8. Applications

8.1 Computer Vision

Computer vision is an interdisciplinary scientific field that deals with how computers can gain high-level understanding from digital images or videos. From the perspective of engineering, it seeks to understand and automate tasks that the human visual system can do.

Computer vision tasks include methods for acquiring, processing, analysing and understanding digital images, and extraction of high-dimensional data from the real world in order to produce numerical or symbolic information, e.g. in the forms of decisions. Understanding in this context means the transformation of visual images (the input of the retina) into descriptions of the world that make sense to thought processes and can elicit appropriate action. This image understanding can be seen as the disentangling of symbolic information from image data using models constructed with the aid of geometry, physics, statistics, and learning theory.

The scientific discipline of computer vision is concerned with the theory behind artificial systems that extract information from images. The image data can take many forms, such as video sequences, views from multiple cameras, multi-dimensional data from a 3D scanner, or medical scanning device. The technological discipline of computer vision seeks to apply its theories and models to the construction of computer vision systems.

Sub-domains of computer vision include scene reconstruction, object detection, event detection, video tracking, object recognition, 3D pose estimation, learning, indexing, motion estimation, visual servoing, 3D scene modelling, and image restoration.

8.1.1 Recognition

A technology in the field of computer vision for finding and identifying objects in an image or video sequence. Humans recognise a multitude of objects in images with little effort, despite the fact that the image of the objects may vary somewhat in different view points, in many different sizes and scales or even when they are translated or rotated. Objects can even be recognised when they are partially obstructed from view. This task is still a challenge for computer vision systems. Many approaches to the task have been implemented over multiple decades.

one or several pre-specified or learned objects or object classes can be recognised, usually together with their 2D positions in the image or 3D poses in the scene. Blippar, Google Goggles and LikeThat provide stand-alone programs that illustrate this functionality.

8.1.2 Detection

Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos. Well-researched domains of object detection include face detection and pedestrian detection. Object detection has applications in many areas of computer vision, including image retrieval and video surveillance.

Every object class has its own special features that helps in classifying the class –for example all circles are round. Object class detection uses these special features. For example, when looking for circles, objects that are at a particular distance from a point (i.e. the centre) are sought. Similarly, when looking for squares, objects that are perpendicular at corners and have equal side lengths are needed. A similar approach is used for face identification where eyes, nose, and lips can be found and features like skin colour and distance between eyes can be found.

8.1.3 Pose Estimation

The specific task of determining the pose of an object in an image (or stereo images, image sequence) is referred to as pose estimation. The pose estimation problem can be solved in different ways depending on the image sensor configuration, and choice of methodology. Three classes of methodologies can be distinguished:

- Analytic or geometric methods: Given that the image sensor (camera) is calibrated and the mapping from 3D points in the scene and 2D points in the image is known. If also the geometry of the object is known, it means that the projected image of the object on the camera image is a well-known function of the object's pose. Once a set of control points on the object, typically corners or other feature points, has been identified, it is then possible to solve the pose transformation from a set of equations which relate the 3D coordinates of

the points with their 2D image coordinates. Algorithms that determine the pose of a point cloud with respect to another point cloud are known as point set registration algorithms, if the correspondences between points are not already known.

- Genetic algorithm methods: If the pose of an object does not have to be computed in real-time a genetic algorithm may be used. This approach is robust especially when the images are not perfectly calibrated. In this particular case, the pose represent the genetic representation and the error between the projection of the object control points with the image is the fitness function.
- Learning-based methods: These methods use artificial learning-based system which learn the mapping from 2D image features to pose transformation. In short, this means that a sufficiently large set of images of the object, in different poses, must be presented to the system during a learning phase. Once the learning phase is completed, the system should be able to present an estimate of the object's pose given an image of the object.

8.1.4 Tracking

Video tracking is the process of locating a moving object (or multiple objects) over time using a camera. It has a variety of uses, some of which are: human-computer interaction, security and surveillance, video communication and compression, augmented reality, traffic control, medical imaging and video editing. Video tracking can be a time-consuming process due to the amount of data that is contained in video. Adding further to the complexity is the possible need to use object recognition techniques for tracking, a challenging problem in its own right.

The objective of video tracking is to associate target objects in consecutive video frames. The association can be especially difficult when the objects are moving fast relative to the frame rate. Another situation that increases the complexity of the problem is when the tracked object changes orientation over time. For these situations video tracking systems usually employ a motion model which describes how the image of the target might change for different possible motions of the object.

8.1.5 Fusion

The image fusion process is defined as gathering all the important information from multiple images, and their inclusion into fewer images, usually a single one. This single image is more informative and accurate than any single source image, and it consists of all the necessary information. The purpose of image fusion is not only to reduce the amount of data but also to construct images that are more appropriate and understandable for the human and machine perception. In computer vision, multisensor image fusion is the process of combining relevant information from

two or more images into a single image. The resulting image will be more informative than any of the input images.

In remote sensing applications, the increasing availability of space borne sensors gives a motivation for different image fusion algorithms. Several situations in image processing require high spatial and high spectral resolution in a single image. Most of the available equipment is not capable of providing such data convincingly. Image fusion techniques allow the integration of different information sources. The fused image can have complementary spatial and spectral resolution characteristics. However, the standard image fusion techniques can distort the spectral information of the multispectral data while merging.

In satellite imaging, two types of images are available. The panchromatic image acquired by satellites is transmitted with the maximum resolution available and the multispectral data are transmitted with coarser resolution. This will usually be two or four times lower. At the receiver station, the panchromatic image is merged with the multispectral data to convey more information.

Many methods exist to perform image fusion. The very basic one is the high-pass filtering technique. Later techniques are based on Discrete Wavelet Transform, uniform rational filter bank, and Laplacian pyramid.

8.2 Natural Language Processing

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyse large amounts of natural language data. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorise and organise the documents themselves.

Challenges in natural language processing frequently involve speech recognition, natural-language understanding, and natural-language generation.

In the early days, many language-processing systems were designed by symbolic methods, i.e., the hand-coding of a set of rules, coupled with a dictionary lookup: such as by writing grammars or devising heuristic rules for stemming.

More recent systems based on machine-learning algorithms have many advantages over hand-produced rules:

- The learning procedures used during machine learning automatically focus on the most common cases, whereas when writing rules by hand it is often not at all obvious where the effort should be directed.
- Automatic learning procedures can make use of statistical inference algorithms to produce

models that are robust to unfamiliar input (e.g. containing words or structures that have not been seen before) and to erroneous input (e.g. with misspelled words or words accidentally omitted). Generally, handling such input gracefully with handwritten rules, or, more generally, creating systems of handwritten rules that make soft decisions, is extremely difficult, error-prone and time-consuming.

- Systems based on automatically learning the rules can be made more accurate simply by supplying more input data. However, systems based on handwritten rules can only be made more accurate by increasing the complexity of the rules, which is a much more difficult task. In particular, there is a limit to the complexity of systems based on handwritten rules, beyond which the systems become more and more unmanageable. However, creating more data to input to machine-learning systems simply requires a corresponding increase in the number of man-hours worked, generally without significant increases in the complexity of the annotation process.

Despite the popularity of machine learning in NLP research, symbolic methods are still (2020) commonly used:

- when the amount of training data is insufficient to successfully apply machine learning methods, e.g., for the machine translation of low-resource languages such as provided by the Apertium system,
- for preprocessing in NLP pipelines, e.g., tokenization, or
- for postprocessing and transforming the output of NLP pipelines, e.g., for knowledge extraction from syntactic parses.

阅读后简要理解计算机视觉和自然语言处理的基本概念和具体任务。

NLP 部分推荐外部资源：https://www.bilibili.com/video/BV1b34y1B7zR?spm_id_from=333.337.search-card.all.click

CV 部分推荐外部资源：https://www.bilibili.com/video/BV1nJ411z7fe?spm_id_from=333.337.search-card.all.click