计算机视觉基础

中国石油大学（华东）
青岛软件学院、计算机科学与技术学院
宫文娟

计算机视觉基础

- 第一章 概论
- 第二章 基础知识
- 第三章 图像分类
- 第四章 图像语义分割
- 第五章 目标检测
- 第六章 识别
- 第七章 目标跟踪
- 第八章 多目视觉
- 第九章 视觉问答

第八章 多目视觉

- 8.1 图像配准
- 8.2 双目图像融合
- 8.3 多目重构

概 述

立体视觉(Stereo Vision)又称为三维视觉，它通过两个或多个相机采集被测目标的图像，并将这些图像调整至同一平面，然后基于其中同一被测特征点所对应像素间的差异来重建三维信息或实现三维测量。

- 研究立体视觉技术，可以将机器视觉系统的应用范围从二维平面扩展到三维环境。
- 立体视觉被广泛用于三维导航、三维定位、目标追踪和机器人研究等工业领域。
  - 例如，双目机器人可以使用三维信息来测量障碍物的尺寸和距离，以进行准确的路径规划。
  - 在目标分拣应用中，使用立体视觉系统可以避免部件被遮挡和照明变化对定位的影响，为机器手臂提供精确的目标位置信息。这就使得立体视觉特别适合那些需要机械手从包装箱或其他容器中分拣出某一特定3D对象的应用。
- 立体视觉系统对于亮度变化和阴影可保持不变性，因此还可用于目标追踪、自动驾驶系统障碍物检测等场合。

概述

多目视觉算法:

8.1. 图像配准

8.2. 双目图像融合

8.3. 多目重建

概述

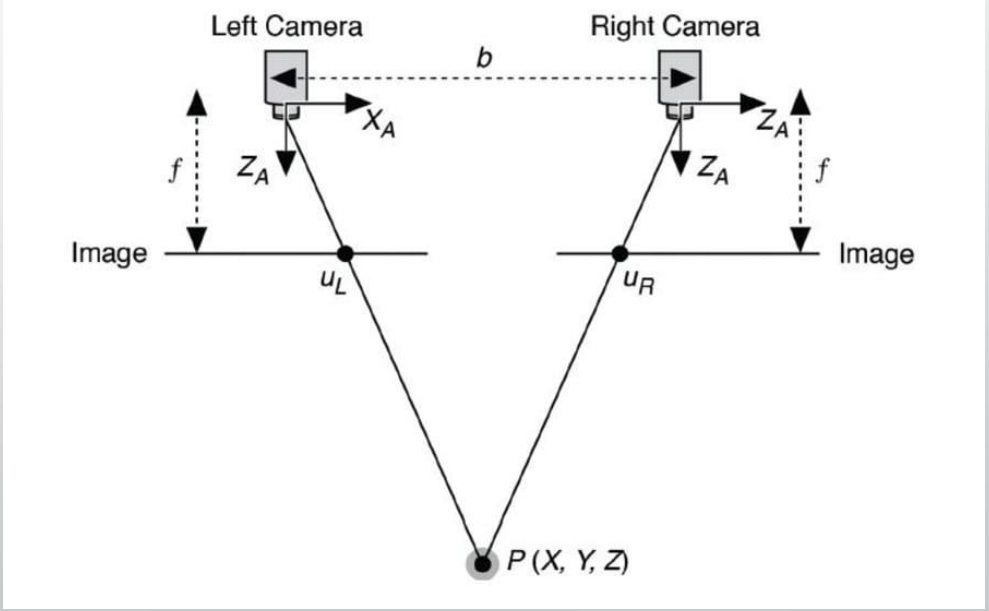
立体视觉(Stereo Vision)

- 基于双相机构建的双目视觉(Binocular Stereo Vision)系统是最小的立体视觉系统。它能完成大多数三维场景下的任务，而多目视觉系统可看作是多个双目系统的组合。

➢ 空间中某个点的深度depth可以计算：

$$depth = f * b/disparity$$

其中，$disparity = u_L - u_R = f * b/z$



典型的立体视觉系统

8.1图像配准

图像配准就是将同一个场景的不同图像转换到同样的坐标系统中的过程。

- 这些图像可以是不同时间拍摄的（多时间配准），可以是不同传感器拍摄的（多模配准），可以是不同视角拍摄的。图像之间的空间关系可能是刚体的（平移和旋转）、仿射的（例如错切），也有可能是单应性的，或者是复杂的大型形变模型。
- 其主要目的是检测输入图像与参考图像之间隐藏的关系，这种关系通常用坐标变换矩阵表示。因此，图像配准本质上可以设计为一个优化问题。

➢ 图像配准在许多实际应用中起着至关重要的作用：

- 在遥感方面如多光谱分类、环境监测、变化检测、图像拼接、天气预报、生成超分辨率图像以及将信息集成到地理信息系统(GIS)中有着广泛的应用；
- 在医学方面，包括计算机X线体层照相术(CT)和核磁共振数据以获得更完整的病人信息，综合分析不同的疾病，实现不同疾病的多模态分析，如监测肿瘤进化，治疗验证，并置病人的数据与解剖图集；
- 在地图制图学中，用于地图更新；在计算机视觉中用于目标定位、自动质量控制和运动跟踪。

8.1图像配准

根据图像采集的方式，可以将图像配准的应用分为以下几类：

1. 多视图分析：从多个视点捕获相似对象或场景的图像，以更好地表示被扫描对象或场景。例子包括图像拼接和从立体视觉中恢复形状。
2. 多时间分析：同一物体/场景的图像在不同的时间被捕获，通常是在不同的条件下，以跟踪在获取的连续图像之间出现的物体/场景的变化。例如运动跟踪，跟踪肿瘤的生长。
3. 多模态分析：利用不同的传感器获取同一物体/场景的图像，将不同来源的信息进行合并，得到物体/场景的细节信息。例子包括集成来自具有不同特征的传感器的信息，提供与光照无关的更好的空间和光谱分辨率；组合传感器捕获的解剖信息，如磁共振影像(MRI)、超声波或CT传感器获取的功能信息，如正电子发射断层扫描(PET)，单光子发射计算机断层扫描(SPECT)或磁共振波谱(MRS)研究和分析癫痫、阿尔茨海默病、抑郁症等其他疾病。

- 图8-1显示了一个MEG-MRI联合配准，这是一个多模态配准的例子。

8.1图像配准

实例

- 图8-1显示了一个MEG-MRI联合配准，这是一个多模态配准的例子。
  - 顶部黄色的点表示大脑图像轴向视图中的解剖标志或基准点(解剖信息)。
  - 底部粉红色的点代表脑磁图传感器的位置。
  - 绿色的点代表头皮脑电图传感器的位置。
  - 这些脑磁图(MEG)和脑电图(EEG)数据包含了功能信息，底部的图像显示了共配的大脑图像

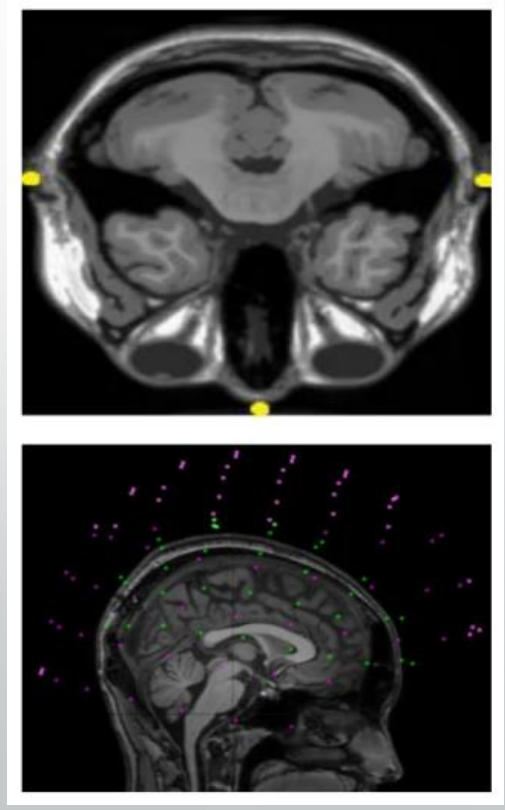
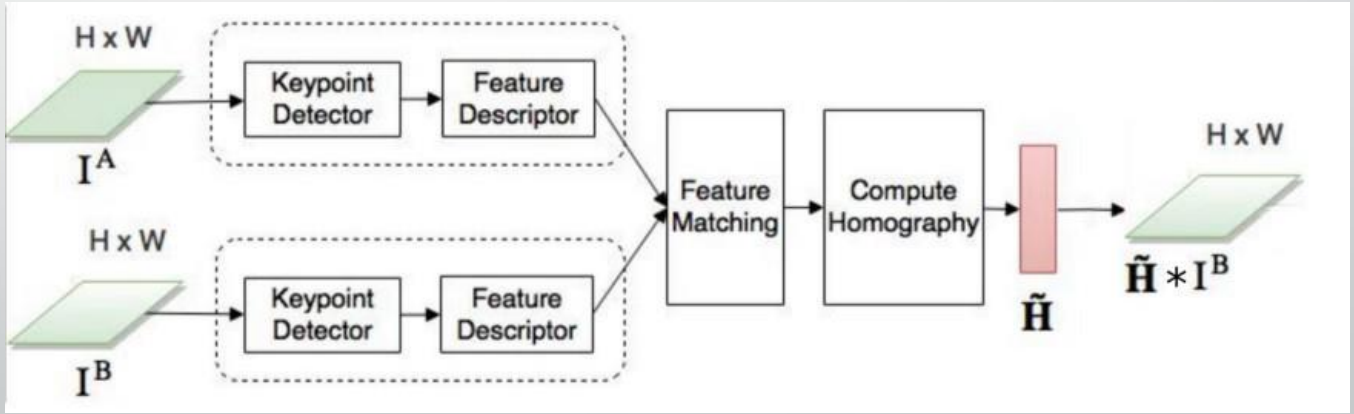


图8-1 多模态MRI-MEG配准

8.1 图像配准

8.1.1 传统的基于特征的方法

- 早期的图像配准主要使用基于特征的方法。这些方法有三个步骤：
  1. 关键点检测和特征描述
  2. 特征匹配
  3. 图像变换
- 简单的说，我们选择两个图像中的感兴趣点，将参考图像（reference image）与待配准图像中的等价感兴趣点进行关联，然后变换待配准图像使两个图像对齐。



利用基于特征的方法对齐可同调变换对齐的两幅图像

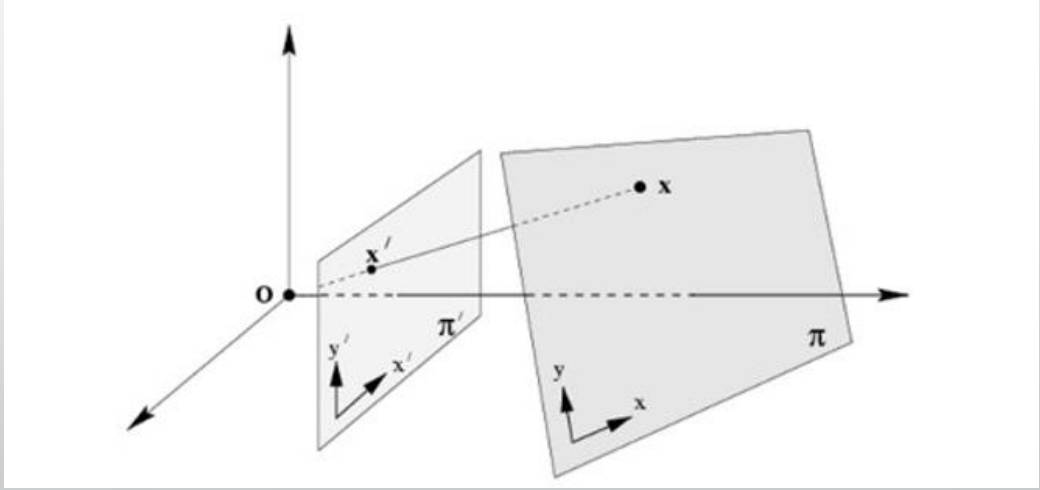
8.1 图像配准

8.1.1 传统的基于特征的方法

- 齐次坐标：给定欧式空间中的一个点(x,y)，对于任意的非零实数Z，三元组(xZ,yZ,Z)被称为该点的齐次坐标。
- 单应变换：

$$s\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = H\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}\begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

- 单应变换的几何意义：

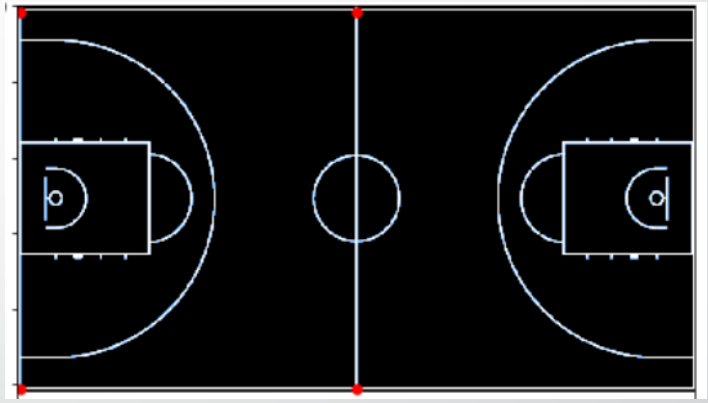


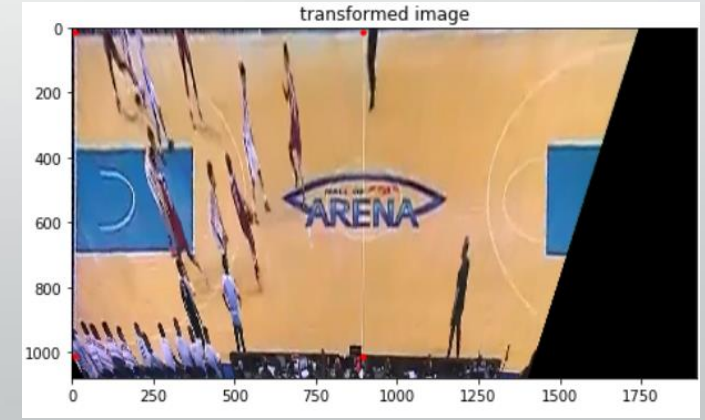
一个场景由不同视角投影得到两幅图像π和π'。场景中的同一个元素在这两幅图像中分别对应x和x'，这两点之间的变换就是单应变换。

8.1图像配准

8.1.1 传统的基于特征的方法

- 实例：





8.1 图像配准

8.1.1 传统的基于特征的方法

1. 关键点检测和特征描述

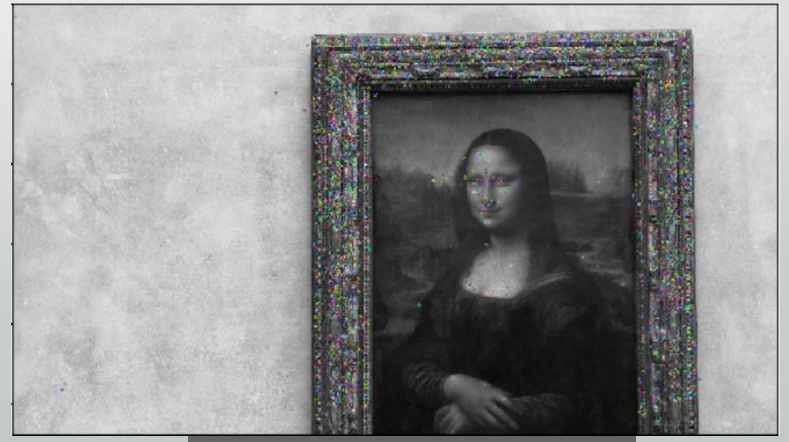
- 关键点就是感兴趣点，它表示图像中重要或独特的内容（边角，边缘等）。每个关键点由描述符表示关键点基本特征的特征向量。许多算法使用关键点检测和特征描述：

① SIFT（Scale-invariant feature transform）是用于关键点检测的原始算法，但它不能免费用于商业用途。SIFT特征描述符对于均匀缩放，方向，亮度变化和对仿射失真不变的部分不会发生变化。

② SURF（Speeded Up Robust Features）是一个受SIFT启发的探测器和描述符。它的优点是非常快。它同样是有专利的。

③ ORB（Oriented FAST and Rotated BRIEF）是一种快速的二进制描述符，它基于 FAST（Features from Accelerated Segment Test）关键点检测和 BRIEF（Binary robust independent elementary features）描述符的组合。它具有旋转不变性和对噪声的鲁棒性。它由OpenCV实验室开发，是SIFT有效的免费替代品。

④ AKAZE（Accelerated-KAZE）是KAZE快速版本。它为非线性尺度空间提供了快速的多尺度特征检测和描述方法，具有缩放和旋转不变性。



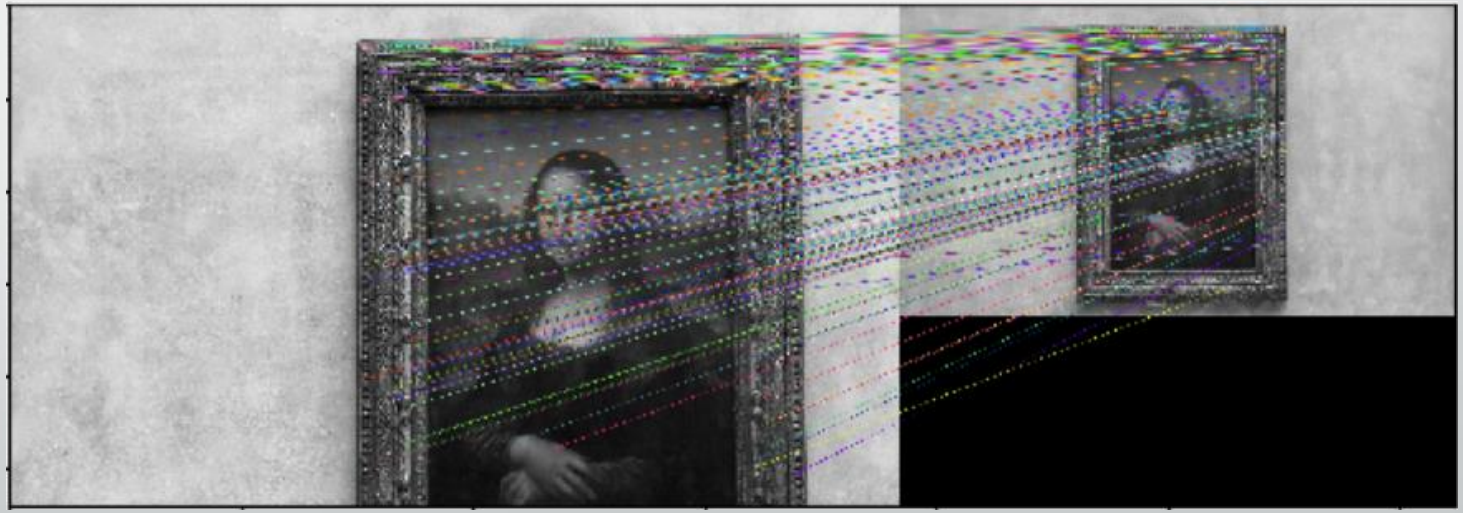
提取的图像特征点

8.1图像配准

8.1.1 传统的基于特征的方法

1. 关键点检测和特征描述
2. 特征匹配

- 一旦在一对图像中识别出关键点，我们就需要将两个图像中对应的关键点进行关联或“匹配”。其中一种方法是KNN匹配。这个方法计算每对关键点之间的描述符的距离，并返回每个关键点的k个最佳匹配中的最小距离。



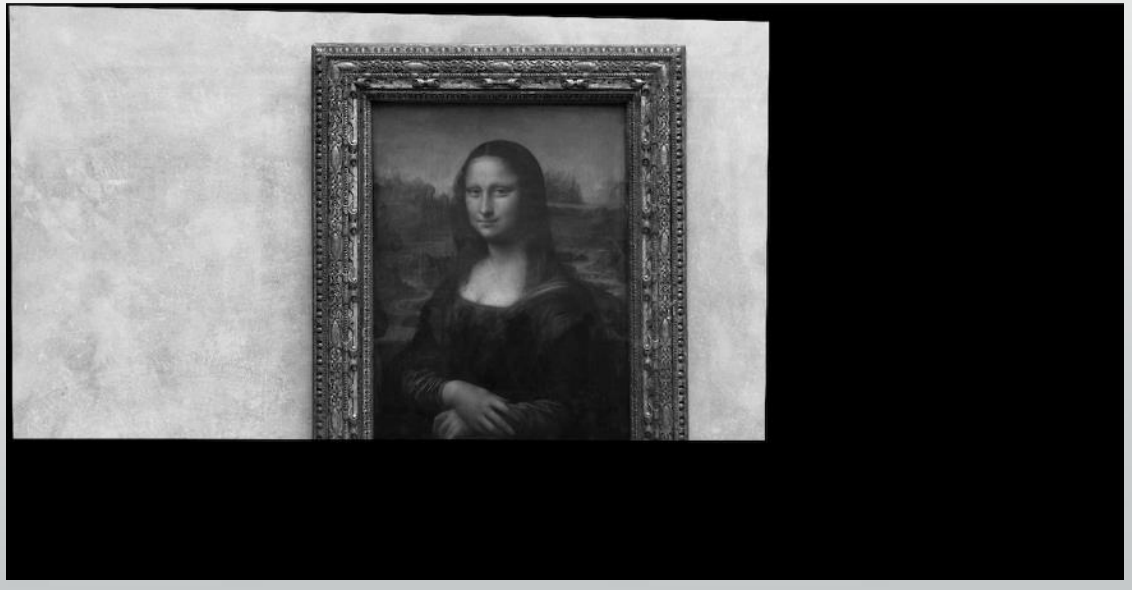
匹配的图像特征点

8.1图像配准

8.1.1 传统的基于特征的方法

1. 关键点检测和特征描述
2. 特征匹配
3. 图像变换

- 在匹配至少四对关键点之后，我们就可以将一个图像转换为另一个图像，称为图像变换（image warping）。空间中相同平面的两个图像通过单应性变换（Homographies）进行关联。Homographies是具有8个自由参数的几何变换，由3x3矩阵表示图像的整体变换（与局部变换相反）。因此，为了获得变换后的感测图像，需要计算Homographies矩阵。
- 为了得到最佳的变换，我们需要使用RANSAC算法检测异常值并去除。



变形后的待配准图像

8.1 图像配准

8.1.2 基于深度学习的图像配准

1. 特征提取

- 深度学习用于图像配准的第一种方式是用于特征提取。卷积神经网络设法获得越来越复杂的图像特征并进行学习。
- 2014年，Dosovitskiy等人提出了一种通用的特征提取方法，使用未标记的数据训练卷积神经网络。这些特征的通用性使转换具有鲁棒性。这些特征或描述符的性能优于SIFT描述符以匹配任务。
- 2014年以来，研究人员将这些网络应用于特征提取的步骤，而不是使用SIFT或类似算法。
- 2018年，Yang等人开发了一种基于相同思想的非刚性配准方法。他们使用预训练的VGG网络层来生成一个特征描述符，同时保留卷积信息和局部特征。这些描述符的性能也优于类似SIFT的探测器，特别是在SIFT包含许多异常值或无法匹配足够数量特征点的情况下。

8.1 图像配准

8.1.2 基于深度学习的图像配准

1. 特征提取
2. 单应变换学习

- 利用神经网络直接学习几何变换对齐两幅图像，而不仅仅局限于特征提取。

8.1图像配准

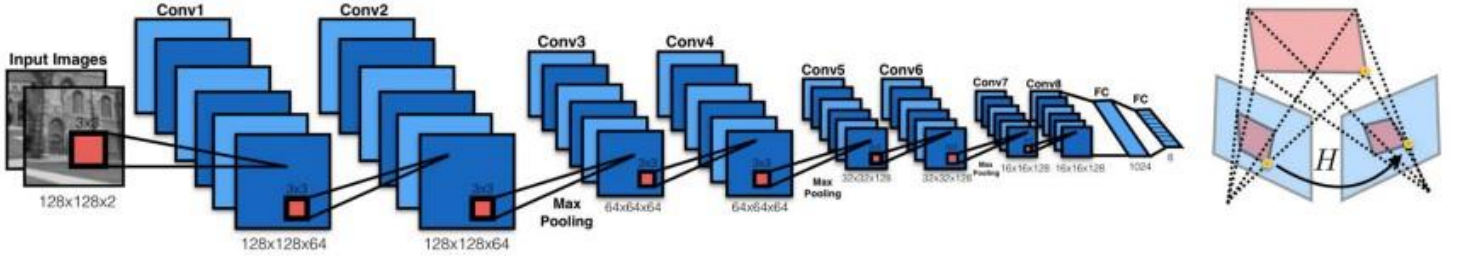
8.1.2 基于深度学习的图像配准

1. 特征提取

2. 单应变换学习

2.1 监督学习

- 在2016年，DeTone等人提出了深度图像单应变换估计方法[1]，提出了HomographyNe回归网络，这是一种VGG风格模型，可以学习两幅相关图像的单应性变换。该算法具有以端到端的方式同时学习单应性变换和CNN模型参数的优势。



[1]. Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, Deep Image Homography Estimation, arXiv, 2016.

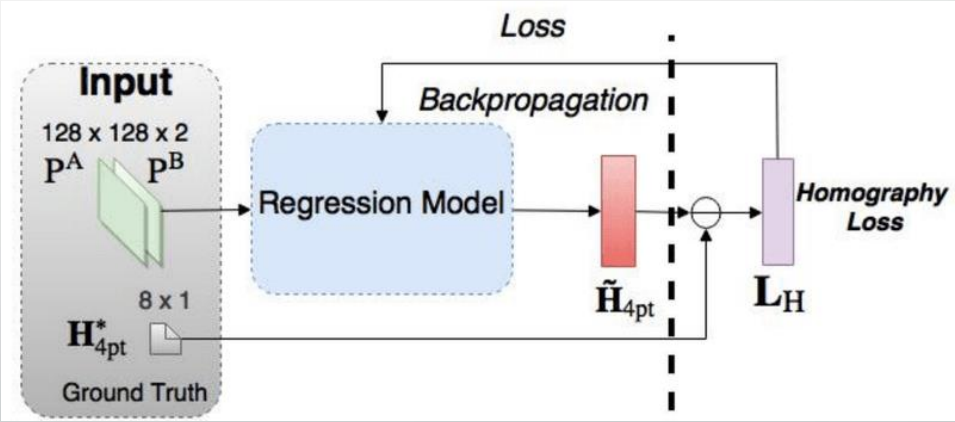
8.1 图像配准

8.1.2 基于深度学习的图像配准

1. 特征提取
2. 单应变换学习

2.1 监督学习

- 网络产生八个数值作为输出。以监督的方式进行训练，并计算输出和真实单应性之间的欧几里德损失。



8.1 图像配准

8.1.2 基于深度学习的图像配准

1. 特征提取
2. 单应变换学习

- 研利用神经网络直接学习几何变换对齐两幅图像，而不仅仅局限于特征提取。

  2.1 监督学习

  - 与其他有监督方法一样，该单应性估计方法需要有标记数据。虽然很容易获得真实图像的单应性，但在实际数据上要昂贵得多。

[1]. Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, Deep Image Homography Estimation, arXiv, 2016.

8.1图像配准

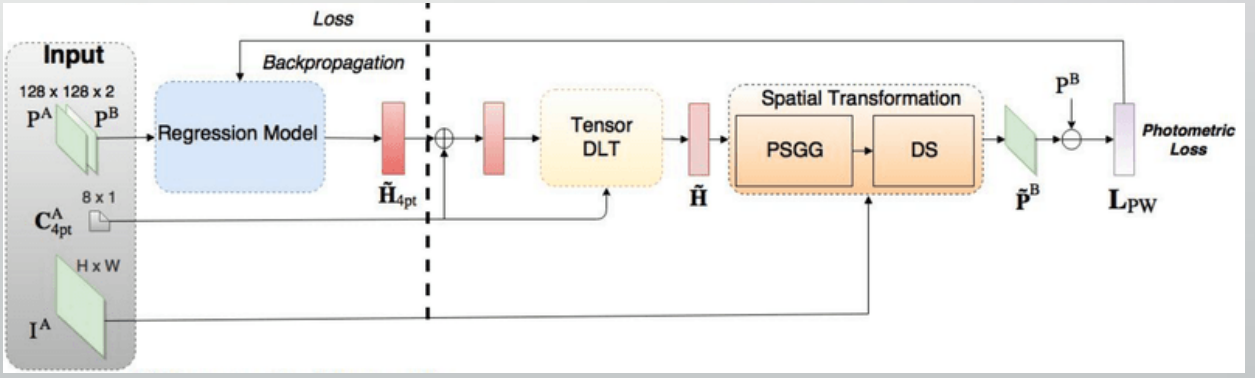
8.1.2 基于深度学习的图像配准

1. 特征提取
2. 单应变换学习

- 研利用神经网络直接学习几何变换对齐两幅图像，而不仅仅局限于特征提取。

2.2 无监督学习

- 基于这个想法，Nguyen等人提出了无监督的深度图像单应性估计方法[2]。他们保留了相同结构的CNN，但是使用适合无监督方法的损失函数：不需要人工标签的光度损失（photometric loss）函数。相反，它计算参考图像和感测变换图像之间的相似性。
- 他们的方法引入了两种新的网络结构：张量直接线性变换和空间变换层。可以使用回归模型输出的单应性参数获得变换后的感测图像，然后我们使用它们来计算光度损失。

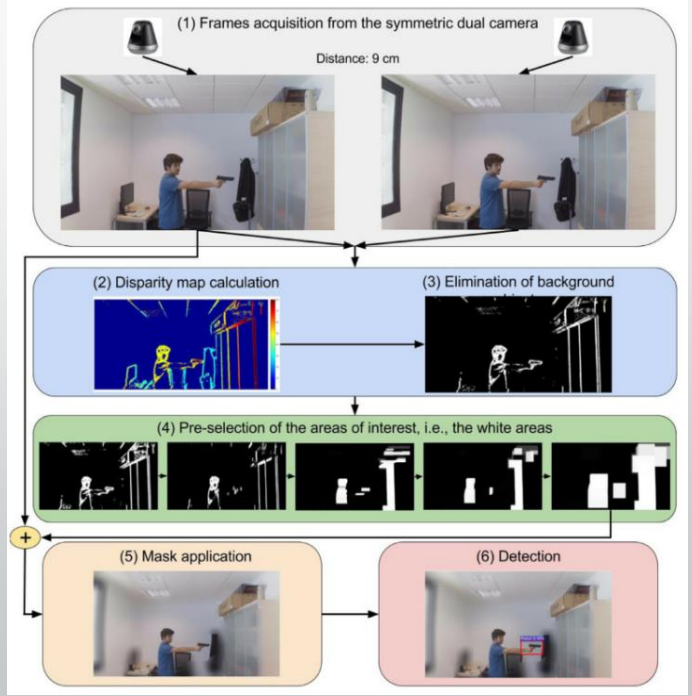


[2]. Ty Nguyen , Steven W. Chen, Shreyas S. Shivakumar, Camillo J. Taylor, Vijay Kumar, Unsupervised Deep Homography: A Fast and Robust Homography Estimation Model, ICRA 2018.

8.2双目图像融合

双目视觉（Binocular Stereo Vision）

- 基于双相机构建的双目视觉（Binocular Stereo Vision）系统是最小的立体视觉系统。
  - 双目视觉系统与生物的眼睛类似，不仅可以获取场景的平面图像信息，还能计算被测目标的距离和相对深度信息。
  - 它能完成大多数三维场景下的任务，而且多目视觉系统可看作是多个双目系统的组合，因此对双目系统的研究就成了立体视觉系统的研究重点。
- Roberto Olmos等人提出了一种双目图像融合方法[3]。



[3]. Olmos R ,Tabik S , Castillo A , et al. A binocular image fusion approach for minimizing false positives in handgun detection with deep learning[J]. Information fusion, 2019, 49: 271-280.

8.2双目图像融合

双目视觉（Binocular Stereo Vision）

- 双目图像融合方法[3]计算基于对称双目视觉方法的视差图，并使用这些信息来预先选择感兴趣的区域，推断在场景中更可能发生的动作。

■ 这项工作是第一次使用对称双目视觉和视差图来减少视频中物体检测的误报数量。

- 此方法具体分为以下五个步骤：

1. 从对称双摄像头获取帧

  - 一般情况下，双摄像头系统需要一个外部开关来同步捕捉帧的时间。然而，这项工作的目的是建立一个只基于相机而没有外部开关的系统。将相机的视野设置为中轴平行，两个相机镜头中心之间的距离设置为9厘米。双摄像头系统是用一个2.4厘米×2.4厘米正方形的棋盘图像来制作的。

[3]. Olmos R ,Tabik S , Castillo A , et al. A binocular image fusion approach for minimizing false positives in handgun detection with deep learning[J]. Information fusion, 2019, 49: 271-280.

8.2双目图像融合

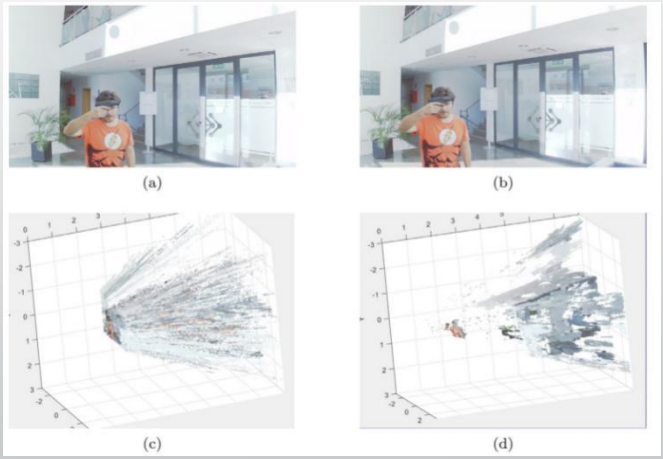
双目视觉（Binocular Stereo Vision）

- 此方法具体分为以下五个步骤：

  1. 从对称双摄像头获取帧

  2. 视差图计算

  - 文章评估了两种算法，块匹配(BM)算法和半全局块匹配(SGBM)算法。一般情况下，这些算法计算的是左侧相机拍摄的图像像素区域与右侧相机拍摄的图像像素区域之间的距离。
  - BM和SGBM算法利用这些信息来估计摄像机和场景中物体之间的距离。这种估计的结果以视差图的形式表示。



根据左(a)和右(b)图像中获取的信息，采用BM算法(c)和SGBM算法(d)计算出的视差图。

8.2双目图像融合

双目视觉（Binocular Stereo Vision）

- 此方法具体分为以下五个步骤：

  1. 从对称双摄像头获取帧
  2. 视差图计算
  3. 背景对象的消除

  - 在计算了视差图之后，选择了一个有限的距离来确定位于这个距离后面的物体。场景中极限距离的选择取决于当前场景的尺寸。距离是由双摄像头系统计算出来的。

8.2双目图像融合

双目视觉（Binocular Stereo Vision）

- 此方法具体分为以下五个步骤：

  1. 从对称双摄像头获取帧
  2. 视差图计算
  3. 背景对象的消除
  4. 预先选择感兴趣的领域

- 双目视差图允许从相机中分辨出更远和更近的物体，因此这些物体可以根据它们的距离被消除。但在实际应用中，由于场景中光照的影响和原始图像的质量较低，生成的视差图存在一定的缺陷，使得消除过程十分困难。由此得到的视差图包含了一定程度的噪声，并且在几个边界处显示出不连续。为了对片段进行清理和改进，采用了一系列形态二元运算。

8.2双目图像融合

双目视觉（Binocular Stereo Vision）

- 此方法具体分为以下五个步骤：

1. 从对称双摄像头获取帧
2. 视差图计算
3. 背景对象的消除
4. 预先选择感兴趣的领域
5. 掩码的应用和检测过程

- 从上一步得到的掩码应用于原始图像，原始图像中与掩码的白色区域对应的部分被保留了下来，而原始图像中与遮罩的黑色区域相对应的部分是模糊的。之后，将探测器应用于整个图像，探测器将只聚焦在感兴趣的区域。

8.3多目重构

多目视觉

- 基于一个场景的进行视图的重建时，通常有两个及两个以上的视图。

➢ 例如，我们可能使用单个移动摄像机拍摄的视频序列来构建3D模型，或者等效地使用静态摄像机拍摄的刚性移动对象的视频序列(图8-6)。这个问题通常被称为从运动图像中恢复三维结构（structure from motion , SfM）或多视图重建（multi-view reconstruction）。这是一种从摄像机在不同视点拍摄的多幅图像中恢复场景三维结构的方法。

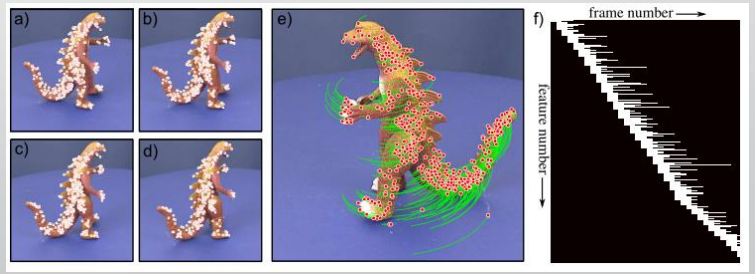


图8-6 从运动图像恢复三维结构

8.3多目重构

8.3.1立体匹配算法

- 双目视觉主要利用左右相机得到的两幅校正图像找到左右图像的匹配点，然后根据几何原理恢复出环境的三维信息。但该方法难点在于左右相机图片的匹配，匹配的不精确都会影响最后算法成像的效果。
- 多目视觉采用三个或三个以上摄像机来提高匹配的精度，但需要消耗更多的时间，实时性也更差。在双目/多目视觉中常见的匹配算法有：

1. SGM算法
   - 半全局立体匹配算法（Semi-Global Matching, SGM）是一种基于逐像素匹配的方法，该方法使用互信息来评价匹配代价，并通过组合很多一维的约束来近似一个全局的二维平滑约束。
2. 半全局块匹配(SGBM)

8.3多目重构

8.3.2三维重建方法

- Chrischoy等人[4]使用深度卷积神经网络，从大量的训练数据中学习物体到物体底层三维形状的映射，而不是在对物体尝试匹配合适的三维形状并尽可能地适应它。
- 受早期使用机器学习来学习2D到3D映射以进行场景理解的研究启发，文章提出了数据驱动方法来解决在给定数量的对象类别中仅从单个图像恢复对象形状的难题。
- 该方法第一次利用深度神经网络，以端到端的方式，从数据中自动学习适当的中间表示，从极少监督的单个图像中恢复近似的3D对象，从而实现重建。

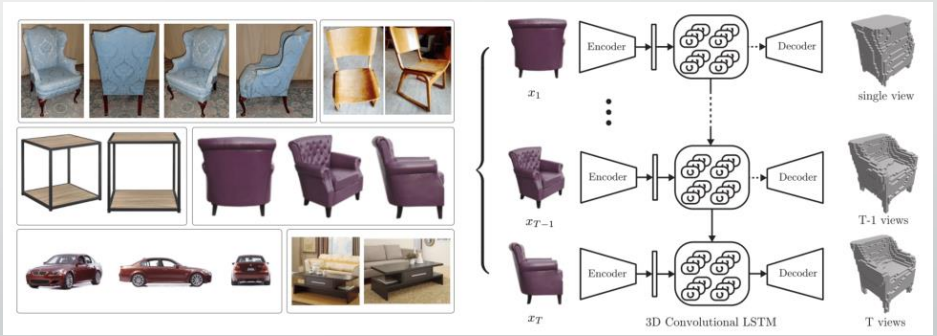


图8-7 (a)待重构的物体 (b)网络架构概述

[4]. Choy C B, Xu D, Gwak J Y, et al. 3D-R2N2: a unified approach for single and multi-view 3d object reconstruction[C]. //European conference on computer vision, October 11-14, The Netherlands, 2016, 628-644 .

8.3多目重构

8.3.2三维重建方法

- 网络由三部分组成：

1. 一个二维卷积神经网络(2D-CNN)
   - 从任意角度给出一个或多个物体的图像，2D-CNN首先将每个输入图像x编码为低维特征T(x)。
2. 一种新的体系结构命名为3D卷积LSTM (3D-LSTM)
   - 根据给定编码后的输入，新提出的3D卷积LSTM (3D-LSTM)单元做出两种类型的单元更新，一种是有选择地更新其单元状态，另一种为关闭输入门从而保留状态。
3. 3D Deconvolutional神经网络(3D-DCNN)
   - 3D-DCNN解码LSTM单元的隐藏状态，并生成3D概率体元重建。网络的整体架构如图8-7所示。

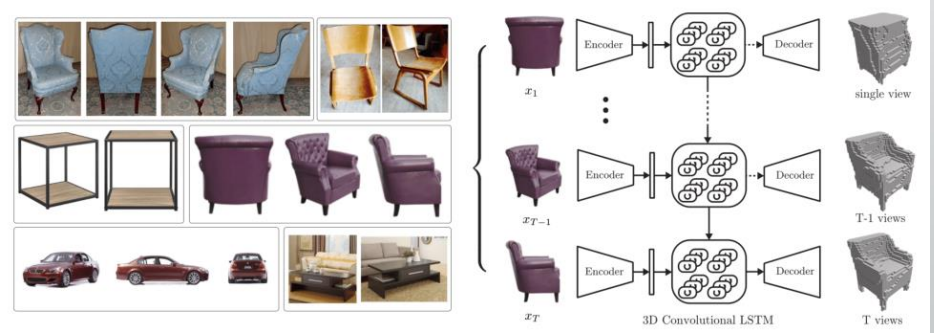


图8-7 (a)待重构的物体 (b)网络架构概述

[4]. Choy C B, Xu D, Gwak J Y, et al. 3D-R2N2: a unified approach for single and multi-view 3d object reconstruction[C]. //European conference on computer vision, October 11-14, The Netherlands, 2016, 628-644 .