Neurocomputing 253 (2017) 24–33

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Learning shape retrieval from different modalities

Hedi Tabia [a,*], Hamid Laga [b,c]

[a] ETIS/ENSEA, University of Cergy-Pontoise, CNRS, UMR 8051, Cergy, France
[b] School of Engineering and IT, Murdoch University, Australia
[c] Phenomics and Bioinformatics Research Centre, University of South Australia

ARTICLE INFO

Article history:
Received 30 April 2016
Revised 29 December 2016
Accepted 23 January 2017
Available online 8 March 2017

Keywords:
Multimodal 3D retrieval
Convolutional Neural Networks
3D shape
Object retrieval
Sketch retrieval

ABSTRACT

We propose in this paper a new framework for 3D shape retrieval using queries of different modalities, which can include 3D models, images and sketches. The main scientific challenge is that different modalities have different representations and thus lie in different spaces. Moreover, the features that can be extracted from 2D images or 2D sketches are often different from those that can be computed from 3D models. Our solution is a new method based on Convolutional Neural Networks (CNN) that embeds all these entities into a common space. We propose a novel 3D shape descriptor based on local CNN features encoded using vectors of locally aggregated descriptors instead of conventional global CNN. Using a kernel function computed from 3D shape similarity, we build a target space in which wild images and sketches can be projected via two different CNNs. With this construction, matching can be performed in the common target space between same entities (sketch–sketch, image–image and 3D shape–3D shape) and more importantly across different entities (sketch-image, sketch-3D shape and image-3D shape). We demonstrate the performance of the proposed framework using different benchmarks including large scale SHREC 3D datasets.

1. Introduction

The rapid growth of the World Wide Web has raised the need and interest in developing efficient tools that search in large data collections in order to find relevant information. While text, images and videos were, and still are, the dominant type of visual information that is commonly shared, 3D model collections started to become part of the web. This is due in part to the emergence of commodity devices and easy-to-use modeling tools, but also to their importance to many fields of science including engineering, architecture, medicine, biology and digital entertainment industry. As a consequence, the need for efficient 3D search tools is growing. Such needs, however, may vary among different categories of users. Some, for example, may require access to 3D models using textual queries. Others may want to use sketches, images, or even other 3D models to query 3D data collections. Thus, a 3D search engine should provide a mechanism that enables users to search for relevant 3D models using as query (a combination of) different modalities. Here, by modalities we mean 3D meshes, 2D images (such as the ones available on the internet) and hand-drawn sketches. This, however, requires narrowing the gap between different modals or fusing these multimodal representations, which is still a very challenging and active topic in machine learning in general and in 3D shape analysis in particular.

In this paper, we address the issue of using multimodal representations for 3D shape classification and retrieval. Although there has been several papers that proposed mechanisms for shape retrieval, most of them consider only one modality [1,2,10]. Recently, Li et al. [3] proposed a framework in which they addressed the joint combination of two different modalities. To do so, they first compute an embedding from 3D model similarities based on handcrafted features. They then project images into the embedding using Convolutional Neural Network (CNN), which allows them computing distances between images and 3D models.

In this paper, we propose the use of three types of modalities; 3D shapes, 2D images and sketches. We then aim at discovering the explicit as well as the implicit relationships, which can be nonlinear, between the various modalities used for shape retrieval. For this end, we project the three modalities into a common k-dimensional space $\mathbb{T}$ in such a way that similar entities, treated as points in $\mathbb{T}$, regardless of whether they are 3D models, images, or sketches, will be close to each other in the target space $\mathbb{T}$. In order to construct the target space $\mathbb{T}$, we first start by computing shape signatures from a collection of 3D objects using deep CNN. We then propose to compute a mapping function P that maps the 3D shape signatures to some common specific space such that the

* Corresponding author.
E-mail address: hedi.tabia@ensea.fr (H. Tabia).

http://dx.doi.org/10.1016/j.neucom.2017.01.101
0925-2312/

dot product between mapped signatures is as close as possible to an original kernel representing the similarity between two shapes. Once the target space $\mathbb{T}$ of the shape signatures is constructed, we project onto it, via a different CNN architecture, two other modalities, namely images and sketches, such that similar objects independently whether they are represented as 3D models, images or sketches, can be easily identified.

We show that by using this framework, one can achieve good performance in multimodal 3D shape retrieval. We demonstrate the utility and performance of the proposed approach for retrieving 3D shapes using different modalities and validate our approach using various benchmarks including large scale SHREC 3D datasets.

1.1. Related works

In this paper, we propose a framework for embedding multiple entities into a common target space in which similarities in both the same as well as cross entities can easily be computed. We study three different entities, namely, hand-drawn sketches, 2D images and 3D mesh models. Several papers have studied this problem in the context of image and video retrieval using multiple modalities such as text, audio, images, and videos [4–6]. In this section, we survey the methods that are most related to our work including 3D shape retrieval, image based shape retrieval and sketch based shape retrieval. Please refer to [7] for a review of the state-of-the-art in image retrieval.

3D shape retrieval. 3D shape retrieval has been a very active field of research in the past decade. We refer the reader to some of the recent surveys on the topic [2,8–12]. In this section, we do not aim to provide a full taxonomy of the topic. We instead view the existing methods from the perspective of the way the features that have been used are computed. Early methods, for instance, used handcrafted descriptors in order to characterize the geometric and topological attributes of 3D shapes [13–22]. Their performance, however, is often limited, particularly when dealing with medium or large databases, since they do not capture the essence or semantics of the shapes being indexed. More recent methods used learned descriptors involving either supervised or unsupervised feature extraction [23–26]. Approaches that use Convolutional Neural Networks (CNN) help discover automatically efficient shape representations from voxelized [23] or depth [25] data.

While there is a large amount of research dedicated to designing efficient shape descriptors, little attention has been given to the way 3D databases are queried. Most of the existing techniques use 3D models as queries in order to retrieve other 3D models that are similar. This is, however, not practical since often users want to retrieve 3D models for which they only have either an image or an abstract description, which can be transposed into a hand-drawn sketch.

Cross-modality 3D shape retrieval. Retrieving 3D shapes using 2D images or 2D/3D sketches as search queries requires a mechanism for comparing descriptors computed on images with descriptors computed on 3D models. This is not a straightforward process since images and 3D models have different representations and thus lie in two different spaces. Moreover, the features that can be extracted from 2D images are often different from those that can be computed from 3D models and thus, direct comparison is not feasible. Daras and Axenopoulos [27] used global description method for comparing 3D shapes and 2D images. More recently, Li et al. [3] proposed a joint embedding (using CNN) of images and 3D shapes into a common space in which they can compute a similarity between both entities. The embedding space in [3] is constructed from 3D shape similarities computed from handcrafted descriptors [28].

Methods that used sketch-based 3D shape retrieval aimed at finding a mechanism for mapping 3D shapes into a space in which it can be compared with 2D sketches, see [2,29] for a survey. Some techniques extract silhouettes from 3D models by projecting them into a binary image. The projected silhouettes are then compared with the 2D sketches. This technique has recently been used in different works such as Kanai [30], Li et al. [31], and Aono and Iwabuchi [32]. From the literature, we also can find the contour or outline feature view which has also been used as a series of points where the surface blends sharply and becomes invisible to the viewer [32–35]. Other techniques use suggestive contour feature views [33] to build sketch-based 3D model retrieval systems [31,36–39]. Suggestive contours are contours in the nearby views, that is, they will become contours after slightly rotating the model. Wang et al. [40] have recently proposed to learn feature representations using CNN and suggestive contour views for sketch based shape retrieval.

1.2. Contributions

Fig. 1 overviews the approach proposed in this paper. The contributions of this work are three-fold. First, we design and develop an efficient framework for multimodal shape querying that involves 3D shapes (3D meshes), 2D images and hand-drawn sketches. Second, we propose a novel 3D shape descriptor based on local CNN features encoded using vectors of locally aggregated descriptors (VLAD) instead of conventional global CNN. Third, using a kernel function computed from 3D shape similarity, we finally build a target space in which wild images and sketches can be projected through two different CNNs learned from 3D shapes. After the target space construction, matching can be performed in the common space between the same entities (sketch–sketch, image–image and 3D shape–3D shape) and more importantly across different entities (sketch-image, sketch-3D shape and image-3D shape). Note that, the proposed framework can be also used for single entity retrieval using multiple entities as well as multiple entities retrieval using single or multiple entities.

2. 3D shape description

Prior to feature extraction, we normalize the pose and scale of the 3D objects in order to ensure invariance to translation and scaling. Here, we do not perform the pose normalization for rotation because the locations of local features are completely ignored in our method. We first translate the 3D objects to their center of mass and then scale them so that their minimum bounding sphere is of radius 1 [41]. We then represent a 3D object using a set of 2D views captured by virtual cameras distributed uniformly around the object. In order to capture all the important features of the object, we use a large number of views (100 in our implementation) and capture 2D images of size 256×256. In the following, we present our method used for extracting and encoding CNN features for 3D shape description. We start by describing the method for extracting features and then present the VLAD encoding scheme used to aggregate the CNN features into a compact representation.

2.1. CNN based local features

Recently deep learning has successfully been used in many computer vision applications [42]. Specifically, CNN with a deep network has effectively outperformed handcrafted features and shallow methods in learning complicated 2D structures of an input image [43–45]. In this work, we propose the use of local CNNs [46,47] for generic feature extraction from 2D views rendered from a 3D shape. We use the pre-trained AlexNet [43] CNN. The network contains eight layers with weights; the first five are convolutional and the remaining three are fully connected. The first convolutional layer filters the $224 \times 224 \times 3$ input image with 96