Computers & Graphics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Computers & Graphics

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

Special Issue on CAD/Graphics 2017

Joint analysis of shapes and images via deep domain adaptation

Q1 Zizhao Wu[a,*], Yunhui Zhang[a], Ming Zeng[b], Feiwei Qin[c], Yigang Wang[a]

[a] School of Media and Design, Hangzhou Dianzi University, China
[b] Software School, Xiamen University, China
[c] School of Computer Science, Hangzhou Dianzi University, China

ARTICLE INFO

Article history:
Received 15 June 2017
Revised 7 July 2017
Accepted 9 July 2017
Available online xxx

Keywords:
3D shape recognition
Joint analysis
Cross-modal retrieval
Convolutional neural network
Domain adaption

ABSTRACT

3D shapes and 2D images usually contain complementary information for each other, and thus joint analysis of both of them will benefit some problems existing in different domains. Leveraging the connection between 2D images and 3D shapes, it is potential to mine lacking information of one modal from the other. Stemming from this insight, we design and implement a CNN architecture to jointly analyze shapes and images even with few training data guidance. The core of our architecture is a domain adaptation algorithm, which builds up the connection between underlying feature spaces of images and shapes, then aligns and correlates the intrinsic structures therein. The proposed method facilitates the recognition and retrieval tasks. Experiments on the shape recognition tasks show that our approach has superior performance under the difficult setting: zero-shot learning and few-shot learning. We also evaluate our method on the retrieval tasks, and demonstrate the effectiveness of the proposed method.

1. Introduction

Q2 With the rapid development of Internet, massive data in multiple modalities such as images, videos, and 3D models are emerging. These heterogeneous data are usually associated to represent the same entity. For example, one can generate images, videos, and 3D models respectively to depict an object in terms of the shape, the color, the texture or the motion. The explosive increase of multimedia data has brought the challenge of how to effectively recognize, retrieve and organize these resources. Significant efforts have been devoted for these tasks, however, most of such efforts handle these modalities of data separately, and do not take full advantage of the complementary information that exists in different domains. This is especially the case in the computer graphics and the computer vision communities, where the differences in properties of viewpoint, lighting, background, occlusion, as well as data representation are most prevalent, hence hinder cross-domain analysis.

In recent years, a few works [1–3] have been proposed to address the problem of joint analysis between 3D shapes and 2D images. These works have demonstrated the great potential for solving some difficult tasks in one domain by exploiting full advantage of the complementary information in the other, which include cross-view image retrieval, cross-modal retrieval, text based shape retrieval, 3D repository and 2D repository filtering, 3D model alignment, 3D shape recognition, etc. We note that most of these works employed the Convolutional Neural Network (CNN) to learn the feature vectors for shapes and images, and achieved remarkable performance in many fields. These algorithms usually represent each 3D shape as a set of rendered images from different views around the model, in order to overcome the gap of data representation between 3D shapes and CNN models. However, with the exception of Li et al. [3], they all treat the rendered images from shapes and 2D images without distinction, which will lead the problem of domain bias [4], due to the great discrepancy in appearance between them.

In this work, we present a novel architecture for joint analysis of shapes and images, our method overcomes the discrepancy between images of different domains by introducing a domain adaptation algorithm, which leverages knowledge across domains. Specifically, a joint source and target convolutional neural network architecture is introduced to learn the feature representations of different domains. Domain adaptation is carried out in the learning process, aiming at fusing the feature representations of different domains into a shared latent space. The shared feature space is semantically meaningful, i.e. data of nearby points hold similar semantic information, regardless the modalities they belong to.

An appealing feature of the proposed algorithm is its capacity of leveraging information from one domain to the other, which facilitates the task of dealing with insufficient training data in one domain. As a result, our architecture can be used to few-shot learning [5], when a small amount of target labeled data is available from

* Corresponding author.
E-mail address: wuzizhao@163.com (Z. Wu).

http://dx.doi.org/10.1016/j.cag.2017.07.013

each category, and zero-shot learning [6], when a small amount of target labeled data is available from a subset of the categories. We first evaluated our approach for the recognition on some popular data sets, the results show that our method matches state-of-the-art performance while requiring less training data of the target. Furthermore, we demonstrate that our method has the ability to correctly predict object category labels for unseen categories, i.e. zero-shot classification, by leveraging the knowledge across domains, which is hard for the state-of-the-art recognition methods. We also evaluated the performance of our approach on cross-modal retrieval of shapes and images, and demonstrated its effectiveness.

The main contributions of this paper include (1) we introduce a novel algorithm for joint analysis of 3D shapes and 2D images; and (2) to the best of our knowledge, we are the first to address the few-shot learning and the zero-shot learning problems in the 3D shape recognition field.

The remainder of this paper is organized as follows. We review the related work in Section 2, then present the overview of the proposed algorithm in Section 3. The details of our domain adaptation algorithm is described in Section 4. We show some experimental results on benchmark datasets in Section 5, followed by conclusions and future work in Section 6.

2. Related work

Our method is related to prior work on shape descriptors, deep learning, domain adaptation and joint image-shape embedding, which we briefly discuss below.

2.1. Shape descriptor

Shape descriptor is an informative representation of the shape, aiming at facilitating tasks such as shape matching, recognition, retrieval, and so on. A large variety of shape descriptors has been developed in the computer graphics community. Most of the earlier shape descriptors focus on geometric properties of the shape such as shape contexts [7,8], shape distribution [9], local diameter [10], volume descriptors [11], spherical harmonics [12], conformal factors [13], Heat Kernel Signature (HKS) [14]. Some view-based descriptors also received widely attention, Cyr and Kimia [15] utilize multi-view projections to identify 3D objects and their poses. Chen et al. [16] propose Light Field Descriptor (LFD), which extracts Fourier descriptors from a set of 2D projections of views.

Instead of designing features according to human prior knowledge, discriminative feature learning provides an alternative way to characterize shapes. This especially benefits from the fast development of deep learning techniques [17], where the learned features lead to significant performance boost in classification and recognition tasks [18,19].

It is only very recent that a few works attempt to tackle 3D shape related problems via deep learning methods, such as classification, recognition and retrieval. Wu et al. [20] work with volumetric representation of 3D shapes, obtaining good results of shape classification on Princeton ModelNet [20]. Zhu et al. [21] utilize Auto-Encoder to learn 3D shape feature with multi-view depth images, leading to accurate 3D shape retrieval. One of the limitations of using 3D volumes as input is the loss in detail when shapes are voxelized. Su et al. [22] propose multi-view CNN(MVCNN) for 3D shape recognition where the features of multiple views are integrated with an extra CNN. Generating shape descriptors based on multiple views can be time-consuming and challenging for real-time retrieval. Bai et al. [23] propose real-time shape retrieval, using GPU acceleration and two inverted files (GIFT).

Our work on 3D shape recognition is similar to MVCNN in that we use deep CNN model to learn shape descriptors. It differs by the fact that our CNN model considers cross-domain input while MVCNN considers only one modality, which strengthens our method to handle with few-shot learning problem and even zero-shot learning problem.

2.2. Convolutional Neural Networks

Recently, Convolutional Neural Networks (CNNs) have been shown to be extremely effective for a variety of visual recognition tasks [18,19,24]. Though many CNN architectures have been proposed, such as AlexNet [18], GoogleNet [25], VGG [26], the network structure of AlexNet remains a popular structure, which consists of five convolutional layers with two fully-connected layers followed by a softmax layer to predict the class label. The network is capable of generating useful feature representations by learning low level features in early convolutional layers and accumulating them to high level semantic features in the latter layers.

There are several deep learning frameworks that efficiently implement the above popular networks, such as Berkeley Caffe [27] and Google Tensorflow [28]. We use Caffe in this paper.

2.3. Domain adaptation

Domain adaptation establishes knowledge transfer from the labeled source domain to the unlabeled target domain by exploring domain-invariant structures [29]. Recent studies have shown that deep neural networks can learn more transferable features for domain adaptation [30–32], which produce breakthrough results on some domain adaptation datasets. However these methods solve the problem of domain adaptation within the same modality. It is unclear how this can be done when moving across modalities. To address this issue, some notable approaches focus on the problem of jointly embedding or learning representations from multiple modalities into a shared feature space to improve learning [33] or enabling zero-shot learning [34,35].

Our work is primarily motivated by Tzeng et al. [36], that introduces a deep CNN model for the domain adaptation problem. We introduce this architecture to facilitate our task for joint analysis of shapes and images, and further make some optimizations to the original model.

2.4. Joint image-shape analysis

Multi-modal feature learning has been researched thoroughly over the past years [33], whereas only few works have addressed the problem in the computer graphics field in recent years. Herzog et al. [1] suggest a new method for structuring multi-modal representations of shapes and keywords, and adds images and sketches to the mix. This method builds the embedding mainly upon the class co-relation and hand-crafted descriptors. Hueting et al. [2] introduce a system for joint images and shapes processing to

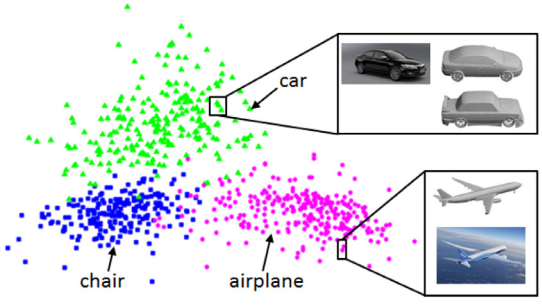

Fig. 1. Our method on joint analysis of shapes and images, and learn to obtain a joint representation in a shared latent space.

Please cite this article as: Z. Wu et al., Joint analysis of shapes and images via deep domain adaptation, Computers & Graphics (2017), http://dx.doi.org/10.1016/j.cag.2017.07.013