# 样本库设计方案

Nercar

# 目录

# 一、 设计目标

大数据时代，数据本身就是核心竞争力。但我们有史以来忽视了数据的价值，采集的图像数据束之高阁，尘封在历史的硬盘中。或者在不同历史时期，由不同的历史人物，反复进行了独立的图像样本库的整理工作，包括现场项目实施和学生课题。造成历史循环的同时，也无法积累经验和数据。测试算法时没有一个统一的数据平台进行平行比较，仅靠经验而非统计设计算法，如同粗制劣造的黑火药，在面对西方列强的 TNT（算法）时，恐重现八国联军入侵的杯具。

回顾历史，发现之所以无法建立一套统一的样本库体系，除了组织上的原因外，需要解决如下技术问题：

1) 不同软件开发工具，样本库格式不统一；

2) 不同项目有不同使用需要，样本库内容不统一；

3) 同一项目不同历史时期，会建立多个样本库。

本项目的目的是：建立一套通用性强、便于进行版本管理的样本库管理系统。通过如下方法解决上述问题：

1) 使用通用性强的 Json 格式的样本库文件（支持 Matlab，C++语言），支持导出为子文件夹+图像文件 形式的样本库（支持 Halcon、CVB 导入）；

2) 使用本地化的 Git 版本管理工具，便于管理样本库的不同历史版本；

3) 便于从数据库导出样本文件的工具、便于编辑管理样本库的工具。

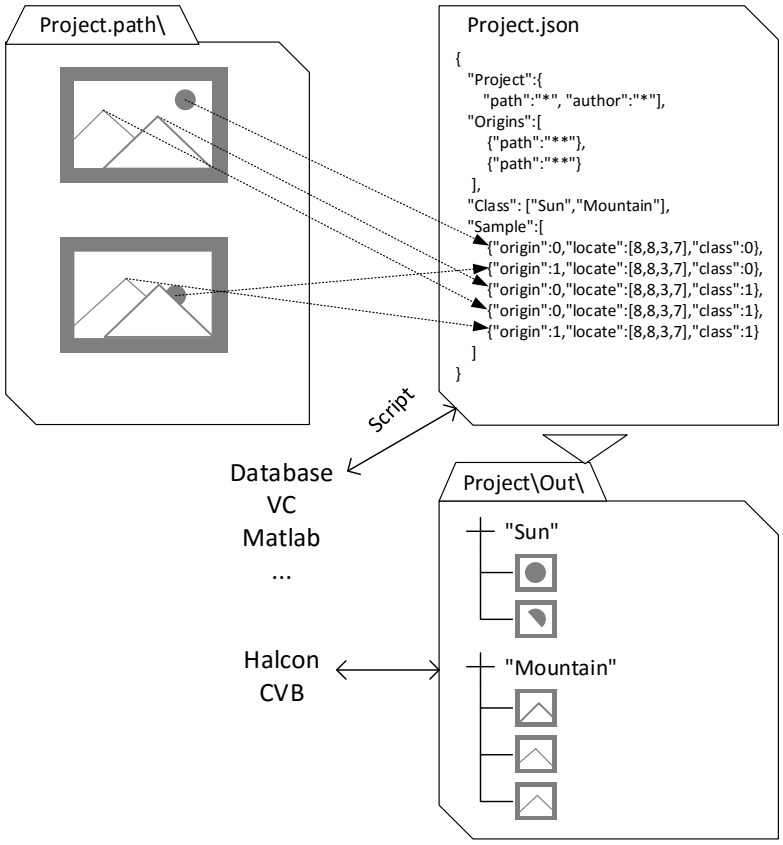
# 二、 项目实现

## 1. 数据格式

从分类器需求上分析：CVB 支持固定长宽尺寸的局部样本图像；Halcon 支持一定范围内可变长宽局部样本图像；卷积神经网络需要整体图像和局部样本坐标。同时预处理算法需要整幅图像进行处理。因此，有必要存储整幅尺寸的原始图像。

使用 Json 格式的样本库文件，记录局部样本图像的类型、源图像路径、尺寸和位置。Json 格式便于多种语言解析，编写导入导出脚本，同时拓展性强便于后续增加新功能。

基于上述原因，设计样本库数据结构为：




图一：数据格式

## 2. 模块构成

该系统涉及样本库服务器、客户端、项目服务器三者间的数据交互：

1) 样本库服务器上设有 Git 服务器服务以及相应的 Git 网络版本库，通过 VPN 服务与公网联通。

2) 客户端上设有 Git 本地版本库，通过 Git 版本管理工具用于管理本地版本库并与 Git 服务器进行推送/更新操作；同时客户端上设有样本库编辑程序，用于编辑和生成样本库的不同版本；算法测试程序用于训练、测试不同版本的样本库。

3) 项目服务器上需要通过一个数据库导出工具，将项目缺陷和图像数据库中的样本图像转换为图一所示的数据格式，保存到本地版本库中生成新的版本。

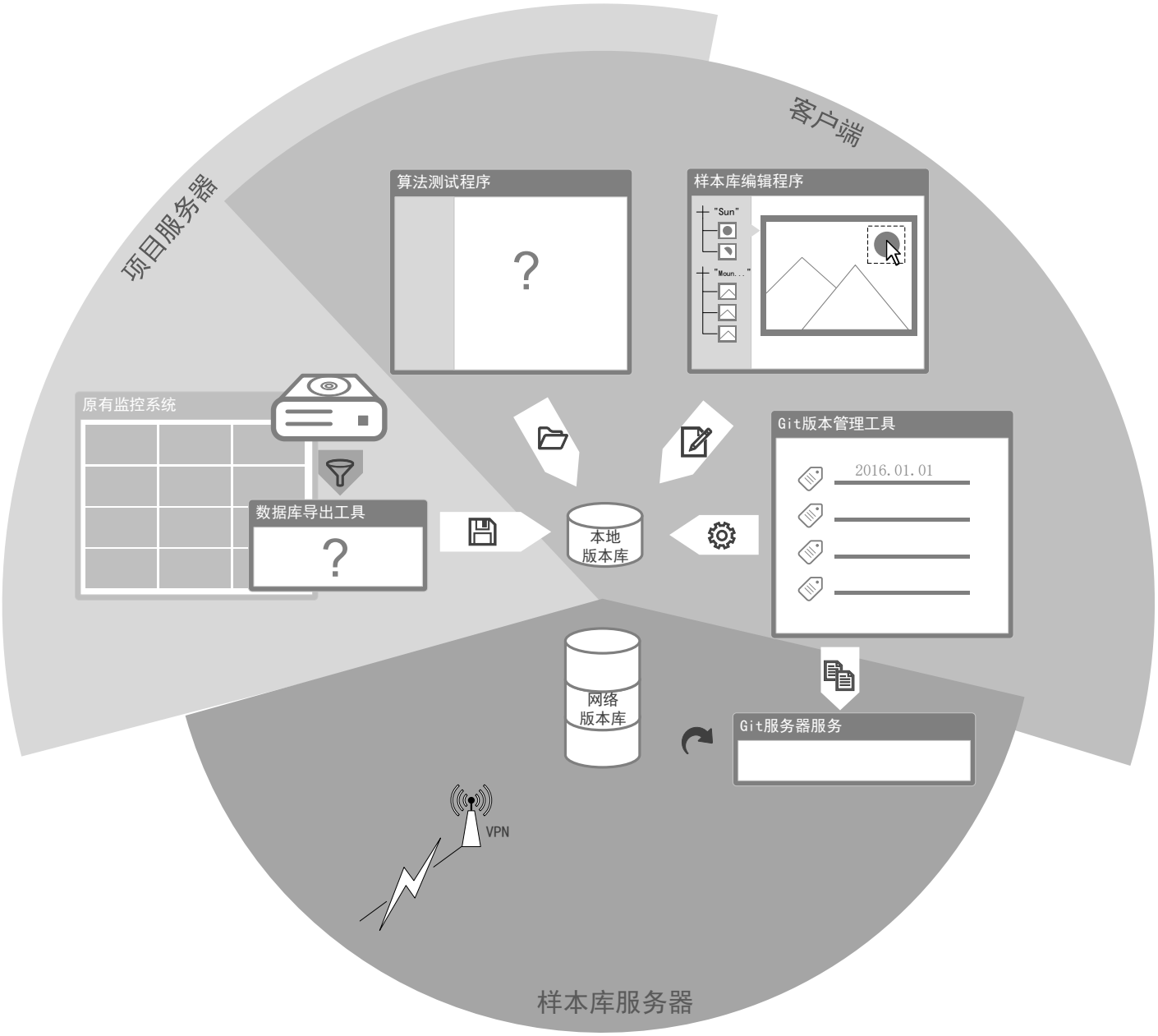
# 三、 进度管理

1) Git 服务器服务使用 Gitblit 搭建，Git 版本管理工具使用 SourceTree，均已搭建完成并测试成功。

2) 样本库编辑程序使用 Html5 完成，正在调试中。

3) 算法测试程序未完成，需要熟悉各平台算法。

4) 数据库导出工具未完成，需要 1.读取项目数据库并转为图一所示数据格式；2.实现 img 格式图像转换，实现双通道独立图像转换为 bmp24 格式彩色图像；3.可集成如项目用户界面中。


项目服务器
客户端
算法测试程序
?
样本库编辑程序
"Sun"
"Moun..."
原有监控系统
数据库导出工具
?
本地
版本库
Git版本管理工具
2016.01.01
网络
版本库
Git服务器服务
VPN
样本库服务器

图二：模块构成

## 四、 参与