Online Indexing Structure for Big Image Data used for 3D Reconstruction

Konstantinos Makantasis[1], Yannis Katsaros[2], Anastasios Doulamis[3] and Matthaios Bimpas[3]
[1]*Technical University of Crete, Chania, Greece*
[2]*EXUS Software Ltd., London, U.K.*
[3]*National Technical University of Athens, Athens, Greece*

Keywords: Feature Matching and Indexing, 3D Reconstruction, Image and Video Retrieval, Image-based Modeling.

Abstract: One of the main characteristics of Internet era is the free and online availability of extremely large collections of images. Although the proliferation of millions of shared photos provide a unique opportunity for cultural heritage e-documentation, the main difficulty is that Internet image datasets are unstructured. For this reason, this paper aims to describe a new image indexing scheme with application in 3D reconstruction. The presented approach is capable, on the one hand to index images in a fast and accurate way and on the other to select form an image dataset the most appropriate images for 3D reconstruction, improving this way reconstruction computational time, while simultaneously keeping the same reconstruction performance.

1 INTRODUCTION

Internet era is characterized by extremely large collections of images available over the web that depict not only contemporary events but also historic incidents and cultural heritage assets. These data are being captured from individual users and usually are located on distributed and heterogeneous databases. Although, the proliferation of millions of shared photographs provides a unique opportunity for cultural heritage e-documentation, which includes retrieval, filtering, indexing and finally exploitation of visual information, there are limited technological tools and research methods that meet this purpose.

The main difficulty in using Internet image collections lies in the fact that the stored image content is unstructured. Simple text-based queries are inefficient for handling unstructured visual content, since images' textual descriptions may be quite different of what they are actually depicting. On the one hand, human centric textual annotation of images is an arduous and inconsistent task due to the complexity of visual content and the subjective perception of humans in interpreting it, and on the other autogenerated geo-location tags suffer from low precision since geo-information does not interpret what is actually depicted.

Our research exploits unstructured Internet image collections stored on distributed multimedia platforms to obtain e-documentation of cultural heritage objects through 3D reconstruction. The main difficulty towards this direction is that there are several outliers, images whose visual content is quite dissimilar with the requested cultural heritage object, in the retrieved dataset. The existence of outliers deteriorates the performance and exponentially increases the computational time of 3D reconstruction. While there exists 3D reconstruction algorithms (Wu et al., 2011; Wu et al., 2012), which present robustness against noisy data, their computational complexity significantly increases with respect to the size of input, making direct implementation practically impossible under a cost effective manner. To make things worse, the volumes of the image data, which are stored over distributed Web repositories are extremely huge and varying imposing high computational challenges to any meta-algorithm that exploits these data for real-time application scenarios.

To address this difficulty, in this paper, we propose an incremental structure scheme able to online index, through the calculation of the visual distance, each new incoming image datum with respect to already indexed image volumes in a fast and accurate way. In this way, we are able to online organize retrieved image data under a computationally efficient manner. The proposed online indexing structure allows for an efficient implementation of meta-algorithms

Makantasis, K., Katsaros, Y., Doulamis, A. and Bimpas, M.
Online Indexing Structure for Big Image Data used for 3D Reconstruction.
DOI: 10.5220/0005852207050714
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 705-714
ISBN: 978-989-758-175-5

that can incrementally process big and varying image volumes. In this paper, a content-based filtering approach is presented suitable for selecting appropriate geometric varying images for 3D reconstruction purposes. In particular, our approach exploits the online structure indexing mechanisms to appropriately organize new incoming image data and then adopts geometric properties in a multi-dimensional image manifold (maximize the geometric volume of image points) to select those data that optimize 3D reconstruction operation.

1.1 Previous Works

Content Based Image Retrieval (CBIR) tools are based on a visual matching process, in order to retrieve images from large repositories. They use image filtering and clustering algorithms to appropriately organize images into groups of similar visual properties discarding, therefore, noisy information. A CBIR scheme requires the user to provide a query image to the system. The query image acts as a reference image, whose visual information is encoded. Then, the system responses by retrieving those images from a database, that present high visual similarity with respect to the reference one.

Towards this direction Murthy *et al.* (Murthy et al., 2010) propose a two stage image retrieval procedure based on the color properties of a reference image. Starting from an initial image set, most of the images are filtered by applying hierarchical clustering. Then, k-means is applied on filtered data to get better favored image results. However, the efficiency of this approach inherently depends on camera properties and environmental conditions of the scene at the time the photo was taken. Chum *et al.* in (Chum et al., 2007) present a system, whose objective is to retrieve all instances of a query object in a large image database. The authors employ, behind visual similarities, a vocabulary tree for indexing and query expansion. Similar to the previous approaches the system presented by Philbin *et al.* in (Philbin et al., 2007) enables the user to select an object of interest within a reference image and then it returns a ranked list of images that contain the selected object. Kekre *et al.* in (Kekre et al., 2011) develop image signatures based on image color properties. Signatures are used to create clusters which are represented by codebooks stored in a database. Each new query image is compared against the existing codebooks in order to estimate the most relevant visual matching. The main drawback of the aforementioned approaches is that they require a reference image or an object of interest to carry out the retrieval process. On the contrary, our method eliminates outliers and organizes the retrieved results under an *unsupervised* framework.

Simon *et al.* in (Simon et al., 2007) focus on visual clustering implemented through an optimization approach that selects a number of canonical scene views for constructing a scene summary. However, the authors assume that an image set that represents the scene is pre-constructed. In contrast, our approach is responsible for creating this set.

Besides, visual information, the description of *"digital born"* media is enhanced by textual information, such as automatically generated geo-tags and camera exif data (Yiakoumettis et al., 2014; Doulamis et al., 2012). The works of (Papadopoulos et al., 2010; Arampatzis et al., 2011; Kalantidis et al., 2011) exploit geo-tagging and annotation to improve the retrieval performance. Particularly, the work of (Papadopoulos et al., 2010) describes an image analysis algorithm that automates the detection of landmarks from large multimedia databases in order to improve content-consumption experience. The idea of geo-clustering is also exploited by the work of (Zheng et al., 2009) for retrieving landmark images. This approach combines geo-information along with hierarchical agglomerative clustering to obtain dense geographic clusters. Due to the fact that the retrieved set contains a lot of image outliers visual clustering is performed to eliminate noisy images. Agarwal *et al.* in (Agarwal et al., 2011) use geo-tagged images and assume multiple different views of the same object in each of these datasets. Then, they create a vocabulary tree for indexing and query expansion to cluster together similar images. Although, the aforementioned approaches are useful for CBIR applications, where the aim is to extract similar images upon a query, they present many shortcomings when they apply for 3D reconstruction scenarios.

1.2 Our Contribution

Initially, the on-line indexing structure is constructed with the aim to scale large image volumes. For this reason, a pre-defined number of landmark images are selected to represent as much as possible the image data points. Particularly, for every image, local descriptors are extracted to encode its visual content. In this paper, the ORB (Rublee et al., 2011) descriptor is utilized. Then, an image graph is constructed the vertices of which correspond to the images while the edges to a pairwise image similarity matching. The cMDS algorithm (Cox and Cox, 2008) is adopted to relate the pairwise similarity of the images with respect to Euclidean distances. Therefore we are able to represent an image as a point in a multi-dimensional

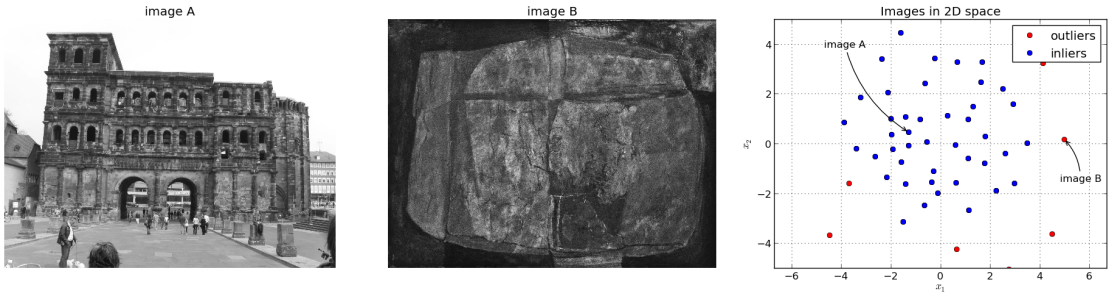


Figure 1: Example of two images that were retrieved by using the textual query "Porta Nigra" and their projection on a 2D manifold. Their coordinates were computed by using the distance between them, which was established by local descriptor pair-wise similarity matching. Image A that depicts the monument is positioned in a high density area, while Image B, which is an outlier, is positioned in a low density area.

manifold. In this multi-dimensional manifold, image landmarks guarantees that the distance of the new incoming image with respect to the remaining indexed ones is able to be computed both computationally efficient under a constant time of operations and effectively.

A textual query and/or geo-information are used to find a subspace in the initial indexed image data that share the same textual and geo-location information. Then, the position of images on the manifold is a clear indicator of how close the visual content of two images is, see Fig.1. The distribution of the retrieved images on the manifold is expected to form i) a compact hyperspace on which images depicting the same object are located and ii) low density areas containing image outliers. In order to develop a robust indexing structure image outliers must be eliminated. Towards this direction space's density property can be exploited through the application of a density based clustering algorithm such as SOS (Janssens et al., 2012). We choose SOS due to its property to compute the probability that a data point is an outlier. Outlier probabilities are favorable to unbounded outliers scores and to hard classification of data, because they allow to select an appropriate and rational threshold for outliers selection.

Having discriminated image data to the compact subspace against the image outliers, the next step is to incrementally extract a set of images that are most suitable for 3D reconstruction. A 3D reconstruction engine exploits different geometric perspectives of an object. For this reason, redundant information can be considered as those images presenting similar geometric views of the object to be reconstructed. The incremental set creation enables us to feed the 3D reconstruction engine with the minimum required number of appropriate geometric views of an object so as to achieve a targeted precise reconstruction at a given scale. The selection technique is based on the fact that the volume contained by a simplex formed by the most representative images is larger than any other simplex volume formed by any other combination of images (Winter, 1999).

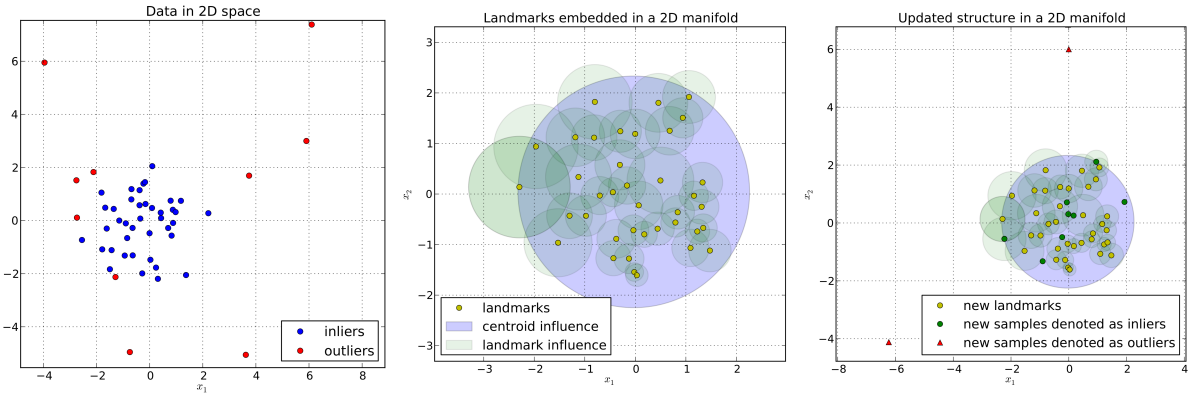
The rest of the paper is organized as follows: Section 2 presents how images are modeled as points on a multi-dimensional manifold. Section 3 focuses on the indexing structure and Section 4 describes the representatives selection technique. Section 5 presents the experimental framework and Section 6 concludes this work.

2 IMAGES AS MULTI-DIMENSIONAL MANIFOLD POINTS

This section presents our approach to encode visual information of an initial retrieved image dataset. We assume that N images, $I^{(1)}, I^{(2)}, ..., I^{(N)}$, are retrieved from web multimedia repositories using geo-location information and textual metadata. Initially, through the adoption of local visual descriptors we represent images' content and then we formulate the similarity/distance between pairs of images. Finally, by exploiting cMDS algorithm we relate the space of distances with the space of Gram matrices, which are used to compute image coordinates onto a multi-dimensional manifold over which each image is represented.

2.1 Geometric Invariant Visual Content Modeling

In this paper, we choose to use ORB descriptor (Rublee et al., 2011) for encoding images' visual content. Our choice is justified by the fact that, on the one hand, ORB performs better than SURF (Bay et al., 2006) and, on the other, it performs as well as SIFT (Lowe, 2004), while being almost two orders of mag-



(a) (b) (c)

Figure 2: (a) Images projected in a 2D manifold. Their coordinates were computed by using their pair-wise distances. (b) Inliers were selected as landmarks and defined a new 2D subspace. (c) New samples (red triangles and green circles) are indexed/projected according to landmarks. Green circles correspond to new samples denoted as inliers, while red triangles correspond to new samples denoted as outliers. New samples that fall into the region of influence of centroid or a landmark are denoted as inliers. In the first case the indexing structure remains as it is, while in the second it is updated.

nitude faster. ORB builds on the FAST keypoint detector (Rosten and Drummond, 2006) and the BRIEF descriptor (Calonder et al., 2010) and addresses their limitations by adding an accurate orientation component to FAST and by incorporating a method for decorrelating BRIEF features under a rotation invariant framework.

To be more specific, for each image pixel p_c, which has been denoted by FAST detector as a corner pixel, a bit-string is adopted from a set of n binary tests $T = \{\tau_1, \tau_2, \cdots, \tau_n\}$, where n is a predefined scalar parameter of the algorithm. The n binary tests take place in an image patch $l(p_c)$ around pixel p_c as follows:

$$\tau_i(l(p_c); q, r) = \begin{cases} 1 & if\ I(r) > I(q) \\ 0 & if\ I(r) \leq I(q) \end{cases}. \quad (1)$$

In Eq(1), variables q, r stands for two pixels within the patch $l(p_c)$, while $I(q)$ and $I(r)$ correspond to image intensities at pixels q and r respectively. Based on the outcome of the n binary tests a feature that describes the patch $l(p_c)$ of image I is constructed as:

$$f_n^{(I)}(l(p_c)) = \sum_{i=1}^{n} 2^{i-1} \tau_i(l(p_c); q, r). \quad (2)$$

By utilizing the *intensity centroid* corner orientation measure (Rosin, 1999) the orientation angle $\theta(l(p_c))$ of the patch $l(p_c)$ can be computed as:

$$\theta(l(p_c)) = \arctan(m_{01}(l(p_c)), m_{10}(l(p_c))), \quad (3)$$

where $m_{01}(l(p_c))$ and $m_{10}(l(p_c))$ stands for the raw moments of the patch $l(p_c)$.

The projection of the feature vector $f_n^{(I)}(l(p_c))$ of Eq.(2) onto the angle $\theta(l(p_c))$ results in a rotation invariant binary representation vector, $\varphi_n^{(I)}(l(p_c))$, of patch $l(p_c)$. Then, the visual content of an image I is represented by a matrix $\Phi^{(I)} \in \{0,1\}^{K \times n}$:

$$\Phi^{(I)} = [\varphi_n^{(I)}(l(p_1)) \ \cdots \ \varphi_n^{(I)}(l(p_K))]^T, \quad (4)$$

where K is a predefined scalar parameter of ORB descriptor algorithm and stands for the number of detected keypoints in an image.

2.2 Formation Image Graphs

For estimating visual similarity between two images, A and B, their correspondent points have to be computed. Correspondences can be estimated by performing a nearest-neighbor keypoints matching algorithm between every pair of images. Due to the fact that ORB keypoints are described by a binary pattern, multi-probe LSH (Lv et al., 2007) is used exploiting the Hamming distance, D_H.

Let us denote as $k_i^{(A)}$ the i^{th} keypoint of image A, which is described by the vector $\varphi_n^{(A)}(l(p_i))$. Then, the most relevant keypoint $k_{ji}^{(B)}$ of image B with respect to the $k_i^{(A)}$ is obtained by the following relation:

$$j_i = \underset{j=1,2,\ldots K}{\operatorname{argmin}} (D_H(\varphi_n^{(A)}(l(p_i)), \varphi_n^{(B)}(l(p_j)))). \quad (5)$$

Having detected all correspondent points between two images A and B we can form a set

$$M^{(A \to B)} = \{(k_i^{(A)}, k_{ji}^{(B)}) | i = 1, 2, \ldots K\} \quad (6)$$

that contains all keypoints $k_i^{(A)}$, $k = 1,2,...,K$ along with the correspondent points $k_{ji}^{(B)}$.

For every pair of images in the dataset, a *two-way* matching is performed. The set of final matches,

$$M^{(A,B)} = M^{(A\to B)} \cap M^{(B\to A)} \tag{7}$$

between images A and B is defined as the intersection of the sets $M^{(A\to B)}$ and $M^{(B\to A)}$.

Two-way matching compensates inconsistencies caused by the fact that the nearest neighbor of an extracted keypoint in image A may be different from the nearest neighbor of the correspondent keypoint in image B.

Using the $M^{(A,B)}$ set, we define a visual similarity metric between images A and B as:

$$s_{i=A,j=B} = \frac{|M^{(A,B)}|}{K}, \tag{8}$$

where $|M^{(A,B)}|$ refers to the cardinality of $M^{(A,B)}$ set.

The output of the aforementioned process for N images is an $N \times N$ symmetric matrix $\boldsymbol{S}$ with elements $s_{ij} \in [0,1]$, $i,j = 1,2,...,N$. Variable s_{ij} takes value close to zero for two quite dissimilar images and close to one when two images are similar. As $\boldsymbol{D}$ we denote the log version of matrix $\boldsymbol{S}$ so as to similar images receive close to zero while quite dissimilar very high value;

$$\boldsymbol{D} = [d_{ij}] = -\log(\boldsymbol{S}). \tag{9}$$

$\boldsymbol{D}$ is an $N \times N$ symmetric matrix with non negative elements and zeros on the main diagonal.

2.3 Image Graph Projection onto Multi-dimensional Manifold

Let us denote as $\boldsymbol{x}^{(i)} \in \mathbb{R}^{\mu}$ the coordinates of i^{th} image in the μ-dimensional space. The space is defined such that the norm between two points of the space, represented by the coordinates $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$, should be equal to their respective image distance, $d_{ij} = -log(s_{ij})$, defined in Eq.(9). The coordinates of all N images in the dataset can be compactly represented by a matrix $\boldsymbol{X}$;

$$\boldsymbol{X} = [(\boldsymbol{x}^{(1)})\ (\boldsymbol{x}^{(2)})\ \cdots\ (\boldsymbol{x}^{(N)})]^T \in \mathbb{R}^{N\times\mu}. \tag{10}$$

If we define the Gram matrix $\boldsymbol{B} = \boldsymbol{X} \cdot \boldsymbol{X}^T$ of images coordinates, then cMDS algorithm can be used to establish a connection between the space of the distances and the Gram matrix $\boldsymbol{B}$ based on the following theorem (the proof can be found in (Cayton, 2006)).

Theorem 1. *A non-negative symmetric matrix $\boldsymbol{D} \in \mathbb{R}^{N\times N}$, with zeros on the diagonal, is an Euclidean distance matrix if and only if $\boldsymbol{B} \equiv -\frac{1}{2}\boldsymbol{HDH}$, where $\boldsymbol{H} \equiv \boldsymbol{I} - \frac{1}{N}\boldsymbol{1}\boldsymbol{1}^T$, is positive semidefinite. Furthermore, this $\boldsymbol{B}$ will be the Gram matrix for a mean centered configuration with interpoint distances given by $\boldsymbol{D}$.*

In cases where dissimilarity matrix $\boldsymbol{D}$ is not Euclidean the matrix $\boldsymbol{B}$ as described by the above theorem will not be positive definite, and thus will not be a Gram matrix. To handle such cases, cMDS algorithm projects the matrix $\boldsymbol{B}$ onto the cone of positive semi-definite matrices by setting its negative eigenvalues to zero. In order to get matrix $\boldsymbol{X}$, the matrix $\boldsymbol{B}$ is spectrally decomposed into $\boldsymbol{B} = \boldsymbol{UVU}^T$ and then $\boldsymbol{X} = \boldsymbol{UV}^{1/2}$. If we denote as q_i and λ_i for $i = 1,2,...,N$ the eigenvectors and eigenvalues of $\boldsymbol{B}$, then matrix $\boldsymbol{U}$ is a square $N \times N$ matrix whose i^{th} column is the eigenvector q_i of $\boldsymbol{B}$ and $\boldsymbol{V} = [v_{ii}]$ is the diagonal matrix whose elements v_{ii} are the corresponding eigenvalues, i.e. $v_{ii} = \lambda_i$. Finally the dimension μ of the multi-dimensional space is equal to the multiplicity of non-zero eigenvalues of matrix $\boldsymbol{B}$.

3 THE ONLINE IMAGE INDEXING STRUCTURE

The number of available images stored on Internet multimedia repositories is continuously increasing. For this reason, the proposed method focuses on creating an indexing structure capable to process *online* new retrieved images other than those included in the initial dataset. However, in order to develop a robust indexing structure, we must eliminate image outliers and form a set that will contain only the visually similar images.

By using the representation of images as points onto an μ-dimensional space, we can intuitively note that outliers must reside to low spatial density areas, whereas visually similar images must form areas of high spatial density. Exploiting the density property, or in other words, the affinity between image points, the μ-dimensional manifold must be partitioned into two disjoint subspaces, $\mathcal{C}$ and $\bar{\mathcal{C}}$, such as all visually similar images belong to $\mathcal{C}$ and all outliers to $\bar{\mathcal{C}}$.

3.1 Affinity-based Partitioning

An affinity-based approach for selecting outliers is the SOS algorithm (Janssens et al., 2012). This algorithm employs the concept of affinity to quantify the relationship from one image point to another. Based on this relationship an image point is denoted as outlier when all other points have insufficient affinity with it.

By using the distance, d_{ij} defined in Eq.(9), between image points $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$, the affinity between

these points can be defined as:

$$\alpha_{ij} = \begin{cases} e^{-(d_{ij}^2/2\sigma_i^2)} & if\ i \neq j \\ 0 & if\ i = j \end{cases}, \quad (11)$$

where σ_i^2 is scalar variance associated with image point $\boldsymbol{x}^{(i)}$. As shown by Eq.(11) an image point has no affinity with itself and the affinity that the point $\boldsymbol{x}^{(i)}$ has with point $\boldsymbol{x}^{(j)}$ is proportional to the probability density at $\boldsymbol{x}^{(j)}$ under a Gaussian distribution $\mathcal{N}(\boldsymbol{x}^{(i)}, \sigma_i^2)$. For determining the variance σ_i^2 for each image point, SOS uses an adaptive approach. Concretely, it employs the perplexity parameter h, which is used to set adaptively the variances in such a way that each point has h effective neighbors (Hinton and Roweis, 2002). At this point it has to be mentioned that h is the only parameter that SOS algorithm requires to be pre-defined.

Unlike to distance matrix $\boldsymbol{D}$, the affinity matrix $\boldsymbol{A} = [\alpha_{ij}]$ is not symmetric. By using the affinity distribution $\boldsymbol{\alpha}_i = [\alpha_{i1}\ \alpha_{i2}\ ...\ \alpha_{iN}]$ for the point $\boldsymbol{x}^{(i)}$, a discrete probability distribution $\boldsymbol{b}_i$ that shows the probability that point $\boldsymbol{x}^{(i)}$ chooses any one of the other points as its neighbors, is defined as

$$\boldsymbol{b}_i = [b_{i1}\ b_{i2}\ ...\ b_{iN}] \ \text{ where } \ b_{ij} = \frac{\alpha_{ij}}{\sum_{k=1}^{N} \alpha_{ik}}. \quad (12)$$

The probability distribution $\boldsymbol{b}_i$ corresponds to the *normalized* affinity $\boldsymbol{\alpha}_i$.

After the estimation of probability distribution $\boldsymbol{b}_i$ the probability the image point $\boldsymbol{x}^{(i)}$ to be denoted as outlier can be estimated by the following theorem (the proof can be found in (Janssens et al., 2012)).

Theorem 2. *If α_{ij} is the affinity that data point $\boldsymbol{x}^{(i)}$ has with data point $\boldsymbol{x}^{(j)}$ and b_{ij} is the normalized affinity between these two points, then the probability that data point $\boldsymbol{x}^{(i)}$ belongs to the outliers class, $\bar{C}$, is given by:*

$$p(\boldsymbol{x}^{(i)} \in \bar{C}) = \prod_{j \neq i} (1 - b_{ji}). \quad (13)$$

The above theorem states that the probability that an image point $\boldsymbol{x}^{(i)}$ belongs to the outliers class, $\hat{C}$, is the probability that this point is never chosen as a neighbor of the other image points.

For N images, the output of SOS algorithm can be compactly represented by a vector $\boldsymbol{\rho} \in \mathbb{R}^N$.

$$\boldsymbol{\rho} = [p(\boldsymbol{x}^{(1)} \in \bar{C})\ \ ...\ \ p(\boldsymbol{x}^{(N)} \in \bar{C})]^T. \quad (14)$$

Using Eq.(14) the set Q that will contain the coordinates of the inlier images can be defined as

$$Q = \{\boldsymbol{x}^{(i)} \mid \rho_i < \theta\} \ \text{ for } \ i = 1, 2, ...N. \quad (15)$$

In Eq.(15) ρ_i stands for the i^{th} element of $\boldsymbol{\rho}$ and θ is a probability threshold to discriminate image outliers than inliers.

3.2 Indexing Structure Initialization

Let us define the set $\mathcal{L} = \{\boldsymbol{x}^{(i)} \mid \boldsymbol{x}^{(i)} \in Q\}$, which contains visual similar images' coordinates onto the multi-dimensional space. The image points $\boldsymbol{x}^{(i)} \in \mathcal{L}$ act as landmarks that determine if a new image $\hat{I}$ must be denoted as inlier or outlier. The elements of $\mathcal{L}$ define a space with a centroid, c, whose coordinates are $\boldsymbol{x}^{(c)}$. Regions of influence are defined around the centroid and each one of the landmarks. The region of influence of centroid, R_c, is defined as

$$R_c(\boldsymbol{x}^{(c)}, r_c) = \{\boldsymbol{x} \mid (\boldsymbol{x} - \boldsymbol{x}^{(c)})^T (\boldsymbol{x} - \boldsymbol{x}^{(c)}) \leq r_c\}, \quad (16)$$

where $r_c = \max\{\|\boldsymbol{x}^{(c)} - \boldsymbol{x}^{(i)}\|_2 \mid \boldsymbol{x}^{(i)} \in \mathcal{L}\}$. In a similar way is defined the region of influence of a landmark $\boldsymbol{x}^{(i)}$

$$R_i(\boldsymbol{x}^{(i)}, r_i) = \{\boldsymbol{x} \mid (\boldsymbol{x} - \boldsymbol{x}^{(i)})^T (\boldsymbol{x} - \boldsymbol{x}^{(i)}) \leq r_i\}. \quad (17)$$

In this case r_i is defined as

$$r_i = \min\{\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|_2 \mid \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)} \in \mathcal{L} \text{ and } i \neq j\}. \quad (18)$$

Regions of influence are used, as described in the next subsection, for classifying new retrieved images as inliers or outliers.

3.3 Online Image Indexing

Let us assume that a new image, $\hat{I}$ is retrieved. We define the set $Q_{\hat{I}}$ as:

$$Q_{\hat{I}} = \{I^{(i)} \mid \boldsymbol{x}^{(i)} \in Q\} \quad (19)$$

The distances between $\hat{I}$ and each one of the images $I^{(i)} \in Q_{\hat{I}}$ are computed by the method described in Section 2.

In order to index the new image $\hat{I}$, it has to be projected onto the multi-dimensional geometric space defined by images belonging to $Q_{\hat{I}}$. Let $\hat{\boldsymbol{x}}^{(\hat{I})}$ be the coordinates of image $\hat{I}$ after its projection onto the multi-dimensional space. The objective of assigning coordinates to image $\hat{I}$ is to minimize the distance distortion given by the following relation:

$$e(I^{(i)}, \hat{I}) = \mid d(I^{(i)}, \hat{I}) - \|\hat{\boldsymbol{x}}^{(\hat{I})} - \boldsymbol{x}^{(i)}\|_2 \mid \quad (20)$$

$d(I^{(i)}, \hat{I})$ is the distance between images $I^{(i)}$ and $\hat{I}$ computed by Eq.(9) and $\|\cdot\|_2$ refers to the L^2-norm of a vector. Eq.(20) measures distance distortion by the absolute error.

The problem of assigning coordinates to image $\hat{I}$ can be seen as a typical optimization problem where the following objective function is minimized.

$$\underset{\hat{\boldsymbol{x}}^{(\hat{I})}}{\operatorname{argmin}} \sqrt{\sum_{i=1}^{L} e(I^{(i)}, \hat{I})^2} \quad (21)$$

For estimating the optimal coordinates $\hat{\boldsymbol{x}}^{(\hat{I})}$ we used simplex downhill method. The time for projecting a new image onto an μ-dimensional space is determined by the simplex downhill method. In general simplex downhill with an objective function g takes $O(mD \times f(g))$ time, where $f(g)$ is the cost to evaluate g, D is the number of dimensions and m the number of iterations. In our case, we have $D = \mu$ and $f(g) = L \cdot \mu$, where L stands for the cardinality of Q_I. The second equation holds because we need to calculate the distances between image $\hat{I}$ and each one of the images $I^{(i)} \in Q_I$ in an μ-dimensional space. In all, the time complexity for indexing a new image is $O(mL\mu^2)$.

Having defined the regions of influence for the centroid and each one of the landmarks (Subsection 3.2), a new image, $\hat{I}$ with coordinates $\hat{\boldsymbol{x}}^{(\hat{I})}$, is denoted as inlier only if $\hat{\boldsymbol{x}}^{(\hat{I})} \in R_c$ or $\hat{\boldsymbol{x}}^{(\hat{I})} \in R_i$ for some $i = 1, 2, ..., |\mathcal{L}|$, where $|\mathcal{L}|$ stands for the cardinality of set $\mathcal{L}$.

If $\hat{\boldsymbol{x}}^{(\hat{I})} \in R_c$ the $\mathcal{L}$ set remains as it is, while Q and Q_I sets are updated according to the following relation:

$$Q := Q \cup \hat{\boldsymbol{x}}^{(\hat{I})} \quad \text{and} \quad Q_I := Q_I \cup \hat{I} \tag{22}$$

If $\hat{\boldsymbol{x}}^{(\hat{I})} \in R_i$ for some $i = 1, 2, ..., |\mathcal{L}|$ and $\hat{\boldsymbol{x}}^{(\hat{I})} \notin R_c$ the sets Q and Q_I are updated according Eq.(22), but in this case the set $\mathcal{L}$ is also updated as:

$$\mathcal{L} := \mathcal{L} \cup \hat{\boldsymbol{x}}^{(\hat{I})} - \min\{\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(c)}\|_2 \mid \boldsymbol{x}^{(i)} \in \mathcal{L}\} \tag{23}$$

This adaptation takes place for taking into consideration new images visual content, while at the same time keeping constant the number of landmarks.

4 REPRESENTATIVE OBJECT GEOMETRIC PERSPECTIVES

After the creation of Q and Q_I, we need to select the most representative images corresponding to different geometric perspectives of the cultural heritage object under 3D reconstruction. The representative images are fed as input to a 3D reconstruction algorithm to improve computational time while simultaneously keeping the same reconstruction accuracy.

4.1 Representatives Selection through Simplex Volume Expansion

We assume that the μ-dimensional volume formed by a simplex with vertices specified by the points of the most representative images should be larger than that formed by any other combination of image points. Let us denote as $\boldsymbol{\nu}^{(i)}$ the i^{th} representative image point, as β the number of representative images required to generate, as $Q_R = \{\boldsymbol{\nu}^{(1)}, \boldsymbol{\nu}^{(2)}, ..., \boldsymbol{\nu}^{(\beta)}\} \subseteq Q$ the set that contains the representative images' points and as $\boldsymbol{w}^{(j)}$ the row vector that equals to $\boldsymbol{\nu}^{(j)} - \boldsymbol{\nu}^{(1)}$ for $j = 2, 3, ..., \beta$. Then the volume, $V(Q_R)$, of the simplex whose vertices are the points $\boldsymbol{\nu}^{(i)}$ for $i = 1, 2, ..., \beta$ can be computed as:

$$V(Q_R) = \frac{|\det(\boldsymbol{W}\boldsymbol{W}^T)|^{1/2}}{(\beta - 1)!} \tag{24}$$

where $\boldsymbol{W}$ is an $(\beta - 1) \times \mu$ matrix whose rows are the row vectors $\boldsymbol{w}^{(j)}$.

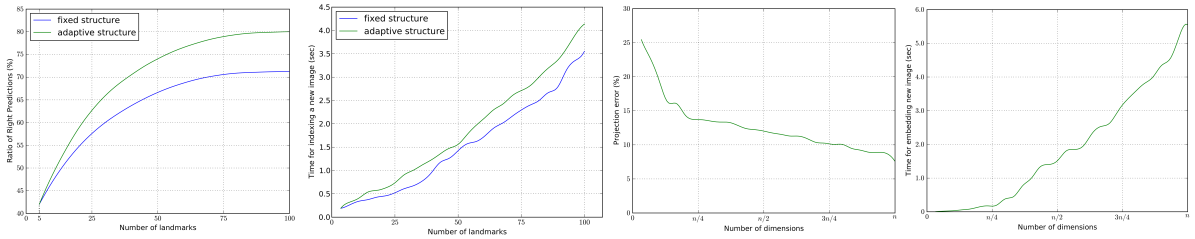
For estimating the most representative images, initially the set Q_R is constructed by randomly selecting β images from set Q and calculate the volume of the simplex formed by the elements of Q_R. Then, an iterative approach is adopted to test every image in the set Q as a candidate representative. To be more specific, each one of the image points of Q_R is replaced, one at a time, with an image point $\hat{\boldsymbol{\nu}}$ from Q that is being tested as candidate representative. Then, the algorithm evaluates if replacing any of the elements of Q_R with the image point being tested results in a larger simplex volume. If this is true, let's say for the point $\boldsymbol{\nu}^{(j)} \in Q_R$, then the $\boldsymbol{\nu}^{(j)}$ point is replaced by the image point $\hat{\boldsymbol{\nu}}$ and the process is repeated again until each image from Q set is evaluated.

For making the selection method *scalable* to large datasets, we follow an incremental approach. Let us assume that β representatives are known. Then, the problem of selecting $\beta + 1$ representatives can be reduced to finding $\beta + 1$ representatives *given* β of them. This way, only the volumes of simplices formed bythe elements of the sets $Q_R \cup \boldsymbol{x}^{(i)}$ for $\boldsymbol{x}^{(i)} \in Q$ need to be evaluated.

5 EXPERIMENTAL RESULTS

In the framework fo this research, we have collected from Internet image repositories images depicting different cultural heritage monuments, such as *Porta Nigra* in Germany, *Parthenon* in Athens and *Descobrimentos* in Lisboa. All these images have been gathered with respect to their textual annotation and geoinformation regardless of the actual type of content they depict. Thus, for each cultural heritage category, a large number of image outliers are encountered.

The evaluation of the presented approach took place in regard to indexing, in terms of accuracy and time complexity, as well as to 3D reconstruction accuracy after the selection of the most representative images. The algorithm was developed in Python and executed on a conventional i5 CPU laptop.



(a) Right predictions (b) New image indexing (c) Projection error (d) New image projection

Figure 3: Diagram (a) shows the ratio of right denotations of new images as inliers or outliers in regard to the number of landmarks, while diagram (b) presents the time required to classify a new image. Diagram (c) shows the projection error when assigning coordinates to new images in regard to the number of dimensions of the space onto which the images are projected. The time required to project a new image onto the multi-dimensional space is presented in (d).

5.1 Indexing Evaluation

In order to evaluate indexing mechanism, we created an indexing structure using a varying number of landmarks. Then, we manually selected one hundred outlier images and one hundred inlier images. These images are fed to the indexing mechanism in order to be classified. Two different versions of the algorithm were tested; using a fixed indexing structure and an adaptive indexing structure. In the first case, the indexing structure remains fixed, while in the latter the adaptation mechanism is enabled and the set of landmarks is updated in order to include new images visual information.

Diagram (a) of Fig.(3) presents the ratio of right denotations of new images as inliers or outliers in regard to the number of landmarks, while diagram (b) at the same figure shows the time required to classify a new image. The version that uses the adaptive indexing structure is presented to outperform the one that uses the fixed structure, due to the fact that it exploits visual information of new images. However it requires more time to classify a new image, as it needs extra time to adapt the indexing structure.

Diagrams (c) and (d) of Fig.(3) present the projection error when assigning coordinates to new images and the time required to project a new image in regard to the number of dimensions of the space onto which the images are projected. The parameter n in x-axis refers to the number of dimensions of the space. In this case parameter n was set to 100 at the same value was set and the number of landmarks used by indexing structure. As shown in diagram (c) the projection error is constantly decreasing as the number of space dimensions is increasing. In diagram (d) the time required to project a new image onto a multi-dimensional space is increasing as the number of space's dimensions is getting larger. This is aligned with the time complexity analysis presented in subsection 3.3.

5.2 Representatives Selection Evaluation

For evaluating our representatives selection approach, we used expert's assessment in order to select from the set Q that contain the visual similar images the n most appropriate for 3D reconstruction: i.e. images correspond to different views of the under reconstruction object.

The set of visually similar images contained N elements, and we selected $n = N/5$ of them as the most representatives, set Q_r. Then, we asked from our representatives selection algorithm to extract $n/5$, $2n/5$, $3n/5$, $4n/5$ and n images from the set Q. The set of extracted images are denoted as $\hat{Q}_i$, where $i \in \{n/5, 2n/5, 3n/5, 4n/5, n\}$ In this framework reconstruction accuracy is defined as $A = |Q_r \cap \hat{Q}_i|/|Q_r|$, where $|\cdot|$ represents the cardinality of a set. By the definition of reconstruction accuracy is obvious that for the cases of $n/5$, $2n/5$, $3n/5$, $4n/5$ and n extracted images, the maximum reconstruction accuracy that can be obtained is 20%, 40%, 60%, 80% and 100% respectively.

Furthermore, we compared our representative selection algorithm with two well known algorithms; K-Means and spectral clustering using normalized cut and min cut. We request from K-Means and spectral clustering algorithms to partition the set Q into $n/5$, $2n/5$, $3n/5$, $4n/5$ and n clusters. Then, from each one of the clusters we selected as representative image, the image that belongs to Q and is closer to centroid than the rest images of the same cluster.

Evaluation results are shown in Fig.(4). As the number of cluster is getting larger, the performances of K-Means and spectral clustering is increasing. This is justified by the fact that as the number of clusters is increasing, each one of them contains fewer elements and thus the probability to select the true representative is increasing. However, our approach outperforms both algorithms in all cases. Fig.(5) shows

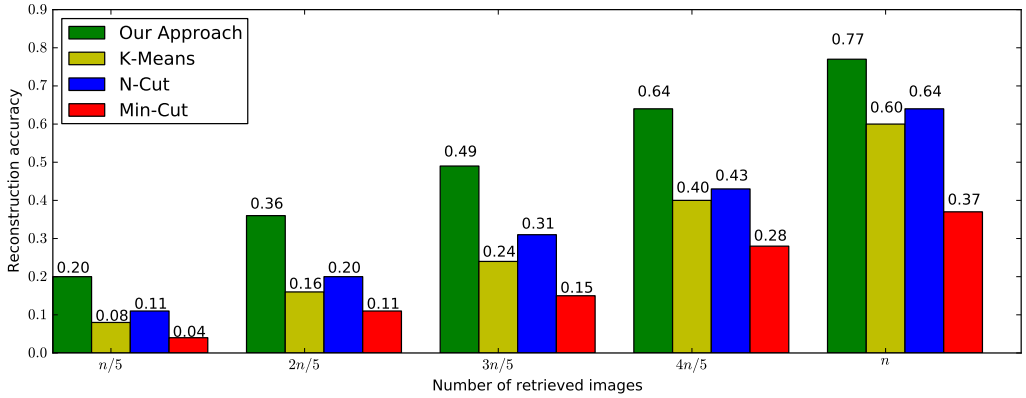


Figure 4: This figure presents reconstruction accuracy in regard to the number of selected representatives.

Figure 5: (a) - (e) show reconstruction results for *"Porta Nigra"* by selecting $n/5$, $2n/5$, $3n/5$, $4n/5$ and n images using our representatives selection approach. (f) shows reconstruction when all images selected by an expert were used.

reconstruction results for *"Porta Nigra"* by selecting $n/5$, $2n/5$, $3n/5$, $4n/5$ and n images using our representatives selection approach.

6 CONCLUSIONS

This paper presents an image indexing approach with application to 3D reconstruction, which is capable to index new images in a *fast* and *accurate* way.

Given a set of images, local descriptors are used to encode images' visual content, which, then, is used for estimating a similarity metric between images. This results in the construction of a similarity matrix. Using this similarity matrix images are represented as points into a multi-dimensional space. Exploiting images' coordinates the indexing structure is initialized by eliminating outliers and forming a set of visually similar images. Then, based on the indexing structure, each new retrieved image can be denoted online as inlier or outlier. Furthermore, an accurate algorithm is described for selecting the most appropriate images for 3D reconstruction; i.e. images that depict different views of the same object.

ACNOWLEDGEMENTS

The research leading to these results has been supported by Marie Curie IAPP project 4D-CH-World: Four Dimensional Cultural Heritage World. Grant agreement number324523.

REFERENCES

Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S. M., and Szeliski, R. (2011). Building rome in a day. *Commun. ACM*, 54(10):105112.

Arampatzis, A., Zagoris, K., and Chatzichristofis, S. A. (2011). Dynamic two-stage image retrieval from large multimodal databases. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 326–337. Springer Berlin Heidelberg.

Bay, H., Tuytelaars, T., and Gool, L. V. (2006). SURF: speeded up robust features. In *Computer Vision ECCV 2006*, number 3951 in Lecture Notes in Computer Science, pages 404–417. Springer Berlin Heidelberg.

Calonder, M., Lepetit, V., Strecha, C., and Fua, P. (2010). BRIEF: binary robust independent elementary features. In *Computer Vision ECCV 2010*, number 6314 in Lecture Notes in Computer Science, pages 778–792. Springer Berlin Heidelberg.

Cayton, L. (2006). Algorithms for manifold learning. Technical Report CS2008-0923, University of California, San Diego, Tech.

Chum, O., Philbin, J., Sivic, J., Isard, M., and Zisserman, A. (2007). Total recall: Automatic query expansion with a generative feature model for object retrieval. In *IEEE 11th Intern. Conf. on Comp. Vision. ICCV*, pages 1–8.

Cox, M. A. A. and Cox, T. F. (2008). Multidimensional scaling. In *Handbook of Data Vis.*, Springer Handbooks Comp.Statistics, pages 315–347. Springer.

Doulamis, N., Yiakoumettis, C., and Miaoulis, G. (2012). On-line spectral learning in exploring 3d large scale geo-referred scenes. In *Progress in Cultural Heritage Preservation*, pages 109–118. Springer.

Hinton, G. E. and Roweis, S. T. (2002). Stochastic neighbor embedding. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in NIPS 15*, pages 833–840.

Janssens, J., Huszar, F., Postma, E., and van den Herik, J. (2012). Stochastic outlier selection. Technical Report TiCC TR 2012-001, Tilburg University, Netherlands.

Kalantidis, Y., Tolias, G., Avrithis, Y., Phinikettos, M., Spyrou, E., Mylonas, P., and Kollias, S. (2011). VIRaL: visual image retrieval and localization. *Multimedia Tools and Applications*, 51(2):555–592.

Kekre, D. H. B., Sarode, T. K., Thepade, S. D., and Vaishali, V. (2011). Improved texture feature based image retrieval using kekres fast codebook generation algorithm. In *Thinkquest*, pages 143–149. Springer India.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Lv, Q., Josephson, W., Wang, Z., Charikar, M., and Li, K. (2007). Multi-probe LSH: efficient indexing for high-dimensional similarity search. In *P33rd Inter. Conf on VLDB*, VLDB '07, page 950961, Vienna, Austria.

Murthy, V. S. V. S., Kumar, S., and Rao, P. S. (2010). Content based image retrieval using hierarchical and k-means clustering techniques. *Intern. Journal of Engineering Science and Technology*, 2(3).

Papadopoulos, S., Zigkolis, C., Kompatsiaris, Y., and Vakali, A. (2010). Cluster-based landmark and event detection on tagged photo collections. *IEEE Multimedia*.

Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conf. on Comp. Vision and Pattern Recognition. CVPR*, pages 1–8.

Rosin, P. (1999). Measuring corner properties. *Computer Vision and Image Understanding*, pages 291–307.

Rosten, E. and Drummond, T. (2006). Machine learning for high-speed corner detection. In *Computer Vision ECCV 2006*, number 3951 in Lecture Notes in Computer Science, pages 430–443. Springer Berlin Heidelberg.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). ORB: an efficient alternative to SIFT or SURF. In *2011 IEEE Intern. Conf. on Comp. Vision (ICCV)*, pages 2564–2571.

Simon, I., Snavely, N., and Seitz, S. M. (2007). Scene summarization for online image collections. In *IEEE 11th Intern. Conf. on Comp. Vision. ICCV*, pages 1–8.

Winter, M. E. (1999). N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data. volume 3753, pages 266–275.

Wu, C., Agarwal, S., Curless, B., and Seitz, S. (2011). Multicore bundle adjustment. In *IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR)*, pages 3057–3064.

Wu, C., Agarwal, S., Curless, B., and Seitz, S. (2012). Schematic surface reconstruction. In *IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR)*, pages 1498–1505.

Yiakoumettis, C., Doulamis, N., Miaoulis, G., and Ghazanfarpour, D. (2014). Active learning of users preferences estimation towards a personalized 3d navigation of geo-referenced scenes. *GeoInformatica*, 18(1):27–62.

Zheng, Y.-T., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T.-S., and Neven, H. (2009). Tour the world: Building a web-scale landmark recognition engine. In *IEEE Conf. on Comp. Vision and Pattern Recognition. CVPR*, pages 1085–1092.