IET Computer Vision

Research Article

IET Journals
The Institution of Engineering and Technology

Towards path-based semantic dissimilarity estimation for scene representation using bottleneck analysis

ISSN 1751-9632
Received on 4th September 2018
Revised 13th February 2019
Accepted on 17th May 2019
E-First on 4th December 2019
doi: 10.1049/iet-cvi.2018.5560
www.ietdl.org

Lijuan Xu[1], Laura Dempere-Marco[2], Fan Wang[1], Zhihang Ji[1], Xiaopeng Hu[1] ✉

[1]*School of Computer Science and Technology, Dalian University of Technology, Dalian, 116024, People's Republic of China*
[2]*Department of Engineering, Universitat de Vic-Unviersitat Central de Catalunya, Vic (Barcelona), 08500, Spain*
✉ *E-mail: xphu@dlut.edu.cn*

Abstract: In natural images, it remains challenging to estimate dissimilarities between image elements for scene representation due to gradual variations of illuminations, textures or clutters. To tackle this problem, we utilise a path-based bottleneck analysis method that captures the semantic information between image elements to measure the dissimilarity. By integrating both the spatial continuity and feature consistency into the understanding of the semantic information, we detect the bottlenecks on the proposed double-S path to define the bottleneck distance, which demonstrates a favourable capability of grouping image elements that follow a similar pattern and separating different ones. In the experiments, the method is proved to be robust to noises and invariant to changing illumination and arbitrary scales in natural images. Tests on some challenging datasets validate the advantage of applying the path-based bottleneck distance in image ranking and salient object detection.

1 Introduction

Scene representation is a well-studied problem in computer vision and has benefited various applications including background modelling [1], image segmentation [2, 3], image retrieval [4–6], data clustering [7, 8] and salient region detection [9–12]. The estimation of dissimilarity between two elements is a fundamental question to be considered for understanding the middle-level semantic information in natural scenes and has attracted tremendous attention in past few decades [6, 8, 13–19]. Although there have been successes with previous approaches, the discrimination of the within- and between-class dissimilarities in scene representation remains a challenge owing to factors including gradual variations of illumination and clutters.

There has been considerable research on the estimation of the similarity/dissimilarity between visual elements in the scene. In computer vision, the distance metric is one of the commonly applied methods [20]. The pairwise Euclidean distance or Manhattan distance has been widely adopted to measure the feature difference between different samples. However, it has been suggested that they can be noise-sensitive and may not be appropriate for data analysis since the above two-distance metrics are based on the assumption that the data distribution is Gaussian or Exponential from the perspective of the maximum likelihood [4, 21, 22]. Therefore, finding a suitable distance metric remains a problem when the distribution is irregular or even unknown.

The geodesic distance (GD) has been applied to capture the intrinsic geometry manifold of the underlying data, which is computed efficiently by finding shortest paths on a connected graph with edges connecting neighbouring objects [23]. Nevertheless, when gradual changes caused by the varying illumination or noises (outliers), the GD metric can be semantically misleading for separating two classes since small gaps are accumulated along the path within the same class [24]. Omer and Werman [18] then applied the bottleneck geodesic metric to measure the affinity between image elements. By comparing the robust histogram density differences between feature points to define the bottleneck value on the shortest path, they achieve better results than the GD in terms of the segmentation performance. However, the shortest path it relies on mainly focuses on the overall connectivity instead of the local smoothness along the path, which may result in the absent of the meaningful changing when gradual variation occurs. Fischer *et al.* [2] proposed a path-based dissimilarity measure (or similarity in [3]) to emphasise the intra-cluster connectedness property for data clustering, based on the observation that two objects which are assigned to the same cluster are either similar or there exists a chain of mediating objects which the two consecutive objects in the chain are similar. If two objects belong to the same cluster, the dissimilarity is defined as the largest edge cost on the minimal intra-cluster path connecting the two objects on a full graph. However, simply applying the largest edge cost to estimate the dissimilarity may miss some important spatial topological relations and feature variations of objects between and within clusters. Strand *et al.* [25] introduced the minimum barrier distance that constitutes the smallest interval value containing all nodes along the path, to measure the dissimilarity for a pair of nodes. In fact, the minimum barrier distance may exaggerate the small difference in uniformly textured regions since it only uses the values between the maximum and minimum.

In addition, some other metrics have been studied to estimate the similarity/dissimilarity between image elements in different applications. Observing the facts in clustering that spectral theory shows favourable performance in finding clusters of arbitrary shapes and structures, but fail to handle datasets with cluttered backgrounds, and the dominant sets can deal with the background noises well, but tend to favour compact groups, Zemene and Pelillo [8] propose the path-based dominant-set clustering method which applies the path-based similarity measure for structure simulation and dominant-set for noise control. For saliency detection, Guan *et al.* [26] consider that the two-stage strategy applied in the graph-based manifold ranking model ignores the correlation between background and foreground cues. Instead of it, they propose a one-stage method to simultaneously optimise the saliency of the background and foreground regions. Hu *et al.* [27] propose a deep level set learning network to achieve compact and uniform saliency maps with accurate boundaries by modelling the semantic properties from the deep network. Liu *et al.* [28] regard the saliency detection as the two-stage graph optimisation process from a coarse-to-fine perspective, where the first stage uses a weighted joint robust sparse representation model to provide a coarse level map, and the second stage integrates a new regionally spatial consistency with the traditional adjacently spatial consistency to refine the coarse saliency map, thus assuring the uniform saliency assignment in complex scenes. Zhang *et al.* [12] efficiently combine the Laplacian sparse subspace clustering and unified low-rank representation to extract large-size salient objects

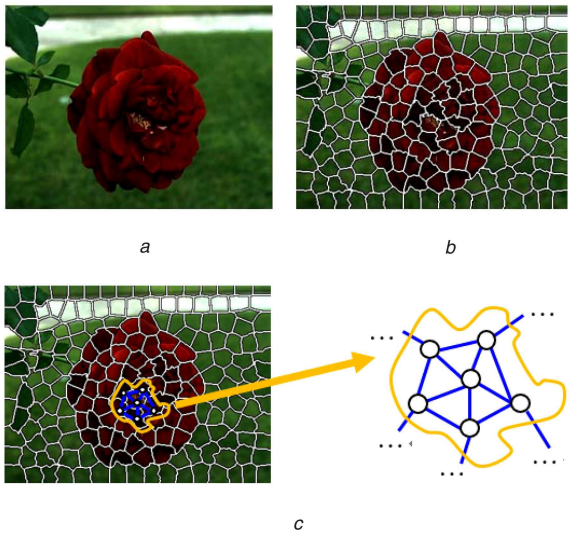

Fig. 1 *Illustration of constructing the graph representation from the image* ***(a)*** Is an input image. (b) The elements are superpixels generated by SLIC segmentation. (c) The construction of the undirected weighted graph is shown, where each superpixel is connected to its k nearest neighbours in the spatial domain, and the weights of connected edges are assigned to the original differences in feature space

in cluttered images, based on the relations (similar saliency values, representation coefficients and reconstruction errors) between spatially adjacent superpixels within the same cluster. In the field of the scene representation, Wan *et al.* [29] propose the fully convolutional representation learning model to use the structural and visual information (RLSV) in scene graphs via jointly structural and visual embedding for predicting new relations and completing scene graphs. Dai *et al.* [19] introduce an integrated deep relational network to tackle the problem of reasonably recognising the relationships among objects, and exploiting the statistical dependencies between them. Elhoseiny *et al.* [30] think that all facts of images such as objects, attributes, actions and interactions can benefit the uniform and generalisable understanding of the visual scenes. Based on this, they propose a structured way to simultaneously learn the visual facts in the image. In image retrieval, Lu *et al.* [31] propose a new deep hashing method for image retrieval, which utilise the deep network for training the target hash code generated from the relations between images of different contents. Observing the shortcomings of the squared Euclidean loss function applied by the vector quantisation techniques in efficient similarity search, Guo *et al.* [6] introduce a novel vector quantisation to enhance both the robustness and generalisation for similarity search. Wu *et al.* [32] propose a self-supervised deep multimodal hashing method for large-scale cross-media search by learning unified hash codes and deep hash functions, and in the meantime, a new discrete optimisation strategy-binary gradient descent to improve the efficiency of training.

This study proposes a path-based bottleneck analysis method, which depends on both the data manifold and the feature consistency, to measure the semantical dissimilarity between image elements of natural images. The method begins with constructing an undirected graph to simulate the intrinsic geometric relationships between different elements, followed by a path analysis to generate a double-S path between each pair of elements. The dissimilarity is then determined by the bottlenecks on the path and is capable of expressing the semantic correlations between image elements, which shows promising strength in minimising intra-class dissimilarities and maximising extra-class dissimilarities regardless of the scales and distributions of image regions. Different from some clustering-like methods where within-class dissimilarities are ideally neglected, the proposed approach is able to preserve certain important semantic information within class. The experimental results demonstrate the high accuracy and robustness of the proposed path-based bottleneck analysis in applications of image ranking and saliency estimation, and, especially, its capacity to handle images with gradual variations of illumination and clutters.

In conclusion, the main contributions of our method are summarised as follows: (1) The path-based bottleneck analysis method relying on both the data manifold and feature consistency is proposed to highlight the semantic relations between visual elements for image representation. (2) The double-S path is generated for each pair of elements in the image and we then apply the bottleneck detection approach to accumulate critical edge weights on the double-S path to provide a reliable distance measure for estimating the dissimilarities between visual elements. In addition, the experimental results demonstrate the performance of utilising the proposed scene representation for image ranking and salient region detection.

The rest of this paper is organised as follows: Section 2 presents the path-based bottleneck detection for semantic dissimilarity estimation. Section 3 exhibits the experiments on natural images with applications to image ranking and salient region detection. The conclusions are given in Section 4.

2 Proposed method

The emphasis of this section and the major contribution of this paper are the path generation and bottleneck detection to establish the semantic correlations between image elements. The proposed method extracts superpixels and uses them as the basic elements for image description and representation. The dissimilarity is then determined by the detected bottlenecks on the path and utilised for scene representation.

2.1 Graph construction

As shown in Fig. 1, an undirected weighted graph $G = (V, E)$ is constructed to describe relationships between image elements, where V is a set of nodes and E is a set of edges. For computational efficiency and perceptual representation, the nodes are visually homogeneous superpixels generated by applying the simple linear iterative clustering (SLIC) algorithm [33]. The emphasis of the graph construction is to determine which elements are neighbours and the weights of connected edges. In the study, each image element is connected to its k nearest neighbours in the spatial domain to preserve the local topology. The weight w_{ij} of each edge e_{ij} between connected neighbouring nodes is assigned to their feature difference in the feature space to represent the local consistency. In fact, the semantic information about spatial topology and feature difference between image elements is embedded in this representation and can be extracted by the proposed path-based bottleneck analysis.

To limit the computation complexity, we apply the k-regular graph where each node is connected to a maximum of k neighbours, and each node is denoted by a feature vector $F = \{f_1, f_2, \ldots, f_n\}$. In this study, F is defined as the average colour of pixels that belong to the superpixel in the CIE-Lab space, that is $F = \{L, a, b\}$. The corresponding weight w_{ij} for each edge e_{ij} that connects nodes i and j in the graph is calculated by the weight function analogous to the Gaussian kernel [2, 34].

$$w_{ij} = \exp\left(\| F_i - F_j \|^2 / 2\sigma^2\right) \tag{1}$$

2.2 Path-based bottleneck detection

In the scene analysis, a dissimilarity measure is required to capture the semantic relevance between image elements. It is preferable that it is small if the elements belong to the same class and large if not. In this study, we are concerned with two aspects about the semantic dissimilarity: (1) the spatial topology and (2) the feature consistency. The first aspect is coded in the weighted graph by connecting the spatially neighbouring objects. The second aspect specifies whether the appearances of image elements are similar in the feature space. The dissimilarity is then defined by a path-based bottleneck distance obtained via path generation and bottleneck detection to signify spatial continuity and feature consistency in the image.

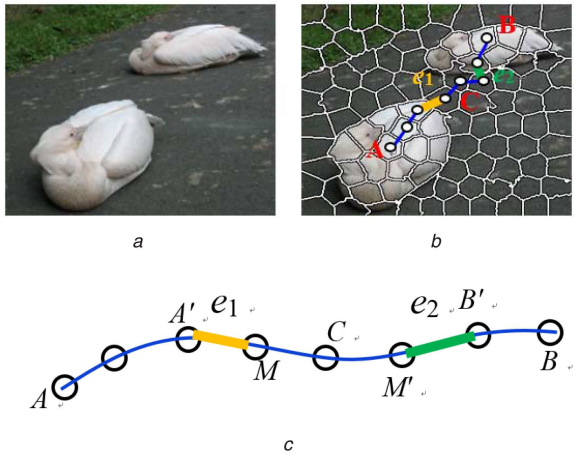

Fig. 2 *A, B and C are three nodes on a path, and node C is in-between A and B*

(a) Original image, ***(b)*** A path passing node A to B through node C, ***(c)*** The path between node A and B

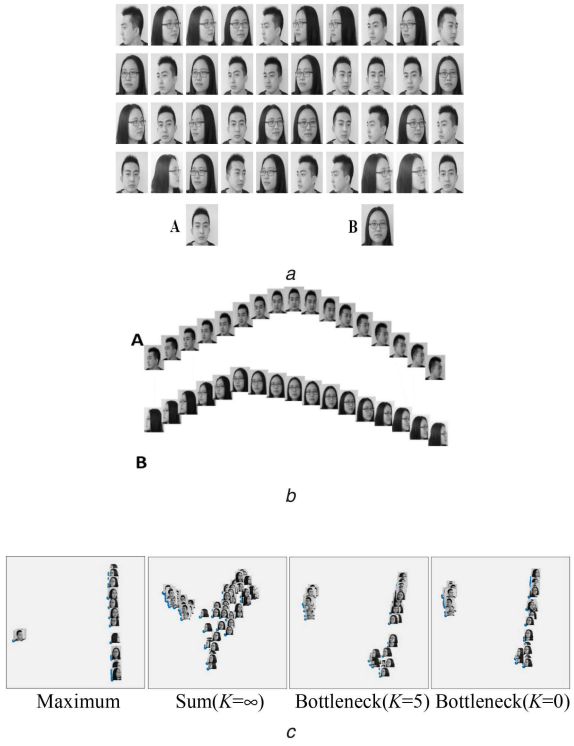

Fig. 3 *Illustration of image ranking*

(a) Shows photos of two people A and B, gradually turning their heads from left to right, ***(b)*** The ideal ranking results, ***(c)*** The comparisons among the maximum, sum and bottleneck method with varying K values

2.2.1 Bottleneck detection: As shown in Fig. 2*b*, although having similar features, nodes A and B should be treated as different objects since they belong to different image regions. Such a difference can be expressed by the dissimilarity measurement to preserve important spatial topological relations of image regions. To capture this semantic information between nodes in a graph, a dissimilarity measure defined on a path should satisfy (2).

$$\begin{cases} \mathrm{Dis}(A \to B) \geq \mathrm{Dis}(A \to C) \\ \mathrm{Dis}(A \to B) \geq \mathrm{Dis}(C \to B) \end{cases} \tag{2}$$

where node C is a mediating node on the path between nodes A and B. To estimate the dissimilarity between A and B, the maximum and sum of weights of path edges are two commonly used methods to define the path distance.

(1) *The maximum method:* In [2], Fischer *et al.* applied the edge with the largest weight on a path to measure the path-based dissimilarity. As demonstrated in Fig. 2*c*, e_1 and e_2 have the largest weights of edges on path $A \to C$ and $C \to B$, respectively. If we assume $w_{e_1} \geq w_{e_2}$, the maximum-of-weights is

$$\begin{aligned} \mathrm{Dis}_{\max}(A, B) &= \max(\mathrm{Dis}_{\max}(A, C), \mathrm{Dis}_{\max}(C, B)) \\ &= \max(w_{e_1}, w_{e_2}) \\ &= w_{e_1} \end{aligned} \tag{3}$$

It is evident that the maximum-of-weights is a special case that satisfies (2). However, such a distance is not able to provide detailed information about semantic relevance between image elements. That is, we cannot in turn infer the relationships between nodes just according to the distance itself. For instance in Fig. 2, given the values of $\mathrm{Dis}_{\max}(A, B)$, $\mathrm{Dis}_{\max}(A, C)$ and $\mathrm{Dis}_{\max}(C, B)$, we cannot tell if B or C is the middle node of the path. The reason is that we totally ignore the relations between B and C by treating $\mathrm{Dis}_{\max}(A, B) = \mathrm{Dis}_{\max}(A, C) = w_{e_1}$. In addition, as shown in (3), $\mathrm{Dis}_{\max}$ is strictly dependent on boundaries between image regions. As a result, vague boundaries of image regions may provide confusing information (small distances between clusters) for computing $\mathrm{Dis}_{\max}$.

(2) *The sum method:* $\mathrm{Dis}_{\mathrm{sum}}$, the sum of all weights on a path, is another widely applied method to measure the path distance. For example, the shortest distance is defined as the sum of all the weights on the shortest path.

$$\mathrm{Dis}_{\mathrm{sum}}(p) = \sum_{p} w_{p[h]p[h+1]} \tag{4}$$

where $p[h]$ denotes the hth node on the path p. The sum-of-weights also satisfies (2). However, it is liable to accumulate a large number of minor weights that are inside a single image region, resulting in misleading semantic information (large distances within class). The example of such a problem will be illustrated as follows.

Fig. 3*a* shows photos of two people A and B gradually turning their heads from left to right. Our goal is to rank all these images like the pattern shown in Fig. 3*b*. That is, the dissimilarities between the head photos of the one person should be small even if the images are taken from different views. So that

$$\begin{aligned} &\forall a, a' \in A, \forall b, b' \in B \\ &\begin{cases} \mathrm{Dis}(a, a') \leq \mathrm{Dis}(a, b) \\ \mathrm{Dis}(b, b') \leq \mathrm{Dis}(a, b) \end{cases} \end{aligned} \tag{5}$$

(5) indicates that the distance between head images should be smaller if they are taken from one person and larger otherwise, preferably for image clustering and segmentation. However, if we utilise the sum strategy to measure the dissimilarity in this case, it is probable that it will accumulate the slight variations between neighbouring images, and lead to incorrectly large dissimilarities between images taken from the one person.

(3) *The bottleneck method*: To overcome the aforementioned problems of the maximum and sum methods, we introduce the bottleneck distance that accumulates necessary information on a path to provide a reliable distance measure. Given a path p with N edges in a weighted graph, its bottlenecks [35] are defined as the edges that connect adjacent semantically-different image regions. As a result, we have a bottleneck set $\mathrm{BS}\{b(1), b(2), b(3), \ldots, b(m)\}$ and a non-bottleneck set $\mathrm{NBS}\{nb(1), nb(2), nb(3), \ldots, nb(n)\}$, where $m + n = N$. Thus, the path is partitioned into $m + 1$ sub-paths $\{p_1, p_2, p_3 \ldots p_{m+1}\}$ by the bottlenecks. Each sub-path p_i has a number of k_i non-bottlenecks with weights $\{w_{nb(p_{i,1})}, w_{nb(p_{i,2})}, w_{nb(p_{i,3})}, \ldots, w_{nb(p_{i,k_i})}\}$ sorted in a descending order, where $k_1 + k_2 + k_3 + \cdots + k_i = n$ and $1 \leq i \leq m + 1$. The $\mathrm{Dis}_{\mathrm{bottleneck}}$ (bottleneck distance) of the path p is defined as

$$\mathrm{Dis}_{\mathrm{bottleneck}}(p) = \sum_{i=1}^{m} w_{b(i)} + \sum_{i=1}^{m+1} \Big(\sum_{j=1}^{k_i} f(j) * w_{nb(p_{i,j})} \Big) \qquad (6)$$
$$f(j) = \mathrm{e}^{(-\alpha/\min(K,k_i))*j+\beta}$$

The first term in (6) is the accumulated weights of the bottleneck edges we have detected, reflecting the sharp or distinct changes between image elements along the path. In the second term, we use the function $f(j)$ to indicate the limitation of non-bottlenecks in their contribution to the bottleneck distance, *i.e.* the effective number of non-bottlenecks involved in the estimation of the bottleneck distance cannot exceed a fixed threshold K. If $K = 0$, $\min(K, k_i) \to \infty$, $f(j) \to 0$, which means that all non-bottlenecks are neglected. Taking the path in Fig. 2 for example, it is obvious that e_1 and e_2 are the two-detected bottlenecks to partition the path $A \to B$ into three sub-paths $(A \to A', M \to M', B' \to B)$. First, we add the weights of the bottleneck edges to reflect the sharp variations. Then for the sub-path $M \to M'$ belonging to the ground region, if we detect there are small and gradual changes between consecutive edges, the usage of the $f(j)$ function can help weight some meaningful non-bottlenecks for preserving some within-class variations. As shown in Fig. 3*c*, the experimental results indicate that the number of non-bottlenecks involved in the estimation of the bottleneck distance can benefit representing the relationships between visual elements when there exist gradual changes.

In this study, the bottleneck detection is achieved based on the following facts:

(i) Since the bottlenecks of a path are edges that connect neighbouring semantically-different image segments, their weights are generally larger than those of non-bottlenecks and can be regarded as the outliers in robust statistics. Table 1 demonstrates the designed algorithm for bottleneck detection on a path by applying the traditional outlier detection method.
(ii) In a path, there are usually fewer bottlenecks than non-bottlenecks.
(iii) If (C, D) is a detected bottleneck edge on a path between A and B, the calculation of $\mathrm{Dis}_{\mathrm{bottleneck}}(A, B)$ can be decomposed into computing $\mathrm{Dis}_{\mathrm{bottleneck}}(A, C)$ of the left sub-path and $\mathrm{Dis}_{\mathrm{bottleneck}}(D, B)$ of the right sub-path, as indicated in Fig. 4. We have

$$\mathrm{Dis}_{\mathrm{bottleneck}}(A, B) = w_{C,D} + \mathrm{Dis}_{\mathrm{bottleneck}}(A, C) + \mathrm{Dis}_{\mathrm{bottleneck}}(D, B) \qquad (7)$$

(iv) Specially, the maximum and sum approaches can be regarded as two-special cases of the bottleneck method if we use the edge with the largest weight or all edges as the bottlenecks, respectively.

2.2.2 Path generation: For the gradually changing of illuminations or clutters in natural images, the path generation for dissimilarity estimation in scene representation should be capable of imitating this kind of smoothness pattern to optimise the connection cost that highlights important contextual relevance between image elements. According to the proximity, similarity and continuity principles of Gestalt laws of grouping in human vision, people tend to group elements together if they are close, similar or connected (or continuous) to each other [36–38]. As a result, sharp abruption should be avoided in the optimisation process for path generation.

To capture the smoothness pattern between image elements by applying the above principles, we propose the smoothest path on the weighted-neighbouring graph, inspired by the minimal intra-cluster path [2, 3]. The minimal intra-cluster path emphasises the intra-cluster connectedness property and is defined in a full graph where every object is connected with every other objects in the same cluster. To avoid large steps or sharp abruptions on a path, we need to minimise the largest step or weight in the path, *i.e.*

$$P_{\mathrm{smoothest}}(A, B) = \min_{p \subset P_{A \to B}} \{ \max_{p[h], p[h+1] \in p} \{ w_{p[h]p[h+1]} \} \} \qquad (8)$$

Table 1 Algorithm for bottleneck detection

Input: a path p on the weighted graph

(1) Sort the edges on the path p in descending order, then compute the maximum weight max-$w(p)$ and median weight med-$w(p)$ from all weights w_i of the edges on p.

(2) ave-$w(p)$ = average(w_i): for all $w_i < \lambda$*med-$w(p)$.

(3) If max-$w(p)$ > δ*ave-$w(p)$, then a bottleneck is detected, go to (4); else there is no bottleneck on p, end.

(4) Use the detected bottleneck to split p into the left and right sub-paths. Go to (1) to detect bottlenecks for these sub-paths.

Output: bottlenecks

Note: λ and δ are set to 3.0 and 2.5, respectively, in this study.

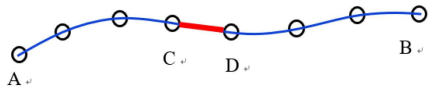

Fig. 4 *Recursive algorithm of the bottleneck detection on a path*

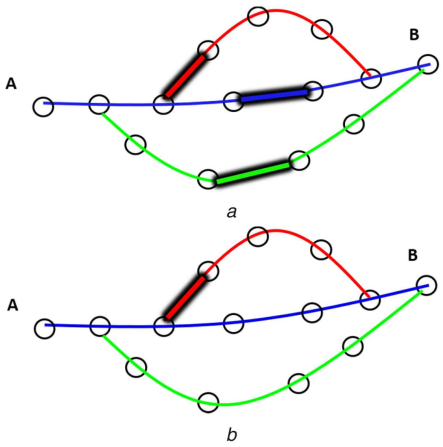

Fig. 5 *There exist three possible paths between node A and B, and signified by blue, red and green, respectively*
(a) Edges that have the largest weight for each path are highlighted by bold lines, ***(b)*** The red path is selected to be the smoothest path because it has the minimal weight among three highlighted edges

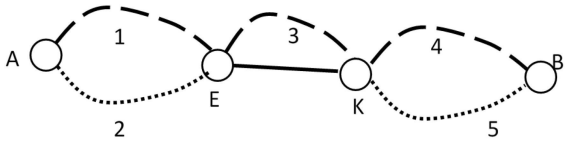

Fig. 6 *T is a MST of graph G where P(A…1…E…3…K…4…B) is the path for nodes A and B in T. We assume that there exists another smoothest path P'(A…2…e(E–K)…5…B) for nodes A and B, e(E–K) is an edge in P' but does not belong to T*

where $P_{A \to B}$ denotes the set of all paths connecting nodes A and B in the graph, $p[h]$ denotes the hth node on the path p from A to B.

Fig. 5 demonstrates an example of identifying the smoothest path. Given the high-computation load, it is impossible to traverse the whole graph for finding the smoothest path that connects each pair of pixels in the image. In this case, how to fast and accurately generate these smoothest paths remains a big challenge [39]. We have proved that the path on a minimum spanning tree (MST) of the graph is a smoothest path in Theorem 1, thus, the problem of traversing all paths to get the smoothest path can be solved by extracting the MST for the graph. The time complexity of the fastest implementation of Prim's algorithm by using a Fibonacci heap is $O(|E| + |V|\log|V|)$.

Theorem 1: A path in a MST of a weighted graph is a smoothest path.

Proof: If e(E–K) is added into T, the path (e(E–K)...3...E) forms a loop in T (Fig 6).

(i) Since e(E–K) does not belong to T, e(E–K) has the largest weight in loop. Otherwise, a new MST T' for graph G can be obtained by deleting the edge with the largest weight in path 3 and the weight of T' is smaller than that of T. It contradicts the fact that T is a MST for graph G.
(ii) Since e(E–K) is the largest weight edge in loop, path P' is not smoother than path (A…2…E…3…K…5…B).
(iii) Repeat (2) to replace all edges which do not belong to T in path P' until the path P for node A and node B in T is obtained. According to (2), path P' is not smoother than path P.□

For illustration, we take the mountain climbing for example to show the generation of the smoothest path between A and B by (8). In addition, the shortest path is adopted here for comparison. However, unlike the smoothest path, identifying the shortest path between two objects A and B is the problem of minimising the sum of the weights of edges regardless how large each step or weight is, leading large image gradients along the path, as shown in Fig. 7*b*.

Since the smoothest path seems to show satisfactory strength in simulating the phenomena of gradual changes in natural images. However, it can be excessively long by following the pattern of smoothness immoderately. In this case, the shortest path could take short cuts by passing through necessary boundaries of image regions, resulting in valuable information for the aforementioned bottleneck detection. As shown in Fig. 8, due to the small colour variations inside the flower region, the generated smoothest path between A and B is too long inside the region, while the shortest path is more meaningful.

To overcome the problem of the excessive long path demonstrated in Fig. 8, we need to control the length of the smoothest path. In this study, we integrate the shortest path into the smoothest path to reduce the path length, and generate a double-S path (smooth and short path), that is

$$\mathrm{Dis}_{\mathrm{sum}}(P_{\mathrm{double-S}}) \leq (1+\alpha)\mathrm{Dis}_{\mathrm{sum}}(P_{\mathrm{shortest}}), \alpha > 0 \tag{9}$$

Accordingly, the double-S path between A and B is achieved via an optimisation process, *i.e.*

$$P_{\mathrm{double-S}}(A,B) = \min_p\{\max_{p[h],\,p[h+1]\in p}\{w_{p[h]p[h+1]}\}\}$$
$$\text{for all } p \subset P_{A\to B} \wedge \mathrm{Dis}_{\mathrm{sum}}(p) \leq (1+\alpha)\mathrm{Dis}_{\mathrm{sum}}(P_{\mathrm{shortest}}), \alpha > 0 \tag{10}$$

Since the computational cost of applying (10) to generate a double-S path is extremely high, it is necessary to design a quick approximation method. In this study, the double-S path is generated by an iteration to replace the sub-paths of the smoothest path with their corresponding shortest sub-paths until (10) is satisfied, which is illustrated in Fig. 9. Table 2 presents the pseudo code of the designed algorithm. It first generates the smoothest and shortest paths for the two nodes, respectively, followed by an iteration to replace the sub-paths if the smoothest path is longer than a threshold. In fact, the experiments demonstrate that the smoothest path and the shortest path for every pair of nodes in a graph are mostly the same. As a result, only a small number of sub-paths require to be replaced.

3 Experiments

In the experiments, we apply the bottleneck analysis on the proposed double-S path to estimate the dissimilarities for scene representation. To demonstrate the power for visual scene analysis, the applications of image ranking and salient object detection are introduced in this section.

3.1 Analysis

3.1.1 Parameters setting: In this study, there are four parameters required to be pre-defined, respectively, λ and δ for the bottleneck detection, K for the bottleneck distance calculation and α for the double-S path generation. The optimal values of these four parameters are determined by both the theoretical analysis and experimental results.

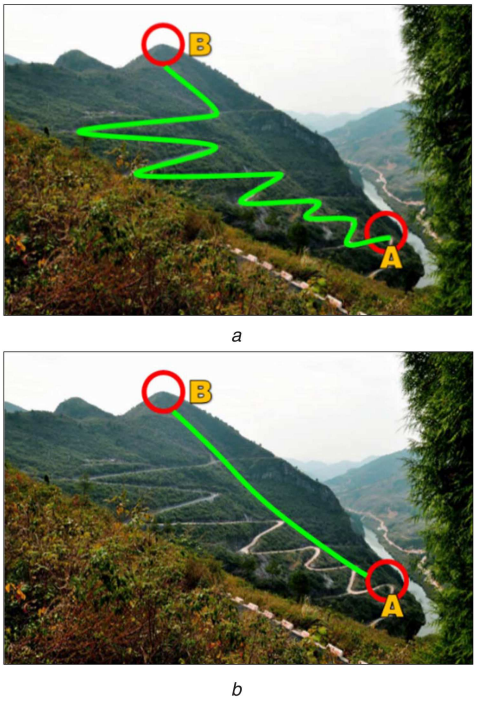

Fig. 7 *Mountain climbing*
(a) Smoothest path identified by minimising the largest step, ***(b)*** Shortest path with the shortest distance regardless how large each step is

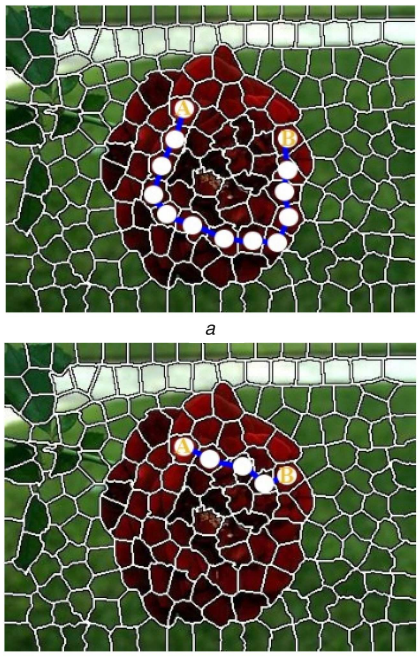

Fig. 8 *Illustration of difference between the smoothest and the shortest paths*
(a) Smoothest path, ***(b)*** Shortest path

λ and δ are two parameters applied to detect the bottlenecks on the path. Since the number of bottlenecks are much fewer than the non-bottlenecks, and the weights of bottlenecks are much larger than the non-bottlenecks on the path, so we regard the bottlenecks as the outliers in robust statistics obeying a normal distribution. For a normal distribution, the probability of data in the range $[-2\sigma, 2\sigma]$ is 95.44%, and in the range $[-3\sigma, 3\sigma]$ is 99.74%. Thus, the data which don't belong to this scope are outliers. To achieve a confidence interval larger than 95%, we set λ and δ to 3.0 and 2.5, respectively, in this study.

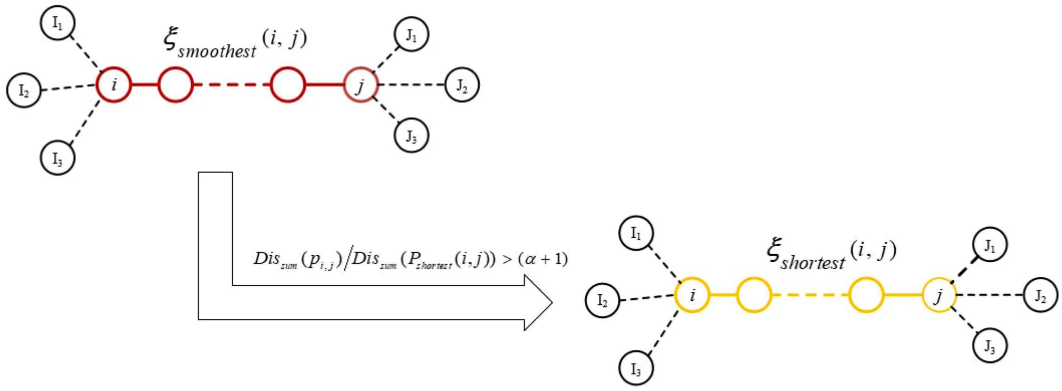

Fig. 9 *Illustration of generating a double-S path*

Table 2 Algorithm for generating a double-S path

Input: a weighted graph G and α
(1) Calculate the shortest and smoothest paths set, $P_{shortest}$ and $P_{smoothest}$ for a pair of nodes. (2) Calculate the sum-of-distance for the paths $P_{shortest}$ and $P_{smoothest}$, and get the corresponding values Sum($P_{shortest}$), Sum($P_{smoothest}$).
(3) Initialise $P_{double\text{-}S} = P_{smoothest}$, Sum($P_{double\text{-}S}$) = Sum($P_{smoothest}$).
(4) Set R = Sum($P_{double\text{-}S}$)/Sum($P_{shortest}$).
(5) If $R > (\alpha + 1)$ Replace the smoothest path with the corresponding shortest path. End
Output: the double-S path $P_{double\text{-}S}$.

Parameter K in (6) is a threshold to determine the contributions of the non-bottlenecks for estimating the bottleneck distance. If $K = 0$, it means only the detected bottlenecks are used to measure the bottleneck distance; If $K > 0$, it means some non-bottlenecks are added into the bottleneck distance when there exist gradual changes between consecutive image elements. The GD is a special case for incorporating all non-bottlenecks on the path. Fig. 3 shows the experimental results of setting different values for K. The final setting of the parameter is determined by the above analysis and experimental results. In this study, we set K to 5 for accumulating the weights of non-bottlenecks along a sub-path if there exist gradual changes between consecutive edges.

Parameter α in (10) is used to decide whether the length of the path is excessive long for a pair of nodes in the process of the double-S path generation. In this study, we use the sum of all edges' weights instead of the number of edges to approximate the length of the path. If the sum of all edges' weights on the path is much larger than that of the shortest path (the shortest distance has the minimal weights for all paths between each pair of nodes on the graph), we consider it is excessive long. Thus, we set $\alpha > 1$ and choose the value 3 in the study.

3.1.2 Complexity analysis: Here, we mainly discuss the complexity time of the bottleneck detection and double-S path algorithms.

For the reclusive algorithm of detecting all bottlenecks along a path with n edges, we first sort these edges in a descending order by the quicksort method with the average time $O(n\log n)$. To find the maximum and median edge values on the path, the time complexity is $O(1)$. Finally, the process of detecting the bottleneck on a sub-path can be calculated in $O(n)$ by using the computed-ranking index. Consequently, the overall computational cost is $O(n\log n)$ ($O(n\log n) + O(1) + O(n)$).

For the generation of the double-S path between a pair of nodes, we should build the MST and the shortest path by applying the Prim and Dijkstra first. With the use of the Fibonacci heap, the time complexity can be $O(|E| + |V|\log|V|)$. If the smoothest path has excessive long path length, it should be replaced by its corresponding shortest path, and the cost is $O(1)$. In summary, the time complexity of generating the double-S path is $O(|E| + |V|\log|V|)$ ($O(|E| + |V|\log|V|) + O(1)$).

3.2 Applications

3.2.1 Natural scene representation: To validate the performance of the bottleneck distance metric for scene representation, we apply some natural images with gradual changes of illuminations, scales or textures. For comparison, we choose four distance metrics, Euclidean distance, GD, $Dis_{bottleneck}$ on the smoothest path, and $Dis_{bottleneck}$ on the double-S path to verify the representation. In addition, the multi-dimensional scaling (MDS) technique is introduced to achieve the visual results. First, we use the algorithm presented in Section 2 to construct a weighted graph and estimate the four different dissimilarity matrixes between nodes on the graph, then the resulted matrixes are projected into a two-dimensional feature space by applying the MDS method. For each node (x, y) in the obtained feature space, we first normalise it into 0–255 and finally, use these normalised coordinates as the colour vector (i.e. $R = x$, $G = y$, and $B = 125$) to fill each element in the image. The image elements painted by similar colours mean close semantic relations estimated in the feature space.

Figs. 10*b–d* shows the visual results of using the Euclidean distance, GD and $Dis_{bottleneck}$ distance on the smoothest, respectively. The results indicate that these distances have a certain capability to depict the relations between image elements. However, each of them has its own drawbacks. Fig. 10*b* describes the two swans in the fifth row, two buildings in the seventh and two trees in the last row with similar colours, indicating that the spatial topological relations between image elements are neglected by the Euclidean distance. In Fig. 10*c*, a single image region such as the grass in the first row and sky in the third and seventh rows are painted by different colours. The problem arises from the accumulation of variations in the same image region. Fig. 10*d* shows the results of applying $Dis_{bottleneck}$ on the smoothest path. By definition, the smoothest path strictly avoids sharp gradients between image regions. As a result, if the boundary is not clear, $Dis_{bottleneck}$ can provide misleading information. For example, in the fourth row of Fig. 10*d*, the vague boundary of the dog's tail results in the mis-segmentation of the dog, and in the sixth row, the same water region is painted by different colours. The proposed $Dis_{bottleneck}$ distance on the double-S path achieves the best results on the representation of the relation and differences between image elements, leading to a meaningful representation of images. As shown in Fig. 10*e*, two swans have different colours, the grass and sky regions are painted uniformly in the same colour, the dog and the diver are correctly segmented, the similar buildings and trees are painted differently, while the varying sky regions are described with the same colour.

3.2.2 Image ranking: Image ranking, aims to automatically rank all the images based on their relationships hidden in the dataset without a source image. As shown in Fig. 3, our main goal is to rank the head photos from two persons effectively. That is, the semantic relevance between images of one person is much smaller

Fig. 10 *Comparison of the different distance metrics on natural images by using MDS*
(a) Original images, ***(b)*** Euclidean distance, ***(c)*** GD on the shortest path, ***(d)*** $\text{Dis}_{\text{bottleneck}}$ on the smoothest path, ***(e)*** $\text{Dis}_{\text{bottleneck}}$ on the double-S path with $\alpha = 3$

than that of different persons, which should satisfy (5). From the perspective of ranking, we hope that some necessary dissimilarities are preserved to differentiate the images of one person as they change a lot from left to right, so that

$$\forall\, x, y, z \in A \cup B \wedge \exists \text{path}\ \ p\ , \text{passing}\ \ \text{from}\, x\ \ \text{to}\ \ z\, \text{through}\ \ y \begin{cases} \text{Dis}(x, y) \le \text{Dis}(x, z) \\ \text{Dis}(y, z) \le \text{Dis}(x, z) \end{cases} \tag{11}$$

(11) shows that the differences between images are preserved to a certain degree, even if they belong to one person.

In this study, the 698 face images of one person with illumination and pose changes provided by Joshua *et al.* [23] are applied to show the performance of our semantic dissimilarity estimation in image ranking. Furthermore, we rank a new data set with three clusters by mixing the images from the above two sets together in the experiments. In addition, we also chose the Extend Yale B dataset which contains 2414 face images taken from 38 persons and each person owns 64 photos with varying gestures and illuminations for testing. For visual assessment, we randomly select 128 face images from two persons and 256 images from four persons. First, all images are down-sampled as 64*64 pixel-vectors, and each image is treated as a node to construct the k-regular graph. For visualisation of the ranking results, the MDS technique has been applied to construct a low-dimensional embedding of the given vectors while faithfully preserving the inter-vector distances in a two-dimensional feature space.

For comparison, we choose Euclidean distance, GD, $\text{Dis}_{\text{bottleneck}}$ on the smoothest path, and $\text{Dis}_{\text{bottleneck}}$ on the double-S path to demonstrate the performance of ranking the images within- and between-clusters. For the Euclidean distance (shown in Fig. 11*a*), it is obvious that the patterns of image changes (illumination, head angles etc.) are revealed clearly for the two-cluster sets. However, the images of multiple persons are not evidently separated into

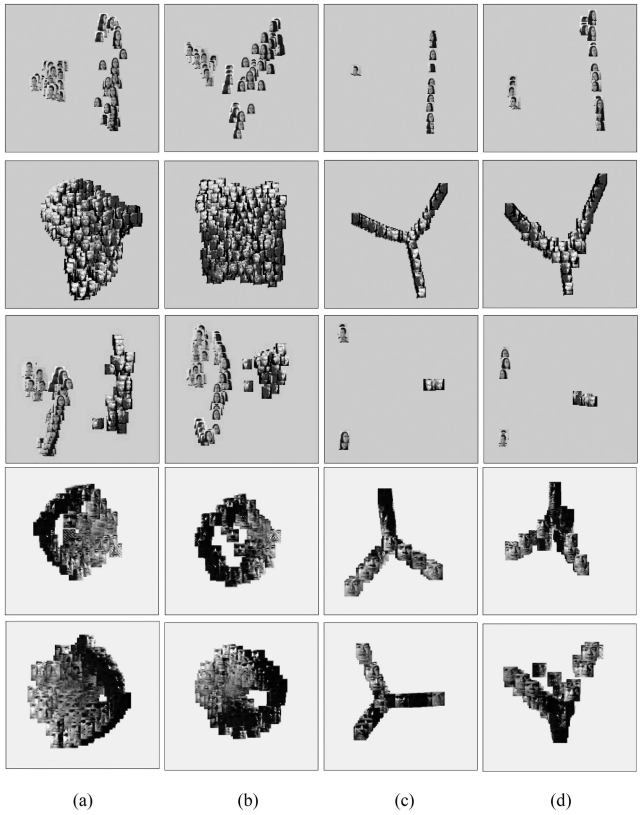

Fig. 11 *Comparison of the different distance metrics by MDS*
(a) Euclidean distance, ***(b)*** GD on the shortest path (Isomap [23]), ***(c)*** $\text{Dis}_{\text{bottleneck}}$ on the smoothest path, ***(d)*** $\text{Dis}_{\text{bottleneck}}$ on the double-S path with $\alpha = 3$

different clusters. The GD estimates the dissimilarity between images by calculating the sum of weights along the shortest path. It may be unreliable to use such a distance for image clustering since the accumulated weights or differences in the same cluster can be larger than that between different clusters, see Fig. 11*b*. Fig. 11*c* indicates that the $\text{Dis}_{\text{bottleneck}}$ on the smoothest path is capable of clustering, and photos are well grouped into clusters of different people. However, since images of one person are projected into the same position, the relations within a cluster are mostly neglected by this method. Similarly, Fig. 11*d* shows that the proposed bottleneck detection on the double-S path also has the capability of clustering. In addition, the experimental results validate the description of the patterns of image changes within each cluster, signifying the capacity of the proposed bottleneck method to highlight extra-class differences and, in the meantime, preserving important intra-class dissimilarities to a certain extent. In the Extend Yale B dataset, for face image ranking by the Euclidean distance and GD, it is difficult to tell the differences of different persons as the ranked images of different persons are mixed with each other. In the proposed method, we note that images with bright illumination and varying gestures are well separated, and images with dark illumination of different persons are mixed. The reason is that these images with dark lights are not able to provide enough information for differentiation.

3.2.3 Salient object detection: Salient region detection is closely related to the selective process in human vision [40] and aims to locate interesting regions or objects in images. Psychological and perceptual research has demonstrated that image contrast is the most influential factor in visual saliency [10, 41]. However, most of the computational models simply use the feature difference to measure the contrast and ignore the semantic relations between image elements [9, 11, 34].

Fig. 12 demonstrates the experimental results on six natural images, which contain round, square, elongated and V-shaped salient objects with non-linear scales. By visual assessment, the proposed path-based bottleneck distance segmentation of salient objects and uniformity inside salient regions, regardless of the

Fig. 12 *Saliency maps generated from different distance metrics*
(a), ***(b)*** Original and ground-truth images, ***(c)***, ***(d)***, ***(e)*** Estimated saliency obtained by using Euclidean, geodesic and bottleneck distance, respectively

gradual illumination variations and distribution of the salient regions.

For a more comprehensive assessment, the receiver operating characteristic (ROC) curve, precision-recall (PR) curve, mean absolute error (MAE) and F_β^w (weighted- F_β) measures are adopted for evaluating the performance of saliency detection. In this work, we introduce the minimum barrier distance based saliency model [42] (MBD), which is the latest distance-based method that outperforms all others on the benchmark datasets, to define the saliency. To demonstrate the effectiveness of our bottleneck distance in estimating the image contrast, we respectively replace the minimum barrier distance with our bottleneck distance and the commonly used GD in the MBD model. The resulted path-based bottleneck distance (PBD) model and GD are then performed on the traditional ASD dataset for comparison with the MBD. In the field of saliency detection, ASD is a widely used dataset that includes 1000 images with accurate human labelled segmentation masks, and is chosen for its wide application in testing almost all saliency models. Fig. 13 shows the comparisons of the three-distance metrics in terms of ROC and PR curves, MAE and F_β^w scores on the ASD dataset. The experimental results show that the proposed path-based distance achieves the best performance of the lowest error and the highest weighted- F_β measure score.

Since the ASD dataset tends to contain images with distinct objects surrounding by clean backgrounds, therefore, we choose the challenging datasets ECCSD and DUT_ORMON, which respectively contain 1000 and 5000 semantically meaningful and structurally complex natural images. In addition, some state-of-the-art approaches (GS_SP [24], GMR [43], RBD [44], MC [45], DSR [46]) are selected to verify the performance. Figs. 14 and Fig. 15 show the overall ROC curves, PR curves, MAE and F_β^w values obtained by using all images in the ECCSD and DUT_ORMON datasets. The experimental results illustrate the high performance of the proposed path-based bottleneck analysis for saliency estimation (PBS).

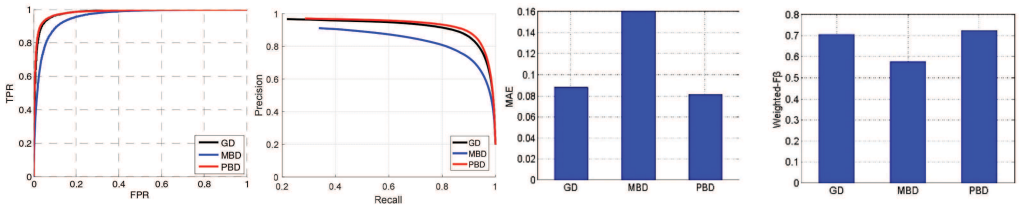

Fig. 13 *ROC curves, PR curves, MAE values and F_β^w scores obtained by applying the GD, MBD and PBD, respectively, on the ASD dataset*

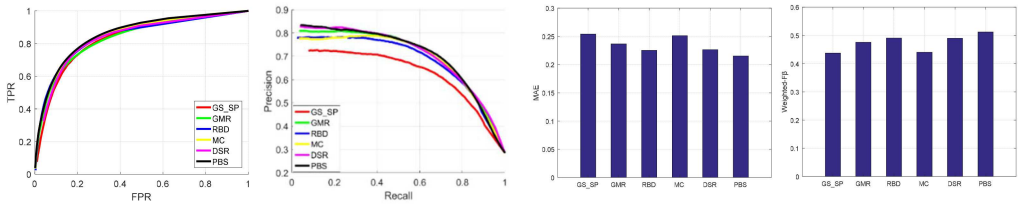

Fig. 14 *ROC curves, PR curves, MAE values and F_β^w scores obtained by applying the GS_SP, GMR, RBD, MC, DSR and PBS on challenging the ECCSD dataset*

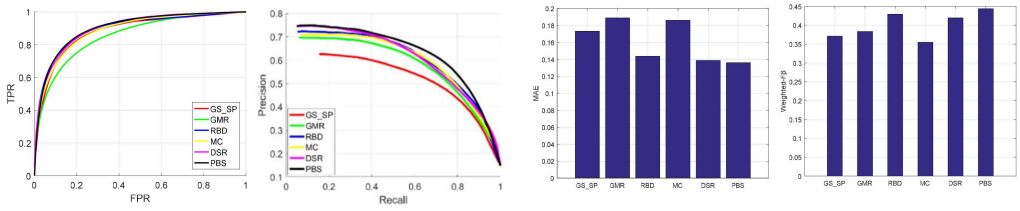

Fig. 15 *ROC curves, PR curves, MAE values and F_β^w scores obtained by applying the GS_SP, GMR, RBD, MC, DSR and PBS on challenging the DUT_ORMON dataset*

4 Conclusion

Estimating semantic dissimilarities between image elements for scene representation remains a challenge due to the high complexity and uncertainty in natural images and other factors including noises, image blurs or gradual variations of illumination and textures. It is preferable that the distance metric is able to integrate the appearance differences and spatial distribution for dissimilarity estimation. To solve the problem, we propose the path-based bottleneck analysis by detecting the bottlenecks on the double-S path. Our method expresses high performance of representing semantically-meaningful topological relations between image elements and, in the meantime, preserving their important dissimilarities. A recursive algorithm for robust bottleneck detection and an approximate algorithm for path generation are designed in this study. The experimental results demonstrate the strengths of the proposed method for scene representation in applications such as image ranking and saliency detection.

The main challenge of this research is to capture the intrinsic image patterns or relationships between image elements which are encoded in an undirected graph, and the corresponding decoding should be robust to the noises, changes of illumination and uncertainty in natural images. The method proposed in this study presents an opportunity to tackle these challenges in a systematic way and can be applied in applications such as background modelling, image coding and segmentation where grouping relevant image elements and extracting important topological information are necessary.

5 Acknowledgments

This research was supported by the 'National Natural Science Foundation of China' (No. 61272523, No. 61471084), 'the National Key Project of Science and Technology of China' (No. 2011ZX05039-003-4) and 'the Fundamental Research Funds for the Central Universities' (No. DUT15QY33).

6 References

[1] Barnich, O., Van Droogenbroeck, M.: 'Vibe: A universal background subtraction algorithm for video sequences', *IEEE Trans. Image Process.*, 2011, **20**, (6), pp. 1709–1724

[2] Fischer, B., Zöller, T., Buhmann, J.M.: 'Path based pairwise data clustering with application to texture segmentation'. IEEE Proc. in Computer Vision and Pattern Recognition, Kauai, USA, 2001, pp. 235–250

[3] Chang, H., Yeung, D.: 'Robust path-based spectral clustering with application to image segmentation'. Tenth IEEE Int. Conf. on Computer Vision, Beijing, China, 2005, vol. 1, pp. 278–285

[4] Yu, J., Tian, Q., Amores, J., *et al.*: 'Toward robust distance metric analysis for similarity estimation'. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, New York, USA, 2006, vol. 1, pp. 316–322

[5] Yu, J., Amores, J., Sebe, N., *et al.*: 'Distance learning for similarity estimation', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2008, **30**, (3), pp. 451–462

[6] Guo, Y., Ding, G., Han, J.: 'Robust quantization for general similarity search', *IEEE Trans. Image Process.*, 2018, **27**, (2), pp. 949–963

[7] Xing, E., Ng, A., Jordan, M., *et al.*: 'Distance metric learning with application to clustering with side-information', *Adv. Neural. Inf. Process. Syst.*, 2003, **15**, pp. 505–512

[8] Zemene, E., Pelillo, M.: 'Path-based dominant-set clustering'. Int. Conf. on Image Analysis and Processing, Genova, Italy, 2015, pp. 150–160

[9] Itti, L., Koch, C., Niebur, E.: 'A model of saliency-based visual attention for rapid scene analysis', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1998, **20**, pp. 1254–1259

[10] Derrick, P., Klinton, L., Ernst, N.: 'Modeling the role of salience in the allocation of overt visual attention', *Vis. Res.*, 2002, **42**, (1), pp. 107–123

[11] Yan, Q., Xu, L., Shi, J., *et al.*: 'Hierarchical saliency detection'. Proc. of Computer Vision and Pattern Recognition, Portland, USA, 2013, pp. 1155–1162

[12] Zhang, Q., Liu, Y., Zhu, S., *et al.*: 'Salient object detection based on super-pixel clustering and unified low-rank representation', *Comput. Vis. Image Underst.*, 2017, **161**, pp. 51–64

[13] Thureson, J., Carlsson, S.: 'Appearance based qualitative image description for object class recognition'. ECCV, Prague, Czech Republic, 2004, pp. 518–529

[14] Zheng, J., Tsuji, S.: 'Generating dynamic projection images for scene representation and understanding', *Comput. Vis. Image Underst.*, 1998, **72**, (3), pp. 237–256

[15] Kadir, T., Brady, M.: 'Saliency, scale and image description', *Int. J. Comput. Vis.*, 2001, **45**, (2), pp. 83–105

[16] Schneider, W.: 'Visual-spatial working memory, attention, and scene representation: A neuro-cognitive theory', *Psychol. Res.*, 1999, **62**, (2–3), pp. 220–236

[17] Lou, Y., Favaro, P., Soatto, S., *et al.*: 'Nonlocal similarity image filtering'. Image Analysis and Processing (ICIAP 2009), Vietri sul Mare, Italy, 2009, pp. 62–71

[18] Omer, I., Werman, M.: 'The bottleneck geodesic: computing pixel affinity'. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, New York, USA, 2006, pp. 1901–1907

[19] Dai, B., Zhang, Y., Lin, D.: 'Detecting visual relationships with deep relational networks'. IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, USA, 2017, pp. 3298–3308

[20] Yang, L., Jin, R.: 'Distance metric learning: A comprehensive survey', Michigan State University, 2006, 2

[21] Itti, L., Koch, C.: 'Feature combination strategies for saliency-based visual attention systems', *J. Electron. Imaging*, 2001, **10**, (1), pp. 161–169

[22] Zakai, M.: 'General distance criteria', *IEEE Trans. Inf. Theory*, 1964, **January**, pp. 94–95

[23] Tenenbaum, J., De Silva, V., Langford, J.: 'A global geometric framework for nonlinear dimensionality reduction', *Science*, 2000, **290**, (5500), pp. 2319–2323

[24] Wei, Y., Wen, F., Zhu, W., *et al.*: 'Geodesic saliency using background priors'. ECCV, Florence, Italy, 2012, pp. 29–42

[25] Strand, R., Ciesielski, K., Malmberg, F., *et al.*: 'The minimum barrier distance', *Comput. Vis. Image Underst.*, 2013, **117**, (4), pp. 429–437

[26] Guan, Y., Jiang, B., Xiao, Y., *et al.*: 'A new graph ranking model for image saliency detection problem'. IEEE 15th Int. Conf. on Software Engineering Research, Management and Applications (SERA), London, UK, 2017

[27] Hu, P., Shuai, B., Liu, J., *et al.*: 'Deep level sets for salient object detection'. CVPR, Honolulu, USA, 2017

[28] Liu, Y., Han, J., Zhang, Q., *et al.*: 'Salient object detection via two-stage graphs', *IEEE Trans. Circuits Syst. Video Technol.*, 2018, **29**, (4), pp. 1023–1037

[29] Wan, H., Luo, Y., Peng, B., *et al.*: 'Representation learning for scene graph completion via jointly structural and visual embedding'. IJCAI, Stockholm, Sweden, 2018, pp. 949–956

[30] Elhoseiny, M., Cohen, S., Chang, W., *et al.*: 'Sherlock: scalable fact learning in images'. AAAI, San Francisco, USA, 2017, pp. 4016–4024

[31] Lu, X., Song, L., Xie, R., *et al.*: 'Deep binary representation for efficient image retrieval', *Adv. Multimedia*, 2017, **2017**

[32] Wu, G., Han, J., Lin, Z., *et al.*: 'Joint image-text hashing for fast large-scale cross-media retrieval using self-supervised deep learning', *IEEE Trans. Ind. Electron.*, 2018, **66**, (12), pp. 9868–9877

[33] Achanta, R., Shaji, A., Smith, K., *et al.*: 'SLIC superpixels compared to state-of-the-art superpixel methods', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, **34**, pp. 2274–2282

[34] Borji, A., Cheng, M., Jiang, H., *et al.*: 'Salient object detection: A benchmark', *IEEE Trans. Image Process.*, 2015, **24**, (12), pp. 5706–5722

[35] Kaibel, V., Peinhardt, M.A.F.: 'On the bottleneck shortest path problem', Konrad-Zuse-Zentrum für Information stechnik, 2006

[36] Rubin, J., Kanwisher, N.: 'Topological perception: holes in an experiment', *Percept. Psychophys*, 1985, **37**, pp. 179–180

[37] Chen, K.: 'Adaptive smoothing via contextual and local discontinuities', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (10), pp. 1552–1567

[38] Chen, L.: 'The topological approach to perceptual organization', *Vis. Cognit.*, 2005, **12**, (4), pp. 553–637

[39] Vassilevska, V.: 'Efficient algorithms for path problems in weighted graphs', ProQuest, 2008

[40] Palmer, S.: '*Vision science: photons to phenomenology*' (MIT press, Cambridge, MA, 1999)

[41] Reinagel, P., Zador, A., Zador, R.: 'Natural scene statistics at the center of gaze', *Comput. Neural Syst.*, 1998, **10**, (4), pp. 341–350

[42] Tu, W., He, S., Yang, Q., *et al.*: 'Real-time salient object detection with a minimum spanning tree'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016, pp. 2334–2342

[43] Yang, C., Zhang, L., Lu, H., *et al.*: 'Saliency detection via graph-based manifold ranking'. CVPR, Portland, USA, 2013, pp. 3166–3173

[44] Zhu, W., Liang, S., Wei, Y., *et al.*: 'Saliency optimization from robust background detection'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Columbus, USA, 2014, pp. 2814–2821

[45] Jiang, B., Zhang, L., Lu, H., *et al.*: 'Saliency detection via absorbing markova chain'. CVPR, Portland, USA, 2013, pp. 1665–1672

[46] Li, X., Lu, H., Zhang, L., *et al.*: 'Saliency detection via dense and sparse reconstruction'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Portland, USA, 2013, pp. 2976–2983