Empirical Evaluation of Dissimilarity Measures for 3D Object Retrieval with Application to Multi-Feature Retrieval

Robert Gregor, Andreas Lamprecht
Ivan Sipiran, Tobias Schreck
University of Konstanz, Germany

Benjamin Bustos
Department of Computer Science
University of Chile, Chile

***Abstract*—A common approach for implementing content-based multimedia retrieval tasks resorts to extracting high-dimensional feature vectors from the multimedia objects. In combination with an appropriate dissimilarity function, such as the well-known L_p functions or statistical measures like χ^2, one can rank objects by dissimilarity with respect to a query. For many multimedia retrieval problems, a large number of feature extraction methods have been proposed and experimentally evaluated for their effectiveness. Much less work has been done to systematically study the impact of the choice of dissimilarity function on the retrieval effectiveness.**

Inspired by previous work which compared dissimilarity functions for image retrieval, we provide an extensive comparison of dissimilarity measures for 3D object retrieval. Our study is based on an encompassing set of feature extractors, dissimilarity measures and benchmark data sets. We identify the best performing dissimilarity measures and in turn identify dependencies between well-performing dissimilarity measures and types of 3D features. Based on these findings, we show that the effectiveness of 3D retrieval can be improved by a feature-dependent measure choice. In addition, we apply different normalization schemes to the dissimilarity distributions in order to show improved retrieval effectiveness for late fusion of multi-feature combination. Finally, we present preliminary findings on the correlation of rankings for dissimilarity measures, which could be exploited for further improvement of retrieval effectiveness for single features as well as combinations.

I. Introduction

Content-based indexing approaches are a requirement to cope with large amounts of multimedia documents. They support retrieval and cluster analysis applications based on the similarity between documents. Feature-based approaches are popular to this end due to their often simple and robust implementations. Also, feature vectors can be used with spatial index structures for scalable similarity search. However, features typically only encode very low-level properties of the multimedia objects. Therefore, careful experimentation is needed to determine suitable features types, extraction parameters and the dissimilarity measure used for comparing them. This is usually done by benchmark data sets which define test documents and similarity judgments (i.e. ground truth). While to date, many different feature types have been proposed for multimedia retrieval, the choice of dissimilarity functions has been researched to substantially less extent. Few systematic studies on the effect of the dissimilarity function for content-based retrieval have been conducted. One example is [20], where a set of dissimilarity functions including L_p norms and information-theoretic measures have been evaluated. The study showed that the choice of dissimilarity has an impact on the effectiveness of retrieval, classification and segmentation tasks and that the choice of the best function also depends on properties of the data. Motivated by this previous study, we experimentally compare the effectiveness of a larger number of dissimilarity measures for 3D shape retrieval tasks. Based on our study involving several data sets and different types of 3D features, we also find that the choice of dissimilarity function significantly affects the effectiveness of 3D Retrieval and depends on the features type it is applied to. Extending this experimental setup, we subsequently study the impact of different dissimilarity measures and several distance normalization schemes when multiple features are aggregated to compute similarity. Our findings can be used to improve the retrieval effectiveness of multi-feature retrieval systems.

The remainder of this paper is structured as follows. In Section II we describe related work on feature-based multimedia retrieval and evaluation. In Section III we state the research questions of this study and the experimental setup it motivates. Subsequently, we present and assess our findings in Section IV. In Section V we discuss limitations of the study and, following a side result of our study, outline a new technique for rank-based aggregation. Finally, Section VI concludes.

II. Background and Related Work

Algorithms for multimedia retrieval have been developed for different types of multimedia data, for example images [10], music/motion [18], video [17], and 3D data [23]. A common approach for tackling this problem is to represent each multimedia document with one or several high-dimensional feature vectors, which are constructed by analyzing the content of the multimedia document. If the query is also a multimedia document (query-by-example), its feature vectors are extracted and then matched with those from the repository. For convenience, a dissimilarity measure can be defined as a distance between the feature vectors extracted from the documents, effectively implementing the multimedia retrieval task as a k-nearest neighbor search.

In the area of 3D shape retrieval, this approach has been used extensively. Initially, feature vectors were designed to globally describe the 3D shape [8] and usually compared by

978-1-4673-6870-4/15/$31.00 © 2015 IEEE

L_1 or L_2. More recently, a large amount of research has been concentrated on defining local features for 3D models, which take into account local shape characteristics for computing several feature vectors for each model. Relying on these, more complex retrieval tasks like part-to-whole similarity search in 3D objects, where the goal is to find a partial matching between the query and the 3D repository [22], have been proposed.

Techniques, such as Bag of Features(BoF), Fisher Vector Encoding or Vector of Locally Aggregated Descriptors[4] see increasing adoption in various fields of multimedia retrieval to encode a single global feature vector from a set of such local features. While there have been studies that assess the effectiveness of such encoding techniques (e.g. for image retrieval [9]), a choice beyond L_1 and L_2 has not been subject to research.

Orthogonally, given the number of different 3D feature types that have been proposed along with their specific advantages and limitations, the question arises, how multiple feature types could be combined to obtain more effective retrieval methods. Studies have shown that combinations can improve overall retrieval effectiveness, e.g., when combining features of different kinds (e.g., view- and extent-based features [24]) or locality (e.g., global and local features [21]). Features of different type can be combined in several ways, including concatenation of individual feature vectors to form a larger vector [24], possibly followed by a dimensionality reduction step like Principal Components Analysis [14] or feature selection [16]. Also, multiple features can be considered for retrieval by aggregating the rankings which each of the original feature vectors gives aggregation of ranks [11]. Feature combination and rank aggregation may also be useful to adapt the similarity search to specific preferences of a given user in relevance feedback-type schemes [5]. However in the work mentioned above, the impact of the choice of dissimilarity measure has not been subject to systematic evaluation.

At the same time, more foundational research on the behavior of L_p dissimilarity measures in high dimensional spaces[3], [12] implies the question if L_1 and L_2 could be outperformed by other dissimilarity measures in applications such as 3D retrieval. To address this gap, we conduct a first systematic study on dissimilarity measures within the context of 3D shape retrieval, that takes into account fractional distances as well as information theoretic dissimilarity measures.

III. Experimental Setup

In this section, we describe the main research questions that we address in this work and in turn detail the test data and methodology we devised to answer the questions.

A. Problem Statement

Our experiments are motivated by the following problems:

P_1 How do different dissimilarity measures compare to each other in terms of effectiveness of 3D retrieval? Is there empirical indication to consider other measures than the ubiquitous L_1 and L_2?

P_2 Is there any dependency between the measures and the types of data or features on the other hand?

P_3 Are there any patterns, that we could then exploit to improve the effectiveness of a multi-feature retrieval system?

B. 3D Features

For our study, we consider a range of 3D shape features, which characterize the global shape of a 3D object by certain low-level measurements. The features can be grouped as follows (please refer to the abbreviations in the later results and to [7] for details):

View-based features are obtained from 2D images of the 3D objects. They comprise features from depth-buffer images (DBF) and from silhouette images (SIL).

Surface-based features characterize measurements obtained from the object surface. They comprise SD2 (histogram of distances between surface point pairs), and GRAY (measures of surface normals). *Extent-based* features describe the normalized spatial extend of a 3D object. They comprise COR (distances of surface triangles from the center of mass), as well as RSH and IRAY (distances of surface point samples from center of mass, represented in the frequency or spatial domain, respectively).

Volume-based features are computed from a volumetric representation of an object. They comprise RIN and VOX (aggregates of a Voxel-based object representation), VOL (object volume measured along radial partitioning scheme), 3DDFT (VOX in the frequency domain), and H3D (object volume in spherical harmonics representation [13]).

Moments-based features refer to statistical moments over shape measurements. They include RMOM (moments of IRAY measures) and PMOM (moments of mesh triangle centroids).

Hybrid features combine several basic features in a joint feature vector. They comprise CPX (combining RSH and GRAY), and DSR (combining DBF, SIL and RSH).

According to their encoding, we can group SD2, COR, RI, RMOM, PMOM as histograms- or distributional measures and GRAY, IRAY, VOX, VOL as feature vectors that show locality of the feature vector dimensions, where nearby dimensions represent measures from close shape points. Note that this selection of global shape features is not complete regarding the large amount of 3D global shape descriptors proposed in the literature. However, it includes features from main feature categories often used in 3D retrieval (i.e., view-, surface-, extent- and volume-based ones).

C. Datasets

For our experiments we used three standard datasets for evaluation of 3D retrieval tasks. The *Princeton Shape Benchmark* (PSB) [2] consists of generic 3D shapes collected from the web, separated into 907 training and 902 test objects categorized into 90 classes. We only computed descriptors for the training objects (note that all of our descriptors do not require training). The *Engineering Shape Benchmark* (ESB) [15] consists of 867 mechanical engineering shapes categorized into 42 classes. The *Konstanz Shape Benchmark* (KNDB) [8] consists of 1838 generic shapes collected from the web. 472 of which are categorized into 55 classes.

D. Dissimilarity Measures

Table I lists the definitions of the measures used in our experiments. In the following we refer to the abbreviations introduced there. To test Minkowski-like L_p measures, we use

Measure	Definition
Minkowski	$L_p(x,y) = \left(\sum_i \lvert x_i - y_i \rvert^p \right)^{1/p}$
Chi square	$\chi^2(x,y) = \frac{1}{2} \sum_i \frac{(x_i - y_i)^2}{(x_i + y_i)}$
Kullback-Leibler Divergence	$KL(x,y) = \sum_i x_i \log \frac{x_i}{y_i}$
Jeffrey Divergence	$JD(x,y) = KL(x,y) + KL(y,x)$
Jensen-Shannon Divergence	$JSD(x,y) = KL(x, \frac{1}{2}(x+y)) + KL(y, \frac{1}{2}(x+y))$
Quadratic Chi	$QC_m^A(x,y) = \sqrt{\sum_{ij} p_i \times q_j \times A_{ij}}$ $p_i = \left(\frac{(x_i - y_i)}{(\sum_c (x_c - y_c) A_{ci})^m} \right)$ $q_j = \left(\frac{(x_j - y_j)}{(\sum_c (x_c - y_c) A_{cj})^m} \right)$
Kolmogorov Smirnov Test	$KS(x,y) = \min_i(\lvert \hat{x_i} - \hat{y_i} \rvert), \hat{x_i} = \sum_{j \leq i} x_i$
Cramer von Mises Criterion	$CvM(x,y) = \sum_i (\hat{x_i} - \hat{y_i})^2, \hat{x_i} = \sum_{j \leq i} x_i$
Weighted Mean Variance	$WMV(x,y) = \frac{\lvert \mu(x) - \mu(y) \rvert}{\lvert \sigma(\mu(x), \mu(y)) \rvert} + \frac{\lvert \sigma(x) - \sigma(y) \rvert}{\lvert \sigma(\sigma(x), \sigma(y)) \rvert}$ $\mu(x) = \frac{1}{n} \sum_i^n x_i \qquad \sigma(x) = \sqrt{\frac{1}{n} \sum_i^n (x_i - \mu(x))^2}$

TABLE I. THE SET OF DISSIMILARITY MEASURES USED IN OUR EXPERIMENTS.

$p \in \{0.1, 0.3, 0.6, 0.9\}$ to obtain fractional distances (i.e. semi-metrics) and $p \in \{1, 2\}$ to obtain Manhattan and Euclidean distance (i.e. regular metrics). A detailed description of most of the dissimilarity measures can be found in [20]. Definition and evaluation of Quadratic Chi (QC), which is intended for histograms, can be found in [19]. It is worth mentioning that there, QC outperforms L_1, L_2, JSD, JD, KL and even Earth Movers Distance (EMD) when applied to BoF descriptors based on local 2D image features (i.e. SIFT). Note that EMD is not included in our comparison due to its exponential time complexity. When assuming constant time computation of log and power functions, all of the tested measures share $O(n)$ time complexity, except for QC, KS and CvM with $O(n^2)$, where n denotes feature dimensionality.

E. Performance Measure and Result Visualization

We determine the effectiveness of different 3D features and dissimilarity measures by *R-precision* (RP, also called first-tier precision) measure [5], [2], $RP \in [0, \dots, 1]$. It gives the precision of a nearest neighbor ranking $R(q)$ of size $r = |R(q)|$ with respect to a query q, where r is equal to the number of objects in the benchmark relevant to q. This measure typically correlates strongly with similar measures of precision in Information Retrieval like average precision or harmonic mean. In the following, we compare the RP results using tables per benchmark, showing the obtained measures for features (rows) and dissimilarity measures (columns).

F. Combinations of Features

From previous work, it is well known that relying in multiple features can substantially improve retrieval effectiveness over individual features. Hence after identifying the performance for individual features and dissimilarity measures, we test multi-feature combinations. To fuse multiple features, there are many strategies (see sec II and V). A simple approach is direct aggregation of the individual distances (i.e. late fusion). However, this approach is sensitive to different distribution characteristics in the respective feature spaces. Furthermore, it is well known (*"Curse of Dimensionality"*[3][12]), that the distribution of the distances, obtained by a dissimilarity measure, can be heavily dependent on the dimensionality of the feature space, which usually varies between different feature types. Both effects can, to limited extent, be addressed by normalizing the distances before aggregation. We test aggregation with several normalization schemes. Let f_i denote individual features in feature set F, which contains features of same type extracted for each object in a benchmark. Let $d : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ be a dissimilarity measure, where $d_{i,j}$ is the distance of f_i to f_j and $D_F = \{d_{i,j} : \forall f_i, f_j \in F \wedge i \neq j\}$ the set of all pairwise distances with respect to d and F:

- *mean*-normalization: $d_{i,j}^{\mu} = \frac{d_{i,j}}{\mu(D_F)}$
- *max*-normalization: $d_{i,j}^{max} = \frac{d_{i,j}}{max(D_F)}$
- α-normalization: $d_{i,j}^{\alpha} = \frac{d_{i,j}}{\tau_{\alpha,D_F}}$ with α set to 0.1 and $F_{D_F}(\tau_{\alpha,D_F}) = \alpha$, where F_{D_F} denotes the cumulative distribution over all $d_{i,j} \in D_F$. Details in [6].

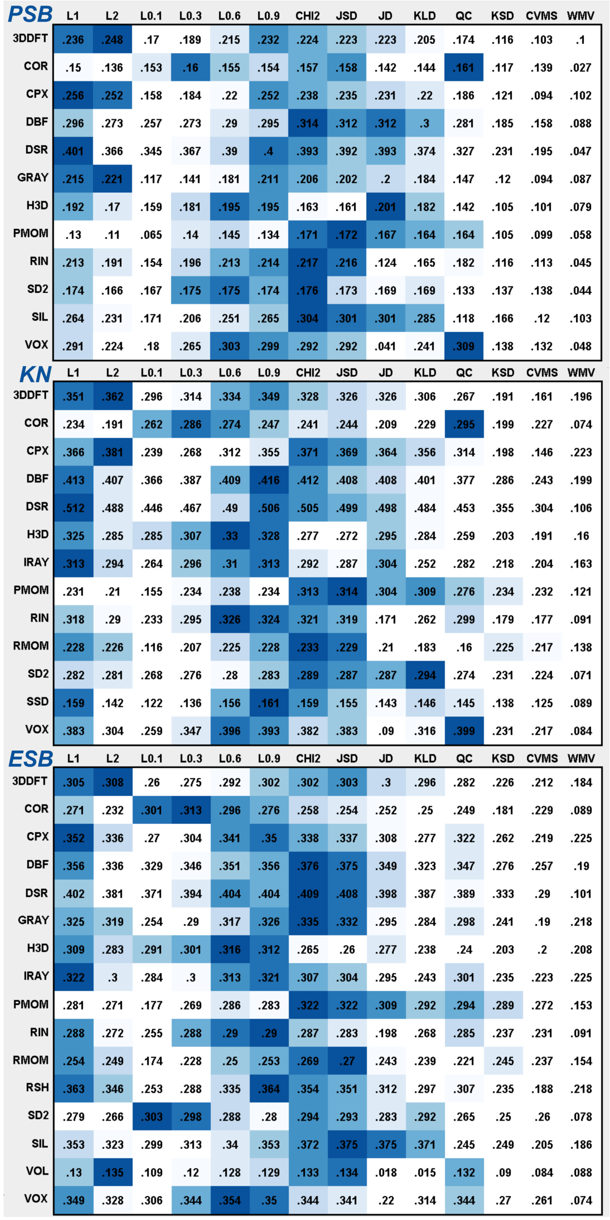
So far, combination schemes usually fuse distances obtained from applying a single metric over all features. We test combinations of 3-6 feature vectors (from PSB and KNDB) for each metric and compare this to combinations relying on the respective best metric per feature vector as identified according to P_1 and P_2. For both, normalization is applied before aggregation.

IV. RESULTS

Fig. 1 shows the results for single features in combination with each of the dissimilarity measures whereas results for multi-feature retrieval are provided by Fig. 2. Overall α- and max-normalization outperformed mean-normalization by a magnitude of 10^{-2} in R-precision. The differences between α- and max-normalization appear to be data dependent and are mostly in the magnitude of 10^{-3} and below. Overall, α-normalization performs better for KNDB and worse for PSB. There are no consistent patterns with respect to measures, except for (KLD), where for both benchmarks and in the majority of cases, α-normalization leads to better results by about .02 to .03. We provide R-precision values based on α-normalization here. Detailed comparisons can be found in the supplementary material [1]. In the following, we interpret the results in terms of the research questions as formulated in Section III-A.

P_1: Comparison of similarity metrics

L_1 can be considered a robust and among the overall best choices concerning single feature retrieval. However in only 19 of the total 41 test cases, it is among the top 3 scoring measures for an individual feature. Furthermore, if we compare all measures by the number of achieved top rankings over the 41 feature/benchmark combinations, L_1 (6) is preceded by χ^2 (9), on par with L_2 and followed by JSD, $L_{0.9}$, $L_{0.6}$ and QC (4). In pairwise comparison, L_1 is outperformed by $L_{0.9}$ in 23, by χ^2 in 21 and by Jensen-Shannon Divergence (JSD)



PSB	L1	L2	L0.1	L0.3	L0.6	L0.9	CHI2	JSD	JD	KLD	QC	KSD	CVMS	WMV
3DDFT	.236	.248	.17	.189	.215	.232	.224	.223	.223	.205	.174	.116	.103	.1
COR	.15	.136	.153	.16	.155	.154	.157	.158	.142	.144	.161	.117	.139	.027
CPX	.256	.252	.158	.184	.22	.252	.238	.235	.231	.22	.186	.121	.094	.102
DBF	.296	.273	.257	.273	.29	.295	.314	.312	.312	.3	.281	.185	.158	.088
DSR	.401	.366	.345	.367	.39	.4	.393	.392	.393	.374	.327	.231	.195	.047
GRAY	.215	.221	.117	.141	.181	.211	.206	.202	.2	.184	.147	.12	.094	.087
H3D	.192	.17	.159	.181	.195	.195	.163	.161	.201	.182	.142	.105	.101	.079
PMOM	.13	.11	.065	.14	.145	.134	.171	.172	.167	.164	.164	.105	.099	.058
RIN	.213	.191	.154	.196	.213	.214	.217	.216	.124	.165	.182	.116	.113	.045
SD2	.174	.166	.167	.175	.175	.174	.176	.173	.169	.169	.133	.137	.138	.044
SIL	.264	.231	.171	.206	.251	.265	.304	.301	.301	.285	.118	.166	.12	.103
VOX	.291	.224	.18	.265	.303	.299	.292	.292	.041	.241	.309	.138	.132	.048

KN	L1	L2	L0.1	L0.3	L0.6	L0.9	CHI2	JSD	JD	KLD	QC	KSD	CVMS	WMV
3DDFT	.351	.362	.296	.314	.334	.349	.328	.326	.326	.306	.267	.191	.161	.196
COR	.234	.191	.262	.286	.274	.247	.241	.244	.209	.229	.295	.199	.227	.074
CPX	.366	.381	.239	.268	.312	.355	.371	.369	.364	.356	.314	.198	.146	.223
DBF	.413	.407	.366	.387	.409	.416	.412	.408	.408	.401	.377	.286	.243	.199
DSR	.512	.488	.446	.467	.49	.506	.505	.499	.498	.484	.453	.355	.304	.106
H3D	.325	.285	.285	.307	.33	.328	.277	.272	.295	.284	.259	.203	.191	.16
IRAY	.313	.294	.264	.296	.31	.313	.292	.287	.304	.252	.282	.218	.204	.163
PMOM	.231	.21	.155	.234	.238	.234	.313	.314	.304	.309	.276	.234	.232	.121
RIN	.318	.29	.233	.295	.326	.324	.321	.319	.171	.262	.299	.179	.177	.091
RMOM	.228	.226	.116	.207	.225	.228	.233	.229	.21	.183	.16	.225	.217	.138
SD2	.282	.281	.268	.276	.28	.283	.289	.287	.287	.294	.274	.231	.224	.071
SSD	.159	.142	.122	.136	.156	.161	.159	.155	.143	.146	.145	.138	.125	.089
VOX	.383	.304	.259	.347	.396	.393	.382	.383	.09	.316	.399	.231	.217	.084

ESB	L1	L2	L0.1	L0.3	L0.6	L0.9	CHI2	JSD	JD	KLD	QC	KSD	CVMS	WMV
3DDFT	.305	.308	.26	.275	.292	.302	.302	.303	.3	.296	.282	.226	.212	.184
COR	.271	.232	.301	.313	.296	.276	.258	.254	.252	.25	.249	.181	.229	.089
CPX	.352	.336	.27	.304	.341	.35	.338	.337	.308	.277	.322	.262	.219	.225
DBF	.356	.336	.329	.346	.351	.356	.376	.375	.349	.323	.347	.276	.257	.19
DSR	.402	.381	.371	.394	.404	.404	.409	.408	.398	.387	.389	.333	.29	.101
GRAY	.325	.319	.254	.29	.317	.326	.335	.332	.295	.284	.298	.241	.19	.218
H3D	.309	.283	.291	.301	.316	.312	.265	.26	.277	.238	.24	.203	.2	.208
IRAY	.322	.3	.284	.3	.313	.321	.307	.304	.295	.243	.301	.235	.223	.225
PMOM	.281	.271	.177	.269	.286	.283	.322	.322	.309	.292	.294	.289	.272	.153
RIN	.288	.272	.255	.288	.29	.29	.287	.283	.198	.268	.285	.237	.231	.091
RMOM	.254	.249	.174	.228	.25	.253	.269	.27	.243	.239	.221	.245	.237	.154
RSH	.363	.346	.253	.288	.335	.364	.354	.351	.312	.297	.307	.235	.188	.218
SD2	.279	.266	.303	.298	.288	.28	.294	.293	.283	.292	.265	.25	.26	.078
SIL	.353	.323	.299	.313	.34	.353	.372	.375	.375	.371	.245	.249	.205	.186
VOL	.13	.135	.109	.12	.128	.129	.133	.134	.018	.015	.132	.09	.084	.088
VOX	.349	.328	.306	.344	.354	.35	.344	.341	.22	.314	.344	.27	.261	.074

Fig. 1. Comparison of average R-precision for a selection of global 3D Shape descriptors and the PSB, KNDB and ESB benchmarks. Rows correspond to feature vector types, columns to metrics. For each row, the best performing measure is color coded with dark blue. Color coding transitions to white over the next 6 best performing measures.

in 18 cases. Thus, for our tests, we can state that χ^2 and $L_{0.9}$ have an equal if not better robustness than L_1 across varying features and data. The robustness for L_2 is not among the top 5, in direct comparison, it falls behind L_1 (34), $L_{0.9}$ (33), χ^2 (30), JSD (30) and $L_{0.6}$ in 22 cases. Note that for R-precision the groups of χ^2,JSD and $L_1, L_{0.9}, L_{0.6}$ show high correlation over all features and data sets. In conclusion our tests indicate, that, without taking into account specific knowledge about the data set or feature type, it is worthwhile to consider other metrics besides L_1 and especially L_2. From our set of dissimilarity measures, in particular χ^2, $L_{0.9}$ and JSD show very good results. In test cases where Quadratic Chi (QC) performs best, either $L_{0.6}$ or $L_{0.3}$ are only marginally worse, but have lower time complexity. Choosing p within $(0.3, 0.6)$ could lead to equal performance for this test cases. Due to the overall low R-precision for KLD, KSD, $CVMS$ and WMV, they will not be discussed in the following sections.

P_2: Dependency between metrics, feature types and dataset

Concerning overall dataset dependencies, the top measures identified in the previous section show consistent in feature specific behavior over PSB and KN dataset and to slightly lesser extent also for ESB. The information theoretic measures χ^2 and JSD perform well for most of the histogram features such as PMOM, RIN, SD2, which often perform sub par with L_1. For KN, where categories are more easily separated (overall higher R-precision), L_1 is roughly on par for several of those features.

According to [12], L_p measures with larger p perform better with growing equality of variance among the individual dimensions within a feature set. Apparently, this is backed by L_2 performing best for 3DDFT and VOL. Both, the Fourier-Transform(FT) and the radial partitioning scheme are prone to equally distributed aliasing noise over all frequencies or partitions when features are extracted from slightly different objects. However, this is contradicted by the performance of the FT based DBF and SIL. L_2 is among the worst performing L_p measures for features where slight changes in objects do not affect multiple dimensions equally such as COR or VOX (i.e. 3DDFT without FT). Note that there are almost no features where L_2 and lower fractional distances are both among the top performing measures.

Concerning [3], the Curse of Dimensionality does not seem to affect larger p with major impact. The high-dimensional VOL(486) and DSR(472), do not generally perform worse with L_2, the low dimensional PMOM(52) and COR(30) are not better with lower fractional distances. The top score of $L_{0.1}$ for SD2 in ESB could result out of the often very heterogeneous distribution of vertices over the surface of CAD models. In the case of ESB, vertex sampling density on the surface is likely characteristic to some of the categories so that respective SD2 features reflect the category by very sparsly distributed peaks in the histogram at varying indices. QC is best for COR and VOX features of PSB and KN. Both COR an VOX have strongly expressed localities at adjacent or identical indices. Note that lower fractional distances show good performance here as well.

P_3: Performance of combined distance metrics

From the previous, we have seen that while on average, L_1 is among the overall best metrics, it is often outperformed by other measures for specific features. Referring to Fig. 2, the results show dependence to dataset, the number as well as the types of features combined. Over both datasets and almost all cases, top performance is either achieved by fusion of the individual best measures (26), χ^2 (17), or JD (10), whereas L_1 trails with only 4 top scores. However it often does not fall behind by large margin. Again $L0.9$ shows similar behavior to $L1$ but is almost consistently outperformed in direct comparison. While JSD performs close to χ^2 for the combinations tested on the PSB dataset (similar to the results for single features), it does not perform well for a larger number of combined features and most other combination tested over KNDB. Surprisingly JD, which overall is clearly outperformed for single features, now often outperforms JSD.

PSB	BEST	L1	L2	L0.1	L0.3	L0.6	L0.9	CHI2	JSD	JD	KLD	QC	KSD	CVMS	WMV
CPX/DBF/DSR/SIL/VOX	.4129	.4099	.3696	.3462	.386	.4056	.4104	.4211	.4177	.3695	.3985	.385	.3029	.2535	.1701
CPX/DBF/DSR/VOX	.41	.4085	.3718	.338	.3843	.4061	.4068	.4163	.4125	.3393	.3978	.4034	.2948	.251	.1483
DBF/DSR/SIL/VOX	.4089	.4036	.3608	.3508	.3927	.4075	.406	.4192	.4157	.3431	.3909	.3776	.2919	.2505	.1325
CPX/DSR/SIL/VOX	.4094	.402	.3611	.3209	.3736	.4	.4019	.41	.4068	.3324	.3852	.3586	.2753	.2285	.1518
CPX/DBF/SIL/VOX	.4066	.3956	.3524	.3155	.366	.3865	.3926	.4125	.4081	.3367	.3871	.3743	.2907	.241	.1857
CPX/DBF/VOX	.3973	.384	.343	.2836	.3447	.3807	.3826	.3998	.396	.2864	.3729	.385	.2687	.227	.1568
DBF/SIL/VOX	.394	.3787	.3222	.3029	.3563	.3799	.3781	.3959	.3928	.2778	.3662	.356	.2547	.2112	.1355
CPX/SIL/VOX	.3938	.3824	.3336	.2674	.3334	.3731	.3805	.39	.3868	.2723	.3667	.3331	.2551	.2006	.1628
CPX/DSR/VOX	.4106	.4101	.3698	.3059	.3619	.3992	.4076	.3988	.3963	.2746	.3737	.3784	.2662	.2254	.1236
DSR/SIL/VOX	.4023	.4016	.349	.3277	.3787	.4052	.4041	.4114	.4076	.2871	.3801	.3564	.2612	.2259	.1057
DBF/DSR/VOX	.4077	.4111	.3741	.3508	.3951	.4152	.4118	.419	.4173	.3074	.392	.4092	.2854	.2508	.0975
CPX/DBF/DSR	.3805	.3756	.3493	.3278	.3452	.364	.3704	.3873	.3854	.3858	.3714	.3542	.2642	.2283	.1312
DBF/DSR/SIL	.3771	.367	.3388	.3336	.3487	.3626	.3653	.39	.3881	.3881	.3723	.3144	.2702	.231	.1248
CPX/DBF/DSR/SIL	.3919	.3784	.348	.3305	.346	.3677	.3751	.3988	.3968	.3968	.3825	.3346	.2808	.2281	.1678
CPX/DBF/SIL	.3788	.3576	.3259	.2994	.3177	.3403	.3507	.3872	.385	.384	.3734	.3153	.2651	.2052	.1953
CPX/DSR/SIL	.3752	.3641	.3345	.2937	.3189	.3481	.3611	.3817	.3769	.3772	.3622	.2861	.2427	.1921	.1463

KN	BEST	L1	L2	L0.1	L0.3	L0.6	L0.9	CHI2	JSD	JD	KLD	QC	KSD	CVMS	WMV
3DDFT/COR/CPX/DBF/H3D/IRAY	.4919	.4887	.4627	.4482	.4741	.4815	.4795	.4684	.4572	.4916	.4717	.4627	.4169	.3594	.2937
3DDFT/COR/CPX/DBF/H3D	.4904	.4821	.461	.4361	.4706	.4741	.4768	.478	.4699	.4845	.4629	.4548	.4029	.358	.2859
COR/CPX/DBF/H3D/IRAY	.4842	.4766	.4578	.4365	.4685	.4757	.4724	.4597	.4477	.4799	.4632	.4438	.3974	.3458	.278
COR/CPX/DBF/H3D	.4944	.4904	.4553	.4207	.459	.4813	.4821	.4609	.4502	.4784	.4511	.4371	.3835	.3388	.2737
3DDFT/COR/DBF/H3D/IRAY	.4826	.4776	.452	.4402	.4703	.4779	.4703	.4482	.4382	.467	.4501	.4398	.3749	.3327	.272
3DDFT/COR/CPX/H3D/IRAY	.4837	.4747	.4517	.4217	.4571	.4712	.4732	.444	.4355	.4684	.4511	.4345	.364	.3193	.284
CPX/DBF/H3D/IRAY	.4659	.4686	.4451	.4173	.4435	.4596	.4637	.4449	.4331	.4782	.4605	.4172	.3821	.3139	.3093
CPX/DBF/H3D	.4739	.4806	.4526	.4051	.4437	.4629	.473	.4599	.4508	.4828	.4549	.4163	.3599	.3067	.3093
3DDFT/CPX/DBF/H3D	.4762	.477	.4605	.4174	.448	.4613	.4696	.477	.4694	.4858	.4603	.4373	.3897	.3252	.3172
3DDFT/CPX/DBF/H3D/IRAY	.4715	.4706	.4552	.4329	.4557	.4642	.4684	.46	.4474	.4852	.4668	.4366	.4052	.3344	.3216
3DDFT/CPX/DBF/IRAY	.4462	.4478	.4344	.421	.4328	.4379	.443	.4476	.4416	.4721	.4581	.4376	.3893	.3249	.2996
CPX/DBF/IRAY	.4374	.4385	.4212	.3945	.4136	.4294	.4354	.4386	.4327	.4593	.4506	.4174	.369	.3046	.2893
3DDFT/COR/CPX/IRAY	.4525	.4428	.4195	.4046	.4293	.4396	.438	.4385	.4288	.4482	.4391	.4418	.354	.3101	.2542
3DDFT/COR/DBF/IRAY	.4625	.4529	.4248	.4271	.4452	.4475	.4491	.432	.4268	.4443	.4285	.439	.3545	.3195	.2313
3DDFT/COR/DBF/H3D	.4661	.4632	.4363	.4284	.4554	.4592	.4553	.4443	.4345	.4515	.4261	.4254	.3596	.3251	.2454
COR/DBF/H3D/IRAY	.4802	.4724	.4409	.4277	.4586	.47	.4649	.4311	.4166	.4531	.4322	.4184	.3609	.3239	.25
3DDFT/COR/CPX/H3D	.4747	.4712	.4412	.4056	.4376	.4609	.465	.4439	.4355	.4545	.4327	.413	.3411	.3049	.272
3DDFT/DBF/H3D	.4507	.4533	.4359	.4013	.4331	.4462	.4482	.4335	.4189	.4425	.4189	.3911	.3463	.3018	.2785
3DDFT/COR/H3D/IRAY	.4704	.4609	.4253	.4103	.4493	.4601	.4551	.4175	.4053	.4415	.4214	.4006	.3339	.3008	.2543
COR/CPX/H3D/IRAY	.4561	.4532	.4169	.4011	.4384	.4503	.4514	.4078	.3979	.4382	.4203	.3993	.3359	.311	.2444
COR/DBF/H3D	.4621	.4637	.4215	.4048	.4436	.4603	.4533	.4151	.4065	.4322	.4045	.402	.3303	.3087	.2233
COR/DBF/IRAY	.4618	.4462	.4079	.4057	.4355	.4469	.4419	.4074	.4008	.4124	.4037	.4186	.3267	.3103	.2064
3DDFT/COR/DBF	.4416	.4288	.4004	.4062	.431	.4315	.4239	.4323	.4288	.4193	.4064	.4339	.3333	.3146	.1951
3DDFT/COR/CPX	.4544	.4326	.4148	.373	.3984	.4193	.427	.4508	.4464	.4254	.4228	.4386	.3119	.2879	.2341
3DDFT/DBF/IRAY	.4324	.4311	.4126	.4048	.4142	.423	.4271	.4091	.4029	.4374	.4195	.3929	.328	.2812	.2753
3DDFT/CPX/IRAY	.4312	.4236	.4177	.3851	.4014	.4156	.4198	.4138	.405	.4445	.4264	.3951	.329	.2725	.2928
3DDFT/CPX/H3D	.4553	.4579	.4399	.3796	.4115	.4321	.4468	.4339	.4252	.4487	.4218	.3843	.3166	.2621	.3074
3DDFT/CPX/H3D/IRAY	.4571	.4552	.4385	.409	.4365	.4494	.4522	.4295	.4174	.4647	.4484	.4067	.3517	.2893	.3112
3DDFT/DBF/H3D/IRAY	.4674	.4644	.4419	.4194	.4478	.4574	.4602	.4237	.4136	.4588	.4369	.4077	.3623	.3041	.3016
DBF/H3D/IRAY	.461	.4529	.4307	.404	.4326	.4482	.4478	.4012	.3885	.4452	.4194	.376	.3419	.2913	.2766
3DDFT/H3D/IRAY	.4462	.4412	.4137	.3877	.4179	.4335	.4341	.3845	.3717	.4249	.4001	.3658	.3109	.2659	.2843
CPX/H3D/IRAY	.4216	.4275	.394	.3771	.4048	.42	.424	.3855	.3712	.4247	.4042	.3651	.3155	.2755	.2773
3DDFT/COR/H3D	.4337	.433	.3883	.3903	.4253	.4342	.4298	.3906	.3829	.407	.3951	.3747	.2912	.2735	.2257
3DDFT/COR/IRAY	.441	.4264	.3929	.3885	.4126	.4236	.4254	.3969	.3917	.4042	.397	.4024	.3022	.272	.2143
COR/CPX/IRAY	.4307	.4141	.3802	.3691	.3992	.4182	.4174	.395	.3887	.3972	.3911	.398	.3199	.2911	.214
COR/CPX/H3D	.4453	.441	.4022	.3609	.4044	.4316	.4346	.401	.3915	.4169	.3935	.3818	.2971	.2838	.2227
COR/H3D/IRAY	.4389	.4293	.3751	.3869	.4237	.43	.4225	.365	.3505	.3908	.375	.3624	.2989	.2876	.203
COR/CPX/DBF/IRAY	.4649	.4535	.4294	.4151	.4463	.4547	.4532	.4506	.4411	.4548	.445	.4487	.3874	.3392	.2481
3DDFT/COR/CPX/DBF/IRAY	.4649	.4616	.4377	.4348	.4513	.4543	.4551	.4615	.4544	.475	.4619	.467	.3996	.35	.2627
3DDFT/COR/CPX/DBF	.4651	.4601	.4375	.4223	.4446	.4552	.4533	.4803	.4763	.4688	.4536	.476	.393	.3484	.2484
COR/CPX/DBF	.464	.4583	.4328	.4066	.4354	.4524	.4553	.4723	.4681	.4543	.4374	.4714	.3623	.3272	.2356
3DDFT/CPX/DBF	.4468	.4537	.4385	.4071	.4185	.4356	.4476	.4788	.4761	.4765	.4667	.4527	.3741	.3184	.2959

Fig. 2. Comparison of average R-precision for multi-feature combinations. The top sections shows combinations over the PSB and the bottom section over the KN benchmark. Rows correspond to feature combinations, columns to dissimilarity measures. Note that the first column combines the individually best performing measures for each feature, whereas the remaining columns a uniform measure was fused.

When not considering the specifics of the individual combinations, late fusion of the best measures and L_1 can overall be considered the most robust choices for our test cases, while L_1 mostly trails the former. This indicates, that using individually best measures should overall be preferred over L_1. For specific combinations, fusion of either χ^2, JD and in rare cases also fractional distances could provide substantial improvement over combining the individually best measures.

V. Limitations and Extensions

We pragmatically considered a specific data type (3D objects) and a selection of global descriptors. Recently, local descriptors have come into focus and should be studied as well as global features that can be derived from them (see sec. II). However, given the characteristics of certain local features, our findings could serve as a first indication that likely in conjunction with different measures than L_1 or L_2 their retrieval performance could be improved.

We chose a range of representative dissimilarity measures with reasonably low time-complexity. It is worthwhile to note that in our tests, measures with $O(1)$ are on par or outperform the (smaller) set of $O(n)$ dissimilarity measures. However, the set of tested measures could be extended (e.g. Canberra and Bray-Curtis distance). Even though the performance margin between the top performing distance measures in our test is often small, their implementation effort is not high either. For other retrieval or kNN-search applications, even small improvements in effectiveness might also be worthwhile if they can be achieved with low effort.

We do not provide detailed results for any (inherently platform-specific) assessments of practical computation time. In early experiments with our mostly Java-based testing infrastructure, aggregated runtime of all pairwise distance computation for the $O(1)$ based measures was easily overshadowed by background activity, albeit carefully chosen JVM heap size, JIT, garbage collection parameters and an otherwise idle system.

For many retrieval applications in general, feature extraction for a single query object often easily overshadows the computational cost of determining feature dissimilarity to each object in the database. On the other hand, when the size of the database is very large, the efficiency of spatial indexing techniques should be taken into account as well (e.g. metric indexing or locality sensitive hashing). Also, we could only study a selection of the many approaches to feature combination. We focused mainly on distance-based aggregation in late fusion. We discarded rank-based aggregation, after initial results indicated significantly worse performance. While the tested DSR and CPX features are an example of early fusion, this set could clearly be extended. Beyond specific 3D Features, the evaluation methodology for feature type and data dependence could be improved. E.g. inspired by the results for P_2, we are convinced that determining correlation measures between statistical moments of a feature set (e.g. histograms of variance among each dimension) and relative dissimilarity measure performance could further improve our assessment. Alternatively a Bonferroni-Test could be applied. Additionally, analysis could be conducted separately for individual object categories. In turn this could provide useful indication of whether query object dependent choices of dissimilarity measure should be subject to further research. On the other hand z-score normalization,

Finally, as an additional experiment, we performed a visual

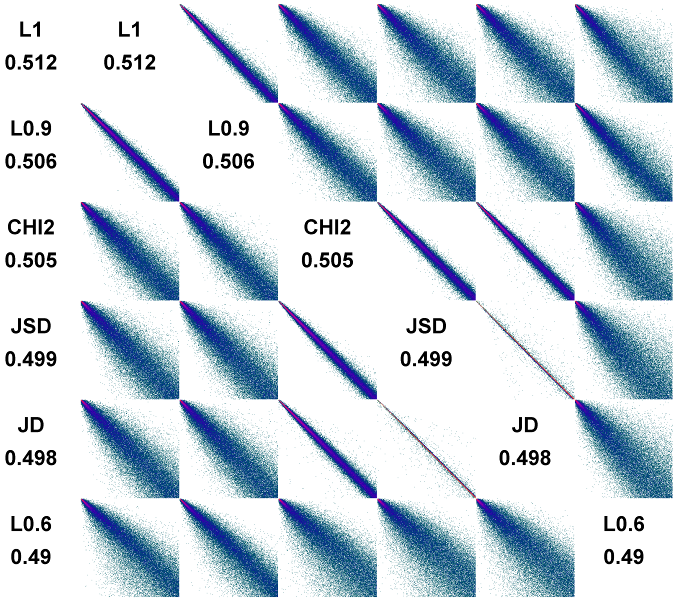


Fig. 3. Matrix of pairwise ranking comparison between best performing dissimilarity measures for the DSR descriptor on the KNDB dataset. Each tile compares all obtained rankings of relevant objects for two measures, starting from the top ranks (top and left of tile) to the 0.5 quantile (bottom and right of tile). E.g. over all query objects from KNDB, JSD and JD show by far the highest correlation in their rankings of relevant objects. This is reflected by a clearly expressed main diagonal within the tile. Note that while R-precision of e.g. L_1 and χ^2 only differs by 0.007, a large number of relevant objects is ranked differently. The effect increases from the top ranks up to the 0.5 quantile.

analysis of the correlation of rankings in pairwise comparison of different dissimilarity measures. In Figure 3 we elaborate an example while more visualizations are provided in the supplementary material [1]. Our results show that although in pairwise comparison, measures often show similar and good R-precision levels, they provide significantly different (uncorrelated) rankings. This is a strong indication that different metrics manage to retrieve different relevant objects and hence potentially complementary result sets. This warrants further investigation, as it might be possible to heuristically combine multiple of such ranking (e.g. also over single features) to improve precision and recall.

VI. Conclusions

Our experiments provide extensive and novel empirical results for the impact of various dissimilarity measures over a broad selection of global 3D features and their late fusion. We found strong empirical indication, that for single feature 3D retrieval, other metrics besides L_1 and L_2 very often provide better performance. Furthermore we showed that determining the best performing measures for individual feature vectors has practical relevance for improving the performance of late-fused multi-feature 3D retrieval. As side results, we found indication that α- and max-normalization outperform mean-normalization over a large number of test runs and outlined a possible opportunity to jointly exploit multiple dissimilarity measures over single features for improving retrieval performance.

Acknowledgments

This work was partially funded by the European Commission in context of the FP7 STREP Project PRESIOUS, grant no. 600533 (http://www.presious.eu/). It was also partially funded by Fondecyt (Chile) Project 1140783.

References

[1] *Supplementary material.* http://dissimilarity-measure-eval.dbvis.de.

[2] The princeton shape benchmark. In *Proceedings of the Shape Modeling International*, pages 167–178, 2004.

[3] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory ICDT 2001*, volume 1973 of *Lecture Notes in Computer Science*, pages 420–434. 2001.

[4] Relja Arandjelovic and Andrew Zisserman. All About VLAD. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1578–1585, 2013.

[5] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England, 2011.

[6] J. Barrios and B. Bustos. Competitive content-based video copy detection using global descriptors. *Multimedia Tools and Applications*, 62(1):75–110, 2013.

[7] B. Bustos, D. Keim, D. Saupe, T. Schreck, and D. Vranic. An experimental effectiveness comparison of methods for 3d similarity search. 6(1):39–54, 2006.

[8] B. Bustos, D. A. Keim, D. Saupe, T. Schreck, and D. Vranic. Feature-based Similarity Search in 3D Object Databases. *ACM Computing Surveys (CSUR)*, 37(4):345–387, 2005.

[9] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. *Procedings of the British Machine Vision Conference 2011*, (1):76.1–76.12, 2011.

[10] R. Datta, D. Joshi, J. Li, and J.Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):1–60, 2008.

[11] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the Web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622. ACM, 2001.

[12] D. François, V. Wertz, and M. Verieysen. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):873–886, 2007.

[13] T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, D. Dobkin, and D. Jacobs. A search engine for 3D models. *ACM Transactions on Graphics*, 22(1):83–105, January 2003.

[14] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kauffman, 2nd edition, 2006.

[15] S. Jayanti, Y. Kalyanaraman, N. Iyer, and K. Ramani. Developing an engineering shape benchmark for CAD models. *Computer-Aided Design*, 38(9):939 – 953, 2006.

[16] H. Liu and H. Motoda, editors. *Computational Methods of Feature Selection*. Chapman and Hall/CRC, Boca Raton, October 2007.

[17] J. Liu, Z. Huang, H. Cai, H.T. Shen, C. W. Ngo, and W. Wang. Near-duplicate video retrieval: Current research and future trends. *ACM Comput. Surv.*, 45(4):44:1–44:23, 2013.

[18] M. Müller. *Information Retrieval for Music and Motion*. Springer, 2007.

[19] O. Pele and M. Werman. The quadratic-chi histogram distance family. In *ECCV*, 2010.

[20] J. Puzicha, Y. Rubner, C. Tomasi, and J. M. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. In *ICCV*, pages 1165–1172, 1999.

[21] T. Schreck, M. Scherer, M. Walter, B. Bustos, S. M. Yoon, and A. Kuijper. Graph-based combinations of fragment descriptors for improved 3d object retrieval. In *Proc. ACM Multimedia Systems Conference*, pages 23–28, 2012.

[22] I. Sipiran, R. Meruane, B. Bustos, T. Schreck, B. Li, Y. Lu, and H. Johan. A benchmark of simulated range images for partial shape retrieval. *The Visual Computer*, 30(11):1293–1308, 2014.

[23] J. W. Tangelder and R. C. Veltkamp. A survey of content based 3d shape retrieval methods. *Multimedia Tools Appl.*, 39(3):441–471, September 2008.

[24] D. Vranic. DESIRE: a composite 3d-shape descriptor. In *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, ICME*, pages 962–965, 2005.