

WHO AND WHERE: PEOPLE AND LOCATION CO-CLUSTERING

Zixuan Wang

Stanford University
zxwang@stanford.edu

Jinyun Yan

Rutgers, New Brunswick
jinyuny@cs.rutgers.edu

ABSTRACT

In this paper, we consider the clustering problem on images where each image contains patches in people and location domains. We exploit the correlation between people and location domains, and proposed a semi-supervised co-clustering algorithm to cluster images. Our algorithm updates the correlation links at the runtime, and produces clustering in both domains simultaneously. We conduct experiments in a manually collected dataset and a Flickr dataset. The result shows that the such correlation improves the clustering performance.

Index Terms— Co-clustering, Face Recognition, Location Recognition

1. INTRODUCTION

Given a large corpus of images, we want to cluster them such that images semantically related are grouped in one cluster. Semantics of an image refer to the information that image carries. For example, the face on the image is usually used to identify *who*. The background of the image refers to the location *where* the person was. All components together can convey what story has happened. In our case, we focus on two entities: *who* and *where*.

While there has been considerable work on automatic face recognition [2, 23] in images and even a modest effort on location recognition [4, 3], the coupling of the two is basically unexplored. An image which contains both people and location implies the co-occurrence of instances in two domains. For example, multiple photos taken at the same private location increase the confidence that similar faces on those photos are from a same person. Within a short time window, the same person on several photos indicates the affinity of locations.

Our framework, shown in Fig. 1, consists of two domains: people and locations. We take into account the inter-relation between two domains to enhance clustering in each domain. Three types of relations in people and location domains are considered: (1) *people-people* (2) *location-location* (3) *people-location*. A set of image patches is extracted and described in each domain. The similarity between patches within each domain is defined based on the visual appearances. The co-occurrence constraints are satisfied if patches from two domains appear in a same image. This relationship reflects the consistency of clustering results which is not embodied from visual appearances in a single domain.

We formulate the clustering task as an optimization problem which aims to minimize the within cluster distances and maximize the consistency across domains. We show this problem can convert to the semi-supervised kernel k-means clustering similar to [12]. However, we generate clustering results for two domains at the same time. During the iterative clustering process, constraints across domains and within domains keep updated. The main idea is that the

clustering result in one domain can aid the clustering in the other domain. We validate our approach with photos gathered from personal albums and a set of public photos crawled from Flickr.

Our contributions are threefold: 1) we propose a co-clustering algorithm for image clustering, focusing on people and locations; the algorithm couples both domains and explores underlying cross-domain relations; 2) our algorithm can simultaneously produce the clustering results of people and location, and outperforms clustering separately on each domain and the baseline co-clustering algorithm; 3) our algorithm is formulated as an optimization problem, which can be solved by through semi-supervised kernel k-means. It is robust and converges fast in practice.

2. RELATED WORK

Face is an important kind of visual objects in images, which is crucial to identify people. In recent years, there have been a lot of efforts in face detection [22], recognition [2, 23] and clustering [1]. The basic idea is to either represent a face as one or multiple feature vectors, or parameterize the face based on some template or deformable models. In addition to treating faces as individual objects, some researchers have been seeking for help from context information, such as background, people co-occurrence, etc. Davis et al. [5] developed a context-aware face recognition system that exploits GPS-tags, timestamps, and other meta-data. Lin et al. [14] proposed a unified framework to jointly recognize the people, location and event in a photo collection based on a probabilistic model.

Most location clustering algorithms are relying on the bag of words model [17]. Large-scale location clustering has been recently demonstrated in [13, 24], which use the GPS information to reduce the large-scale task down into a set of smaller tasks. Hays et al. [9] proposed an algorithm for estimating a distribution over geographic locations from the query image using a purely data-driven scene matching approach. They leveraged a dataset of over 6 million GPS-tagged images from the Flickr. When the temporal data is available in the corpora, it also helps to localize sequences of images.

However, clustering in people and location domains are usually treated as separate tasks. Location patches in photos with faces are not well exploited. While GPS information of the photo is not easily accessible, we propose a co-clustering algorithm, which simply use patches of the photo itself to discover the correlation in these two domains.

3. OUR APPROACH

In this section, we present the co-clustering framework to simultaneously cluster images in people and location domains. We have two major steps. The first step is pre-processing. We extract face and location patches from the corpus of images, and compute the visual features. The next step is co-clustering. The people-people,

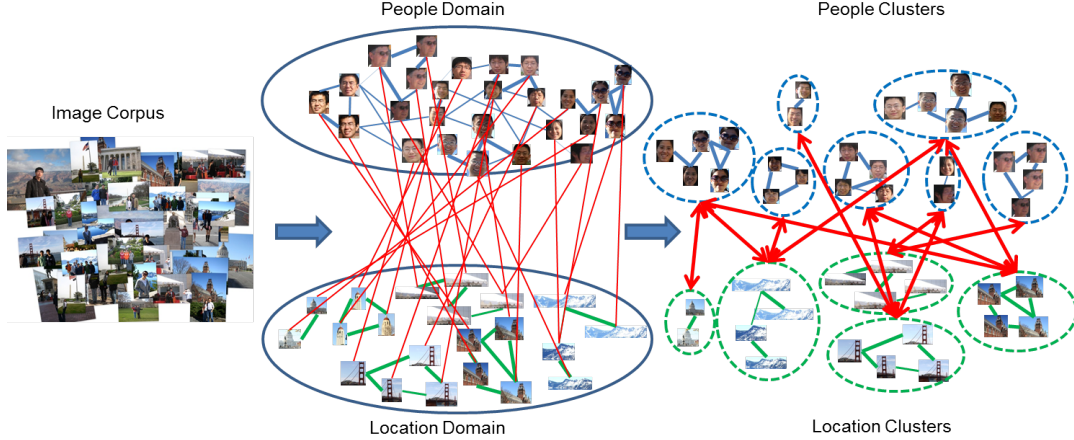


Fig. 1. The framework of people and locations co-clustering. The red lines represent the co-occurrence relations.

location-location and people-location relations are generated and updated. We describe the detail for each step below.

3.1. Pre-processing

We here describe how to extract features from people and location domains, and discover the relations between both domains.

People Domain: We use Viola-Jones face detector [22] to extract face patches from an image. To obtain high accuracy, a nested detector is applied to reduce the false positive rate. Every face will have a corresponding face patch. All face patches are normalized to the same size. We adopt the algorithm in [21] to detect seven facial landmarks from each extracted face patch. For each input face patch, four landmarks (outer eye corners and mouth corners) are registered to the pre-defined positions using the perspective transform. Then all seven facial landmarks are aligned by the computed perspective transform. For each landmark, two SIFT descriptors of different scales are extracted to form the face descriptor. We build a face graph over all face patches in the image collection. In the graph, each vertex represents a face patch. The weight of the edge is the similarity of face descriptors of two face patches.

Location Domain For each image, Hessian affine covariant detector [16] is used to detect interest points. The SIFT descriptor [15] is extracted on every interest point. The method similar to the work of Heath et al. [10] is used to discover the shared locations in the image collection. The content-based image retrieval [17] is applied to find top related images, and avoid quadratic pairwise comparisons. Lowe’s ratio test [15] is used to find the initial correspondences. RANSAC [7] is used to estimate the affine transform between a pair of images and compute feature correspondences between images. For every location patch, two types of features are extracted: a bag of visual words [20] and a color histogram. The bag of words descriptor summarizes the frequency that prototypical local SIFT patches occur. It captures the appearance of component objects. For images taken in an identical location, this descriptor will typically provide a good match. The color histogram characterizes certain scene regions well. These two types of features are concatenated to represent the location patch. A location graph is built similarly to the face graph. Each vertex in the graph represents a location patch. The weight of the edge is the similarity of location descriptors of two location patches.

Inter-relations across Domains To co-cluster across the people and location domains, several basic assumptions are made as fol-

lows.

Cannot Match Link. One person cannot appear twice in one image. Therefore, there is a *cannot match link* between a pair of face patches in the same image. Here we do not consider the exceptions like the photo collage or mirrors in the image. If two locations are far away according to the ground truth e.g. GPS signals, and two face patches appear in these two locations during a short time period, there is a *cannot match link* between this pair of patches. This assumption comes from that people cannot teleport within a short time period, for example, one people cannot appear in San Francisco and in New York within an hour.

Must Match Link. Two location patches are connected by a *must match link* if there is an affine transform found between them in the location graph construction. Because the links verified by RANSAC have high accuracy, we trust that they connect patches in the same location. Two location patches are connected by a *must match link* if they appear in the same image. Two different buildings may appear in the same image, therefore, in our setting, one location is defined as an area which may contain different backgrounds. Two location patches are connected by a *must match link* if they co-occur with the same people within a short time period. This assumption also comes from the fact that people cannot move too fast.

Possible Match Link. Two face patches that appear in the same location but not in the same image probably belong to the same people, due to the strong co-occurrence between the location and the face. This is true if the place has special meaning to the person, for example, his/her home or office, where he/she visits frequently. However, the assumption is not always true. For example, at tourist attractions, every people would take photos there. Therefore, the locations do not contribute much for the clustering in people domain. A weight is needed for the locations to distinguish the public locations and private locations. Private location is more helpful for clustering in people domain, while public location will introduce many noise.

3.2. Problem Formulation

We formulate our people and location co-clustering as an optimization problem. Given a set of feature vectors $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$, the goal of the standard k-means in each domain is to find a k -way disjoint partitioning (S_1, \dots, S_k) such that the following objective is mini-

mized:

$$f_{\text{kmeans}} = \sum_{c=1}^k \sum_{\mathbf{x}_i \in S_c} \|\mathbf{x}_i - \mathbf{m}_c\|^2 \quad (1)$$

where \mathbf{m}_c is the cluster center of c . The matrix E is defined as pairwise squared Euclidean distances among the data points, such that $E_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2$. We introduce an indicator vector \mathbf{z}_c for the cluster S_c .

$$\mathbf{z}_c(i) = \begin{cases} 1 & \text{if } i \in S_c \\ 0 & \text{if } i \notin S_c \end{cases} \quad (2)$$

where $\mathbf{z}_c^T \mathbf{z}_c$ is the size of cluster S_c , and $\mathbf{z}_c^T E \mathbf{z}_c$ gives the sum of E_{ij} over all \mathbf{x}_i and \mathbf{x}_j in S_c . Now the matrix \tilde{Z} is defined such that the c th column of \tilde{Z} is equal to $\mathbf{z}_c / (\mathbf{z}_c^T \mathbf{z}_c)^{1/2}$. \tilde{Z} is an orthonormal matrix, $\tilde{Z}^T \tilde{Z} = I_k$. Let N_F be the number of face patches and N_L be the number of location patches. k_F is the number of face clusters and k_L is the number of location clusters. By considering the relations between people and location domains, we write the objective as:

$$\begin{aligned} \text{minimize} \quad & f_F + f_L - f_{FL} - f_{LF} \\ \text{subject to} \quad & f_F = \text{tr}(\tilde{Z}_F^T E_F \tilde{Z}_F), \quad \tilde{Z}_F^T \tilde{Z}_F = I_{k_F}, \\ & f_L = \text{tr}(\tilde{Z}_L^T E_L \tilde{Z}_L), \quad \tilde{Z}_L^T \tilde{Z}_L = I_{k_L}, \\ & f_{FL} = \sum_{i=1}^t \text{tr}(M_i^T M_i), \quad M_i = \tilde{Z}_F^T C_{FL}^T T_i \tilde{Z}_L, \\ & f_{LF} = \text{tr}(N^T N), \quad N = \tilde{Z}_F^T C_{FL}^T P \tilde{Z}_L. \end{aligned} \quad (3)$$

E_F and E_L are pairwise squared Euclidean distance matrices in people and location domains. To integrate the must match constraints and cannot match constraints, the distance of the must match link is set to 0 and the distance of the cannot match link is set to $+\infty$. f_F and f_L with constraints $\tilde{Z}_F^T \tilde{Z}_F = I_{k_F}$ and $\tilde{Z}_L^T \tilde{Z}_L = I_{k_L}$ are the standard k-means optimization problems in people and location domains respectively.

The binary people-location co-occurrence matrix C_{FL} is defined as: the i th column of C_{FL} is the location patches that co-occur with the face patch i . For example, if the first column of C_{FL} is $(0, 0, 1, 0, 1, 0, \dots)^T$, which means the first face patch co-occurs with the third and the fifth location patches in the same image.

$C_{FL} \tilde{Z}_F$ is a clustering of location patches which is based on the face clustering result \tilde{Z}_F . Our goal is to maximize the consistency between the location clustering \tilde{Z}_L and $C_{FL} \tilde{Z}_F$. Location patches are weighted differently to reflect different semantic meanings of the people and location interactions. It is not difficult to discover the similarity between the definitions of f_{FL} and f_{LF} except the weight matrix T_i and P . f_{FL} optimizes the consistency that locations co-occur with the same people during a short time period should be one location. T_i is a $N_L \times N_L$ binary diagonal matrix that non-zero entries on the diagonal indicate these location patches are taken within a short time period. For example, $T_i = \text{diag}(0, 1, 0, 1, 1, \dots)$ means the second, the fourth and the fifth location patches have similar timestamps. There are t time constraints that are automatically learned from the meta-data of images.

f_{LF} optimizes the consistency that private locations are useful to identify people. P is a $N_L \times N_L$ diagonal weight matrix. It defines a score for each location patches. The private locations have larger weights and the private locations have small weights. The diagonal matrix P is defined as:

$$P_{ii} = \frac{\log(k_F / N_{FL_i})}{\log(k_F)} \quad (4)$$

where P_{ii} approximates 0 at public locations such as landmarks and it is approximate 1 at private locations. L_i is the location cluster that i belongs to. N_{FL_i} is the number of people appear in location L_i .

3.3. Alternative Optimization

The optimization problem (3) is not convex when the optimization variables involve \tilde{Z}_F and \tilde{Z}_L . Therefore, we use the alternative optimization by fixing variables in one domain and optimize on other variables and do this iteratively. When fixing variables, e.g. \tilde{Z}_F . The problem becomes a semi-supervised kernel k-means problem, which can be solved easily. We solve the problem following this sequence until the convergence: $\tilde{Z}_L \rightarrow \tilde{Z}_F \rightarrow P \rightarrow \tilde{Z}_L \rightarrow \tilde{Z}_F \rightarrow P \rightarrow \dots$. The first \tilde{Z}_F and \tilde{Z}_L are computed using the standard kernel k-means without cross-domain relations. After the initial clustering results are known, the weight matrix P can be computed using equation (4) and in the following iteration, the semi-supervised kernel k-means is used to integrate the cross-domains relations.

3.3.1. Semi-supervised Kernel K-means

We now briefly describe the existing semi-supervised kernel k-means algorithm [12]. The objective is written as the minimization of:

$$\sum_{c=1}^k \sum_{\mathbf{x}_i, \mathbf{x}_j \in S_c} \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{|S_c|} - \sum_{\substack{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{M} \\ c_i = c_j}} \frac{2w_{ij}}{|S_c|} + \sum_{\substack{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C} \\ c_i = c_j}} \frac{2w_{ij}}{|S_c|} \quad (5)$$

where \mathcal{M} is the set of must match link constraints, \mathcal{C} is the set of cannot match link constraints, w_{ij} is the penalty cost for violating a constraint between \mathbf{x}_i and \mathbf{x}_j , and c_i refers to the cluster label of \mathbf{x}_i . The first term in this objective function is the standard k-means objective function, the second term is a reward function for satisfying must match link constraints, and the third term is a penalty function for violating cannot match link constraints. The penalties and rewards are normalized by cluster size: if there are two points that have a cannot match link constraint in the same cluster, we will penalize higher if the corresponding cluster is smaller. Similarly, we will reward higher if two points in a small cluster have a must match link constraint. Thus, we divide each w_{ij} by the size of the cluster that the points are in.

Let A be the similarity matrix $A_{ij} = \mathbf{x}_i^T \mathbf{x}_j$ and let \tilde{A} be the matrix such that $\tilde{A}_{ij} = A_{ii} + A_{jj}$. Then, $E = \tilde{A} - 2A$. By replacing E in the trace minimization, the problem is equivalent to the minimization of $\text{tr}(\tilde{Z}^T (\tilde{A} - 2A - 2W) \tilde{Z})$. We calculate $\text{tr}(\tilde{Z}^T \tilde{A} \tilde{Z})$ as $2\text{tr}(A)$, which is a constant and can be ignored in the optimization. This leads to a maximization of $\text{tr}(\tilde{Z}^T (A + W) \tilde{Z})$. If we define a matrix $K = A + W$, our problem is expressed as a maximization of $\text{tr}(\tilde{Z}^T K \tilde{Z})$ and is mathematically equivalent to unweighted kernel k-means [6].

3.3.2. Alternative Optimization

If \tilde{Z}_F is fixed and \tilde{Z}_L is optimized. The objective f_{LF} can be written as:

$$f_{LF} = \text{tr}(\tilde{Z}_L^T P^T C_{FL} \tilde{Z}_F \tilde{Z}_F^T C_{FL}^T P \tilde{Z}_L) \quad (6)$$

The objective f_{FL} can be written as:

$$f_{FL} = \sum_{i=1}^t \text{tr}(\tilde{Z}_L^T T_i^T C_{FL} \tilde{Z}_F \tilde{Z}_F^T C_{FL}^T T_i \tilde{Z}_L) \quad (7)$$

We obtain the following optimization problem:

$$\begin{aligned}
& \text{maximize} && \text{tr}(\tilde{Z}_L^T(2A_L + \sum_{i=1}^t W_{Li} + Q_L)\tilde{Z}_L) \\
& \text{subject to} && W_{Li} = T_i^T C_{FL} \tilde{Z}_F \tilde{Z}_F^T C_{FL}^T T_i, \\
& && Q_L = P^T C_{FL} \tilde{Z}_F \tilde{Z}_F^T C_{FL}^T P, \\
& && \tilde{Z}_L^T \tilde{Z}_L = I_{k_L}.
\end{aligned} \tag{8}$$

where A_L is the affinity matrix in the location domain. This optimization problem can be solved by setting the kernel matrix $K_L = 2A_L + \sum_{i=1}^t W_{Li} + Q_L$. Similarly, if \tilde{Z}_L is fixed and \tilde{Z}_F is optimized. Using the fact that $\text{tr}(AB) = \text{tr}(BA)$ we can rewrite the f_{LF} and f_{FL} , and obtain the following optimization problem:

$$\begin{aligned}
& \text{maximize} && \text{tr}(\tilde{Z}_F^T(2A_F + \sum_{i=1}^t W_{Fi} + Q_F)\tilde{Z}_F) \\
& \text{subject to} && W_{Fi} = C_{FL}^T T_i \tilde{Z}_L \tilde{Z}_L^T T_i^T C_{FL}, \\
& && Q_F = C_{FL}^T P \tilde{Z}_L \tilde{Z}_L^T P^T C_{FL}, \\
& && \tilde{Z}_F^T \tilde{Z}_F = I_{k_F}.
\end{aligned} \tag{9}$$

where A_F is the affinity matrix in the face domain. This optimization problem can be solved by setting the kernel matrix $K_F = 2A_F + \sum_{i=1}^t W_{Fi} + Q_F$.

4. EVALUATIONS

We conduct experiments on two datasets to validate our approach. The first dataset contains images collected from personal albums with labeled ground truth. The second one uses a larger dataset crawled from online photo service: Flickr. We choose K-means with constraints [11] as the baseline algorithm. We also compare the performance of clustering on the each single domain by normalize cut [19] and Kmeans without any constraint. We use the RandIndex [18] to evaluate the performance of the clustering.

4.1. Personal Albums

This dataset contains 111 images collected from personal albums. In total it has 11 people and 13 locations. In the location domain, the top 50 image candidates are selected for the pairwise geometric verification. For each image, the bounding box of matched interest points is extracted as the location patch. A bag of words histogram (1000 visual words), 256-bin color histogram are extracted from each location patch. The dimension size of features in the location domain is 1, 256. We use a weight ratio of 1 : 1 for BoW:color features. All feature vectors are $L2$ normalized. In total, there are 146 face patches and 266 location patches.

In the dataset, each image associates a timestamp in the Exif header. The mean-shift [8] is used to cluster images in the time sequence and a matrix T_i is defined for each cluster of images. We cluster the face and location patches using the normalized cuts based on their appearance features as the baseline. K-means with constraints are also compared by adding the initial must match links and cannot match links in each domain. Figure 2 shows results in the people domain and the location domain.

From Figure 2, we observe the steady improvement on the clustering results when the number of clusters is larger than 2. The k-means with constraints are quite sensitive to the number of clusters. The best RandIndex values of methods across all K values are ordered as: Co-clustering, k-means-with-constraints and Normalize cuts. The values for these methods except Co-clustering do not vary

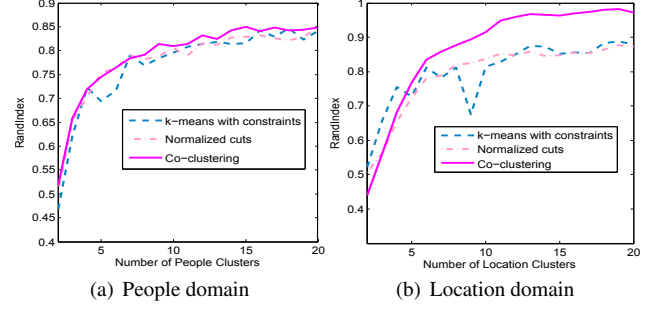


Fig. 2. RandIndex on the personal dataset.

much. The performance gain of Co-clustering in the location domain is very significant. It's mainly resulted from the must match link within the location domain. For the people domain, the difference in the clustering performance is very big, however, the steady increase over K is still promising.

4.2. Online Photo Sets

Dataset preparation: We use 140 names of public figures to query Flickr and filter out images without geo-location information. In total, we collect 53,800 images. We then filter out images without faces. The ground truth of the people domain is obtained directly from names. The ground truth of the location is obtained by clustering the longitude and latitude associated with images. We use the agglomerative clustering to discover location clusters. We consider each geo-location data including the longitude and the latitude as a point in the two dimensional space. In this dataset, we set the number of locations to be 100.

Figure 3 shows RandIndex values on the people domain and the location domain comparing k-means, Normalized cuts and Co-clustering over different K values. The improvement is not as big as that in the personal album dataset. It is mainly caused by the noise of the image set. The ground truth of the location domain is clustered by geo-location information which is not necessary equal to the location in the image. The ground truth of the people domain could also contain noise e.g. different people with the same name may appear together within one cluster. One future work is to find efficient algorithm to deal with the noise.

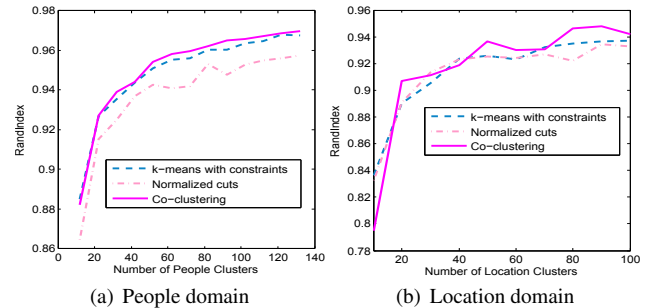


Fig. 3. RandIndex on the Flickr dataset.

5. CONCLUSION

We present a novel algorithm to co-cluster the people and location simultaneously. The relations across domains are used to enhance the clustering in single domain. We validate our approach using two datasets, and the experiment show that our algorithm performs better

than clustering in the single domain and the baseline co-clustering algorithm.

6. REFERENCES

- [1] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, E. Learned-Miller, and D. Forsyth. Names and faces in the news. *CVPR*, 2004.
- [2] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. *CVPR*, 2010.
- [3] C.-Y. Chen and K. Grauman. Clues from the beaten path: Location estimation with bursty sequences of tourist photos. In *CVPR*, 2011.
- [4] D. Chen, G. Baatz, Köser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices. 2011.
- [5] M. Davis, M. Smith, J. Canny, N. Good, S. King, and R. Janakiraman. Towards context-aware face recognition. In *ACM MM*, 2005.
- [6] I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: spectral clustering and normalized cuts. In *KDD*, 2004.
- [7] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981.
- [8] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 1975.
- [9] J. Hays and A. Efros. IM2GPS: estimating geographic information from a single image. *CVPR*, 2008.
- [10] K. Heath, N. Gelfand, M. Ovsjanikov, M. Aanjaneya, and L. Guibas. Image webs: Computing and exploiting connectivity in image collections. In *CVPR*, 2010.
- [11] S. R. Kiri Wagstaff, Claire Cardie and S. Schroedl. Constrained k-means clustering with background knowledge. In *ICML*, 2000.
- [12] B. Kulis, S. Basu, I. Dhillon, and R. Mooney. Semi-supervised graph clustering: A kernel approach. In *ICML*, 2005.
- [13] X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *ECCV*, 2008.
- [14] D. Lin, A. Kapoor, G. Hua, and S. Baker. Joint people, event, and location recognition in personal photo collections using cross-domain context. In *ECCV*, 2010.
- [15] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.
- [16] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 2004.
- [17] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.
- [18] W. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 1971.
- [19] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 2000.
- [20] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.
- [21] M. Uříčář, V. Franc, and V. Hlaváč. Detector of facial landmarks learned by the structured output SVM. In *VISAPP*, 2012.
- [22] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.
- [23] Q. Yin, X. Tang, and J. Sun. An associate-predict model for face recognition. In *CVPR*, 2011.
- [24] Y.-T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. Tour the world: building a web-scale landmark recognition engine. In *CVPR*, 2009.