From 2D photography to 3D object retrieval : contributions and benchmarking

Méthodes 2D-3D pour l'indexation et la recherche d'images

Thibault Napoléon
Hichem Sahbi

2009D008

Février 2009

Département Traitement du Signal et des Images
Groupe TII : Traitement et Interprétation des Images

From 2D Photography to 3D Object Retrieval: Contributions and Benchmarking

Méthodes 2D-3D pour l'Indexation et la Recherche d'Images

Thibault Napoléon THIBAULT.NAPOLEON@TELECOM-PARISTECH.FR
Institut Télécom; Télécom ParisTech
CNRS LTCI UMR 5141, France

Hichem Sahbi HICHEM.SAHBI@TELECOM-PARISTECH.FR
CNRS LTCI UMR 5141
Télécom ParisTech, France

Abstract

3D retrieval has recently emerged as an important boost for 2D search techniques, by its several complementary aspects, for instance, enriching views in 2D image datasets, overcoming occlusion and serving in many real world applications such as photography, art, archeology and geo-localization.

In this paper, we introduce a complete "2D photography to 3D object" retrieval framework which, given a (collection of) picture(s) of the same scene or object, allows us to retrieve the underlying similar objects in a database of 3D models. The contributions of our method include (i) a generative approach for alignment which is able to find canonical views consistently through scenes/objects and (ii) an efficient but effective matching method used for ranking. The results are reported through a benchmarking consortium and evaluated/compared by a *third-party*, showing clearly the good performance of our framework with respect to the other participants.

Résumé

L'indexation 3D connaît un intérêt croissant par rapport aux méthodes 2D classiques grâce à ses nombreux aspects complémentaires, par exemple l'enrichissement des vues des objets reconnus/recherchés, ce qui permet de prendre en compte les occultations. L'intérêt est aussi pratique notamment en photographie, l'art et la géo-localisation.

Dans cet article, on introduit une nouvelle approche d'indexation dite "2D-vers-3D" qui, étant donnée une (ou plusieurs) images de même objets/scènes, permet de retrouver des modèles similaires dans une base d'objets 3D. La contribution de notre méthode inclut (i) une approche d'alignement capable de trouver des vues canoniques d'un objet 3D ainsi (ii) qu'une méthode d'appariement rapide et précise utilisée pour la classification. Les résultats sont évalués et comparés à d'autres méthodes dans le cadre d'une compétition internationale qui montre clairement les bonnes performances de notre approche par rapport aux autres participants.

Keywords: Multi-View Photography Indexing and Retrieval, 3D Object Recognition, Dynamic Programming.

1. Introduction

3D object recognition and retrieval recently gained a big interest (NIST, 2008) because of the limitation of the 2D approaches which clearly suffer several drawbacks including the lack of information (due for instance to occlusion), pose sensitivity, illumination changes, etc. This is also due to the exponential growth of storage and bandwidth on the Internet, the increasing needs for services from 3D content providers (museum institutions, car manufacturers, etc.) and the easiness in collecting gallery sets[1] as computers are now equipped with highly performant, easy to use, 3D scanners and graphic facilities for real-time modeling, rendering and manipulation. Nevertheless, at the current time, functionalities including retrieval of 3D models are not yet sufficiently precise in order to be available for large usage.

Almost all the 3D retrieval techniques are resource (time and memory) demanding prior to achieve recognition and ranking, as they operate on massive amount of data and they require many upstream steps including object alignment, 3D-to-2D projections and normalization. However and when no hard runtime constraints are expected, 3D search engines offer real alternatives and substantial gains in performance, with respect to only image-based retrieval approaches mainly when the relevant informations are appropriately extracted and processed (see for instance Bimbo and Pala (2006).)

Existing 3D object retrieval approaches can either be categorized into those which operate directly on the 3D content and those which extract or already have "2.5D" or 2D contents (stereo-pairs or multiple-views of images, artificially rendered 3D objects, silhouettes, etc.) Comprehensive surveys on 3D retrieval can be found in (Tangelder and Veltkamp, 2004; Shilane et al., 2004; Zaharia and Prêteux, 2004; Bustos et al., 2004; Bimbo and Pala, 2006; Biasotti et al., 2006a). Existing state of the art techniques may also be categorized depending on the fact that they require a preliminary step of alignment and those which operate directly by extracting global, possibly invariant, signatures on the 3D models such as Zernike's 3D moments (Novotni, 2003). The latter are extracted using salient characteristics on 3D, "2.5D" or 2D shapes and ranked according to similarity measures. Structure-based approaches, presented in Tung and Schmitt (2005); Tierny et al. (2006b); Hilaga et al. (2001); Tierny et al. (2006a), encode topological shape structures and make it possible to compute efficiently, without pose alignment, similarity between two global or partial 3D models. Funkhouser and Shilane (2006); Biasotti et al. (2006b) introduced two methods for partial shape-matching able to recognize similar sub-parts of objects represented as 3D polygonal meshes. The methods in (Funkhouser et al., 2003; Saupe and Vranic, 2001; Kazhdan et al., 2003) use spherical harmonics in order to describe shapes, where rotation invariance is achieved by taking only the power spectrum of the harmonic representations and discarding all "rotation dependent" informations. Other approaches include those which analyze 3D objects using analytical functions/transforms (Zarpalas et al.,

[1] Event though in a chaotic way because of the absence of consistent alignments of 3D models.

2006; Laga et al., 2006) and also those based on learning (Ohbuchi and Kobayashi, 2006).

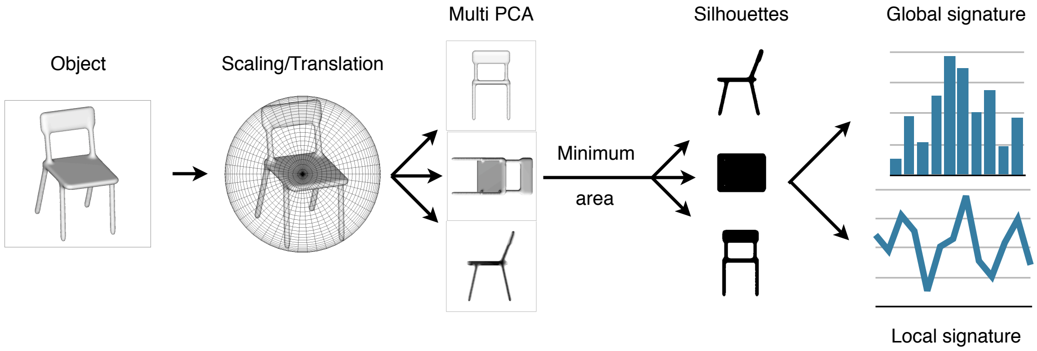

Figure 1: This figure shows alignment and feature extraction process. First, we compute the smallest enclosing ball of a 3D object. Then, we combine PCA with the minimal visual hull in order to align the underlying 3D model. Finally, three silhouettes of a canonical view are extracted prior to feature extraction.

An other family of 3D object retrieval approaches belongs in the frontier between 2D and 3D querying paradigms; for instance the method in (Papadakis et al., 2008), is based on extracting and combining spherical 3D harmonics with "2.5D" depth information features and in (Mahmoudi and Daoudi, 2002; Filali-Ansary et al., 2007) are based on selecting characteristic views and encoding them using the curvature scale space descriptor. Other "2.5D" approaches (Chaouch and Verroust-Blondet, 2007) are based on extracting rendered depth lines (as in Vranic (2004); Chaouch and Verroust-Blondet (2006); Ohbuchi et al. (2003)), resulting from vertices of regular dodecahedrons and matching them using dynamic programming. Chen (2003) proposed a 2D method based on Zernike's moments that provides the best results on the Princeton shape benchmark (Shilane et al., 2004). In this method, rotation invariance is obtained using the light-field approach where all the possible permutations of several dodecahedrons are used in order to cover the space of view points around an object.

1.1 Motivations

Due to the compactness of global 3D object descriptors, their performance in capturing the inter/intra class variabilities are known to be poor in practice. In contrast, local geometric descriptors, even though computationally expensive, achieve relatively good performance and capture inter/intra class variabilities (including deformations) better than global ones (see Section 5). *The framework presented in this paper is based on local features and also cares about computational issues while keeping advantages in terms of precision and robustness.*

Our *target* is searching 3D databases of objects using one or multiple 2D views; this scheme will be referred to as "2D-to-3D". We define our *probe* set as a collection of single or multiple views of the same scene or object while our *gallery* set corresponds to a large set of 3D models. A query, in the probe set, will either be (i) multiple views of the same object, for instance stereo-pair, or (ii) a 3D object model processed in order to extract several views; so ending with the "2D-to-3D" querying paradigm in both cases (i+ii). Gallery data are also processed in order to extract several views for each 3D object (see Section 2).

At least two reasons motivate the use of the 2D-to-3D querying paradigm:

- The difficulty of getting "3D query models" when only multiple views of an object of interest are available. This might happen when 3D reconstruction techniques (Jin et al., 2005) fail or when 3D acquisition systems are not available. 2D-to-3D approaches should then be applied instead.
- 3D gallery models can be manipulated via different similarity and affine transformations, in order to generate multiple views which fit the 2D probe data, so "2D-to-3D" recognition and retrieval paradigm can be achieved.

1.2 Contributions

This paper is a novel 2D-to-3D retrieval framework with the following contributions:

(i) A new generative approach for aligning 3D objects and extracting their canonical 2D views is introduced. The method is based on combining principal component analysis (PCA) with the minimal visual hull (see Section 2). Given a 3D object, alignment is achieved by minimizing its the visual hull with respect to the pose parameters (translation, scale, etc.) We found in practice that this clearly outperforms the usual PCA and makes the retrieval process invariant to several transformations including rotation, reflection, translation and scaling.

(ii) Afterwards, robust and compact contour signatures are extracted using the set of canonical 2D views[2]. Our signature is an extension of the multi-scale curve representation first introduced in (Adamek and O'Connor, 2004) and it is based on computing a set of convexity/concavity coefficients on the contours of the (2D) object views. We also consider global histogram descriptors which capture global distributions of these coefficients in order to perform pruning and speedup the whole search process.

(iii) Finally, ranking is performed using our variant of dynamic programming which considers only a subset of possible matches thereby providing a considerable gain in performance for the same amount of errors.

Figures 1, 2 show our whole proposed matching and retrieval framework which was benchmarked in the international Shrec'09 contest on structural shape retrieval. This framework achieves very encouraging performance and outperforms almost all the participating runs.

[2] As contours are strongly correlated, only a subset of them are used.

In the remainder of this paper, we consider the following notation; let X be a random variable standing for the 3D coordinates of vertices in any 3D object. For a given object, we assume that X is drawn from an existing but unknown probability distribution P. Let us consider $\mathcal{P}_n = \{X_1, ..., X_n\}$ as n realizations of X, forming a 3D object model. $\mathcal{P}_n$ or $\mathcal{P}$ (resp. $\mathcal{G}_m$ or $\mathcal{G}$) will be used to denote a 3D object belonging to the probe (resp. the gallery) set while $\mathcal{O}$ is a generic 3D object either belonging to the probe or the gallery set. Without any loss of generality 3D models are characterized by a set of vertices which may be meshed in order to form a closed surface or a compact manifold of 2 intrinsic dimensions. Other notations and terminologies will be introduced as we go through different sections of this paper which is organized as follows. Section 2 introduces the alignment and pose estimation process. Section 3 presents the global and the local multi-scale contour convexity/concavity signatures. The matching process together with pruning strategies are introduced in Section 4, ending with experiments and comparison on the recent Shrec'09 international benchmark in Section 5.

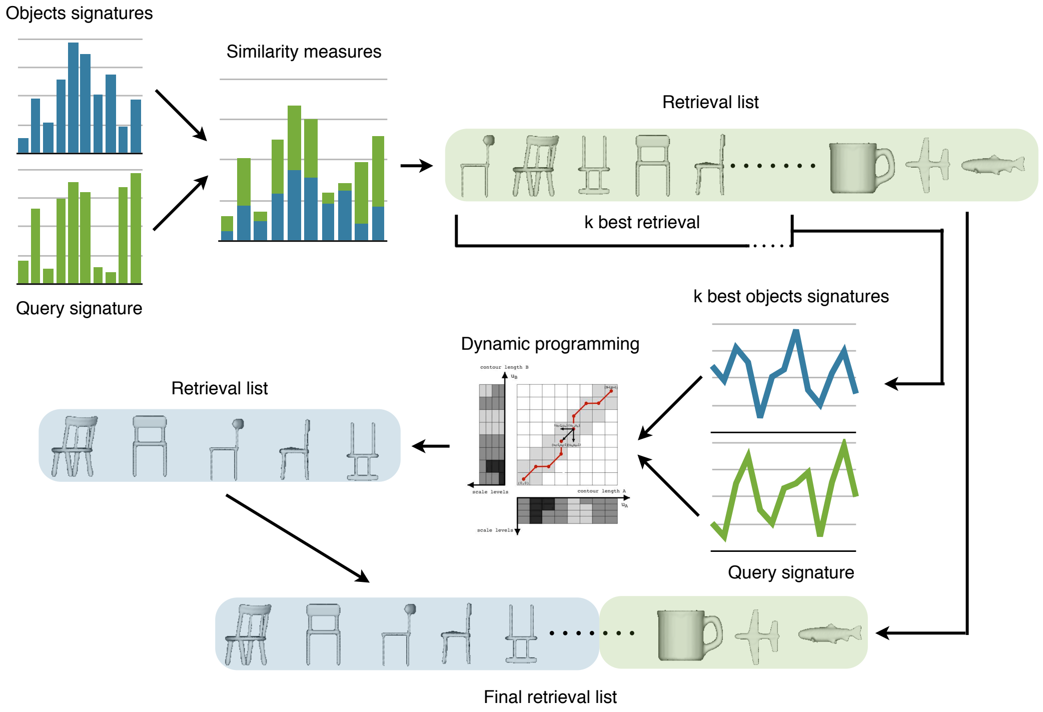

Figure 2: This Figure shows an overview of the matching framework. First, we compute many distances between the global signature of the query and all objects in the database. According to these distances, we create a ranked list. Then, we search the best matching between the local signature of the query and the top k ranked objects.

2. Pose Estimation

The goal of this step is to make retrieval invariant to 3D object transformations (including scaling, translation, rotation and reflection) and also to generate multiple views of 3D models both in the probe and the gallery sets. Pose estimation consists in finding the parameters of the above transformations (denoted resp. $s \in \mathbb{R}$, $(t_x, t_y) \in \mathbb{R}^2$, $(\theta, \rho, \psi) \in \mathbb{R}^3$ and $(r_x, r_y, r_z) \in \{-1, +1\}^3$) by aligning 3D models to fit into canonical poses and the underlying 2D multiple views will be referred to as the *canonical views* (see Figure 1).

Let us consider $\Theta = (s, t_x, t_y, \theta, \rho, \psi, r_x, r_y, r_z)$ and given a 3D object $\mathcal{O}$, our alignment process is *generative* i.e., based on varying and finding the optimal parameters Θ which minimize the following visual hull criterion(Fischer and Gartner, 2004):

$$\hat{\Theta} \;=\; \arg\min_{\Theta} \sum_{v \in \{xy,xz,yz\}} f\Big([\mathbf{P}_v \circ \mathbf{T}_\Theta](\mathcal{O})\Big) \qquad (1)$$

$\mathbf{P}_v$, $v \in \{xy, xz, yz\}$, denote respectively the "3D-to-2D" parallel projections on the xy, xz, yz canonical 2D planes, characterized by their normals (resp. $n_{xy} = (0\ 0\ 1)'$, $n_{xz} = (0\ 1\ 0)'$, $n_{yz} = (1\ 0\ 0)'$) and $\mathbf{T}_\Theta = \mathbf{F}_{r_x,r_y,r_z} \circ \mathbf{\Gamma}_s \circ \mathbf{R}_{\theta,\rho,\psi} \circ \mathbf{t}_{t_x,t_y}$ denotes the global alignment transformation resulting from the combination of translation, rotation, scaling and reflection. The visual hull in (1) is defined *as the sum of the projection areas of* $\mathcal{O}$ *using* $\mathbf{P}_v \circ \mathbf{T}_{\hat{\Theta}}$. Let $\mathcal{H}_v(\mathcal{O}) = [\mathbf{P}_v \circ \mathbf{T}_{\hat{\Theta}}](\mathcal{O}) \subset \mathbb{R}^2$, $v \in \{xy, xz, zy\}$, here $f \in \mathbb{R}^{\mathcal{H}_v(\mathcal{O})}$ provides this area on each silhouette of a 2D canonical view.

It is clear that the objective function (1) is difficult to solve as one needs to recomputed, for each possible Θ the underlying visual hull. So it becomes clear that parsing the domain of variation of Θ makes the search process tremendous; furthermore, no gradient descent can be achieved, as there is no guarantee that f is continuous w.r.t, Θ. Instead, we restrict the search by considering the following procedure:

Translation and scaling: $\mathbf{t}_{t_x,t_y}$ and $\mathbf{\Gamma}_s$ are recovered simply by centering and rescaling the 3D points in $\mathcal{O}$ so they fit inside an enclosing ball of unit radius.

Rotation: $\mathbf{R}_{\theta,\rho,\psi}$ is taken as one of the four possible candidates matrices including (i) identity[3], or one of the transformation matrices resulting from PCA either on (ii) gravity centers (iii) vertices or (iv) face normals, of $\mathcal{O}$. The three cases (ii), (iii), (iv) will be referred to as PCA, continuous PCA (CPCA) and normal PCA (NPCA) respectively (Vranic, 2004; Vranic et al., 2001).

Axis Reordering and Reflection: this step consists in *re-ordering and reflecting* the three projection planes $\{xy, xz, yz\}$, in order to generate all the possible 2D canonical views, resulting into 48 possible partitions, i.e., 3! (for reordering) × 2^3 (for reflection). The domain of six possible reordering includes $\{xy, xz, yz\}$, $\{xz, xy, zy\}$, $\{yx, yz, xz\}$, $\{yz, yx, zx\}$, $\{zx, zy, xy\}$, $\{zy, zx, yx\}$ while the set of possible reflections contains $\{x'y', x'z', y'z'\}$, here

[3] The initial object pose is assumed to be the canonical one.

$x' = r_x\ x,\ y' = r_y\ y,\ z' = r_z z$. Reflection makes it possible to consider mirrored views of objects, while reordering allows us to permute the principal orthogonal axes of an object and therefore permuting the underlying 2D canonical views.

For each combination taken from the "48 reflections and reordering × 4 possible rotations" (see explanation earlier), the objective function (1) is evaluated and the combination (i.e., the underlying Θ) which minimizes this function is kept as the best transformation. When fixing only the best rotation, 48 possible reflections/reordering are also used in order to generate 48 possible mirrored and reordered axes and accordingly 48 canonical views for each object $\mathcal{P}_n$, $\mathcal{G}_m$ in both the probe and the gallery sets. This generative querying approach covers many canonical views and improves retrieval results as will be discussed later.

3. Multi-View Object Description

Again, for each object $\mathcal{O}$ we extract 48 canonical views each one has three images which correspond to the projection of $\mathcal{O}$ on the three underlying canonical planes (see Section 2). Then, each image is processed in order to extract and describe external contours using a variant of (Adamek and O'Connor, 2004). Our description is based on a multi-scale analysis which extracts convexity/concavity coefficients on each contour. Since the latter are strongly correlated through many views of a given object $\mathcal{O}$, we describe our contours on canonical 2D views containing only three images. This reduces redundancy and also speedups the whole feature extraction and matching process (see Figure 3).

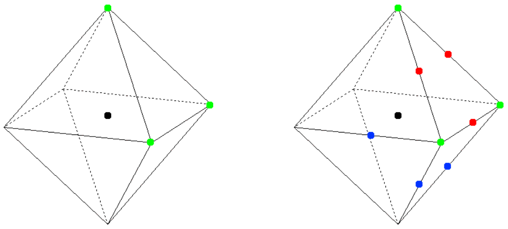

Figure 3: This Figure shows view points when capturing images/silhouettes of 3D models. The left-hand side picture shows the three view-points corresponding to the three PCA axes while the right-hand side one, contains also six bi-sectors. The latter provides better view-point distribution over the unit sphere.

In practice, each contour, denoted C, is sampled with N (2D) points ($N = 100$) and processed in order to extract the underlying convexity/concavity coefficients at K different scales ($K = 10$) (Adamek and O'Connor, 2004). Contours are iteratively filtered (10 times) using a Gaussian kernel with an increasing scale parameter $\sigma \in \{1, 2, ..., \sigma_K\}$. Each curve C will then evolve into 10 different smooth silhouettes. Let us consider a parameterization of C using the curvilinear abscissa u as $C(u) = (x(u),\ y(u))$, $u \in [0, N-1]$, let us also denote C_σ as a smooth version of C resulting from the application of the Gaussian kernel with a scale σ (see Figure 4).

We use simple convexity/concavity coefficients as local descriptors for any 2D point $\mathbf{p}_{u,\sigma}$ on C_σ ($\mathbf{p}_{u,0} = C(u)$). Each coefficient is defined as the signed amount of shift of $\mathbf{p}_{u,\sigma}$ between two consecutive scales σ and $\sigma-1$. Put differently, a convexity/concavity coefficient denoted $\mathbf{d}_{u,\sigma}$ is taken as $\|\mathbf{p}_{u,\sigma} - \mathbf{p}_{u,\sigma-1}\|_2$, here $\|r\|_2 = \sum_i^d r_i^2$ denotes the L_2 norm.

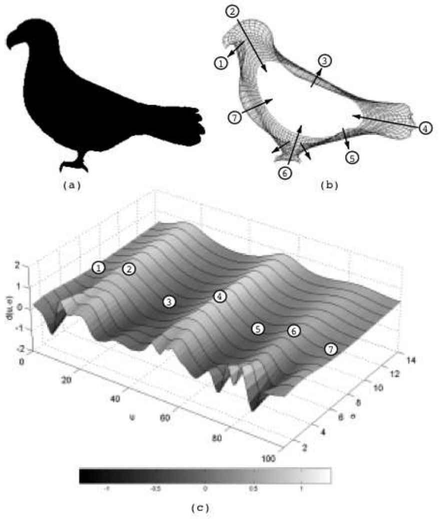

Figure 4: Example of extracting the MCC shape representation: original shape image (a), filtered versions of the original contour at different scale levels (b), final MCC representation for $N = 100$ contour points and $K = 14$ scale levels (c).

4. Coarse-to-Fine Matching

4.1 Coarse Pruning

A simple *coarse* shape descriptor is extracted both on the gallery and probe sets. This descriptor quantifies the distribution of convexity and concavity coefficients through 2D points belonging to different silhouettes of a given object $\mathcal{O}$ taken at the 48 different canonical views (see Section 2). This coarse descriptor is a multi-scale histogram containing 100 bins as the product of 10 scales of the Gaussian kernel (see Section 3) and 10 quantification values for convexity/concavity coefficients. Each bin of this histogram counts, through all the view-point silhouettes of $\mathcal{O}$, the frequency of the underlying convexity/concavity coefficients. This descriptor is poor in terms of its discrimination power, but efficient in order to reject almost all the false matches while keeping candidate ones when ranking the gallery objects w.r.t the probe ones (as shown in the experiments).

4.2 Fine Matching by Dynamic Programming

Given two objects $\mathcal{P}$, $\mathcal{G}$ respectively from the probe and the gallery sets, and the underlying silhouettes/curves $\{C_i\}$, $\{C_j^{'}\}$ taken from two canonical 2D views. Each canonical view is indexed by the pose parameters $\Theta_{\mathcal{P}}$ (resp. $\Theta_{\mathcal{G}}$) of $\mathcal{P}$ (resp. $\mathcal{G}$) (see Section 2). A global scoring function is defined between $\mathcal{P}$, $\mathcal{G}$ as the average sum of matching pseudo distances involving all the silhouettes $\{C_i\}$, $\{C_j^{'}\}$ taken from the canonical views (of $\Theta_{\mathcal{P}}$, $\Theta_{\mathcal{G}}$) as

$$S(\mathcal{P},\mathcal{G}) \;=\; \frac{1}{N_s}\sum_{i=1}^{N_s}\mathbf{DSW}(C_i,C_i^{'}), \tag{2}$$

here N_s is the number of silhouettes per object (in practice, $N_s = 3$ or 9, see Section 5). For matter of robustness, we consider the expectation of $S(\mathcal{P},\mathcal{G})$ through all the possible 48 canonical 2D views of $\mathcal{P}$.

Silhouette matching is performed using dynamic programming; given the two curves C_i, $C_i^{'}$ a matching pseudo distance, denoted **DSW**, is obtained as a sequence of operations (substitution, insertion and deletion) which transforms C_i into $C_i^{'}$ (Bellman, 1966). Given $N = 100$ sample points from C_i, $C_i^{'}$, and the underlying local convexity/concavity coefficients $F, F' \subset \mathbb{R}^K$, the **DSW** pseudo-distance is defined as

$$\mathbf{DSW}(C_i,C_i^{'}) \;=\; \frac{1}{N}\sum_{u=1}^{N}\Big\|F(u)-F'(g(u))\Big\|_1, \tag{3}$$

here $\|r\|_1 = \sum_i |r_i|$ denotes the L_1-norm, $F(u) \in F$ and $g : \{1,...,N\} \rightarrow \{1,...,N\}$ is the dynamic programming matching function, which assigns for each curvilinear abscissa u in C_i its corresponding abscissa $g(u)$ in $C_i^{'}$. Given the distance matrix $D_{uu'} = \|F(u) - F'(u')\|_1$, the matching function g is found by selecting a path in D which minimizes the number of operations (substitution, deletion and insertion) in order to transform C_i into $C_i^{'}$ while preserving the ordering assumption (i.e., if u is matched with u' then $u+1$ should be matched only with $u'+l$, $l > 0$). We introduced a variant of the standard dynamic programming version which, instead of examining all the possible matches, considers only those which belong in a diagonal band of D, i.e., l is allowed to take only small values (see Figure 5).

Dynamic programming pseudo-distance provides a good discrimination power and can capture the intra-class variations better than the global distance (shown in Section 4.1). Nevertheless, it is still computationally expensive but when combined with coarse pruning the whole process is significantly faster and also precise (see Table 1).

5. Experiments

5.1 Databases

In order to evaluate the robustness of the proposed framework, we used the Watertight dataset of the Shrec'07 benchmark as a ground truth. This dataset contains 400 "3D" objects represented by seamless surfaces (without defective holes or gaps). The models of this

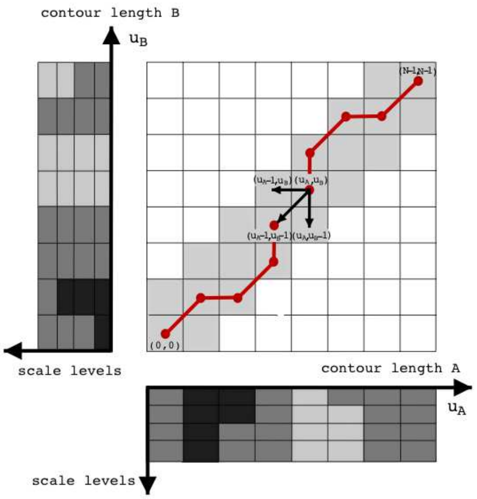

Figure 5: This Figure shows dynamic programming used in order to find the global alignment of two contours.

database have been divided into 20 classes each one contains 20 objects. The 3D models were taken from two sources: the first one is a deformation of an initial subset of objects (octopus, glasses,...), while the second one is a collection of original 3D models (chair, vase, four leg...). Each 3D object belongs to a unique class among different semantic concepts with strong variations including human, airplane, chair, etc. For instance, the human class contains persons with different poses and appearances ”running, seating, walking, etc.”, globally the database is very challenging.

5.2 Benchmarking

We evaluated our method using recall-precision. Precision is defined as the fraction of relevant retrieved objects over the cardinality of the display while recall is defined as the fraction of relevant retrieved objects over the total number of relevant 3D models in the dataset. A plot that appears shifted up indicates better retrieval results. In addition to recall/precision plot, we use several quantitative statistics to evaluate the results:

- The nearest neighbor (NN) which represents the fraction of the first nearest neighbors which belong to the same class as the query.

- The first-tier (FT) and the second-tier (ST): These measures give the percentage of objects in the same class as the query that appear in the k best matches. For a given class $\mathcal{C}$ containing $|\mathcal{C}|$ objects, k is set to $|\mathcal{C}|-1$ for the first-tier measure while k is set to $2(|\mathcal{C}|-1)$ for second-tier (ST).

- Finally, we use the discounted cumulative gain (DCG) measure which gives more importance to well ranked models. Given a query and a list of ranked objects, we define for each ranked object a variable r_i equal to 1 if its class is equal to the class of the query and 0 otherwise. The DCG measure is then defined as:

$$\begin{array}{rcl} \mathbf{DCG}_i & = & \mathbf{DCG}_{i-1} + \dfrac{r_i}{\log_2(i)} \quad \text{if} \quad (i \neq 1) \\ & & r_i \quad \text{otherwise} \end{array} \tag{4}$$

We take the expectation of these measures on the entire database, i.e., by taking all the possible object queries.

5.3 Performances and Discussion

Figure 6, table 1 (row 2,3) illustrate precision recall and the statistics defined earlier. We clearly see that our new pose estimation method, defined in Section 2, gives better results compared to the classical PCA approach for alignment. Again our pose estimation model makes it possible to extract several canonical 2D views and for each one we compared results using either three or nine 2D images per canonical view (see results in figure 7 and rows 2,3 of table 1.)

In order to control/reduce the runtime to process and match local signatures, we used our pruning approach based on the global signature discussed in Section 4.1. The parameter k allows us to control the trade off between robustness and speedup of the retrieval process. A small value of k gives real-time (online) responses with an acceptable precision while a high value requires more processing time but gives better retrieval performance. Figure 8 shows the NN, FT, ST and DCG measures for different pruning thresholds k. Rows 3, 5 and 6 of table 1 show different statistics for $k = 0$, 50 and 400.

	NN (%)	FT (%)	ST (%)	DCG (%)	Average Runtime Per Query
3 views, PCA, pruning with $k = 50$	95.5	60.7	71.2	86.3	1.1 (s)
3 views, our pose, pruning with $k = 50$	95.5	61.4	73.3	86.7	1.1 (s)
9 views, our pose, pruning with $k = 50$	94.5	64.3	74.3	87.8	3.2 (s)
3 views, our pose, pruning with $k = 0$	89	56.1	71.3	83.4	0.003 (s)
3 views, our pose, pruning with $k = 400$	95.5	62.4	74.3	87	8.2 (s)

Table 1: Performance on the Shrec'09 challenge.

Comparisons of our approach with respect to different methods/participants are available and were generated by a third party in the Shrec'09 Structural Shape Retrieval contest. This dataset contains 200 objects and results were evaluated on 10 queries. The performance of this shape retrieval contest were measured using 1^{st} (10 objects) and 2^{nd} (30 objects) tier precision and recall, presented as the F-measure. This is a global measure which provides us with the overall retrieval performance.

We submitted in this benchmark four runs:

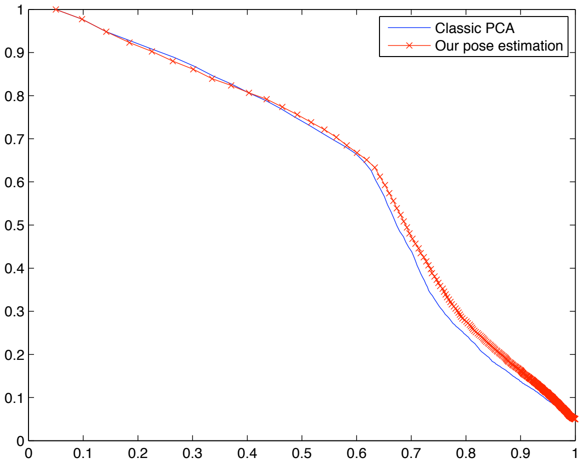

Figure 6: This Figure shows precision versus recall using our framework with 3 silhouettes per canonical view (pruning threshold $k = 50$). Classical PCA pose estimation method is shown in blue while our method in red.

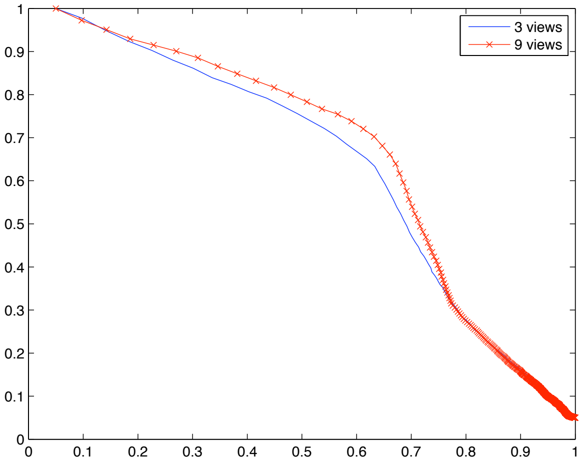

Figure 7: This Figure shows comparison of precision versus recall (with our pose estimation method + pruning threshold $k = 50$), using 3 silhouettes (in blue) and 9 silhouettes (in red).

- Run 1 (MCC1): 9 silhouettes and pruning threshold $k = 0$. The average runtime for each query is 0.01 s.

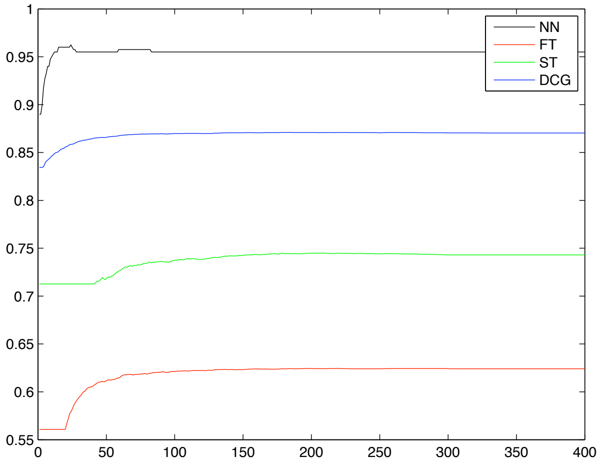

Figure 8: This figure shows the evolution of the NN, FT, ST and DCG measures (in %) w.r.t. the pruning size k. We found that $k = 75$ makes it possible to reject almost all the false matches in the gallery set. We found also that the CPU runtime scales linearly with respect to k.

- Run 2 (MCC2): 9 silhouettes and pruning threshold $k = 50$. The average runtime for each query is 3.2 s.
- Run 3 (MCC3): 9 silhouettes and pruning threshold $k = 200$. The average runtime for each query is 12 s.
- Run 4 (MCC4): 3 silhouettes and pruning threshold $k = 50$. The average runtime for each query is 1 s.

We can see in the Figures 9 and 10 that the third run of our method (shown in pink) outperforms the others for the first tier and is equivalent to the Compound SID CMVD 3 for the second tier. The results four the second run are similar to the SIFT 3D (shown in yellow) and to the Compound SID CMVD 1 methods (shown in green) with efficient runtime performances.

6. Conclusion

We introduced in this paper a novel and complete framework for 2D-to-3D object retrieval. The approach makes it possible to extract canonical views using a generative approach combined with principal component analysis. The underlying silhouettes/contours are matched using dynamic programming in a coarse-to-fine way which makes the search process efficient and also effective.

One of the major drawbacks of dynamic programming resides in the fact that it is not a metric, so one cannot benefit from lossless acceleration techniques which provide precise

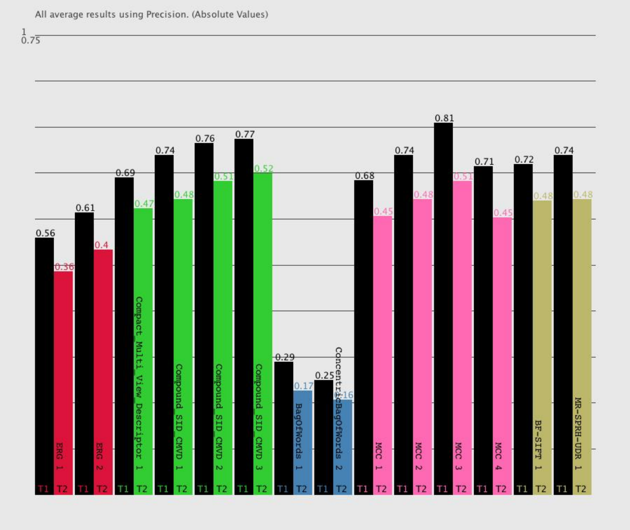

Figure 9: This Figure shows the precision using two tier: 10 (shown in black) and 20 (shown in color). Our results are shown in pink/black. These results can be checked in the Shrec'09 challenge home pages.

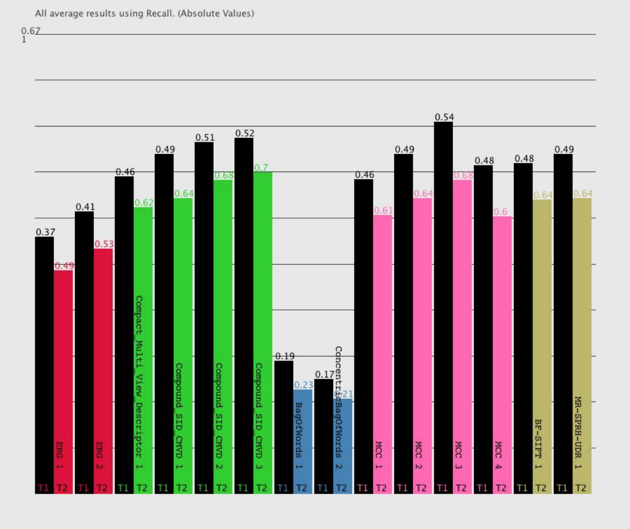

Figure 10: This Figure shows the recall using two tier: 10 (shown in black) and 20 (shown in color). Our results are shown in pink/black. These results can be checked in the Shrec'09 challenge home pages.

results and efficient computation. One possible extension is to tackle this issue by introducing new matching approaches that allow us to speedup the search process while keeping high precision.

Acknowledgments

This work was supported by the European Network of Excellence KSpace and the French National Research Agency (ANR) under the AVEIR Project.

References

T. Adamek and N. E. O'Connor. A multiscale representation method for nonrigid shapes with a single closed contour. *IEEE Trans. Circuits Syst. Video Techn*, 14(5):742–753, 2004.

R. Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

S. Biasotti, D. Giorgi, S. Marini, M. Spagnuolo, and B. Falcidieno. A comparison framework for 3d object classification methods. *Multimedia Content Representation, Classification and Security : International Workshop, MRCS 2006*, sep 2006a.

Silvia Biasotti, Simone Marini, Michela Spagnuolo, and Bianca Falcidieno. Sub-part correspondence by structural descriptors of 3d shapes. *Computer-Aided Design*, 38(9):1002 – 1019, 2006b. ISSN 0010-4485. doi: DOI: 10.1016/j.cad.2006.07.003. Shape Similarity Detection and Search for CAD/CAE Applications.

Alberto Del Bimbo and Pietro Pala. Content-based retrieval of 3d models. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):20–43, 2006. ISSN 1551-6857. doi: http://doi.acm.org/10.1145/1126004.1126006.

Benjamin Bustos, Daniel Keim, Dietmar Saupe, Tobias Schreck, and Dejan Vranic. An experimental comparison of feature-based 3d retrieval methods. In *3DPVT '04: Proceedings of the 3D Data Processing, Visualization, and Transmission, 2nd International Symposium*, pages 215–222, Washington, DC, USA, 2004. IEEE Computer Society. ISBN 0-7695-2223-8. doi: http://dx.doi.org/10.1109/3DPVT.2004.30.

M. Chaouch and A. Verroust-Blondet. A new descriptor for 2d depth image indexing and 3d model retrieval. *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, 6:VI –373–VI –376, 16 2007-Oct. 19 2007. ISSN 1522-4880. doi: 10.1109/ICIP.2007.4379599.

Mohamed Chaouch and Anne Verroust-Blondet. Enhanced 2d/3d approaches based on relevance index for 3d-shape retrieval. In *Shape Modeling International'06*, Matsushima, jun 2006.

D.Y. Chen. *Three-dimensional model shape description and retrieval based on lightfield descriptors.* Ph.d., NTU CSIE, June 2003.

Tarik Filali-Ansary, Mohamed Daoudi, and Jean-Philippe Vandeborre. A bayesian 3d search engine using adaptive views clustering. In *IEEE Transactions On Multimedia*, jan 2007.

Kaspar Fischer and Bernd Gartner. The smallest enclosing ball of balls: Combinatorial structure and algorithms. *International Journal of Computational Geometry and Applications (IJCGA)*, 14:341–387, 2004.

Thomas Funkhouser and Philip Shilane. Partial matching of 3d shapes with priority-driven search. In *Symposium on Geometry Processing*, jun 2006.

Thomas Funkhouser, Patrick Min, Michael Kazhdan, Joyce Chen, Alex Halderman, David Dobkin, and David Jacobs. A search engine for 3D models. *ACM Transactions on Graphics*, 22(1):83–105, 2003.

Masaki Hilaga, Yoshihisa Shinagawa, Taku Komura, and Tosiyasu L. Kunii. Topology matching for fully automatic similarity estimation of 3d shapes. In *SIGGRAPH*, pages 203–212, 2001.

H. Jin, S. Soatto, and A. Yezzi. Multi-view stereo reconstruction of dense shape and complex appearance. *Intl. J. of Computer Vision*, 63(3):175–189, 2005.

Michael M. Kazhdan, Thomas A. Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3d shape descriptors. In *SGP '03: Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 156–165, Aire-la-Ville, Switzerland, Switzerland, 2003. Eurographics Association.

Hamid Laga, Hiroki Takahashi, and Masayuki Nakajima. Spherical wavelet descriptors for content-based 3d model retrieval. In *International Conference on Shape Modeling and Applications (SMI06)*, 2006.

Said Mahmoudi and Mohamed Daoudi. 3d models retrieval by using characteristic views. In *ICPR (2)*, pages 457–460, 2002.

NIST. Shape retrieval contest on a new generic shape benchmark. *itl.nist.gov/iad/vug/sharp/benchmark/shrecGeneric/*, 2008.

Marcin Novotni. 3d zernike descriptors for content based shape retrieval. In *In The 8th ACM Symposium on Solid Modeling and Applications*, pages 216–225. ACM Press, 2003.

Ryutarou Ohbuchi and Jun Kobayashi. Unsupervised learning from a corpus for shape-based 3d model retrieval. In *ACM MIR 2006*, Santa Barbara, CA, USA, oct 2006.

Ryutarou Ohbuchi, Masatoshi Nakazawa, and Tsuyoshi Takei. Retrieving 3d shapes based on their appearance. In Nicu Sebe, Michael S. Lew, and Chabane Djeraba, editors, *MIR '03: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 39–45. ACM, 2003.

P. Papadakis, l. Pratikakis, S. Perantonis, T. Theoharis, and G. Passalis. ShrecÕ08 entry: 2d/3d hybrid. *Shape Modeling and Applications, 2008. SMI 2008. IEEE International Conference on*, pages 247–248, June 2008. doi: 10.1109/SMI.2008.4547990.

Dietmar Saupe and Dejan V. Vranic. 3d model retrieval with spherical harmonics and moments. In Bernd Radig and Stefan Florczyk, editors, *Proceedings of the 23rd DAGM-Symposium on Pattern Recognition*, volume 2191 of *Lecture Notes in Computer Science*, pages 392–397. Springer, 2001.

Philip Shilane, Patrick Min, Michael Kazhdan, and Thomas Funkhouser. The princeton shape benchmark. In *SMI '04: Proceedings of the Shape Modeling International 2004*, pages 167–178, Washington, DC, USA, 2004. IEEE Computer Society. ISBN 0-7695-2075-8. doi: http://dx.doi.org/10.1109/SMI.2004.63.

J.W.H. Tangelder and R.C. Veltkamp. A survey of content based 3d shape retrieval methods. *Shape Modeling Applications, 2004. Proceedings*, pages 145–156, June 2004. doi: 10.1109/SMI.2004.1314502.

Julien Tierny, Jean-Philippe Vandeborre, and Mohamed Daoudi. 3d mesh skeleton extraction using topological and geometrical analyses. In *14th Pacific Conference on Computer Graphics and Applications (Pacific Graphics 2006, ACM/SIGGRAPH sponsored)*, pages 85–94, Taipei, Taiwan, oct 2006a.

Julien Tierny, Jean-Philippe Vandeborre, and Mohamed Daoudi. Invariant high-level reeb graphs of 3d polygonal meshes. In *3rd IEEE International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT'06)*, Chapel Hill, North Carolina, USA, jun 2006b.

Tony Tung and Francis Schmitt. The augmented multiresolution reeb graph approach for content-based retrieval of 3d shapes. *International Journal of Shape Modeling*, 11(1): 91–120, 2005.

Dejan V Vranic. *3D Model Retrieval*. PhD thesis, University of Leipzig, 2004.

Dejan V Vranic, Dietmar Saupe, and J Richter. Tools for 3d-object retrieval: Karhunen-loeve transform and spherical harmonics. In J-L Dugelay and K Rose, editors, *Proceedings of the IEEE 2001 Workshop Multimedia Signal Processing*, pages 271–274, Budapest, Hungary, sep 2001.

Titus Zaharia and Françoise Prêteux. 3d versus 2d/3d shape descriptors: A comparative study. In *in SPIE Conf. on Image Processing: Algorithms and Systems*, volume 2004, 2004. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.96.4410`.

Dimitrios Zarpalas, Petros Daras, Apostolos Axenopoulos, Dimitrios Tzovaras, and Michael G. Strintzis. 3d model search and retrieval using the spherical trace transform. In *EURASIP Journal on Advances in Signal Processing*, 2006.

Dépôt légal : 2009 – 1^er trimestre
Imprimé à l'Ecole Nationale Supérieure des Télécommunications – Paris
ISSN 0751-1345 ENST D (Paris) (France 1983-9999)

TELECOM ParisTech

Institut TELECOM - membre de ParisTech

46, rue Barrault - 75634 Paris Cedex 13 - Tél. + 33 (0)1 45 81 77 77 - www.telecom-paristech.frfr

Département TSI