Mounira Taileb, Sami Touati

[[1]](#footnote-2)

NOHIS-Tree: High-Dimensional Index Structure for Similarity Search

*Abstract*—In Content-Based Image Retrieval systems it is important to use an efficient indexing technique in order to perform and accelerate the search in huge databases. The used indexing technique should also support the high dimensions of image features. In this paper we present the hierarchical index NOHIS-tree (Non Overlapping Hierarchical Index Structure) when we scale up to very large databases. We also present a study of the influence of clustering on search time. The performance test results show that NOHIS-tree performs better than SR-tree. Tests also show that NOHIS-tree keeps its performances in high dimensional spaces. We include the performance test that try to determine the number of clusters in NOHIS-tree to have the best search time.

*Keywords*—High-dimensional indexing, k-nearest neighbors search.

# INTRODUCTION

T

here is a need for content-based image retrieval systems in many fields such as geography, security and criminal picture identification system. To implement a CBIR system two areas are involved which are image processing and similarity search. Image processing is the automatic description of the images; it consists of extracting from an image its visual properties (form, color, texture…). These properties are represented as multidimensional vectors called descriptors [1]. To find the images similar to a query image, a similarity search such k- nearest neighbors is made for each descriptor of the query image. Many similarity searches are carried out: as many searches as descriptors characterizing the image query. The use of a data structure to index the descriptors database is essential.. The need for an efficient indexing technique becomes a major concern for two reasons: the large quantity of images and the high dimensionality of descriptors.

Existing indexing techniques work well with low-dimensional descriptors. However, their performance degrades when the dimension of descriptors increases. This phenomenon is known as the *curse of dimensionality* [2].

The rest of this paper is organized as follows. Section 2 discusses the related work. Section 3 describes the indexing technique NOHIS-tree. Section 4 evaluates performance of NOHIS-tree; it is compared to sequential scan, PDDP- tree and SR-tree. SR-tree has been chosen because it is a multidimensional index proposed recently which still attracts the attention [3]. Finally, section 5 concludes the paper.

# Related Work

Obtaining a high-dimensional index can be made by using traditional techniques of indexing such as R-tree [4], or by using a clustering algorithm to form clusters or groups of descriptors, and the clusters are supported by a hierarchical structure, as an example BIRCH use CF-tree [5], DBSCAN use R\*-tree [6] and X-tree [7]. Many high-dimensional index structures have been proposed, the most known and used are data-partitioning based index structure such as SS-tree [8], SR-tree [9], X-tree, considered as extensions of R-tree, and space-partitioning based index structure such as k-d-B-tree [10], hB-tree [11], and LSDh-tree [12] derived from kd-tree [13]. The R-tree-based index structures suffer from overlapping between bounding regions and the low fanouts, this influence negatively on the results of query processing. The kd-tree based index structures drawbacks are essentially the no guarantee of using allocated space; this led to the consultation of few populated or empty clusters.

Taking into account drawbacks cited above, and with an aim to accelerate nearest neighbors’ search, the high dimensional index technique called NOHIS was proposed [14]. It is composed of two phases:

- The first offline phase consists in gathering descriptors in clusters; the clustering algorithm used is the Principal Direction Divisive Partitioning (PDDP) [15]. It’s one of the divisive hierarchical clustering algorithms; it divides data recursively into two sub-clusters by using the hyper-plane orthogonal to the principal direction derived from the covariance matrix and passing through the cluster center to divide. A binary not balanced tree is obtained at the end of the clustering process. The contribution consists of using minimum bounding rectangle (MBR) avoiding overlap, MBRs are directed according to the principal direction (principal component) used in the clustering algorithm to divide a cluster into two sub-clusters. We call NOHIS-tree, the tree obtained by using PDDP, in which we use non overlapping rectangles.

- The second online phase is the step during which the interrogation of the obtained index is made by a nearest neighbors search carried out on the NOHIS-tree. The search algorithm is adapted to the used MBRs.

# the nohis-tree

The existing multi-dimensional indexing techniques can be divided in two groups according to the partitioning strategy,

1. **M. Taileb is with Information Technology department, King Abdul-Aziz University, Jeddah, KSA (e-mail: mtaileb@kau.edu.sa).**

   **S. Touati, is with King Saud University, Riyadh, KSA (e-mail:** [**stouati@ksu.edu.sa**](mailto:stouati@ksu.edu.sa)**).**

   Acknowledgment: The authors would like to thank the Information Technology department of Faculty of Computing and Information Technology, King Abdul-Aziz University, Jeddah, KSA, and the research center in the college of computer and information sciences, King Saud University, for their financial support to complete this study. [↑](#footnote-ref-2)