Constructing Robust Affinity Graphs for Spectral Clustering

Xiatian Zhu[1], Chen Change Loy[2], Shaogang Gong[1]
[1]Queen Mary University of London, London E1 4NS, UK
[2]The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
xiatian.zhu@qmul.ac.uk, ccloy@ie.cuhk.edu.hk, s.gong@qmul.ac.uk

Abstract

Spectral clustering requires robust and meaningful affinity graphs as input in order to form clusters with desired structures that can well support human intuition. To construct such affinity graphs is non-trivial due to the ambiguity and uncertainty inherent in the raw data. In contrast to most existing clustering methods that typically employ all available features to construct affinity matrices with the Euclidean distance, which is often not an accurate representation of the underlying data structures, we propose a novel unsupervised approach to generating more robust affinity graphs via identifying and exploiting discriminative features for improving spectral clustering. Specifically, our model is capable of capturing and combining subtle similarity information distributed over discriminative feature subspaces for more accurately revealing the latent data distribution and thereby leading to improved data clustering, especially with heterogeneous data sources. We demonstrate the efficacy of the proposed approach on challenging image and video datasets.

1. Introduction

Spectral clustering is a popular clustering method [14, 15, 24, 25], which exploits the eigen-structure of a data affinity graph to partition data into disjoint subsets of similar samples. The performance of spectral clustering heavily relies on the goodness of the data affinity graph as it defines an approximation to the pairwise distances between data samples. In most contemporary techniques, the data affinity graph, *e.g.* a kNN graph, is constructed from a pairwise similarity matrix measured between samples. The notion of data similarity is often intimately tied to a specific metric function, typically the ℓ_2-norm (or the Euclidean metric) measured considering the whole feature space, with a Gaussian kernel to enforce locality.

Defining pairwise similarity for effective spectral clustering is fundamentally challenging [10] given complex data that are often of high dimension, heterogeneous, while no prior knowledge or supervision is available. Trusting all available features blindly for measuring pairwise similarities and constructing data graphs is susceptible to unreliable and/or noisy features, particularly so for real-world visual data, *e.g.* images and videos where signals can be intrinsically inaccurate and unstable owing to uncontrollable sources of variation, changes in illumination, context, occlusion and background clutters [7]. Moreover, confining the notion of similarity to the ℓ_2-norm metric implicitly imposes unrealistic assumption on complex data structures that do not necessarily possess the Euclidean behaviour.

Our goal is to infer robust pairwise similarity between samples so as to construct more meaningful affinity graphs for improved spectral clustering. To this end, we formulate a unified and generalised data similarity inference framework based on the unsupervised clustering random forest with three innovations. (1) Instead of considering the complete feature space as a whole, the proposed model is designed to avoid less informative features by measuring between-sample proximity via discriminative feature subspaces, yielding similarity graphs that better express the underlying semantic structure in data. (2) We relax the Euclidean assumption for data similarity inference by following the information-theoretic definition of data similarity presented in [11], which states that different similarities can be induced from a given sample pair if distinct propositions are taken or different questions are asked about data commonalities. Motivated by the same idea, our model derives pairwise similarities of arbitrary sample pairs from an exhaustive set of comparative tests, using different feature variables with distinct inherent semantics as criteria. Such subtle similarities distributed over discriminative feature subspaces are combined automatically and effectively for producing robust pairwise affinity matrices. (3) The pairwise affinity matrix generated by the proposed model automatically possesses the local neighbourhood. Thus, no additional Gaussian kernel is needed to enforce locality.

We demonstrate the effectiveness of the proposed approach on both image and video datasets. Specifically, we show the advantages of using the proposed affinity

1063-6919/14 $31.00 © 2014 IEEE
DOI 10.1109/CVPR.2014.188

graph learning model for clustering challenging visual data when compared against both the baseline and the state-of-the-art methods including the Euclidean-distance-based k nearest neighbour (kNN) [23], Dominant Neighbourhoods (DN) [16], Consensus of kNN (cons-kNN) [18], as well as non-metric based unsupervised manifold forests [4, 17, 26].

2. Related Work

A large body of work has been conducted on spectral clustering with focus on different aspects and applications [20, 15, 14, 25, 5, 24, 8, 19]. In general, existing approaches to improving spectral clustering performance can be classified into two paradigms: (1) How to improve data grouping when the method of generating a data affinity matrix is fixed [20, 15, 24]. For example, Xiang and Gong [24] propose to identify informative and relevant eigenvectors of a data affinity matrix; (2) How to construct robust affinity graphs so as to improve the clustering results using standard spectral clustering algorithms [25, 23, 16, 18]. Our approach is related to the second paradigm.

Approaches to adapting to the local data structures for improving the robustness of affinity graphs have been proposed [25, 23]. Particular focus has been spent on learning an adaptive scaling factor σ for the Gaussian kernel (also known as radial basis function or heat kernel) $\exp\left(-\frac{\text{dist}^2(\mathbf{x}_i,\mathbf{x}_j)}{\sigma^2}\right)$, when computing the similarity between samples $\mathbf{x}_i$ and $\mathbf{x}_j$. These methods, however, are still susceptible to the presence of noisy and irrelevant features.

To mitigate the above issue, Pavan and Pelillo [16] propose a graph-theoretic algorithm for forming tight neighbourhoods via selecting the maximal cliques (or maximising average pairwise affinity), with the hope of constructing graphs with fewer false affinity edges between samples. More recently, a kNN based graph generation method is proposed in [18] where the consensus information from multiple kNNs is used for discarding noisy edges and identifying strong local neighbourhoods. In contrast to all the aforementioned methods that blindly trust all available variables, the proposed graph inference method exploits discriminative and informative features for measuring more robust data pairwise similarities. The resulting affinity matrix is thus more robust against noisy real-world visual data.

Random forest-based affinity graph construction has been attempted in [21, 4, 26]. The intuition is that tree leaf nodes contain discriminative data partitions, which could be exploited for generating robust affinity graphs. We show that the above approaches are special cases of our affinity inference method. Specifically, we propose a generalised model, which is not only capable of learning discriminative feature subspaces for robust affinity graph construction as in previous methods, but also able to further exploit the hierarchical structure of random forest to better capture subtle and weak data proximity.

3. Robust Affinity Graph Construction

The proposed affinity graph construction approach is built upon clustering random forests, which are an unsupervised form of random forests. A clustering forest is an ensemble of T_{clust} binary decision trees learned independently from each other, each with a training set $X^t \subset X$ drawn randomly from the whole training dataset $X = \{\mathbf{x}_i\}_{i=1}^N, \mathbf{x}_i \in \mathbb{R}^d$, where N denotes the sample number in X and d the feature dimension of data sample. The proposed model has a few important merits:

1. Our model is purely unsupervised without requiring any ground truth annotations, since it is based on clustering forests rather the more popular supervised classification or regression random forests [2, 4].
2. By virtue of the random subspace feature selection during training forests, the pairwise affinity matrix generated by our model is less susceptible to corruption of noisy and irrelevant features.
3. Each decision tree in the forest hierarchically encodes an exhaustive set of comparative tests or split functions, which implicitly define different notions of between-sample similarities. Our model is capable of extracting and combining these subtle similarities at distributed discriminative subspaces for learning robust pairwise affinity matrices.

Below, we first briefly describe how to train individual decision trees of a clustering forest, with particular focus on its discriminative feature selection (Sec. 3.1). We then discuss how to derive robust pairwise similarities from the trained forest (Sec. 3.2).

3.1. Clustering Decision Tree Training

Each decision tree of a clustering forest contains a set of internal (or split) and leaf (or terminal) nodes organised in a hierarchical fashion. Every internal node is associated with a question or split function, which attempts to partition the arriving training data into left or right child nodes. By adopting the pseudo two-class algorithm [2, 13], the training of a clustering forest can be accomplished using a similar strategy of learning a classification forest. Specifically, the learning of a clustering/classification forest involves the optimisation of a binary split function in every split node. The binary split function is defined as

$$\text{h}(\mathbf{x}, \boldsymbol{\vartheta}) = \begin{cases} 0, & \text{if } \mathbf{x}_{\vartheta_1} < \vartheta_2, \\ 1, & \text{otherwise.} \end{cases} \quad (1)$$

This split function is parameterised by two parameters $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2)$: (i) a feature dimension $\vartheta_1 \in \{1, \ldots, d\}$, and (ii) a feature threshold $\vartheta_2 \in \mathbb{R}$. All arrival samples S of a split node s will be channelled to either the left l or right r child nodes, according to the output of Eqn. (1).

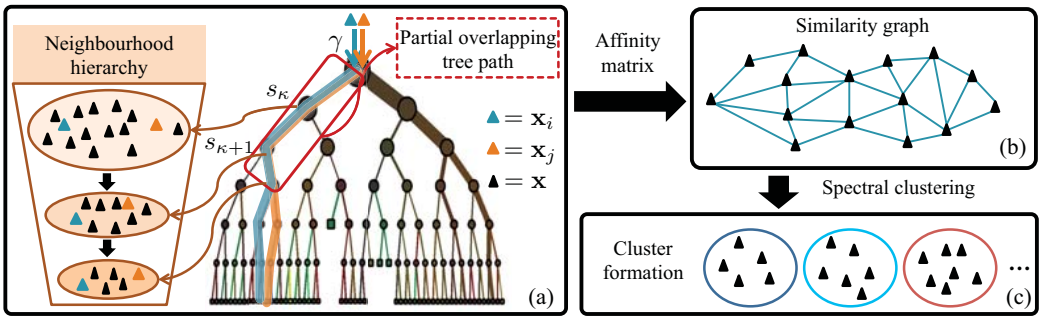


Figure 1. The pipeline of data clustering, with focus on the hierarchical neighbourhoods along a tree path in a clustering tree, which are formed by selecting and employing discriminative features. We exploit the hierarchical tree structures and neighbourhoods for robust data pairwise similarity inference.

The optimal split parameter $\boldsymbol{\vartheta}^*$ is chosen via

$$\boldsymbol{\vartheta}^* = \underset{\Theta}{\operatorname{argmax}}\, \Delta\mathcal{I}, \tag{2}$$

where $\Theta = \{\boldsymbol{\vartheta}^i\}_{i=1}^{m_{\text{try}}(|S|-1)}$ represents a parameter set over m_{try} randomly selected features. The cardinality of a set is given by $|\cdot|$. Typically, a greedy search strategy is exploited to identify $\boldsymbol{\vartheta}^*$. The information gain $\Delta\mathcal{I}$ is formulated as

$$\Delta\mathcal{I} = \mathcal{I}_s - \frac{|L|}{|S|}\mathcal{I}_l - \frac{|R|}{|S|}\mathcal{I}_r, \tag{3}$$

where L and R denote the sets of data routed into l and r, and $L \cup R = S$. The information criterion $\mathcal{I}$ can be either the entropy or the Gini impurity [3]. In this study, we use the Gini impurity due to its simplicity and efficiency.

By doing so, an internal node s selects the most *discriminative* (*i.e.* maximising the information gain) feature from m_{try} candidates as its split variable and exploits it to partition the training data S. This process is repeated throughout the whole tree training stage until some stopping criterion is satisfied, *e.g.* the number of training samples S arriving at a node is equal to or smaller than a threshold ϕ. After the node splitting process stops, leaf nodes are formed. Importantly, each internal node is attached to an identified discriminative feature as its split variable.

3.2. Structure-Aware Robust Affinity Inference

The above training procedure allows us to partition data with very complex distributions at the discovered discriminative feature subspaces. Each split function (Eqn. (1)) encodes a different notion of between-sample similarity, defined by its split variable and threshold.

To quantify data similarities for generating a robust pairwise affinity matrix, we propose a *structure-aware affinity inference model (ClustRF-Strct)* based on clustering random forest. The model takes into account the whole tree hierarchical structures, *i.e.* a tree path from the root until leaf nodes traversed by data samples $\mathbf{x}$ (Fig. 1-(a)). Specifically, given the t-th clustering tree, we channel a sample pair $(\mathbf{x}_i, \mathbf{x}_j)$ from the root node γ until reaching their respective leaf nodes ℓ^i and ℓ^j. Subsequently, two tree paths composed by the root node γ, internal and leaf nodes can be generated:

$$\mathcal{P}^i = \{\gamma, s_1^i, \dots, s_\kappa^i, \dots, \ell^i\}, \tag{4}$$
$$\mathcal{P}^j = \{\gamma, s_1^j, \dots, s_\kappa^j, \dots, \ell^j\}, \tag{5}$$

with s_κ^i and s_κ^j denoting the κ-th internal nodes traveled by $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively.

Intuitively, a sample pair $(\mathbf{x}_i, \mathbf{x}_j)$ is considered dissimilar if they are split at the very beginning, *e.g.* from the root node γ. On the other hand, if the samples travel together passing the same set of internal nodes till the identical leaf node, *i.e.* $\mathcal{P}^i = \mathcal{P}^j$, their similarity is high. Beyond the two extreme cases above, there exist intermediate similarities: let λ the length of which $\mathcal{P}^i$ and $\mathcal{P}^j$ overlaps (Fig. 1-(a)), *i.e.*

$$\begin{cases} s_\kappa^i = s_\kappa^j & \text{if } \kappa = \{1, \dots, \lambda\}, \\ s_\kappa^i \neq s_\kappa^j & \text{if } \kappa = \{\lambda+1, \dots\}, \\ \ell^i \neq \ell^j. \end{cases} \tag{6}$$

Clearly, a larger value in λ signifies more split tests both samples $(\mathbf{x}_i, \mathbf{x}_j)$ have gone through together, implying higher similarity shared between them. A lower value in λ suggests subtle and weak similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. To capture different strengths of data similarities, we derive a principled and generalised tree structure aware data pairwise similarity inference method, ClustRF-Strct, as

$$a_{i,j}^t = \frac{\sum_{\kappa=1}^{\lambda} w_\kappa}{\sum_{\kappa=1}^{M} w_\kappa}, \tag{7}$$

where $M = \max(|\mathcal{P}^i|, |\mathcal{P}^j|) - 1$, and w_κ is the weight assigned to the corresponding tree node (*i.e.* either s_κ or ℓ) on the longer tree path. Note that the root node γ is not considered in computing the similarity since all samples share the same root node. The pairwise similarity $a_{i,j}^t$ defines the individual elements of a tree-level affinity matrix $A^t \in \mathbb{R}^{N\times N}$. To combine consensus from multiple decision trees in the forest, we generate the final smooth affinity matrix $A \in \mathbb{R}^{N\times N}$ as

$$A = \frac{1}{T_{\text{clust}}} \sum_{t=1}^{T_{\text{clust}}} A^t. \tag{8}$$

ClustRF-Strct is regarded as a generic affinity inference model since distinct strategies of defining node weights w_i can produce different affinity graph construction methods/instantiations, as we will describe below.

3.2.1 Variant I - The Binary Affinity Model

We show that the methods proposed in [4, 17, 26] are special cases of the proposed ClustRF-Strct. All these methods share the same mechanism in estimating a pairwise similarity matrix using a clustering random forest. We name these methods collectively as *the binary affinity inference model (ClustRF-Bi)*, since they derive pairwise affinity based only on whether or not (binary) two samples fall into the same leaf node of a tree.

Prior to discussing their relationship to our approach, we review the underlying mechanism of ClustRF-Bi in measuring pairwise similarity between data samples given a learned clustering forest. Recall that each individual tree of a forest partitions the training samples at its leaves $\ell(\mathbf{x})$: $\mathbb{R}^d \rightarrow \mathbb{L} \subset \mathbb{N}$, where ℓ represents a leaf node index and $\mathbb{L}$ refers to the set of all leaves in a given tree. For each tree, the ClusterRF-Bi model first computes a tree-level $N \times N$ affinity matrix A^t with elements defined as

$$a^t_{i,j} = \exp^{-\text{dist}^t(\mathbf{x}_i, \mathbf{x}_j)}, \quad \text{with} \tag{9}$$

$$\text{dist}^t(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 0, & \text{if } \ell(\mathbf{x}_i) = \ell(\mathbf{x}_j), \\ +\infty, & \text{otherwise.} \end{cases} \tag{10}$$

With Eqn. (9), the ClustRF-Bi assigns the maximal similarity $a^t_{i,j} = 1$ to a sample pair $(\mathbf{x}_i, \mathbf{x}_j)$ if $\mathcal{P}^i = \mathcal{P}^j$ (*i.e.* completely overlapping), and the minimum similarity $a^t_{i,j} = 0$ to them otherwise, regardless of any partial overlap in their tree paths. This formulation is equivalent to setting $w_\kappa = 0$ for every internal node, $w_\kappa = 1$ for all leaf nodes in Eqn. (7). Hence, this mechanism is a special case of our ClustRF-Strct. A potential problem with ClustRF-Bi is that it may lose the weak and subtle proximity of sample pairs proportional to the degree of path overlap. We will show in our experiments in Sec. 5 that considering only completely overlapping path pairs, *i.e.* $\mathcal{P}^i \backslash \mathcal{P}^j = \emptyset$, as in ClustRF-Bi, is not sufficient for producing satisfactory data clusters.

3.2.2 Variant II - The Uniform Structure Model

To address the limitation of ClustRF-Bi in losing weak similarity between data samples, we propose to consider the non-completely-overlapping path pairs as well while measuring tree-level data similarities using the proposed ClustRF-Strct model. In particular, we treat all tree nodes as uniformly important by setting $w_\kappa = 1$ in Eqn. (7). Therefore, Eqn. (7) can be rewritten as

$$a^t_{i,j} = \frac{\lambda}{\max(|\mathcal{P}^i|, |\mathcal{P}^j|) - 1}. \tag{11}$$

We call this model as *ClustRF-Strct-Unfm*. With Eqn. (11), all partially overlapped path pairs also contribute to the similarity estimation between samples. As shown in the experiments (Sec. 5), this new formulation captures weak data similarities encoded in the tree structures, and thus is capable of better revealing the underlying data structure than the conventional ClustRF-Bi model.

3.2.3 Variant III - The Adaptive Structure Model

The ClustRF-Strct-Unfm is capable of capturing subtle and weak data proximity through exploiting the path sharing mechanism of sample pairs in the hierarchical structure of the forest. Nevertheless, the uniform node weighting implies an implicit assumption that all tree nodes (*e.g.* s_κ or ℓ) are equally important in defining similarity. In reality this may not be true, particularly with data of complex distributions, since different nodes reside at distinct layers of the tree hierarchy with dissimilar properties, *e.g.* the size and structure of the arrival training samples. To characterise such node (or data subset) properties, we propose an *adaptive structure-aware affinity inference (ClustRF-Strct-Adpt)*.

The ClustRF-Strct-Adpt model exploits the *hierarchical neighbourhood* formed in each clustering tree (see Fig. 1-(a)). Our notion of hierarchical neighbourhood generalises the idea presented in [12]. Specifically, [12] only regards samples sharing the same tree terminal node as neighbours. We extend the neighbourhood notion to the whole tree hierarchy. Imagine a situation where a target sample $\mathbf{x}_t$ traverses in a tree hierarchy from the root node until some arbitrary internal node s_κ. Some other samples $S_\kappa \setminus \mathbf{x}_t$ have also gone through the same tree path and fall onto the same internal node s_κ with $\mathbf{x}_t$. These samples form a neighbourhood with $\mathbf{x}_t$ on node s_κ in the tree hierarchy.

Samples that form a hierarchical neighbourhood have passed through the same set of split functions (Eqn. (1)) associated with each tree node. Intuitively, the deeper the hierarchical neighbourhood is formed, the higher the similarity shared among the samples in the same neighbourhood, since those samples have survived and are still connected after identical discriminative split tests (Eqn. (1)). Motivated by this observation, we assign each tree node s_κ with a scale-adaptive weight (Eqn. (7)) as

$$w_\kappa = \frac{1}{|S_\kappa|}. \tag{12}$$

Consequently, we assign larger weights to deeper tree nodes, since $|S_\kappa| > |S_{\kappa+1}|$. As such, ClustRF-Strct-Adpt estimates similarity between a sample pair $(\mathbf{x}_i, \mathbf{x}_j)$ via

$$a^t_{i,j} = \frac{\sum_{\kappa=1}^{\lambda} \left(\frac{1}{|S_\kappa|}\right)}{\sum_m \left(\frac{1}{|S^b_m|}\right) + \frac{1}{|\Lambda^b|}}, \tag{13}$$

Table 1. Datasets for experiments, with examples in Figure 2.

Dataset	# Clusters	# Features	# Samples
Image Segmentation [1]	7	19	2310
CMU-PIE [22]	10	1024	1000
USAA [6]	8	14000	1466
ERCe [26]	6	2672	600

where

$$\hat{b} = \underset{b \in \{i,j\}}{\operatorname{argmax}} |\mathcal{P}^b|, \tag{14}$$

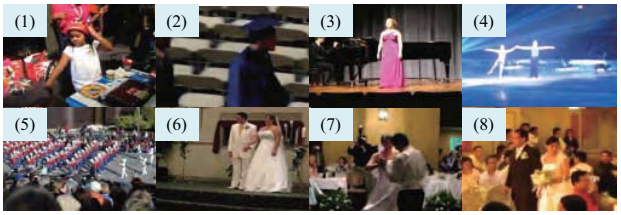
i.e. the cumulated neighbourhood size of the longer tree path is utilised as the normalisation factor, and $\Lambda^{\hat{b}}$ denotes the set of data samples reaching into the leaf node $\ell^{\hat{b}}$. Similar to Eqn. (11), a maximum similarity is assigned to sample pairs that share the same leaf node. Nevertheless, the tree node similarity weight is no longer distributed linearly along the forest hierarchy as in Eqn. (11), but in a non-linear way adaptive to the size of hierarchical neighbourhood.

4. Experimental Setting

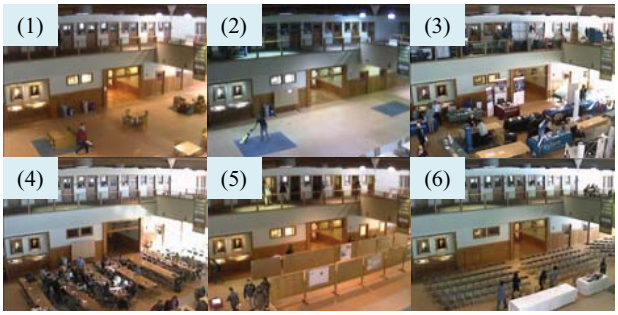
Datasets – A wide range of visual datasets are utilised for evaluating the proposed model: (1) Image Segmentation [1]: a scene image dataset from the UCI repository, including 7 types of different outdoor scenes: Brickface, Sky, Foliage, Cement, Window, Path, and Grass. The objective is to partition image patches into the above seven types. (2) CMU-PIE [22]: a face image dataset drawn from CMU-PIE. It comprises 10 different persons selected in random, each with 100 images of near frontal poses and various expressions and lighting conditions (Fig. 2-(a)). We aim to group together all the face images from the same person on this dataset. (3) USAA [6]: a YouTube video dataset. This dataset features common social group activities where unconstrained space of objects, events and interactions makes them intrinsically complex and challenging to detect (Fig. 2-(b)). The goal is to cluster these video clips into 8 groups each with coherent semantics, *e.g*. the same social activity. (4) ERCe [26]: a visual surveillance video dataset. The dataset is challenging because of various types of physical events characterised by large changes in the environmental setup, participants, and crowdedness, as well as intricate activity patterns. This dataset consists of 600 video clips from 6 campus events, each with 100 samples (Fig. 2-(c)). Our purpose is to classify the ERCe video clips into the six events.

Features – For Image Segmentation, USAA, and ERCe, we use the same features as provided by [1], [6] and [26]. Specifically, for Image Segmentation, we use the low-level visual features from image patches, *e.g*. colour, pixel intensity. These appearance features may be unreliable and noisy, especially given outdoor scenes. As to USAA, the resulting high-dimensional (14000-D) feature vectors are drawn from three heterogeneous modalities, namely static

(a) CMU-PIE [22]: each row corresponds to one person.



(b) USAA [6]: (1) Birthday Party, (2) Graduation, (3) Music Performance, (4) Non-music Performance, (5) Parade, (6) Wedding Ceremony, (7) Wedding Dance, (8) Wedding Reception.



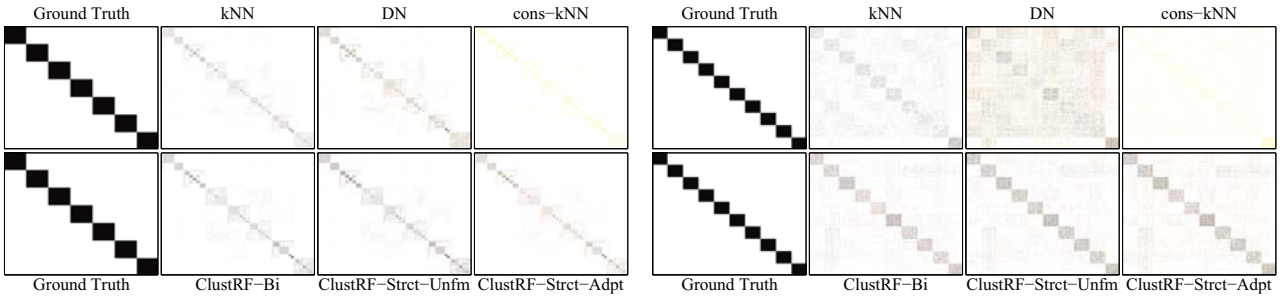
(c) ERCe [26]: (1) Student Orientation, (2) Cleaning, (3) Career Fair, (4) Group Study, (5) Gun Forum, (6) Scholarship Competition.

Figure 2. Examples from CMU-PIE [22], USAA [6], ERCe [26] datasets.

appearance, motion and auditory. The data samples from ERCe are also of high-dimensional (2672-D), involving heterogeneous feature types, *e.g*. colour histogram (RGB and HSV), optical flow, local texture, holistic image appearance, object detection. With CMU-PIE, we first normalise and crop the face images into 32×32 in spatial resolution, and their raw pixel values are then employed as the representation. Such a representation is affected by large differences in illumination, facial expression, and head pose. All data features are scaled to the range of $[-1, 1]$. To initially remove less-informative features on the high-dimensional datasets, *e.g*. CMU-PIE, USAA and ERCe, we perform PCA on them and the first 30 dominant components are used as the final representation. The same sets of feature data are used across all methods for fair comparison.

Baselines – We compare the proposed affinity graph learning model ClustRF-Strct with:

1. *k Nearest Neighbours (kNN)* [23]: the most traditional affinity graph construction method using the Euclidean distance on the input feature space. To convert an Euclidean distance matrix D into an affinity graph A, we compute each element in A as $a_{i,j} =$



(a) Image Segmentation [1] (b) CMU-PIE [22]

Figure 3. Qualitative comparison of the affinity graphs generated by different methods. Better to view by Zoom-In.

$\exp(-\text{dist}^2_{i,j}/\sigma^2_{i,j})$ with σ_{ij} the adaptive kernel size that is computed as the mean distance of M-nearest neighborhoods as in [23]. We will evaluate the sensitivity of M on the clustering performance in Sec. 5.

2. *Dominant Neighbourhoods (DN)* [16]: a tight affinity graph learning approach. To reduce the amount of potentially noisy edges in a given Euclidean affinity graph, the DN model attempts to identify sparse and compact neighbourhoods through selecting only the maximal cliques in the input graph.
3. *Consensus of kNN (cons-kNN)* [18]: the state-of-the-art affinity graph construction method. For selecting strong local neighbourhoods, the consensus information collected from various neighbourhoods in a provided kNN graph is exploited by this algorithm for producing a more robust affinity graph.
4. *ClustRF-Bi* [4, 17, 26]: the clustering random forest binary affinity model (Sec. 3.2.1). This method exploits discriminative features identified during the training of clustering forests to construct data affinity graphs. The resulting affinity graphs can thus be less-sensitive to noisy features, compared to the Euclidean-metric-based methods, *e.g.* kNN, DN and cons-kNN.

Evaluation metrics – We use the widely adopted adjusted Rand Index (ARI) [9] as the evaluation metric, with the range of $[-1, 1]$. ARI measures the agreement between the clustering results and the ground truth in a pairwise fashion, with higher values indicating better clustering quality. For all experiments involving clustering forest based models, *i.e.* ClustRF-Bi, ClustRF-Strct-Unfm, and ClustRF-Strct-Adpt, we report the ARI values averaged over 5 trials.

Implementation details – The number of trees T_{clust} in a clustering forest is set to 1000. We observed stable results given a larger forest size. This observation agrees with [4]. We set m_{try} (see Eqn. (2)) to $\sqrt{d}$ with d the feature dimensionality of the input data and employ a linear data separation [4] as the split function (see Eqn. (1)). The value of ϕ is obtained through cross-validation on each dataset.

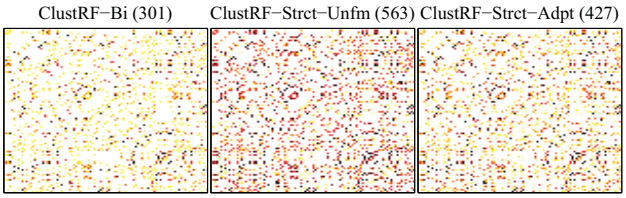


Figure 4. Comparison on cluster forest based models: the pairwise affinity between different face images from the same person (CMU-PIE [22]). The numbers in the parentheses are the summation of all pairwise similarities induced by the corresponding method. Larger is better.

5. Evaluations

5.1. Evaluation of Affinity Graph

We first examine the data affinity graphs, which could qualitatively reflect how effective a neighbourhood graph construction method is. Figure 3 depicts some example affinity matrices generated by all comparative models.

It can be observed that ClustRF-Strct-Unfm and ClustRF-Strct-Adpt produce affinity matrices with more distinct block structure and less false edges compared with others. This suggests the superiority of the proposed models in learning the underlying semantic structures in data, potentially leading to more compact and separable clusters. A number of noisy pairwise edges are found in the affinity graphs yielded by ClustRF-Strct-Unfm than those by ClustRF-Strct-Adpt. This is a consequence of not considering the goodness of hierarchical neighbourhoods in ClustRF-Strct-Unfm (Sec. 3.2.2), leading to less accurate induced data similarities in comparison to ClustRF-Strct-Adpt. This observation shows the effectiveness of the proposed adaptive weighting mechanism in suppressing noisy or inaccurate features on learning data sample proximity.

We now examine and discuss the characteristics of affinity matrices constructed by other baselines. It is observed from Fig. 4 that compared to the ClustRF-Strct models, ClustRF-Bi has the tendency to underestimate the similarity of sample pairs that actually originate from the same clusters. This is owing to that ClustRF-Bi only assumes data similarity on the completely overlapped tree path pairs,

Table 2. Sensitivity of M: the clustering results of different methods given varying values of M in terms of AUC, with M the parameter used for computing the adaptive Gaussian kernel size during the process of converting a Euclidean distance matrix into an affinity graph (see Sec. 4).

Dataset	Image Segmentation [1]					CMU-PIE [22]					USAA [6]					ERCe [26]				
M	20	40	60	80	100	20	40	60	80	100	20	40	60	80	100	20	40	60	80	100
kNN [23]	34.8	36.2	37.6	37.8	37.9	4.4	4.4	4.9	4.8	4.7	3.5	3.1	3.3	3.6	3.6	45.9	48.1	52.1	52.7	51.8
DN [16]	38.3	29.1	34.7	37.2	37.2	3.0	2.3	2.4	3.0	3.5	2.6	2.3	2.5	2.0	1.7	51.0	52.1	49.9	18.3	25.6
cons-kNN [18]	34.9	36.8	35.8	36.8	35.9	4.0	4.4	4.3	4.3	4.2	3.8	3.8	3.8	3.8	3.9	49.2	52.1	52.0	52.0	55.7
ClustRF-Bi [4, 17, 26]	39.5					19.8					4.5					56.1				
ClustRF-Strct-Unfm	40.7					**22.9**					4.7					59.3				
ClustRF-Strct-Adpt	**41.8**					20.5					**5.7**					**60.4**				

and thus loses subtle and weak data proximity (Sec. 3.2.1). Given intrinsically ambiguous datasets with unreliable features, incomplete overlapping path pairs can often occur as samples of the same categories may only share similarity in some feature subspaces. In such cases, ClustRF-Bi shall perform poorly as compared to our ClustRF-Strct models, as we shall show next.

With kNN, DN, and cons-kNN, affinity graphs with indistinct block structure are observed, with a mix of large quantity of faulty edges. In contrast to ClustRF-Bi that is 'overly reluctant' in assigning data proximity to sample pairs, the Euclidean distance based methods go to the other extreme by blindly believing all available features and therefore tend to introduce false data proximity.

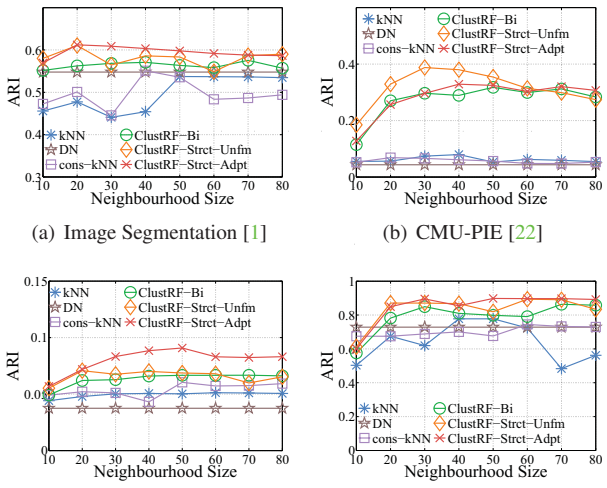
5.2. Evaluation of Clustering Performance

In this experiment, we quantitatively evaluate data clustering performance of different graph construction methods by applying the spectral clustering algorithm [25] on their affinity graphs as discussed in Sec. 5.1.

It is observed from Fig. 5 and Table 2 ClustRF-Strct-Unfm and ClustRF-Strct-Adpt outperform baseline methods, *e.g.* by as much as $>$**125**% and $>$**120**% relative improvement against kNN, $>$**190**% and $>$**180**% against DN, $>$**130**% and $>$**125**% against the state-of-the-art cons-kNN, $>$**5**% and $>$**10**% against the discriminative-feature-based model ClustRF-Bi in terms of the area under the ARI curve averaged over all the datasets. This is in line with the observations in Fig. 3. Importantly, we find that ClustRF-Strct-Unfm and ClustRF-Strct-Adpt significantly outperform the Euclidean distance based methods on CMU-PIE. This can be due to the capability of our model of capturing and aggregating subtle data proximity distributed over discriminative feature subspaces, thus suitable to handle ambiguous and unreliable features caused by variation in illumination, face expression or pose on the CMU-PIE data. A large improvement margin is also observed on the USAA dataset with data collected from heterogeneous sources. All these evidences suggest the superior capability of our model in dealing with high-dimensional data and heterogeneous sources for generating robust affinity graphs.

As shown in Fig. 5, ClustRF-Strct-Unfm is more likely to suffer when the size of neighbourhood k increases, whilst ClustRF-Strct-Adpt behaves more stably. The tendency is likely to be caused by the relatively noisier affinity matrix induced by ClustRF-Strct-Unfm, as we observed in Sec. 5.1. The results further justify the importance of considering neighbourhood-scale-adaptive weighting on tree nodes (Sec. 3.2) for suppressing data noise.

The Euclidean-distance-based models produce the poorest results over all the datasets. Inaccurate and noisy features are potential causes. For example, the face images from the CMU-PIE dataset are intrinsically ambiguous owing to large variations in illumination and expressions (Fig. 2-(a)). The extracted features are therefore unreliable. Similar situations are observed on other datasets. The cons-kNN model attempts to circumvent this problem via searching for consensus from multiple kNNs. Nevertheless this is proved challenging, particularly when a large quantity of potential noisy edges exist in the given kNN due to the unreliable input data, leading to possibly inconsistent neighbour votes from multiple kNNs. DN is likely to suffer from the same problem as the maximal cliques in the given affinity graph is no longer trustworthy. This interpretation is further supported by the fact that for all kNN, cons-kNN and DN, the clustering performance changes dramatically with the varying settings of neighbourhood size k, *e.g.* on Image Segmentation and ERCe. That is, a large amount of



(a) Image Segmentation [1] (b) CMU-PIE [22]

(c) USAA [6] (d) ERCe [26]

Figure 5. ARI curve: comparison between different methods on the spectral clustering performance given different scales of neighbourhood k. The neighbourhood size M used on computing the adaptive Gaussian kernel size is fixed to 20.

inaccurate edges in the affinity graphs lead to the requirement of a more careful neighbourhood size selection, so as to trade-off between the true and false data similarities.

By exploiting discriminative features, the ClustRF-Bi model suffers less from noisy data, and produces better results than the Euclidean-distance-based methods. However, it is inferior to the proposed ClustRF-Strct variants, since it is not capable of capturing subtle data pairwise similarity encoded in partially overlapped path pairs.

Sensitivity of M – Here we evaluate the sensitivity of M on kNN, DN and cons-kNN. The parameter M is employed to estimate the adaptive Gaussian kernel size for converting a Euclidean distance matrix into a similarity graph [23] (Sec. 4). Note that ClustRF-Bi, ClustRF-Strct-Unfm and ClustRF-Strct-Adpt are free from M since they directly derive affinity graphs from the learned forests, rather than from distance matrices which require a Gaussian kernel to enforce locality. It is evident from Table 2 that for all the Euclidean-distance-based affinity graph learning models, a careful selection of adaptive Gaussian kernel size can produce better clustering results. However, their best results are still worse than those by clustering forest based models, due to the limitation in handling intrinsically noisy and irrelevant feature data. Importantly, the proposed ClustRF-Strct model gains superior performance to other baselines in all cases.

6. Conclusion

We have presented a novel generalised and unsupervised approach to constructing more robust and meaningful data affinity graphs for improving spectral clustering, particularly with data of high dimension and from heterogeneous sources. Instead of blindly trusting all available variables, we adopt an information-theoretic definition on data similarity and derive affinity graphs through capturing and combining subtle and weak data pairwise proximity distributed in discriminative feature subspaces identified during the training stage of clustering forests. Furthermore, the affinity graphs constructed by our model naturally possess the local neighbourhood, with no need of Gaussian kernel. Extensive experiments on clustering challenging visual datasets have demonstrated the superiority of the proposed affinity inference model over the state-of-the-art models. Beyond spectral clustering, our model can also benefit other applications, *e.g.* manifold ranking.

References

[1] A. Asuncion and D. Newman. UCI machine learning repository, 2007. 5, 6, 7

[2] L. Breiman. Random forests. *ML*, 45(1):5–32, 2001. 2

[3] L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and regression trees*. Chapman & Hall/CRC, 1984. 3

[4] A. Criminisi. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2-3):81–227, 2012. 2, 4, 6, 7

[5] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nystrom method. *TPAMI*, 26(2):214–225, 2004. 2

[6] Y. Fu, T. Hospedales, T. Xiang, and S. Gong. Learning multimodal latent attributes. *TPAMI*, 2013. 5, 7

[7] S. Gong, C. C. Loy, and T. Xiang. Security and surveillance. In *Visual Analysis of Humans*, pages 455–472. Springer, 2011. 1

[8] H.-C. Huang, Y.-Y. Chuang, and C.-S. Chen. Affinity aggregation for spectral clustering. In *CVPR*, 2012. 2

[9] L. Hubert and P. Arabie. Comparing partitions. *J CLASSIF*, 2(1):193–218, 1985. 6

[10] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010. 1

[11] D. Lin. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304, 1998. 1

[12] Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. *J AM STAT ASSOC*, 101(474):578–590, 2006. 4

[13] B. Liu, Y. Xia, and P. S. Yu. Clustering through decision tree construction. In *CIKM*, 2000. 2

[14] U. Luxburg. A tutorial on spectral clustering. *STAT COMPUT*, 17(4):395–416, 2007. 1, 2

[15] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2002. 1, 2

[16] M. Pavan and M. Pelillo. Dominant sets and pairwise clustering. *TPAMI*, 29(1):167–172, 2007. 2, 6, 7

[17] Y. Pei, T.-K. Kim, and H. Zha. Unsupervised random forest manifold alignment for lipreading. In *ICCV*, 2013. 2, 4, 6, 7

[18] V. Premachandran and R. Kakarala. Consensus of k-nns for robust neighborhood selection on graph-based manifolds. In *CVPR*, 2013. 2, 6, 7

[19] O. Shamir and N. Tishby. Spectral clustering on a budget. In *AISTATS*, pages 661–669, 2011. 2

[20] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 22(8):888 –905, 2000. 2

[21] T. Shi and S. Horvath. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, pages 118–138, 2006. 2

[22] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. *TPMAI*, 25(12):1615–1618, 2003. 5, 6, 7

[23] J. Wang, S.-F. Chang, X. Zhou, and S. Wong. Active microscopic cellular image annotation by superposable graph transduction with imbalanced labels. In *CVPR*, 2008. 2, 5, 6, 7, 8

[24] T. Xiang and S. Gong. Spectral clustering with eigenvector selection. *PR*, 41(3):1012–1029, 2008. 1, 2

[25] L. Zelnik-manor and P. Perona. Self-tuning spectral clustering. In *NIPS*, 2004. 1, 2, 7

[26] X. Zhu, C. C. Loy, and S. Gong. Video synopsis by heterogeneous multi-source correlation. In *ICCV*, 2013. 2, 4, 5, 6, 7