Semi-Supervised Clustering by Selecting Informative Constraints

P Vidyadhar Rao, C.V Jawahar

Center for Visual Information Technology

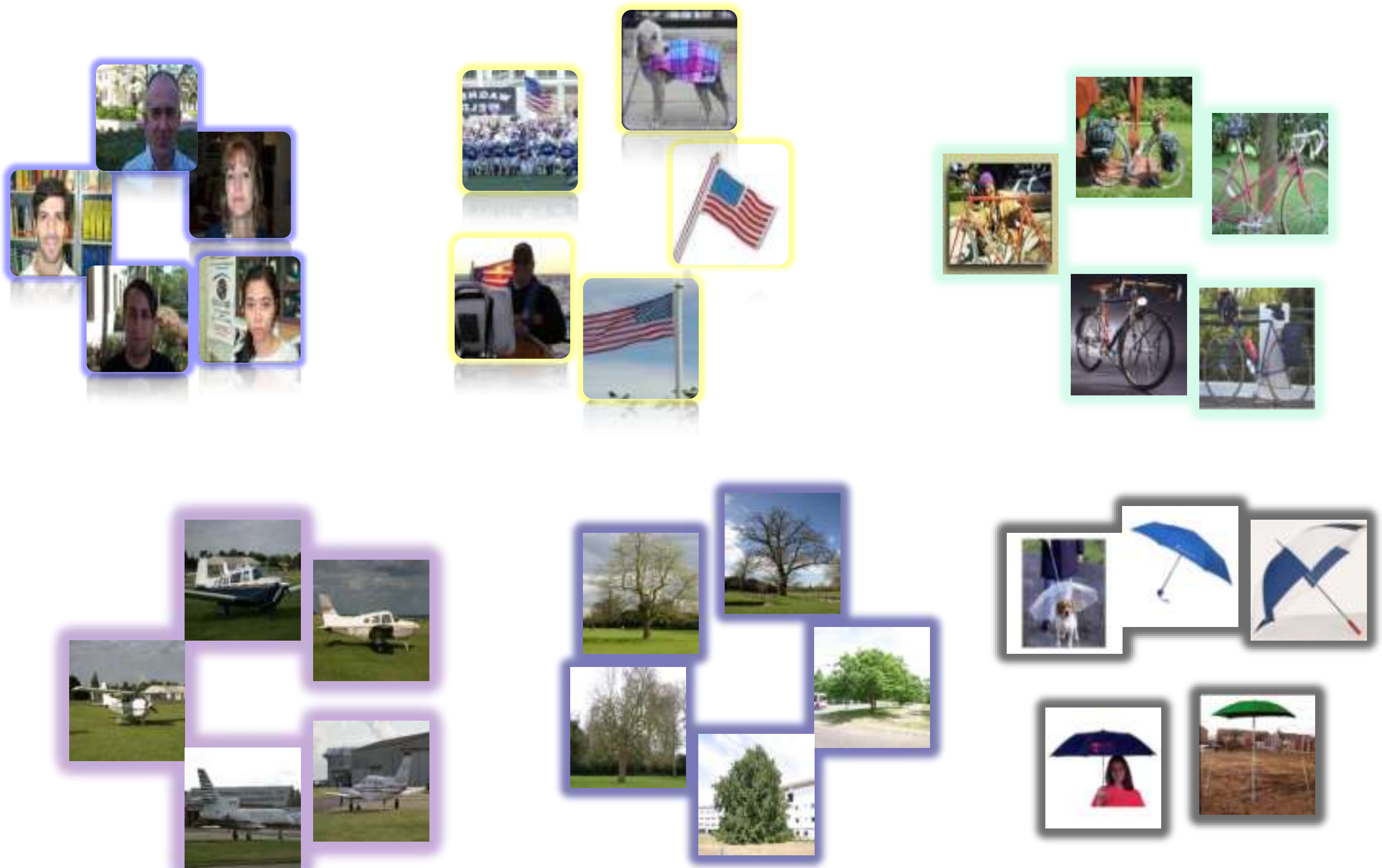
IIIT Hyderabad

Clustering Problem

- Input: A set of objects X

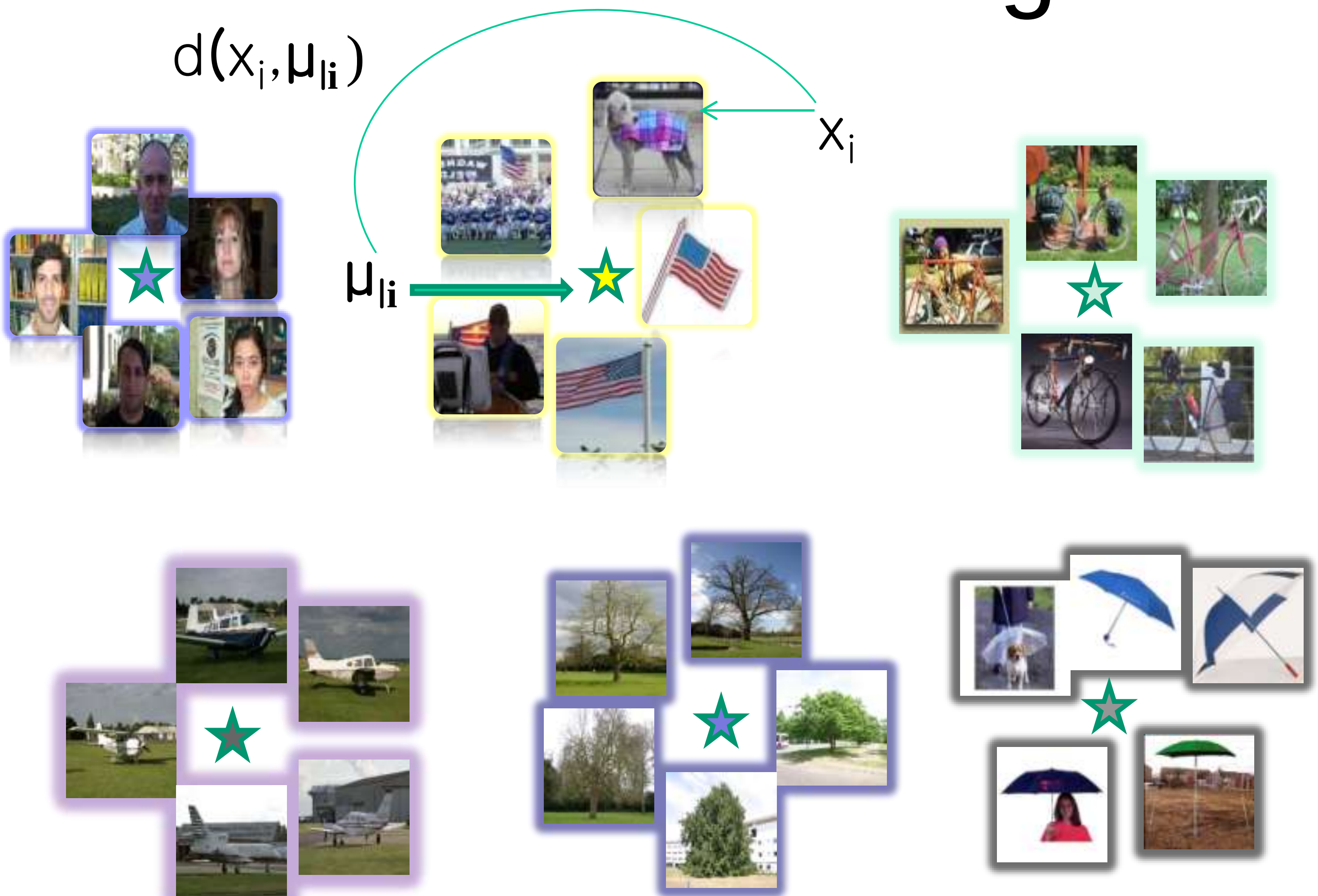
IIIT Hyderabad

Clustering Problem



- Output: A set of clusters(subsets)

Kmeans Clustering



- Objective: min $\Sigma ||x_i - \mu_{li}||_2$

Clustering in multi-view Environment

Clustering in multi-view Environment

Person

Umbrella

Clustering in multi-view Environment

Dog **Women**

Clustering in multi-view Environment

Person

Umbrella

Dog **Women**

- Distance function is not known in unsupervised setting.

Clustering in multi-view Environment

Objective

$$\min \phi_A(X) = \sum_{x_i \in X} d_A(x_i, \mu_{l_i}) \tag{1}$$

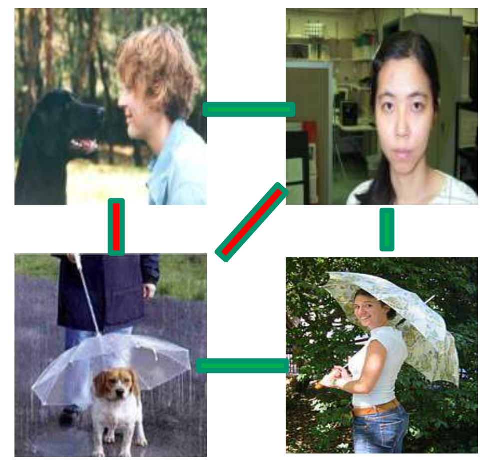
where $d_A(x, y) = (x - y)^T A(x - y)$ and A is a $d \times d$ PSD matrix.

- Distance metric, **A**, reflects relationships b/w objects.
- Kmeans with Euclidean distance metric: **A = I**;
- Choice of distance function decides the quality of clustering.
- **Problem:** Learn distance function and partitions

Semi-Supervised Clustering

- Supervision offers instance level constraints like must-link and cannot-link constraints.

Partially labeled Image Databases



A case of Conflicting Constraints

- However, it is not a good idea to derive partitions strictly satisfying every constraint!

Selecting Informative Constraints

- Need to exploit partially labeled data and/or (dis)similarity constraints to construct more useful distance function.

Informativeness

Amount of information in the constraint set that the algorithm cannot determine on its own.

Coherence

Amount of agreement with in the constraints themselves, with respect to a given distance metric

- *When more informative constraints are under the learned metric, the more likely they are to improve clustering.*

Metric Learning

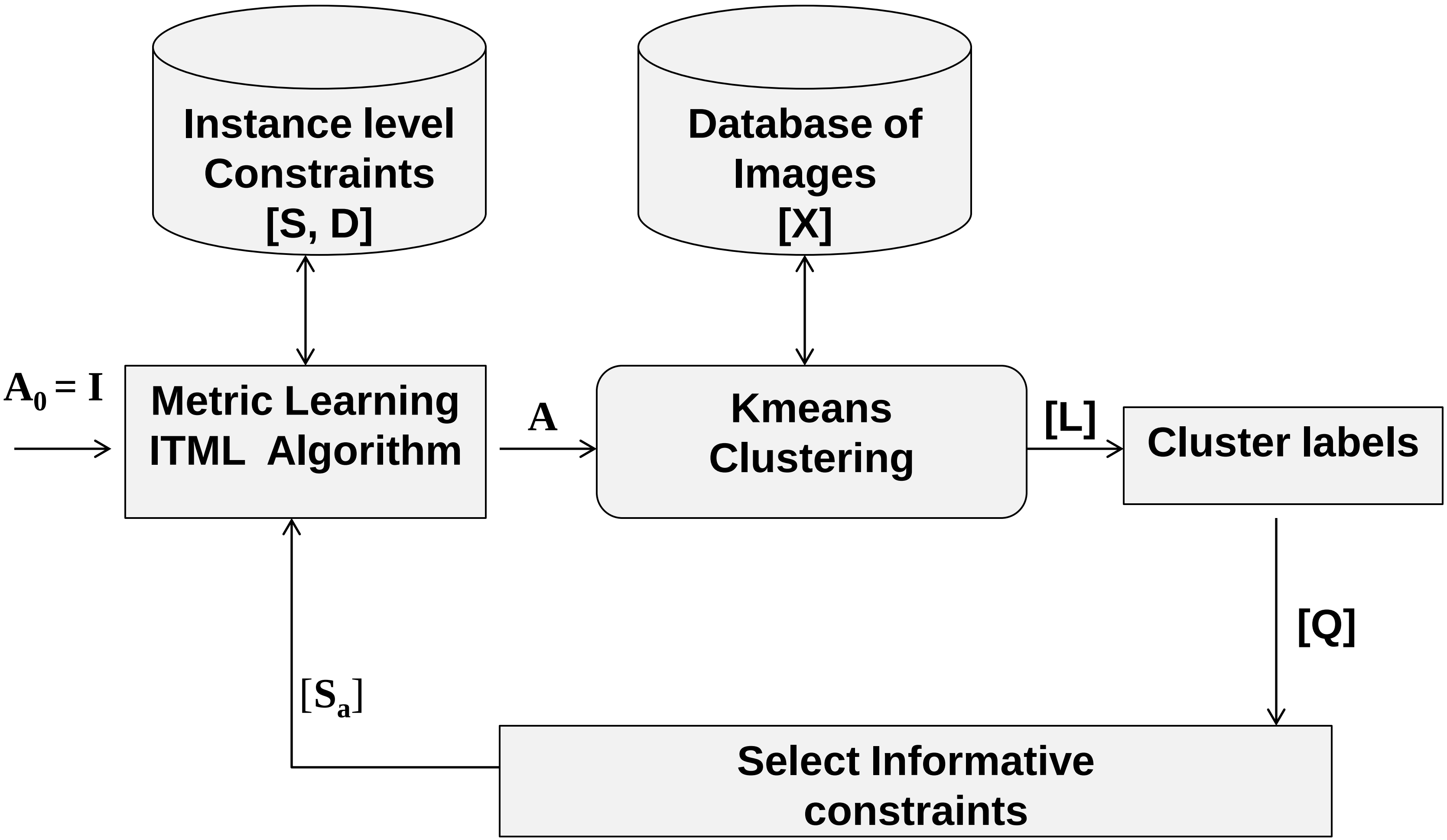
Objective

$$
\begin{aligned}
\min \quad & D_{ld}(A, A_0) \\
\text{s.t.} \quad & A \succeq 0 \\
& d_A(x_i, x_j) \leq u \qquad (i,j) \in S \\
& d_A(x_i, x_j) \geq v \qquad (i,j) \in D
\end{aligned} \tag{2}
$$

where $D_{ld}(A, A_0) = tr(AA_0^{-1}) - \log det(AA_0^{-1}) - d$; v and u are large and small values, respectively.

- Enforce simple **distance constraints** on instance level constraints.
- ITML algorithm solves Eq.(2) to learn metric under these constraints.

Our Approach



Our Approach

Semi-Supervised Clustering by Selecting Informative Constraints

Initialization: $A_0 = I$; $S_u = S \cup D$; $S_a = \{\}$; $k = 0$;
repeat
    for $(i, j) \in S_u$ **do**
        $A^{ij} \leftarrow \mathrm{ITML}(X,\ A_0,\ S_a \cup (i,j),\ u, v)$;
        $(Q^{ij}, I^{ij}) \leftarrow \mathrm{Kmeans}(X,\ A^{ij},\ K)$;
    end for
    $(i^*, j^*) \leftarrow \arg\max_{ij}(Q^{ij})$;
    $A_0 \leftarrow A^{i^* j^*}$;
    $S_a \leftarrow S_a \cup (i^*, j^*)$;
    $S_u \leftarrow S_u \setminus (i^*, j^*)$;
    k_{++};
until $(k \leq t)$

- Complexity: $O(t.|S_u|)$ metric learning and clustering operations.

Implementation Methods

- Variations to our semi-supervised clustering(SSC) algorithm
  - *SSC-rand:* Metric learned from randomly selected constraints
  - *SSC-OLDML:* Metric learned with most recently obtained metric as prior for ITML
  - *SSC-active:* Metric learned from active constraint set.

Experiments

- Image Datasets
  - 10 MNIST handwritten digits: 1000 images
  - 11 objects from Caltech-256: 550 images
  - 20 objects from MSRC2: 600 images

Experiments

- Image Representation
  - Digit Images: Normalized to a 20x20 pixelbox and 400 pixel values are used.
  - Object Images: 800 visual words are extracted using SIFT descriptors.
- Performance Evaluation

$$\phi_A(.) \qquad \begin{aligned} \text{Precision} &= \frac{\#PairsCorrectlyPredictedInSameCluster}{\#TotalPairsPredictedInSameCluster} \\ \text{Recall} &= \frac{\#PairsCorrectlyPredictedInSameCluster}{\#TotalPairsInSameCluster} \\ F_1\text{-Score} &= \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision} + \text{Recall}} \\ \text{Rand Index} &= \frac{\#PairsCorrectlyPredicted}{\#TotalPairs} \end{aligned}$$

Results

- Performance measured in terms of $\phi_A(.)$

Dataset	Algorithm			
	K-means	SSC-rand	SSC-OLDML	SSC-active
MNIST	37380	36562	61474	**34726**
Caltech-256	2.665	2.565	2.618	**2.020**
MSRC2	2.059	2.275	3.344	**1.991**

- SSC methods show improvement over Kmeans.
- SSC-active performs better than K-Means, SSC-rand and SSC-OLDML.

Results

- Performance measured in terms of RandIndex

Dataset	Algorithm				
	K-means	SSC-rand	SSC-OLDML	MPCK-means	SSC-active
MNIST	0.875	0.881	0.861	0.862	**0.921**
Caltech-256	0.769	0.758	0.827	**0.841**	0.807
MSRC2	0.892	0.895	0.881	0.859	**0.904**

- SSC-active performs better than K-Means, SSC-rand and SSC-OLDML
- SSC-active performs better than MPCK-Means on two datasets.

Results

- Performance measured in terms of F_1-score

Dataset	Algorithm				
	K-means	SSC-rand	SSC-OLDML	MPCK-means	SSC-active
MNIST	0.410	0.434	0.334	0.377	**0.621**
Caltech-256	0.150	0.156	0.195	**0.249**	0.215
MSRC2	0.155	0.162	0.128	**0.226**	0.203

- SSC-active performs better than K-Means, SSC-rand and SSC-OLDML
- SSC-active performs close to MPCK-Means on object datasets and outperforms on digit dataset.

Qualitative Results

- Cars from Caltech-256

Qualitative Results

- Cycles from Caltech-256

Qualitative Results

- Books from MSRC2

Qualitative Results

- Water from MSRC2

Qualitative Results

- Flowers from MSRC2

Conclusions

- SSC-active requires only few informative constraints
- *Informativeness* is ensured via pairwise distance constraints.
- *Coherence* is ensured by selecting constraints using SSC-active.
- SSC-active always performed better than unsupervised K-Means unlike MPCK-Means

Future work

- Efficiently select informative constraints using active learning strategies.
- Scalability of our algorithm for large dimension image representations.

Thank You!

Questions?

Backup slides from here

Qualitative Results

- Cars from MSRC2

Qualitative Results

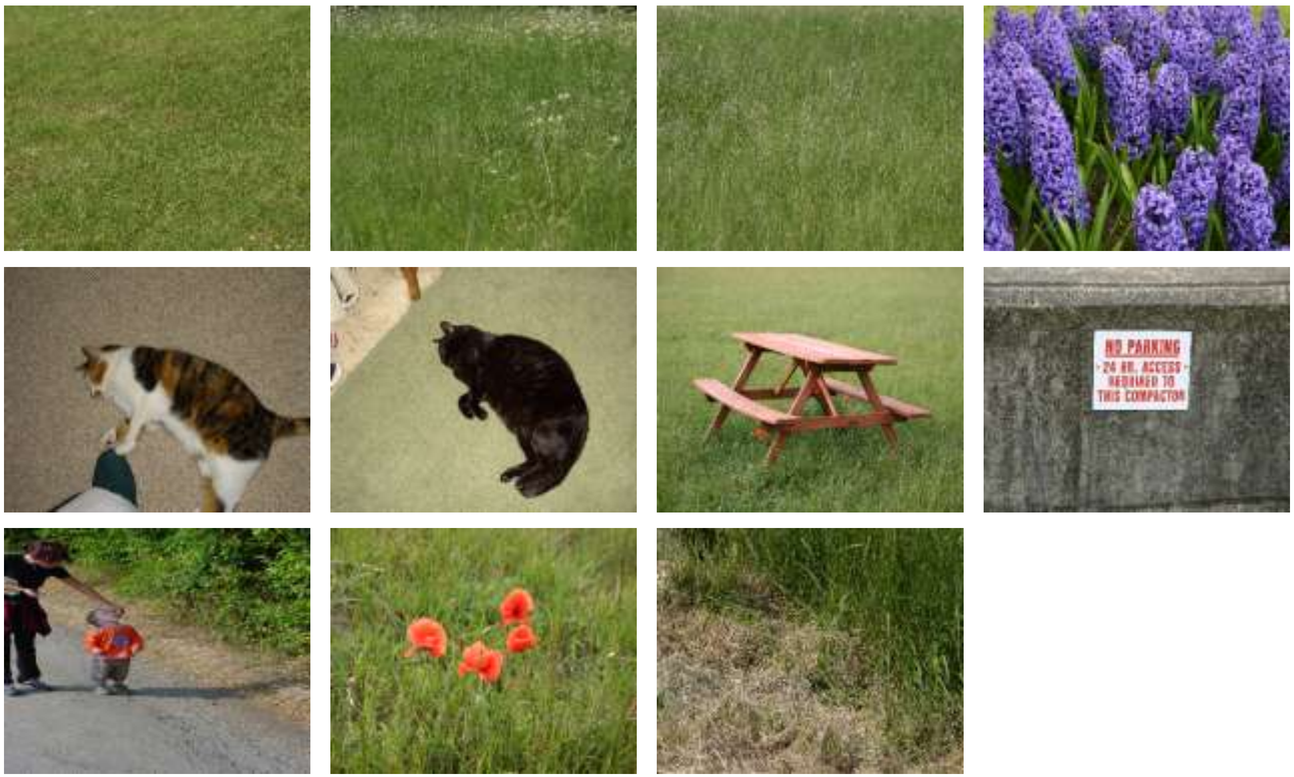
- void cluster from MSRC2

Qualitative Results

- void cluster from MSRC2

Qualitative Results

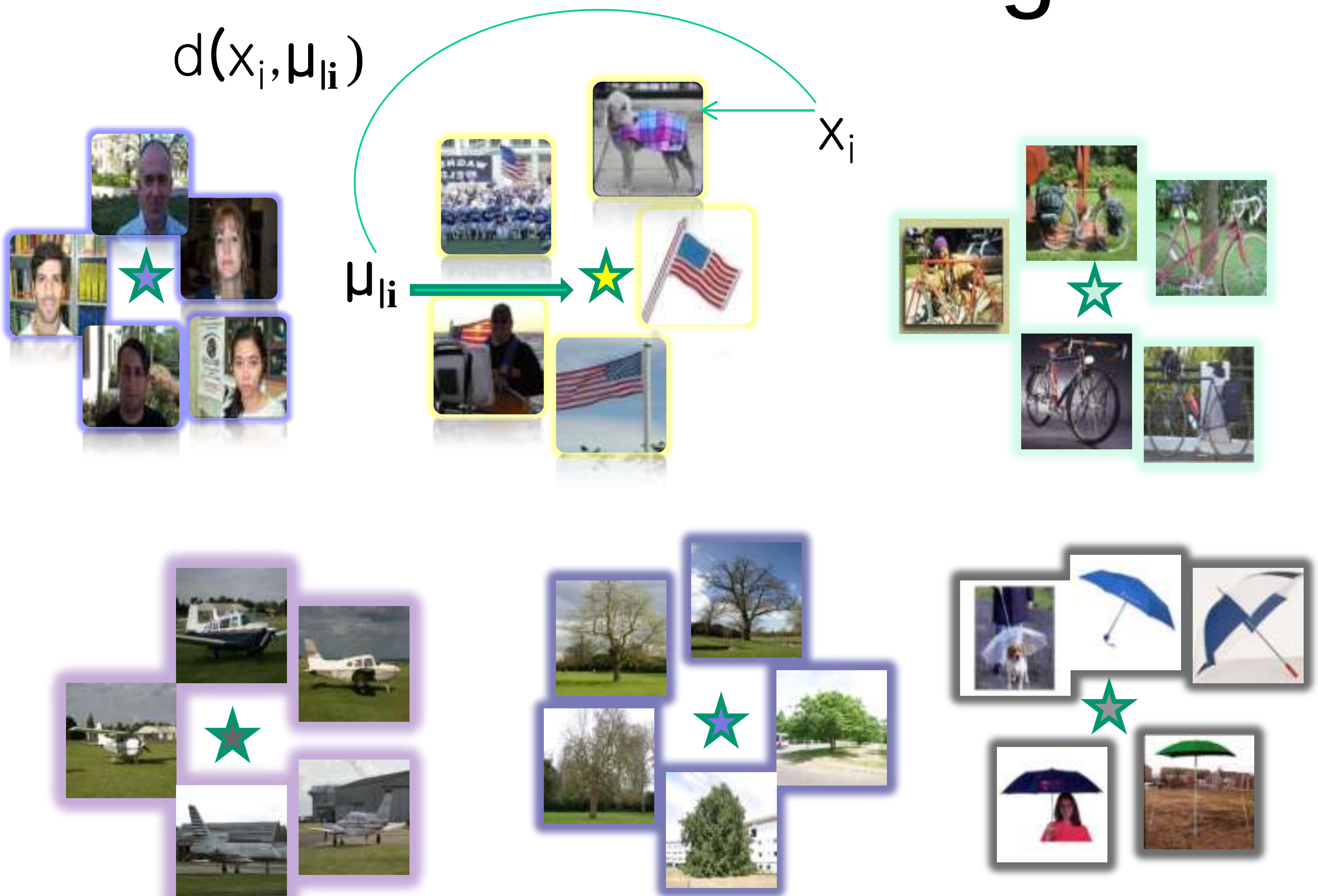
- void cluster from MSRC2



Qualitative Results

- Cycles from MSRC2

Kmeans Clustering



- Objective: min Σ $|| x_i - \mu_{li} ||_2$

Selecting Informative Constraints

- *When more informative constraints are under the learned metric, the more likely they are to improve clustering.*

Objective

Learn A w.r.t the given constraints such that $\phi_A(X) < \phi_I(X)$.

- Our Approach
  - Iteratively solve for Eq.(1) and Eq.(2)
  - Incrementally select informative constraints
  - Learn metric using informative constraints
  - Use learned metric for Kmeans clustering.

Semi-Supervised Clustering

- Supervision offers instance level constraints like must-link and cannot-link constraints.

Partially labeled Image Databases

Fully labeled Image Databases

- However, it is not a good idea to derive partitions strictly satisfying every constraint!