Final Project List

Computer Graphics – Final Project

Prepared by: Anum Masood (TA)

10/1/2016

Contents

Project List for Computer Graphics – 2016

1. Efficient Deep Learning for Stereo Matching

Description:

Reconstructing scene is the key element for 3D imaging in many applications like robotics and self-driving cars. To improve this process, 3D sensors such as LIDAR are commonly employed. Utilizing cameras is an attractive alternative, as it is typically a more cost-effective solution. However, despite decades of research, estimating depth from a **stereo pair** is still an open problem and research is still being done in this field. Dealing with occlusions, large saturated areas and repetitive patterns are some of the remaining challenges. Very recently, convolutional networks have also been exploited to learn how to match for the task of stereo estimation. Current approaches learn the parameters of the matching network by treating the problem as binary classification. Given a patch in the left image, the task is to predict if a patch in the right image is the correct match. The goal is to propose a matching network which is able to produce very accurate results in less than a second of GPU computation.

Reference Paper:

http://www.cs.toronto.edu/~urtasun/publications/luo_etal_cvpr16.pdf

2. Theory and Practice of Structure-from-Motion using Affine Correspondences

Description:

Tools and techniques for obtaining information about the geometry of 3D scenes from 2D images have also been a challenging research field. This task is challenging because the image formation process is not generally invertible: from its projected position in a camera image plane, a scene point can only be recovered up to a one-parameter ambiguity corresponding to its distance from the camera. Hence, additional information is needed to solve the reconstruction problem. Structure from motion techniques are used in a wide range of applications including photogrammetric survey, the automatic reconstruction of virtual reality models from video sequences, and for the determination of camera motion. One possibility is to exploit prior knowledge about the scene to reduce the number of degrees of freedom. For example parallelism and co-planarity constraints can be used to reconstruct simple geometric shapes such as line segments and planar polygons from their projected positions in individual views. Another possibility is to use corresponding image points in multiple views. Given its image in two or more views, a 3D point can be reconstructed by triangulation. Affine Correspondences (ACs) are more informative than Point Correspondences (PCs) that are used as input in mainstream algorithms for Structure-from-Motion (SfM).

Reference Paper:

http://arthronav.isr.uc.pt/~carolina/files/CVPRsubm.pdf

3. Weakly supervised object boundaries

Description:

Learning based boundary detection methods require extensive training data. Since labeling object boundaries is one of the most expensive types of annotations, there is a need to relax the requirement to carefully annotate images to make both the training more affordable and to extend the amount of training data. In this paper we propose a technique to generate weakly supervised annotations and show that bounding box annotations alone suffice to reach high-quality object boundaries without using any object-specific boundary annotations. Weak supervision techniques are required to achieve the top performance on the object boundary detection task, outperforming by a large margin the current fully supervised state-of-the-art methods.

Reference Paper:

https://arxiv.org/pdf/1511.07803.pdf

4. Face Alignment Across Large Poses

Description:

Face alignment, which fits a face model to an image and extracts the semantic meanings of facial pixels, has been an important topic in CV community. However, most algorithms are designed for faces in small to medium poses (below 45 degree), lacking the ability to align faces in large poses up to 90 degree. The challenges are three-fold: Firstly, the commonly used landmark-based face model assumes that all the landmarks are visible and is therefore not suitable for profile views. Secondly, the face appearance varies more dramatically across large poses, ranging from frontal view to profile view. Thirdly, labeling landmarks in large poses is extremely challenging since the invisible landmarks have to be guessed. A solution to the three problems in a new alignment framework is required, in which a dense 3D face model is fitted to the image via convolutional neutral network (CNN). A 3D model has to be synthesized for large-scale training samples in profile views to solve the problems of data labeling.

Reference Paper:

https://arxiv.org/abs/1511.07212

5. Deep Interactive Object Selection

Description:

Interactive object selection (also known as interactive segmentation) has become a very popular research area over the past years. It enables users to select objects of interest accurately by interactively providing inputs such as strokes and bounding boxes. The selected results are useful for various applications such as localized editing and image/video composition. There are many algorithms proposed to solve this problem. Previous algorithms require substantial user interactions to estimate the foreground and background distributions. Deep-learning-based algorithm is required which has a much better understanding of objectiveness and thus can reduce user interactions to just a few clicks. In this project an algorithm is to be proposed which can transform user-provided positive and negative

clicks into two Euclidean distance maps which are then concatenated with the RGB channels of images to compose (image, user interactions) pairs.

Reference:

https://arxiv.org/pdf/1603.04042.pdf

6. Learning Relaxed Deep Supervision for Better Edge Detection

Description:

Relaxed deep supervision (RDS) along with convolutional neural networks can be used for improved edge detection. The conventional deep supervision utilizes the general ground-truth to guide intermediate predictions while the hierarchical supervisory signals with additional relaxed labels to consider the diversities in deep neural networks. First of all the relaxed labels have to be captured from simple detectors (e.g. Canny).Then they are to be merged with the general ground-truth to generate the RDS. Finally the RDS would be used to supervise the edge network following a coarse-to-fine paradigm. These relaxed labels can be seen as some false positives that are difficult to be classified. These false positives in the supervision are considered, and are to be removed in order to achieve high performance for better edge detection. The lack of training images can be compensated by capturing coarse edge annotations from a large dataset of image segmentations to pre-train the model.

Reference:

http://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Liu_Learning_Relaxed_Deep_CVPR_2016_paper.pdf

7. Two Illuminant Estimation and User Correction Preference

Description:

There is a solution required for the problem of white-balance correction when a scene contains two illuminations. This is a two step process:

1) Estimate the two illuminants
2) Correct the image

Existing methods attempt to estimate a spatially varying illumination map, however, results are error prone and the resulting illumination maps are too low resolution to be used for proper spatially varying white balance correction. In addition, the spatially varying nature of these methods make them computationally intensive. However this problem can be effectively addressed by not attempting to obtain a spatially varying illumination map, but instead by performing illumination estimation on large sub-regions of the image. The proposed technique should be able to detect when distinct illuminations are present in the image and accurately measure these illuminants. If the spatially varying image correction is accurately done then user study could be later performed to see whether there is a preference for how the image should be corrected when two illuminants are present, but only a global correction can be applied.

Reference:

http://www.cse.yorku.ca/~mbrown/pdf/cvpr2016_two_illuminant.pdf

8. Connecting Images and Natural Language

Description:

Intelligent agents require the ability to perceive their environments, understand their high-level semantics, and communicate with humans. While computer vision has recently made great strides in visual recognition, the predominant paradigm is to predict one or more fixed visual categories for each image. Recent advances have allowed us to significantly expand the vocabulary of computer vision systems by treating natural language as a label space. Recent progress in areas such as image-sentence ranking, (dense) image captioning and visual Q&A has enabled the researchers to propose novel techniques for connecting the images and the natural language.

Reference:

https://cs.stanford.edu/people/karpathy/main.pdf
http://machinelearning.wustl.edu/mlpapers/paper_files/ICML2011Socher_125.pdf

9. Multiview Image Completion with Space Structure Propagation

Description:

Image completion synthesizes image structure to fill in larger missing areas. Multiview image completion is a branch of image completion, focusing on image completion of multiple photographs. Single Image Completion prioritized patches to be filled, and greedily propagated the patches from the known source region to the target region. This problem is solved with an expectation-maximization (EM) optimization using an image pyramid. Since the nearest neighbor field (NNF) search is the computational bottleneck of image completion, therefore by introducing an approximated search algorithm the whole process was accelerated. Other proposed techniques include the gradients of images for patch-based synthesis and alpha-blending of multiple patches. Multiple image completion consists of three steps: preprocessing, structure propagation, and structure-guided completion. Multiview image completion method provides geometric consistency among different views by propagating space structures. Since a user specifies the region to be completed in one of multiview photographs casually taken in a scene, it is necessary to have a method to complete the set of photographs with geometric consistency by creating or removing structures on the specified region. One solution could be to incorporate photographs to estimate dense depth maps. Initially complete color as well as depth from a view, and then facilitate two stages of structure propagation and structure-guided completion. Structure propagation will optimize space topology in the scene across photographs, while structure-guide completion will enhance, and complete local image structure of both depth and color in multiple photographs with structural coherence by searching nearest neighbor fields in relevant views.

Reference:

http://vclab.kaist.ac.kr/cvpr2016p1/CVPR2016_Multiview_Inpainting.pdf

10. Situation Recognition: Visual Semantic Role Labeling for Image Understanding

Description:

Situation recognition generalizes activity recognition and human-object interaction, using the assignment of roles to define how actors, objects, substances, and locations participate in activities. Situation recognition, a problem that involves predicting activities along with actors, objects, substances, and locations and how these pieces fit together (semantic roles). It is the problem of producing a concise summary of the situation an image depicts including: (1) the main activity (e.g., clipping), (2) the participating actors, objects, substances, and locations (e.g., man, shears, sheep, wool, and field) and most importantly (3) the roles these participants play in the activity (e.g., the man is clipping, the shears are his tool, the wool is being clipped from the sheep, and the clipping is in a field). FrameNet can be used to define a large space of possible situations and collect a large-scale dataset containing over 500 activities, 1,700 roles, 11,000 objects, 125,000 images, and 200,000 unique situations. Structured prediction baselines could be used in activity-centric images, situation-driven prediction of objects and independent object & activity recognition.

Reference:

https://pdfs.semanticscholar.org/7904/1876e322ce1287faf88b6e75dd94c7375b3f.pdf

11. Shallow and Deep Convolutional Networks for Saliency Prediction

Description:

Prediction of salient areas in images has been traditionally addressed with hand-crafted features based on neuroscience principles. The problem has to be addressed with a completely data-driven approach by training a convolutional neural network (convnet). The learning process is formulated as a minimization of a loss function that measures the Euclidean distance of the predicted saliency map with the provided ground truth. The recent publication of large datasets of saliency prediction has provided enough data to train end-to-end architectures that are both fast and accurate. Firstly the end-to-end CNNs trained and tested for the purpose of saliency prediction. Objective is to compute saliency maps that represent the probability of visual attention on an image, defined as the eye gaze fixation points.

Reference:

https://arxiv.org/pdf/1603.00845.pdf

12. Video-Story Composition via Plot Analysis

Description:

People have the natural desire to capture and store personal experiences and memories. Today, we are able to record our activities more easily with decreasing cost of cameras and media storages. Moreover, with the success of smart phones and applications, photos and videos have become omnipresent in our daily lives. Consequently, people tend to capture photos and record videos without a limited storage burden and process them later. Unfortunately, manual post-processing of these contents is usually tedious, and thus a need for an automatic summarization of contents has arisen leading to various

researchers to work on this topic. The problem of composing a story out of multiple short video clips taken by a person during an activity or experience is challenging. Inspired by plot analysis of written stories, proposed method could generates a sequence of video clips ordered in such a way that it reflects plot dynamics and content coherency. That is, given a set of multiple video clips, our method composes a video which we call a video-story. Metrics are defined on scene dynamics and coherency by dense optical flow features and a patch matching algorithm. Using these metrics, an objective function is defined for the video-story. To efficiently search for the best video-story, Branch-and-Bound algorithm can be used which guarantees the global optimum.

Reference:

http://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Choi_Video-Story_Composition_via_CVPR_2016_paper.pdf

13. Efficient 3D Room Shape Recovery from a Single Panorama

Description:

A 360″ full-view indoor panorama is obtained easily using cameras nowadays but to recover the 3D room shape from this panorama is a challenge. Several methods are available to solve this problem, either by adopting the Indoor World model, which consists of a single floor, a single ceiling, and vertical walls or by estimating a cuboid shape that fits the room layout. Intrinsically, most of these approaches work in a discretized manner, that is, the results are selected from a set of candidates based on certain scoring functions. The generation rules of the candidates limit the scope of these algorithms. An algorithm can automatically infer a 3D shape from a collection of partially oriented superpixel facets and line segments. The core part of the algorithm is a constraint graph, which includes lines and superpixels as vertices, and encodes their geometric relations as edges.T o perform 3D reconstruction based on the constraint graph by solving all the geometric constraints as constrained linear least-squares is to be done. The selected constraints used for reconstruction can be identified using an occlusion detection method with a Markov random field.

Reference:

http://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Yang_Efficient_3D_Room_CVPR_2016_paper.pdf

14. Field Model for Repairing 3D Shapes

Description:

Recent advances of 3D acquisition devices and 3D scene reconstruction research have enabled large-scale acquisition of 3D scene data and this has raised a demand on 3D data analysis. However, it often happens that the 3D data cannot be obtained at high quality, even by recent reconstruction methods. Specifically, the 3D surfaces are missing and/or broken and this phenomenon causes difficulties for many sequential tasks such as 3D object detection and recognition, shape analysis, and scene understanding. Repairing missing and broken surfaces thus plays a critical role and deserves in-depth study. In this paper, we focus on repairing incomplete 3D shapes. This problem can be also referred to

as shape completion. Probably the objects are not occluded, i.e. they can be fully observed in RGB/RGB-D images. However, this assumption does not mean that objects can be completely reconstructed. Existing shape completion approaches make use of geometric information represented at either low-level or highlevel. Low-level geometry describes local structures, e.g. local smoothness, and can be used to fill small holes on broken surfaces. A field model for repairing 3D shapes constructed from multi-view RGB data is to be proposed. Specifically, we represent a 3D shape in a Markov random field (MRF) in which the geometric information is encoded by random binary variables and the appearance information is retrieved from a set of RGB images captured at multiple viewpoints. The local priors in the MRF model capture the local structures of object shapes and are learnt from 3D shape templates using a convolutional deep belief network.

Reference:

http://sonhua.me/pdf/nguyen-repairing-cvpr16.pdf

15. Efficiently Creating 3D Training Data for Fine Hand Pose Estimation

Description:

While many recent hand pose estimation methods critically rely on a training set of labelled frames, the creation of such a dataset is a challenging task that has been overlooked so far. As a result, existing datasets are limited to a few sequences and individuals, with limited accuracy, and this prevents these methods from delivering their full potential. A semi-automated method is proposed for efficiently and accurately labeling each frame of a hand depth video with the corresponding 3D locations of the joints: The user is asked to provide only an estimate of the 2D reprojections of the visible joints in some reference frames, which are automatically selected to minimize the labeling work by efficiently optimizing a sub-modular loss function. Exploit spatial, temporal, and appearance constraints to retrieve the full 3D poses of the hand over the complete sequence.

Reference:

https://arxiv.org/pdf/1605.03389.pdf

16. Temporal Multimodal Learning in Audiovisual Speech Recognition

Description:

Robust Automatic Speech Recognition (ASR) has been the key to the natural human-computer interfaces in most cases, but it's challenged by the noisy environments. One example of such an environment is street, where the traffic noise makes it very hard for recognizing the speech. Considering that vision is free of audio noise and can provide complemented information to audio in the noisy condition. In view of the advantages of deep networks in producing useful representation, the generated features of different modality data (such as image, audio) can be jointly learned using Multimodal Restricted Boltzmann Machines (MRBM). Recently, audiovisual speech recognition based the MRBM has attracted much attention, and the MRBM shows its effectiveness in learning the joint representation across audiovisual modalities. However, the built networks have weakness in modeling the multimodal

sequence which is the natural property of speech signal. Temporal multimodal deep learning architecture, named as Recurrent Temporal Multimodal RBM (RTMRBM),that models multimodal sequences by transforming the sequence of connected MRBMs into a probabilistic series model. Compared with existing multimodal networks, it's simple and efficient in learning temporal joint representation.

Reference:

http://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Hu_Temporal_Multimodal_Learning_CVPR_2016_paper.pdf

17. Multi-Oriented Text Detection with Fully Convolutional Networks

Description:

Multi-Oriented Text Detection with Fully Convolutional Networks is an approach for text detection in natural images. Both local and global cues are taken into account for localizing text lines in a coarse-to-fine procedure. First, a Fully Convolutional Network (FCN) model is trained to predict the salient map of text regions in a holistic manner. Then, text line hypotheses are estimated by combining the salient map and character components. Finally, another FCN classifier is used to predict the centroid of each character, in order to remove the false hypotheses. The framework is general for handling text in multiple orientations, languages and fonts.

Reference:

https://arxiv.org/pdf/1604.04018.pdf

18. Structure from Motion with Objects

Description:

Factorization methods for Structure from Motion (SfM) deliver highly efficient solutions for the simultaneous calibration and 3D reconstruction using image point trajectories/matches. To reconstruct the position of rigid objects and to jointly recover affine camera calibration solely from a set of object detections in a video sequence is a challenge for researchers. In practice, this work can be considered as the extension of Tomasi and Kanade factorization method using objects. Instead of using points to form a rank constrained measurement matrix, form a matrix with similar rank properties using 2D object detection proposals. In detail, first fit an ellipse onto the image plane at each bounding box as given by the object detector. The collection of all the ellipses in the dual space is used to create a measurement matrix that gives a specific rank constraint. This matrix can be factorised and metrically upgraded in order to provide the affine camera matrices and the 3D position of the objects as an ellipsoid. Moreover, full 3D quadric can be recovered thus giving additional information about object occupancy and 3D pose. Finally, 2D points measurements can be seamlessly included in the framework to reduce the number of objects required.

Reference:

https://pavisdata.iit.it/data/cvpr2016a/2016_cvpr_sfmwo.pdf

19. Convolutional Pose Machines

Description:

Pose Machines provide a sequential prediction framework for learning rich implicit spatial models. In this work we show a systematic design for how convolutional networks can be incorporated into the pose machine framework for learning image features and image-dependent spatial models for the task of pose estimation. The contribution of this paper is to implicitly model long-range dependencies between variables in structured prediction tasks such as articulated pose estimation. We achieve this by designing a sequential architecture composed of convolutional networks that directly operate on belief maps from previous stages, producing increasingly refined estimates for part locations, without the need for explicit graphical model-style inference. The characteristic difficulty of vanishing gradients during training by providing a natural learning objective function that enforces intermediate supervision, thereby replenishing back-propagated gradients and conditioning the learning procedure.

Reference:

https://arxiv.org/pdf/1602.00134.pdf

20. Using Deep Learning for Pulmonary Nodule Detection & Diagnosis

Description:

Revolutionary image recognition technique deep learning can be used for detection of malignant pulmonary nodules. Deep learning technique is based on deep neural network. We report results of the initial findings and performance of deep neural nets using a combination of various choice parameters. Classification accuracy, sensitivity and specificity of the network performance is assessed for various combinations of convolutional layers.

Reference:

http://gkmc.utah.edu/winter2016/sites/default/files/webform/abstracts/WCBI2016_Paper.pdf