The 12th Asian Conference on Computer Vision, Singapore, Nov. 1-5, 2014

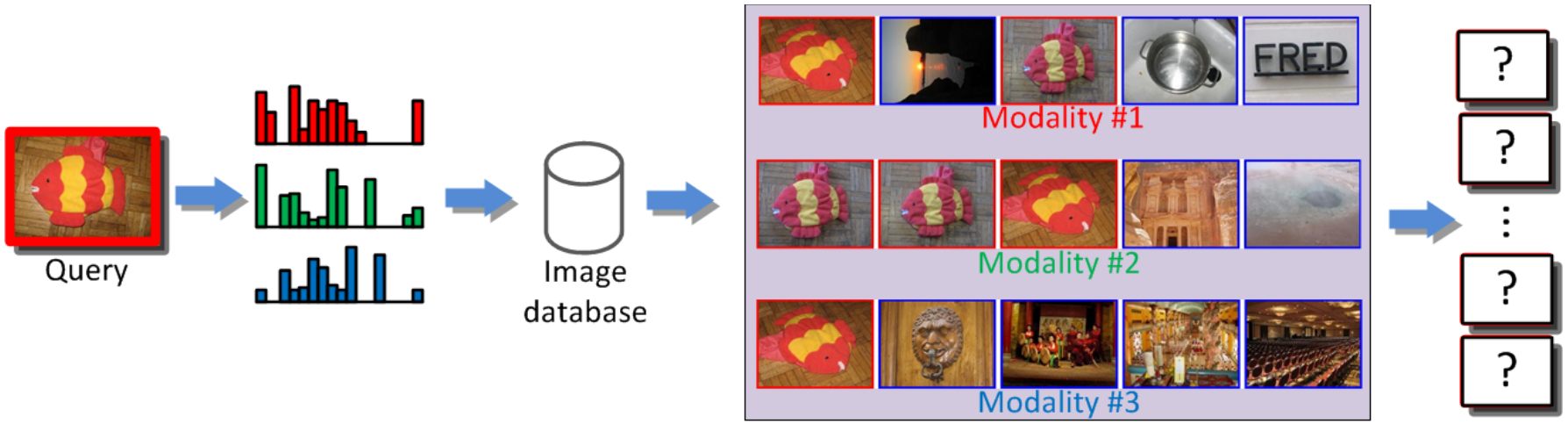
Submodular Reranking with Multiple Feature Modalities for Image Retrieval

Presenter: Zhuolin Jiang

Fan Yang[1], Zhuolin Jiang[2] and Larry S. Davis[1]
[1]UMIACS, University of Maryland, College Park, MD
[2]Noah's Ark Lab, Huawei Technologies

Motivation

- Given a query image represented by multiple feature modalities, how to improve retrieval quality by using these modalities?



- Concatenating multimodal features into a long feature vector is infeasible, when dimension of feature vectors is very high
- Instead, we fuse the image retrieval results from multiple modalities based on *submodularity*

Our Approach

- We define a *submodular objective function* for reranking images retrieved by multiple feature modalities, which consists of an information gain term and a relative ranking consistency term.
- It can be efficiently optimized by a simple greedy algorithm, which gives a near-optimal solution with a (1-1/e)-approximation bound[§].
- It can be easily extended to other generic information retrieval tasks with multiple independent ranked lists returned by heterogeneous and non-visual features.

[§]G. Nemhauser, , L. Wolsey and M. Fisher, An Analysis of Approximations for Maximization Submodular Set Functions - II, Mathematical Programming, 1978.

Preliminaries

- Set function

$$F : 2^E \rightarrow R$$

E: the ground set

F: set function

A_1

$F(A_1) = 4.5$

A_2

$F(A_2) = 6.4$

Preliminaries

- Submodular Set Function

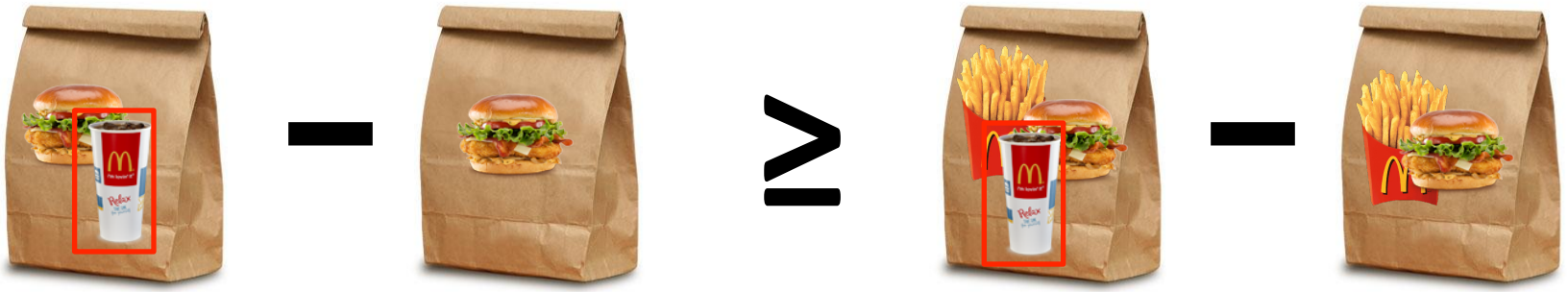
$$F : 2^E \rightarrow R$$

$$F(A_1 \cup \{a\}) - F(A_1) \geq F(A_2 \cup a) - F(A_2)$$

$$A_1 \subseteq A_2 \subseteq E \qquad a \in E \backslash A_2$$

diminishing return property

$F(A_1 \cup \{a\}) - F(A_1) \geq F(A_2 \cup \{a\}) - F(A_2)$



Submodular Reranking

- The submodular objective set function for reranking task is:

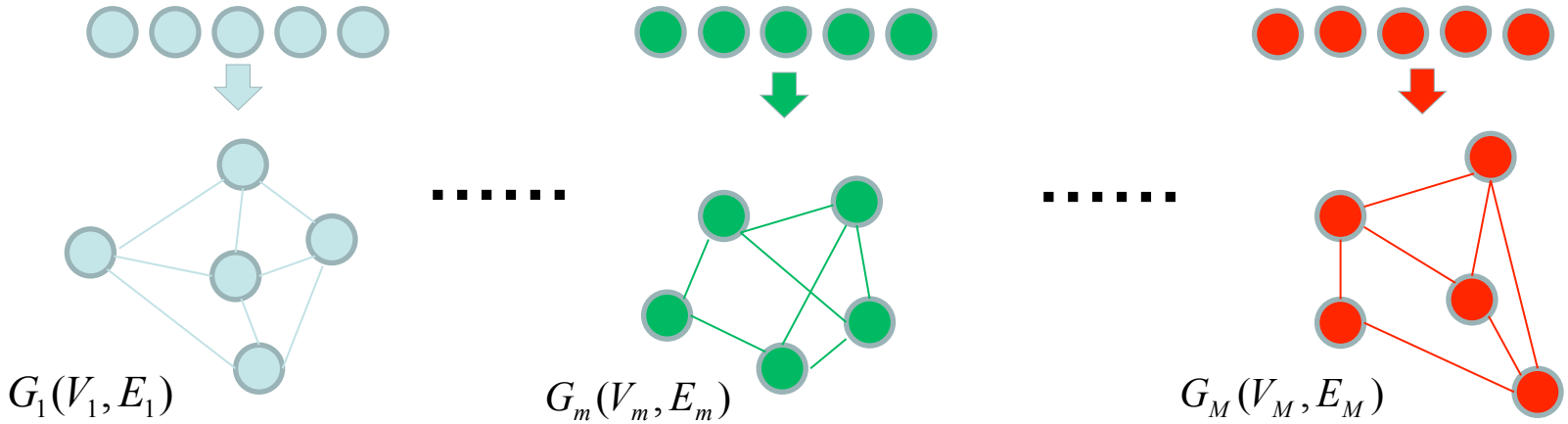
$$\max_{\mathcal{S}} \quad R(\mathcal{S}) + \lambda T(\mathcal{S})$$

$$s.t. \quad \mathcal{S} \subseteq \mathcal{V}, |\mathcal{S}| \leq K_s$$

  - S are selected images from *V*, $V = V_1 \cup V_2 \cup ... \cup V_M$ are the images retrieved by M feature modalities; K_s is the largest number of selected images;
  - R(S) is an information gain term, which selects a group of images that are similar to the query and closely related to each other;
  - T(S) is a relative ranking consistency term, which selects images that have consistent relative ranks across modalities and are similar to the query but only found by a few modalities.

Information Gain

- Given *M* initial ranked lists of retrieved images for a query image, each of which contains *K* images, we represent each initial ranked list as an undirected graph.



We denote $V = V_1 \cup V_2 \cup ... \cup V_M$ as the union of all nodes.

Our aim is choosing a subset of nodes *S* from *V* which are most similar to the query image.

Information Gain

Start from a single graph G_m , the information gain is defined as

$$F_m(\mathcal{S}) = H(\mathcal{V}_m \backslash \mathcal{S}) - H(\mathcal{V}_m \backslash \mathcal{S} | \mathcal{S})$$

where S is the selected nodes from *V* ; $V_m \backslash S$ is the set V_m with *S* removed. $H(V_m \backslash S)$ is the entropy of unselected nodes. $H(V_m \backslash S \mid S)$ is the conditional entropy of unselected nodes based on S.

By the random walk model, mathematically, we have

$$H(\mathcal{V}_m \backslash \mathcal{S}) = -\sum_{v \in \mathcal{V}_m \backslash \mathcal{S}} p_m(v) \log p_m(v)$$

marginal probability of v being similar to the query

and

$$H(\mathcal{V}_m \backslash \mathcal{S} | \mathcal{S}) = -\sum_{v \in \mathcal{V}_m \backslash \mathcal{S}, s \in \mathcal{S}} p_m(v, s) \log p_m(v|s)$$

$p_m(v,s) = p_m(v \mid s) p_m(s)$

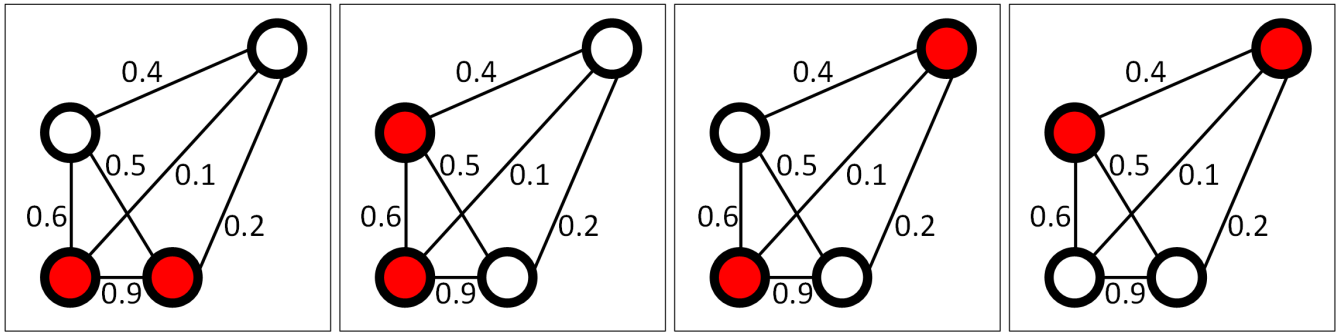
transition probability of walking from s to v

Information Gain

We simply summing up the information gains of the individual graphs to have the complete information gain term

$$R(\mathcal{S}) = -\sum_{m}\Big(\sum_{v\in\mathcal{V}\setminus\mathcal{S}} p_m(v)\log p_m(v) - \sum_{v\in\mathcal{V}\setminus\mathcal{S},s\in\mathcal{S}} p_m(v,s)\log p_m(v|s)\Big)$$

$R(S)$ **is submodular and non-decreasing.** Maximizing $R(S)$ tends to select a group of images that are similar to the query and closely related to each other.



$F_m(S)$=0.402 $F_m(S)$=0.389 $F_m(S)$=0.351 $F_m(S)$=0.351

Red dots are selected subset. The marginal probability of all nodes is set to ¼.

Relative Ranking Consistency

- **Relative ranking** between two images v_i and v_j is defined as

$$rr_m(v_i, v_j) = |r_{m,v_i} - r_{m,v_j}|, \quad v_i, v_j \in \mathcal{V}$$

r_{m,v_i} is the ranking (position) of image I_i in the *m*-th ranked list.

- **Relative ranking consistency measure** across all the ranked lists is

$$\mathcal{C}(v_i, v_j) = \frac{1}{Z} \sum_{m,m' \in M, m \neq m'} 1 - \frac{\min(rr_m, rr_{m'})}{K}$$

K is the number of retrieved images and Z=Mx(M-1)/2 is a normalization factor

– If two images vi and vj are ranked similarly across multiple modalities, they have high RRC scores and similar ranks in the reranked list.

– If an image is highly similar to the query but only highly ranked by a small number of modalities, it still has relatively high RRC score.

Relative Ranking Consistency

- The complete relative ranking consistency term T(S) is defined as:

$$T(\mathcal{S}) = (1-q)\sum_{s=1}^{|\mathcal{S}|} q^s \cdot \frac{1}{s} \sum_{v_i, v_j \in \mathcal{S}, r_{v_i} < r_{v_j} = s} \mathcal{C}(v_i, v_j)$$

  - q is a pre-defined decay weight parameter, so that a higher ranked image contributes more to the function value.

- $T(S)$ **is a submodular and non-decreasing set function.** Maximizing $T(S)$ leads to a set of images which are highly ranked and similarly ranked with each other in the initial ranked lists.

Optimization

- The objective function $Q(S) = R(S) + \lambda T(S)$ is *submodular* and *monotonically Increasing*. It can be solved by a simple greedy algorithm.
- Maximizing a submodular function with a uniform matroid constraint yields a (1-1/e) approximation to the optimal solution.

Input: Graphs $\{\mathcal{G}_1, ..., \mathcal{G}_M\}$, initial ranked lists $\{\mathbf{r}_1, ..., \mathbf{r}_M\}$, K_s and λ.
Output: Reranked list $\mathbf{r}$ and final retrieved images $\mathcal{S}$.
Initialization: $\mathcal{S} \leftarrow \emptyset$, $\rho^{cur} \leftarrow 0$, $\mathbf{r} \leftarrow \mathbf{0}$
while $|\mathcal{S}| < K_s$ **do**
  $a^* = \underset{\mathcal{S} \cup \{a\} \in \mathcal{V}}{\arg\max}\ Q(\mathcal{S} \cup \{a\}) - Q(\mathcal{S})$
  if $Q(\mathcal{S} \cup \{a^*\}) \leq Q(\mathcal{S})$ **then**
    break;
  end if
  $\rho^{cur} \leftarrow \rho^{cur} + 1$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{a^*\}$; $r_{a^*} \leftarrow \rho^{cur}$
end

[§]G. Nemhauser, L. Wolsey and M. Fisher, An Analysis of Approximations for Maximization Submodular Set Functions - II, Mathematical Programming, 1978.

Experimental Results

- Evaluated Datasets:
  - ✓ ***Holidays***: 1491 image from 500 categories, where the first image in each category is used as a query.
  - ✓ ***Ukbench***: 10200 images from 2550 objects or scenes.
  - ✓ ***Oxford*** and ***Paris***: 5062 and 6412 photos of famous landmarks, respectively. Both datasets have 55 queries, where multiple queries are from the same landmark.
- Visual Features:
  - ✓ boW vectors from Hessian affine + SIFT descriptor using single assignment and approximate k-means (AKM). Standard tf-idf weighting is used.
  - ✓ 1192-dimension GIST feature.
  - ✓ 4000-dimension HSV color feature with 40 bins for H and 10 bins for S and V components.

Experimental Results

- Evaluation Criteria:
  - ✓ Mean average precision-MAP (in %) (*For Holidays, Oxford and Paris datasets*)
  - ✓ Average top 4 hits(N-S score) (*For Ukbench dataset*)
- Comparisons with state-of-the-art approaches. IG and RRC denote the results using only information gain term or using only relative ranking consistency term.

Datasets	BoW [32]	GIST [33]	Color	**Ours**	[10]	[7]	[8]	[18]	[32]	[16]	[34]	[35]	[19]	IG	RRC
Holidays	77.2	35.0	55.8	**84.9**	84.6	-	75.1	83.9	-	78.0	82.1	76.2	61.4	83.9	73.1
UKbench	3.50	1.96	3.09	**3.78**	3.77	3.45	-	3.64	3.67	3.56	-	3.52	3.36	3.75	3.54
Oxford	67.4	24.2	8.5	**74.3**	-	66.4	54.7	68.5	81.4	-	78.0	75.2	41.3	68.5	33.0
Paris	69.3	19.2	8.4	**74.8**	-	-	-	-	80.3	-	73.6	74.1	-	64.6	39.2

  - ✓ [10] also combines these three features (but of better performance) based on graph fusion.

Experimental Results

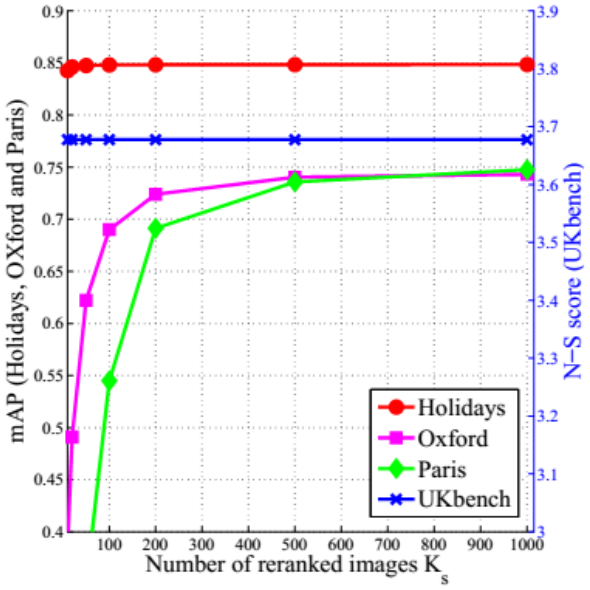
- Comparisons with other rank aggregation baseline approaches (mean rank, median rank, geometric mean rank, robust rank and borda count). Runtime (in seconds) of reranking 1000 images for a single query using direct greedy evaluation and lazy greedy evaluation is shown in the right-most columns.

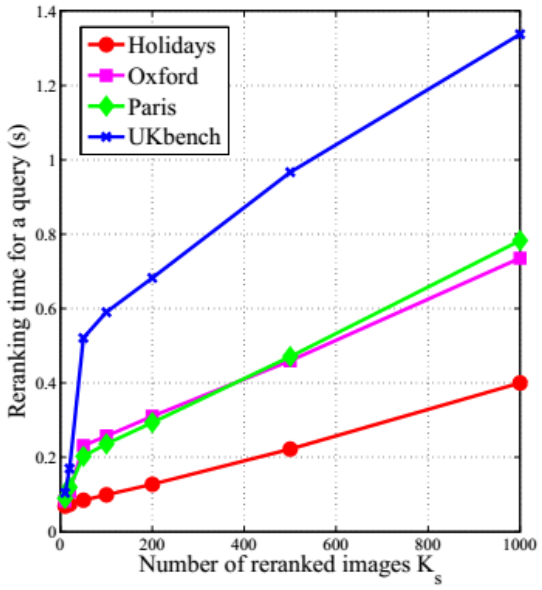
Datasets	Mean [36]	Median [37]	Geo-mean [37]	Robust [38]	Borda [37]	Ours	direct	lazy	speed-up
Holidays	59.2	71.7	76.4	71.5	59.2	**84.9**	16.5	0.40	41x
UKbench	2.89	3.47	3.50	3.33	2.89	**3.78**	55.7	1.34	42x
Oxford	18.6	34.7	40.5	35.6	18.6	**74.3**	38.3	0.74	52x
Paris	24.4	38.5	46.6	39.8	24.4	**74.8**	43.1	0.78	55x

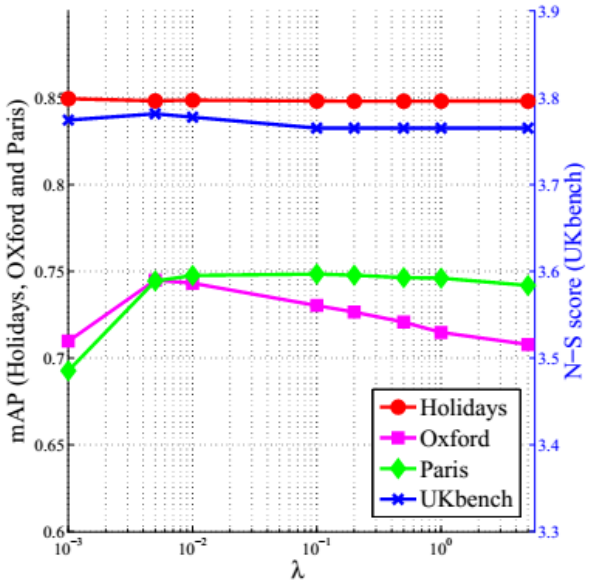
  - Our submodular reranking clearly outperforms other rank aggregation algorithms that do not use the inter-relationships amongst multiple ranked lists.
  - By use lazy evaluation, we further improve the speed by over 40 times.

Parameter Analysis

- Performances with different parameters







Left: Change of mean-average-precision performance (mAP) with respect to K_s
Middle: Average reranking time for a single query with respect to K_s
Right: Change of mAP with respect to λ

Conclusions

- We address the problem of reranking images with multiple feature modalities based on submodularity.
- We incorporate the information gain and relative ranking consistency into the objective function, which can effectively exploit the relationships of image pairs and multiple ranked lists at both the coarse level and the fine level.
- The objective function can be efficiently optimized by a simple greedy algorithm, which can provide a performance-guaranteed solution.
- It can be easily extended to other generic retrieval tasks with multiple independent ranked lists returned by multiple feature modalities.
- **Our source code will be released soon!**

References

- Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR. (2007) 1–8
- Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: ECCV. (2008) 304–317
- Zhang, S., Yang, M., Cour, T., Yu, K., Metaxas, D.N.: Query specific fusion for image retrieval. In: ECCV. (2012) 660–673
- Jegou, H., Douze, M., Schmid, C.: On the burstiness of visual elements. In: CVPR. (2009) 1169–1176
- Jegou, H., Chum, O.: Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening. In: ECCV. (2012) 774–787
- Wang, X., Yang, M., Cour, T., Zhu, S., Yu, K., Han, T.X.: Contextual weighting for vocabulary tree based image retrieval. In: ICCV. (2011) 209–216
- Aslam, J.A., Montague, M.H.: Models for metasearch. In: SIGIR. (2001) 275–284

References

- Qin, D., Gammeter, S., Bossard, L., Quack, T., Gool, L.J.V.: Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In: CVPR. (2011)
- Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42 (2001) 145–175
- Qin, D., Wengert, C., Gool, L.V.: Query adaptive similarity for large scale object retrieval. In: CVPR. (2013)
- Shen, X., Lin, Z., Brandt, J., Avidan, S., Wu, Y.: Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In: CVPR. (2012) 3013–3020
- Dwork, C., Kumar, R., Naor, M., Sivakumar, D.: Rank aggregation methods for the web. In: WWW. (2001) 613–622
- Kolde, R., Laur, S., Adler, P., Vilo, J.: Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics 28 (2012) 573–580
- Zheng, J., Jiang Z., Chellappa R., Phillips J.: Submodular Attribute Selection for Action Recognition in Video. In: NIPS. (2014)

References

- Kim, G., Xing, E.P., Li, F.F., Kanade, T.: Distributed cosegmentation via submodular optimization on anisotropic diffusion. In: ICCV. (2011) 169–176
- Krause, A., Cevher, V.: Submodular dictionary selection for sparse representation. In: ICML. (2010) 567–574
- Jiang,Z.,Zhang,G.,Davis,L.S.:Submodular dictionary learning for sparse coding. In: CVPR. (2012) 3418–3425
- Jiang, Z., Davis, L.S.: Submodular salient region detection. In: CVPR. (2013) 2043–2050
- Zhu, F., Jiang, Z., Shao, L.: Submodular object recognition. In: CVPR. (2014)
- Cao, L., Li, Z., Mu, Y., Chang, S.F.: Submodular video hashing: a unified framework towards video pooling and indexing. In: ACM Multimedia. (2012) 299–308
- Liu, M., Tuzel, O., Ramalingam, S., Chellappa, R.: Entropy rate superpixel segmentation. In: CVPR. (2011)
- Jegelka,S.,Bilmes,J.:Submodularity beyond submodular energies: Coupling edges in graph cuts. In: CVPR. (2011) 1897–1904

Thank you!

Questions?