On (not) indexing quadratic form distance by metric access methods

Tomáš Skopal
Tomáš Bartoš
Jakub Lokoč

SIRET Research Group
Faculty of Mathematics and Physics
Charles University in Prague

Overview

- Quadratic Form Distance (QFD)
- **QMap model**
  - QFD to L_2 Space Transformation
- QMap and MAMs
  - Experimental evaluation
- Conclusion

Technical Background

- Content-based similarity searching
  - Pair-wise similarity
  - High-dimensional (feature) vectors
- Fast query processing
  - Efficiency
  - Effectiveness

Quadratic Form Distance (QFD)

- Similarity measuring:

$$QFD_A(\mathbf{u},\mathbf{v}) = \sqrt{(\mathbf{u}-\mathbf{v})\mathbf{A}(\mathbf{u}-\mathbf{v})^T}$$

- **u**, **v** feature vectors (1×n)
- **A** similarity matrix/QFD matrix (n×n)
  - positive definite ($zAz^T > 0$)
  - static / dynamic correlations
  - data independent

Quadratic Form Distance (QFD)

- Applications
  - QBIC project (Querying Images by content)
  - 2D & 3D shapes
  - Protein structures
  - *MindReader*
- Advanced
  - SQFD (Signature QFD)

Indexing QFD

- Transformation approaches
  - QBIC system
- Lower-bounding (e.g. Faloutsos et. al 1994)
  - Contractive reduction techniques
  - SVD / KLT decompositions
- Combination (e.g. Hafner et. al 1995)
  - Transformation to k-dimensional L_p space

Motivation

- Metric Access Methods (MAMs)
  - Effective/efficient similarity searching
  - Reduce distance computations
  - Complexity depends on distance function
- QFD is considered as **expensive** – $O(n^2)$
  - Indexing needed
- We show the transformation of QFD
  - Obtain **cheaper** distance function – $O(n)$

QFD vs. L_2 comparison (correlated dimensions)

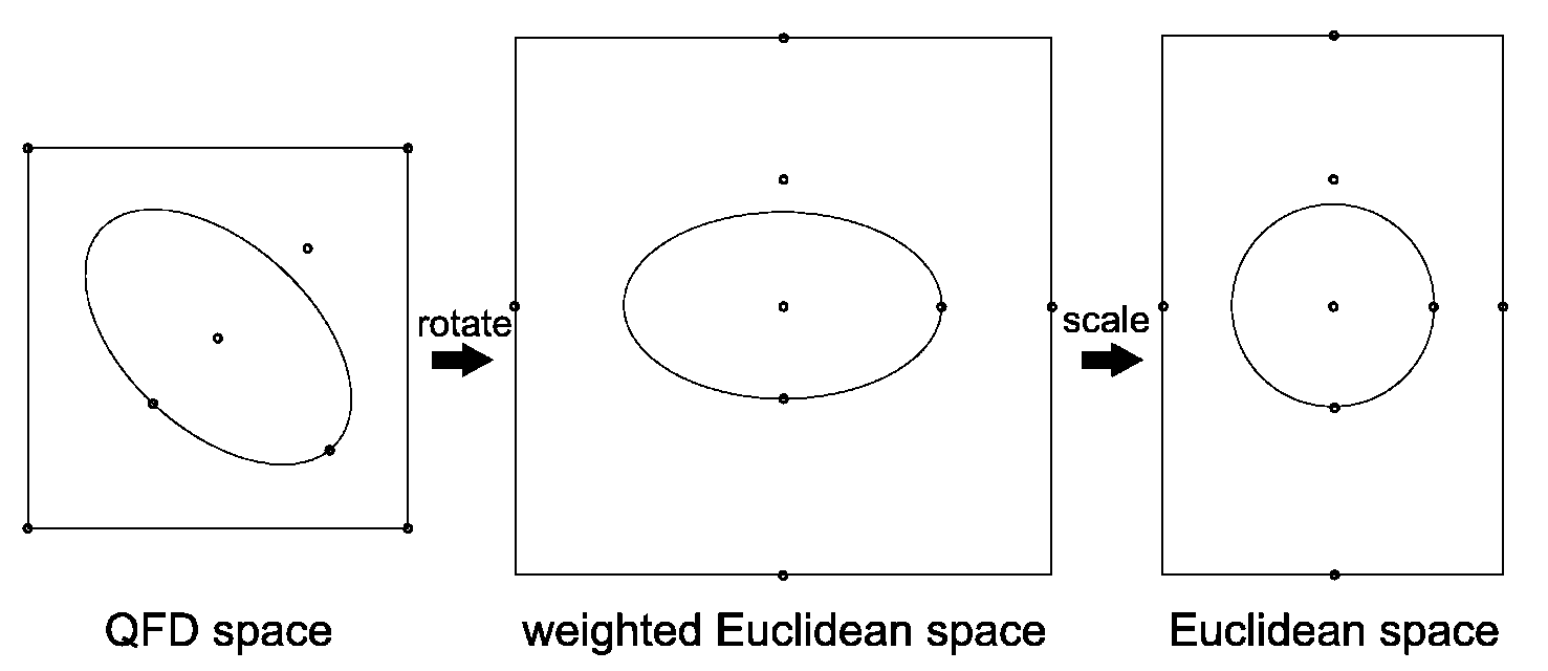
QFD	Euclidean (L_2) distance
Correlated dimensions ✔	Independent dimensions ✘
Expensive – $O(n^2)$ ✘	Cheap – $O(n)$ ✔

QMap model

- Transform QFD space
  - $\mathbf{L_2}$ instead of **QFD**
  - Preserving distances (homeomorphism)

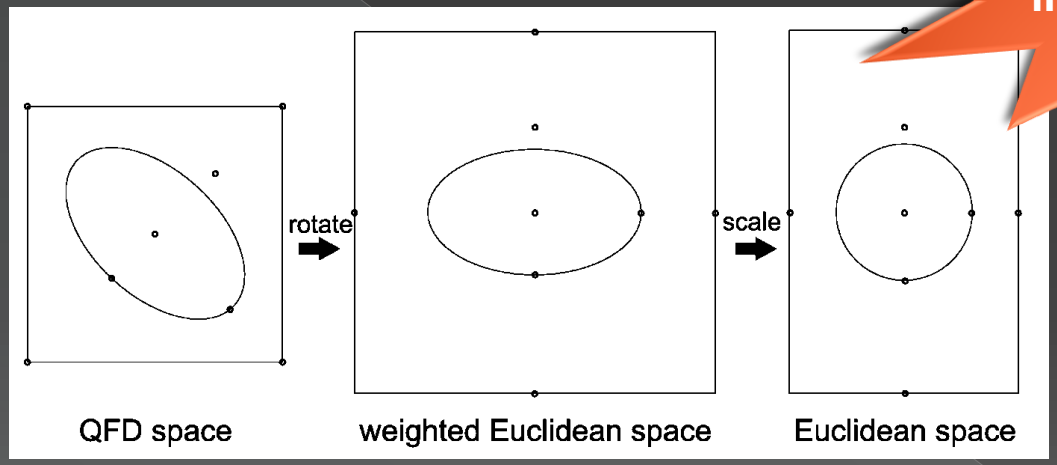
QMap model

- Transform QFD space
  - L_2 instead of **QFD**
  - Preserving distances (homeomorphism)



QMap model

- **Rotate** (weighted L_2 space)
  Scale (L_2 space)



Linear transformation

- Transformation matrix **B**
  - obtained by <u>Cholesky decomposition:</u>
    $\mathbf{A} = \mathbf{B}\mathbf{B}^T$

QFD to L_2 Space Transformation

1. $QFD_A(\mathbf{u},\mathbf{v}) = \sqrt{(\mathbf{u}-\mathbf{v})\mathbf{A}(\mathbf{u}-\mathbf{v})^T}$
2. Cholesky decomposition: $\mathbf{B}\mathbf{B}^T = \mathbf{A}$
3. $QFD(\mathbf{u},\mathbf{v}) = \sqrt{(\mathbf{u}-\mathbf{v})\mathbf{B}\mathbf{B}^T(\mathbf{u}-\mathbf{v})^T}$
4. $QFD(\mathbf{u},\mathbf{v}) = \sqrt{[(\mathbf{u}-\mathbf{v})\mathbf{B}][(\mathbf{u}-\mathbf{v})\mathbf{B}]^T}$
5. $QFD(\mathbf{u},\mathbf{v}) = \sqrt{(\mathbf{u}\mathbf{B}-\mathbf{v}\mathbf{B})(\mathbf{u}\mathbf{B}-\mathbf{v}\mathbf{B})^T}$
6. $L_2(\mathbf{u}',\mathbf{v}') = \sqrt{(\mathbf{u}'-\mathbf{v}')(\mathbf{u}'-\mathbf{v}')^T}$

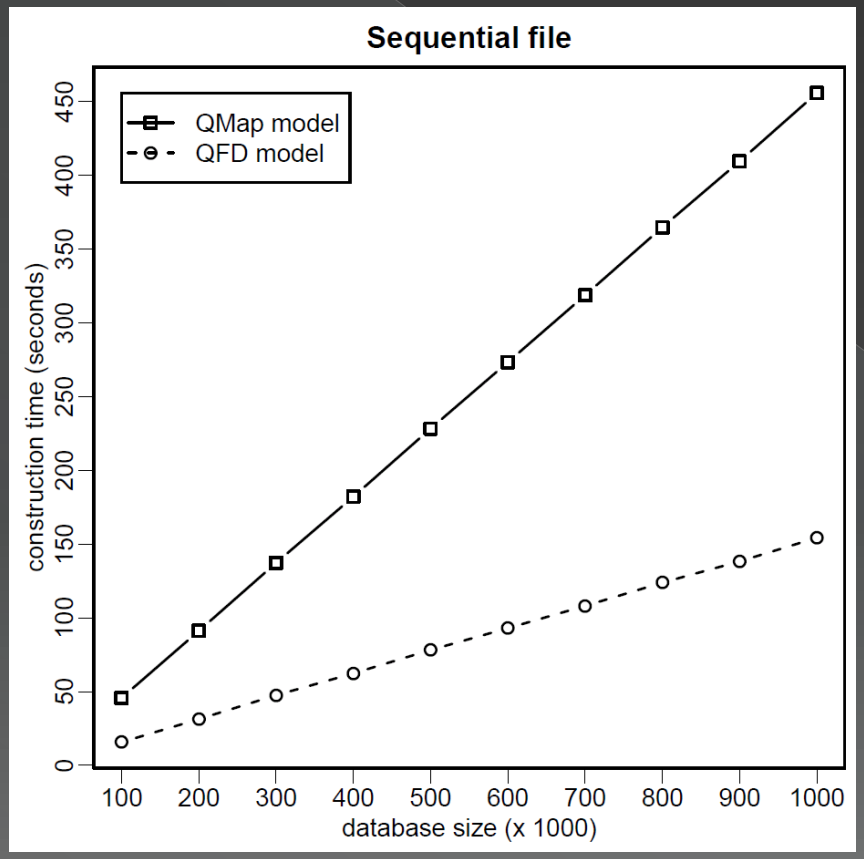
Experiments

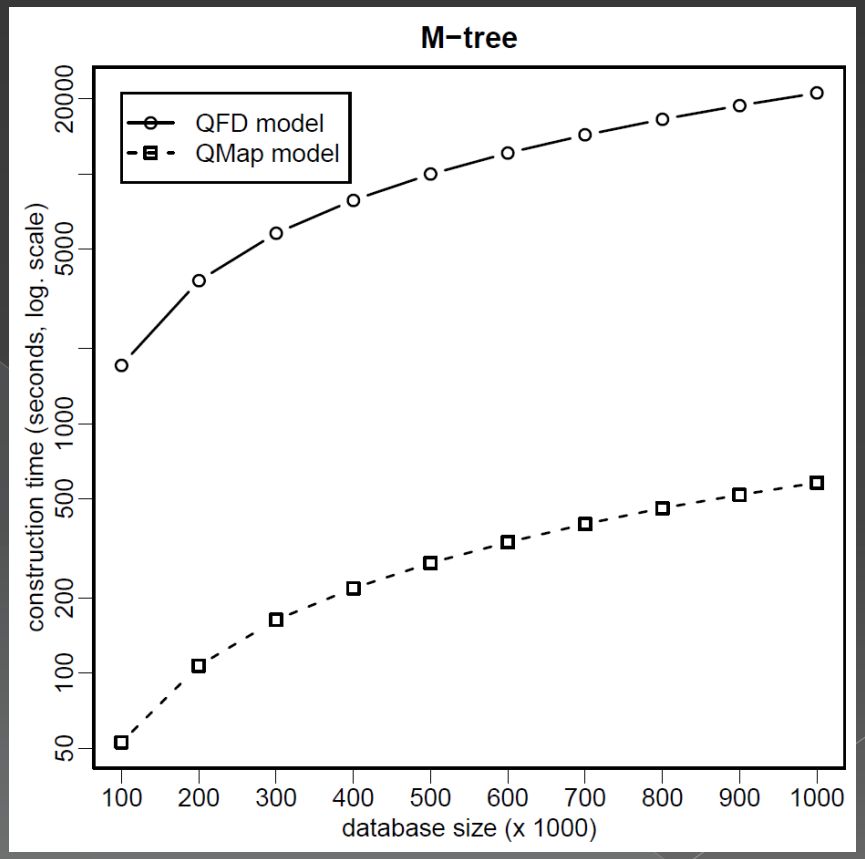
- Application of QMap in MAM
  - Sequential (SEQ) file
  - Pivot Table
  - M-tree
- **1,000,000** images (**512** dimensional RGB histogram)
- Actions
  - Indexing
  - Querying

Indexing - time complexity results

Method (model)	"Winner"
SEQ file (QFD) SEQ file (QMap)	**QFD** ✗
Pivot Table (QFD) Pivot Table(QMap)	**QMap** ✓
M-tree (QFD) M-tree (QMap)	**QMap** ✓

Indexing - Experiments

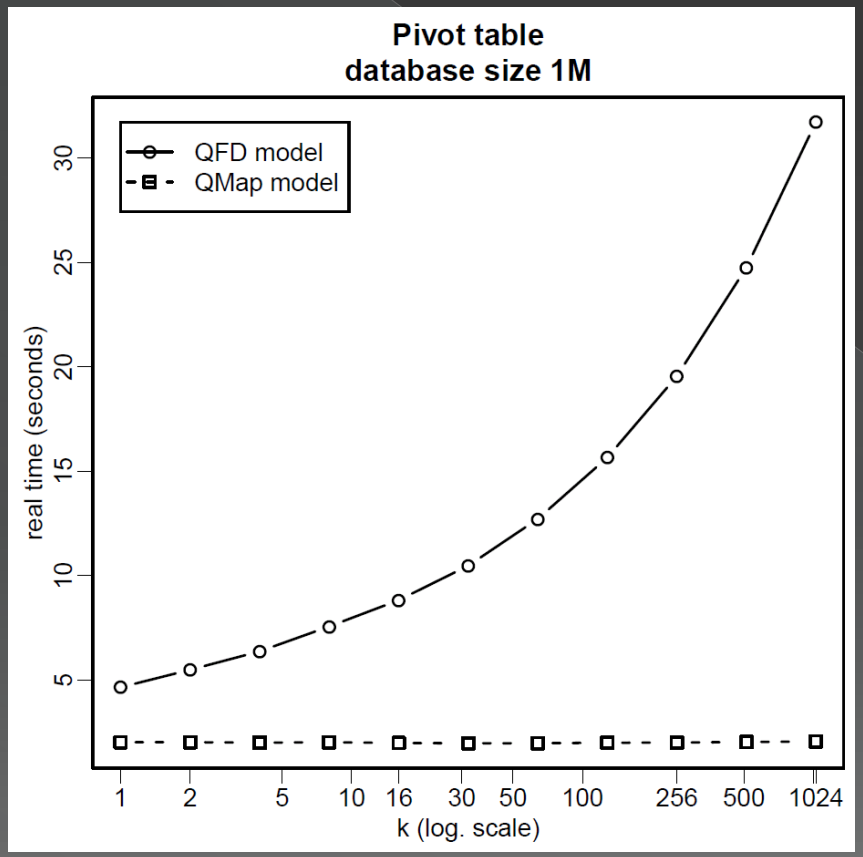


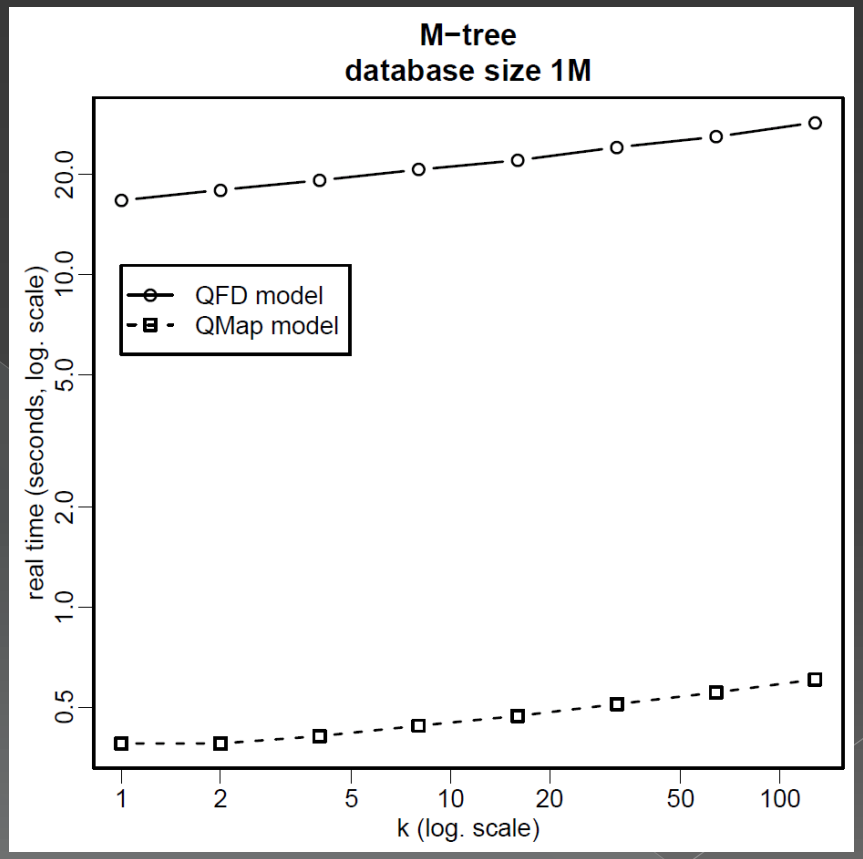


Querying - time complexity results

Method (model)	"Winner"
SEQ file (QFD) SEQ file (QMap)	**QMap** ✔
Pivot Table (QFD) Pivot Table(QMap)	**QMap** ✔
M-tree (QFD) M-tree (QMap)	**QMap** ✔

Querying- Experiments (kNN)





Contributions

- QMap model
  - Space transformation: **QFD** $\longrightarrow$ $\mathbf{L_2}$
  - Distance-preserving (homeomorphic)
  - Data-independent
  - Output is explicitly formulated
- QMap model is **separated** from the usage of any access methods
  - Superior performance

Thank you for your attention.

SIRET Research Group

http://siret.ms.mff.cuni.cz