判别学习卷积特征描述子，应用于人脸认证

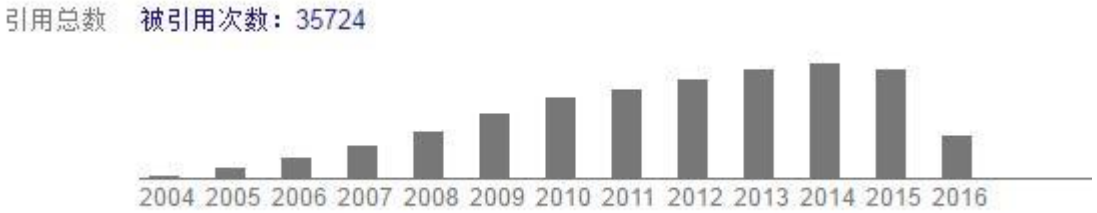
一、选题的意义

特征描述子广泛应用在计算机视觉领域，如运动恢复结构（SfM），三角化(Triangulation)，3D重建，全景图拼接，图像特征点匹配，图像匹配，虚拟现实（VR），全息投影，图像搜索，目标识别，目标追踪等等。经典的特征描述子如SIFT，SURF等对计算机视觉的发展起到了重要作用。当前的特征特征描述子需要精心的手工设计，主要表示图像的低级特征，容易受场景光照变化，拍摄视角差异，传感器设置不同的影响。

一、选题的意义

从谷歌学术引用量看行业趋势

Distinctive image features from scale-invariant keypoints. [2004]



浅层特征

引用量已经趋于稳定

Imagenet classification with deep convolutional neural networks. [2012]

深度特征

引用量快速上升

一、选题的意义

下面先通过计算机视觉方面的应用介绍特征描述子的重要性。

- 运动恢复结构（SfM），也可以称为稀疏重建。
- 三角化（ Triangulation ）。
- 3D重建。
- 全景图拼接。
- 图像特征点匹配。
- 图像匹配。
- 图像搜索。
- 视频检索。
- 人脸认证。

一、选题的意义

运动恢复结构（稀疏重建）

多视角图像

相机位姿+场景结构

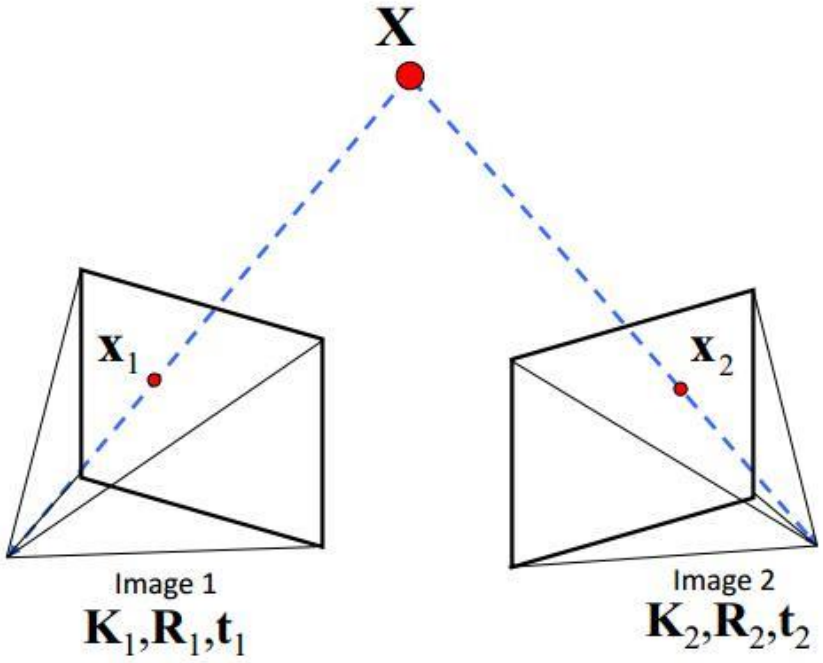
网络上众多罗马斗兽场的图像，如何寻找有重叠区域的相邻图像进行稀疏重建？

局部特征描述子

一、选题的意义

三角化（ Triangulation ）

已知x、K、 R、 t, 求X

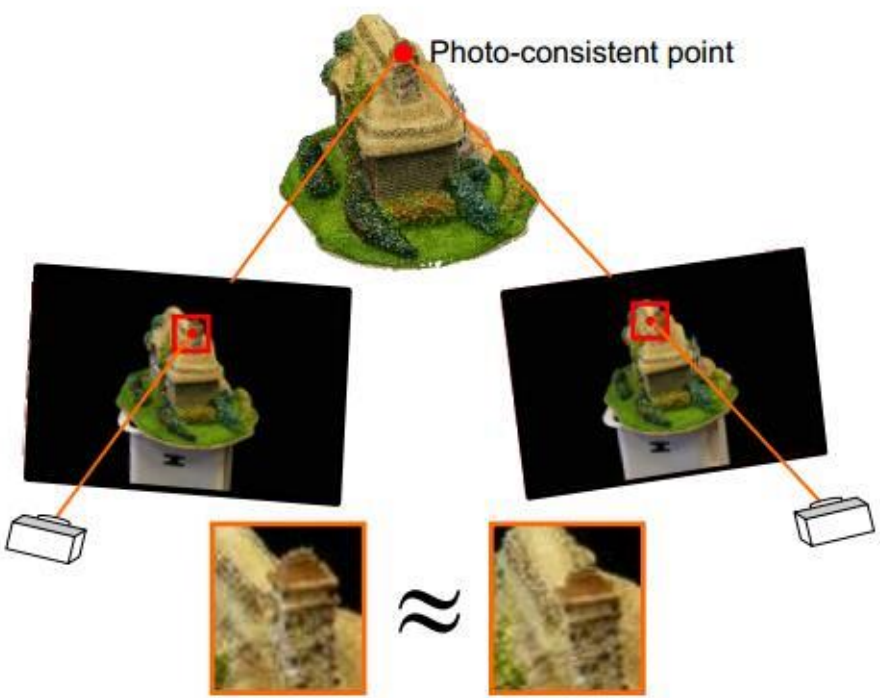


如何判断点x1和点x2是同一个点？

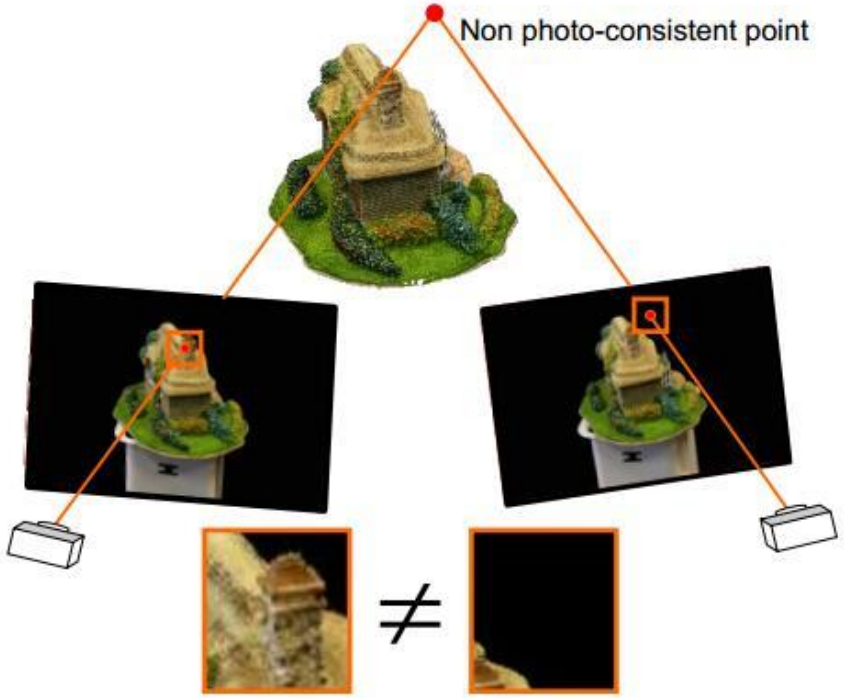
SIFT描述子
SURF描述子

一、选题的意义

三角化（ Triangulation ）



同一个点



不同点

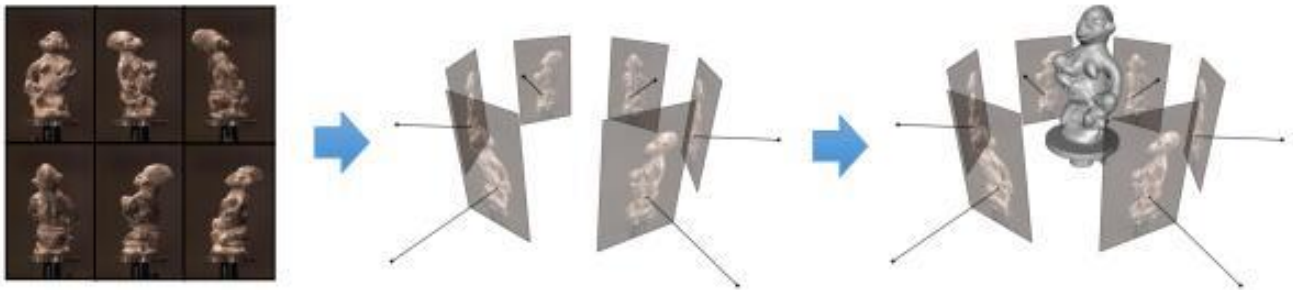
一、选题的意义

Triangulation 的应用

DJI PHANTOM 4

一、选题的意义

3D 重建



多视角图像　　相机位姿　　3D模型

如何从多视角图像序列中获取相邻视角的图像，并计算相机精确位姿？

特征描述子

一、选题的意义

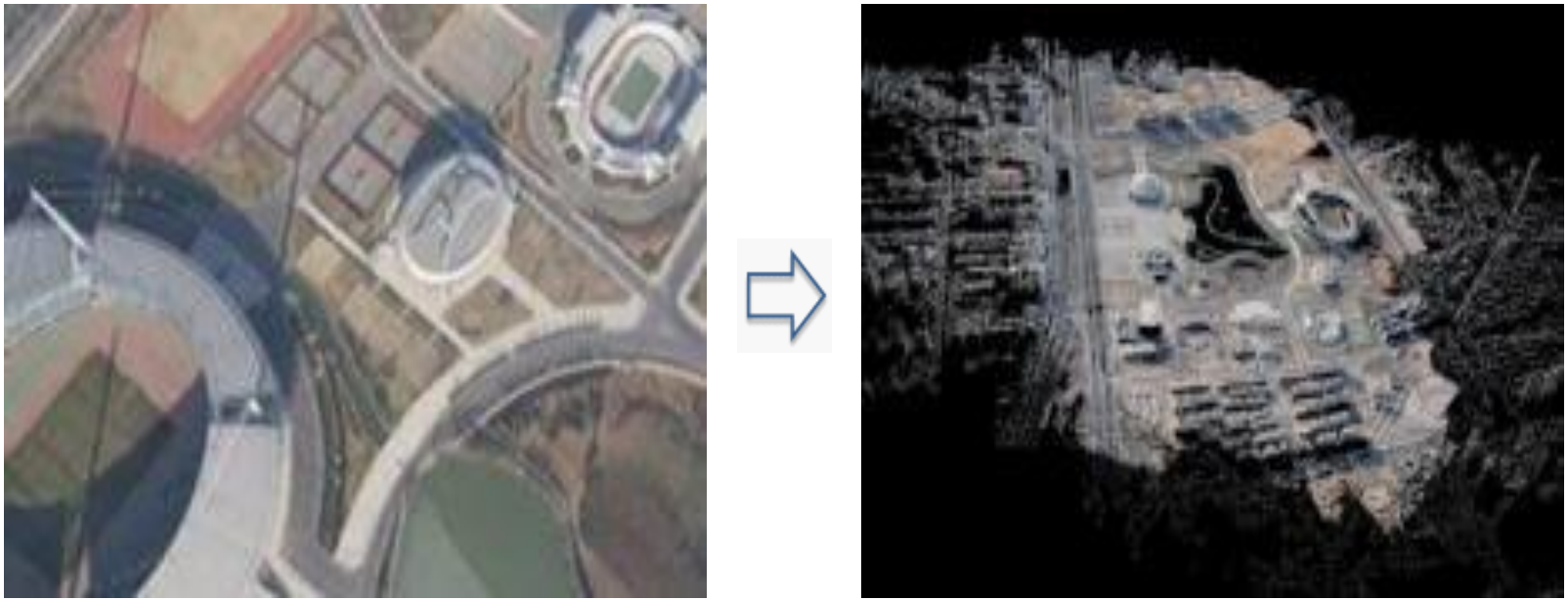
3D 重建应用-古建筑数字保护



颐和园石舫（第二行图为重建结果）

一、选题的意义

3D 重建应用-三维地图生成



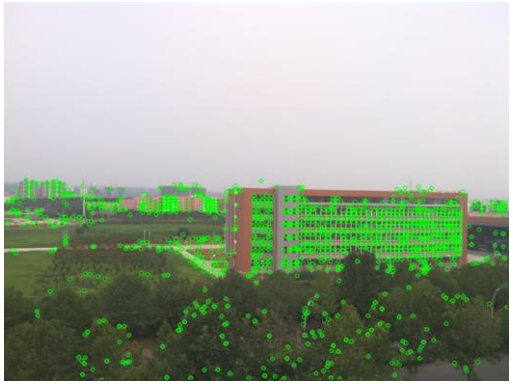
通过载人机或无人机航拍图像自动获取高精度三维地形数据，
可应用于智慧城市建设。

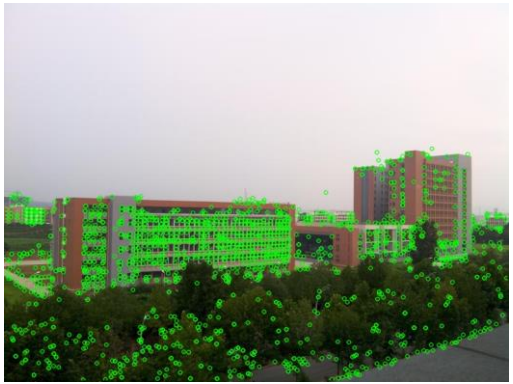
一、选题的意义

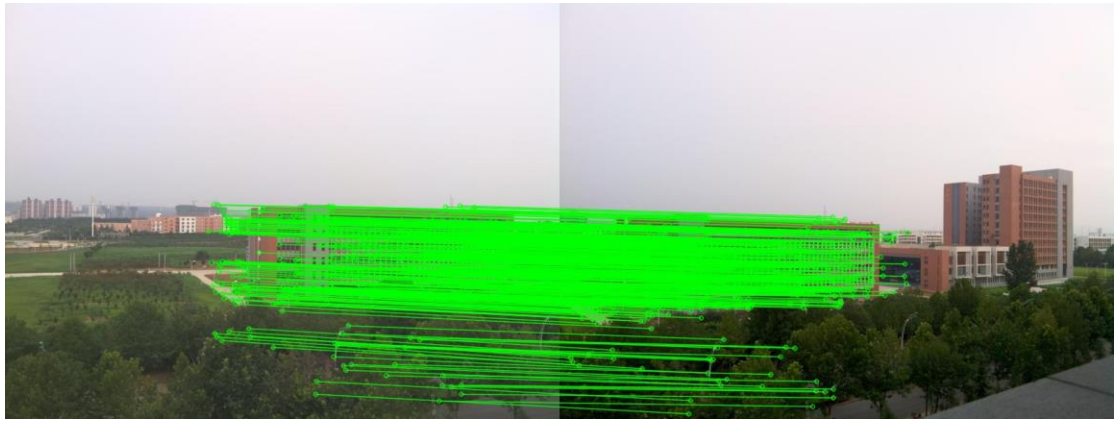
使用移动设备的3D重建

一、选题的意义

图像特征点匹配







通过特征点描述子向量匹配特征点，因此特征描述子必须具有尺度不变，旋转不变，一定范围内的视角不变，对光照变化不敏感等特性

一、选题的意义

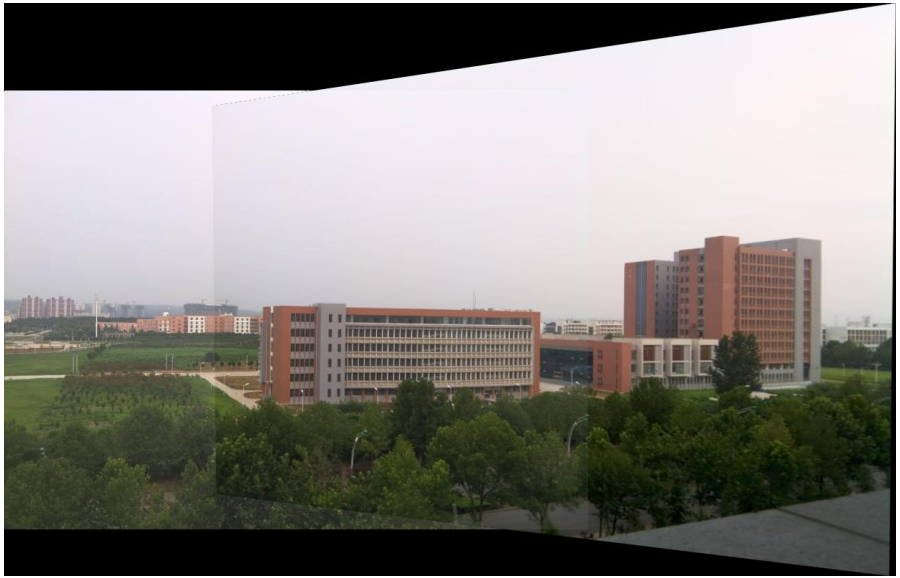
全景图拼接

通过对多幅航拍图像或者无人机拍摄的图像进行全景拼接，获取大范围广角地面图像，如何获取多福图像的重叠区域？

图像特征描述子

一、选题的意义

全景图拼接



对移动手持设备获取的图像进行拼接，
显著扩大可视化范围

一、选题的意义

图像匹配

非匹配图像对

匹配图像对

如何判断两幅图像是否是同一个场景？

局部图像描述子

一、选题的意义

图像匹配

光照，对比度，尺度不同，如何增强匹配性能？

一、选题的意义

图像搜索

Data Base

Input

Output

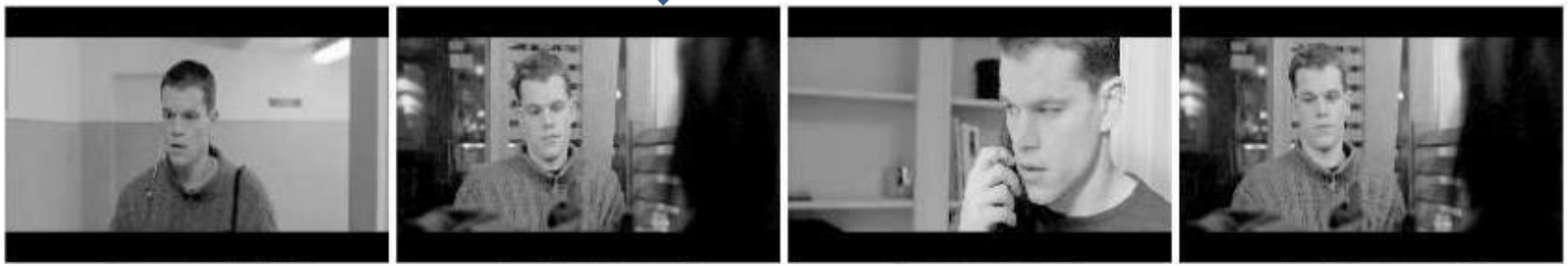
通过使用局部不变特征描述子进行图像搜索

一、选题的意义

视频检索

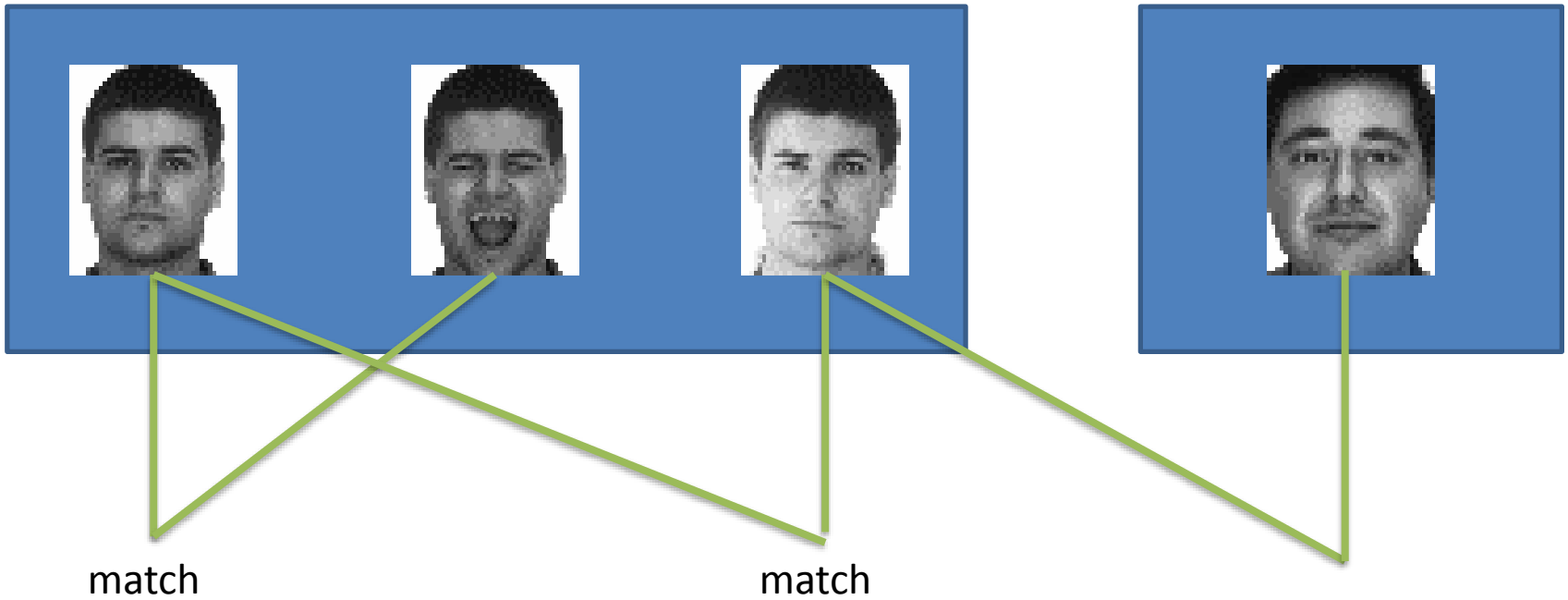
查询输入

Top 4 输出结果



一、选题的意义

人脸认证



如何使学习到的特征描述子对光照变化，局部变形不敏感？

使用深度神经网络学习局部的，抽象的高级特征。

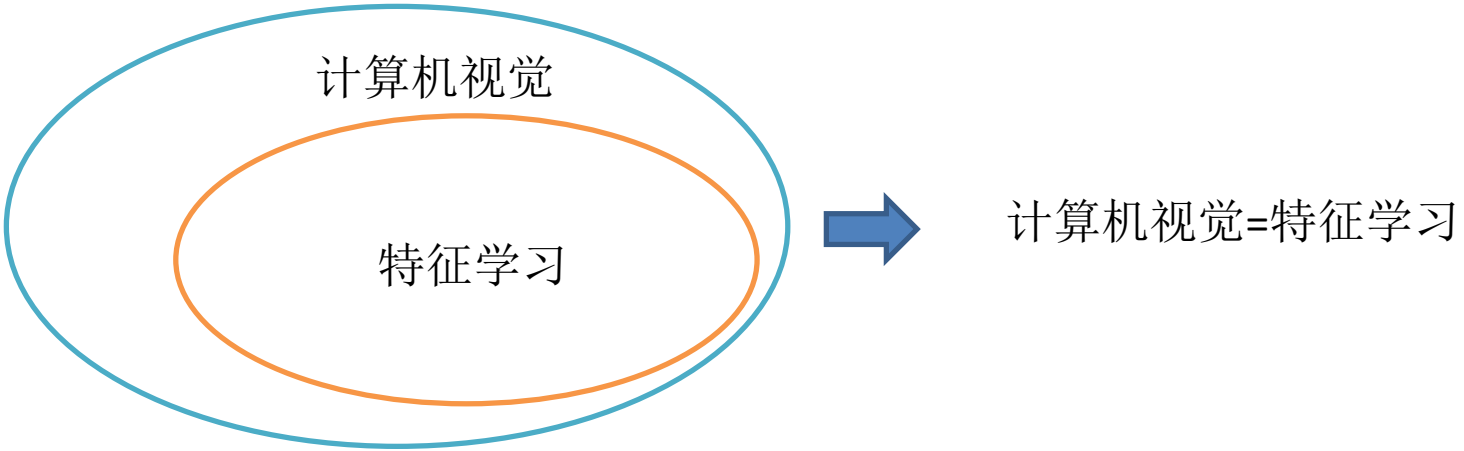
一、选题的意义

从上面介绍的例子看出，图像描述子已经广泛应用在计算机视觉领域，而且会对人们的生活方式产生非常深远的影响。大数据时代已经带来，也给我们带来了全新的图像描述子获取方法：深度神经网络。

通过深度神经网络学习区别的特征描述子，相比传统手工精心设计的描述子，深度特征描述子能够表示更加抽象的特征，对场景光照变换，视角变换，图像传感器设置不同，色差，局部区域变形具有较高的不变性。

一、选题的意义

目前，国内在深度学习领域做得比较出色的公司有百度，阿里巴巴，腾讯，地平线，格林深瞳等。其中，百度识图推出了一系列图像搜索产品，如相同检索，人脸检索，相似检索，图像识别，识图APP等。百度这些产品背后都离不开基于深度神经网络的区别特征学习。



一、选题的意义

我们从深度卷积神经网络成功应用于大规模图像分类任务中吸取灵感，利用机器学习方法构建卷积神经网络学习图像特征描述子，并应用于人脸验证。但是我们的框架并不仅仅局限于人脸认证。在这个框架下，可以学习到图像的深度区别特征描述子，并应用特征描述子完成其他计算机视觉任务，如图像搜索，特征点匹配，相似人脸检索等。

接下来第二部分介绍我们的算法框架。

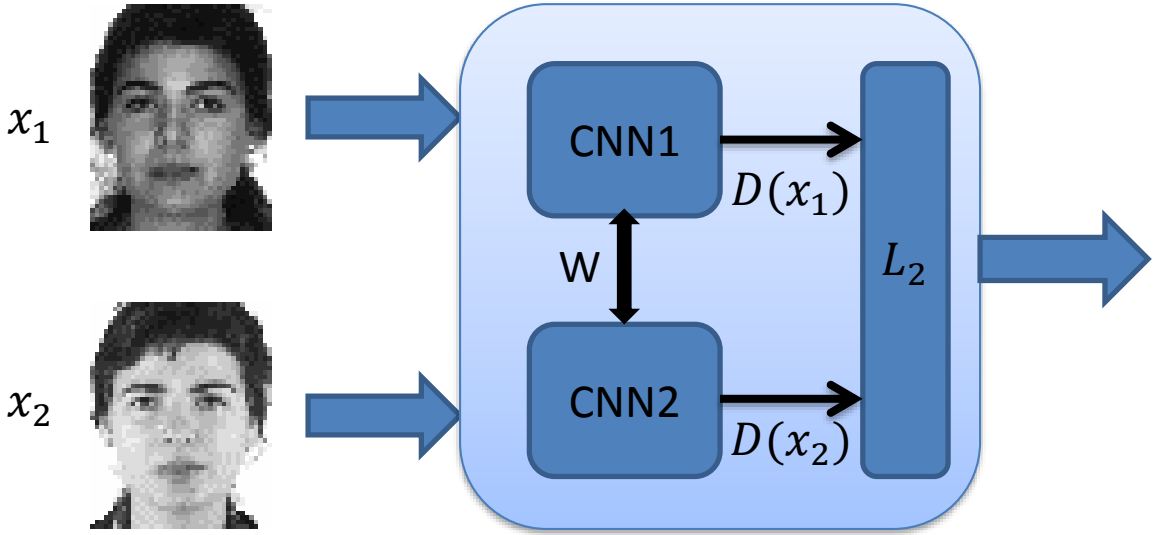
二、技术方案

开发平台

计算机型号	联想Lenove Y460c(PC)
CPU	Intel(R) Core(TM) i3-380M @2.53GHz
GPU	NVIDIA GeForce GT 425M独立显卡
内存	6GB(DDR3)
显存	1GB(DDR3)
操作系统	Windows 7(64位)旗舰版
开发环境	Matlab R2012a

二、技术方案

原理框图



CNN1和CNN2共享同样的参数，算法中CNN1和CNN2是同一个CNN网络

$D(x_1)$和$D(x_2)$表示网络学习到的输入图像的区别特征，$L(x_1, x_2)$表示网络训练过程中的损失函数。

二、技术方案

技术思路

由于我们希望网络能够学习到图像特征的区别表示，对于人脸验证即同一个人在不同时刻，不同光照条件，不同表情下获取的图像需要有尽可能相似的特征表示。因此，我们通过最小化损失函数强迫两幅同一个人脸图像特征之间的欧几里得距离尽可能小。同时，对于两幅不是同一个人的图像，我们通过最小化损失函数强迫其特征之间的欧几里得距离尽可能大。

令$E(x_1, x_2)=\|D(x_1)-D(x_2)\|_2$，表示两幅图像的特征之间的欧几里得距离，根据上面的分析，损失函数的选择必须满足下面的条件：

1. 如果x_1和x_2是同一个人脸图像，则$L(x_1, x_2)$是$E(x_1, x_2)$的单调递增函数。
2. 如果x_1和x_2是不同人脸的图像，则$L(x_1, x_2)$是$E(x_1, x_2)$的单调递减函数。

二、技术方案

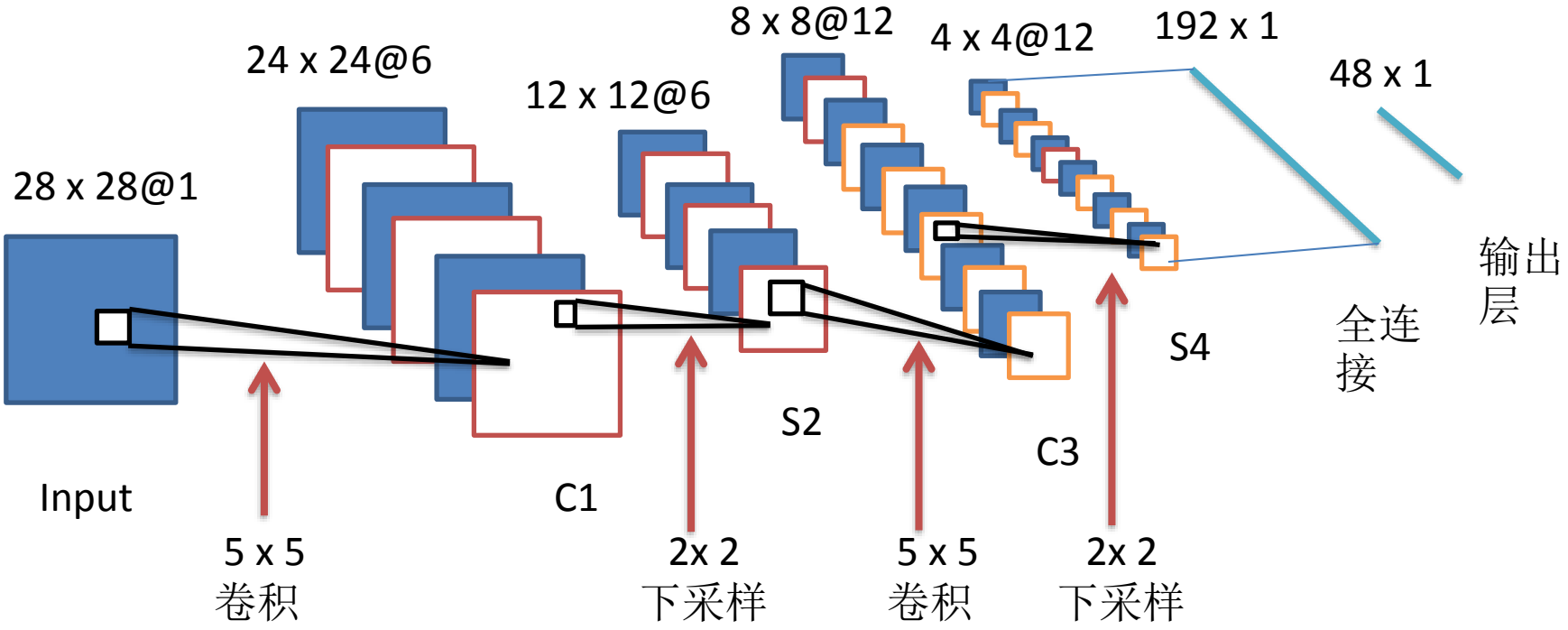
损失函数

$$L(x_1, x_2)=\begin{cases} \|D(x_1)-D(x_2)\|_2, & x_1和x_2指的是同一个人 \\ \max(0, C-\|D(x_1)-D(x_2)\|_2), & x_1和x_2指的不是同一个人 \end{cases}$$

C是一个常量，表示欧氏距离$\|D(x_1)-D(x_2)\|_2$的上界，实验中C值取为1，取得了较好的效果。从上面的损失函数可以看出，如果输入两幅图像是同一个人，网络会不断学习减少这两个图像的特征之间的欧式距离。如果输入的两幅图像不是同一个人，网络会学习增大这两个图像的特征之间的欧式距离。

二、技术方案

CNN网络架构



输出层节点个数设置为48，因此，特征描述子向量的维度是48。

二、技术方案

CNN网络架构

C1层和C3层的激活函数使用sigmoid函数，S2层和S4层采用mean-pooling策略，即对邻域内值求平均。

C1层：5*5*6+6=156

S2层：0

C3层：12*(5*5*6)+12=1812

S4层：0

输出层：192*48+48=9264

网络所有参数个数：156+1812+9264=11232

二、技术方案

数据集和数据预处理

网络上获取的AR数据集包含120个人，每人14幅图像，其中男性60人，女性60人。每个人的14幅图像之间存在光照变化，表情变化等，非常接近现实中情况。我们选择其中50个男性和50个女性的数据集作为训练数据集，剩下的10位男性和10位女性的数据集作为测试数据。原始图像的分辨率为40 x 50，为了加快算法的训练过程和简化模型的复杂度，我们下采样所有数据集为28 x 28。我们承认，下采样丢失了细节信息，肯定会增大算法的误差率。但是，在目前的模型下，我们的算法已经表现出了出色的性能。我们相信，通过输入原始数据并加大数据量，同时增加网络的复杂度，一定会获得更好的结果。

二、技术方案

AR数据集训练样本展示

二、技术方案

训练和测试数据集生成

不同于通常的网络训练过程，在我们的模型训练过程中输入的是一对图像，这两幅图像或者是同一个人，或者是不同的人，因此我们必须从训练集合中生成足够的同一个人的图像对（一致图像对）和不是一个人的图像对（非一致图像对）。在测试图像集合中也这样做。

前面讲到，训练集合共100人，每人14幅图像，因此对于每个人的图像共有C_{14}^{2}组合构成一致图像对，因此训练集共有一致图像对数为C_{14}^{2}*100=9100。然而，不同人的组合方式是巨大的，此数值是14*（99*14）*100/2=970200。因此，为了减少非一致图像对的个数，对每个人的每幅图像，首先从剩余的99人中的随机选择50个人，然后从这50个人各自随机选择一副图像，这样对于任意一副图像和其他50幅可以组成50个不同的非一致图像对，最终非一致图像对个数是100*14*50=70000。

二、技术方案

训练和测试数据集生成

从上面可以看出训练数据集中非一致图像对个数大约是一致图像对个数的7倍，在实验过程中我们发现这个比例关系是合理的。由于非一致图像对是通过随机采样获取的，因此存在采样重叠，然而一致图像对不会出现重叠，因此非一致图像对个数多于一致图像对个数也是必须的。

为了评估训练的模型，我们必须生成测试数据集，这些测试数据集是由前面提到的不在训练集中的另外20人组成的。不同于训练集中非一致图像对个数多于一致图像对个数，测试集中一致图像对个数和非一致图像对个数大约相等。一致图像对个数是C_{14}^{2}*20=1820。不同于训练集中随机采样50人，测试集中随机采样8人，因此采样重叠率大大降低，可知测试集中非一致图像对个数是20*14*8=2240。

二、技术方案

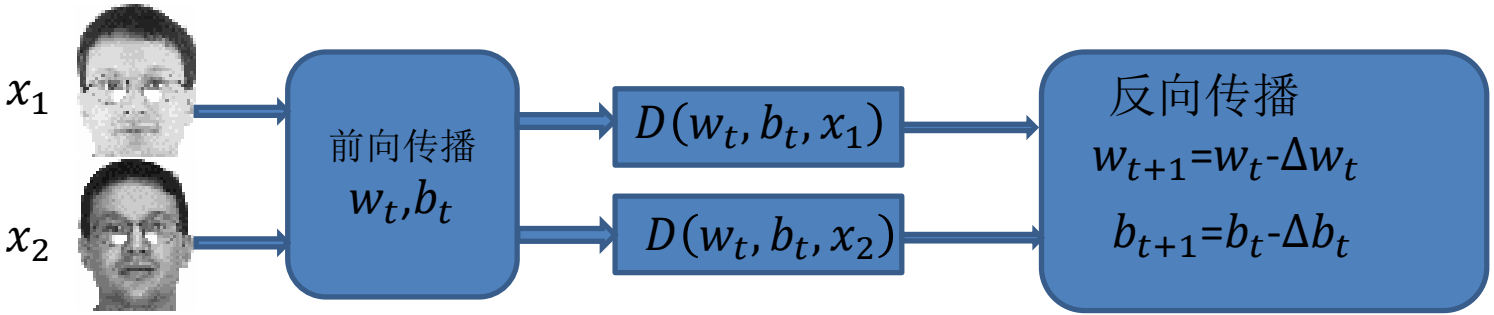
训练和测试数据集生成

	一致图像对个数	非一致图像对个数	总数
训练集	9100	70000	79100
测试集	1820	2240	4060

训练集和测试集数据是相互独立的，测试集中的数据并未在训练集中出现过。

二、技术方案

网络训练



$$\Delta w_t = \gamma * \frac{\partial L(w_t, b_t, x_1, x_2)}{\partial w_t} + \mu * w_t + \alpha * \Delta w_{t-1}$$

$$\Delta b_t = \gamma * \frac{\partial L(w_t, b_t, x_1, x_2)}{\partial b_t}$$

其中，γ表示学习系数，该模型设置为0.01。μ表示权重衰减项，该模型设置为0.0001。α表示冲量系数，该模型设置为0.9。mini-batches设置为50。

二、技术方案

特征描述子向量归一化

不同于训练过程中的损失函数直接计算两个图像描述子向量之间的欧式距离，在测试阶段，我们首先各自归一化两幅图的描述子向量，然后计算归一化之后的向量之间的欧式距离。

$$D'(x_1)=\frac{D(x_1)}{\|D(x_1)\|_2} \qquad D'(x_2)=\frac{D(x_2)}{\|D(x_2)\|_2}$$

二、技术方案

判决条件

$$dis = \|D'(x_1) - D'(x_2)\|_2$$

IF dis<th

　　x_1和x_2是同一个人。

ELSE

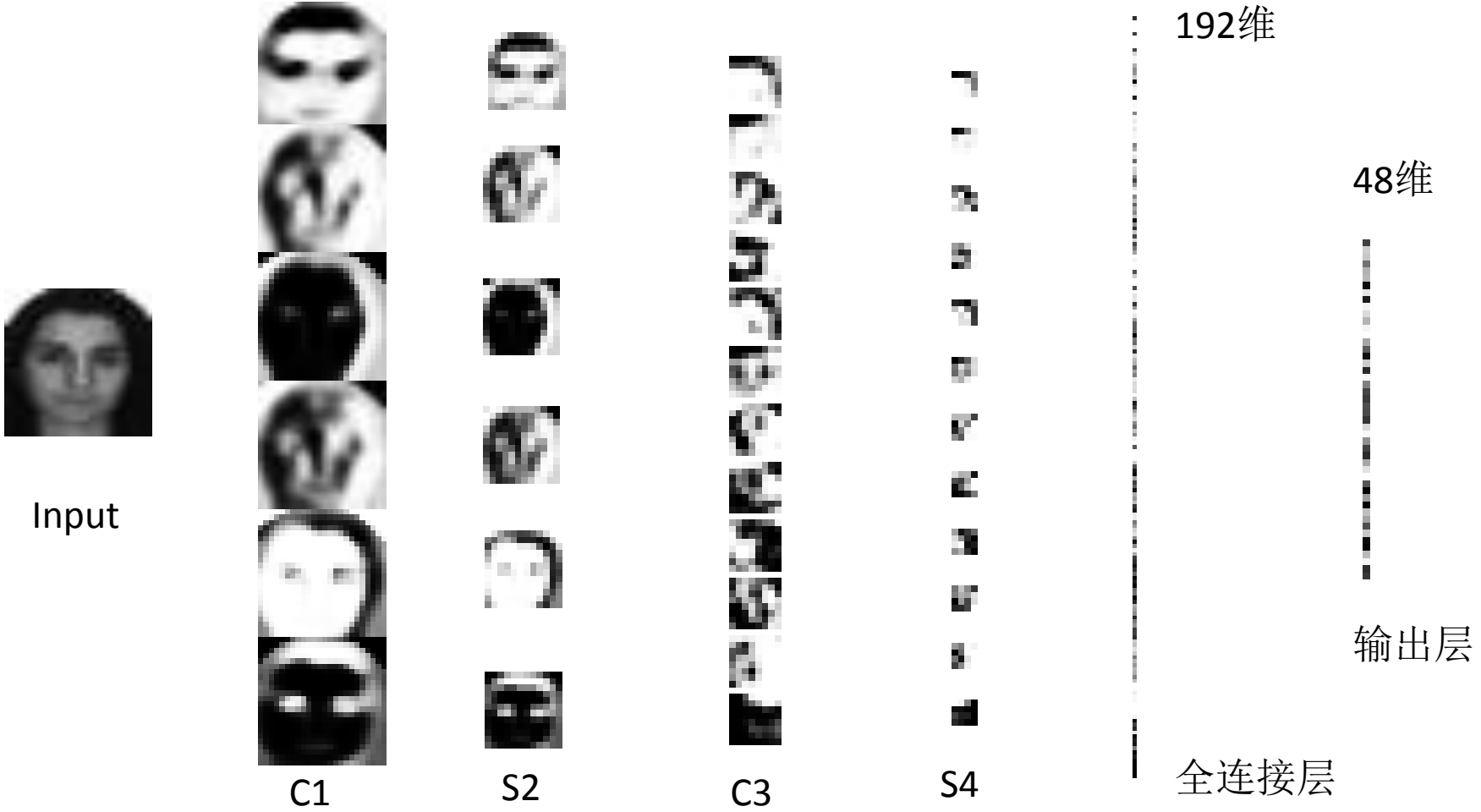
　　x_1和x_2不是同一个人。

End IF

其中dis表示归一化描述子向量之间的欧式距离，th是设置的阈值。

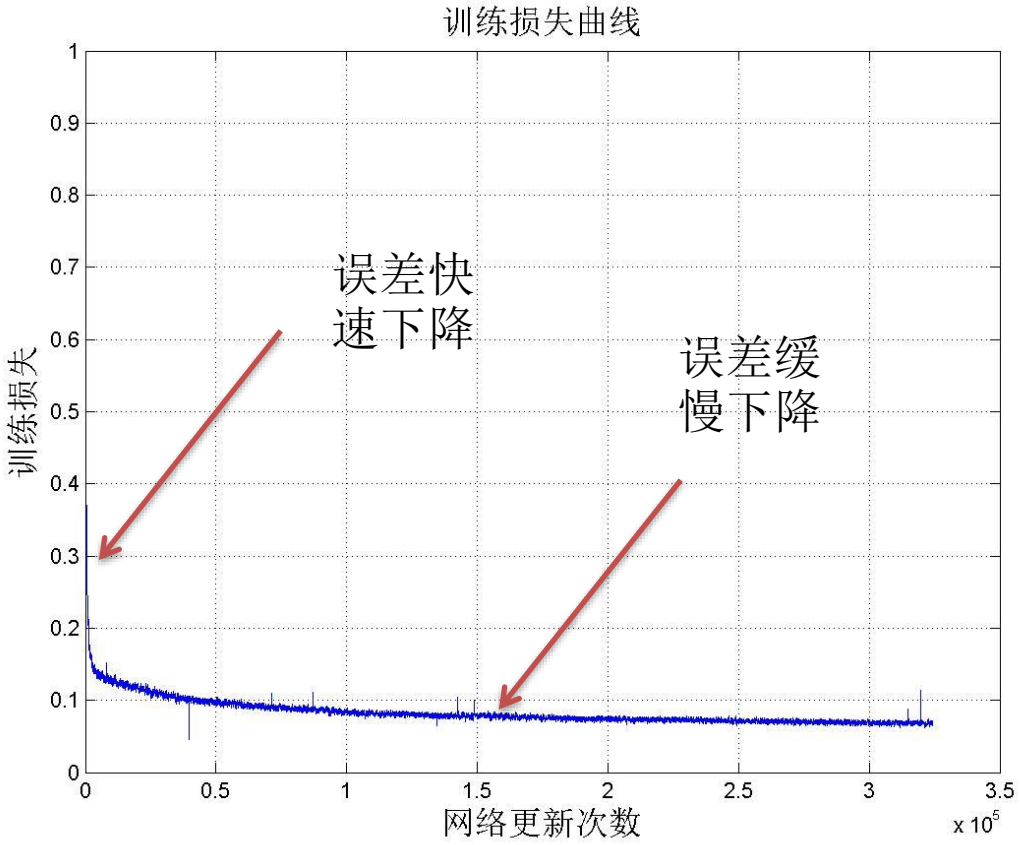
三、实验结果

网络对于特定输入的内部状态



三、实验结果

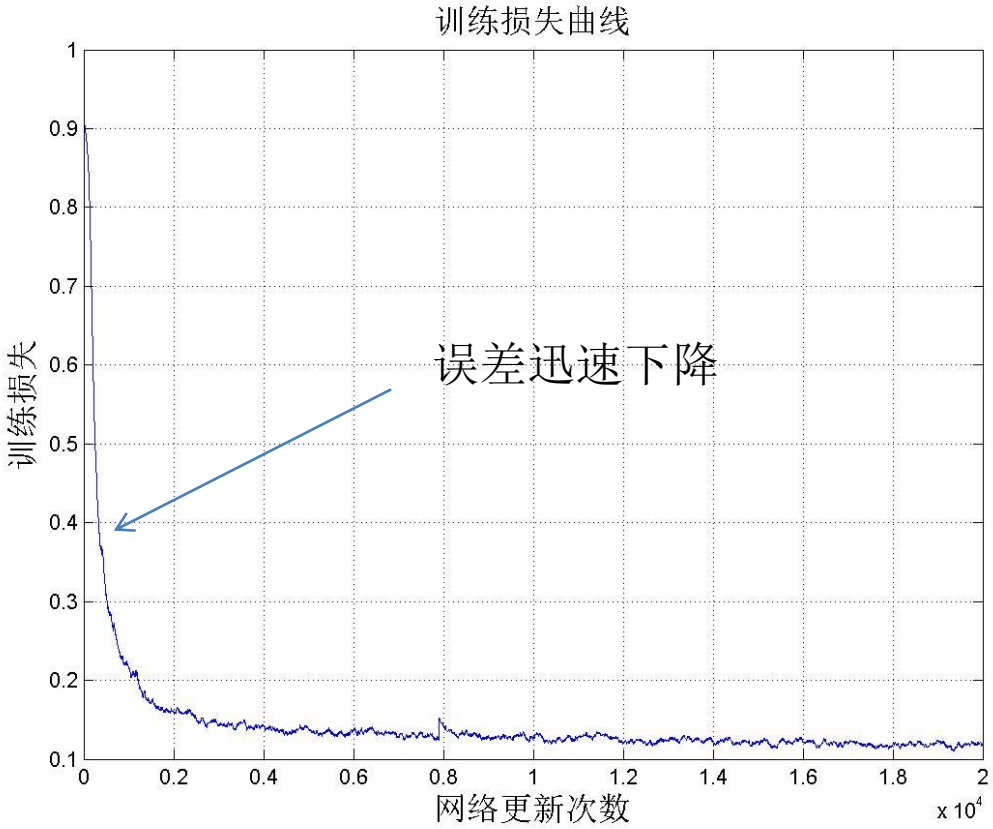
训练集损失曲线



共迭代205次

三、实验结果

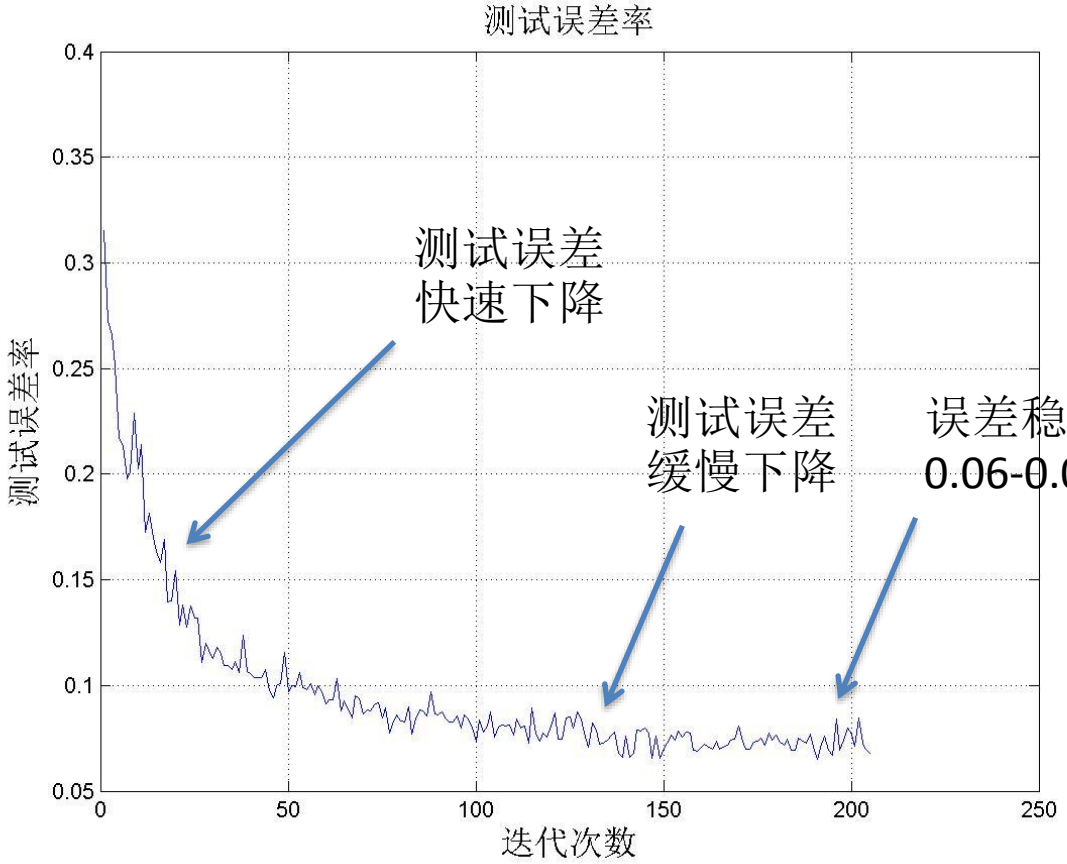
训练集损失曲线局部放大



为了清晰显示迭代初期误差曲线，进行了局部截图

三、实验结果

训练过程中在测试集上的误差率

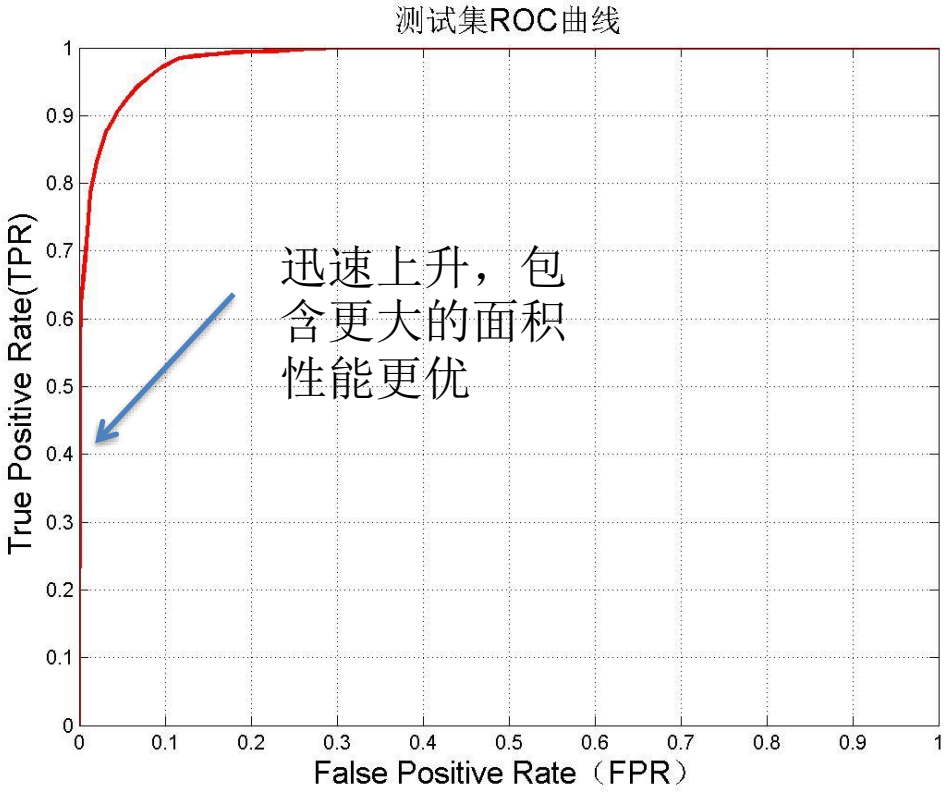


算法共迭代了205次

训练过程中设置阈值为0.18

三、实验结果

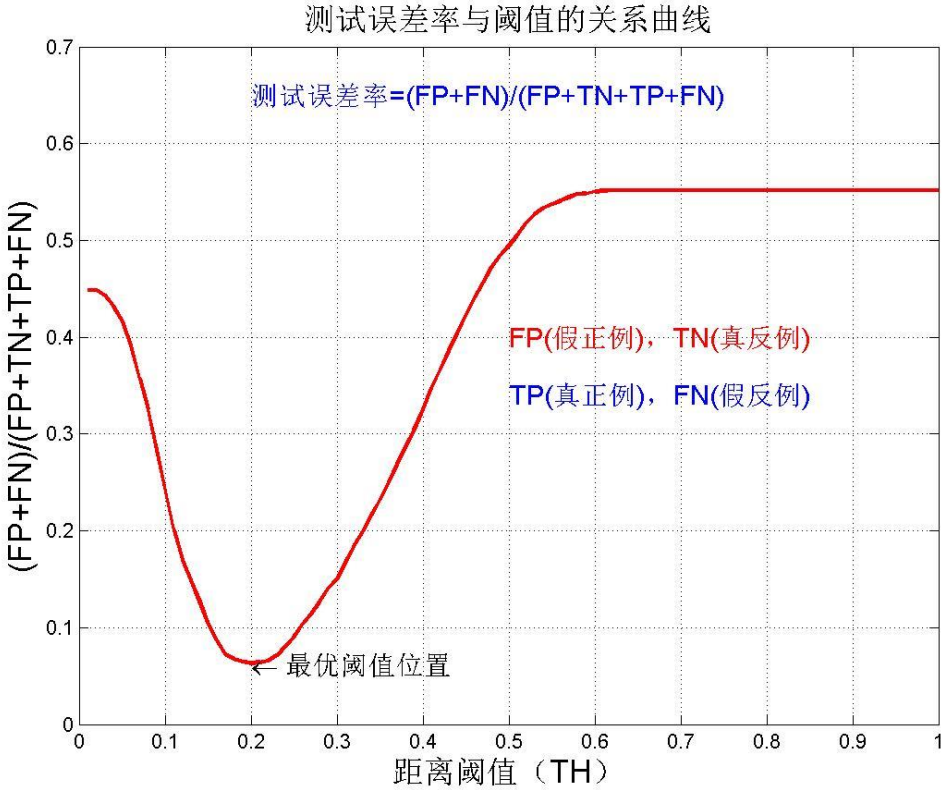
测试集ROC曲线



ROC曲线包含的面积越大性能越好

三、实验结果

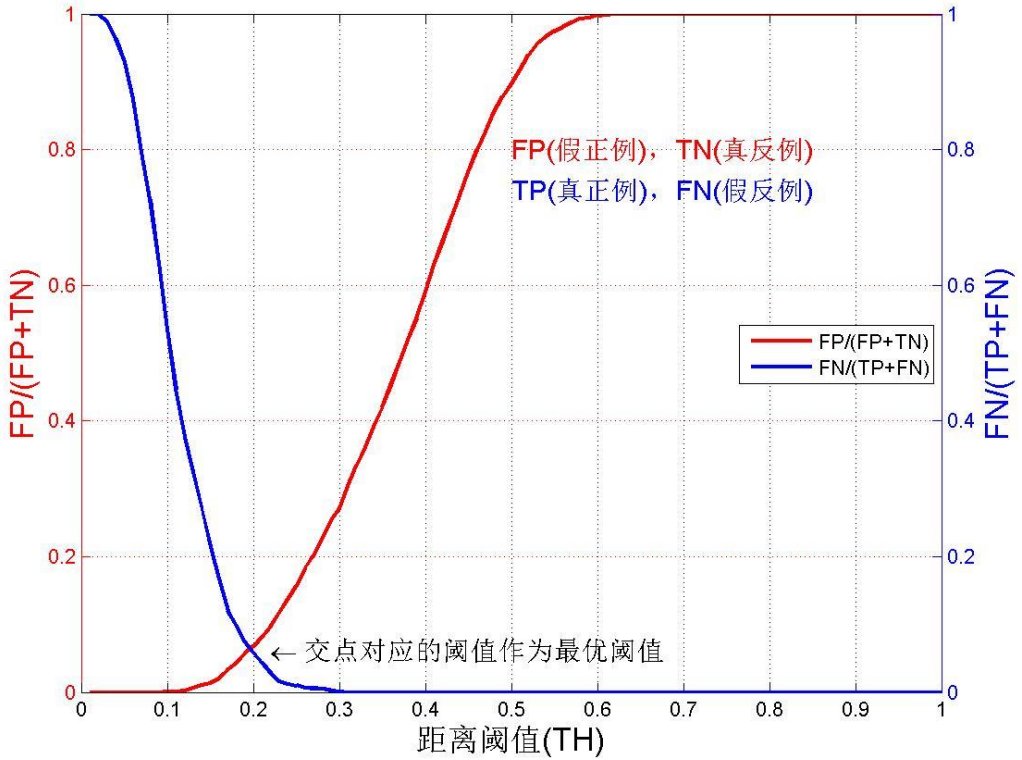
最优阈值选择



从曲线可以看出最优阈值应该选择为0.2

三、实验结果

最优阈值选择



红色曲线表示实际是不一致人脸对但是被判断为一致人脸对的个数占测试样本集中实际是不一致人脸对的总个数的百分比。

蓝色曲线表示实际是一致人脸对但是被判断为非一致人脸对的个数占测试样本集中实际是一致人脸对的总个数的百分比。

当然，两个百分比都需要尽量小，因此0.2是合适的，也符合前面给出的阈值结果。

三、实验结果

阈值对比

阈值	TP（真正例）	FN（假反例）	FP（假正例）	TN（真反例）	错误率
0.16	1514	306	46	2194	8.67%
0.18	1641	179	96	2144	6.77%
0.20	1715	105	152	2088	6.33%
0.22	1768	52	216	2024	6.60%

假反例指的是，实际是正例但是被误判为反例。假正例指的是，实际是反例但是被误判为正例

三、实验结果

运行时间

我们的模型运行在Matlab 2012a 环境上，模型共迭代了205次，共花费时间18.3小时，每次迭代花费平均时间是321.3秒。

测试样本数为4060，测试共花费时间7.5秒，每对样本测试时间平均值为7.5/4060=1.85毫秒。可以看出我们的算法在验证阶段运行速度是非常快的，这得益于我们的模型简单，可调参数少。由于模型简单，可调参数少，我们的模型非常节省内存，能在离线状态下独立运行，同时速度非常快。因此，我们的模型非常适合运行在低功耗嵌入式智能设备上。

三、实验结果

实际是一个人，判断结果也是一个人

光照变化
欧氏距离：0.0946

表情变化
欧氏距离：0.0977

表情不同
欧氏距离：0.0898

表情变化
欧氏距离：0.0982

光照变化
欧氏距离：0.0832

表情不同
欧氏距离：0.1080

三、实验结果

实际是一个人，判断结果也是一个人

光照变化
欧氏距离：0.0862

表情变化
欧氏距离：0.1151

表情不同
欧氏距离：0.0877

光照变化
欧氏距离：0.1675

表情变化
欧氏距离：0.1421

表情和光照不同
欧氏距离：0.1172

三、实验结果

实际不是一个人，判断结果也不是一个人

欧氏距离：0.4410

欧氏距离：0.2871

欧氏距离：0.3081

欧氏距离：0.2889

欧氏距离：0.3879

欧氏距离：0.3252

三、实验结果

优点与改进

从上面的实验结果可以看出，我们的模型性能非常优越，在测试数据集上达到了6.33%的错误率。同时在验证阶段算法的运行效率也是非常高的。下面总结出我们的模型的优点和需要改进之处。

优点

1、结构简单，可调参数少，需要训练数据集少，占用内存少。
2、验证速度快，功耗低，能在离线状态下在本地快速完成计算。
3、模型不仅仅适用于人脸认证任务，计算机视觉中的很多任务都适用。
4、验证精度高，能处理常见情形，如光照变化，表情变化等。

改进

1、在保证验证速度的前提下，增加模型的复杂度，提高输入图像的分辨率，提高特征学习的泛化能力。
2、增加训练数据量，增加模型深度，提高模型的数据表示能力。
3、使用成熟的开源深度学习工具，结合GPU加快模型的训练。
4、使用C++替换matlab。
5、不断降低测试错误率。

结论和展望

我们使用卷积网络模型学习图像区别的特征表示，训练数据集被分为一致图像对和非一致图像对，通过优化损失函数减小一致图像对的特征距离，同时增大非一致图像对的特征距离。通过人脸认证任务验证了我们的模型的有效性。在人脸图像存在较大光照变换，局部表情变化的情况下，我们的模型表现出了出色的性能。由于我们的模型可调参数少，结构简单，非常适合运行在移动设备上，如智能手机，儿童玩具等等。但是我们的模型不仅仅局限于人脸认证任务，计算机视觉中很多任务我们的模型都可以胜任，如图像搜索，图像匹配等。

同时，由于我们的模型使用低分辨率图像，造成了图像细节丢失，区别学习能力下降。未来，我们将不断改进我们的网络，采用更加先进和复杂的模型，同时增加训练数据量。