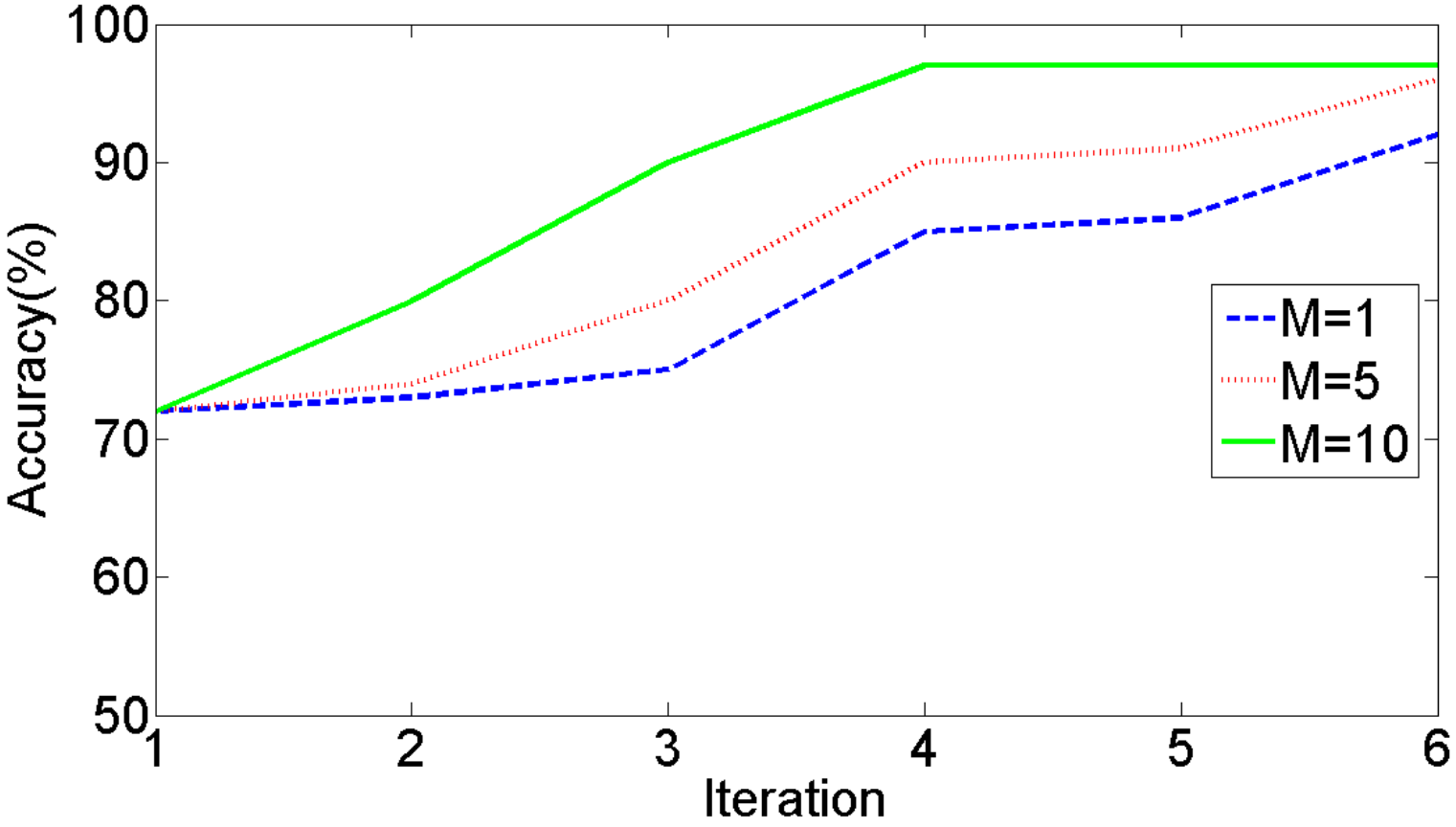
Table 2. Accuracy of *Caltech101* dataset

Approach	Accuracy
Cai [4]	64.9%
Baseline min-Hash	21.54%
Sig min-Hash	59.7%

Result

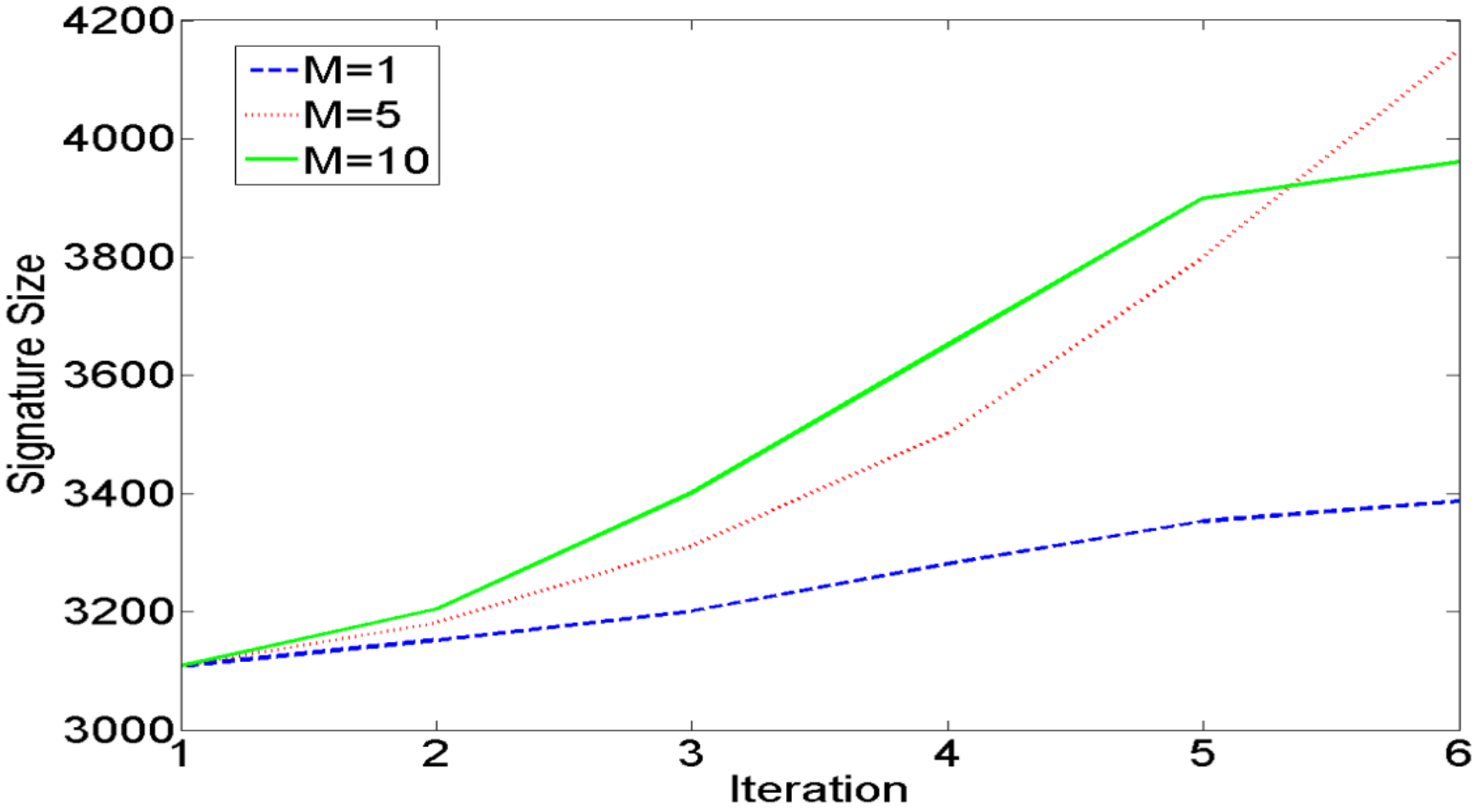
- YouTube Video Dataset
  - 1200 real life video of 11 different categories
  - To form the initial image signature, a histogram is computed using the features by Gilbert et al [10]
    - These consist of compound corner classifiers trained on the KTH dataset [23]
  - Initial signature size 3108
  - 36 groundtruth videos 6 iterations M=5; accuracy 97%
    - Other approaches use 1121 groundtruth video

Result



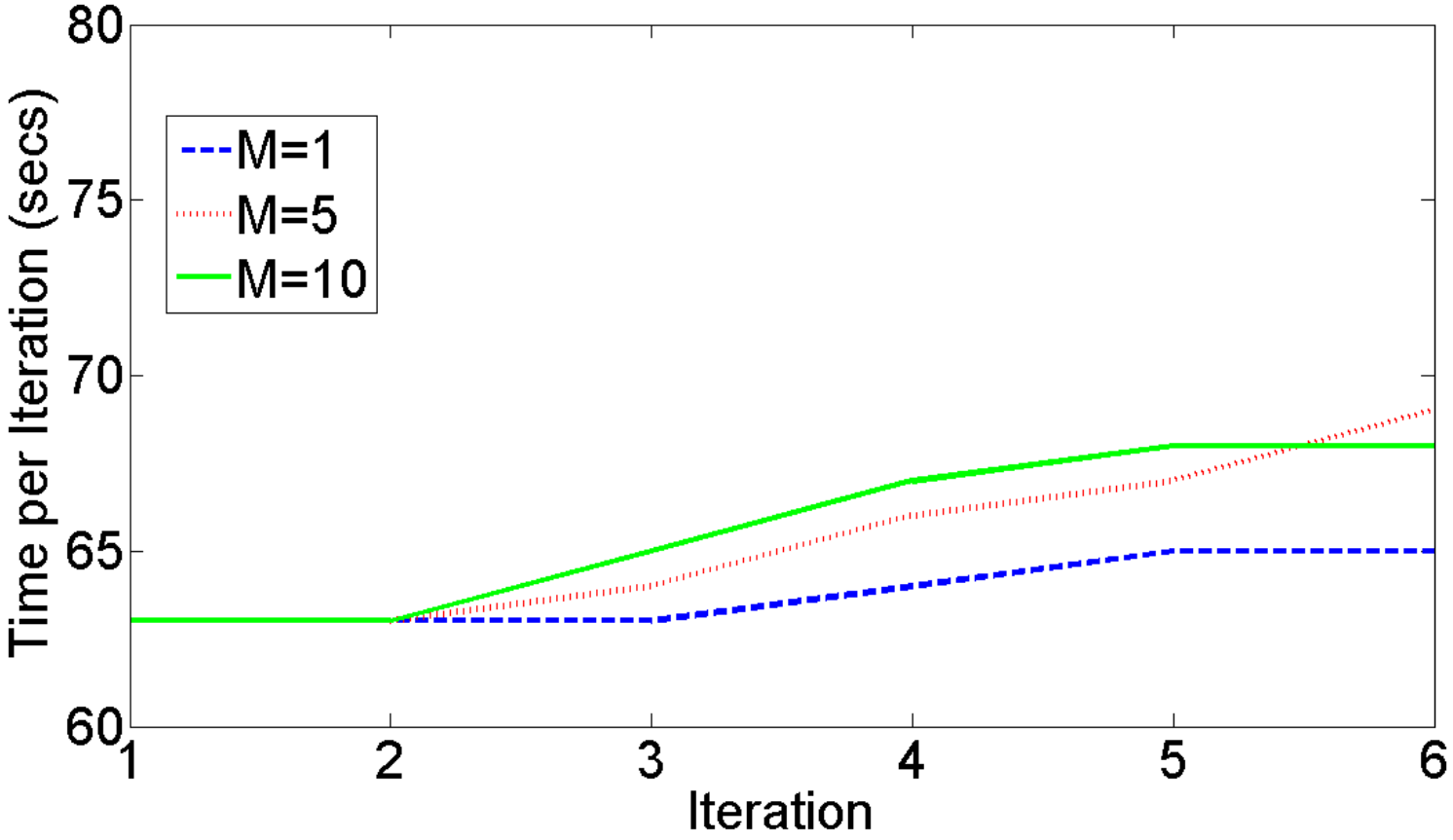
YouTube Dataset Accuracy with respect to iteration level, and varyingM

Result



Signature size for YouTube, with varyingM

Result



YouTube Time taken per iteration for varyingM

Result

Table 3. Accuracy of *YouTube* dataset

Approach	Accuracy
Cinbis [14]	75.2%
Liu [18]	71.2%
Bregonzio [2]	63.1%
Baseline min-Hash	56.4%
Sig min-Hash	79.7%

Result

- KTH Dataset
  - 25 people performing each of the 6 actions, 4 times; giving 599 video sequences
  - 8 people for training, and 8 people testing
  - For initial signature, same features as used for YouTube dataset are employed
  - 7 iteration, M=5 only 42 labeled videos instead of 192 videos used, 91.2% accuracy

Result

Table 4. Accuracy of *KTH* dataset

Approach	Accuracy
Schüldt [23]	71.71%
Laptev [16]	91.8%
Laptev [16]	91.8%
Wang [26]	92.1%
Gilbert [10]	95.7%
Baseline min-Hash	44.3%
Signature min-Hash APriori	91.2%

Result

- Hollywood2 Dataset
  - 12 action classes, 600000 frames 7 hours of video seq.
  - HoG/HoF descriptors using the interest point detection method of [15], and construct a visual dictionary using K-Means with K = 4000 visual words and train a SVM classifier. The classifier response is used as the input for the image signature

Result

Table 5. Accuracy of *Hollywood2* dataset

Approach	Accuracy
Marszalek [19]	35.5%
Han [12]	42.1%
Wang [26]	47.7%
Gilbert [10]	50.9%
Baseline min-Hash	26.9%
Signature min-Hash APriori	43.2%

Computational Cost

Table 6. Computational Time of datasets

Dataset	Dataset Size	Img Sig Size	Iter Time
Image Scene	100	53	1 sec
Caltech101	5050	2150	30 sec
YouTube	1200	3108	63 sec
KTH	768	1204	25 sec
Hollywood2	884	4503	45 sec

Conclusion

- Image signature is used for clustering similar videos and images
- A generic, efficient and iterative algorithm for interactively clustering is presented
- Min-Hash and APriori data mining approaches are employed for video and images

Thanks...