基于深度学习的人脸视频检索系统——终期报告

景宸琛：2120161005 李丝雨：2120161010
王雨佳：2120161061 张力嘉：2120161077

摘要

给出一个人物的人脸视频，人脸视频检索的目标是从视频数据库中寻找包含该人物的视频。人脸视频检索的挑战性问题主要在于人脸类内差异较大以及巨大的时间和空间复杂度。为了解决这些问题，我们通过深度卷积神经网络（deep CNN）学习出具有判别性且紧致的人脸表达来实现人脸视频检索。网络包含特征提取与哈希学习两个模块，特征提取模块用于对视频的每一帧提取判别性特征，哈希学习用于将高维特征映射到低维海明空间以适用于检索任务。网络将两个模块融入统一的优化框架以获得最相容的特征提取器与哈希函数。

1. 背景

随着多媒体技术及计算机网络技术的迅速发展，多媒体已经广泛地应用于多个领域，如公共信息业、教育、商业及娱乐等等。人们可以在互联网上轻而易举的获取海量的视频数据。然而，有了相关视频不等于就找到了目标信息，查找视频、分析视频的工作常常会耗用大量的时间和人力。如何在海量视频中更方便、更省力地查找到相关信息呢?

随着用户需求越来越强烈，视频检索技术也得以快速发展。检索技术源于互联网发展需求。各类搜索引擎，如 Baidu、Google、Bing 以及 Yahoo 等都是以此技术为基础的。随着网络带宽不断的提高，人们可以更加快捷地将自己采集到的各种多媒体信息进行共享，或者进行多媒体信息的交互，越来越多的信息通过视频等多媒体的形式展现在互联网中，这对以图像、视频为代表的多媒体信息检索技术提出了越来越高的要求。

之前比较通用的方法是采用文本注释图像和视频信息，以基于文本的数据库管理系统进行图像和视频检索。但文本注释方法对大量的信息不仅费力而

且力不从心，对于在存储的视频节目中寻找指定的视频片断这样的应用需求，比如特定节目内容的搜索、定位就更加困难，基本只能靠人工的观看、识别和记录。因此，多媒体信息检索技术的发展意义重大，已经吸引了国内外研究者的广泛注意。

在多媒体信息检索技术中，人脸视频检索是其中一项有着广阔的应用前景的技术。人脸视频检索指的是，给定一个人的视频，在人脸视频数据库中检索包含该人物的所有视频。随着可获取的视频数据越来越多，人脸视频检索技术也愈发重要。例如，在嫌疑人追踪场景下，我们可以使用通过监控摄像头拍下的嫌疑人的视频，在海量的监控视频中检索该嫌疑犯的所有视频，从而实现嫌疑犯的快速追踪和定位。

人脸视频检索的难点在于，人脸类内差异较大以及巨大的时间和空间复杂度。由于姿势、光照、表情、衣着、背景以及人物朝向等等因素，数据集中的人脸视频的类内差异非常大。这表明人脸的表达一定要对类内的差异具有鲁棒性并在类别间具有判别力。除此之外，为了提高检索速度以及减少空间占用，人脸表达必须尽量紧凑。现存的基于视频的人脸识别算法用成千上万甚至更多个浮点数表示人脸特征，具有极大的时间和空间复杂度，以至于不适用于人脸视频检索的任务。

为了解决这些问题，本实验通过深度卷积神经网络（deep CNN）学习出具有判别性且紧致的人脸表达来实现人脸视频检索。

2. 目的

人脸视频检索是一个具有大量应用且极具吸引力的研究领域。但是人脸类内差异较大以及巨大的时间和空间复杂度一直困扰着研究者们。本实验旨在通过深度卷积神经网络（deep CNN）学习出具有判别性且紧致的人脸表达，并学习哈希函数将高维特征投影到低维海明空间，并建立一个人脸视频检索系统。

3. 环境

Ubuntu 14.04.5 LTS

Intel(R) Core(TM) i7-4930K CPU @ 3.40GHz

32.00 GB 内存

深度学习框架：caffe

软件：MATLAB R2014a

4. 方法

为了实现人脸视频检索，本实验将哈希学习融入深度卷积神经网络建立了一个用来学习判别性且紧致的人脸表达的端到端的深度学习系统。图 4-1 展示了本实验的网络模型及整体的训练步骤。

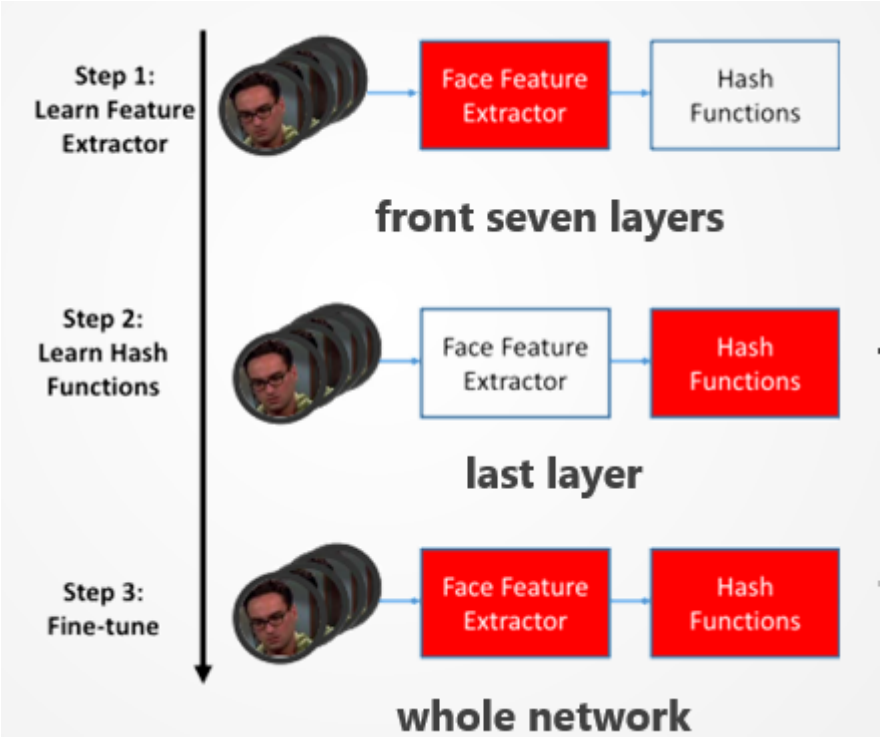

图 4-1 模型及训练步骤

4.1 学习特征提取器

ImageNet 数据集包含 1000 类别的 1,200,000 张图片。为了更好的初始化，本实验在 ImageNet 数据集上进行网络的预训练。为了方便，本实验使用公开的已在 ImageNet 数据集上训练好的 AlexNet 模型。AlexNet 包含卷积层、归一化层、线性层、ReLU 激活层和池化层。为了简单我们用L_{1-5}来代表

5 个卷积层，L_{6-8}来代表 3 个线性层。L_7输出 4096 维的特征，L_8特征的维数为 1000，L_8之后是一个 softmax 分类器以生成分类的概率分布。之前的研究表明使用L_7层的 4096 维特征好于使用手工特征。在模型中，L_{1-7}用于特征提取。

4.2 预训练哈希函数

通过预训练的 deep CNN，获得了丰富的中层特征。但是，此特征依然是高维的，因此时间和空间复杂度较高。为了生成用于检索的有效并且紧致的表达。进一步使用判别式二值码（DBC）方法来将神经网络生成的高维特征投影到更低维的二值空间。在二值空间，每位只能是 1 或-1，因此两个向量的距离可以通过每位的异或操作简单地计算，因此节省了大量的时间。除此之外，二值特征空间的低维特性也保证了较低的空间复杂度。

用 X 代表获取的 C 类的特征，维数为 t，样本总数为 n。优化目标是学习将 X 从特征空间投影到二值空间的哈希函数 W。如公式（1），

$$\mathbf{b}_i = sgn(\mathbf{W}^\top \mathbf{x}_i), \forall \mathbf{x}_i \in \mathbf{X}, \tag{1}$$

其中 sgn()代表符号函数，正值返回 1，负值返回-1。b_i代表x_i的编码。生成的所有编码组成一个二值矩阵B$\epsilon R^{s\times n}$，其中 s 远小于 n。W$\epsilon R^{t\times s}$包含 s 个哈希函数。

在学习哈希函数时，DBC 考虑了两个约束条件：判别性和稳定性。首先，为了解决类内差异大的问题，哈希编码必须具有判别性，即同类样本应该具有相似的编码，异类样本能够被很好的分离开。其次，哈希编码的稳定性也应该被考虑在内。每一个都可以被视为特征空间的一个超平面，与 SVM 相同，稳定性意味着超平面之间的距离要尽可能大。因此，最后的哈希函数学习模型如（2）式。

$$\min_{w,\xi,L,B} \frac{1}{2} \sum_{c\in\{1:C\}} \sum_{m,n\in c} d(B_m, B_n) + \gamma \sum_{s\in\{1:k\}} \|w^s\|^2$$
$$+\lambda_1 \cdot \sum_{\substack{i\in\{1:N\}\\ s\in\{1:k\}}} \xi_i^s - \frac{\lambda_2}{2} \sum_{\substack{c'\in\{1:C\}\\ p\in c'}} \sum_{\substack{c''\in\{1:C\}\\ c'\neq c'', q\in c''}} d(B_p, B_q) \tag{2}$$
$$s.t.\ \ l_i^s(w^{s'} x_i) \geq 1 - \xi_i^s \quad \forall i \in \{1:N\}, s \in \{1:k\}$$
$$b_i^s = (1 + sign(w^{s'} x_i))/2 \quad \forall i \in \{1:N\}, s \in \{1:k\}$$
$$\xi_i^s > 0 \quad \forall i \in \{1:N\}, s \in \{1:k\}$$

目标函数（2）是一个非凸函数，所以找到全局最小值是不实际的。幸运的是，局部最小值已经可以生成有效的哈希编码。实验中使用优化算法如下：

Algorithm 1 Optimization

Input: $X = [x_1, ...x_N]$ (x_i is low-level feature vector for i^{th} image).
Output: B ($B_i = [b_i^1, b_i^2, ..., b_i^k]$ is binary code for i^{th} image).
1: Initialize B by: $B \leftarrow$ Projection of X on first k components of $PCA(X)$
2: Binarize B: $B \leftarrow (1 + sign(B))/2$
3: **repeat**
4: Optimizing for B in $\min_B \frac{1}{2} \sum_{c\in\{1:C\}} \sum_{m,n\in c} d(B_m, B_n)$ $-\frac{\lambda_2}{2} \sum_{c'\in\{1:C\}p\in c'} \sum_{c''\in\{1:C\}c'\neq c'', q\in c''} d(B_p, B_q)$ (see supplementary materials for details)
5: $l_i^s \leftarrow (2b_i^s - 1)\ \forall i\ i \in \{1:N\}, \forall s\ s \in \{1:k\}$
6: Train k linear-SVMs to update $w^s\ \forall s,\ s \in \{1:k\}$ using L as training labels (l_i^s : label for i^{th} image when training s^{th} split)
7: $b_i^s \leftarrow (1 + sign(w^{s^T} x_i))/2\ \ \forall i, s\ i \in \{1:N\},\ s \in \{1:k\}$
8: **until** Convergence on optimization 1

4.3 微调

我们最终的微调过程将特征提取和哈希函数学习融入了一个统一的优化框架从而构建了一个端到端的系统。在学习系统中，特征提取与哈希过程是最相容的，也就是说，学习到的特征是为哈希来服务的，并且哈希编码的性能可以指导人脸特征的学习。这样，一对具有语义相似性的人脸将有相似的紧致哈希表达，不同类的人脸表达之间会有很大的距离。为了实现这个目的，本实验使用可以反映人脸检索任务中目标的三元组排序损失（triplet ranking loss）。

假定深度卷积神经网络描述复杂的非线性映射 g：I→$\{+1, -1\}^s$，其中 I

代表人脸图像空间，所以一个人脸图像 q 可以表达为一个 s 位的二值表达 g(q)。三元组排序损失反应的是样本之间的相对相似度，即“与 $\hat{q}$ 相比，人脸 q 与 $\tilde{q}$ 更相似”。由式（3）定义：

$$\begin{aligned} &l\left(g(\mathbf{q}), g(\tilde{\mathbf{q}}), g(\hat{\mathbf{q}})\right) = \\ &\max\left(d\left(g(\mathbf{q}), g(\tilde{\mathbf{q}})\right) - d\left(g(\mathbf{q}), g(\hat{\mathbf{q}})\right) + \delta, 0\right), \end{aligned} \tag{3}$$

其中，δ 是一个控制距离的参数。用 Q 代表获取的 C 类的特征，微调阶段的目标函数如式（4）：

$$\max_{\mathbf{w}^*, \mathbf{w}} \sum_{i=1}^{C} \sum_{\substack{\mathbf{q}, \tilde{\mathbf{q}} \in \mathbf{Q}_i \\ \mathbf{q} \neq \tilde{\mathbf{q}}}} \sum_{j \neq i, \hat{\mathbf{q}} \in \mathbf{Q}_j} l\left(g(\mathbf{q}), g(\tilde{\mathbf{q}}), g(\hat{\mathbf{q}})\right), \tag{4}$$

其中，W 和 W^* 分别代表最后一层和之前层的参数。本实验使用反向传播算法对网络进行微调。

5. 数据集介绍

本实验使用 ICT-TV 数据集来评价使用的方法。ICT-TV 数据集包含两个源于美剧（生活大爆炸和越狱）的大规模视频集。这两部美剧的拍摄风格非常不同。生活大爆炸是一部仅有 5 名主角的情景喜剧，大多数情景在室内，每集大概 20 分钟。与之不同，越狱的大多数场景在室外拍摄，每集大概 42 分钟，包含很大的光照变化。所有的视频都来自两部美剧的第一季：生活大爆炸 17 集，越狱 22 集。人脸视频的数目分别是 4,667 和 9,435。视频的每帧都是 150×150 的图像。图 5-1 和图 5-2 分别是两个视频集中部分视频的示意图。

图 5-1 BBT 数据集

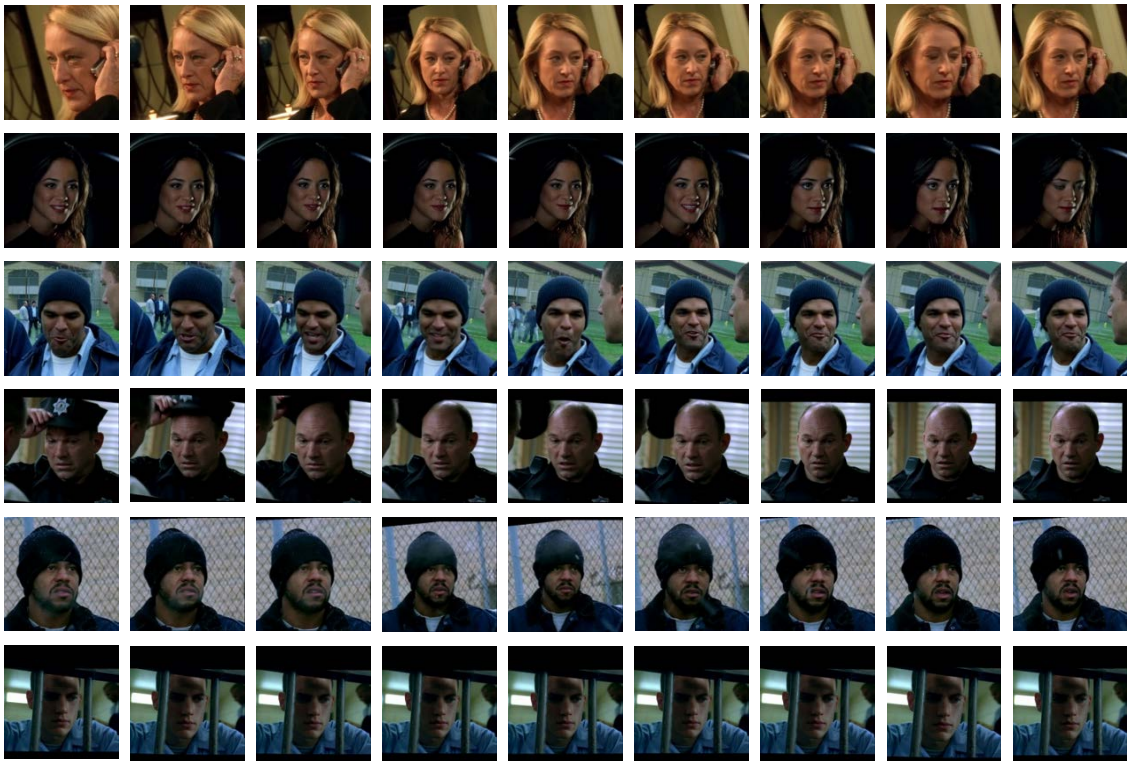

图 5-2 PB 数据集

6. 实验结果

本实验将使用的方法与三种哈希方法进行比较:LSH(局部敏感哈希), ITQ（迭代量化），KSH(核监督哈希)。对于每一个美剧数据集，随机选取每位演员的 10 个人脸视频来训练哈希函数，剩下的用来测试。测试时随机选取每个主演的 10 个视频作为查询集，其余的人脸视频作为数据库集。为了评价网络的表达能力，使用了两种评价准则：MAP 和 PR 曲线，所有方法使用相同的训练集和测试集。

在实验中，网络的输入为人脸图像，输出为二值哈希表达，并且每位都可以视为人脸的一个视觉属性。将人脸视频看做一系列人脸图像的集合。给定一个人脸视频，每一帧都被输入进 deep CNN 并得出一个二值表达。所有的二值表达通过硬投票混合，然后可以得到一个用于检索的视频二值表达。

最终的结果如表 6-1 和图 6-1 所示。

数据集	方法	8 位	16 位	32 位	64 位	128 位	256 位
BBT	LSH	0.4302	0.5301	0.6874	0.7486	0.8541	0.8761
	ITQ	0.8419	0.9019	0.8889	0.9130	0.9252	0.9345
	KSH	0.8338	0.9116	0.9388	0.9441	0.9430	0.9435
	Ours	0.9430	0.9525	0.9445	0.9628	0.9563	0.9625
PB	LSH	0.1308	0.1299	0.1906	0.2672	0.3487	0.4264
	ITQ	0.3571	0.4450	0.5074	0.5337	0.5370	0.5332
	KSH	0.5028	0.6155	0.6313	0.7041	0.7227	0.7456
	Ours	0.4873	0.5320	0.5869	0.5988	0.6184	0.6438

表 6-1 不同编码长度时两个数据集上的 MAP

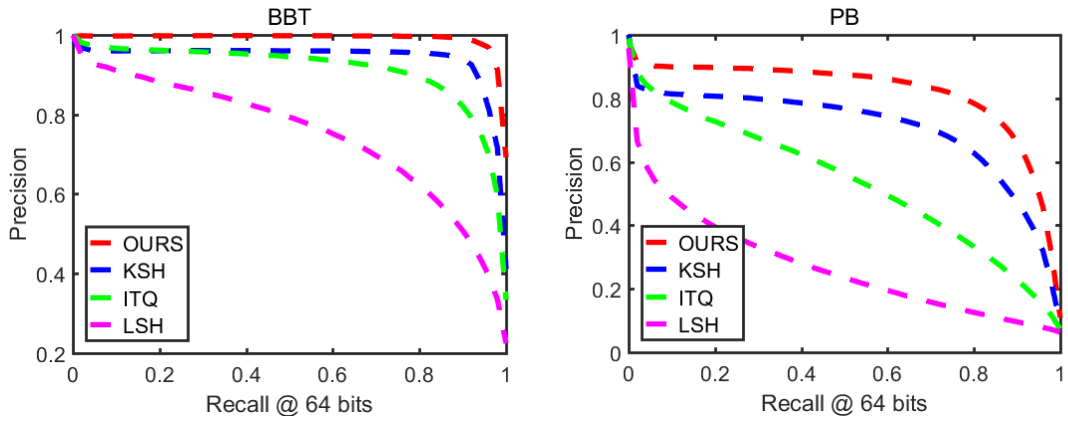

图 6-1 哈希码长为 64 位时两个数据集的 PR 曲线

实验结果显示，我们的方法将提取特征和哈希学习融入一个统一的网络并取得了更好的性能。

7. 结论

本实验使用深度卷积神经网络建立了一个人脸视频检索系统。此网络将特征提取和哈希学习融入了一个统一的优化框架以保证特征提取器可以最好的兼容于接下来的哈希学习。为了更好的初始化网络，特征提取器通过在 ImageNet 数据集上预训练的 AlexNet 的前七层来初始化，另外，我们使用判别二进制编码方法对哈希函数进行预训练，同时，初始化之后，网络被微调以获得更好的性能。