Intelligent Visual Media Processing: When Graphics Meets Vision

Ming-Ming Cheng[1], Qi-Bin Hou[1], Song-Hai Zhang[2], Paul L. Rosin[3,1]

[1] *CCCE & CS, Nankai University* [2] *TNList, Tsinghua University* [3] *Cardiff University*

E-mail: {cmm.thu,andrewhoux}@gmail.com; shz@tsinghua.edu.cn; rosinpl@cardiff.ac.uk

Received submit date

Abstract The computer graphics and computer vision communities have been working closely together in recent years, and a variety of algorithms and applications have been developed to analyze and manipulate the visual media around us. There are three major driving forces behind this phenomenon: i) the availability of big data from the Internet has created a demand for dealing with the ever increasing, vast amount of resources; ii) powerful processing tools, such as deep neural networks, provide effective ways for learning how to deal with heterogeneous visual data; iii) new data capture devices, such as the Kinect, bridge between algorithms for 2D image understanding and 3D model analysis. These driving forces have emerged only recently, and we believe that the computer graphics and computer vision communities are still in the beginning of their honeymoon phase. In this work we survey recent research on how computer vision techniques benefit computer graphics techniques and vice versa, and cover research on analysis, manipulation, synthesis, and interaction. We also discuss existing problems and suggest possible further research directions.

Keywords Computer graphics, computer vision, survey, scene understanding, image manipulation

1 Introduction

Computer graphics and computer vision begin with inverse problems. Traditional computer graphics starts with geometric models and produces photorealistic images, with emphasis on interaction, synthesis, *etc.* As illustrated in Fig. 1, traditional computer vision starts with input image sequences and produces geometric models, with an emphasis on semantic understanding, matching, *etc.* The trend that these two fields are converging has been noticed since the 1990's [1]. More and more computer graphics researchers are trying to use vision techniques to help create and manipulate visual scenes as efficiently as possible [2]. Using computer graphics techniques to help solving vision problems is also becoming popular [3–5].

To date, billions of internet images, videos and 3D models have been created and are shared on the internet everyday [6]. Such big visual data have hastened a variety of image/video/geometry analysis and manipulation applications, by providing ever existing vast amount of resources which enable novel applications that are otherwise impossible by traditional methods. On one hand, enabling smart computer graphics tools to intelligently create compelling results with minimal user interaction requires computer vision techniques to extract semantic components and knowledge from the huge volume of available data. *e.g.* deep convolutional neural networks [7], continually boost state-of-the-art performance for a wide range of tasks, but typically rely on expensive, large scale, human labeled data to learn from. To overcome this bottleneck computer graphics techniques can be developed to automatically help learning algorithms to collect training examples. The bond between computer graphics and computer vision has been further blurred by the emergence of RGBD image capturing devices, such as Microsoft Kinect, Intel RealSense, Apple PrimSense, *etc.*

Footnotes

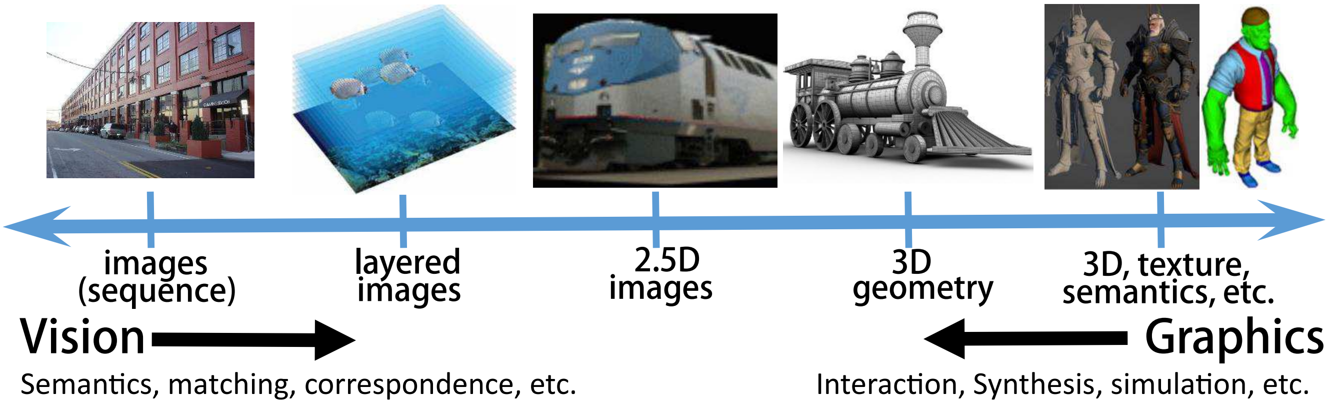

images (sequence)
layered images
2.5D images
3D geometry
3D, texture, semantics, etc.
Vision
Semantics, matching, correspondence, etc.
Graphics
Interaction, Synthesis, simulation, etc.

Fig. 1. Graphics and vision spectrum: traditional graphics starts on the right with more geometry based, while traditional vision starts from more image based on the left. Currently, graphics and vision tends to fuse together, with emphasis on interaction and semantics understanding respectively.

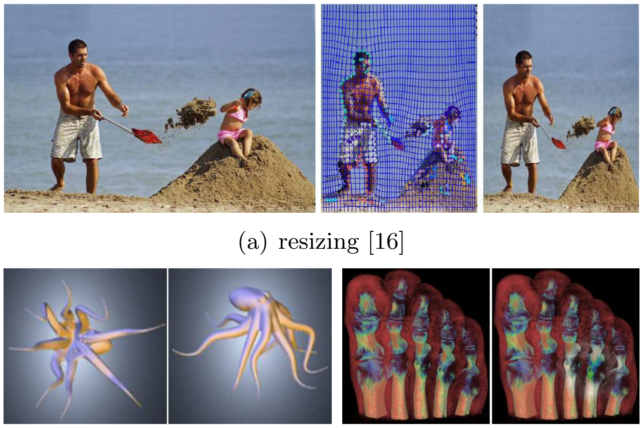
The RGBD images directly associate image and geometry processing algorithms, making productive collaboration between computer graphics and computer vision much easier.

In this paper, we survey recent research on how computer vision techniques benefit computer graphics techniques and vice versa. These topics include saliency aware media processing (Sec. 2), content understanding for smart image manipulation (Sec. 3), depth estimation and 3D modeling (Sec. 4), and data synthesis for visual learning (Sec. 5). We also discuss existing problems and suggest possible further research directions (Sec. 6).

2 Saliency Aware Media Processing

The concept of saliency originates in the study of human perception, and relates to how some parts of the scene appear to be more important than others. The computation of saliency is normally considered to be primarily a bottom-up (and therefore general purpose) process, based on local image features such as colour and contrast [8–11]. Computer vision widely uses saliency, as it provides a lightweight means to identify the most informative and important areas in a scene, such as the foreground objects. Another category of the use of saliency is to help analyse the quality of images generated by image and video compression and processing algorithms. For instance, the artifacts created by compression need to be quantified in a perceptually aware manner, and so saliency is used as a convenient proxy [12]. Many algorithms have been developed for salience detection, and readers are referred to the recent surveys for more details [13–15].

There are also many instances in graphics that can benefit from employing saliency to predict human perception. One category is the set of applications which manipulate an image or 3D model, and incur some error during that process, *e.g.* image resizing [16] or mesh simplification [17]. Better results will be obtained if the errors can be restricted to the non-salient parts of the data rather than the salient parts. Another category is when some part of the data is to be enhanced by amplification, *e.g.* boosting image intensities [18] or surface curvature [19]. Restricting the amplification to the salient regions tends to produce less confusing and more attractive results.



(a) resizing [16]

(b) view selection [20]

(c) visualization [21]

Fig. 2. Saliency aware media processing. Images are reproduced from the corresponding references.

2.1 Content aware resizing

When displaying image content at different sizes and aspect ratios, content distortion is a common phenomenon. A smart way for enhancing the user experience is to make sure that prominent objects should be kept similar to their original contents and any distortion should be restricted to less important regions.

The seam carving approach [22] was an early classic work in content aware image resizing. It works by greedily removing/inserting one-dimensional seams passing through regions which are estimated, via saliency detection, to be of less importance. Wang *et al.* [23] further improved the speed issue and overcame jagged edges via continuous optimization instead of discrete seam carving. Inspired by conformal energy in geometry processing, Zhang *et al.* [16] proposed a real-time convex optimization solution with a closed form solution, see Fig. 2(a). Several authors have extended the bottom-up salience measure to incorporate higher-level aspects, *e.g.* object semantics [24] and symmetry [25]. Image re-targeting has also been extended to deal with image enlarging [26], stereo images [27], video sequences [28] and stereoscopic 3D video [29].

Resizing 3D models, while requiring the important structure of the underlying models to be retained as much as possible, is of great importance. Significant research effort has gone into this area in order to easily place 3D models into different scenes. Miao and Lin [30] constructed a quadratic energy function, that incorporated an edge sensitivity measure, to help guide salient feature-preserving model resizing, Jia *et al.* [31] designed a region-based descriptor to compute the saliency of each region based on its contrast to neighboring regions and a hierarchical method for computing saliency. They showed that by optimizing a global energy function on the mesh, visually appealing mesh resizing results can be obtained.

2.2 Shape simplification and enhancement

Mesh saliency was first introduced by Lee *et al.* [20], which used a center-surround operator on Gaussian-weighted mean curvatures at multiple scales. They used a weighting map derived from the computed saliency map to guide the order of vertex pair contractions to produce mesh simplification, and showed their superiority to other methods, see Fig. 2(b).

Song *et al.* [32] also proposed a mesh saliency method for mesh simplification, which incorporated the Conditional Random Field (CRF) framework with a saliency detection process. In this approach, a multi-scale representation for meshes is first generated and then a CRF is adopted to detect saliency regions using neighborhood consistency. Zhao *et al.* [17], provided an alternative approach for mesh simplification using mesh saliency [32]. They produced a saliency map by diffusing the shape index field with the non-local means filter. Recently, Castelló *et al.* [33] presented a view based method for surface simplification using mesh saliency. They first defined a new simplification error metric to improve the visual quality of the simplified models and then used viewpoint saliency as a weighting factor of the quality of the viewpoint.

Enhancing shape signatures so that important features could be highlighed for viewing and artistic reasons also requires estimation of mesh saliency. In [34], Miao *et al.* developed a saliency guided shading scheme for shape depiction by incorporating the visual saliency measure of a polygonal mesh into the normal enhancement operation. Due to the introduction of the visual saliency measure of the 3D shape, this approach can adjust the illumination and shading to enhance the geometric salient features of the underlying model by dynamically perturbing the surface model. In [19], Miao *et al.* presented a visual saliency based shape depiction scheme for relief surface. They combined three different bottom-up feature maps and defined a new multi-channel salience measure. By incorporating this salience measure into an exaggeration operation, a saliency-guided shape depiction scheme was developed. Understanding salient features is also been used to preserve important shape features during mesh deformation [35].

2.3 Visualization

Visualization aims to guide the observer's attention to the relevant aspects of the representation. Therefore, it is important to model aspects of the human visual system, and saliency provides a simple approach to doing so.

Kim *et al.* [21] designed a visual saliency based operator to help enhance selected regions of a volume, see Fig. 2(c). They plugged the operator into an existing visualization pipeline and showed that based on the center-surround mechanisms of the human visual system, saliency-guided enhancement for volume visualization was effective and could be applied in several contexts. Besides, Jänicke and Chen [18] proposed a metric to measure the quality of a visualization. They believed that the distribution of saliency over a visualization image could be thought of as an important measure of the quality of the visualization. Meanwhile, they provided an approach to compute such a metric for a visualization image in the context of a dataset.

Semmo *et al.* [36] used salience to control the use of different graphic styles and levels of detail for visualizing a given view of a 3D city model, in order to direct the viewer's gaze to the most important information. Salient regions were rendered with photorealistic graphics, while non-salience regions were rendered with non-photorealistic graphics, which provided image abstraction. The different rendering styles were combined in a seamless manner using alpha blending.

2.4 3D printing

3D printing as an additive manufacturing work recently has been applied to a wide range of applications on account of its ability to facilitate rapid fabrication of objects of any shape. Therefore, without doubt, it is one of the hot topics in graphics.

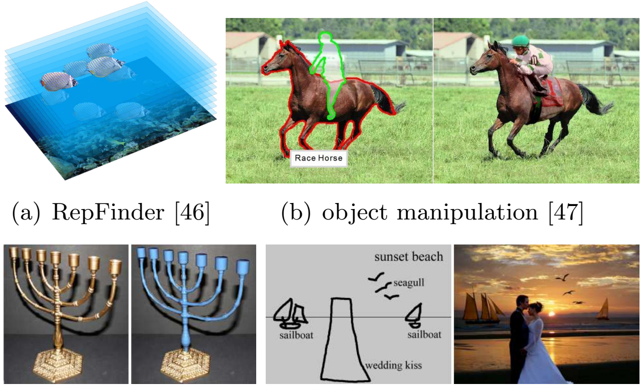
Song et al. [37] presented a voxelization-based method for 3D printing which dispenses with connectors, glue, and screws while proposing to connect the printed 3D parts by 3D interlocking. The object is decomposed into a set of initial 3D interlocking parts. To improve their aesthetic property, these cutting seams are refined by swapping voxels among adjacent 3D parts so as to avoid putting cutting seams across salient parts. The salience of boundary voxels is estimated via a 3D mesh salience [20] measure.

In [38], Wang *et al.* present an adaptive width slicing scheme for 3D printing systems. In order to reduce the printing time while at the same time maintaining the visual quality of the printing results, they optimise a cost function involving those two factors, the latter being computed using a saliency based metric. Furthermore, they gain greater efficiencies by developing a saliency based segmentation approach to partition an object into subparts, and then optimize the slicing of each subpart separately.

3 Content Understanding for Smart Manipulation and Synthesis

While most existing computer graphics tools, *e.g.* Adobe PhotoShop and Autodesk Maya, mainly support low-level operations and are typically employed for touch-up or local enhancement of visual content [39, 40], high level image editing techniques that allow users to specify large scale meaningful changes using simple interactions have recently gained great research attention [41–44]. Psychologists believe that humans process and organize visual information based on relations between scene structures [45]. Allowing the user to manipulate content at the level of objects in the scene, while being aware of scene structure, is an attractive editing modality that is aligned with our mental data representation.

However, to mimic real world user experience with physical environments and to enable object level manipulation, we need to understand the content in the visual data and overcome four major challenges: i) visual data are composed of ungrouped elements, *e.g.* pixels and polygons, rather than semantic objects; ii) recovering geometry information about how objects are arranged in 3D is often an ill-posed problem and unlikely to be solved in the near future; iii) correlations between objects are hard to infer but are critical to maintain realism during the editing processing; iv) semantic constraints about how objects should behave after user adjustments requires not only information about the target being manipulated, but also prior knowledge that exists in human experience and big internet data.



(a) RepFinder [46] (b) object manipulation [47]

(c) 3-Sweep [43] (d) image montage [48]

Fig. 3. Content understanding for smart manipulation and synthesis. Images are reproduced from the corresponding references.

3.1 Smart Manipulation

With an increased level of content understanding provided by computer vision techniques, visual media manipulation tools could more intelligently infer user intentions, thus reducing the requirement of precise user input and tedious interactions.

In [46], the RepFinder system detects approximately repeated objects and builds dense correspondences between them, to enable object level manipulation whilst preserving correlations among the repetitions, see Fig. 3(a). Goldberg *et al.* [47] proposed a data-driven approach to interactively manipulating objects in a photograph using related objects obtained from internet images, see Fig. 3(b). By matching the candidate object with user input strokes, the system automatically finds candidate objects from the internet, enabling a range of novel editing experiences that is impossible with low-level operations (*e.g.* removing part of an object to reveal its interior). Lu *et al.* [49] further enables object level manipulation for timeline editing of video contents.

Understanding object shapes and their perspective relations is also crucial for high level image manipulation experience. Zheng *et al.* [50] explore user interaction to creates partial scene reconstructions based on cuboid-proxies structures. Such partial scene structure allows a range of intuitive image edits, so that users only need to provide high-level semantic hints and the system ensures plausible operations that mimic real-world behavior, which are otherwise difficult to achieve. In [43], the 3-Sweep system further uses general cylinders and cuboid structure to understand the components of the shape, their projections, and relationships, see Fig. 3(c). Besides object geometry, rough scene geometry is also important for high level image editing applications. Iizuka *et al.* [51] proposed a system in which the user can move objects in an image whilst ensuring that object size and object overlap are automatically adjusted. This is achieved by estimating the perspective structure of the scene in a single image with the assistance of user-drawn strokes. Estimating object shape and scene geometry from a single image is inherently an ill posed problem. The success of these methods such as [51–54] typically rely on user interactions (*e.g.* strokes [54] and bounding boxes [55]) and simplifying assumptions (*e.g.* cuboid-proxies [50] and general cylinders [43]).

High level graphics applications which rely on semantic meanings [56] or scene geometry of complex objects [44, 57] often require information that does not explicitly exist in a single image. Knowledge acquired from large collections of visual data are useful for obtaining plausible results by resolving ambiguity and uncertainty. In the ImageSpirit [56] system, Cheng *et al.* proposed treating nouns as object labels and adjectives as visual attribute labels. This allows novel verbal interaction based on semantic knowledge learned from a set of images with dense object class and attribute labels. Kholgade *et al.* [44] proposed to leverage the structure and symmetry in stock 3D models for estimating illumination and completing the hidden parts of an object seen in a single photograph. Huang *et al.* [57] jointly analysed web images and shape collections for single view reconstruction. Such joint analysis regularizes the optimization formulation and stabilizes correspondence estimation, thus enabling reconstruction of different objects using a smaller collection of existing 3D models.

3.2 Visual Content Synthesis

Chen *et al.* [48] developed an interesting system named Sketch2Photo that was capable of automatically converting a simple freehand sketch, along with a few text label annotations, into a realistic picture, see Fig. 3(d). Due to the fact that the pictures are found by searching the Internet, many inappropriate results may be produced. In order to overcome this drawback, a filtering scheme is used to eliminate inappropriate images, and an image blending algorithm is adopted to find an optimal combination of discovered images.

In [58], the PoseShop system was proposed for constructing a segmented human image database that was used to synthesise personalized comic-strips. By employing computer vision techniques, only minimal manual intervention was required. Segmentation followed by further filtering [48] was able to produce four hundred thousand segmented human characters of sufficient quality. The images were analysed so as to automatically provide clothes descriptions that can be used by the user alongside the text attributes to query the database when constructing the comic-strips. Tanahashi *et al.* [59] proposed an efficient framework for storyline visualization from streaming video data. Hasegawa and Saito [60] presents a method for synthesis stroboscopic image from video sequence for sports analysis.

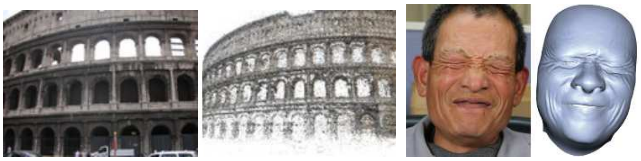
Lalonde *et al.* [61] developed a system that can insert new objects into existing photographs. A new automatic algorithm is presented so as to improve the object segmentation and blending, estimate true 3D object size and orientation, and estimate scene lighting conditions. Moreover, an intuitive user interface is provided, which is able to make object insertion much faster.

In [62], Xu *et al.* presented a system that could automatically convert a freehand sketch drawing containing multiple objects into a semantically valid and well arranged scene composed of 3D models. By performing co-retrieval and co-placement of 3D models, the amount of user intervention needed for sketch-based 3D modeling is greatly decreased.

Chia *et al.* [63] designed a new colorization system that can colorize grayscale photos with less manual labor. The user provides a semantic text label and selects an automatically generated foreground object segmentation, and this system can automatically download and filter suitable relevant images using a new filtering method. These then provide reference images that are suitable for driving the colorization process.

4 Depth Estimation and 3D Modeling

Scene modeling from imagery data is one of the main tasks of both computer vision and computer graphics, and thus also the point at which the above two fields merge or diverge. Many analysis methods which originate in the graphics domain, such as 3D geometry analysis, are introduced into depth estimation and 3D modeling to produce much more accurate 3D geometric data of the scenes. Thus, this section describes applications in both graphics and vision that use techniques such as structure from motion to recover geometry and also synthesise imagery.



(a) building Rome in a day [64] (b) facial capture [65]

Fig. 4. Depth estimation and 3D modeling. Images are reproduced from the corresponding references.

4.1 Modeling 3D Scenes

Unlike active scene modeling systems, such as structured light projectors, vision based modeling aims at creating a 3D model of the real world by simply taking images of it mainly using stereo matching. Structure from motion (SfM) is a passive modeling technique that simultaneously estimates 3D scene structure and camera poses from 2D image sequences. Although the problem of SfM was proposed several decades ago [66], it was not until recently that progress became dramatic due to the advances in computing performance [67]. Applications based on SfM also occur in scene reconstruction and 3D object modeling.

Snavely *et al.* developed a photo browser [68] which takes unstructured collections of photos of sites as input and computes the viewpoint of each photo as well as a sparse 3d point cloud of the scene. The results enable the user to explore the photos in 3D space. Later Agarwal *et al.* presented a system named 'Building Rome in a Day' [64], see Fig. 4(a). The system can handle an extremely large quantity of photos (*e.g.* the results returned by Google when searching for a city). Frahm *et al.* [69] introduced a dense 3D reconstruction system which is able to deal with about 3 million internet images within the span of a day on a single PC with a GPU. Recently, Fuhrmann *et al.* implemented the 'Multi-View Environment' [70], an end-to-end image-based geometry reconstruction tool which takes the photos of a scene as input and produces a textured surface mesh as the result.

Various applications can be developed using vision based scene modeling and point cloud matching and rendering. Ceylan *et al.* [71] coupled structure-from-motion and 3D symmetry detection for urban facades. The recovered symmetry information along with the 3D geometry enables image editing operations maintaining consistency across the images. Kopf *et al.* [72] proposed an algorithm to create videos with smooth camera motion from first-person videos, which are captured during sports and thus suffer from erratic camera shake. This work employs SfM to estimate the camera pose for each frame and re-renders the video using a smooth camera path.

Since SfM can recover the structure of large scaled scenes, it can be exploited for positioning. Recent studies have developed algorithms to recognize the location of the query image from the point cloud produced by SfM. Tan *et al.* [73] presented a monocular SLAM (Simultaneously Localization and Mapping) system which uses a special keyframe representation and updating method to handle dynamic environment. Li *et al.* [74,75] proposed an approach to use a sparse transform to the joint estimation of 3D shapes and motions, while using wa avelet basis to fit 3D shape trajectory. The system demonstrated robust performance when handling nonrigid target with occlusion.

4.2 Facial performance

Facial expression plays a critical role in almost all aspects of human interaction and face-to-face communication. As such, face and facial performance modeling has long been considered a grand challenge in the field of computer graphics and vision. Using special equipment, such as facial markers [76], camera arrays [77], and structured light projectors [78], enables the capture of high fidelity 3D facial geometry, which are crucial to be captured especially in film and game production.

Recently, techniques have been developed which are more suitable for consumer-level capture approaches [79]. They do not require such special equipment, but instead are based on the co-modeling of 3D geometry and 2D landmarks in videos of facial expressions. Cao *et al.* [80] present a fully automatic approach to real-time facial tracking and animation with a single video camera, which can reach the same level of robustness and accuracy as demonstrated in RGBD-based algorithms. This method introduces a Displaced Dynamic Expression (DDE) model that simultaneously represents the 3D geometry of the user's facial expressions and the 2D facial landmarks which correspond to semantic facial features in video frames. By learning a generic regression model from public image datasets, this approach can be applied to arbitrary video cameras to infer accurate 2D facial landmarks as well as the 3D facial shape without any training. Cao *et al.* [65] further developed facial tracking system that captures human performance with high fidelity in realtime, see Fig. 4(b).

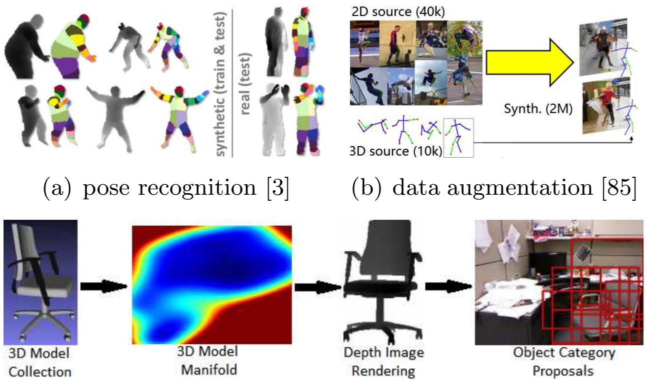
4.3 Human Motion Capture

Motion capture is the process of recording the movement of people (animals or jointed rigid structures in general), which is one of the main demands of scene modeling. It is mainly used in connection with capturing large scale body movements, which are the movements of the head, arms, torso and legs. Motion capture is widely used in education, training, sports and recently computer animation for television, cinema, and video games, virtual reality, which are mainly in the graphics domain. Although traditional methods are often based on the capture and processing of active or passive sensors, i.e. acoustic, inertial, LED, magnetic or reflective markers, the vision based approaches allow in principle for touch-free capture and gradually being introduced into graphics and VR applications. Recently, 4D Performance Capture (4DPC) [81] has been introduced to capture shape, appearance and motion of the human body from multi-view videos. It derives a sequence of reconstructed 3D meshes with temporally consistent vertices and topology, which capture detailed surface dynamics plus associated video that can be projected onto the mesh. Making use of 4DPC data, Huang *et al.* [82] proposed a skeleton driven character animation by motion graph path optimization and a learnt part-based Laplacian surface deformation model.

Recent works focus on motion and appearance control to reproduce character animations, and use machine learning. Xia *et al.* [83] present a novel solution for real-time generation of stylistic human motion that automatically transforms unlabeled, heterogeneous motion data into new styles using an online learning algorithm that automatically constructs a series of local mixtures of autoregressive models (MAR) to capture the complex relationships between styles of motion. Pons-Moll [84] propose a new model called Dyna that is learned from examples and is able to produce realistic soft-tissue motions for a wide range of body shapes and motions.

5 Synthesize Big Data for Visual Learning

In recent years there has been an increased demand for data in computer vision. This is due in part to the widespread use of machine learning, as well as the increased emphasis within computer vision on large scale, rigorous testing. Consequently researchers are looking for efficient means with which to acquire or generate such large training and test tests.



(a) pose recognition [3]

(b) data augmentation [85]

(c) shape manifold [5]

Fig. 5. Synthesize big data for visual learning. Images are reproduced from the corresponding references.

Databases of 3D models provide examples from which we can learn models of scenes. Such 3D models provide rich information from which vision algorithms can learn from, such as shape, surface normal, materials, lighting, viewpoint, perspective, occlusions, *etc.* The problem is whether such synthesised data are of sufficient quality to be useful for computer vision algorithms, and so care needs to be taken to provide realistic characteristics such as noise and natural variations. This section provides three examples that use synthesized data for visual learning.

5.1 Pose Recognition

Human pose recognition from videos and images has been widely studied for decades. How to estimate human pose fast and reliably is challenging. This subsection will review some advanced pose recognition approaches using synthesized data.

Shotton *et al.* [3] proposed a real-time human pose recognition approach that transformed the difficult pose recognition task into a simple pixel-level classification problem by presenting an intermediate representation in terms of body parts, see Fig. 5(a). For training data, they designed a randomized rendering pipeline that randomly selected a set of parameters, such as height, weight, camera noise, *etc.* and then used computer graphics methods to render depth and body part images from 3D meshes. In the learning process, they employed simple depth comparison features that were 3D translation invariant and used randomized decision forests. With a huge database of synthesized image pairs very deep forests can be trained without overfitting.

In [86], Shotton *et al.* introduced two efficient approaches, body part classification (BPC) and offset joint regression (OJR), to predict the 3D positions of body joints from a single depth image. A similar rendering method as done in [3] was used for generating synthetic data that includes fully labeled training data, alongside real hand-labeled depth images, and test data. Both BPC and OJR use decision forests and simple depth-invariant image features. But differently, the BPC approach tries to infer a set of surface body parts that are aligned with the joints of interest, while the OJR approach tries to directly estimate the positions of interior body joints.

Rogez and Schmid [85] designed an image-based synthesis engine that combined image regions from different images to augment images and used the resulting images to train a CNN for 3D pose prediction, see Fig. 5(b). Their image-based synthesis engine is composed of two parts. A MoCop-guided image mosaicing is first used to stitch images patches together and then a pose-aware blending process is performed to improve the quality and erase patch seams. With training data, an end-to-end CNN is adopted for 3D body pose classification.

5.2 Object Detection

Object detection is one of the most challenging tasks in computer vision, and has made great success in recent years. Synthesized datasets from depth images further promote its development.

Song *et al.* [87] proposed to use depth maps for object detection. They developed a 3D detector to help overcome various impediments to recognition, such as the variations of texture, illumination, shape, clutter, *etc.* The training data is a collection of synthetic depth maps that are obtained by rendering 3D CAD models from hundreds of viewpoints. During depth rendering, features are extracted from the 3D point cloud, followed by an Exemplar-SVM classifier [88].

Peng *et al.* [89] used synthetic images to investigate the invariance of deep CNNs to various low-level cues and presented their own CNN for object detection. Given some 3D CAD models for each object, a set of synthetic 2D images are generated by simulating a variety of low-level cues, including shape, surface color, reflectance, and location, *etc.* They showed that if a model had been trained for detection task, it was unnecessary to incorporate synthetic images with simulated cues.

In [90], Gupta *et al.* used semantically rich image and depth features to do object detection. To generate more data for training and fine-tuning their network, they rendered the full 3D synthetic CAD object models from various viewpoints to produce synthesized scenes. At each pixel from the depth image they extracted three channels: horizontal disparity, height above ground, and the angle with respect to gravity. A modified R-CNN framework is used to produce rich features and to perform object detection.

Zheng *et al.* [5] generate object detection proposals by using compact 3D shape manifold. A low dimensional Gaussian Process Latent Variable Shape Space is trained. Then, shape variations are sampled from this manifold and then used for the training process, see Fig. 5(c).

5.3 Object recognition

2D object recognition has made great progress because of the development of deep networks. With the appearance of advanced devices that produce 3D point clouds, there is an increasing amount of works [91–93] that focus on developing 3D recognition using 3D convolutional networks.

Wu *et al.* [91] designed a convolutional deep belief network to model the joint probabilistic distribution over 3D voxel data. In order to train the deep network, a large-scale 3D CAD model dataset is generated by mapping each voxel to a binary tensor according to whether the voxel is inside the mesh surface. Synthetic data has also been demonstrated as a powerful tool for generating large scale annotated data for training text recognition neural networks [94, 95].

Wohlhart and Lepetit [93] introduced the efficient and scalable Nearest Neighbor search in a descriptor space to perform object recognition. They used a mixture of synthetic and real world data for training. The latter was created by regularly sampling viewpoints over a half-dome over the object mesh, and rendering the object in RGBD an empty background using Blender. A convolutional network is used to directly map the raw image patch to a compact and discriminative descriptor. They also used the Euclidean distance to evaluate the similarity between descriptors.

6 Discussion and Conclusion

We have reviewed a variety of recent studies in which computer graphics and computer vision techniques benefit each other. On the one hand, advanced vision techniques provide powerful tools for understanding and providing salient features, object segmentation, 3D geometry, scene perspective, semantic meanings, *etc.* With an improved degree of scene understanding, a number of image manipulation tools could be made more intelligent, by being aware of important object parts, being able to perform manipulations at the object level, or being able to guess user intentions. We note that there are still few large scale benchmarks for comparing the performance of different graphics applications that use vision techniques. This prevents the systematic study and boosting of performance that is often observed in pure computer vision work. With the rapid development of vision techniques, especially recent deep learning methods, we believe more and more vision analysis will becoming robust enough to support ever more vision applications.

On the other hand, graphics techniques have also been explored for synthesis of big visual data for pose recognition, object detection, object recognition, *etc.* There are also many analysis methods which originate in the graphics domain, such as 3D geometry analysis, which have been introduced into depth estimation and 3D modeling to produce much more accurate 3D geometric data of the scenes, or capture human motion and facial performance. However, although growing very quickly, the amount of graphics techniques that have been used in vision is still much less than in the other direction. More research effort is required to assist the creation of training data, the generation of candidate detections, assistance with the modeling process, *etc.*

Both the graphics and vision communities require total scene understanding for a variety of real world tasks. Such semantic understanding typically involves various individual tasks, which are highly correlated. To date, the majority of the research has been devoted to research in one or two tasks. Although such research is typically very deep, it is not broad enough to consider many of the vision and graphics tasks jointly, which would potentially enable a lot more cues to be exploited than is used in a typical computer vision or graphics system. Recently some pioneering work has jointly explored 3D modeling, object segmentation, user interaction, online learning, and camera localization [96–98]. Although these novel systems can only deal with simple visual scenes, and support a limited amount of scene understanding, they lead the way to a bright future using total scene understanding via jointly discovering, reconstructing, interacting and learning in the environment.

Acknowledgments

We would like to thank the anonymous reviewers for their useful feedback. This research was sponsored by NSFC (NO. 61572264), Huawei Innovation Research Program (HIRP), CAST young talents plan, and Tianjin Short-term Recruitment Program of Foreign Experts.

References

[1] Lengyel J. The convergence of graphics and vision. *Computer*, 1998, 31(7):46–53.

[2] Kang S B. Vision for graphics. In *Joint Intl. Symp. on Computer Vision*, 2007, pp. 1–12.

[3] Shotton J, Sharp T, Kipman A, Fitzgibbon A, Finocchio M, Blake A, Cook M, Moore R. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 2013, 56(1):116–124.

[4] Xiao J. Graphics for vision: Learning to see using big 3d synthetic data, 2015. CAD Keynote Talk.

[5] Zheng S, Prisacariu V A, Averkiou M, Cheng M M, Mitra N J, Shotton J, Torr P H, Rother C. Object proposals estimation in depth image using compact 3d shape manifolds. In *GCPR*, 2015, pp. 196–208.

[6] Meeker M. Internet trends 2014-code conference. *Glokalde*, 2014, 1(3).

[7] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521(7553):436–444.

[8] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 1998, (11):1254–1259.

[9] Cheng M M, Mitra N J, Huang X, Torr P H S, Hu S M. Global contrast based salient region detection. *IEEE TPAMI*, 2015, 37(3):569–582.

[10] Qi W, Cheng M M, Borji A, Lu H, Bai L F. SaliencyRank: Two-stage manifold ranking for salient object detection. *Computational Visual Media*, 2015, 1(4):309–320.

[11] Wu X, Du M, Chen W, Wang J. Salient object detection via region contrast and graph regularization. *Science China Information Sciences*, 2016, pp. 1–14.

[12] Zhang W, Borji A, Wang Z, Le Callet P, Liu H. The application of visual saliency models in objective image quality assessment: A statistical evaluation. *IEEE TNNLS*, 2016, 27(6):1266–1278.

[13] Borji A, Cheng M M, Jiang H, Li J. Salient object detection: A benchmark. *IIEEE TIP*, 2015, 24(12):5706–5722.

[14] Borji A, Cheng M M, Jiang H, Li J. Salient object detection: A survey. *arXiv preprint arXiv:1411.5878*, 2014.

[15] Han J, Liu N, Zhang D. Visual saliency detection and applications: A survey. *Frontiers of Computer Science*, 2017.

[16] Zhang G X, Cheng M M, Hu S M, Martin R R. A shape-preserving approach to image resizing. *Computer Graphics Forum*, 2009, 28(7):1897–1906.

[17] Zhao Y, Liu Y. Patch based saliency detection method for 3d surface simplification. In *IEEE ICPR*, 2012, pp. 845–848.

[18] Jänicke H, Chen M. A salience-based quality metric for visualization. In *Computer Graphics Forum*, volume 29, 2010, pp. 1183–1192.

[19] Miao Y W, Feng J Q, Wang J R, Pajarola R. A multi-channel salience based detail exaggeration technique for 3d relief surfaces. *JCST*, 2012, 27(6):1100–1109.

[20] Lee C H, Varshney A, Jacobs D W. Mesh saliency. In *ACM TOG*, volume 24, 2005, pp. 659–666.

[21] Kim Y, Varshney A. Saliency-guided enhancement for volume visualization. *IEEE TVCG*, 2006, 12(5):925–932.

[22] Avidan S, Shamir A. Seam carving for content-aware image resizing. In *ACM TOG*, volume 26, 2007, p. 10.

[23] Wang Y S, Tai C L, Sorkine O, Lee T Y. Optimized scale-and-stretch for image resizing. *ACM TOG*, 2008, 27(5):118.

[24] Zhang L, Wang M, Nie L, Hong L, Rui Y, Tian Q. Retargeting semantically-rich photos. *IEEE Trans. on Multimedia*, 2015, 17(9):1538–1549.

[25] Wu H, Wang Y S, Feng K C, Wong T T, Lee T Y, Heng P A. Resizing by symmetry-summarization. *ACM TOG*, 2010, 29(6):159.

[26] Zhang F, Zhang X, Qin X Y, Zhang C M. Enlarging image by constrained least square approach with shape preserving. *JCST*, 2015, 30(3):489–498.

[27] Li B, Duan L Y, Lin C W, Huang T, Gao W. Depth-preserving warping for stereo image retargeting. *IEEE TIP*, 2015, 24(9):2811–2826.

[28] Jain E, Sheikh Y, Shamir A, Hodgins J. Gaze-driven video re-editing. *ACM TOG*, 2015, 34(2):21.

[29] Liu Y, Sun L, Yang S. A retargeting method for stereoscopic 3d video. *Computational Visual Media*, 2015, 1(2):119–127.

[30] Miao Y, Lin H. Visual saliency guided global and local resizing for 3d models. In *Intl. Conference on CAD/Graphics*, 2013, pp. 212–219.

[31] Jia S, Zhang C, Li X, Zhou Y. Mesh resizing based on hierarchical saliency detection. *Graphical Models*, 2014, 76(5):355–362.

[32] Song R, Liu Y, Zhao Y, Martin R R, Rosin P L. Conditional random field-based mesh saliency. In *IEEE ICIP*, 2012, pp. 637–640.

[33] Castelló P, Chover M, Sbert M, Feixas M. Reducing complexity in polygonal meshes with view-based saliency. *Computer Aided Geometric Design*, 2014, 31(6):279–293.

[34] Miao Y, Feng J, Pajarola R. Visual saliency guided normal enhancement technique for 3d shape depiction. *Computers & Graphics*, 2011, 35(3):706–712.

[35] Zhao Y, Lu S, Qian H, Yao P. Robust mesh deformation with salient features preservation. *Science China Information Sciences*, 2016, pp. 1–9.

[36] Semmo A, Trapp M, Kyprianidis J E, Döllner J. Interactive visualization of generalized virtual 3d city models using level-of-abstraction transitions. In *Computer Graphics Forum*, volume 31, 2012, pp. 885–894.

[37] Song P, Fu Z, Liu L, Fu C W. Printing 3d objects with interlocking parts. *Computer Aided Geometric Design*, 2015, 35:137–148.

[38] Wang W, Chao H, Tong J, Yang Z, Tong X, Li H, Liu X, Liu L. Saliency-preserving slicing optimization for effective 3d printing. In *Computer Graphics Forum*, volume 34, 2015, pp. 148–160.

[39] Criminisi A, Pérez P, Toyama K. Region filling and object removal by exemplar-based image inpainting. *IEEE TIP*, 2004, 13(9):1200–1212.

[40] Adams A, Gelfand N, Dolson J, Levoy M. Gaussian kd-trees for fast high-dimensional filtering. In *ACM TOG*, volume 28, 2009, p. 21.

[41] Simakov D, Caspi Y, Shechtman E, Irani M. Summarizing visual data using bidirectional similarity. In *IEEE CVPR*, 2008, pp. 1–8.

[42] Shamir A, Avidan S. Seam carving for media retargeting. *Communications of the ACM*, 2009, 52(1):77–85.

[43] Chen T, Zhu Z, Shamir A, Hu S M, Cohen-Or D. 3-sweep: Extracting editable objects from a single photo. *ACM TOG*, 2013, 32(6):195.

[44] Kholgade N, Simon T, Efros A, Sheikh Y. 3d object manipulation in a single photograph using stock 3d models. *ACM TOG*, 2014, 33(4):127.

[45] Koffka K. *Principles of Gestalt psychology*, volume 44. Routledge, 2013.

[46] Cheng M M, Zhang F L, Mitra N J, Huang X, Hu S M. RepFinder: finding approximately repeated scene elements for image editing, 2010.

[47] Goldberg C, Chen T, Zhang F L, Shamir A, Hu S M. Data-driven object manipulation in images. *Computer Graphics Forum*, 2012, 31:265–274.

[48] Chen T, Cheng M M, Tan P, Shamir A, Hu S M. Sketch2Photo: internet image montage. *ACM TOG*, 2009, 28(5):124.

[49] Lu S P, Zhang S H, Wei J, Hu S M, Martin R R. Timeline editing of objects in video. *IEEE TVCG*, 2013, 19(7):1218–1227.

[50] Zheng Y, Chen X, Cheng M M, Zhou K, Hu S M, Mitra N J. Interactive images: cuboid proxies for smart image manipulation. *ACM TOG*, 2012, 31(4):99–1.

[51] Iizuka S, Endo Y, Hirose M, Kanamori Y, Mitani J, Fukui Y. Object repositioning based on the perspective in a single image. In *Computer Graphics Forum*, volume 33, 2014, pp. 157–166.

[52] Rong Y, Zheng Y, Shao T, Yang Y, Zhou K. An interactive approach for functional prototype recovery from a single rgbd image. *Computational Visual Media*, 2016, 2(1):87–96.

[53] Wu J, Rosin P L, Sun X, Martin R R. Improving shape from shading with interactive tabu search. *JCST*, 2016, 31(3):450–462.

[54] Zhao H L, Nie G Z, Li X J, Jin X G, Pan Z G. Structure-aware nonlocal optimization framework for image colorization. *JCST*, 2015, 30(3):478–488.

[55] Cheng M M, Prisacariu V A, Zheng S, Torr P H, Rother C. Densecut: Densely connected crfs for realtime grabcut. In *Computer Graphics Forum*, volume 34, 2015, pp. 193–201.

[56] Cheng M M, Zheng S, Lin W Y, Vineet V, Sturgess P, Crook N, Mitra N J, Torr P. ImageSpirit: verbal guided image parsing. *ACM TOG*, December 2014, 34(1):3:1–3:11.

[57] Huang Q, Wang H, Koltun V. Single-view reconstruction via joint analysis of image and shape collections. *ACM TOG*, 2015, 34(4):87.

[58] Chen T, Tan P, Ma L Q, Cheng M M, Shamir A, Hu S M. Poseshop: human image database construction and personalized content synthesis. *IEEE TVCG*, 2013, 19(5):824–837.

[59] Tanahashi Y, Hsueh C H, Ma K L. An efficient framework for generating storyline visualizations from streaming data. *IEEE TVCG*, 2015, 21(6):730–742.

[60] Hasegawa K, Saito H. Synthesis of a stroboscopic image from a hand-held camera sequence for a sports analysis. *Computational Visual Media*, 2016, 2(3):277–289.

[61] Lalonde J F, Hoiem D, Efros A A, Rother C, Winn J, Criminisi A. Photo clip art. *ACM TOG*, 2007, 26(3):3.

[62] Xu K, Chen K, Fu H, Sun W L, Hu S M. Sketch2Scene: sketch-based co-retrieval and co-placement of 3d models. *ACM TOG*, 2013, 32(4):123.

[63] Chia A Y S, Zhuo S, Gupta R K, Tai Y W, Cho S Y, Tan P, Lin S. Semantic colorization with internet images. In *ACM TOG*, volume 30, 2011, p. 156.

[64] Agarwal S, Snavely N, Simon I, Seitz S M, Szeliski R. Building rome in a day. In *IEEE ICCV*, 2009, pp. 72–79.

[65] Cao C, Bradley D, Zhou K, Beeler T. Realtime high-fidelity facial performance capture. *ACM TOG*, 2015, 34(4):46.

[66] Longuet-Higgins H C. A computer algorithm for reconstructing a scene from two projections. *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms, MA Fischler and O. Firschein, eds*, 1987, pp. 61–62.

[67] Ding C, Liu L. A survey of sketch based modeling systems. *Frontiers of Computer Science*, 2016, 10(6):985–999.

[68] Snavely N, Seitz S M, Szeliski R. Photo tourism: exploring photo collections in 3d. In *ACM TOG*, volume 25, 2006, pp. 835–846.

[69] Frahm J M, Fite-Georgel P, Gallup D, Johnson T, Raguram R, Wu C, Jen Y H, Dunn E, Clipp B, Lazebnik S et al. Building rome on a cloudless day. In *ECCV*, 2010, pp. 368–381.

[70] Fuhrmann S, Langguth F, Moehrle N, Waechter M, Goesele M. MVE – an image-based reconstruction environment. *Computers & Graphics*, 2015, 53:44–53.

[71] Ceylan D, Mitra N J, Zheng Y, Pauly M. Coupled structure-from-motion and 3d symmetry detection for urban facades. *ACM TOG*, 2014, 33(1):2.

[72] Kopf J, Cohen M F, Szeliski R. First-person hyper-lapse videos. *ACM TOG*, 2014, 33(4):78.

[73] Tan W, Liu H, Dong Z, Zhang G, Bao H. Robust monocular SLAM in dynamic environments. In *IEEE ISMAR*, 2013, pp. 209–218.

[74] Li K, Yang J, Jiang J. Nonrigid structure from motion via sparse representation. In *ICME*, 2014.

[75] Li K, Yang J, Jiang J. Nonrigid structure from motion via sparse representation. *IEEE Trans. on cybernetics*, 2015, 45(8):1401–1413.

[76] Huang H, Chai J, Tong X, Wu H T. Leveraging motion capture and 3d scanning for high-fidelity facial performance acquisition. In *ACM TOG*, volume 30, 2011, p. 74.

[77] Zhang L, Snavely N, Curless B, Seitz S M. Spacetime faces: High-resolution capture for˜ modeling and animation. In *Data-Driven 3D Facial Animation*, pp. 248–276. Springer, 2008.

[78] Beeler T, Hahn F, Bradley D, Bickel B, Beardsley P, Gotsman C, Sumner R W, Gross M. High-quality passive facial performance capture using anchor frames. *ACM TOG*, 2011, 30(4):75.

[79] Chen K, Lai Y K, Hu S M. 3d indoor scene modeling from rgb-d data: a survey. *Computational Visual Media*, 2015, 1(4):267–278.

[80] Cao C, Hou Q, Zhou K. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM TOG*, 2014, 33(4):43.

[81] Casas D, Tejera M, Guillemaut J Y, Hilton A. Interactive animation of 4d performance capture. *IEEE TVCG*, 2013, 19(5):762–773.

[82] Huang P, Tejera M, Collomosse J, Hilton A. Hybrid skeletal-surface motion graphs for character animation from 4d performance capture. *ACM TOG*, 2015, 34(2):17.

[83] Xia S, Wang C, Chai J, Hodgins J. Realtime style transfer for unlabeled heterogeneous human motion. *ACM TOG*, 2015, 34(4):119.

[84] Pons-Moll G, Romero J, Mahmood N, Black M J. Dyna: A model of dynamic human shape in motion. *ACM TOG*, 2015, 34(4):120.

[85] Rogez G, Schmid C. Mocap-guided data augmentation for 3d pose estimation in the wild. *arXiv preprint arXiv:1607.02046*, 2016.

[86] Shotton J, Girshick R, Fitzgibbon A, Sharp T, Cook M, Finocchio M, Moore R, Kohli P, Criminisi A, Kipman A et al. Efficient human pose estimation from single depth images. *IEEE TPAMI*, 2013, 35(12):2821–2840.

[87] Song S, Xiao J. Sliding shapes for 3d object detection in depth images. In *ECCV*, pp. 634–651. Springer, 2014.

[88] Malisiewicz T, Gupta A, Efros A A. Ensemble of exemplar-SVMs for object detection and beyond. In *IEEE ICCV*, 2011, pp. 89–96.

[89] Peng X, Sun B, Ali K, Saenko K. Learning deep object detectors from 3d models. In *IEEE ICCV*, 2015, pp. 1278–1286.

[90] Gupta S, Girshick R, Arbeláez P, Malik J. Learning rich features from RGB-D images for object detection and segmentation. In *Computer Vision–ECCV 2014*, pp. 345–360. Springer, 2014.

[91] Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, Xiao J. 3D ShapeNets: a deep representation for volumetric shapes. In *IEEE CVPR*, 2015, pp. 1912–1920.

[92] Maturana D, Scherer S. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IEEE IROS*, 2015, pp. 922–928.

[93] Wohlhart P, Lepetit V. Learning descriptors for object recognition and 3d pose estimation. In *IEEE CVPR*, 2015, pp. 3109–3118.

[94] Jaderberg M, Simonyan K, Vedaldi A, Zisserman A. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014.

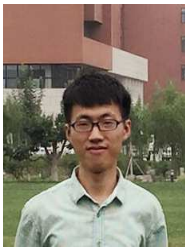
[95] Zhu Y, Yao C, Bai X. Scene text detection and recognition: Recent advances and future trends. *Frontiers of Computer Science*, 2016, 10(1):19–36.

[96] Valentin J, Vineet V, Cheng M M, Kim D, Shotton J, Kohli P, Nießner M, Criminisi A, Izadi S, Torr P. SemanticPaint: interactive 3d labeling and learning at your fingertips. *ACM TOG*, 2015, 34(5):154.

[97] Xu K, Huang H, Shi Y, Li H, Long P, Caichen J, Sun W, Chen B. Autoscanning for coupled scene reconstruction and proactive object analysis. *ACM TOG*, 2015, 34(6):177.

[98] Tateno K, Tombari F, Navab N. When 2.5D is not enough: Simultaneous reconstruction, segmentation and recognition on dense SLAM. In *IEEE ICRA*, 2016, pp. 2295–2302.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012. Then he was a research fellow for 2 years with Prof. Philip Torr in Oxford. He is now an associate professor at Nankai University. His research interests include computer graphics, computer vision, and image processing. He has received the Google PhD fellowship award, the IBM PhD fellowship award, and the new PhD Researcher Award from Chinese Ministry of Education.



Qibin Hou is currently a Ph.D. Candidate at the College of Computer Science and Control Engineering, Nankai University. His research interests include deep learning, image processing, and computer vision.

Song-Hai Zhang obtained his Ph.D. in 2007 from Tsinghua University. He is currently an associate professor of computer science at Tsinghua University, China. His research interests include image and video processing, geometric computing.

Paul L. Rosin is a professor at the School of Computer Science & Informatics, Cardiff University. His research interests include the representation, segmentation, and grouping of curves, knowledge-based vision systems, early image representations, low level image processing, machine vision approaches to remote sensing, methods for evaluation of approximation algorithms, medical and biological image analysis, mesh processing, non-photorealistic rendering and the analysis of shape in art and architecture.