Computational Visual Media
DOI 10.1007/s41095-xxx-xxxx-x Vol. x, No. x, month year, xx–xx

Article Type (Research/Review)

Unsupervised Random Forest for Affinity Estimation

Yunai Yi[1], Diya Sun[1], Peixin Li[1], Tae-Kyun Kim[2], Tianmin Xu[3], and Yuru Pei[1] ✉

Abstract This paper presents an unsupervised clustering random-forest-based metric for affinity estimation of the large and high-dimensional data. The combinational criteria for the node splitting in forest construction is feasible to handle the rank-deficiency when measuring the clustering compactness. The binary forest-based metric is extended to continuous metrics by exploiting both the common traversing path and the smallest shared parent node. The proposed forest-based metric is feasible and efficient for affinity estimation by passing down data pairs in the forest with a limited number of decision trees. A pseudo-leaf-splitting (PLS) algorithm is introduced to account for spatial relationships, which regularises affinity measures and relieves inconsistent leaf assignments. The random-forest-based metric with the PLS facilitates the establishment of consistent and point-wise correspondences. The proposed method has been applied to automatic phrase recognition using color and depth videos and point-wise correspondence. Extensive series of experiments demonstrate the effectiveness of the proposed method in affinity estimation compared with the state-of-the-art.

Keywords Affinity estimation, forest-based metric, unsupervised clustering forest, pseudo-leaf-splitting.

1 Introduction

Affinity estimation is an essential step in various computer vision and image processing tasks. The affinity estimation of motion trajectories, for example, is utilized in motion segmentation [12, 43] and action recognition [50]. The automatic phrase recognition employs the trajectory affinity to define motion patterns in color and depth videos [38]. The point-to-point affinity and shape correspondence are essential for attribute transfer and data reuse [8, 30, 37, 44, 45], as well as the shape comparisons in morphological studies [9, 39]. It is, however, computationally non-trivial to estimate pairwise affinities for a large-scale dataset, where the complexity grows quadratically dependent on the cardinality of the dataset. Some distance metrics, such as the earthmover distance, make the computation cost higher for high-dimensional data. This paper presents the unsupervised random-forest-based metric for efficient affinity estimation, demonstrating its efficacy on automatic phrase recognition and point-wise correspondence of a shape corpus.

The random forest has gained popularity in computer vision for decades, being well-known for its scalability and real-time testing as a valuable generalization to unseen data [19, 21, 24, 27, 35, 47, 54]. The clustering random forest works in an unsupervised fashion [10, 18, 34, 46, 55, 58, 59] to estimate the underlying data distribution and affinity without prior labels. Alzubaidi et al. [2] utilized the density forest [18] with a Gaussian distribution assumption in tree nodes, where the clustering compactness was measured by the covariance matrix. However, the zero-valued determinant in the case of rank-deficiency makes the criteria invalid. The combinational node splitting criteria as an integration of the trace-based distribution measurement and the scatter index [38] are feasible to handle the rank-deficiency for optimal node splitting.

Recently, a series of researches address the forest-based metric for affinity estimation. The cascaded clustering forest (CGF) was proposed to refine the voxel-wise affinity by iteratively updated geodesic coordinates [40] with a set of clustering models. Mixed metric random forest (MMRF) utilized the self-learning of data distributions for matching consistencies across images [41], taking advantage of the weak labeling and classification criterion

1 Key Laboratory of Machine Perception (MOE), Department of Machine Intelligence, Peking University, Beijing 100871, China. E-mail: yrpei@pku.edu.cn.
2 Department of Electrical and Electronic Engineering, Imperial College London, London, UK.
3 School of Stomatology, Stomatology Hospital, Peking University, Beijing 100081, China.

Manuscript received: ; accepted:

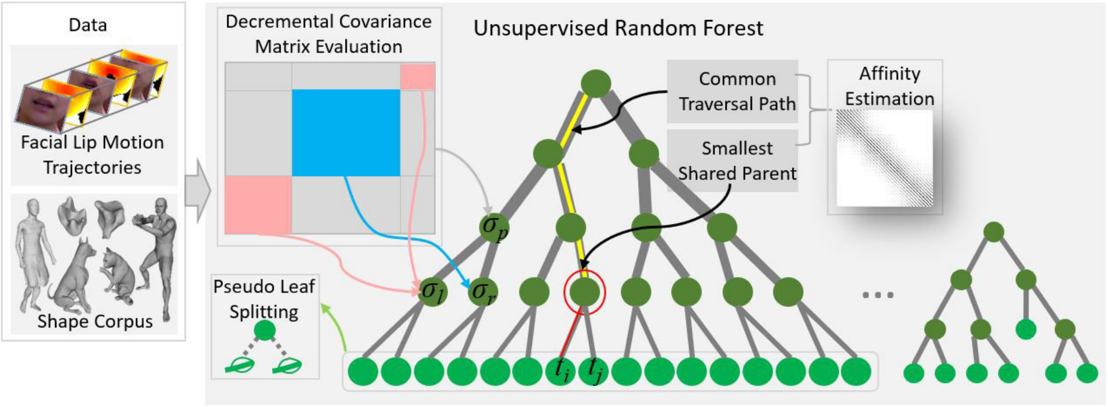

Fig. 1 The proposed unsupervised random forest-based metric for affinity estimation. The forest-based continuous metric is defined by utilizing both the length of the common traversing path and the cardinality of the smallest shared parent node. A pseudo-leaf-splitting algorithm is proposed to account for spatial relationships, regularising affinity measures and inconsistent leaf assignments. The decremental covariance matrix evaluation technique is used to ease the learning complexity.

to optimize node splitting. The oblique clustering forest (OCF) [48] extended the splitting criterion from traditional orthogonal hyperplanes to oblique hyperplanes, reducing the tree depth and the model complexity. The spatial consistent (SC) clustering forest employed a data-dependent learning guarantee of unsupervised clustering randomized trees [42]. The above clustering forests introduce additional computations, such as cascaded clustering models [40], the fine-tuning with the penalized weighting of the classification entropy [41], the dominant principal component and regressions [48], and the data-dependent learning guarantee for tree pruning [42], to improve performances on data clustering and affinity estimation. In contrast, this work does not introduce additional computation costs to construct the clustering forest. Instead, we extend the binary forest-based metric to a continuous one for affinity estimation. In light of the observation that training the unsupervised clustering forest is typically more time-consuming than the supervised classification forest due to the entropy estimation of the high-dimensional data, the decremental covariance matrix evaluation technique is introduced to avoid the assessment of covariance matrices from scratch and ease the learning complexity.

Affinities are measured efficiently by the hierarchical clustering forests, in contrast to the learning-based feature fusion for the affinity graph by the iterative optimization of convex problems [32]. Two points are intuitively assumed to be similar in case that they arrive at the same leaf. The generalized forest-based metric is derived by the average affinities from individual trees. The forest-based binary metric has been used to measure data similarity [18, 58]. The continuous affinity measure has been proposed based on the common traversal path from the root to leaf nodes as well as the node cardinality on the path [59]. To relieve the weight computation on the traversal path, we present a forest-based metric as a linear combination of normalized common-traversal-path-based and the smallest-shared-parent-based metrics. The proposed metric takes into account both the unbalanced data distribution and partial similarity. Once given the pairwise affinities of a dataset, it is straightforward to compute the low-dimensional embedding. Ganapathi-Subramanian et al. [22] constructed a joint latent embedding function as a combination of diffusion embedding and a linear mapping for descriptor transport in a shape corpus, where the nonlinear embedding function relied on the predefined feature descriptors. The paper addresses the forest-based metric and affinity estimation. The embedding is conducted by the existing multi-dimensional scaling (MDS) algorithm [1], which is computed based on affinity estimation without explicit representation learning.

This work introduces a pseudo-leaf-splitting (PLS) algorithm to handle the inconsistent leaf assignments, since the random forest built upon independent data points is insufficient to accommodate global data structures. The random-forest-based metric with the PLS regularises the point-wise correspondences. The proposed PLS technique differs from existing methods [25, 26, 36, 44] in that it bridges the gap between separate point-wise correspondence and consistency refinements. The deep learning-based methods have been used for shape correspondence [7, 23, 33, 52], which are learned from prior ground truth correspondence or the metric space alignment. The 3DN [52] and the 3D-coded [23] were the unsupervised end-to-end network to infer global displacements fields between a shape and the template, utilizing Chamfer and earthmover distance-based loss functions. The FMNet [33] optimized a feature extraction network via a low-dimensional spectral map. The ADD3 used anisotropic diffusion-based spectral feature descriptors [7]. The FMNet [33] and ADD3 [7] are learned in a supervised manner, requiring prior ground truth correspondence. Unlike deep neural network-based descriptor learning, this work exploits unsupervised forest-based metric learning for point-wise correspondence.

This paper presents a combined forest-based metric and a PLS regularization scheme to improve the forest-based metric for affinity estimation, as shown in Fig. 1. The main contributions of this work are: (1) the continuous forest-based metric is presented by exploiting both the common traversing path and the cardinality of the smallest shared parent node, enabling efficient and effective affinity estimation of the large and high-dimensional data. (2) The PLS scheme is proposed to regularise the forest-based metric to account for the global spatial and structural relationships, relieving inconsistent leaf assignments. (3) The proposed method has gained success in affinity estimation of facial trajectories and 3D points, enabling efficient and automatic phrase recognition and consistent correspondence of 3D shape corpus compared with the state-of-the-art.

2 Unsupervised Random Forest

Given the unlabeled dataset $T = \{t_i | i = 1, ..., N\}$, a set of trees are trained independently. The unsupervised density forest estimates the underlying data distribution using a Gaussian distribution assumption [18]. The combinational node splitting criteria integrate a trace-based distribution measurement and a scatter index [38]. The objective function I of the j-th node with data T_j is defined as follows.

$$I = -\sum_{i=l,r} \frac{m_{T_j^i}}{m_{T_j}} \ln\left(tr\left(\sigma(T_j^i)\right)\right) + \lambda \frac{\|\mu_l - \mu_r\|_\infty}{\sum_{i=l,r} \phi(T_j^i, \mu_i)}, \tag{1}$$

where $tr(\cdot)$ is the matrix trace. σ denotes the covariance matrix of the Gaussian distribution. $m_{T_j^i}$ denotes the size of the left or the right children nodes, and m_{T_j} the parent node size. $\phi(T_j^i, \mu_i) = \max_{t \in T_j^i} \|t - \mu_i\|_\infty$. μ_l and μ_r are the centroids of the left and right child nodes. The constant λ is set to 50 empirically.

Since the covariance matrices need to be repeatedly evaluated when given the randomly selected parameters, it is time-consuming to evaluate the covariance matrix σ from scratch for the optimal splitting parameters when building the forest. This work introduces a decremental covariance matrix evaluation technique (see Appendix A). The complexity of the covariance matrix evaluation is reduced from $O(m\rho^2)$ to $O(\rho)$ by the decremental technique, where m is the cardinality of the node. ρ denotes the data dimensionality. The trace evaluation complexity is reduced to $O(\kappa\rho)$ given κ randomly selected parameters.

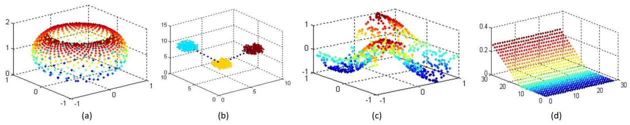

Fig. 2 Toy datasets. (a) $punctured\,sphere$, (b) $3D\,clusters$, (c) $twin\,peaks$, and (d) $corner$.

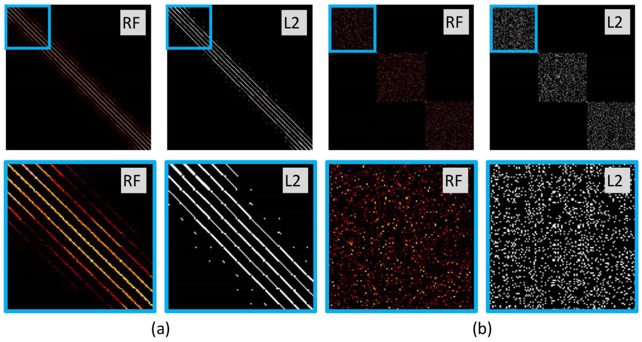

Fig. 3 The affinity matrices obtained by the proposed forest-based metric and the $L2$-norm followed by kNN on (a) $corner$ and (b) $3D\,clusters$ datasets.

3 Forest-based Affinity Estimation

3.1 Binary Forest-Based Metric

The forest leaves L define a partition of the training data. When feeding an instance t to a tree, it will finally reach a leaf $\ell(t) \in L$, after a sequence of binary tests stored in the branch nodes. When the instances are assigned to the same leaf node, they are assumed to be similar and the pairwise affinity is set to 1, and 0 otherwise. The symmetric affinity matrix $\mathcal{A}$ is defined as a weighted combination of $\mathcal{A}_k$ from independent trees.

$$\mathcal{A} = \frac{1}{n_T} \sum_{k=1}^{n_T} \mathcal{A}_k, \tag{2}$$

where n_T is the tree number. Since only points inside one leaf node are considered to be similar, the affinity matrix from the random forest automatically accounts for neighboring relationships. Thus, $\mathcal{A}$ can be viewed as a geodesic affinity matrix of the original dataset. On the contrary, when using the L_2 distance metric, there is no prior on local neighboring relationships. The kNN-like algorithm is needed to find neighbors from the pairwise distance matrix with additional time costs.

The affinity matrix obtained by the binary metric is often relatively sparse since only the point pairs sharing the same leaf node are assumed to be similar. Generally speaking, the leaf node should not be too small to account for the affinity of the dataset. Moreover, randomized trees are required to provide enough similar candidate points in leaf nodes.

3.2 Continuous Forest-Based Metric

Aside from the binary affinity, a continuous forest-based metric is proposed based on the common path $\mathbb{P}_{ij}$ of two instances t_i and t_j when they traverse from the root to leaves $\ell(t_i)$ and $\ell(t_j)$. The distance $d_{cp}(t_i, t_j)$ is computed by the common path as follows:

$$d_{cp}(t_i, t_j) = \frac{\nu_{ij} - |\mathbb{P}_{ij}|_o}{\nu_{ij}}, \tag{3}$$

where $\nu_{ij} = \max(\nu_i, \nu_j)$ is the maximum depth of $\ell(t_i)$ and $\ell(t_j)$. $|\cdot|_o$ returns the cardinality of a set. If two instances reach the same leaf node, the distance is zero. Otherwise, the distance is set to 1 when the two instances have no common path. The binary affinity definition is a special case of Eq. (3) by setting the common path to null in case the instances do not reach the same leaf. However, there is no guarantee that the decision tree is balanced for an arbitrary dataset. In this case, the similarity is defined based on the cardinality of the data stored in the smallest shared parent (SSP) node $T_{p_{ij}}$ of $\ell(t_i)$ and $\ell(t_j)$.

$$d_{sp}(t_i, t_j) = \frac{|T_{p_{ij}}|_o - \zeta_{ij}}{|T_r|_o - \zeta_{ij}}, \tag{4}$$

where $\zeta_{ij} = \min(|\ell(t_i)|_o, |\ell(t_j)|_o)$ is the minimum leaf size of $\ell(t_i)$ and $\ell(t_j)$. When t_i and t_j go into the same leaf node, the SSP node $T_{p_{ij}}$ is the leaf itself, and distance d_{sp} is zero. On the other hand, when the shared parent node is the largest one, i.e. the root node T_r, d_{sp} is set to 1. In case that the leaf size n_l is selected as the termination criterion of the tree growth, the above SSP-based metric can be simplified as $d_{sp}(t_i, t_j) = \vartheta(|T_{p_{ij}}|_o - n_l)$, where the normalization constant $\vartheta = (|T_r|_o - n_l)^{-1}$. For the unbalanced data distribution, the distance between two instances in the small cluster is shorter than that in the large cluster based on the above definition in Eq. (4). It is rational considering two instances are likely to be far apart in the large cluster. Compared with the adaptive forest-based metric [59], here the cardinality of the SSP node is used to determine the affinity without the weight computation in the shared traversal path. The combined forest-based metric d_f is defined as a linear fusion of the common path-based d_{cp} and the SSP-based d_{sp}.

$$d_f = w_{cp} d_{cp} + w_{sp} d_{sp}, \tag{5}$$

where the constant weight $w_{cp} + w_{sp} = 1$. The entry in the affinity matrix $\mathcal{A}$ is defined as $\mathcal{A}_{ij} = 1 - d_f(t_i, t_j)$.

Proposition 1. *The functions defined in Eq. (3), Eq. (4), and Eq. (5) are non-negative metrics with following properties:*

- *Identity:* $d(t_i, t_i) = 0$;
- *Positivity:* $d(t_i, t_j) \geq 0$;
- *Symmetry:* $d(t_i, t_j) = d(t_j, t_i)$;
- *Triangle inequality:* $d(t_i, t_k) \leq d(t_i, t_j) + d(t_j, t_k)$.

The proof of Proposition 1 is given in Appendix B. The above binary, the common-path-based, the SSP-based, and the combined distance metrics are applied to a set of toy data (see Fig. 2 and Fig. 3). The difference $e_{\mathcal{A}}$ between the affinity matrices $\mathcal{A}$ computed by the clustering forest-based metrics and $\mathcal{A}_{L_2}$ by the L_2 norm and the kNN is shown in Fig. 4.

$$e_{\mathcal{A}} = \frac{\|\mathcal{A} \oplus \mathcal{A}_{L_2}\|_F^2}{n_{\mathcal{A}}}, \tag{6}$$

where $\oplus$ is the *xor* of matrix entries. $n_{\mathcal{A}}$ is the size of $\mathcal{A}$. $\|\cdot\|_F$ is the Frobenius norm. The combined random-forest-based metric can achieve the lower $e_{\mathcal{A}}$ than the binary, the common-path-based, and the SSP-based metrics. All metrics can reduce the difference $e_{\mathcal{A}}$ when enlarging the forest size. The Dice similarity metric [20] e_I is used to compare the k nearest neighbors obtained by the proposed metrics with those by the L2 norm as shown in Fig. 5. The nearest neighbors obtained by the combined random-forest-based metric are more consistent with the L2 metric than other metrics. We observe that the consistency increases with the enlarging forest size. Moreover, when enlarging the forest size, the performance of the binary random-forest-based metric can be similar to that of the combined metric (see Fig. 4(a) and Fig. 5(a)), because a large number of randomized decision trees tend to provide enough neighboring candidates.

The look-up of feature values and the comparison with the thresholds when traversing trees are very fast and negligible in time. Although the cost of pairwise distances of small subsets or sampled point pairs is much lower than the dense pairwise distance computation, the kNN-graph-based method is time-demanding for the high-dimensional dataset. The proposed forest traversal and leaf assignments have a linear complexity regarding the data size. More importantly, the time complexity of our method has no relations with the dimensionality, which is desirable for the high dimensional data. In the extreme case of forest-based metric, i.e., the binary metric, there are no multiplication operations in the affinity estimation. Since the instances in the same leaf node are assumed to be similar, the complexity depends on the number of the leaf nodes, and there is no pairwise distance computation by the binary forest-based metric. As to the continuous metrics, such as d_{sp}, there are just normalization operations in the affinity estimation.

4 Pseudo Leaf Splitting

It is efficient to acquire the pairwise affinity matrix between datasets by the random-forest-based metric. However, there is no regularization for the point-wise

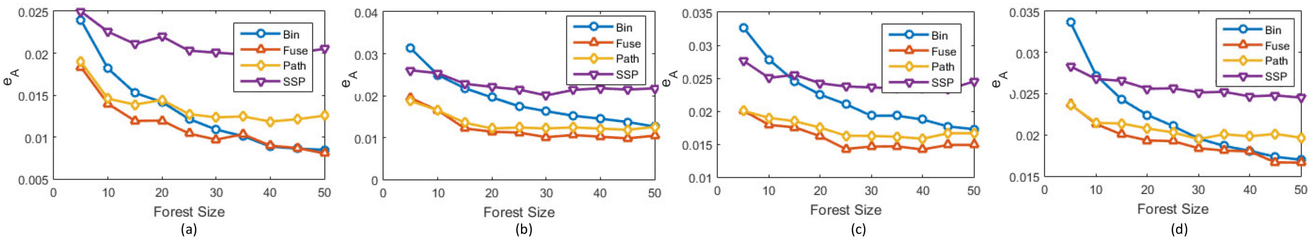

Fig. 4 The affinity matrices difference $e_{\mathcal{A}}$ of the combined forest-based metric (Fuse), the binary metric (Bin), the common-path-based metric (Path), and the SSP-based metrics on four toy datasets, including (a) $corner$, (b) $punctured\, sphere$, (c) $twin\, peaks$, and (d) $3D\, clusters$.

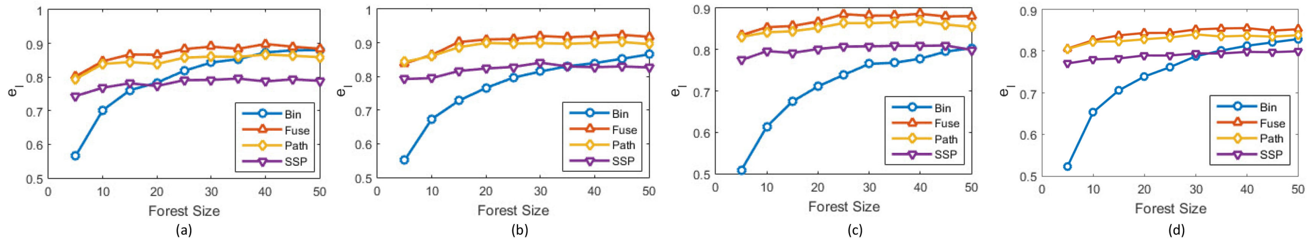

Fig. 5 The Dice similarity e_I of nearest-neighbors obtained by the combined forest-based metric (Fuse), the binary metric (Bin), the common-path-based metric (Path), and the SSP-based metrics on four toy datasets, including (a) $corner$, (b) $punctured\, sphere$, (c) $twin\, peaks$, and (d) $3D\, clusters$.

correspondence because the random forest is built upon the independent feature descriptors without considering the relationship. For instance, when establishing correspondence C between dataset X and Y, the forest-based metric can be used to produce the candidate matching pair $\{(x_i, y_i) \in C | x_i \in X, y_i \in Y\}$. The above correspondence has no guarantee for the relationship preservation, i.e., $g(x_i, x_j) \propto g(y_i, y_j)$ when $(x_i, y_i) \in C$ and $(x_j, y_j) \in C$. g is some function to measure the relationship, e.g. the geodesic distance on 3D mesh surfaces. This work introduces the PLS to handle the lack of affinity regularization in the forest-based metric.

To begin with, the leaf node ℓ^* with the largest span is located as the starting leaf, and

$$\ell^* = \underset{x_i, x_j \in \ell}{\operatorname{argmax}}\, g\left(x_i, x_j\right). \tag{7}$$

The span of starting node ℓ^* is denoted as $\eta^* = \max_{x_i, x_j \in \ell^*} g\left(x_i, x_j\right)$. Generally speaking, the leaf of extreme points can be identified in this way, e.g. the leaf node of the tiptoe in 3D human mesh dataset. The mixed Gaussian model (GMM) is used to fit the point distribution in the leaf node. For simplicity, the dominant mode acquired by the mean shift method [16] is used to represent the leaf. Let μ_ℓ^* denote the center of the dominant mode in ℓ^*. Point $x^* \in \ell^*$ is selected as the seed when

$$x^* = J(X) = \arg\min_{x \in \ell^*} \|x - \mu_\ell^*\|. \tag{8}$$

$J(X)$ returns the seed point of dataset X. The point set belonging to X and ℓ^* is split according to the seed selection. In our system, the seed point is assigned to the left leaflet. The binary test in leaf splitting is defined as

$$\varphi^*(x) = \begin{cases} 1, & \text{if } g(x, x^*) < 0.5\eta^*, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Given the starting leaf node and the seed selection, the leaf splitting is propagated to other leaves. The unprocessed leaves are sorted according to the distance to the seed point $x^* \in \ell^*$. The propagation begins from the nearest leaf node of ℓ^*. Let ℓ_k be the current leaf node. As to point $x \in \ell_k$, the binary test in leaf splitting of dataset X is defined as

$$\varphi_k(x) = \begin{cases} 1, & \text{if } g(x, x^*) \leq 0.5(\eta_{k1} + \eta_{k2}), \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

where $\eta_{k1} = \min_{x \in \ell_k} g(x, x^*)$, and $\eta_{k2} = \max_{x \in \ell_k} g(x, x^*)$. Only the leaf node with the ambiguous correspondence needs to be split, which can be determined simply by checking the span of the leaf node. In case that the span is greater than the predefined threshold, i.e. 10% of the largest span of dataset X in our experiments, the leaf nodes are split. The process of the pseudo-leaf-splitting is shown in Algorithm 1.

The PLS is a general technique to regularise the pairwise affinity obtained by the forest. Here the function g is used to measure the point-wise relationship between points inside a dataset, where the leaflet splitting tests are set according to the span of the dataset. There are no requirements that two sets share the same span when using the forest-based metric and the PLS regularization to establish the point-wise correspondence. The proposed scheme can handle the non-isometrically deformed datasets by using the data-dependent binary tests in Eq. (9) and Eq. (10).

Algorithm 1 Pseudo Leaf Splitting

```
Input: Random forest $RF$, dataset $X$.
Output: Pseudo leaf splitting.
for Each tree in $RF$ do
    Locate starting leaf $\ell^*$ with the largest span (Eq. (7));
    Compute the centroid of the dominant mode in $\ell^*$;
    Get a seed point $x^* \in \ell^*$ (Eq. (8));
    Split leaf node $\ell^*$ as Eq. (9);
    Sort unprocessed leaves by the distance to $x^*$;
    for Each inconsistent leaf node do
        Leaf splitting as Eq. (10);
    end for
end for
```

It is computationally complex to find the consistent correspondence of the shape corpus. Existing techniques coped with the consistent correspondence in the shape corpus by minimizing the overall distortion using the dynamic programming [36], the positive semi-definite matrix decomposition [25], and the functional map networks [26]. The additional refinement is required for consistent correspondence when given the initial pairwise mapping. The gap between the point-wise correspondence of shapes and the consistency refinement can be avoided by taking into account the point distribution in the shape corpus. Different from the example-based classification forest for the shape correspondence [44], there is no need for labeled training data using the proposed forest-based metric.

The correspondence function between surface mesh X^p and X^q is denoted as $\tau_{pq}(x_i^p) = x_j^q$, where affinity $\mathcal{A}_{ij}^{pq} = \max_{x_{j*}^q \in X^q} \mathcal{A}_{ij*}^{pq}$. When given a group of surface meshes, the point-wise correspondence by the PLS is consistent and satisfies the cycle constraints. That is, when $\tau_{pq}(x_i^p) = x_j^q$ and $\tau_{qr}(x_j^q) = x_k^r$, $\tau_{pr}(x_i^p) = x_k^r$. It can be ascribed to the seed selection based on the Gaussian fitting of the dominant mode in ℓ^*. The mapping between starting seed points of X_p and X_q is $\tau_{pq}(x^{p*}) = J_q J_p^{-1}(x^{p*}) = x^{q*}$. It is obvious that the correspondence of seed points satisfies the cycle constraints, where $\tau_{pr}(x^{p*}) = J_r J_q^{-1} J_q J_p^{-1} = J_r J_p^{-1} = x^{r*}$. Taking into account the similarity propagation nature of the PLS, the point-wise correspondence satisfies the cycle constraints.

5 Experiments

Datasets and Metric. The proposed method is applied to the affinity estimation of the uttering datasets, including the KinectVS [38], the OULUVS [56], and the OuluVS2 [4]. The KinectVS consists of twenty subjects uttering twenty phrases six times [38]. The color and depth video data are obtained by *Kinect* with a resolution of 640 × 480. The OULUVS dataset [56] consists of the color videos of twenty subjects uttering ten phrases five times with a resolution of 720 × 576. The OuluVS2 [4] consists of color videos of 53 subjects uttering ten phrases three times with a resolution of 1920 × 1080.

The AAM algorithm [17] is used to extract 35 patch trajectories around lips and jaws as [38], where the shape and texture features of patches are concatenated together to represent the trajectories. In our experiments, the affinity matrix obtained by the forest-based metric is sorted, and r nearest neighbors are viewed as matching candidates of probe trajectories. r is set at 1 (Top-1), 5 (Top-5), and 10 (Top-10) in the affinity evaluation. If the trajectory with the same label as the probe occurs in the candidate set, there is a hit. The trajectory labeling accuracy is computed as n_{hit}/n_{probe}, where n_{hit} and n_{probe} denote the numbers of hits and probe trajectories respectively.

We evaluate the proposed method on the 3D shape corpus, including TOSCA [11], Scape [3], SHREC07-NonSym [11, 25], and Faust datasets [6]. The wave kernel signature (WKS) [5] and the normalized geodesic distance vector are used as the feature descriptor of 3D points. The geodesic distance vector of point x is composed of the geodesic distance between x and all other points on the surface meshes by the fast marching algorithm. The correspondence accuracy of 3D surface meshes X and Y is defined as

$$e_{XY} = \frac{1}{n_X} \sum_{i=1}^{n_X} g(\tau(x_i), \tau'(x_i)), \qquad (11)$$

where τ and $\tau^{'}$ are the estimated and the ground truth point-wise mapping functions. n_X is the point number of X. g is the geodesic distance function. The percentages of correct matchings with a set of geodesic errors, including 0.02, 0.05, 0.10, and 0.16, are reported in our experiments.

5.1 Affinity Estimation

The proposed method is applied to affinity estimation on the facial trajectories and 3D points. We compare the proposed criteria with the classical Gini index [59], the determinant of the covariance matrix [18], and the variance of feature differences [55] on the facial trajectories (Fig. 6 (a, b, c)) and 3D shape datasets (Fig. 6 (e, f, g)). The node splitting criteria based on the determinant of the covariance matrix [18] fail in all datasets due to the rank deficiency of the covariance matrices. The forests built by the Gini index of the dummy set [10, 58, 59] depend on the construction of the synthetic data, being limited to locate the data clusters effectively. The node splitting criteria try to find a feature pair to produce the largest variance of feature difference [55], which do not model the data distribution of children nodes. On the other hand, our splitting criteria handle the data distribution and produce the best results with the Fuse metric. The tree numbers are set to 17 and 50 on the visual

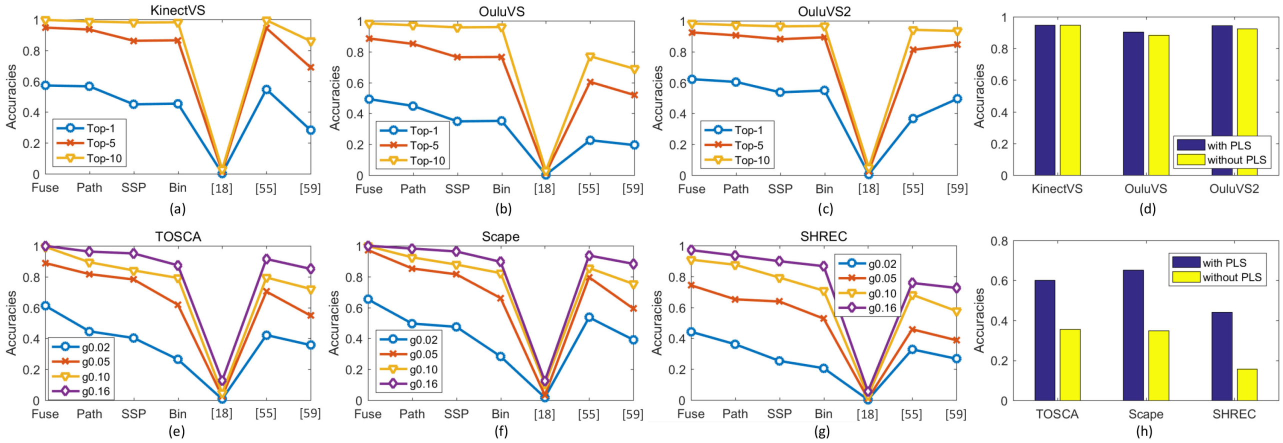

Fig. 6 Labeling accuracies by the combined random-forest-based metric (Fuse), the binary (Bin), the common-path-based (Path), and the SSP-based metrics, the random forests with node splitting criteria of the determinant of the covariance matrix [18], the variance of feature differences [55], and the Gini index [59] on (a) KinectVS, (b) OuluVS, (c) OuluVS2, (e) TOSCA, (f) Scape, and (g) Shrec-NonSym datasets. The Top-5 and g0.02 accuracies with and without the PLS on facial trajectories and 3D points are shown in (d) and (h) respectively.

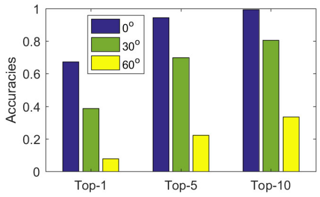

Fig. 7 Labeling accuracies of facial trajectories on OuluVS2 of different camera views including 0^o, 30^o, and 60^o.

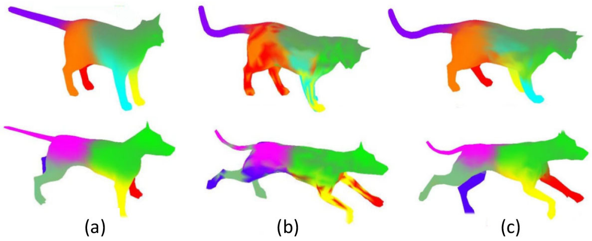

Fig. 8 The pairwise shape correspondence between (a) the reference and the target shapes (b) without and (c) with the PLS regularization.

uttering datasets and 3D shape datasets.

The comparison of different metrics, i.e., the binary (Bin), the common-path (Path), the SSP, and the combined distance metrics (Fuse), on the facial trajectories and 3D points are shown in Fig. 6 (a, b, c) and Fig. 6 (e, f, g). The Fuse metric shows apparently better performance than the binary one, and produces an improvement to the Path and the SSP-based metrics. For two pairs with the common paths of the same length, the one with the smaller SSP is more similar than the other. Both the Path and SSP metrics contribute to the affinity estimation based on the tree traversal in forests.

Fig. 6 (d) and Fig. 6 (h) show the labeling accuracies of the facial trajectories on the KinectVS, OuluVS, OuluVS2, as well as the 3D points matching accuracies on the TOSCA, Scape, and Shrec-NonSym datasets with and without the PLS regularization. The labeling performance based on the affinity estimation with the PLS regularization is better than the one without the PLS on all datasets. Because the shape feature defined as the difference of patch positions in adjacent frames possesses the motion information, the symmetric facial trajectories on the left and right half faces are less likely to be confused. Thus, the improvements with the PLS regularization in the facial trajectory datasets are limited compared with those in 3D shape datasets.

The facial tracker is designed for the frontal faces, and the tracking performance deteriorates when given profile facial images in the OuluVS2 phrase dataset [4]. Fig. 7 shows the effects of the facial landmark tracking on the affinity estimation of trajectories. The less accurate facial landmark tracking in the profile views makes it harder to locate the correct facial trajectories. The facial trajectory labeling accuracy of the frontal view is better than the profile views in the Top-1, Top-5, and Top-10 experiments.

5.2 Dense Correspondence Between Shapes

An unsupervised random forest-based metric with the PLS regularization scheme is employed to estimate the point distribution (Fig. 8). The comparisons of the pairwise correspondence by the proposed method with the functional maps (FM) [37], the blended intrinsic maps (BIM) [30], the coarse-to-fine combinatorial method [45], and the classification random forest (CRF) [44] are shown in Table 1. Similar to [44], we only conduct the experiments on the classes with more than six objects for the enough

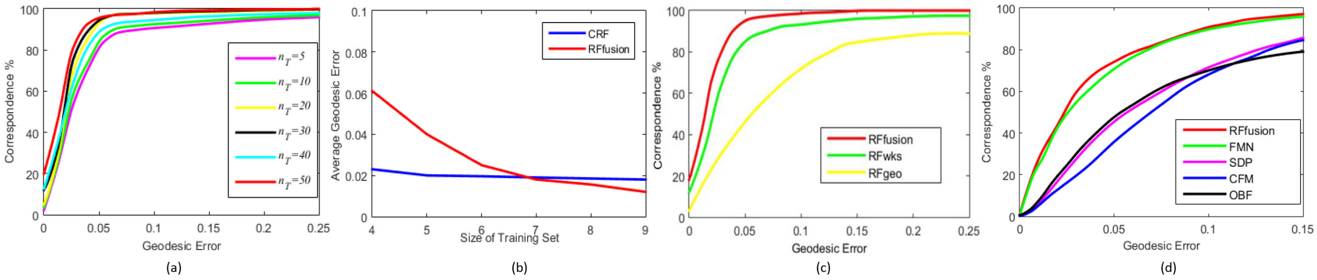

Fig. 9 (a) The correspondence errors with different forest sizes on the TOSCA dataset. (b) The average geodesic errors corresponding to various sizes of training sets of the proposed method (RF$_{fusion}$) and the classification random forest (CRF) [44]. (c) The comparison of pairwise correspondence errors with different feature channels on the human motion data [49]. (d) The comparison of consistent correspondence errors on the SHREC-NonSym dataset by the proposed method, the FMN [26], the SDP [25], the CFM [51], and the OBF [36] methods.

training data of the forest. Here all shapes except the query are used to train the forest. Our method can achieve more than 96% correct matchings within 0.05 geodesic errors. In experiments, the WKS and the geodesic distance vectors are used as the point descriptor. Table 1 illustrates the correspondence accuracy based on the WKS (RF$_{wks}$), the geodesic distance vector (RF$_{geo}$), and the feature fusion (RF$_{fusion}$). In our experiment, the accuracy of the dense correspondence by the RF$_{fusion}$ outperforms those using the RF$_{wks}$ or the RF$_{geo}$. The fusion of the local shape descriptor WKS and the contextual geodesic vector facilitates the searching for the optimal node splitting.

Tab. 1 Comparison of pairwise correspondences by the proposed methods with and without the PLS regularization, the combinatorial [45], the FM [37], the BIM [30], and the CRF [44] on the TOSCA dataset.

Methods	Correspondence (%)			
	g0.02	g0.05	g0.10	g0.16
Combinatorial [45]	24.8	56.0	80.8	90.5
BIM [30]	44.3	84.6	95.7	97.7
FM [37]	66.5	86.8	94.0	96.7
CRF [44]	65.6	94.5	99.1	99.2
RF$_{geo}$	21.9	46.3	71.7	84.2
RF$_{wks}$	44.8	84.1	93.1	96.2
RF$_{fusion}$	**67.3**	**96.5**	**99.4**	**100**
w/o PLS	35.6	63.3	72.5	79.8

The point-wise matching based on the forests with different numbers of trees is shown in Fig. 9 (a). The forest size is larger than that of the supervised CRF [44]. The relatively large number of randomized decision trees are needed to estimate the correspondence in an unsupervised manner. The more training data, the more accurate correspondence can be obtained (see Fig. 9 (b)).

We have applied the proposed method to the motion dataset [49], where the first 10% shapes are used to train the forest. There is no requirement that the training and testing shapes are from the same kind of motions. Our method can achieve more than 95% correct matchings within 0.05 geodesic errors as shown in Fig. 9 (c).

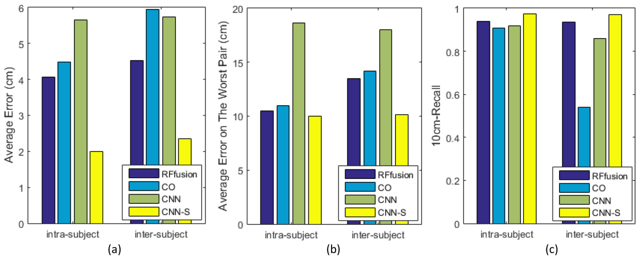

Fig. 10 The comparison in terms of (a) the average errors, (b) the average error on the worst pair, and (c) the 10 cm recall of the proposed method, the CO [13], the CNN [53], and the CNN-S [53] on the Faust dataset.

The proposed method is compared with the convex-optimization-based nonrigid registration (CO) [13] and the CNN classifier-based method [53] on the Faust database [6]. Similar to [13, 53], the correspondence is computed between pairs of meshes from the same subject (intra-subject) or different subjects (inter-subject). Aside from the testing pairs, all other meshes are used to build the random forest. Our method outperforms the CO and the CNN-based methods in the average error, the average error of the worst pair, and the 10-cm recall as shown in Fig. 10. The CNN followed by the non-rigid registration (CNN-S) produced the best results. However, the CNN and the CNN-s were built upon 2D depth maps, where the partial scans and additional registration operations were required.

Fig. 11 and Table 2 show the comparison with the deep learning-based shape correspondence models, including the 3D-coded [23], the FMNet [33], and the ADD3 [7] on the Scape dataset. The proposed forest-based metric with the PLS regularization outperforms the compared supervised and unsupervised deep learning-based models with the matching accuracy of 0.65 vs. 0.48 (3D-coded) and 0.27 (ADD3) regarding the g0.02. The supervised FMNet has

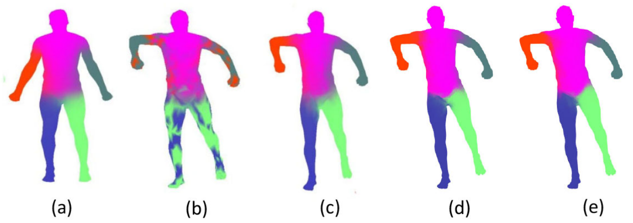

Fig. 11 The comparison with deep learning-based methods. (a) Reference. (b) and (c) are the proposed method without and with the PLS regularization. (d) 3D-coded [23]. (e) FMNet [33].

the best performance, which is learned from the prior ground truth correspondence and the mapping in both the spatial and spectral domains. On the other hand, the proposed approach only requires unsupervised forest-based metric learning for point-wise affinity.

Tab. 2 Comparison of the proposed method with deep learning-based shape correspondence methods on the Scape dataset.

Methods	3D-coded [23]	FMNet [33]	ADD3 [7]	Ours
g0.02	0.48	**0.78**	0.27	0.65

5.3 Consistent Correspondence in Shape Corpus

Aside from the pairwise shape correspondence, the proposed method is compared with existing consistent correspondence methods, including the positive semi-definite matrix decomposition (SDP) [25], the optimization-based framework (OBF) for the distortion minimization [36], the functional map network (FMN) [26], and the consistent functional maps (CFM) [51] on the SHREC-NonSym dataset as shown in Table 3 and Fig. 9 (d). The proposed method takes advantage of the point distribution modeling by the clustering forest and the PLS regularization scheme, outperforming the compared methods with the correspondence accuracies of 44.2 (g0.02) on the Shrec-NonSym dataset.

Tab. 3 Comparison of consistent correspondence by the proposed RF$_{fusion}$ with and without the PLS regularization, the FMN [26], the SDP [25], the CFM [51], and the OBF [36] methods on the Shrec-NonSym dataset.

Methods	Correspondence (%)			
	g0.02	g0.05	g0.10	g0.16
FMN [26]	42.7	70.9	89.8	95.8
SDP [25]	16.9	45.6	71.7	85.7
CFM [51]	12.8	36.0	68.3	84.5
OBF [36]	19.5	47.8	70.2	79.0
RF$_{fusion}$	**44.2**	**74.3**	**90.9**	**97.1**
w/o PLS	15.8	30.6	58.9	69.4

Table 4 illustrates the correspondence by the proposed method, the SDP [25], the OBF [36], and the fuzzy correspondences (FC) [29] on the TOSCA and Scape datasets. The proposed method outperforms the SDP [25] and the OBF [36] with significant margins in the local matching with 0.02 geodesic errors, which means the proposed method has an edge in the matching specificity. As to the 0.16 geodesic errors, the proposed method can realize the full matching as the SDP [25] and the OBF [36] methods.

Tab. 4 Comparison of the matching with 0.02 geodesic errors (g0.02) and 0.16 geodesic errors (g0.16) by the proposed RF$_{fusion}$ with and without the PLS regularization, the SDP [25], the OBF [36], and the FC (♯) [29] on the TOSCA and the Scape datasets.

	Error	SDP [25]	OBF [36]	RF$_{fusion}$	w/o PLS
TOSCA	g0.16	**100**	97.6	**100**	79.8
	g0.02	34.1	37.5	**60.2**	35.6
Scape	g0.16	**100**	**100**	**100**	77.3
	g0.02	41.2	48.6♯ [29]	**65.3**	34.8

As shown in Table 1, 3, and 4, the proposed forest-based metric with the PLS regularization refines the forest-based metric and produces an improvement with a large margin in both pairwise and consistent correspondence in the shape corpus.

5.4 Phrase Recognition

The phrase recognition accuracies of the proposed method (RF$_{fusion}$) on the depth and color videos are illustrated in Fig. 12. The accuracy of subject-independent (SI) experiments is lower than that of subject-dependent (SD) experiments. The performance variations in the SD and the SI experiments can be ascribed to personal speaking characteristics and person-specific texture differences regarding the mustache and the lip shapes. The SI experiments on the frontal phrase set of the OuluVS2 with an average accuracy of 84.8% are comparable to the state-of-the-arts [31, 57] (see Fig. 13 and Table 5). The system based on the deep neural networks produces a large margin improvement [14, 15], where a large number of parameters need to be learned from annotated training data.

Tab. 5 Phrase recognition accuracies on the OuluVS2 dataset.

Methods	Zhou[57]	Lee[31]	Chung [14]	Chung[15]	Ours
Accuracy	73.5	81.1	93.2	**94.1**	84.8

Fig. 14 illustrates the phrase recognition accuracies of each subject on the color videos (RF_{color}) with a patch size of 15 × 15 and the depth videos (RF_{depth}) with a patch size of 7 × 7 of the KinectVS dataset. We set the patch sizes of

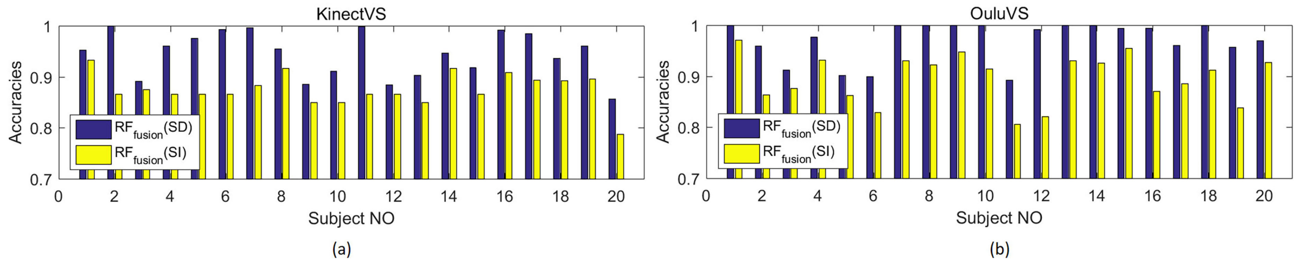

Fig. 12 Phrase recognition accuracies of each subject in (a) KinectVS and (b) OuluVS datasets by the RF$_{fusion}$ in the subjection-dependent (SD) and independent (SI) experiments.

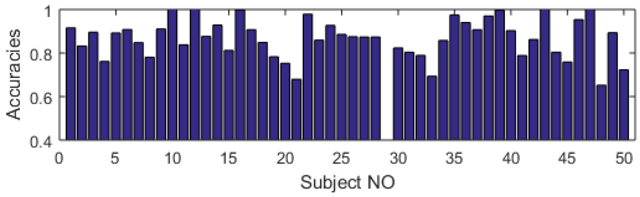

Fig. 13 Phrase recognition accuracies of each subject of the OuluVS2 phrase dataset.

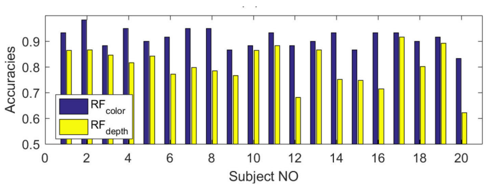

Fig. 14 Phrase recognition accuracies of each subject on the color videos (RF_{color}) with a patch size of 15×15, and the depth videos (RF_{depth}) with a patch size of 7×7 on the KinectVS dataset.

the color and depth videos as [38], where the patch size of depth videos is smaller than the color videos considering the relatively low signal-to-noise ratio of the depth video.

5.5 Comparison with Forest-based Correspondence

The proposed method utilizes the multivariate Gaussian distribution and the clustering forest-based metrics for affinity estimation and correspondence. We compare with the recent work on forest-based metrics, including the OCF [48], the MMRF [41], the SC forest [42], and the classification forest (CLA) [28], on supervoxel-wise correspondence as shown in Table 6. The dataset consists of 150 clinically obtained cone beam CTs (CBCT) [42], where each CBCT is decomposed into 5000 supervoxels. The proposed approach extends the binary forest-based metric to a continuous one, and achieves the Dice similarity coefficient (DSC) of 0.93 on the maxilla, outperforming the MMRF (0.88), the SC (0.89), and the CLA (0.81) using the binary metric. Here the OCF achieves the best performance with the DSC of 0.93 and 0.95 on the mandible and the maxilla. Note that the proposed approach does not introduce additional computational costs to forest construction, in contrast to additional dominant principal component estimations and regressions in the OCF [48].

Tab. 6 Comparisons on supervoxel-wise correspondence by forest-based methods.

	MMRF [41]	SC [42]	OCF [48]	CLA [28]	Ours
Mandible	0.91	0.92	**0.93**	0.88	0.88
Maxilla	0.88	0.89	**0.95**	0.81	0.93

6 Conclusions

We have presented the unsupervised random-forest-based metrics for the affinity estimation of the large and high-dimensional data, taking advantage of both the common traversing path and the smallest shared parent node. The proposed forest-based metric combined with the PLS is feasible to account for the spatial relationship for consistent correspondence. The proposed PLS scheme regularises the forest-based metric and avoids the gap between the the point-wise correspondence and additional consistency refinements inside a shape corpus. The proposed method is applied to phrase recognition using color and depth videos, as well as the point-wise correspondence of 3D shapes. The experiments demonstrate the effectiveness of the proposed method compared with the state-of-the-art.

References

[1] Y. Aflalo, A. Dubrovina, and R. Kimmel. Spectral generalized multi-dimensional scaling. *International Journal of Computer Vision*, 118:380–392, 2016.

[2] L. Alzubaidi, Z. Mohsin, and R. I. Hasan. Using random forest algorithm for clustering. *Journal of Engineering and Applied Sciences*, 13:9189–9193, 01 2018.

[3] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. *ACM Transactions Graphics*, 24(3):408–416, 2005.

[4] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen. Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*, volume 1, pages 1–5, 2015.

[5] M. Aubry, U. Schlickewei, and D. Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. In *IEEE International Conference on Computer Vision Workshops*, pages 1626–1633, 2011.

[6] F. Bogo, J. Romero, M. Loper, and M. Black. Faust: Dataset and evaluation for 3d mesh registration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3794–3801, 2014.

[7] D. Boscaini, J. Masci, E. Rodolà, M. Bronstein, and D. Cremers. Anisotropic diffusion descriptors. *Computer Graphics Forum*, 35, 2016.

[8] D. Boscaini, J. Masci, E. Rodolà, M. M. Bronstein, and D. Cremers. Anisotropic diffusion descriptors. *Computer Graphics Forum*, 35(2):431–441, 2016.

[9] D. Boyer, Y. Lipman, E. Clair, J. Puente, B. Patel, T. Funkhouser, J. Jernvall, and I. Daubechies. Algorithms to automatically quantify the geometric similarity of anatomical surfaces. *Proceedings of the National Academy of Sciences*, 108(45):18221–18226, 2011.

[10] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[11] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. *Numerical geometry of non-rigid shapes*. Springer Science & Business Media, 2008.

[12] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. *Computer Vision–ECCV 2010*, pages 282–295, 2010.

[13] Q. Chen and V. Koltun. Robust nonrigid registration by convex optimization. In *IEEE International Conference on Computer Vision*, pages 2039–2047, 2015.

[14] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103, 2016.

[15] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[16] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[17] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[18] 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, 2011.

[19] A. Criminisi and J. Shotton. *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media, 2013.

[20] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[21] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2188–2202, 2011.

[22] V. Ganapathi-Subramanian, O. Diamanti, and L. Guibas. Modular latent spaces for shape correspondences. *Computer Graphics Forum*, 37, 2018.

[23] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry. 3d-coded: 3d correspondences by deep deformation. In *European Conference on Computer Vision*, 2018.

[24] T. Hengl, M. Nussbaum, M. N. Wright, G. Heuvelink, and B. Gräler. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6, 2018.

[25] Q. Huang and L. Guibas. Consistent shape maps via semidefinite programming. *Computer Graphics Forum*, 32:177–186.

[26] Q. Huang, F. Wang, and L. Guibas. Functional map networks for analyzing and exploring large shape collections. *ACM Transactions on Graphics*, 33(4):36, 2014.

[27] M. Jeung, S.-S. Baek, J. Beom, K. Cho, Y. Her, and K.-S. Yoon. Evaluation of random forest and regression tree methods for estimation of mass first flush ratio in urban catchments. *Journal of Hydrology*, 575:1099–1110, 2019.

[28] F. Kanavati, T. Tong, K. Misawa, M. Fujiwara, K. Mori, D. Rueckert, and B. Glocker. Supervoxel classification forests for estimating pairwise image correspondences. *Pattern Recognition*, 63:561–569, 2017.

[29] V. G. Kim, W. Li, N. J. Mitra, S. Chaudhuri, S. DiVerdi, and T. Funkhouser. Learning part-based templates from large collections of 3d shapes. *ACM Transactions on Graphics*, 32(4):70, 2013.

[30] V. G. Kim, Y. Lipman, and T. Funkhouser. Blended intrinsic maps. *ACM Transactions on Graphics*, 30(4):79, 2011.

[31] D. Lee, J. Lee, and K.-E. Kim. Multi-view automatic lip-reading using neural network. In *ACCV Workshop on Multi-view Lip-reading Challenges*, 2016.

[32] Z. Li, F. Nie, X. Chang, Y. Yang, C. Zhang, and N. Sebe. Dynamic affinity graph construction for spectral clustering using multiple features. *IEEE Transactions on Neural Networks and Learning Systems*, 29:6323–6332, 2018.

[33] O. Litany, T. Remez, E. Rodolà, A. Bronstein, and M. Bronstein. Deep functional maps: Structured prediction for dense shape correspondence. *IEEE International Conference on Computer Vision*, pages 5660–5668, 2017.

[34] B. Liu, Y. Xia, and P. S. Yu. Clustering through decision tree construction. In *International Conference on Information Knowledge Management*, pages 20–29.

[35] F. Moosmann, B. Triggs, F. Jurie, et al. Fast discriminative visual codebooks using randomized clustering forests. In *Conference on Neural Information Processing Systems*, pages 985–992, 2006.

[36] A. Nguyen, M. Ben-Chen, K. Welnicka, Y. Ye, and L. Guibas. An optimization approach to improving collections of shape maps. *Computer Graphics Forum*, 30:1481–1491.

[37] M. Ovsjanikov, M. Ben-Chen, J. Solomon, A. Butscher, and L. Guibas. Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics*, 31(4):30, 2012.

[38] Y. Pei, T.-K. Kim, and H. Zha. Unsupervised random forest manifold alignment for lipreading. In *IEEE International Conference on Computer Vision*, pages 129–136, 2013.

[39] Y. Pei, L. Kou, and H. Zha. Anatomical structure similarity estimation by random forest. In *IEEE International*

Conference on Image Processing, 2016.

[40] Y. Pei, Y. Yi, G. Chen, T. Xu, H. Zha, and G. Ma. Voxel-wise correspondence of cone-beam computed tomography images by cascaded randomized forest. *IEEE International Symposium on Biomedical Imaging*, pages 481–484, 2017.

[41] Y. Pei, Y. Yi, G. Ma, Y. Guo, G. Chen, T. Xu, and H. Zha. Mixed metric random forest for dense correspondence of cone-beam computed tomography images. In *Medical image computing and computer-assisted intervention*, 2017.

[42] Y. Pei, Y. Yi, G. Ma, T.-K. Kim, Y. Guo, T. Xu, and H. Zha. Spatially consistent supervoxel correspondences of cone-beam computed tomography images. *IEEE Transactions on Medical Imaging*, 37:2310–2321, 2018.

[43] S. Rao, R. Tron, R. Vidal, and Y. Ma. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1832–1845, 2010.

[44] E. Rodola, S. Bulo, T. Windheuser, M. Vestner, and D. Cremers. Dense non-rigid shape correspondence using random forests. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4184, 2014.

[45] Y. Sahilliolu and Y. Yemez. Coarse-to-fine combinatorial matching for dense isometric shape correspondence. *Computer Graphics Forum*, 30(5):1461–1470, 2011.

[46] T. Shi and S. Horvath. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 2012.

[47] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, page 7, 2011.

[48] D. Sun, Y. Pei, Y. Guo, G. Ma, T. Xu, and H. Zha. Dense correspondence of cone-beam computed tomography images using oblique clustering forest. In *British Machine Vision Conference*, 2018.

[49] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics*, 27(3):97, 2008.

[50] M. Vrigkas, V. Karavasilis, C. Nikou, and I. Kakadiaris. Matching mixtures of trajectories for human action recognition. *European Conference on Computer Vision*, 19:27–40, 2014.

[51] F. Wang, Q. Huang, and L. Guibas. Image co-segmentation via consistent functional maps. In *IEEE International Conference on Computer Vision*, pages 849–856, 2013.

[52] W. Wang, D. Ceylan, R. Mech, and U. Neumann. 3dn: 3d deformation network. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1038–1046, 2019.

[53] L. Wei, Q. Huang, D. Ceylan, E. Vouga, and H. Li. Dense human body correspondences using convolutional networks. *arXiv preprint arXiv:1511.05904*, 2015.

[54] C. M. Yeilkanat. Spatio-temporal estimation of the daily cases of covid-19 in worldwide using random forest machine learning algorithm. *Chaos, Solitons, and Fractals*, 140:110210 – 110210, 2020.

[55] G. Yu, J. Yuan, and Z. Liu. Unsupervised random forest indexing for fast action search. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 865–872, 2011.

[56] G. Zhao, M. Barnard, and M. Pietikainen. Lipreading with local spatiotemporal descriptors. *IEEE Transactions Multimedia*, 11(7):1254–1265, 2009.

[57] Z. Zhou, X. Hong, G. Zhao, and M. Pietikäinen. A compact representation of visual speech data using latent variables. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1), 2014.

[58] X. Zhu, C. Loy, and S. Gong. Video synopsis by heterogeneous multi-source correlation. In *IEEE International Conference on Computer Vision*, pages 81–88, 2013.

[59] X. Zhu, C. Loy, and S. Gong. Constructing robust affinity graphs for spectral clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1450–1457, 2014.

A Decremental Covariance Matrix Evaluation

Since the covariance matrices need to be evaluated repeatedly when given the random selected parameters, it is time consuming to evaluate the covariance matrix σ from scratch for the optimal splitting parameters when building the forest. Let ρ be the data dimensionality, the time complexity of the covariance matrix construction is $O(\kappa \cdot \min(m_{T_l}^2\rho, m_{T_l}\rho^2) + \kappa \cdot \min(m_{T_r}^2\rho, m_{T_r}\rho^2))$ for κ randomly selected parameters. The complexity of the trace evaluation is $O(\kappa m_{T_l}\rho + \kappa m_{T_r}\rho)$. The decremental evaluation technique of covariance matrices is presented using the fact that the data in each node are a subset of the root node.

Let σ_p, σ_l, σ_r denote the covariance matrices of the parent and two children nodes respectively. The ij-th entry of σ_p is defined as $\sigma_{p_{ij}} = \mathbf{E}((t_i - \mu_p)(t_j - \mu_p)')$. Without losing generality, here the left child node is assumed to be larger than the right one. To begin with, the covariance matrix of the smaller child node, i.e. the right one, is computed. The entry of σ_r is defined as $\sigma_{r_{ij}} = \mathbf{E}((t_i - \mu_r)(t_j - \mu_r)')$. For a point pair (t_i, t_j) belonging to both the parent and the right child nodes, the differences of corresponding entries in σ_p and σ_r are computed as follows.

$$\widetilde{\sigma}_{p_{ij}} - \widetilde{\sigma}_{r_{ij}} = -(t_i + t_j)(\mu_p - \mu_r)' + \|\mu_p\|^2 - \|\mu_r\|^2, \quad (12)$$

where $\widetilde{\sigma}_{p_{ij}} = \sigma_{p_{ij}} \cdot (m_{T_p} - 1)$, and $\widetilde{\sigma}_{r_{ij}} = \sigma_{r_{ij}} \cdot (m_{T_r} - 1)$. Let σ_r^* denote the sub-matrix of σ_p with the columns and rows corresponding to points in the right child node.

The trace of the covariance matrix σ_r of the right child node is derived as

$$tr(\sigma_r) = \frac{tr(\sigma_r^*)(m_{T_p} - 1) + 2\sum_{i=1}^{m_{T_r}} t_i o_r' - m_{T_r}\mathfrak{o}_r}{m_{T_r} - 1}, \quad (13)$$

where o_r is the displacement vector from the centroid of the right child node to the parent, and $o_r = \mu_p - \mu_r$. The right child-related constant $\mathfrak{o}_r = \|\mu_p\|^2 - \|\mu_r\|^2$. Given $tr(\sigma_r)$, the trace of σ_l is computed as follows.

$$tr(\sigma_l) = \frac{tr(\sigma_p)(m_{T_p} - 1) - tr(\sigma_r)(m_{T_r} - 1) + \mathfrak{o}_l}{m_{T_l} - 1}, \tag{14}$$

where $\mathfrak{o}_l = m_{T_p}\|\mu_p\|^2 - m_{T_r}\|\mu_r\|^2 - m_{T_l}\|\mu_l\|^2$.

Once given the randomly selected splitting parameters, the centroids μ_l and μ_r of the left and right children nodes, as well as the norms $\|\mu_l\|$ and $\|\mu_r\|$ are computed. And then, the trace of the covariance matrix of the smaller child node, e.g. the right one, is computed based on the submatrix extracted from the parent node as Eq. (13). The trace of the covariance matrix of the other child node is computed when given $tr(\sigma_p)$ and $tr(\sigma_r)$ as Eq. (14). Since just the traces of the covariance matrices are needed to estimate the information gain in our system, the complexity of the covariance matrix evaluation is reduced from $O(m_{T_l}\rho + m_{T_r}\rho)$ to $O(\rho)$. When given κ randomly selected parameters, the trace evaluation complexity is reduced to $O(\kappa\rho)$.

B Proof of Proposition 1.

We will prove the functions defined in Eq. (3), Eq. (4), and Eq. (5) are metrics as follows.

Eq. (3). Let t_i, t_j, t_k be three input instances and the corresponding leaf nodes as $\ell(t_i), \ell(t_j), \ell(t_k)$. The common paths are denoted as $\mathbb{P}_{ij}$, $\mathbb{P}_{jk}$, and $\mathbb{P}_{ik}$.

Identity: $d_{cp}(t_i, t_i) = (|\mathbb{P}_{ii}|_o - |\mathbb{P}_{ii}|_o)/\nu_{ii} = 0$;

Positivity: Because $|\mathbb{P}_{ij}|_o \leq \nu_i$ and $|\mathbb{P}_{ij}|_o \leq \nu_j$, $|\mathbb{P}_{ij}|_o \leq \nu_{ij}$. Thus, $d_{cp}(t_i, t_j) = (\nu_{ij} - |\mathbb{P}_{ij}|_o)/\nu_{ij} \geq 0$;

Symmetry: $|\mathbb{P}_{ij}|_o = |\mathbb{P}_{ji}|_o$, so $d_{cp}(t_i, t_j) = d_{cp}(t_j, t_i)$;

Triangle inequality: Suppose that $\mathbb{P}_{ij}$ is the longest common path. Then $|\mathbb{P}_{ij}|_o \geq |\mathbb{P}_{ik}|_o$ and $|\mathbb{P}_{ij}|_o \geq |\mathbb{P}_{jk}|_o$. It follows that $|\mathbb{P}_{ik}|_o = |\mathbb{P}_{jk}|_o$ and $d_{cp}(t_j, t_k) = d_{cp}(t_k, t_i) \geq d_{cp}(t_i, t_j)$.

Thus, $d_{cp}(t_j, t_k) \leq d_{cp}(t_i, t_j) + d_{cp}(t_i, t_k)$, $d_{cp}(t_i, t_k) \leq d_{cp}(t_i, t_j) + d_{cp}(t_j, t_k)$, and $d_{cp}(t_i, t_j) \leq d_{cp}(t_i, t_k) + d_{cp}(t_j, t_k)$.

Similarly, when $\mathbb{P}_{jk}$ or $\mathbb{P}_{ik}$ is the longest common path, the triangle inequality property holds.

Eq. (4). *Identity*: $d_{sp}(t_i, t_i) = (|T_{p_{ii}}|_o - |T_{p_{ii}}|_o)/(|T_r|_o - |T_{p_{ii}}|_o) = 0$;

Positivity: Because $|T_{p_{ij}}|_o \geq \zeta_{ij}$ and $|T_r|_o \geq \zeta_{ij}$, $d_{sp}(t_i, t_j) \geq 0$;

Symmetry: $|T_{p_{ij}}|_o = |T_{p_{ji}}|_o$, so $d_{sp}(t_i, t_j) = d_{sp}(t_j, t_i)$;

Triangle inequality: Suppose that $T_{p_{ji}}$ is the smallest shared parent node. It follows that $|T_{p_{ik}}|_o = |T_{p_{jk}}|_o$ and $d_{sp}(t_j, t_k) = d_{sp}(t_k, t_i) \geq d_{sp}(t_i, t_j)$.

Thus, $d_{sp}(t_j, t_k) \leq d_{sp}(t_i, t_j) + d_{sp}(t_i, t_k)$, $d_{sp}(t_i, t_k) \leq d_{sp}(t_i, t_j) + d_{sp}(t_j, t_k)$, and $d_{sp}(t_i, t_j) \leq d_{sp}(t_i, t_k) + d_{sp}(t_j, t_k)$.

Similarly, when $T_{p_{jk}}$ or $T_{p_{ik}}$ is the smallest shared parent node, the triangle inequality property holds.

Eq. (5). Since the function is a weighted combination of two metrics as defined in Eq. (3) and Eq. (4), it is obvious that the function defined in (5) is a metric.