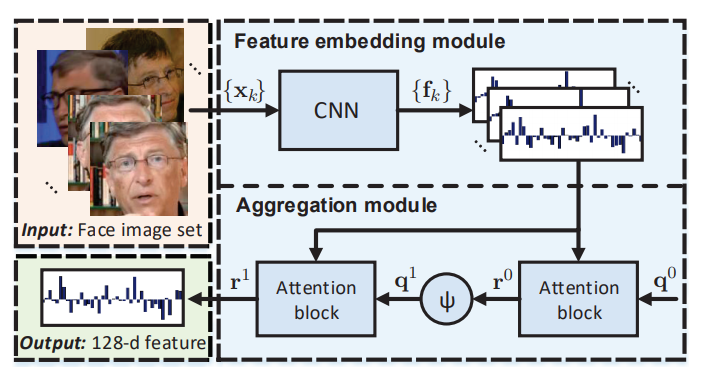
# 视频人脸识别神经聚合网络

## 摘要

文章介绍了视频人脸识别神经聚合网络。这个网络使用一个人的几张照片或者视频作为输入，产生一个紧凑的、固定维度的代表这次识别的特征。整个网络是由两个模块组成。嵌入模块的特征是一个深层次的将每一个脸的图片转化成特征矢量的卷积神经网络。集成模块由两个可以在卷积神经网络中自适应集成特征矢量来产生一个特征的注意块。因为这个注意机制，集合是与图像顺序无关的。我们用没有任何作用信号的标准分类和验证损失来训练视频人脸识别神经聚合网络，并且我们发现网络自动的趋向使用高质量的照片，避免那些模糊的、遮挡的照片。根据在IJB-A、YouTube Face、Celebrity-1000上的人脸识别实验基准，视频人脸识别神经聚合网络表现优于简单的聚集方法并且达到了最准确的精度。

## 介绍

视频人脸识别最近已经越来越火了。相比较照片人脸识别，更多的信息能够从视频中提取出来，因为视频自然的合并了同一个个体的不同动作和照明条件。视频人脸识别的关键是构造一个视频中人脸的合适表现，这样它能有效地把不同帧整合在一起，获取有效的信息同时丢掉干扰信息。



图片1. 我们视频人脸识别的网络架构。所有输入的人脸照片 { Xk } 都被一个带有卷积神经网络的特征嵌入模块处理，产生一个特征矢量 { fk }。这些特征被传到集成模块，产生一个128维度的向量r1代表着输入的图像。产生的集合用来识别。

一个简单方法就是把视频中的人脸组合成一个像从深层神经网络提取的以帧为单位的人脸特征，这是最近大多数神经网络使用的方法。这样的集合综合的获得所有帧的信息。不管怎样，比较两个视频面部识别，一个需要融合所有帧中两个面部中匹配的结果。设n为视频帧数的平均值，一次操作的计算复杂度为O(n2)，这对于大规模的识别是不可取的。此外，这样基于集合的表示将回是O(n)空间复杂度，这要耗费大量的储存和需要有效的索引。

我们觉得需要一个视频层面压缩的，固定大小的特征识别，不管这个视频多大。这样的能力能够允许不需要一帧一帧的去匹配的，而是直接的，固定时间计算相似度和距离。一个直接的解决方法是在每一帧都提取一个特征然后进行某种类型的帧级功能聚合到一起形成视频级别的表现。

最常用的集中策略是平均和最大值计中。这些简单的集中策略在之前的工作中被证实是有校的，我们觉得一个好的集中和聚集策略应该可以自适应的衡量和组合所有帧之间的帧级别的特征。灵感很简单，一个视频（尤其是长的）或者图像集可能包含各种光照、分辨率、姿势的人脸图片，所以一个好的算法应该偏向那些更易判读或者记忆的照片和避免那些质量不高的照片来进行人脸识别。

为此，我们寻找一种自适应加权方案，将视频中的所有帧级功能线性组合在一起，形成一个紧凑而有区别的面部表情。与以前的方法不同，我们不修复权重也不依赖任何特定的启发式方法来设置它们。相反，我们设计了一个神经网络来自适应计算砝码。我们把我们的网络命名为神经聚合网络，其系数可以训练通过正常人脸识别中的监督学习培训任务不需要额外的监督信号。

我们的神经聚合网络由两个可以端到端和可以单独的训练的主要模块组成。第一个是一个功能嵌入模块，作为一个使用深度卷积神经网络模型的帧级特征抽取器。另一个是自适应融合的聚合模块所有视频帧的特征矢量。

我们的神经聚合网络是有聚集技术的主要优点，能够处理任意输入大小并产生顺序不变量。这个网络的关键组件灵感来自神经图灵机工作，两者都应用了注意机制通过访问外部存储器来组织输入。这种机制可以接受任意大小和工作的输入调整每个输入元素通过加权平均，非常重要的是它是有序的独立，可训练的参数。在这项工作中，我们设计了一个简单的两级注意网络结构。与此关注机制关联的块人脸特征集合。

除了建立一个视频级的表现外，神经聚合网络也可以作为一个学科层次功能提取器，用于融合多个数据源。举个例子，你可以给它提供所有可用的图像和视频，或者多个同一物体的视频集成视频，以获得固定尺寸的单一特征。这样，人脸识别系统因为集合不只会有时间和内存上的效率，也表现出卓越的表现，我们将在实验中展示。

我们评估了神经网络的视频人脸验证和识别任务。与基线和其他竞争性的方法相比，我们观察到三个具有挑战性的数据集（包括YouTube人脸数据集、IJB-A数据集和Celebrity-1000数据集）。

最后，我们要指出，我们提出的神经聚合网络可以作为学习内容自适应池的通用框架。因此，它也可以作为其他计算机视觉任务的集成方案。

### 相关工作

基于视频或图像集的人脸识别一直是人们研究的热点。本文研究的输入是一组无次序的人脸图像。这里不考虑利用时间动力学的现有方法。对于基于集合的人脸识别，许多以前的方法都试图用外观子空间或流形来表示人脸图像集合，并通过计算流形相似性或距离来进行识别，这些传统的方法在受限的设置下可以很好地工作，但通常无法处理具有挑战性的、不受约束的场景。存在较大的外观变化。

在不同的轴上，有些方法基于局部特征构建视频特征表示。例如，PEP方法通过提取和聚类局部特征来进行基于局部的表示。视频Fisher矢量面（VF2）描述符使用Fisher矢量编码将不同视频帧的本地特征聚合在一起，形成视频级表示。

近年来，最先进的人脸识别方法已被深度卷积神经网络所控制，对于视频人脸识别，大多数方法要么采用成对帧特征相似性计算，要么采用辅助（平均/最大）帧特征池。这促使我们寻求一种适应性集成方法。

如前所述，这项工作也与神经图灵机和其工作有关。然而，值得注意的是，虽然它们使用循环神经网络（RNN）来处理顺序输入/输出，但是我们的方法中没有RNN结构。我们只借用它们的可区分内存寻址/注意方案来进行特性聚合。

## 神经集成网络

如图1所示，NAN网络以一组人脸图像作为输入，输出一个特征向量作为识别任务的表示。它建立在一个现代的深度CNN模型上，用于帧特征嵌入，通过自适应地将视频中的所有帧聚合成一个紧凑的矢量表示，它让视频人脸识别变得更加强大。

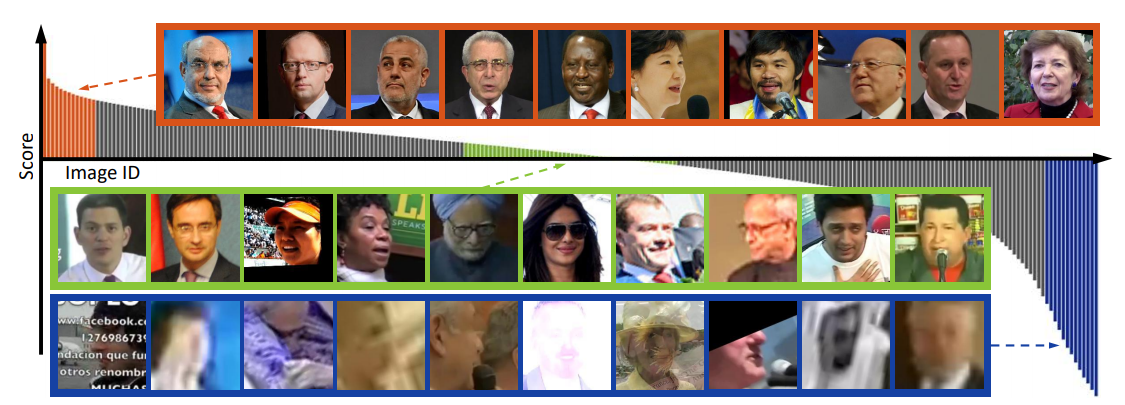


图2。IJB-A数据集中的人脸图像，根据在人脸识别任务中训练的单个注意块的得分（等式2中的e值）进行排序。从得分最高的5%和10%的窗口中对顶行、中行和底行中的面进行采样。分别以中位数和最低5%为中心。

### 特征嵌入模块

我们的NAN图像嵌入模块是一个深度卷积神经网络（CNN），它将视频的每一帧嵌入到人脸特征表示中。为了利用具有高端性能的现代深层CNN网络，本文采用了批标准化（BN）技术的GoogleNet。当然，其他网络架构也同样适用于这里。GoogleNet生成128维图像特征，这些特征首先被标准化为单位向量，然后送入聚合模块。在本文的其余部分中，我们将简单地将使用的谷歌（Googlenet）BN网络称为CNN。

### 集成模块

在n对视频人脸数据上考虑视频人脸识别任务(Xi，yi)in=1，Xi是人脸视频序列或者有Ki张图片的图片集，i.e.Xi = { x1i, x1i, ... xKi }，XiK，k = 1, ... , Ki 是视频中的第K帧，yi是不重复的Xi的ID。每一帧Xik都有一个不重复的标准化的从特征嵌入模块提取的特征表现fik。为了更好的可读性，在剩余的文本中适当的地方，我们省略了上面的索引。我们的目标是利用视频中的所有特征向量生成一组线性权重 { ak }kK=1，所以集成特征表达式是：

r = akfk

这样，聚集特征向量的大小与CNN提取的单个人脸图像特征的大小相同。

显然公式1的关键是它的权重 {ak}。如果 ak = 1/K，公式1会变成简单的平均数，这并不是我们想要在我们实验中表现的。

设计中我们的聚合模块考虑了三个主要原则。首先因为视频数据源因人而异，模块应该能够处理不同数量的图像（即不同的Ki）。第二，聚合应该与图像顺序保持不变，我们更喜欢在图像序列反转或重新混合时结果保持不变。通过这种方式，聚合模块可以处理任意一组图像或视频，而不需要时间信息（例如，从不同互联网位置收集的信息）。第三，在标准人脸识别训练任务中，该模块应能适应输入人脸，并通过监督学习来训练参数。

我们的解决方案受到了[12，32，38]中描述的记忆-注意力机制的启发。其中的想法是使用一个神经模型通过一个可区分的寻址/注意方案来读取外部储存。这些模型通常与循环神经网络（RNN）相结合，以处理顺序输入/输出。虽然我们不需要RNN结构，但是它的内存注意机制适用于我们的聚合任务。在本研究中，我们将人脸特征视为记忆，将特征加权视为记忆寻址过程。我们在聚合模块中使用“注意块”，描述如下。

#### 注意块

注意块从特征嵌入模块中读取所有特征向量，并为其生成线性权重，具体地说，让fk作为面特征向量，然后注意块通过点积用核q对其进行过滤，得到一组相应的意义ek。然后将它们传递给SoftMax运算符以生成正权重。这两种操作可以分别用以下方程式来描述：

ek=qTfk

ak=

可以看出，我们的算法本质上选择了所有特征向量所跨越的凸壳内部的一个点。一项相关工作是[3]，其中每个面图像集都近似于一个凸壳，集相似性定义为两个凸壳之间的最短路径。

这样，输入的数量 {fk} 不会影响聚合r的大小，聚合r与单个特征 {fk} 具有相同的维度。此外，聚集结果与 {fk} 的输入阶数是不变的：根据式1、2、3，排列 {fk} 和 {fk} 对聚集表示r没有影响，并且通过标准反向传播和梯度下降可训练的滤波核q调制注意块。

单注意块-通用面部特征质量测量。我们首先尝试使用一个注意块进行聚合。在这种情况下，矢量q是要学习的参数。它具有与单个特征F相同的尺寸，并作为测量面特征质量的通用先验工具。

我们训练网络对提取的人脸特征进行IJB-A数据集[19]中的视频人脸验证（详见第2.3节和第3节），图2显示了数据集中所有人脸图像的排序分数。可以看出，经过训练后，网络倾向于高质量的面部图像，例如高分辨率和相对简单的背景图像。它通过模糊、遮挡、不正确的曝光和极端的姿势来降低面部图像的权重。表1表明，在验证和识别任务中，网络比平均池基线实现更高的准确性。

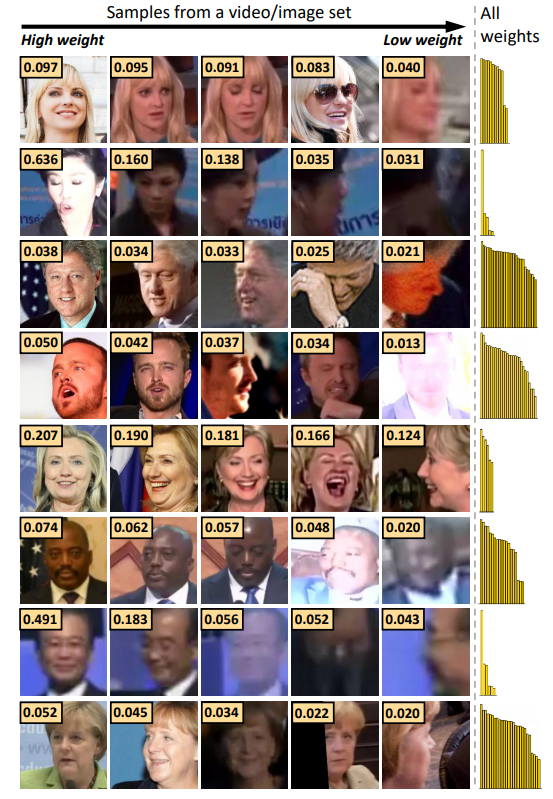


图3。显示由我们的NaN计算的图像集中的图像权重的典型示例。在每行中，从一个图像集中抽取五个面图像，并根据其权重（矩形中的数字）进行排序；最右侧的条形图显示集合中所有图像的排序权重（高度缩放）。

级联两个注意块-内容感知聚合。我们相信一个内容感知的聚合可以执行得更好。其背后的直觉是，人脸图像的变化可以在特征空间中的不同地理位置（对于不同的人）以不同的方式表达，而内容感知聚合可以学习选择对输入图像集的身份更具识别性的特征。为此，我们以如下所述的级联和端到端方式使用两个注意块。

让q0作为第一个注意块的核心，r0作为q0的聚合特性。我们通过一个以r0为输入的传输层自适应地计算第二个注意块的核心q0：

q1 = tanh(Wr0+b)

W和b是矩阵的权重和各神经元的偏压矢量，而且 tanh(x) = 施加双曲正切非线性。由q1生成的特征向量r1将是最终的聚合结果。因此，（q0，w，b）现在是聚合模块的可训练参数。

我们在IJB-A数据集上对网络进行了再次训练，表1表明，该网络比使用单一注意块获得了更好的效果。图3显示了训练网络为不同视频或图像集计算的权重的一些典型示例。

我们目前的NAN全部方法基于在得到所有剩余实验结果，采用了这种级联的双注意块设计（如图1所示）。

### 网络训练

NAN网络可以接受标准配置的人脸验证和识别任务的训练。

#### 训练损失

为了验证，我们构建了一个具有两个NaN共享权重的暹罗神经聚集网络结构[8]，并将平均对比损失最小化[14]：li.j = yi,j || r1i - r1j ||22 + (1-yi,j)max(0, m - ||r1r - r1r ||22)，yi,j = 1 当 (i,j) 相同， 否则yi,j=0。m常量在我们的实验里设置为2。

为了便于识别，我们在NAN的顶部添加了一个完整的连接层，然后是一个SoftMax，并将平均分类损失降到最低。

#### 模块训练

这两个模块可以以端到端的方式同时训练，也可以逐个单独训练。在这项工作中选择后一个选项。具体来说，我们首先用识别任务对CNN进行单个图像的训练，然后根据CNN提取的特征对聚合模块进行训练。更多详情见第3.1节。

我们选择这一单独的训练策略主要是针对两个原因。首先，在本文中，我们将重点分析具有关注机制的聚合模块的有效性和性能。尽管深度CNN在基于图像的人脸识别任务中取得了巨大的成功，但我们对CNN特征集合的了解却很少。第二，训练一个深入的CNN通常需要大量的标记数据。尽管现在可以获得数以百万计的静止图像用于训练[35、28、31]，但收集这些独特的面部视频或布景似乎并不现实。作为我们今后的工作，我们对NAN的训练是永无止境的。

## 实验

本节评估所提议的NAN网络的性能。我们将首先介绍我们的训练细节和基线方法，然后报告三个视频人脸识别数据集的结果：IARPA Janus Benchmark A（IJB-A）[19]、YouTube Face数据集[42]和Celebrity-1000数据集[23]。

### 训练细节

如第2.3节所述，本工作单独训练两个网络。为了训练CNN，我们使用从Internetto获取的大约3M来自5万个身份的面部图像进行基于图像的识别。使用JDA方法[5]，并与LBF方法[29]对齐。输入图像大小为224x224。训练结束后，CNN被固定，我们重点分析了神经聚集模块的有效性。

集成模块针对我们使用标准反向传播和anRMSProp解算器测试的每个视频面数据集进行训练[36]。使用全零参数初始化，即我们从平均池开始。针对每个数据集调整每批大小、学习速率和迭代。由于网络非常简单，图像功能也很紧凑（128-D），所以训练过程非常高效：在台式电脑的CPU上，总共1米图像的5千视频对训练只需要2分钟。

### 基线方法

由于我们的目标是压缩视频人脸，因此我们将结果与简单的聚合策略（如平均池）进行比较。我们还比较了一些集到集的相似性测量，利用图像级别上的成对比较。为了保持简单，我们简单地使用L2特征距离进行人脸识别（所有特征都是标准化的），尽管可以与外部度量学习或模板自适应技术[10]结合起来进一步提高每个数据集的性能。

在基线方法中，CNN+Min L2、CNN+Max L2、CNN+Mean L2和CNN+SoftMin L2根据所有帧对的L2特征距离测量两个视频人脸的相似性。它们需要存储视频的所有图像特征，即具有O（N）空间复杂性。前三种算法分别使用最小值、最大值和平均值对距离，因此对于相似计算具有O（n2）复杂性。CNN+SoftMin L2对应于[24，25，1]等作品中提倡的最软相似性得分。它的计算复杂度为O（m\*n2）。CNN+MaxPool和CNN+AvePool分别是最大池和平均池，沿每个特征维度进行集成。这两种方法以及我们的NAN为每个视频生成128-D特征表示，并计算O（1）时间内的相似性。

### IJB-A数据集结果

IJB-A数据集[19]包含从无约束环境中捕获的人脸图像和视频。它具有全姿态变化和广泛的成像条件变化，因此是非常具有挑战性的。共有500名受试者，共有5397张图片和2042个视频，平均每个受试者有11.4张图片和4.2个视频。我们使用STN[4]人脸检测器检测带有标志的人脸，然后将人脸图像与相似变换对齐。

在这个数据集中，每个训练和测试实例都被称为模板，它由1到190个混合静态图像和视频帧组成。由于一个模板可能包含多个媒体，并且数据集为每个图像提供媒体ID，另一种可能的聚合策略是首先聚合每个媒体中的帧功能，然后聚合模板中的媒体功能[10，30]。该策略也在CNN+Avepool和我们的NAN的工作中进行了测试。请注意，媒体ID在实践中可能并不总是可用的。

我们在1:1人脸验证的比较协议和1:n人脸识别的搜索协议上测试了建议的方法。为了进行验证，报告了真实接受率（TAR）与假阳性率（FAR）。对于鉴定，报告了真阳性鉴定率（TPIR）、假阳性鉴定率（FPIR）和Rank-n精度。表2给出了不同方法的数值结果，图4给出了用于验证的接收机工作特性（ROC）曲线，以及用于识别的累积匹配特性（CMC）和决策误差权衡（DET）曲线。根据10次分割的[19，13]计算指标。

一般来说，CNN+MAXL2、CNN+MINL2和CNN+MAXPOOL在基线方法中表现最差。CNN+SoftMinl2的表现略好于CNN+MaxPool。媒体ID的使用显著提高了CNN+Avepool的性能，但对NAN的提升相对较小。我们相信NAN已经对由一些媒体的劣质图像主导的模板具有了强大的鲁棒性。如果没有媒体聚合，NAN的利润率将远远超过其所有基线，特别是在低远距情况下。例如，在验证任务中，我们的NAN的fars值分别为0.001和0.01，分别为0.860和0.933，将最佳结果的基线误差分别减少约39%和23%。

据我们所知，通过媒体聚合，我们的NAN实现了与以前方法相比的最高性能。它在far=0.1时具有相同的验证tar，标识等级为10 cmc，与最先进的方法[10]相同，但在所有其他指标上都优于它（例如，在far=0.01时为0.881:0.836 tars，在fpir=0.01时为0.817:0.774 tpirs，在fpir=0.01时为0.958:0.928 rank-1精度）。

图3显示了一些加权结果的典型示例。NAN能够选择高质量和更具辨别力的面部图像，同时排斥较差的面部图像。

### Youtube Face数据集结果

然后我们在YouTube人脸（YTF）数据集[42]上测试我们的方法，该数据集是为视频中的无约束人脸验证而设计的。它包含3425个视频，1595个不同的人，视频长度从48到6070帧不等，平均长度为181.3帧。我们提供10对500视频对，并按照标准验证协议报告交叉验证的平均精度。我们再次使用STN和相似性转换来对齐人脸图像。

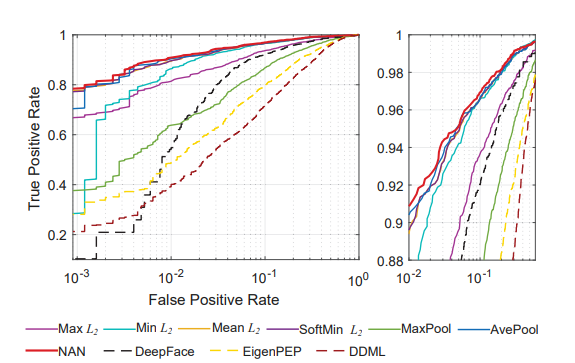


图5。不同方法和我们的NAN在YTF数据集上的平均ROC曲线超过10次分割。

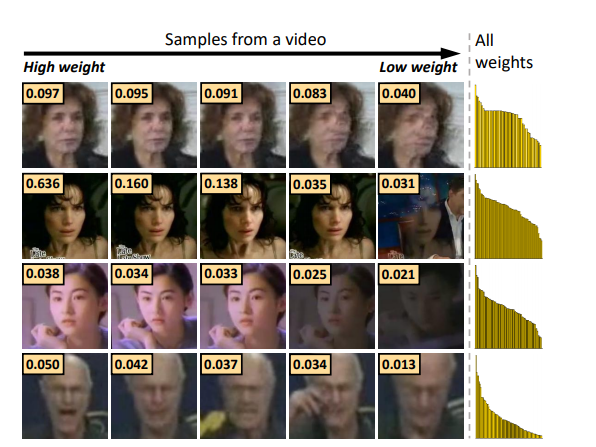


图6。YTF数据集上的典型示例显示了由NAN计算的视频帧的权重。在每一行中，从视频中抽取五帧，并根据其权重（矩形中的数字）进行排序；最右侧的条形图显示所有帧的排序权重（按比例缩放的高度）。

我们的NAN、其基线和其他方法的结果如表3所示，其ROC曲线如图5所示。可以看出，NAN再次超越了其所有基线。与IJB-A上的结果相比，NAN和最佳执行基线之间的差距较小。这是因为此数据集中的面变化相对较小（与图6和图3中的示例进行比较），因此与原始平均池或计算平均l2距离相比，无法提取更多有益信息。

与以前的方法相比，我们的NAN平均精度达到95.72%，减少了12.3%的FaceNet误差。请注意，FaceNet也基于GoogleNet风格的网络，并且使用了每个视频中100帧的所有对的平均相似性（即10000对）。据我们所知，只有vgg面[28]的精度（97.3%）高于我们的精度。然而，这一结果是基于YTF的进一步区分性度量学习，没有它，精度仅为91.5%[28]。

### Celebrity-100数据集结果

Celebrity-100数据集[23]旨在研究基于视频的无约束人脸识别问题。它包含159726个1000人的视频序列，总共2.4米的帧（每个序列15帧）。我们使用提供的5个面部标志来对齐面部图像。此数据集上存在两种类型的协议c打开集和关闭集。有关协议和数据集的更多详细信息，请参见[23]。

关闭设置测试。对于封闭集协议，我们首先在识别丢失的视频序列上训练网络。我们以fc层的输出值为得分，以得分最大的主题为结果。我们还为CNN+Avepool培训了一个线性分类器，对每个视频特征进行分类。由于这些特性是建立在视频序列上的，因此我们称之为“videoaggr”，以将其与下一步要描述的另一种方法区分开来。数据集中的每个主题都有多个视频序列，因此我们可以通过聚合所有训练（库）视频序列中的所有可用图像来为主题构建一个单一的表示。我们称这种方法为SubjectAggr。通过这种方法，可以绕过线性分类器，通过比较特征二级距离即可实现识别。

结果如表4所示。注意[23]和[22]没有使用深度学习，并且没有基于深度网络的方法在此数据集上报告结果。所以我们主要在下面比较基线。从表4和图7（a）可以看出，对于videoaggr和subjectaggr而言，NAN始终优于基线方法。subjectaggr方法在基线上取得了显著改善。有趣的是，subjectaggr导致CNN+Avepool的性能比其videoaggr明显下降。这表明，当在包含多个视频的主题级别上应用幼稚聚合时，这种聚合会变得更糟。然而，我们的NAN可以受益于subjectaggr，产生的结果始终优于或等于videoaggr方法，并提供了相当大的精度提高比基线。这表明我们的NAN在处理大数据变化方面非常有效。  
 开放式测试。然后我们用封闭集协议测试NAN。我们首先在提供的训练视频序列上训练网络。在测试阶段，我们采用前面描述的SubjectAggr方法为每个画廊主题构建高度紧凑的人脸表示。只需比较聚合面表示之间的二级距离即可进行标识。

表5和图7（b）中的结果表明，我们的NAN显著降低了基线CNN+Avepool的误差。这再次表明，在存在较大的面方差的情况下，广泛使用的策略（如平均池聚合和成对距离计算）远不是最佳的。在这种情况下，我们学习的NAN模型显然更强大，它的聚集特征表示更适合于视频人脸识别任务。

## 结论

我们提出了一种用于视频人脸表示和识别的神经聚合网络。它将所有输入帧与一组内容自适应权重融合在一起，从而形成与输入帧顺序不变的紧凑表示。聚合方案简单，计算量小，内存占用小，但经过训练后可以生成高质量的人脸表示。建议的NAN可用于一般视频或集合表示，我们计划将其应用于未来工作中的其他视觉任务。