Regularised Cross-Modal Hashing

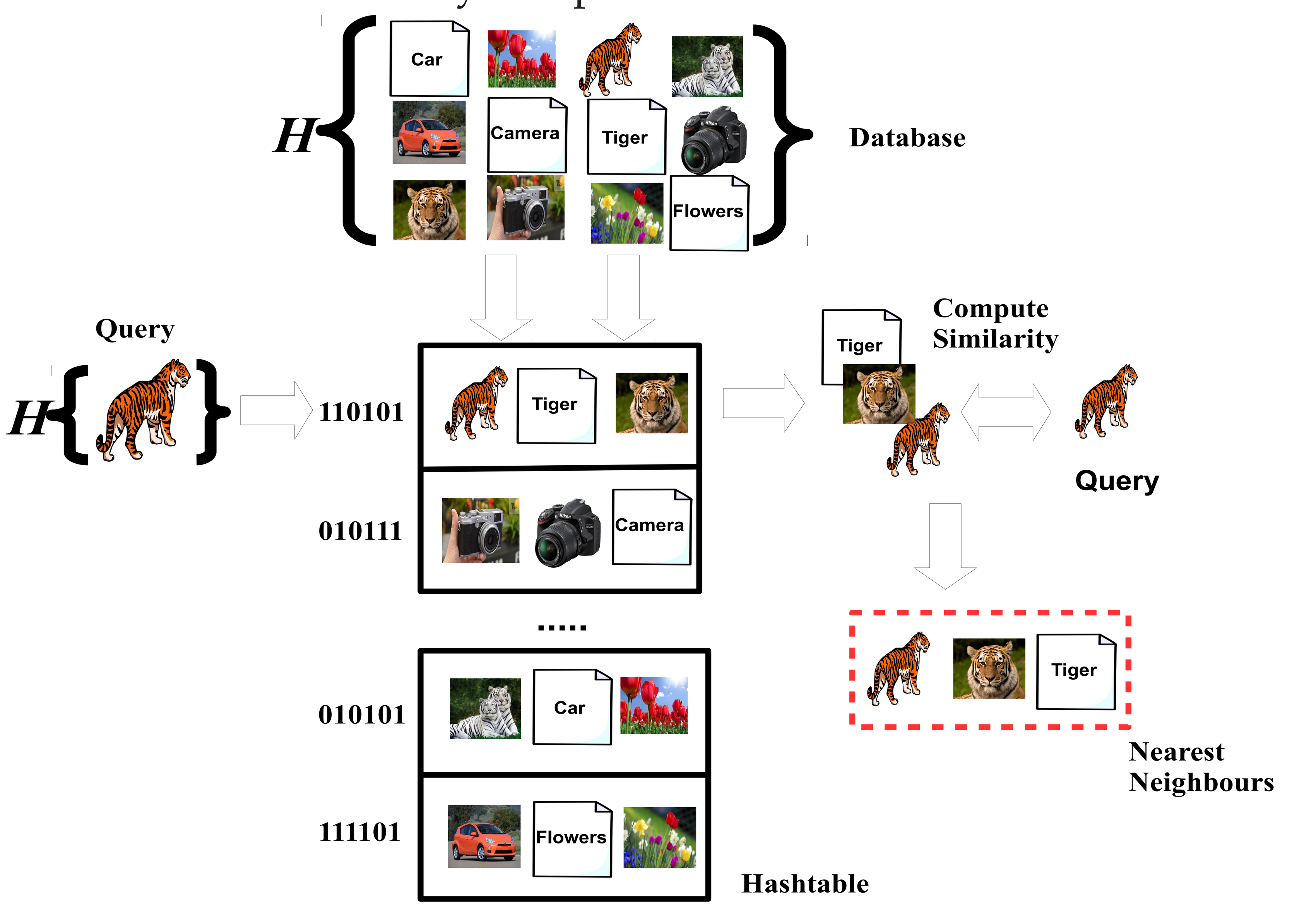
Sean Moran†, Victor Lavrenko
† sean.moran@ed.ac.uk

Multi-modal retrieval using binary hashcodes

- **Research Question:** Can learning binary hashcodes for cross-modal data-points enable efficient and effective multi-modal retrieval?
- **Approach:** we learn **hyperplanes** that split both feature spaces (e.g. text, image descriptors) into buckets so that similar cross-modal data-points fall into the buckets labelled with the same hashcodes.

Hashing-based approximate nearest neighbour search

- **Problem:** Nearest Neighbour (NN) search in multi-modal datasets.
- **Hashing-based approach:**
  - Generate a similarity preserving binary hashcode for query.
  - Use the fingerprint as index into the buckets of a hash table.
  - If collision occurs only compare to items in the same bucket.



- *Hashtable buckets are the polytopes formed by intersecting hyperplanes in the word and image descriptor feature spaces.*
- **This work:** learn hyperplanes to encourage collisions between similar multi-modal data-points.

Regularised Cross-Modal Hashing (RCMH)

- **Step A:** Use LSH [1] to initialise **word** bits $\mathbf{B} \in \{-1,1\}^{N \times K}$
  N: # data-points, K: # bits
- **Repeat for M iterations:**
  - **Step B:** *Graph regularisation*, update the bits of each data-point to be the average of its nearest neighbours

$$\mathbf{B} \leftarrow sgn\left(\alpha\, \mathbf{S}\mathbf{D}^{-1}\mathbf{B} + (1-\alpha)\,\mathbf{B}\right)$$

    * $\mathbf{S} \in \{0,1\}^{N \times N}$: adjacency matrix, $\mathbf{D} \in \mathbb{Z}_{+}^{N \times N}$ diagonal degree matrix, $\mathbf{B} \in \{-1,1\}^{N \times K}$: bits, $\alpha \in [0,1]$: interpolation parameter, sgn: sign function

  - **Step C:** *Word data-space partitioning*, learn hyperplanes that predict the K word bits with maximum margin

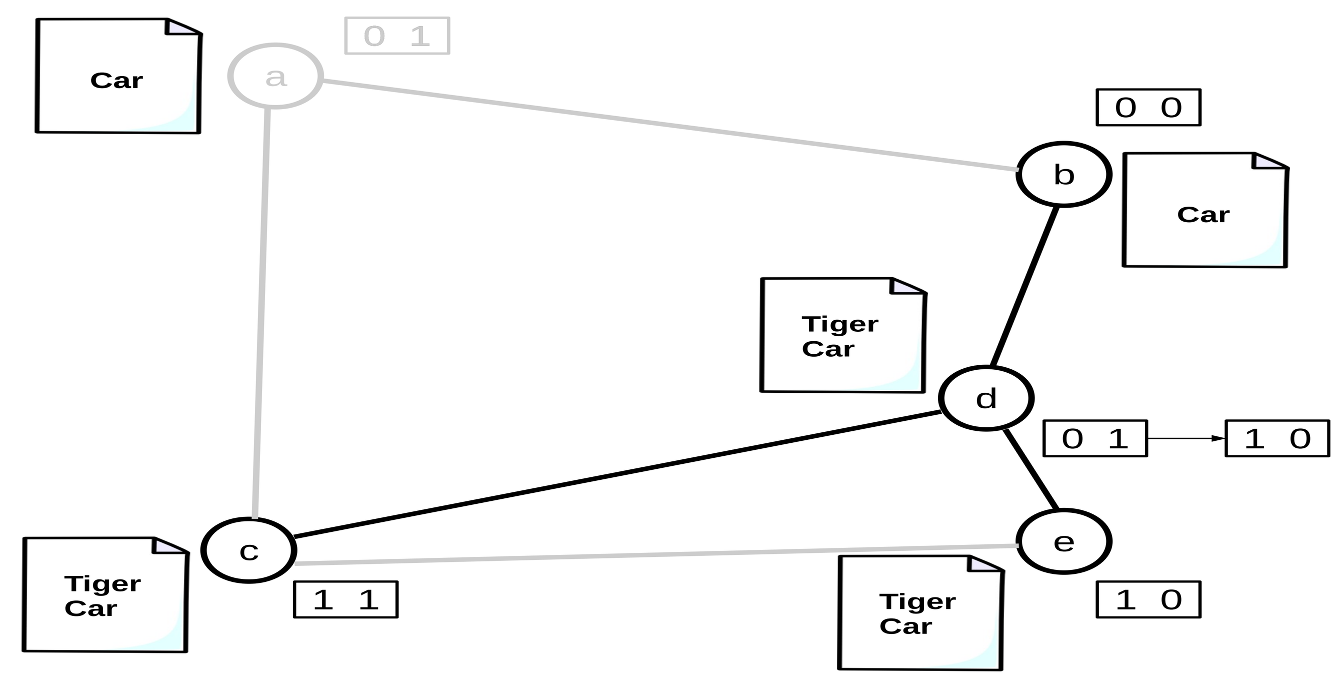
$$\text{for } k = 1 \ldots K: \quad \min \; ||\mathbf{f}_k||^2 + C \sum_{i=1}^{N} \xi_{ik}$$
$$\text{s.t.} \quad B_{ik}(\mathbf{f}_k^{\intercal}\mathbf{a}_i + b_k) \geq 1 - \xi_{ik} \quad \text{for } i = 1 \ldots N$$

    * $\mathbf{f}_k \in \Re^D$: **word** hyperplane, $b_k \in \Re$: bias, $\mathbf{a}_i \in \Re^D$: **word** descriptor, B_{ik}: bit k for **word** data-point $\mathbf{a}_i$, ξ_{ik}: slack variable

  - **Step C:** *Visual data-space partitioning*, learn visual hyperplanes $\mathbf{g}_k$ that predict the K *word* bits B_{ik} with maximum margin
  - **Step D:** Update $\mathbf{B}$: $b_{ik} = sgn(\mathbf{f}_k^{\intercal}\mathbf{a}_i + b_k)$
- Use the learnt hyperplanes $\{\mathbf{f}_k, \mathbf{g}_k\}_{k=1}^{K}$ to generate hashcodes

Step B: Graph Regularisation (bit smoothing)

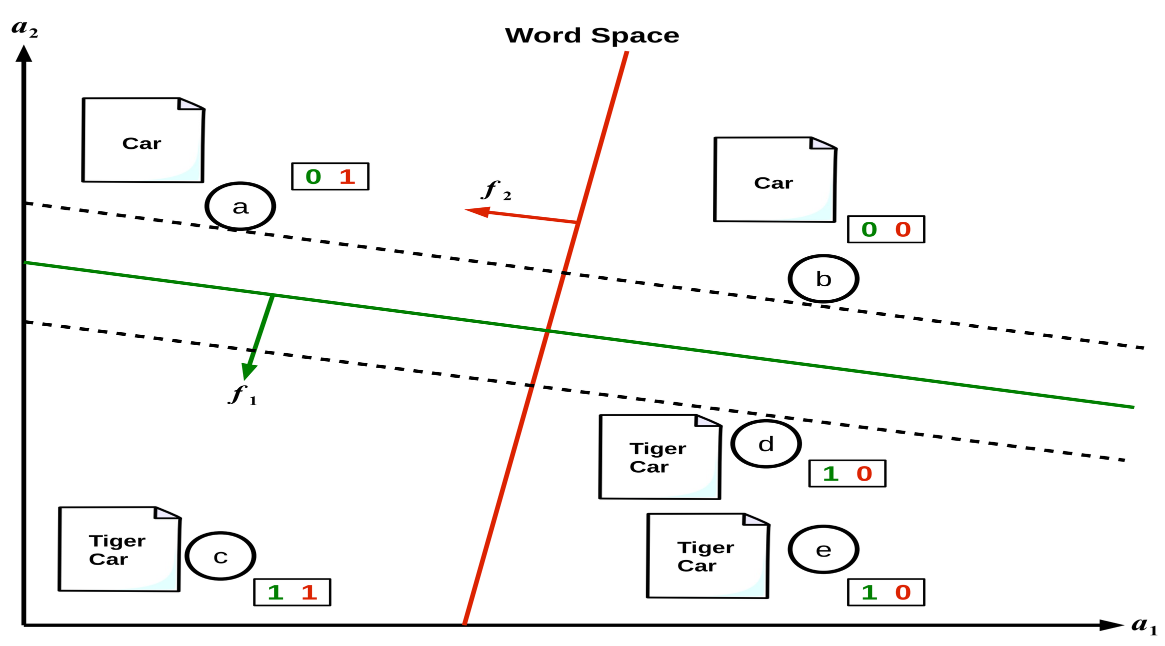
- Set word bits of each data-point to be the average of its neighbours:



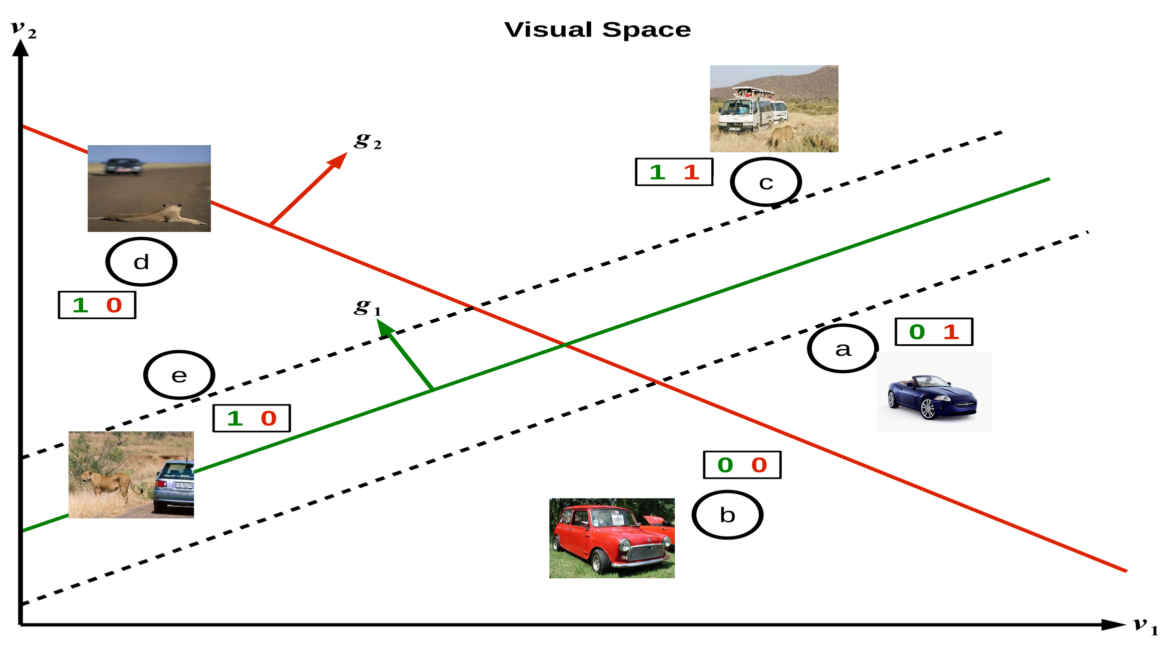
- Data-points are circles, arcs are neighbourhood relationship. Node d's hashcode (01 → 10) is updated to be consistent with b,c,e.

Step C: Learning the hashtable buckets (hyperplanes)

- Word hyperplanes f_1,f_2 learnt using bit 1 (green), bit 2 (red) as labels:

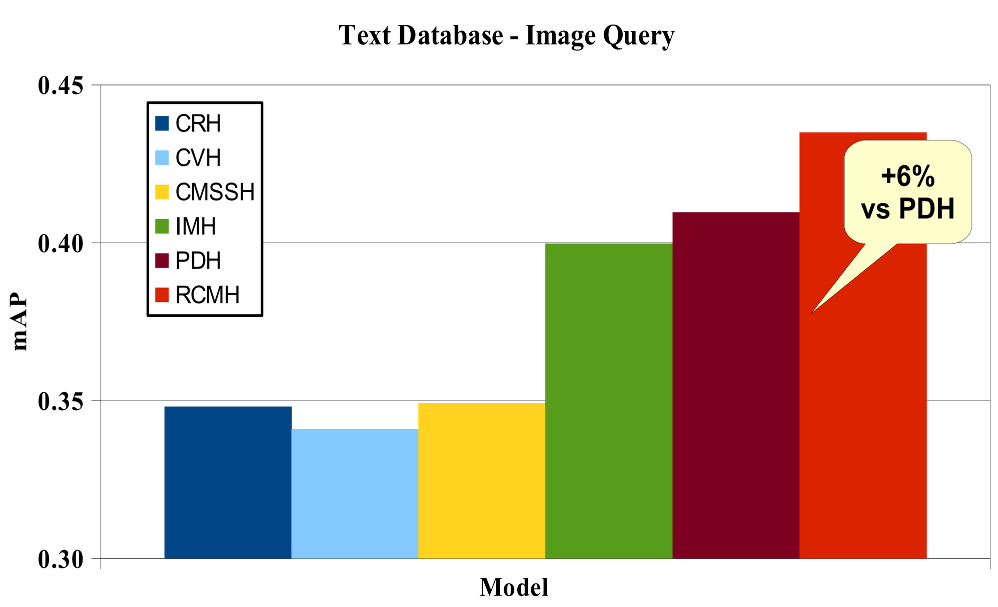


- Visual hyperplanes $\mathbf{g}_1$,$\mathbf{g}_2$ learnt using same regularised bits:



Results: Retrieval effectiveness and efficiency

- Retrieval evaluation on two standard multi-modal (text,image) datasets. Image query used to retrieve documents, and vice-versa.
- Retrieval on NUS-WIDE (left). Timing on Wiki dataset (right).



	Training (s)	Test (s)
RCMH	0.336	0.003
PDH [2]	3.715	0.003
CMSSH [3]	5.172	0.003

- Our model more effective *and* efficient than competitive baselines.

Conclusions and References

- New effective and efficient iterative model for cross-modal hashing
- Hashcode bits smoothed over using adjacency graph are used to learn hashtable buckets (hyperplanes) in word and image space.
- Extend to high volume data stream and cross-lingual retrieval.

References:
[1] Indyk, P., Motwani, R.: Approximate nearest neighbors: Towards removing the curse of dimensionality. In: STOC (1998).
[2] M. Rastegari et al. Predictable dual-view hashing. In ICML'13.
[3] M. M. Bronstein et al. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In CVPR'10.