上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

学士学位论文

BACHELOR'S THESIS

论文题目：图像生成模型与生成式学习

学生姓名：徐亦飞

学生学号：5130309056

专　　业：致远计算机科学班

指导教师：张丽清

学院(系)：致远学院

图像生成模型与生成式学习

摘要

近年来，计算机视觉领域发展迅猛，特别是以卷积神经网络为首的深度学习算法大行其道，完成了一个又一个视觉难题。然而大部分算法都是判别式算法，需要极大的数据量和极其复杂的数据标注。而生成式模型可以规避这些问题。

本文主要研究的是 FRAME 随机场模型。FRAME 随机场模型是一个生成式概率模型，FRAME 是 Filter 过滤器，Random Field 随机场和 Maximum Entropy 最大熵的缩写。模型将图片视为一个马尔科夫随机场，使用了 Gabor 过滤器作为底层表达，并使用类似最大熵原理的方法进行训练。

在对原版 FRAME 随机场模型，其非均一化，稀疏化的拓展进行探究后。本文提出了稀疏 FRAME 随机场模型的层次化拓展，将生成学习与层次化表达学习结合起来，大大加强了模型的适用范围和鲁棒性能。模型将不但允许 FRAME 随机场模型中的 Gabor 过滤器进行位置，大小和角度的小幅度扰动，也将允许部件级别（即 Gabor 过滤器的集合）的扰动。

经过三个精密设计的实验，包括物件检测关键点预测，聚类问题和分类问题，通过不同 FRAME 随机场模型间的纵向对比和其他生成式模型间的横向对比，充分展示了层次化稀疏 FRAME 随机场模型在学习可解释性的模型上的优势。

关键词：机器学习，生成式模型，层次化

Generative Learning on Computer Vision

Abstract

Nowadays, there is a rapid growth in the area of computer vision, especially the deep learning algorithms including Convolution Neural Network. They are capable to achieve dozens of problems in vision. However, most of them are discriminative models, which need a large amount of labeled data. On the contrary, the generative models do not need big data.

This paper will focus on FRAME model. FRAME model, consists of Filter, Random Field and Maximum Entropy, is a generative probabilistic model. The model considers the image as a Markov random field, with the Gabor filter as the under layer expression. It can be learned by a method similar to the maximum entropy principle.

To improve original FRAME model and its extension on inhomogeneous and sparsification, we propose a hierarchical extension of FRAME model, which combines the generative learning and hierarchical representation. It will highly increase the robustness and application scope. The new model will not only allow to shift the locations, orientations and scales of the Gabor filters, but allow to shift the whole part (which is a set of filters) as well.

Three experiments, including object detection, clustering and classification, along with the comparison between different versions of FRAME model and the comparison with different generative and hierarchical model, show that our proposed model is capable of learning meaning and interpretable templates.

Key words: Machine Learning, Generative Model, Hierarchical Representation

目录

第一章　绪论

1.1 研究目的与意义

二十一世纪是大数据的时代。而视觉数据，包括图片和视频数据，在全球互联网流量中的占比超过了七成。人们对处理分析此类数据有极大的需求。这也是近年来计算机视觉领域十分火爆的原因。随着神经网络的发展，计算机目前可以完成大量计算机视觉领域的工作，通过大量样本的监督学习，深度神经网络已经可以在人脸识别，物件检测，分类，检索等特定问题上全方位达到或超越人类的水平。例如目前已经广泛应用在手机，网站登录，甚至银行系统操作的人脸识别就是计算机视觉领域最成功的的应用之一。

然而，目前的神经网络过度依赖于训练样本的质量与数量，极大的限制了网络的应用能力和拓展能力。在样本质量上，通常要求大量特定类型通过极其复杂的人工标注。例如，在人脸检测的应用中，不仅需要性别，年龄，人种的多样性训练样本，更需要在标注时，逐张手工标注人脸及各部位的位置信息；在样本数量上，随着深度神经网络参数的愈发复杂，例如 Resnet 等大型深度神经网络框架拥有一百多层网络和千万量级的参数，训练它所要求的数据数量也需要达到百万量级以上。

为了解决此类问题，出现了许多无监督或半监督学习的模型，即训练的样本无需全部人工标注，只需使用在现实和网络中广泛存在的未标注数据，用以学习模型（半监督则只需要少量标注），其中，生成式模型是最有潜力的方向之一。生成式学习意图通过少量的，没有标签的无监督学习方式，让计算机理解图片，视频或其他数据类型中隐含的语义信息，对其进行建模。

本课题主要研究的是 FRAME 随机场模型，一个发展较为成熟，但相对小众的生成式模型。该模型相对于其他生成式模型，具有更加鲜明的统计意义，极具解释性。深度学习经常令人诟病的一点，是研究者很难指出神经网络学出的特征指代的是什么。而在 FRAME 模型中，提取的特征则非常的明确可解释，一个训练好的模型可以通过多种方式，可视化出模型。通过学习此类模型方法，能够让计算机更好的理解数据，真正达到“机器学习”的学习目的。同时，模型的无监督特性和极少数据需求的特性也大大增加了模型的适用范围。

与此同时，目前领域内的许多高性能的模型都使用了层次化表示方法，自底而上或自顶向下，大大提高了模型的适用范围和效果。层次化表示的思想与人脑的思维方式一致，从局部到整体，从整体到局部是人观察世界的通用手段。在本课题中，我们对 FRAME 随机场模型引入了层次化的特征，以此来提高模型的适用性和鲁棒性。

1.2 研究内容简述

本课题研究的是计算机视觉中的生成模型与生成式学习。我们将主要研究 FRAME 随机场模型及其优化和改进，包括非均一化 FRAME 随机场模型和稀疏 FRAME 随机场模型，然后我们拓展了层次化稀疏随机场模型，并对其进行纵向对比。同时，我们还将选择几个类似的生成式模型进行实现和探索，将它们与 FRAME 随机场模型进行横向对比。

FRAME 随机场模型是一个应用在二维信息上的生成式概率模型。FRAME 是 Filter 过

滤器，Random Field 随机场和 Maximum Entropy 最大熵的缩写。模型最初应用在随机纹理图案上，将图片信息视为均匀的马尔科夫随机场，使用大量的过滤器建模，然后借鉴了最大熵原理进行学习与训练。

为了拓展适用范围至物件图像，引入了非均一化 FRAME 随机场模型，在其中，不同的过滤器拥有不同的权值。对物件类型描述更加重要的部分过滤器拥有更高的权值，例如边缘部分，明确的线条部分等等。非均一化 FRAME 随机场模型也被称为稠密 FRAME 随机场模型，因为模型在各个方位，都有着不同大小，不同角度的过滤器，数量巨大。为了简化模型，降低计算量，引入了稀疏 FRAME 随机场模型，过滤器的定义方式与原先相同，但在稀疏 FRAME 随机场模型中，我们只选择少量的非常重要的过滤器来定义模型。

稀疏 FRAME 随机场模型中，作为过滤器的 gabor 特征可以小规模的变形，因此能够拥有一定的鲁棒性。然而，面对较大范围的扰动，例如头部整体旋转或者肢体的扰动，则会失败。于是，我们拓展了这个模型，增加了层次化，加入了部件级别的表达和变形。我们添加了一层，将多个 gabor 特征的集合作为新的过滤器，使其能够整体的小规模变形，由此大大提高了模型的鲁棒性。我们将改进后的模型称之为层次化稀疏 FRAME 随机场模型。

除了在实现模型的过程中，进行可视化模型的探索之外，我们设计了三个计算机视觉领域的基本任务来评估和对比 FRAME 随机场模型。分别是，物件检索及关键点预测，聚类问题和分类问题。三个问题均采用了无监督学习的方式，输入的数据包括 FRAME 模型参考论文中的已有数据，以及我们自己收集的部分数据。训练的结果将于其他的无监督生成式模型进行对比，展示模型效果。

1.3 相关工作与文献综述

目前在计算机视觉领域对物件图案的高层语义表达较为出色的模型大多为判别式模型，例如参考文献[1,2,3]中描述的层次化学习模型。此类模型往往需要大量的训练样本，同时学习出的模型不具备很好的解释性。

生成式模型方面，最原始的生成式模型包括 GMM 高斯混合模型[4]，朴素贝叶斯模型[4]，LDA 主题模型[5]等等；基于深度学习，则有 GAN 对抗网络[6]，VAE 变分自编码模型[7]等等。在层次化语义表达方面有参考文献[8, 9]，与我们的 FRAME 随机场模型极为相似，也使用了 Gabor 过滤器的组合。我们的工作同时与 And-or 图[10]模型和多层模型[11]相关。为了表达模型中的视觉部件，他们均使用了基于 Gabor 过滤器的模板，为了规避马尔科夫蒙特卡洛 MCMC 的计算量，假设选择的过滤器是正交且独立的。而我们基于 FRAME 随机场模型，并不需要上述假设，所以我们的模型可以更好的通过 MCMC 进行训练和生成，大幅增加了模型的可解释性。

第二章　FRAME 随机场模型的算法与原理

本章节将阐述 FRAME 随机场模型的框架结构，训练算法和原理。为第三章推广层次化稀疏 FRAME 随机场模型打下基础。

在第一节中，我们将回顾最原始的 FRAME 随机场模型[16]，并对后续章节会使用到的名词，算法进行简单的说明。

第二节与第三节将阐述 FRAME 随机场模型后续的发展，分别是非齐次 FRAME 随机场模型[14]和稀疏 FRAME 随机场模型[15]。

2.1 FRAME 随机场模型简介与名词解释

FRAME 随机场模型，即 filter，random field 和 maximum entropy，过滤器，随机场和熵，模型使用了一系列过滤器，借鉴了最大熵原理，将二维信息建模成为一个马尔科夫随机场。原始的 FRAME 随机场模型应用在图案纹理图片上，经过非均一化的拓展后运用在了物件图像中。

FRAME 随机场模型是一个生成式概率模型。接下来，我们分别解释一下什么是生成式模型，马尔科夫随机场，最大熵原理，哈密尔顿蒙特卡洛以及使用到的 Gabor 过滤器。

2.1.1 生成式模型与判别式模型

生成式模型与判别式模型是机器学习领域的两种模型分类。简单来说，生成式模型估计输入数据 x 相对于类别标签 y 的联合概率分布$P(x,y)$，而判别式模型则估计他们的条件概率分布$P(y|x)$。通过贝叶斯公式，生成式模型可以得到判别式模型，反之则不行。

下图中，左图是两个高斯模型，混合后即为高斯混合模型，这是一个较为基本的生成式模型，模型将观测样本视为一个混合的高斯模型，而算法要做的就是将他们找出来。右图则是一个理想化的二类分割的分布，这个分布使用判别式模型能达到很好的效果，而判别式模型算法的目的就是要找到中间的分隔。

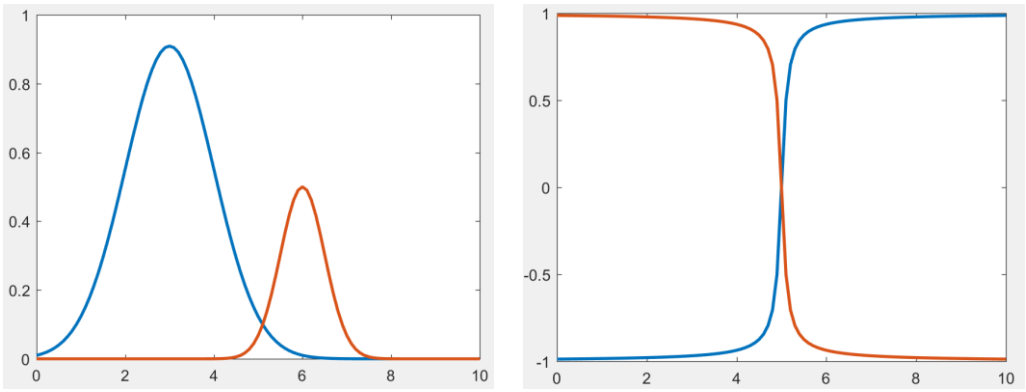


图 2-1 生成式模型与判别试模型

两者的具体特点如下表所示：

表 2-1 生成式模型与判别试模型的区别

	生成式模型	判别式模型
特点	对后验概率建模，从统计角度表示数据分布情况，反应同类数据本身的相似度	寻找不同类别之间的最优分类面，反映的是异类数据之间的差异
优点	1. 携带信息较为丰富 2. 单类问题研究更加灵活 3. 模型可以增量学习 4. 适用于不完整的数据学习	1. 分类便捷灵活 2. 清晰分辨差异特征 3. 性能要求较少，容易学习
缺点	1. 学习和计算过程较为复杂 2. 容易分类是得到误报	1. 不能反映数据本身的特性 2. 变量间的关系不清楚，不可视

2.1.2 Gabor 过滤器

在本文中，我们实验使用的过滤器称之为 Gabor 过滤器。它是一个用于边缘检测的线性过滤器，Gabor 过滤器的频率与方向表示都很类似人类的视觉神经系统，哺乳动物大脑的视觉皮层中的细胞就可以被建模成 Gabor 函数。在实验中，它们在纹理表达和区分上有很好的效果。一个典型的 Gabor 过滤器可视化如图 2-2 所示。

在空间域上，一个二维的 Gabor 过滤器是一个由正弦平面波调制的高斯核函数。即，Gabor 过滤器的脉冲响应由正弦波乘以高斯函数定义。由卷积定律，Gabor 滤波器脉冲响应的傅里叶变换是谐波函数的傅里叶与高斯函数的傅里叶的卷积。滤波器具有表示正交方向的实部和虚部，两个成分可以形成复数或单独使用。

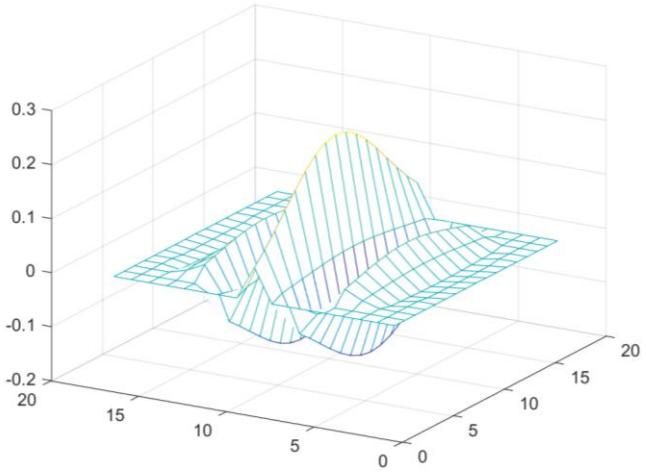


图 2-2 Gabor 过滤器可视化

其中，它的复数形式是：

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma)=\exp\left(\frac{-x'^2+\gamma^2y'^2}{2\sigma^2}\right)\exp\left(i\left(2\pi\frac{x'}{\lambda}+\psi\right)\right) \tag{2-1}$$

实部为：

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma)=\exp\left(\frac{-x'^2+\gamma^2y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\psi\right) \tag{2-2}$$

虚部为：

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma)=\exp\left(\frac{-x'^2+\gamma^2y'^2}{2\sigma^2}\right)\sin\left(2\pi\frac{x'}{\lambda}+\psi\right) \tag{2-3}$$

其中：$x'=x\cos\theta+y\sin\theta$；$y'=-x\sin\theta+y\cos\theta$

2.1.3 （马尔科夫）随机场

在概率论中，由样本空间$\Omega = \{0,1,\dots\}^n$取样构成的随机变量X_i组成的$S = \{X_i, i = 1,\dots,n\}$。若对所有的$\omega \in \Omega$均满足$\pi(\omega) > 0$，则称$\pi$为一个随机场。

而马尔科夫随机场则是满足了马尔科夫性质性质的随机场，马尔科夫性质如下：

$$P(X_{n+1} = x|X_0,\dots,X_n) = P(X_{n+1} = x|X_n) \tag{2-4}$$

也即，后一个随机变量仅与它的前一个变量有关，而与再之前的变量无关。马尔科夫性质是一个在统计学习领域非常重要的性质，它的存在让许多的概率模型的建模，采样成为可能。

在许多的概率模型包括 FRAME 随机场模型中，将二维的图片信息视为一个二维平面上的随机场，也即每一个位置的像素取值按照某种分布赋值。有时候模型也会假定图片是一个马尔科夫随机场，在这种情况下，每个像素的取值只和他边上的相邻像素取值有关，这很好的表现了图像中像素之间的空间关联性，能够有效的描述图像的局部统计特性。

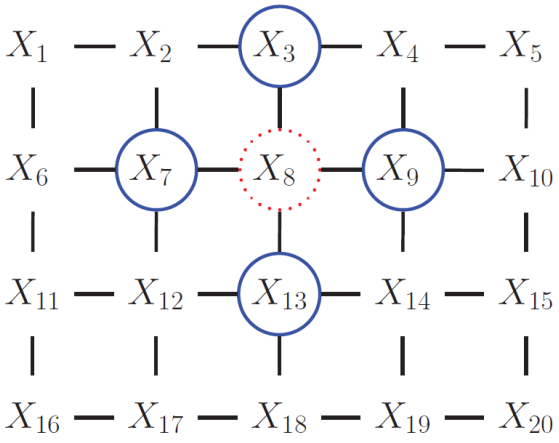


图 2-3 随机场示意图

2.1.4 最大熵原理

熵是统计学中量化一系列随机变量混乱程度的量。最大熵原理是一种选择随机变量统计特性最符合客观情况的准则，在建模一个随机分布时非常重要。有部分观测或先验知识时，在无穷多种可能的分布中，我们一般选择熵最大的分布。这种方法虽然有一定主观性，但却也是最符合客观情况的一种选择。

例如我们知道一个已知函数的期望：$E_p[\phi_n(x)] = \int \phi_n(x)p(x)dx = \mu_n, for\ n = 1,\dots,N$。令$\Omega = \{p(x) \mid E_p[\phi_n(x)] = \mu_n, n = 1,\dots,N\}$，那么最大熵原理认为最优选择是选择一个熵最大的分布：

$$p^*(x) = \arg \max_{E_p[\phi_n(x)]=\mu_n} \left\{ -\int p(x) \log p(x)\, dx \right\} \tag{2-5}$$

根据拉格朗日乘数，

$$p(x;\Lambda) = \frac{1}{Z(\Lambda)} \exp\left(-\sum_{n=1}^{N} \lambda_n \phi_n(x) \right) \tag{2-6}$$

其中$\Lambda = \{\lambda_i, i = 1,\dots,n\}$为拉布朗日算子。

我们不在这里继续演算这个例子，我们可以在这里看到这个最优的概率分布和我们的 FRAME 随机场模型极其相似，我们的模型也蕴含了最大熵原理来选择模型。

2.1.5 哈密尔顿蒙特卡洛方法

哈密尔顿蒙特卡洛方法（HMC）也称为混合蒙特卡洛，是一种马尔科夫蒙特卡洛 MCMC 算法，他利用了梯度信息，使得随机游走能更高的向高概率的区域移动。它基于哈密尔顿动力学的离散化，并引入了 M-H 判断步骤来保证分布的正确性和稳定性。

马尔科夫蒙特卡洛方法 MCMC 就是构造合适的马尔科夫链进行抽样然后使用蒙特卡洛方法进行积分计算。马尔科夫链是符合马尔科夫性质的一系列随机变量（马尔科夫性质就是指随机变量仅与它的前一个变量有关，而与再之前的变量无关），它可以收敛到平稳分布。建立一个以π为平稳分布的马尔科夫链，达到平稳状态后，马尔科夫链的值就相当于在分布$\pi(x)$中抽取样本，所以可以达到随机模拟的目的。蒙特卡洛方法则是统计模拟方法，也即通过大量随机样本来近似进行积分计算，通过抽样近似全局。

HMC 模拟哈密尔顿力学的物理系统进行采样。物理系统中的粒子拥有势能和动能，从而在高维空间中运动。粒子可以通过位置和速度的组合态表示，位置和速度两个变量相互独立，我们可以先获取速度然后获取位置从而进行采样。HMC 大体算法如下：

1. 根据单变量高斯分布采样变量（粒子）的速度。
2. 执行 n 次迭代（即随机游走）获得新的位置状态 X。
3. 通过位置状态 X 的能量高低来决定是否接受他成为新的采样。

2.2 非齐次 FRAME 随机场模型

2.2.1 非齐次 FRAME 随机场模型框架

我们开始尝试建模相同类别的物件图像。令$\{\boldsymbol{I}_m,\ m=1,\dots,M\}$是训练图像的集合，他们同属于一个类型$D$；$B_{x,s,\alpha}$是位置以 x 为中心，大小为 s，角度为$\alpha$的基础过滤器函数，在本文的实验中，我们选择 Gabor 小波函数。我们假设 s 和α在有限和适当的离散范围内取值。内积$\langle \boldsymbol{I}, B_{x,s,\alpha}\rangle$可以被视为图片$\boldsymbol{I}$对过滤器函数$B_{x,s,\alpha}$的响应。我们假设过滤器函数都被规范化到了$\ell_2-\text{norm}$。

非齐次 FRAME 随机场模型的概率分布定义如下：

$$p(\boldsymbol{I};\lambda)=\frac{1}{Z(\lambda)}\exp\left(\sum_{x,s,\alpha}\lambda_{x,s,\alpha}\left(\langle \boldsymbol{I}, B_{x,s,\alpha}\rangle\right)\right)q(\boldsymbol{I}) \tag{2-7}$$

其中，$q(\boldsymbol{I})$是一个已知的参考分布；$\lambda_{x,s,\alpha}()$是不同过滤器的权重函数；$Z(\lambda)$是归一化常数。

$$Z(\lambda)=\int\left[\exp\left(\sum_{x,s,\alpha}\lambda_{x,s,\alpha}\left(\langle \boldsymbol{I}, B_{x,s,\alpha}\rangle\right)\right)\right]q(\boldsymbol{I})d\boldsymbol{I}=E_q\left[\exp\left(\sum_{x,s,\alpha}\lambda_{x,s,\alpha}\left(\langle \boldsymbol{I}, B_{x,s,\alpha}\rangle\right)\right)\right] \tag{2-8}$$

前面提到，原始的 FRAME 随机场模型应用在纹理图案上是一个稳定的马尔科夫随机场，他假设权重函数λ独立于位置 x，只与大小 s 和角度α有关，而非齐次 FRAME 随机场模型则需要分别估计每个不同的权重函数$\lambda_{x,s,\alpha}()$。非参数的估计$\lambda_{x,s,\alpha}()$会导致计算压力过大，在小数据量的训练图像情况下，我们可以参数化它

$$\lambda_{x,s,\alpha}(r)=\lambda_{x,s,\alpha}|r|$$

其中，$r=\langle \boldsymbol{I}, B_{x,s,\alpha}\rangle$，即过滤器响应，等式右边的$\lambda_{x,s,\alpha}$代表响应绝对值的系数。

对于参考分布$q(\boldsymbol{I})$，很多的马尔科夫随机场模型包括原始 FRAME 随机场模型都使用了简单的均匀分布。在本文中，我们使用了高斯白噪声，遵循独立$N(0,\sigma^2)$高斯分布：

$$q(\boldsymbol{I})=\frac{1}{(2\pi\sigma^2)^{|D|/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_x \boldsymbol{I}(x)^2\right) \tag{2-9}$$

其中，|D|是图像中的像素数量，$q(\boldsymbol{I})$本身应是一个最大熵模型，我们希望他能重现图像强度的边缘均值和方差。在本文中，我们标准化了训练样本，将边缘均值置为 0，固定方差$\sigma^2=1$。这个$q(\boldsymbol{I})$可以视为模型的初始值，或者视为前景对象删除的背景残差图像的模型。由此，$p(\boldsymbol{I};\lambda)$可以被写作一个相对于统一度量的指数族模型。

2.2.2 非齐次 FRAME 随机场模型训练算法

FRAME 随机场模型可以视为一个指数族模型的特殊情况，参数$\boldsymbol{\lambda}=(\lambda_{x,s,\alpha},\forall x,s,\alpha)$可以通过训练图像经过最大似然估计学习得出，期望函数是：

$$E_\lambda\left(\left|\langle \boldsymbol{I}, B_{x,s,\alpha}\rangle\right|\right)=\frac{1}{M}\sum_{m=1}^{M}\left|\langle \boldsymbol{I}_m, B_{x,s,\alpha}\rangle\right|,\forall x,s,\alpha \tag{2-10}$$

似然函数是：

$$L(\lambda)=\frac{1}{M}\sum_{m=1}^{M}\log p(\boldsymbol{I}_m;\lambda)=\frac{1}{M}\sum_{m=1}^{M}\sum_{x,s,\alpha}\lambda_{x,s,\alpha}\left|\langle \boldsymbol{I}_m, B_{x,s,\alpha}\rangle\right|-logZ(\lambda)+\frac{1}{M}\sum_{m=1}^{M}\log q(\boldsymbol{I}_m) \tag{2-11}$$

我们通过梯度下降算法最大化似然函数$L(\lambda)$，梯度为：

$$\frac{\partial L(\lambda)}{\partial \lambda_{x,s,\alpha}}=\frac{1}{M}\sum_{m=1}^{M}\left|\langle \boldsymbol{I}_m, B_{x,s,\alpha}\rangle\right|-E_{p(\boldsymbol{I};\lambda)}\left[\left|\langle \boldsymbol{I}, B_{x,s,\alpha}\rangle\right|\right],\forall x,s,\alpha \tag{2-12}$$

其中，$E_{p(\boldsymbol{I};\lambda)}[\left|\langle \boldsymbol{I}, B_{x,s,\alpha}\rangle\right|]$是分布$p(\boldsymbol{I};\lambda)$下$\left|\langle \boldsymbol{I}, B_{x,s,\alpha}\rangle\right|$的期望，也是$\log Z(\lambda)$的导数。

我们通过采样样本的方式估计期望。

$$E_{p(\boldsymbol{I};\lambda)}\left[\left|\langle \boldsymbol{I}, B_{x,s,\alpha}\rangle\right|\right]\approx\frac{1}{\bar{M}}\sum_{m=1}^{\bar{M}}\left|\langle \bar{\boldsymbol{I}}_m, B_{x,s,\alpha}\rangle\right| \tag{2-13}$$

生成图像样本可以通过哈密尔顿蒙托卡罗方法（HMC）采样。与 Gibbs 采样不同，HMC 使用能量函数的梯度来计算，这很符合我们的模型期望。马尔卡夫蒙托卡罗方法中需要计算的能量函数为：

$$U(\boldsymbol{I})=-\sum_{i=1}^{n}\lambda_i\left|\langle \boldsymbol{I}_m, B_{x_i,s_i,\alpha_i}\rangle\right|+\frac{1}{2}|\boldsymbol{I}|^2 \tag{2-14}$$

它的梯度为：

$$\frac{\partial U}{\partial \boldsymbol{I}} = -\sum_{x,s,\alpha} \lambda_{x,s,\alpha} sign\big(\langle \boldsymbol{I}, B_{x,s,\alpha}\rangle\big) B_{x,s,\alpha} + \boldsymbol{I} \tag{2-15}$$

计算需要两层卷积计算（自底向上的卷积，然后是自顶向下的反卷积），可以通过 Matlab 内置的函数通过 GPU 快速得出。初始化成功后，$\bar{M}$个样本可以并行的采样得出。

令$\lambda^{(t)}$是第 t 轮的估计，$\{\bar{\boldsymbol{I}}_m, m = 1, \dots, \bar{M}\}$是通过目前的概率估计$p\big(I, \lambda^{(t)}\big)$采样生成出的图像样本，然后我们可以更新$\lambda$:

$$\lambda_{x,s,\alpha}^{(t+1)} = \lambda_{x,s,\alpha}^{(t)} + \gamma_t \Big(\frac{1}{M}\sum_{m=1}^{M} \big|\langle \boldsymbol{I}_m, B_{x,s,\alpha}\rangle\big| - \frac{1}{\bar{M}}\sum_{m=1}^{\bar{M}} \big|\langle \bar{\boldsymbol{I}}_m, B_{x,s,\alpha}\rangle\big|\Big) \tag{2-16}$$

这既是我们的算法使用的最大似然估计函数。其中γ_t是步长。

对于归一化参数$Z(\lambda)$，我们可以计算前后迭代回合的比值间接得出：

$$\frac{Z\big(\lambda^{(t+1)}\big)}{Z\big((\lambda^{(t)}\big)} = E_{p(\boldsymbol{I};\lambda^{(t)})}\left[\exp\left(\sum_{x,s,\alpha}\Big(\lambda_{x,s,\alpha}^{(t+1)} - \lambda_{x,s,\alpha}^{(t)}\Big) \times \big|\langle \boldsymbol{I}, B_{x,s,\alpha}\rangle\big|\right)\right] \tag{2-17}$$

这里的期望可以直接估计为采样得出的图像样本$\{\bar{\boldsymbol{I}}_m\}$取平均：

$$\frac{Z\big(\lambda^{(t+1)}\big)}{Z\big((\lambda^{(t)}\big)} \approx \frac{1}{\bar{M}}\sum_{m=1}^{\bar{M}} \exp\left(\sum_{x,s,\alpha}\Big(\lambda_{x,s,\alpha}^{(t+1)} - \lambda_{x,s,\alpha}^{(t)}\Big) \times \big|\langle \bar{\boldsymbol{I}}_m, B_{x,s,\alpha}\rangle\big|\right) \tag{2-18}$$

训练开始时，我们令$\lambda^{(0)} = 0, \log Z\big(\lambda^{(0)}\big) = 0$，之后如此迭代：

$$\log Z\big(\lambda^{(t+1)}\big) = \log Z\big(\lambda^{(t)}\big) + \log \frac{Z\big(\lambda^{(t+1)}\big)}{Z\big(\lambda^{(t)}\big)} \tag{2-19}$$

$Z\big(\lambda^{(t)}\big)$的计算基于$p(\boldsymbol{I}; \lambda^{(t)})$的并行马尔科夫链的计算。

2.2.3 非齐次 FRAME 随机场模型训练步骤

在这里，我们重新简明扼要的描述一遍算法的具体流程。

算法步骤 1　训练非齐次 FRAME 随机场模型

算法输入：

一系列的训练图片：$\{\boldsymbol{I}_m,\ \boldsymbol{m} = \boldsymbol{1}, \dots, \boldsymbol{M}\}$

算法输出：

$\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_{x,s,\alpha}, \forall \boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\alpha}\}$ 和 $\boldsymbol{\log Z(\lambda)}$。

1. 创建所有的过滤器$\{\boldsymbol{B}_{x,s,\alpha}, \forall \boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\alpha}\}$。
2. 初始化$\boldsymbol{\lambda}_{x,s,\alpha}^{(0)} \leftarrow \boldsymbol{0}, \forall \boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\alpha}$，$\boldsymbol{\log Z(\lambda^{(0)}) \leftarrow 0}$，$\boldsymbol{t \leftarrow 0}$。
3. 计算训练样本 sum 图：$\boldsymbol{H_{x,s,\alpha}^{obs} \leftarrow \frac{1}{M}\Sigma_{m=1}^{M}\big|\langle I_m, B_{x,s\cdot\alpha}\rangle\big|, \forall x, s, \alpha}$。
4. 初始化生成的采样样本，置为高斯白噪声图：$\boldsymbol{\{\bar{I}_m\} \sim N(0, \sigma^2)}$。

   REPEAT

   5. 通过 HMC，从$\boldsymbol{p(I; \lambda^{(t)})}$中生成采样图片样本$\boldsymbol{\{\bar{I}_m, m = 1, \dots, \bar{M}\}}$。
   6. 计算生成样本 sum 图：$\boldsymbol{H_{x,s,\alpha}^{syn} \leftarrow \frac{1}{\bar{M}}\Sigma_{m=1}^{\bar{M}}\big|\langle \bar{I}_m, B_{x,s\cdot\alpha}\rangle\big|, \forall x, s, \alpha}$

7. 更新 $\boldsymbol{\lambda}_{x,s,\alpha}^{(t+1)} \leftarrow \boldsymbol{\lambda}_{x,s,\alpha}^{(t)} + \boldsymbol{\gamma}_t\left(\boldsymbol{H}_{x,s,\alpha}^{obs} - \boldsymbol{H}_{x,s,\alpha}^{syn}\right), \forall \boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\alpha}$
8. 更新 $\log \boldsymbol{Z}\left(\boldsymbol{\lambda}^{(t+1)}\right) \leftarrow \log \boldsymbol{Z}\left(\boldsymbol{\lambda}^{(t)}\right) + \frac{1}{\bar{M}} \sum_{m=1}^{\bar{M}} \exp\left(\sum_{x,s,\alpha}\left(\boldsymbol{\lambda}_{x,s,\alpha}^{(t+1)} - \boldsymbol{\lambda}_{x,s,\alpha}^{(t)}\right) * \left|\langle \bar{\boldsymbol{I}}_m, \boldsymbol{B}_{x,s,\alpha}\rangle\right|\right)$
9. 令 $\boldsymbol{t} \leftarrow \boldsymbol{t} + \boldsymbol{1}$

UNTIL $\sum_{x,s,\alpha}\left|\boldsymbol{H}_{x,s,\alpha}^{obs} - \boldsymbol{H}_{x,s,\alpha}^{syn}\right| < \epsilon$

2.3 稀疏 FRAME 随机场模型

2.3.1 稀疏化 FRAME 随机场模型框架

非齐次 FRAME 随机场模型使用了大量的过滤器，其概率公式中，(x, s, α)需要遍历所有的位置，大小和角度，使得模型的训练和构造极其复杂。然而，一张图片信息中，大部分的过滤器其实响应有限。于是稀疏化就应运而生了，在稀疏 FRAME 随机场模型中，我们只选择少量的过滤器进行训练和构造。我们这样定义：

$$p(\boldsymbol{I}; \boldsymbol{B}, \lambda) = \frac{1}{Z(\lambda)} \exp\left(\sum_{x,s,\alpha} \lambda_i\left(\langle \boldsymbol{I}, B_{x_i,s_i,\alpha_i}\rangle\right)\right) q(\boldsymbol{I}) \tag{2-20}$$

其中，$B = (B_{x_i,s_i,\alpha_i}, i = 1, \ldots, n)$是选择的过滤器集合，$\lambda = (\lambda_i, i = 1, \ldots, n)$是各自的权重。给定这样的一个过滤器集合，我们可以使用相同的采样估计，训练迭代方式来训练模型，一些最大熵的性质依然成立。

它有着如下的优点：

（1） 它大大加快了训练速度，减少了模型大小。
（2） 因为参数数量变少，模型的鲁棒性和稳定性会更好。
（3） 在进行马尔科夫蒙特卡洛估计时，因为过滤器不再高度相关，所以收敛速度更快。
（4） 在图像重构时，他可以跟线性加法稀疏编码模型联系起来。
（5） 它允许选择的过滤器函数进行位置，大小和角度的微小扰动和变形。

我们展开说明一下第五点，在选择了少量的过滤器之后，我们将$p(\boldsymbol{I}; \boldsymbol{B}, \lambda)$变成一个可以变形的模型，用来适配每个训练图像，也即，每个图像都有一个通过通用的模板 B 经过微小扰动后得到的模板$B_m = (B_{x_i+\Delta x_{m,i}, s_i, \alpha_i+\Delta\alpha_{m.i}}, i = 1, \ldots, n)$。$\Delta x_{m,i}$，$\Delta\alpha_{m,i}$都是一个很小的离散的值（例如可以设定为$\Delta x_{m,i} \in [-3,3]$，即扰动不超过三个像素），当我们在适配$I_m$时，公式中的$B_{x_i,s_i,\alpha_i}$就被替换成为了$B_{x_i+\Delta x_{m,i}, s_i, \alpha_i+\Delta\alpha_{m.i}}$。

2.3.2 稀疏 FRAME 随机场模型的学习算法

此模型的难点是如何选择这些过滤器，并通过这些过滤器来学习模型。刚才在优点中提到，稀疏 FRAME 随机场模型与线性加法稀疏编码模型有很大的联系，我们可以用稀疏编码模型的方式来描述我们的模型：

$$I_m = \sum_{i=1}^{n} c_{m,i} B_{x_i,s_i,\alpha_i} + \epsilon_m \tag{2-21}$$

其中，$c_{m,i}$是线性投影的最小二乘重建系数。ϵ_m是结果残差图像。

介于高斯白噪声背景模型$q(\boldsymbol{I})$是独立的，那么上述模型中：$C_m = (c_{m,i}, i = 1, \dots, n)$服从一个确定的分布：$p_C(C;\lambda)$，$\epsilon_m$则是高斯白噪声的一个投影，与$C_m$独立。那么$I_m$的似然函数可以分解为$C_m$与$\epsilon_m$的似然函数。前者依赖于$\lambda$，而后者则只依赖于残差图像：$\| \epsilon_m \|^2 = \| I_m - \sum_{i=1}^{n} c_{m,i} B_{x_i,s_i,\alpha_i} \|^2$。

这样考虑稀疏 FRAME 随机场模型的话，之前的训练步骤能够一分为二。第一步，我们通过最小化总体最小二乘重构误差来选择过滤器集合$\boldsymbol{B} = (B_{x_i,s_i,\alpha_i}, i = 1, \dots, n)$；第二步，已知了过滤器集合，我们通过使用训练非齐次 FRAME 随机场模型的方法来估计$\boldsymbol{\lambda}$。

第一步：可变形共享稀疏编码的训练

这一步中，我们希望通过训练图像，找出我们所需要的过滤器集合。

对于训练图像集合$\{\boldsymbol{I}_m,\ m = 1, \dots, M\}$，为了训练过滤器集合$\boldsymbol{B} = (B_{x_i,s_i,\alpha_i}, i = 1, \dots, n)$我们最小化：

$$\sum_{m=1}^{M} \| I_m - \sum_{i=1}^{n} c_{m,i} B_{x_i+\Delta x_{m,i},s_i,\alpha_i+\Delta\alpha_{m.i}} \|^2 \tag{2-22}$$

最小化可以通过共享匹配追求算法来完成，该算法可以选择基本函数来同时对多个图像进行编码，同时通过局部最大池化算法来推断局部扰动。最大池化算法如下：

$$(x_i, s_i, \alpha_i) = \arg\max_{x,s,\alpha} \sum_{m=1}^{M} \max_{\Delta x,\Delta\alpha} \left|\langle \epsilon_m, B_{x+\Delta x,s,\alpha+\Delta\alpha} \rangle\right|^2 \tag{2-23}$$

其中，$max_{\Delta x,\Delta\alpha}$就是将$\Delta x_{m,i}$和$\Delta x_{\alpha,i}$局部最大池化的结果。得到$(x_i, s_i, \alpha_i)$后对每个图像，通过选择取到最大值的扰动值作为$\Delta x_{m,i}$和$\Delta x_{\alpha,i}$：

$$\left(\Delta x_{m,i}, \Delta x_{\alpha,i}\right) = \arg\max_{\Delta x,\Delta\alpha} \left|\langle \epsilon_m, B_{x+\Delta x,s,\alpha+\Delta\alpha} \rangle\right|^2 \tag{2-24}$$

第二步 训练稀疏 FRAME 随机场模型

在找到了我们需要的过滤器集合$\boldsymbol{B} = (B_{x_i,s_i,\alpha_i}, i = 1, \dots, n)$后，我们对$I_m$建模，同样使用最大似然估计来估计$\lambda$。似然函数如下：

$$L(I_m|B,\lambda) = \sum_{i=1}^{n} \lambda_i \max_{\Delta x,\Delta\alpha} \left|\langle \boldsymbol{I}, B_{x_i+\Delta x,s_i,\alpha_i+\Delta\alpha} \rangle\right| - \log Z(\lambda) \tag{2-25}$$

这个可以当作模板的匹配分数。我们同样允许过滤器进行位置和方向的小幅扰动，扰动同样由局部最大池化算法进行估计。

在学习算法中，同样的，令$\lambda^{(t)}$是当前回合对λ的估计，$\{\bar{\boldsymbol{I}}_m, m = 1, \dots, \bar{M}\}$是使用 HMC 从$p(\boldsymbol{I};\boldsymbol{B},\lambda^{(t)})$采样生成的图片样本。我们这样更新$\lambda$:

$$\lambda_i^{(t+1)} = \lambda_i^{(t)} + \gamma_t \left(\frac{1}{M} \sum_{m=1}^{M} \max_{\Delta x,\Delta\alpha} \left|\langle \boldsymbol{I}_m, B_{x_i+\Delta x,s_i,\alpha_i+\Delta\alpha} \rangle\right| - \frac{1}{\bar{M}} \sum_{m=1}^{\bar{M}} \left|\langle \bar{\boldsymbol{I}}_m, B_{x_i,s_i,\alpha_i} \rangle\right|\right) \tag{2-26}$$

学习到的模型$p(\boldsymbol{I};\boldsymbol{B},\lambda)$是没有变形过的模板。对于生成的图片样本，没有局部最大池化。局部最大池化只应用在训练样本上用来滤除它们的形状变形。因此，在外观和形状变化之间存在明显的分离。

同样的，我们使用哈密尔顿蒙特卡洛方法（HMC）来采样生成图像，对于 HMC 中能量函数的梯度与式（2-15）类似，为$\sum_i \lambda_i sign\left(\langle \boldsymbol{I}, B_{x_i,s_i,\alpha_i}\rangle\right)B_{x_i,s_i,\alpha_i} + \boldsymbol{I}$。

对于归一化参数$Z(\lambda)$，使用类似的式子计算比值间接求解：

$$\frac{Z\left(\lambda^{(t+1)}\right)}{Z((\lambda^{(t)})} \approx \frac{1}{\bar{M}}\sum_{m=1}^{\bar{M}} \exp\left(\sum_{x,s,\alpha}\left(\lambda_{x,s,\alpha}^{(t+1)} - \lambda_{x,s,\alpha}^{(t)}\right) \times \left|\langle \bar{\boldsymbol{I}}_m, B_{x,s,\alpha}\rangle\right|\right) \tag{2-27}$$

2.3.3 稀疏 FRAME 随机场模型训练步骤

在这里，我们重新简明扼要的描述一遍算法的具体流程。

算法步骤 2　训练稀疏 FRAME 随机场模型

算法输入：

一系列的训练图片：$\{\boldsymbol{I}_m,\ \boldsymbol{m}=\boldsymbol{1},\dots,\boldsymbol{M}\}$

算法输出：

选择的过滤器函数集合$\boldsymbol{B}=(\boldsymbol{B}_{x_i,s_i,\alpha_i},\boldsymbol{i}=\boldsymbol{1},\dots,\boldsymbol{n})$，权重参数$\boldsymbol{\lambda}=\{\boldsymbol{\lambda}_{x,s,\alpha},\forall \boldsymbol{x},\boldsymbol{s},\boldsymbol{\alpha}\}$ 和 $\boldsymbol{\log Z(\lambda)}$。

1. 初始化$\boldsymbol{i}\leftarrow\boldsymbol{0}$。

   REPEAT

   2. 令$\boldsymbol{i}\leftarrow\boldsymbol{i}+\boldsymbol{1}$，选择：$(\boldsymbol{x_i},\boldsymbol{s_i},\boldsymbol{\alpha_i})\leftarrow \boldsymbol{\arg\max}_{x,s,\alpha}\sum_{m=1}^{M}\boldsymbol{\max}_{\Delta x,\Delta\alpha}\left|\langle \boldsymbol{\epsilon}_m, \boldsymbol{B}_{x+\Delta x,s,\alpha+\Delta\alpha}\rangle\right|^2$。
   3. 对于每个 m，给定$(\boldsymbol{x_i},\boldsymbol{s_i},\boldsymbol{\alpha_i})$，得到：$(\boldsymbol{\Delta x_{m,i}},\boldsymbol{\Delta x_{\alpha,i}})\leftarrow \boldsymbol{\arg\max}_{\Delta x,\Delta\alpha}\left|\langle \boldsymbol{\epsilon}_m, \boldsymbol{B}_{x+\Delta x,s,\alpha+\Delta\alpha}\rangle\right|^2$。
   4. 更新参数：$\boldsymbol{c_{m,i}}\leftarrow\langle \boldsymbol{\epsilon_m}, \boldsymbol{B}_{x_i+\Delta x_{m,i},s_i,\alpha_i+\Delta\alpha_{m,i}}\rangle$。
   5. 更新残差项：$\boldsymbol{\epsilon_m}\leftarrow\boldsymbol{\epsilon_m}-\boldsymbol{c_{m,i}}\boldsymbol{B}_{x_i+\Delta x_{m,i},s_i,\alpha_i+\Delta\alpha_{m,i}}$。

   UNTIL $\boldsymbol{i}=\boldsymbol{n}$

6. 初始化$\boldsymbol{\lambda_i^{(0)}}\leftarrow\boldsymbol{0}\ \boldsymbol{for}\ \boldsymbol{i}=\boldsymbol{1},\dots,\boldsymbol{n}$，$\boldsymbol{\log Z(\lambda^{(0)})}\leftarrow\boldsymbol{0}$，$\boldsymbol{t}\leftarrow\boldsymbol{0}$。
7. 计算训练样本 sum 图：$\boldsymbol{H_i^{obs}}\leftarrow\frac{\boldsymbol{1}}{\boldsymbol{M}}\boldsymbol{\Sigma_{m=1}^{M}}\left|\langle \boldsymbol{I_m}, \boldsymbol{B}_{x_i,s_i,\alpha_i}\rangle\right|, \boldsymbol{for}\ \boldsymbol{i}=\boldsymbol{1},\dots,\boldsymbol{n}$。
8. 初始化生成的采样样本，置为高斯白噪声图：$\{\bar{\boldsymbol{I}}_{\boldsymbol{m}}\}\sim\boldsymbol{N}(\boldsymbol{0},\boldsymbol{\sigma^2})$。

   REPEAT

   9. 通过 HMC，从$\boldsymbol{p}(\boldsymbol{I};\boldsymbol{B},\boldsymbol{\lambda^{(t)}})$中生成采样图片样本$\{\bar{\boldsymbol{I}}_{\boldsymbol{m}},\boldsymbol{m}=\boldsymbol{1},\dots,\bar{\boldsymbol{M}}\}$。
   10. 计算生成样本 sum 图：$\boldsymbol{H_i^{syn}}\leftarrow\frac{\boldsymbol{1}}{\bar{\boldsymbol{M}}}\boldsymbol{\Sigma_{m=1}^{\bar{M}}}\left|\langle \bar{\boldsymbol{I}}_{\boldsymbol{m}}, \boldsymbol{B}_{x_i,s_i,\alpha_i}\rangle\right|, \boldsymbol{for}\ \boldsymbol{i}=\boldsymbol{1},\dots,\boldsymbol{n}$。
   11. 更新 $\boldsymbol{\lambda_i^{(t+1)}}\leftarrow\boldsymbol{\lambda_{xi}^{(t)}}+\boldsymbol{\gamma_t}\left(\boldsymbol{H_i^{obs}}-\boldsymbol{H_i^{syn}}\right), \boldsymbol{for}\ \boldsymbol{i}=\boldsymbol{1},\dots,\boldsymbol{n}$。
   12. 更新$\boldsymbol{\log Z}\left(\boldsymbol{\lambda^{(t+1)}}\right)\leftarrow\boldsymbol{\log Z}\left(\boldsymbol{\lambda^{(t)}}\right)+\frac{\boldsymbol{1}}{\bar{\boldsymbol{M}}}\boldsymbol{\Sigma_{m=1}^{\bar{M}}}\boldsymbol{\exp}\left(\sum_i\left(\boldsymbol{\lambda_i^{(t+1)}}-\boldsymbol{\lambda_i^{(t)}}\right)*\left|\langle \bar{\boldsymbol{I}}_{\boldsymbol{m}}, \boldsymbol{B}_i\rangle\right|\right)$
   13. 令 $\boldsymbol{t}\leftarrow\boldsymbol{t}+\boldsymbol{1}$

   UNTIL $\sum_i\left|\boldsymbol{H_i^{obs}}-\boldsymbol{H_i^{syn}}\right|<\epsilon$

第三章　稀疏 FRAME 随机场模型的层次化拓展

在上一章节的稀疏 FRAME 随机场模型中，我们提到该模型中学习到的 Gabor 过滤器允许小范围的扰动，以此来增加模型的适配性。在某些时候，小范围的扰动并不足以适配部件层次的扰动。如果我们只是简单的加大扰动的范围，会导致模板因为扰动过大失去意义。所以我们引入了层次化的推广，使得模板在保证大体适合图片类型的情况下，允许部件层面，也即一个 Gabor 过滤器集合的整体扰动。层次化的语义表达是机器学习中很重要的一个部分，例如参考文献[8,9]和之后实验对比中使用的 DPM 可变动部件模型 [21]。我们的工作将层次化和生成式模型 FRAME 随机场模型结合，结合了两者的优势，大大加强了模型的适用范围。

3.1 层次化稀疏 FRAME 随机场模型框架

我们形式化的描述层次化的推广，推广后的模型概率分布如下：

$$p(\boldsymbol{I};\boldsymbol{H},\lambda)=\frac{1}{Z(\lambda)}\exp\left(\sum_{j=1}^{K}\sum_{i=1}^{n_j}\lambda_i^j\left(\langle \boldsymbol{I},B_{x_i^j,s_i^j,\alpha_i^j}\rangle\right)\right)q(\boldsymbol{I}) \tag{3-1}$$

其中，$H=\left\{\left(B_{x_i^j,s_i^j,\alpha_i^j},i=1,\dots,n_j\right),j=1,\dots,K\right\}$ 代表模型的模板中 K 个部件的集合，每个部件由n_j个过滤器组成。$\lambda=\left\{\left(\lambda_i^j,i=1,\dots,n_j\right),j=1,\dots,K\right\}$是各自的权重。

学习这样一个层次化的模型首先需要想之前训练 FRAME 模型那样找出需要的过滤器集合极其权重，然后找到对应的所属部件分块。在本文中，我们简单的将模板切成$K=d\times d$个方框，作为部件，位于这些方框内的过滤器就分别属于各自的部件。然后，我们允许这些部件进行平移，旋转，拉伸的小范围扰动。

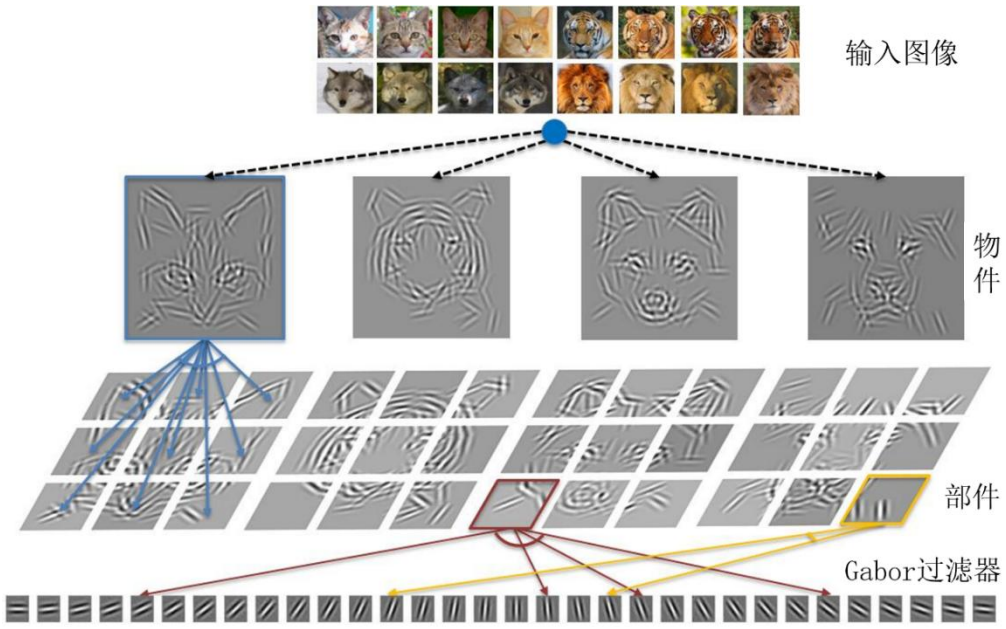


图 3-1 层次化稀疏 FRAME 随机场模型框架结构

图 3-1 展示了层次化稀疏 FRAME 随机场模型的框架，除去输入图像，共有三层。分别是物件层，部件层和过滤器层。每个部件由多个 Gabor 过滤器组成；而每个物件又由多个部件组成。

3.2 使用层次化稀疏 FRAME 随机场模型的推断算法

在阐述如何学习层次化稀疏 FRAME 随机场模型之前，我们先来描述下如何在已知模型的情况下推断物件所在位置。我们需要在后续层次化稀疏 FRAME 随机场模型的学习过程中用到这个算法。这个算法也使得无监督学习成为可能。

假设学习到的层次化稀疏 FRAME 随机场模型为 $H=\{BB^j, j=1,\dots,K\}$,其中，$BB^j=\{B_{x_i^j,s_i^j,\alpha_i^j}, i=1,\dots,n_j\}$为每个部件的过滤器集合。令输入图像为$\boldsymbol{I}$。

为了更直观的说明问题，我们引入了 SUM 图，MAX 图的概念。详细定义如算法所示。

1. Up-1 对所有位置，大小和方向的 Gabor 过滤器计算特征图 SUM1：
$$SUM1(x,s,\alpha)=\left|\left\langle I,B_{x,s,\alpha}\right\rangle\right|,\forall x,s,\alpha \tag{3-2}$$
2. Up-2 使用最大池化算法来考虑对所有 Gabor 过滤器可能的扰动，得出 MAX1：
$$MAX1(x,s,\alpha)=\max_{\Delta x,\Delta\alpha} SUM1(x+\Delta x,s,\alpha+\Delta\alpha),\forall x,s,\alpha \tag{3-3}$$
3. Up-3 对所有位置 X 和部件 j，计算各部件匹配得分矩阵 SUM2：
$$SUM2^j(X)=\sum_{i=1}^{N_j} MAX1\left(X+x_i^j,s_i^j,\alpha_i^j\right)-\log Z\left(\lambda^j\right),\forall j,X \tag{3-4}$$
4. Up-4 再次最大池化，考虑各部件的扰动，得出 MAX2：
$$MAX2^j(X)=\max_{\Delta X} SUM2^j(X+\Delta X),\forall j,X \tag{3-5}$$
5. Up-5 在每个位置 X，计算整个物件的匹配得分矩阵 SUM3：
$$SUM3(\mathrm{X})=\sum_{j=1}^{K} MAX2^j\left(\mathrm{X}+X_j\right),\forall \mathrm{X} \tag{3-6}$$
6. Up-6 取 SUM3 中最大的为最终的最优匹配得分 MAX3：
$$MAX3(X)=\max_X SUM3(X) \tag{3-7}$$
7. Down-1 取到 MAX3 的位置即为模板在图像上的位置：
$$\hat{X}=\arg\max_X SUM3(X) \tag{3-8}$$
8. Down-2 各部件在 MAX2 中取到的位置即为各部件在模板中扰动的相对位置：
$$\Delta X_j=\arg\max_{\Delta X} SUM2^j(\hat{X}+X_j+\Delta X),\forall j \tag{3-9}$$
9. Down-3 在计算 MAX1 中取到的扰动量即为各 Gabor 过滤器的扰动偏差：
$$\left(\Delta x_i^j,\Delta\alpha_i^j\right)=\arg\max_{\Delta x,\Delta\alpha} SUM1(\hat{X}+X_j+\Delta X_j+x_i^j+\Delta x,s_i^j,\alpha_i^j+\Delta\alpha),\forall i,j \tag{3-10}$$

若使用非层次化的稀疏 FRAME 随机场模型，推断部件的过程就可视为推断整个物件。只需去除上述算法的 Up-3 和 Up-4 以及 Down-2，并在 Up-5 中使用 MAX1 代替即可。

3.3 用多层可变形模型看待层次化稀疏 FRAME 随机场模型

我们可以把 H 视为一个可变形的模板，而它里面的每个部件本质上就是一个稀疏 FRAME 随机场模型，其中的过滤器选择方式，权重项的学习方法都是相同的。与此同时，若将部件模板这个过滤器集合本身当做是一个过滤器，用类似的方法合成了最终的物件模板。我们用同样使用稀疏编码模型的方式来描述我们的模型：

$$I_m = \sum_{j=1}^{K} C_{m,j} \boldsymbol{BB}^{j}_{X_j,S_j,A_j} + \epsilon_m \tag{3-11}$$

其中$C_{m,j} = (c_{m,i}, i = 1, \dots, n_j)$是第 j 个部件中一系列过滤器相对于第 m 张图片的参数向量; $\boldsymbol{BB}^{j}_{X_j,S_j,A_j} = (B_{X_j+x_i+\Delta x_i, S_j+s_j, A_j+\alpha_i+\Delta\alpha_i}, i = 1, \dots, n_j)$ 相当于之前过滤器的拓展，拓展成为了一个由过滤器集合组成的部件模板，它的位置在X_j，大小为S_j，角度为A_j (为了和同时存在的过滤器扰动区分，部件的相关下表运用了大写符号)。

因为我们允许部件级别的小范围扰动，所以我们加上扰动因子$(\Delta X_{m,j}, \Delta S_{m,j}, \Delta A_{m,j})$：

$$I_m = \sum_{j=1}^{K} c_{m,j} \boldsymbol{BB}^{j}_{X_j+\Delta X_{m,j}, S_j+\Delta S_{m,j}, A_j+\Delta A_{m,j}} + \epsilon_m \tag{3-12}$$

在本文中，我们对部件层面扰动的默认设定是：$\Delta X_{m,j} \in [-2,2] \times [-2,2]$ 像素，$\Delta S_{m,j} \in \{-0.1, 0, 0.1\}, \Delta A_{m,j} \in [-1,1] \times \pi/16$。

对于部件级别$\boldsymbol{BB}^{j}_{X_j,S_j,A_j}$的对数似然函数与稀疏 FRAME 随机场模型类似：

$$L\left(I_m | \boldsymbol{BB}^{j}_{X_j,S_j,A_j}; \lambda\right) = \sum_{i=1}^{n} \lambda_i \max_{\Delta x, \Delta\alpha} \left|\langle \boldsymbol{I}, B_{x_i+\Delta x, s_i, \alpha_i+\Delta\alpha} \rangle\right| - \log Z(\lambda) \tag{3-13}$$

我们假设各部件内的过滤器不会在扰动后出现重叠的情况，也即，各部件的过滤器 Gabor 函数是互相独立正交的。于是整个模板 H 相对于图片I_m的对数似然函数为：

$$L(I_m | H) = \sum_{j=1}^{K} \max_{\Delta X, \Delta S, \Delta A} L\left(I_m, | \boldsymbol{BB}^{j}_{X_j+\Delta X_{m,j}, S_j+\Delta S_{m,j}, A_j+\Delta A_{m,j}}\right) \tag{3-14}$$

3.4 层次化稀疏 FRAME 随机场模型的学习

现在我们来阐述层次化稀疏 FRAME 随机场模型的学习算法。

目标函数：我们需要通过训练图像$\{\boldsymbol{I}_m, m = 1, \dots, M\}$学习 K 个部件的模板$\{\boldsymbol{BB}^j, j = 1, \dots, K\}$，以及得出部件级别的扰动$(\Delta X_{m,j}, \Delta S_{m,j}, \Delta A_{m,j})$。需要最大化的目标函数就是所有训练图像相对于整个模板$\boldsymbol{H}$的对数似然函数：

$$\sum_{m=1}^{M}\sum_{j=1}^{K} L\left(I_m, |\boldsymbol{BB}^{j}_{X_j+\Delta X_{m,j},S_j+\Delta S_{m,j},A_j+\Delta A_{m,j}}\right) \tag{3-15}$$

为了求得这个函数的最大化，我们使用一个类 EM 的算法，将算法一分为二，多次迭代。

第一步：模型推断

假设层次化稀疏 FRAME 随机场模型$H=\{\boldsymbol{BB}^{j}_{X_j,S_j,A_j}, j=1,\dots,K\}$已经确定，我们需要通过推断过程推理物件模板的所在位置：

$$\hat{X}=\arg\max_{X}\sum_{j=1}^{K}\max_{\Delta X,\Delta S,\Delta A} L(I|\boldsymbol{BB}^{j}_{X+X_j+\Delta X,S_j+\Delta S,A_j+\Delta A}) \tag{3-16}$$

和其部件级别的扰动：

$$(\Delta X_j,\Delta S_j,\Delta A_j)=\arg\max_{\Delta X,\Delta S,\Delta A} L(I|\boldsymbol{BB}^{j}_{\hat{X}+X_j+\Delta X,S_j+\Delta S,A_j+\Delta A}) \tag{3-17}$$

使用上一节提到的推断算法就能高效的得出结果。

第二步：重新训练模型

这一步我们假设物件模板的位置给定，子部件的位置也已经确定，直接使用稀疏 FRAME 随机场模型的学习方法，继续学习模板即可。

3.5 层次化稀疏 FRAME 随机场模型可视化

层次化稀疏 FRAME 随机场模型是一个解释性极强的生成模型。在这一节我们将展示训练过程中层次化 FRAME 随机场模型的可视化输出。

图 3-2 展示了层次化的九个部件组合而成物件的模型可视化，图 3-3 展示了 Gabor 过滤器层面的模型可视化，两者都有 8 个模型，分别训练自八种不同的动物脸部图片，具体数据集参见章节 4-2-3。

图 3-4 展示了狮子这一类型的模板，左边三图是不同大小的 Gaobr 过滤器分别的图像，右边是三个大小结合的图像。

图 3-5 则展示了模板的可变形特性，十张图分别是十张训练图像，它们各自的变形后模板。

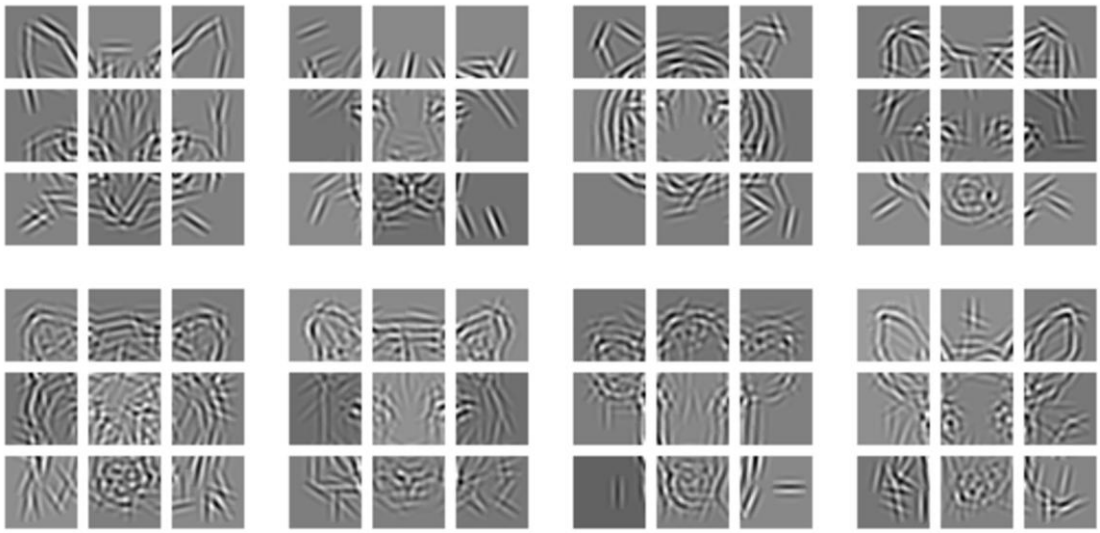


图 3-2 多个模板的层次化稀疏 FRAME 模型可视化

图 3-3 多个模板的 Gabor 过滤器部分可视化

图 3-4 层次化稀疏 FRAME 随机场模型模板可视化

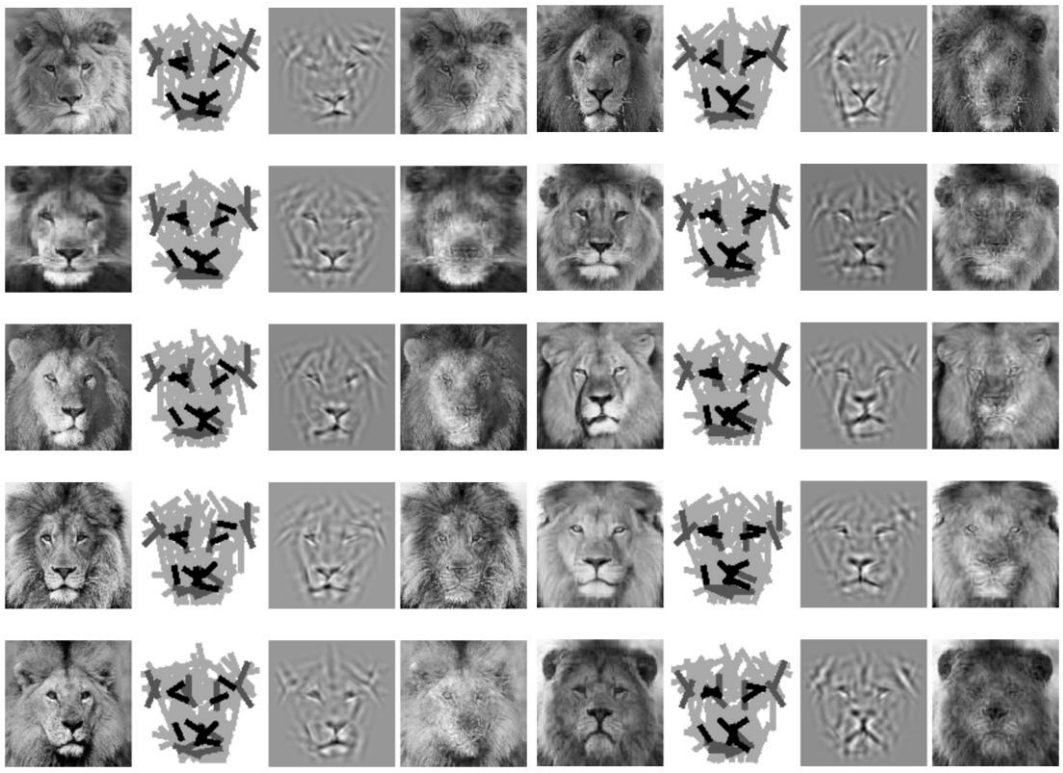


图 3-5 对训练图片变形后的模板可视化

3.6 层次化稀疏 FRAME 随机场模型与卷积神经网络的关系

一个有趣的事实是，我们的层次化稀疏 FRAME 随机场模型和卷积神经网（CNN）络有着很深的关系。它本质上可以被描述为一个三层稀疏链接的卷积神经网络，包含了复杂的最大池化操作，然后人为的将第一层变为我们使用的 Gabor 过滤器，最后通过生成式学习的方法学习它。其中，我们选择出的 Gabor 过滤器集合可以当成是 CNN 中第一层的过滤器，模型学习中涉及到的$\max_{\Delta x,\Delta \alpha}$可以相当于一个池化层，过滤器与图像的内积$\langle \boldsymbol{I}, B_{x_i,s_i,\alpha_i} \rangle$则对应卷基层，$-\log Z$则是偏差项。

所以，我们有可能可以进一步的拓展我们的层次化模型，甚至增大他的层数，提高模型的复杂性，以应对更大数据规模，更加通用化的训练问题。

第四章　实验设计与结果

在本章节中，我们使用三个精心设计的实验，通过纵向对比不同形式的 FRAME 随机场模型和横向对比多种其他生成式模型，展示了层次化稀疏 FRAME 模型的四大优势，可解释性，鲁棒性，无监督特性和极少数据需求特性。

第一个实验是物件检测和关键点预测，实验展示了层次化稀疏 FRAME 模型的鲁棒性和可解释性；第二个实验是聚类问题，突出了模型的无监督特性；第三个实验是分类问题，实验通过将模板作为一个特征提取方式，表明层次化稀疏 FRAME 模型模板的可解释性。

在第一章节，我们首先介绍了将要使用的对比模型；在后续章节，分别描述了实验。

4.1 实验对比模型介绍

4.1.1 And-Or 图算法

And-or 图算法[10]是一个无监督学习的随机可重构模板，他可以生成一组有效的对象模板。And-or 图中，and 节点指代部件的组成，而 or 节点则表示部件的关节和结构变化。

下图是一个简明的 and-or 图的例子，模型学习自 320 张动物图片，实心圆点代表 and 节点，空心的代表 or 节点，矩形的相当于叶子节点，他们是不同的动物脸部内的部件。

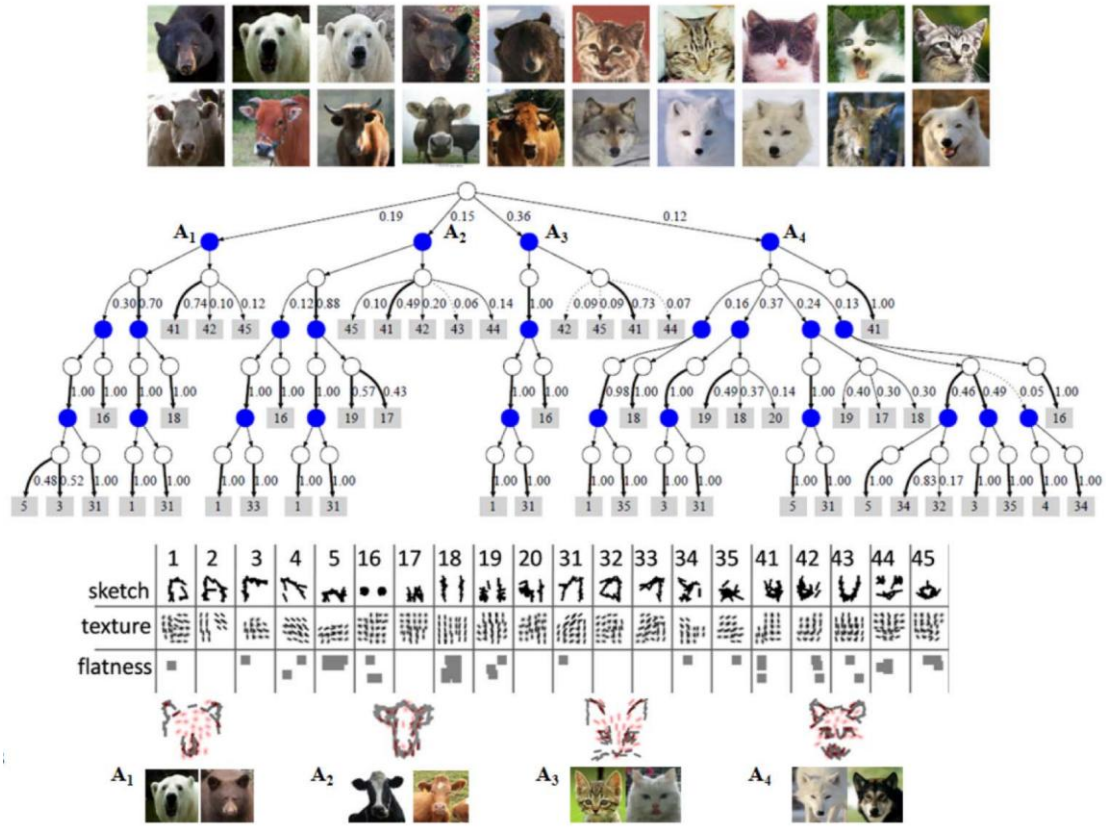


图 4-1 And-or 图算法的框架结构[10]

学习 And-or 图大体由两部分组成：

第一步：对可多次使用的多层词典进行区块追踪。首先，我们使用训练图片的特征相应来建立数据矩阵。数据矩阵的每一行对应一个训练样本，每一列对应一种特征（比如颜色特征，纹理特征等）。一般，数据矩阵的行数很少，但列数很大（百万量级）。数据矩阵被归一化后，响应为 1 的说明图像存在该特征，为 0 则说明不尊在。每个科室部分对应于数据矩阵中的一些矩阵块，即部分的特征，部分的样本。这些特征组成了块中元素的模板。某个元素出现次数越多，说明它越重要。寻找大块的 1 可以使用信息投影原理和最大似然估计。一旦找到它们，我们将更新数据矩阵，用新的部件响应代替这些列。这个步骤重复的进行，直到完整整个图的训练。值得注意的是，通过共同追求对象和零件，可以减少或消除零件的普遍模糊性（即，将对象分割成零件）。

第二步，在 and-or 图上进行图压缩。在区块追踪完毕后，我们将训练图像编码成一系列的部件，记录某部件是否出现，和出现的位置。这样的一个 and-or 图有很多的孩子节点，会有很多的分支，模型的复杂度过大。我们提出一个图压缩算法，有如下两个步骤：

1. 共享部件：例如$(A \cap B) \cup (A \cap C) \rightarrow A \cap (B \cup C)$
2. 合并部件：这个操作将一些分支概率相似的 or 节点合并成一个节点，并重新估计概率。

通过这两个操作，减少了模型复杂度，而损失极小。

4.1.2 混合伯努利模板

这是一个使用 EM 算法学习的混合伯努利模板[19]。模型使用二进制向量来表示图片的特征，其中每个二进制分量指示在图像域的特定小区内是否存在本地特征或结构。每个模板就是一个二进制向量，它通过以一定的概率独立地选择其二进制组件来生成示例。图（）说明了基本思想。图像被分成几个方块区域（图中为 9*9=81 个区域）。有一系列的基本元素轮廓字典可以填充到这些区域中（如图中左边的 18 个基本元素）。每个区域都可能包含一个或多个基本元素，所以对于这个例子来说，二进制向量中有 9*9*18 维，分别说明了在某某区域中是否有某某基本元素。我们使用局部边缘检测，Gabor 过滤器，小波变换或者预训练的分类器来检测某某基本元素是否在一个区域中。在训练这个混合模型时，我们使用了 EM 算法。

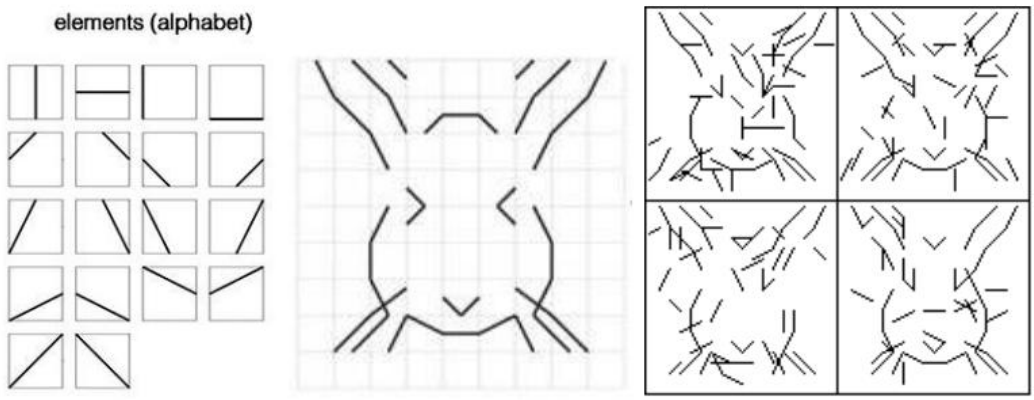


图 4-2 混合伯努利模板框架结构[19]

4.1.3 HOG 特征+K 中心聚类

我们将这个方法作为聚类实验的基准线。实现方法是首先抽取图片的 Hog 特征[20]，然后使用该特征进行 K 中心聚类。

HOG 特征是方向梯度直方图特征，是一种在计算机视觉和图像处理中进行检测的特征描述子。一幅图像中，局部的表征和形状能够被梯度信息或边缘的方向密度分布很好的描述。首先，将图像分成小的连通区域，称为细胞单元，然后采集细胞单元中各像素点的梯度或边缘方向直方图，最后将直方图组合起来成为特征。具体流程如下：

1. 标准化图片的 gamma 空间和颜色空间
2. 计算图片梯度，图片中像素点(x,y)的梯度值和其方向分别为：

$$G(x,y)=\sqrt{G_x(x,y)^2+G_y(x,y)^2}$$
$$\alpha(x,y)=\tan^{-1}\left(\frac{G_y(x,y)}{G_x(x,y)}\right) \tag{4-1}$$

   其中：$G_x(x,y)=H(x+1,y)-H(x-1,y);G_y(x,y)=H(x,y+1)-H(x,y-1)$分别是水平方向和垂直方向的梯度。$H(x,y)$是像素值。
3. 为每个细胞单元构建梯度方向直方图
4. 吧细胞单元组合成大的块，将块内的直方图归一化

最后我们将组成的特征向量进行 k 中心聚类。

K 中心聚类算法是聚类算法中最基本的算法，几乎是所有机器学习教材中的第一课。算法需要达成的目标是，最小化划分簇的平方误差：

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}||x-\mu_i||_2^2 \tag{4-2}$$

其中$\mu_i=\frac{1}{|C_i|}\sum_{x\in C_i}x$是第 i 类中样本的均值向量。直观来看，这个式子刻画了样本围绕其所在类的均值向量的紧密程度，E 越小则说明同类中的相似度更高。

聚类算法如下：

1. 随机选择 k 个样本作为初始均值向$\{\mu_1,\dots,\mu_k\}$。
2. 计算所有样本与 k 个均值向量间的距离，将样本放到均值向量最近的类中.
3. 对于每个类，更新均值向量为目前类中元素的均值
4. 若均值向量有更新，则重新进行第二步，否则算法结束。

4.1.4 可变形部件模型

可变形部件模型 DPM 算法[21]是一种基于部件的检测算法。DPM 本质是一个弹簧形变模型，部件与部件之间互相关联，而每个部件则是一个通过 HOG 特征训练而成的子模型。

HOG 特征即上一节介绍的方向梯度直方图，如下图所示，DPM 在训练动物脸部时，会自动找出若干个部件（在我们的设定中，是 8 个），每个部件是一个框，里面是该部件的

HOG 直方图，通过组合，合成了最终的 DPM 模型。

DPM 的训练使用了隐式 SVM 算法，对于样本 x，隐式 SVM 的目标函数是：

$$f_\beta(x) = \max_{z \in Z(x)} \beta\Phi(x,z) \tag{4-3}$$

其中，β是模型参数的向量，在 DPM 中，代表主过滤器的连接，部件过滤器和变型成本。z 是隐变量，在 DPM 中，代表物件配置，$\Phi(x,z)$是特征向量，包含子窗口的特征金字塔和部件变型特征。

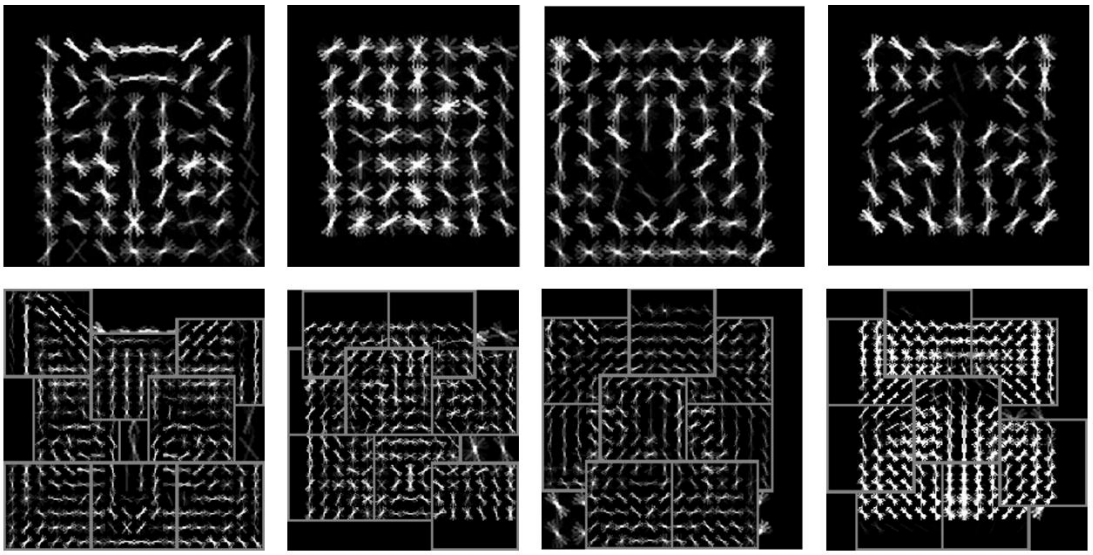


图 4-3 可变形部件模板可视化

4.2 物件检测与关键点预测实验

4.2.1 实验描述

物件检测是计算机视觉领域的一个基本问题。问题非常简单，给定一张图片，找出物件的所在位置。为了更好的表现出模型对物件中部件的变动感知能力，我们引入了关键点预测实验。关键点预测类似人脸识别领域的人脸对齐，给定训练图像的关键点位置，我们要求程序在输出物件位置，大小和角度的同时，预测关键点的位置。

更一般的，对于训练阶段，算法的输入是：

一系列训练图片，他们的位置，大小和角度，以及每张图上一系列关键点的位置。

对于测试阶段，算法的输入是：

一系列的测试图片。图片没有任何标签。

测试阶段算法需要输出的是：

物件所在的位置，大小和角度，以及每个关键点的预测方位。

4.2.2 实验评价方式

对于物件检测框图的评价，我们使用了

$$UoT = \frac{S_{predicted} \cap S_{truth}}{S_{predicted} \cup S_{truth}} \tag{4-4}$$

其中$S_{predicted}$指预测的框选面积，S_{truth}为真实的框选面积。

对于关键点预测的评价，我们首先定义每个关键点在一幅图上的均一化偏移：

$$p_{ij} = \frac{\sqrt{(x_{pre} - x_{gt})^2 + (y_{pre} - y_{gt})^2}}{width} \tag{4-5}$$

其中，width 指当前预测的物件框的宽度。值得注意的是，若算法找不到物件，则定义p_{ij}无限大。

得出均一化偏移后，我们可以画出关于某个关键点的准确率-召回率曲线。我们计算该曲线的曲线下面积 AUC 作为最终的评价标准。在评价时，我们有三种评价方式，第一种是基于单个关键点，结果将每个关键点的准确率-召回率曲线的曲线下面积取平均；第二种是基于部件，结果将人为设定的部件所包含的关键点的均一化偏移求和后，绘制出该部件的准确率-召回率曲线，求其曲线下面积后取平均；第三种是基于整个物件，结果将所有关键点的均一化偏移求和，绘制出整个部件的准确率-召回率曲线，求其曲线下面积。显然，曲线下面积越大，预测精度越高，面积是一个 0-1 的值。

4.2.3 实验数据集

我们选择了动物面部监测数据集，共八个类的动物，每个类别包括 10 个训练图像和 30 个测试图像。对于每个图像，人工标注了大约二十个关键点的位置，包括左右耳各 5 个，左右眼各四个，嘴巴鼻子和下巴有 2-6 个不等的关键点，标注的例子如下图所示：

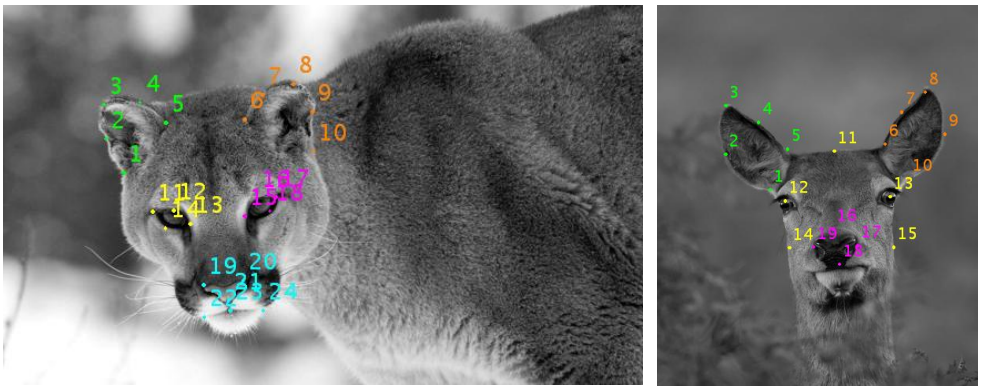


图 4-4 关键点检测实验标注示例

4.2.4 FRAME 随机场模型的实验算法

我们使用在第三章第二节阐述的推断算法，来进行物件检测。推断算法可以得出物件的位置，旋转角度和大小，划出框图。

对于关键点预测，我们首先将关键点与得出的模型模板中的 Gabor 过滤器建立联系。对于第 i 个关键点第 j 张训练图片，我们找到离他最近的过滤器B_{ij}，并记录关键点到该 Gabor 过滤器的相对位置(x_{ij}, y_{ij})。对于每个关键点，多张训练图片会找到多个过滤器，可能不同也可能相同，我们选择他们的众数$mode(B_{ij}|j = 1, \dots, N)$ 作为选定的过滤器，而记录的相对位置为所有训练图片相对于这个过滤器取平均$(\sum_{j=1}^{N} x_{ij}/N, \sum_{j=1}^{N} y_{ij}/N)$。在预测测试图片的关键点位置时，找到这些过滤器（值得注意的是，过滤器的位置会经过扰动平移，所以每张测试图片的过滤器位置都是不同的），叠加之间记录的相对位置平移，即为算法预测的关键点。

4.2.5 实验结果

我们首先展示一些物件检测得到的实例，其中红色的框代表程序预测得到的物件位置；蓝色的框是其中部件的位置。图片三行为一组，从上至下分别是层次化稀疏 FRAME 随机场模型，And-or 图模型，可变形部件模型 DPM 的结果。

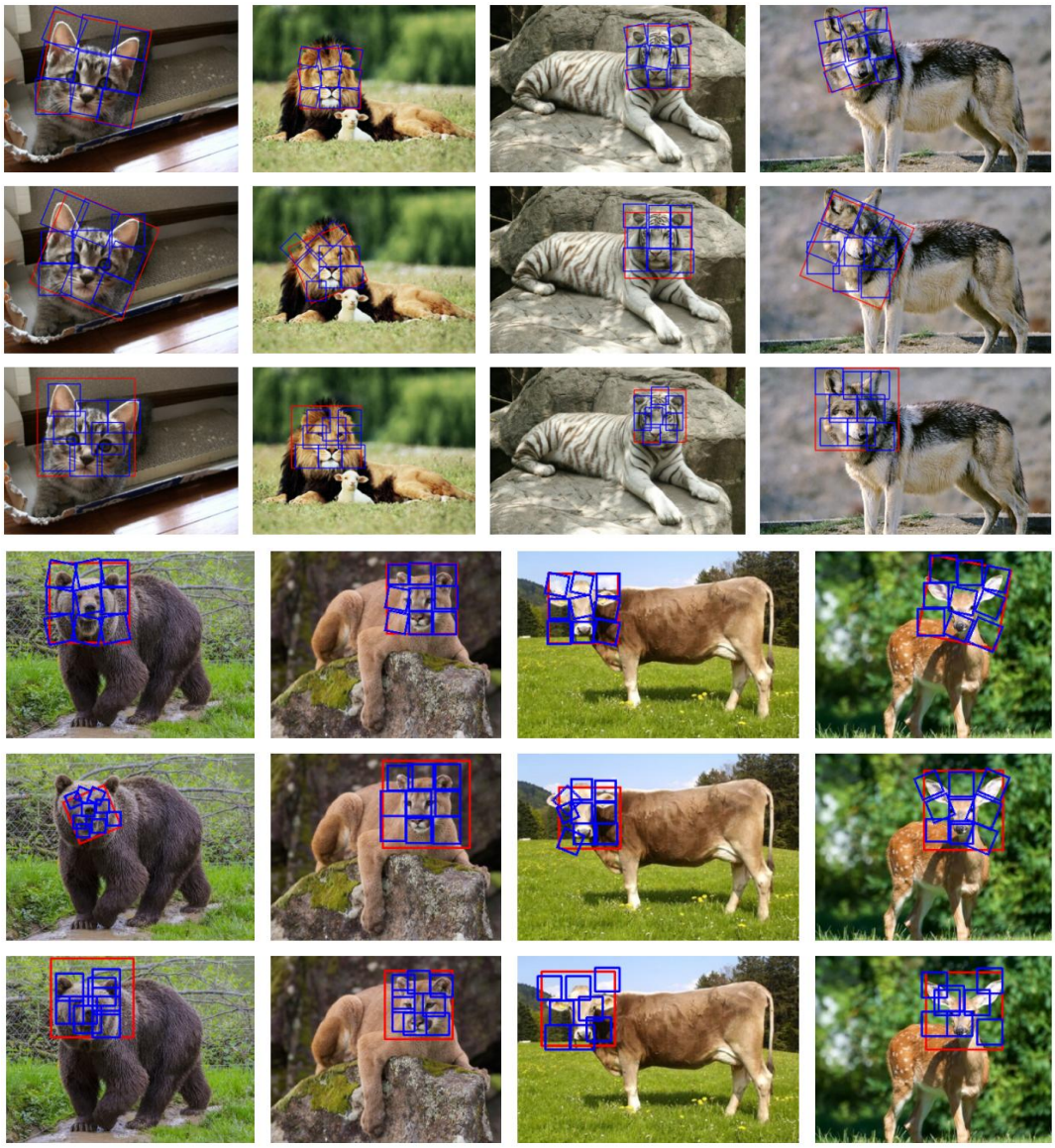


图 4-5 物件检测结果精选

然后我们展示一些比较困难的，其他算法预测失败的例子，同样三行为一组，从上至下分别是层次化稀疏 FRAME 随机场模型，And-or 图模型，可变形部件模型 DPM 的结果：

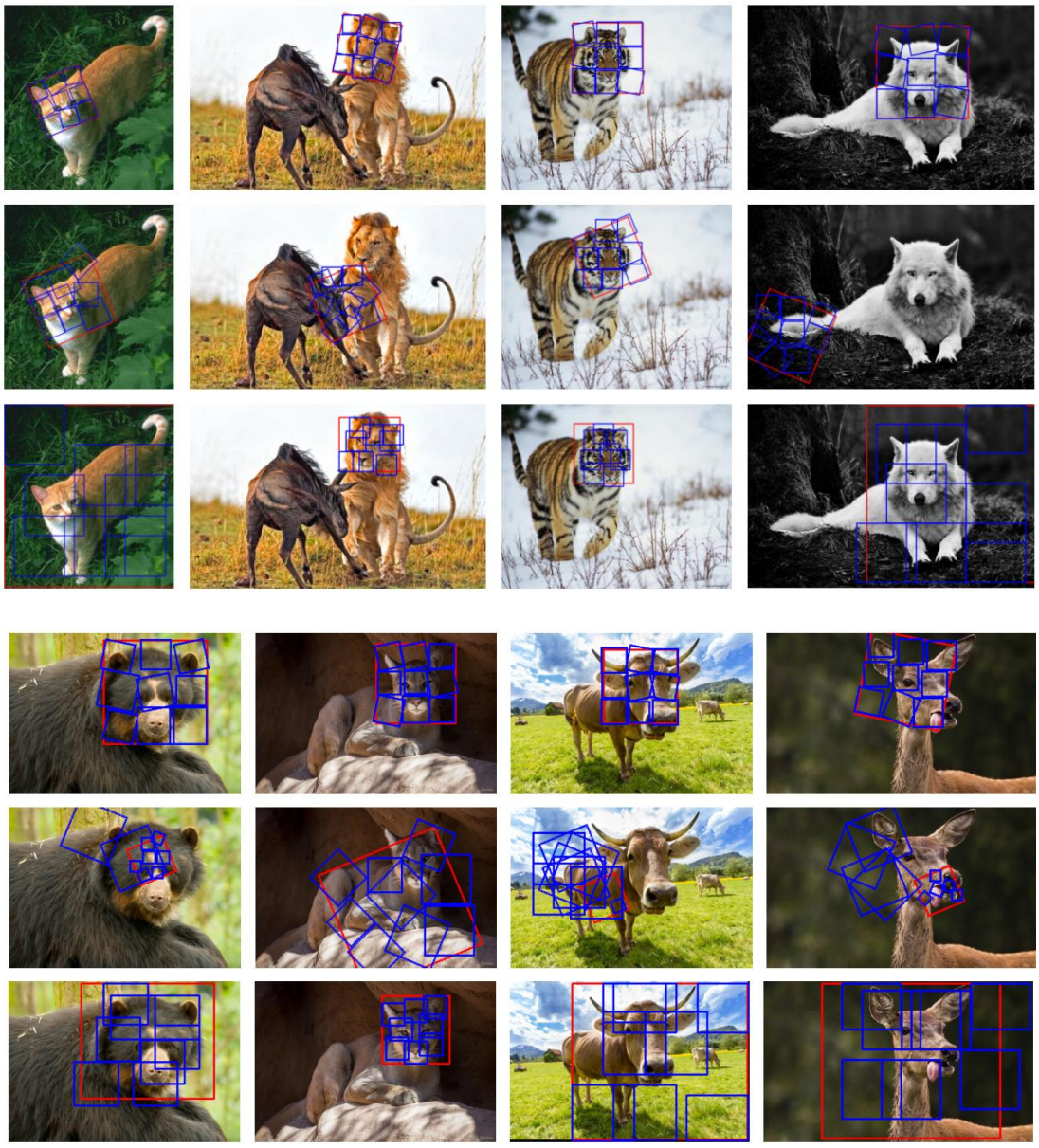


图 4-6 物件检测结果精选续

最后展示一下关键点预测的结果，在这里我们都选择了三种算法都成功检测出的结果，每行为一组，从左至右分别是，关键点的真实位置，层次化稀疏 FRAME 随机场模型，And-or 图模型，可变形部件模型 DPM 的预测结果：（为了更好的展示点的位置，部分图片经过了裁剪）

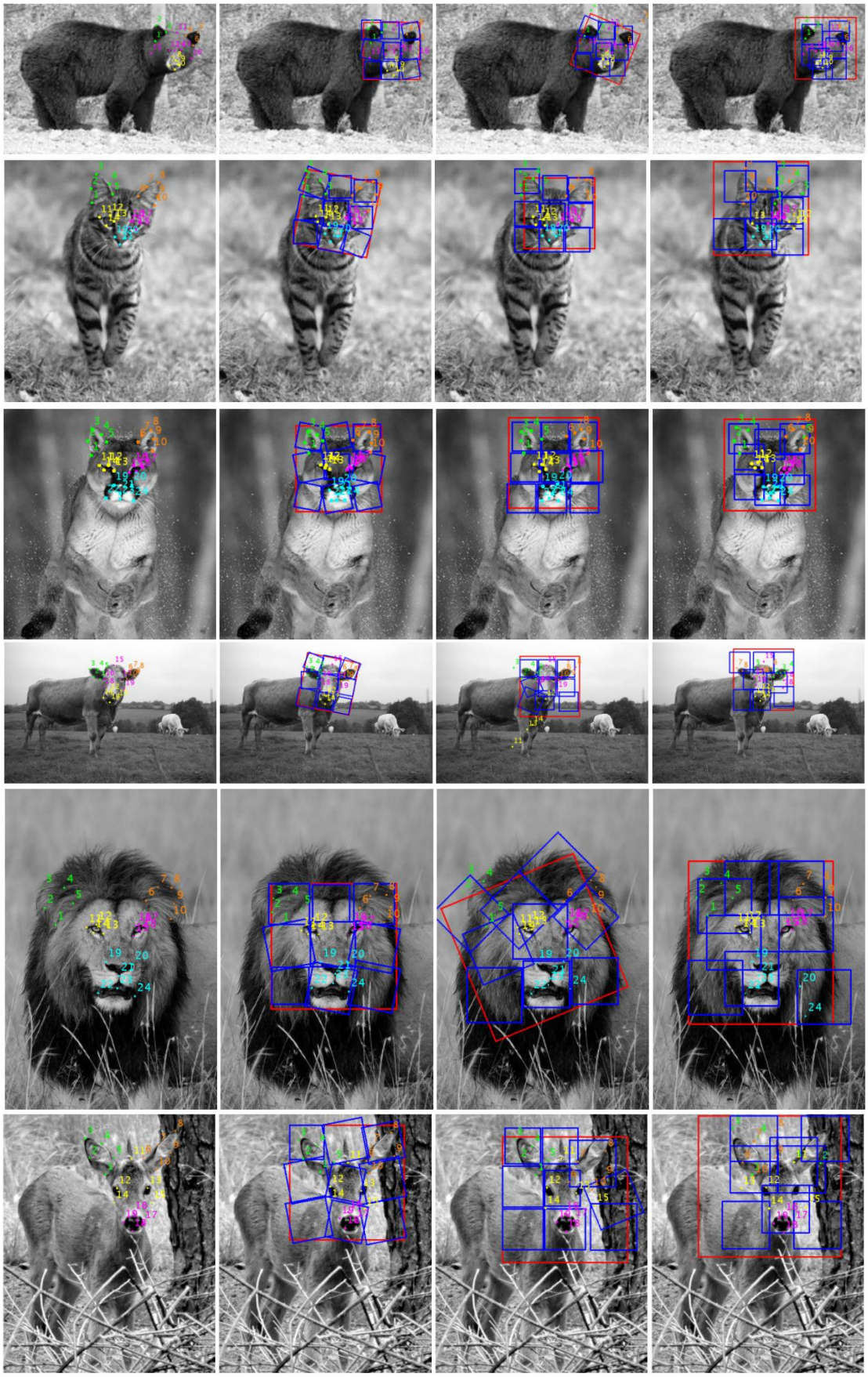


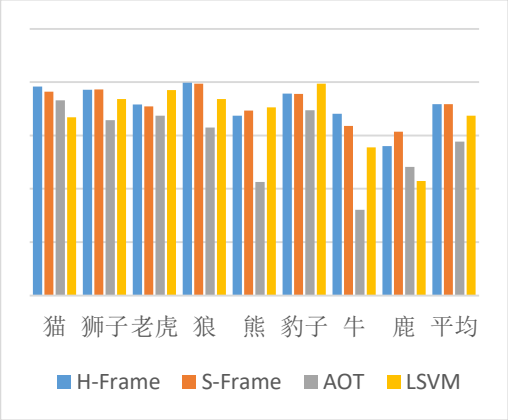
图 4-7 关键点预测结果展示

下面展示物件检测的数值结果，表中和图中的简称分别代表：

1. H-Frame：层次化稀疏 FRAME 随机场模型
2. S-Frame：稀疏 FRAME 随机场模型
3. LSVM：可变型部件模型 DPM 模型（参见 4.1.3）
4. AOT：And-or 图算法（参见 4.1.1）

表 4-1 物件检测 UOT 表

	HFrame	SFrame	AOT	LSVM
猫	**78.41%**	76.45%	73.22%	66.82%
狮子	77.21%	**77.23%**	65.74%	73.68%
老虎	**71.60%**	70.89%	67.46%	76.98%
狼	**79.81%**	79.46%	63.04%	73.66%
熊	67.47%	**69.40%**	42.55%	70.53%
豹子	**75.67%**	75.57%	69.45%	79.47%
牛	**68.15%**	63.60%	32.11%	55.58%
鹿	56.02%	**61.44%**	48.17%	42.88%
平均	**71.79%**	**71.75%**	**57.72%**	**67.45%**



下表展示了关键点实验的数值结果和曲线，首先是基于单个关键点，结果将每个关键点的准确率-召回率曲线的曲线下面积取平均：

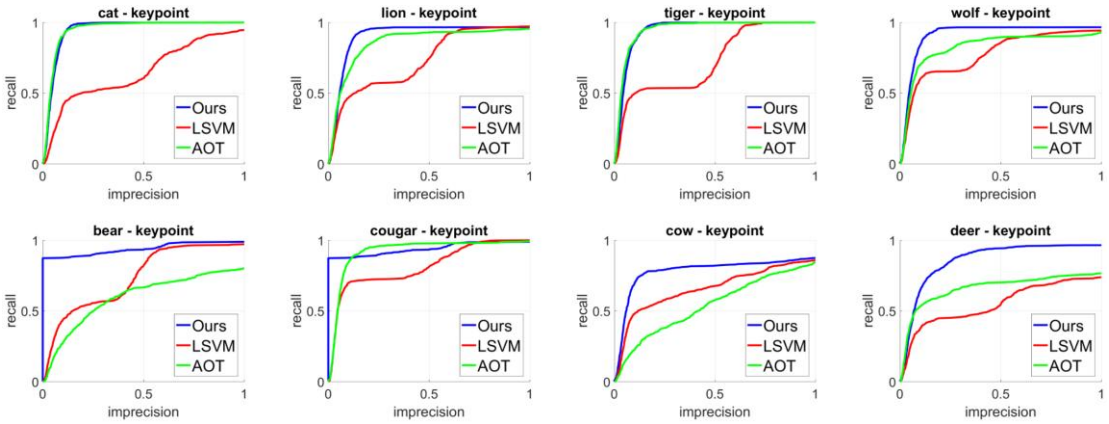
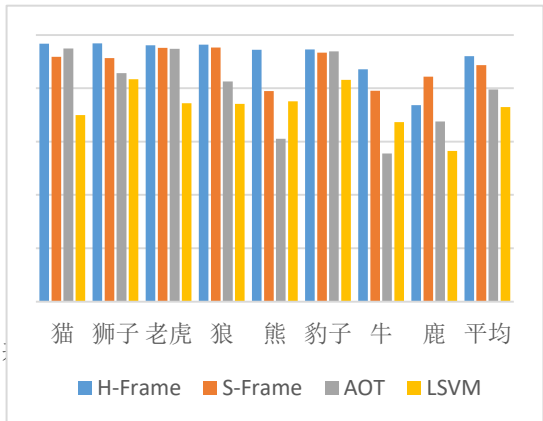


图 4-8 关键点检测点精度均值的准确率-召回率曲线

表 4-2 关键点检测点精度均值的准确率-召回率曲线下面积率和柱状图

	HFrame	SFrame	AOT	LSVM
猫	96.7%	91.8%	94.9%	70.0%
狮子	96.8%	91.3%	85.7%	83.4%
老虎	96.1%	95.1%	94.8%	74.4%
狼	96.4%	95.3%	82.5%	74.1%
熊	94.4%	78.9%	61.1%	75.1%
豹子	94.5%	93.4%	93.8%	83.1%



牛	87.1%	79.1%	55.6%	67.3%
鹿	73.7%	84.4%	67.6%	56.5%
平均	**92.0%**	**88.7%**	**79.5%**	**73.0%**

其次是基于部件，结果将人为设定的部件所包含的关键点的均一化偏移求和后，绘制出该部件的准确率-召回率曲线，求其曲线下面积后取平均：

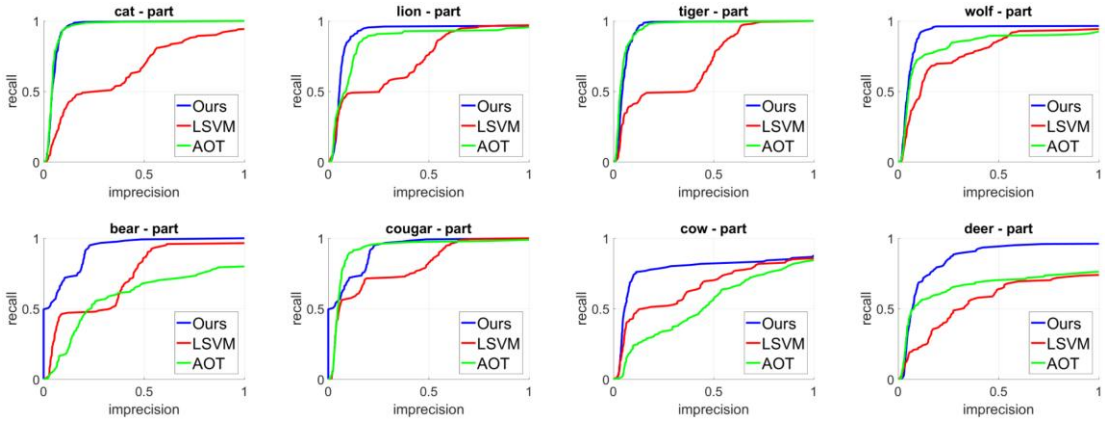
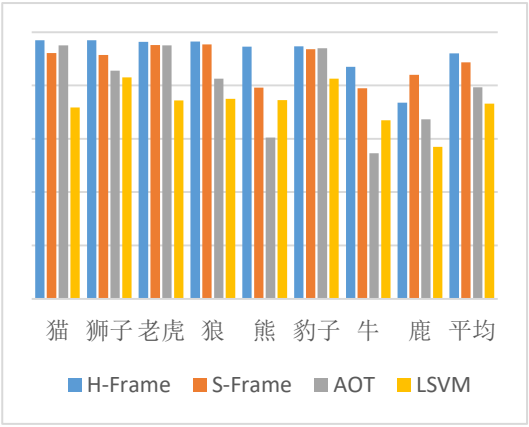


图 4-9 关键点检测部件精度均值的准确率-召回率曲线

表 4-3 关键点检测部件精度均值的准确率-召回率曲线下面积率和柱状图

	HFrame	SFrame	AOT	LSVM
猫	96.9%	92.1%	95.0%	71.8%
狮子	96.9%	91.4%	85.6%	83.0%
老虎	96.3%	95.2%	95.0%	74.4%
狼	96.5%	95.4%	82.6%	75.0%
熊	94.5%	79.2%	60.5%	74.5%
豹子	94.7%	93.6%	93.9%	82.5%
牛	87.0%	78.9%	54.6%	67.0%
鹿	73.5%	84.0%	67.3%	57.0%
平均	**92.0%**	**88.7%**	**79.3%**	**73.2%**



最后是基于整个物件，结果将所有关键点的均一化偏移求和，绘制出整个部件的准确率-召回率曲线：

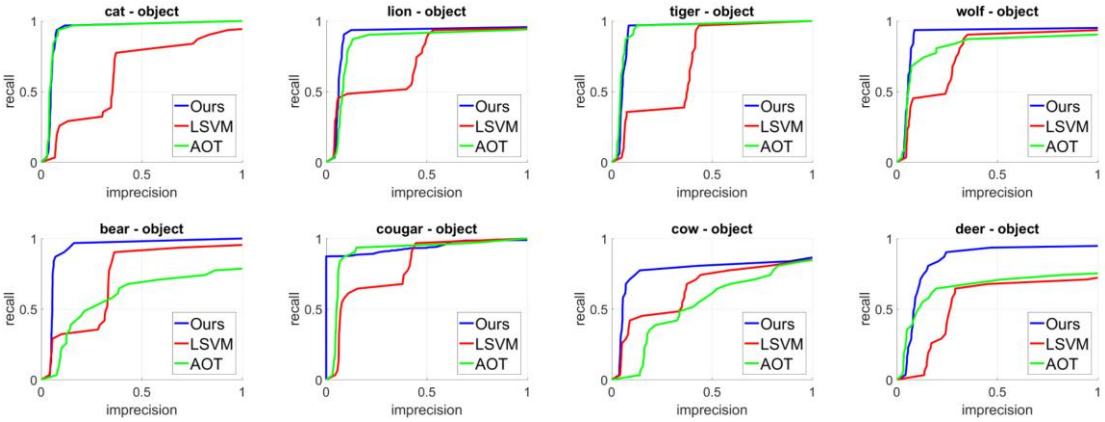
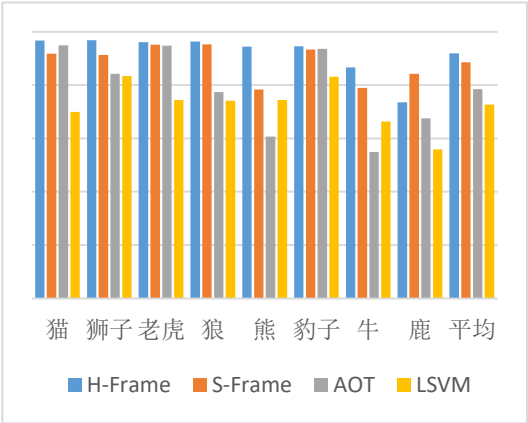


图 4-10 关键点检测物件精度的准确率-召回率曲线

表 4-4 关键点检测物件精度的准确率-召回率曲线下面积率和柱状图

	HFrame	SFrame	AOT	LSVM
猫	96.7%	91.8%	94.9%	70.0%
狮子	96.8%	91.3%	84.2%	83.4%
老虎	96.1%	95.1%	94.8%	74.4%
狼	96.4%	95.3%	77.4%	74.1%
熊	94.4%	78.4%	60.7%	74.4%
豹子	94.5%	93.4%	93.6%	83.1%
牛	86.6%	78.9%	54.9%	66.3%
鹿	73.6%	84.2%	67.5%	55.9%
平均	**91.9%**	**88.6%**	**78.5%**	**72.7%**



4.2.6 实验结果分析与评价

虽然本实验的数据集比较小，但因为是生成式模型的原因，数据集的大小并不影响模型的发挥。从物件检测的效果上来看，两种 FRAME 随机场模型各有优劣，而显著好于其他两种算法。从关键点检测的效果来看，可以看出层次化稀疏 FRAME 随机场模型相较于稀疏 FRAME 随机场模型以及其他两种算法，有着很好的优势。

4.3 聚类分析实验

4.3.1 实验描述

聚类问题是无监督学习领域的一个基本问题。给定一系列数据，在没有其他标签的情况下，需要通过寻找数据间的内在联系，将数据聚类成给定个数个类。

4.3.2 实验数据集

我们使用了 12 个不同的数据集，里面的图片数量分别从 30 到 75 张不等，我们设定的

聚类类别数量也从 2 到 5 类不等，具体数据集如下：

表 **4-5 聚类数据集包含的内容**

数据集编号	类别数量	类型描述	图片数量
1	2	公牛和奶牛	30
2	2	茶杯和茶壶	30
3	2	飞机和直升机	30
4	3	大象，鹿和骆驼	45
5	3	不同形状的钟表	45
6	3	天鹅，老鹰和海鸥	45
7	4	眼睛鼻子耳朵和嘴	60
8	4	不同形状的花	60
9	4	电脑配件	60
10	5	音乐乐器	75
11	5	鹿，猫，狼，老虎和狮子	75
12	5	老虎，豹，斑马，狗和牛	75

4.3.3 实验评价方式

每个数据集中的图片均由人工标注的属于哪个类的标签，标签仅在评价时使用。

对于每个数据集每个算法，我们计算结果的条件精度和条件熵，分别定义如下：

$$\begin{cases} purity(dataset, method) = \sum_{y} p(y) \max_{x} p(x|y) \\ entropy(dataset, method) = \sum_{y} p(y) \sum_{x} -p(x|y) \log(p(x|y)) \end{cases} \tag{4-6}$$

其中，x 是我们标注的标签，y 是算法预测的类别标签。条件精度和条件熵都是 0-1 之间的数，条件精度越大越好，条件熵越少越好。

4.3.4 FRAME 随机场模型的实验算法和参数设定

在第三章中，我们提出了 FRAME 随机场模型的推理算法，可以无监督的从一系列训练图片中训练处模型模板。在聚类问题中，我们一开始随机的将数据分为所需聚类的数量类，然后在这几组数据中分别训练出一个 FRAME 模型模板。例如若需要聚类成五类，则训练五个模型模板。

得到模板后，将所有的数据和所有模板运行推理算法，得到 MAX3 的值，也即该图片属于该模型的分值大小，对于每张图片，选择值最大为当前回合图片的预测类型。然后我们将同样预测类型的图片组成一个新的类，重新使用训练一个 FRAME 模型模板。

由此循环往复，经过 k 回合后（实验证明 k=10 左右时能保持稳定），就能完成聚类，同时得到每个类别的模型模板。

这个算法在实践过程中，会出现初始值敏感的问题，也即一开始随机分类的质量对后续的结果影响很大。在这里，我们通过多初始值并行运算的方法解决了这个问题，我们多次随机，多次运行，然后选取置信度最高的结果作为输出。在运行推理算法时，我们会得

到一个 MAX3 的值来量化算法对答案的置信度，MAX3 的值在不同模型间是可比较的，在完成聚类后，我们将所有图片与其对应预测类型的模型的 MAX3 值求和，选择多个初始值中最大的那个作为算法的最终输出。

下图是每次迭代结束后四类层次化稀疏 FRAME 随机场模型的模型可视化，这个分类是电脑配件数据集，包含主机，键盘，鼠标和显示器四类。

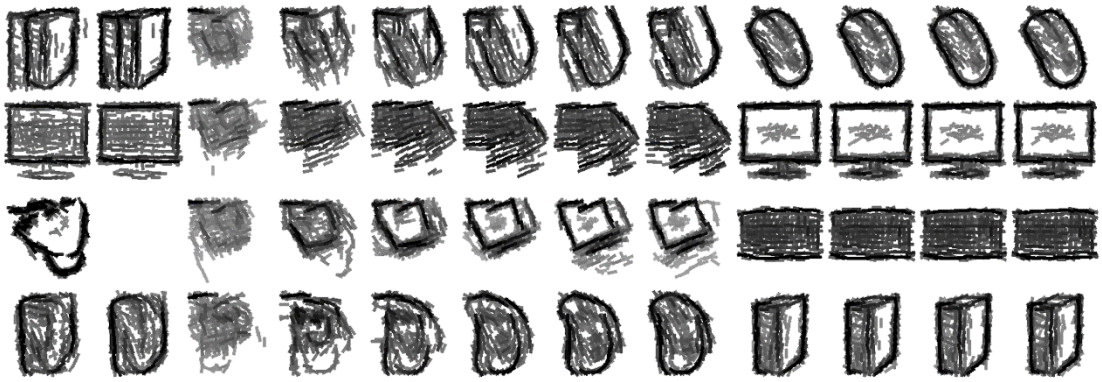


图 4-11 聚类算法每次迭代的可视化输出

下图是二分类问题飞机和直升飞机的模型可视化：

图 4-12 聚类算法每次迭代的可视化输出 2

下图展示了另外两个典型的数据类中完成聚类后的模板：

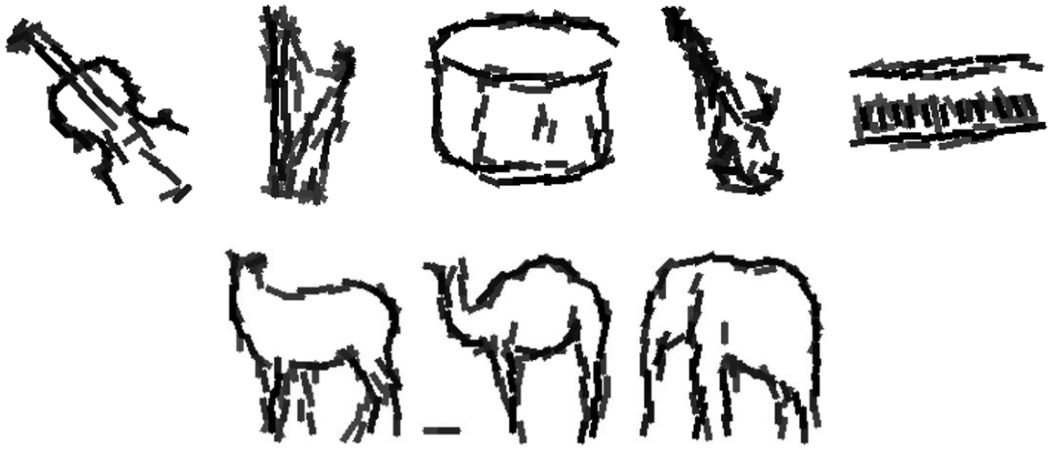


图 4-12 聚类算法结束后的 FRAME 模板

4.3.5 实验结果

表 **4-6 聚类问题条件精度数据表**

数据集	H-Frame	S-Frame	I-Frame	AB	BT-EM	HOG	AOT
1	**0.993**	0.967	0.887	0.667	0.873	0.76	0.813
2	**0.993**	0.98	0.907	0.787	0.82	0.64	0.773
3	**0.993**	0.96	0.973	0.96	0.713	0.793	0.907
4	**0.92**	0.907	**0.92**	0.729	0.72	0.8	0.876
5	**0.996**	0.987	0.982	0.658	0.858	0.84	0.849
6	1	1	1	0.836	0.8	0.933	1
7	**0.92**	0.917	0.85	0.83	0.773	0.807	0.83
8	**0.993**	0.953	0.92	0.903	0.73	0.78	0.77
9	**0.96**	0.893	0.953	0.923	0.85	0.84	0.88
10	**0.907**	0.797	0.883	0.797	0.869	0.715	0.824
11	**0.96**	0.872	0.923	0.888	0.757	0.784	**0.96**
12	**0.909**	0.907	0.88	0.805	0.813	0.768	0.712
平均	**0.962**	0.928	0.923	0.815	0.798	0.788	0.849

表 **4-7 聚类问题条件熵数据表**

数据集	H-Frame	S-Frame	I-Frame	AB	BT-EM	HOG	AOT
1	**0.025**	0.123	0.213	0.585	0.345	0.479	0.371
2	**0.025**	0.165	0.246	0.453	0.404	0.636	0.425
3	**0.025**	0.25	0.082	0.139	0.53	0.434	0.192
4	**0.17**	0.202	0.177	0.594	0.594	0.491	0.305
5	**0.017**	0.05	0.067	0.658	0.302	0.333	0.365
6	**0**	**0**	**0**	0.26	0.355	0.092	**0**
7	**0.14**	0.15	0.208	0.321	0.421	0.272	0.313
8	**0.025**	0.118	0.163	0.176	0.552	0.519	0.346
9	0.106	0.191	**0.067**	0.169	0.28	0.265	0.216
10	**0.22**	0.425	0.286	0.447	0.301	0.516	0.359
11	**0.055**	0.191	0.112	0.225	0.486	0.387	0.064
12	**0.189**	0.222	0.29	0.354	0.459	0.477	0.543
平均	**0.083**	0.174	0.159	0.365	0.419	0.408	0.291

其中表中的算法简称分别代表：

H-Frame：层次化稀疏 FRAME 随机场模型

S-Frame：稀疏 FRAME 随机场模型

I-Frame：稀疏 FRAME 随机场模型的另外一种训练方式

AB：动态基模型[9]（与 FRAME 极其类似一个模型）

BT-EM：使用 EM 算法训练的伯努利模板（参见 4.1.2）

HOG：抽取 HOG 特征后进行 k 中心聚类算法（参见 4.1.3）

AOT：And-or 图算法（参见 4.1.1）

4.3.6 实验结果分析与评价

聚类问题，作为机器学习的基本问题一直是一个老大难问题。虽然这个实验使用的数据量并不大，但数据的变化程度是挺大的，所以聚类难度不小，我认为能达到这样的精度也是非常可喜的。这个实验充分说明，模型在无监督学习方面有着无可比拟的优势，模型非常善于在混乱的无标签数据中找出我们需要的信息进行建模，认识并理解自然界中的种种物体模式。

4.4 类型分类实验

4.4.1 实验描述

分类问题是机器学习领域最基础的问题。几乎所有教材和课程都将分类问题作为机器学习的第一个例子进行讲解。目前计算机视觉领域有着很多能够很好处理分类问题的算法，其中最厉害的当属卷积神经网络 CNN，许多的大型企业早已实现商业化这些算法。我们设计这个实验的目的并不是超越这些成熟的模型。一方面我们的模型本身并不支持超大规模的数据量训练，如果要公平的与卷积神经网络为代表的一系列深度神经网络算法进行对比，只给几十上百个训练数据进行训练，神经网络根本无法学习，几乎可以肯定会出现过拟合的现象，实验结果惨不忍睹也毫无意义；另一方面，我们的 FRAME 随机场模型属于生成式模型，而深度神经网络均属于判别式模型，从原理上就极为不同，没有任何比较的意义。

我们设计这个实验的目的是为了说明 FRAME 随机场模型学习得出的模型模板不仅有着良好的可解释性，更有着鲜明的统计意义。层次化稀疏 FRAME 随机场模型可以被用于无监督的学习分层特征。也即，我们把我们的模型视为一个只需要极少数据需求就可以学习的特征提取器。顶层特征监督学习得出的分类器可以用作分类，这就是我们设计实验的目的。

4.4.2 实验评价方式

分类问题的评价方式非常简单，通过公平的方式抽取特征后，使用相同的 SVM 算法进行分类，对比分类的精度，精度定义如下：

$$accuracy(method) = N_{right}/N_{total} \tag{4-7}$$

其中，N_{right}是正确的测试样本个数，N_{total}是测试样本总数。

4.4.3 实验数据集

我们使用了 LHI 的动物图片数据库，拥有大约两千两百张图片，被分为了 20 个不同的类。每个类型都有丰富的变化，包括翻转，旋转，扰动，甚至是类别下有子分类的明显不同。我们随机的将这些图片分为相同的两部分，一半用于训练，一半用于测试。

二十个分类和数据情况如下表：

表 **4-8 分类数据集包含的内容**

类别	数量	类别	数量	类别	数量	类别	数量	类别	数量
熊	102	猫	161	鸡	100	牛	104	鹿	103
狗	391	鸭	103	鹰	101	象	100	人	100
狮	102	猴	100	鼠	100	熊猫	119	鸽	115
猪	101	兔	100	羊	100	虎	114	狼	100

4.4.4 FRAME 随机场模型的实验算法

对于每个类别，我们使用和上一个聚类模型相同的方法，将每个类别中的训练数据聚类成 5 或 11 类，训练 5 或 11 个模型模板。选择训练这些数量是因为有些类别中不同图片有着极大的不同，例如公母区别，有毛的没毛的，训练多个模型模板可以帮助模型适配不同的子分类，使抽取的特征更加适合图片本身。

然后，我们将每个类别的 5 个或 11 个模型模板，共 20 个类，100 个或 220 个模板组成一个模板集合。在实验过程中，我们曾经尝试过偶数数量的模板，发现模型显著的差于奇数数量，我们也曾经尝试过 3-15 不同的个数发现，5-11 这个区间内的效果较好，故选择了这两个数字作为模型模板的数量。

在抽取图像特征时，对于每个模板，计算内积得到 SUM 图，然后将这些矩阵进行统计参数映射 SPM 算法。SPM 算法将 SUM 图分为 1,4,16 个区域，然后将各自区域的最大值作为最终的特征。

得到所有模板，所有区域的特征后，将这些特征全部送入 SVM 算法中进行训练。在 SVM 时，我们选用了三种 SVM 的方式，分别是逻辑回归，$\ell 1-norm$和$\ell 2-norm$。

4.4.5 实验结果

下表是分类精度，三种方法分别对应的是：

H-Frame：层次化稀疏 FRAME 随机场模型

S-Frame：稀疏 FRAME 随机场模型

AOT：And-or 图模板模型（参见 4.1.1）

表 **4-9 分类实验结果展示**

模板数	SVM 方法	H-Frame	S-Frame	AOT
5	逻辑回归	**74.41%**	71.13%	64.80%
5	$\ell 1-norm$	**74.32%**	70.46%	66.13%
5	$\ell 2-norm$	**74.33%**	70.62%	65.80%
11	逻辑回归	**75.50%**	72.73%	60.96%
11	$\ell 1-norm$	**75.59%**	71.38%	61.07%
11	$\ell 2-norm$	**75.83%**	72.56%	62.54%

平均	**75.00%**	**71.48%**	**63.55%**

下表列出了层次化 FRAME 随机场模型在使用 5 个模板逻辑回归的算法下，各分类的分类成功率。斜对角线的绿色框是各分类的正确率，黄色的数字指的是数据属于左边行名类型，但却被预测成了下面列名的类型的百分比。最后一行的总计指的是有百分之多少的数据被预测成为了这个类别。例如狗的总计是 178%，说明 96%的狗预测成功了，另外有 82%的不属于狗但却预测成为狗的数据存在。通过可视化这个表格，可以看出，模型在 20 个不同的分类中的效果不一，有部分极其成功，也有部分略有欠缺。

表 **4-10** **二十类各类分类成功率**

	熊	猫	鸡	牛	鹿	狗	鸭	鹰	象	人	狮	猴	鼠	熊猫	鸽	猪	兔	羊	虎	狼
熊	80					2					2	6				2				8
猫		91				3	1			1		1	1				1			
鸡			84			2		2			2	2			2		4	2		
牛		2	2	67			2		2	4	2			2		2	2	12	2	
鹿				2	96												2			
狗	3	1	1	1	2	86		1	1	1	2	2		1			1	1	1	1
鸭	2	2				8	75	4		2		2	2		4					
鹰	6		8			2	4	64							16					
象						12		2	82				2		2					
人							2			94		2						2		
狮		2				2		2	2		82		4				2		4	
猴	4		4			14		4			2	62	2		4	2	2			
鼠		4				2						2	84				2	2	2	2
熊猫						2								95	3					
鸽	4					4		2			2				86		2	2		
猪				2		26		2	2		2		8	2		38	10	2	2	4
兔		2				4		2	2		2		6	2		2	76			
羊	4			22	8		2		2	2			2			10		48		
虎		4		2		5			2										87	
狼	10					4							2				2			82
总计	113	108	99	96	106	178	86	85	95	104	98	79	113	102	117	56	106	71	98	97

4.4.6 实验结果分析与评价

对 20 类的复杂图片进行分类，能够达到 75%左右的正确率是一件不容易的事，特别是在训练图片如此少的情况下。这个实验说明，层次化稀疏 FRAME 随机场模型学习出的模板不但解释性强，对图片信息的包含也非常的准确。在实验过程中，关于参数的选择对结果印象也很大，这个实验只经过了很微小的调试，没有修改很多的参数，所以目前的结果可以说远远不是模型的最终水平。

4.5 实验总结

4.5.1 实验难点分析

在层次化 FRAME 随机场模型的实现和改良过程中，出现了许多的问题也克服了许多的问题，在此我简单说明一下遇到的几个主要问题和目前的解决方法。

问题一，关于三大实验的数据集大小的问题。三个实验的数据集都只有大约一百左右的量级，即使作为生成式模型的实验，也是略有不足的。所有的数据都取自之前某些论文的小型数据集再加上了自己的拓展，并没有应用目前领域内通用的数据库例如 imageNet，pascal 或者 celeba，原因是这些数据库数据量过大，内部图片的多样性过于丰富，而它们的标签却对我们的训练没什么用处。结果上来看，我们的层次化稀疏 FRAME 随机场模型很好的完成了学习，测试得到了较为良好的结果。但是需要指出的是，实验数据集数据量较少，其数据的多样化和全面化都无法与目前深度学习系列的最佳水平对比。

问题二，是对初值选取的敏感问题。在聚类问题是我们指出了这个问题，我们发现，聚类一开始随机分配类型的随机种子对后续的结果异常关键，不同的种子导致最终的训练精度差距在 20%以上，虽然我们在实验时通过多初始值训练的方式变相解决了这个问题。但这个问题依旧存在，一方面是因为多初始值多次训练成倍的增长了训练时间，另一方面，使用此类工程上的解决方案并不能本质的解决问题。

问题三，是对不同分类的效果有差异。例如在物件检测实验中，层次化稀疏 FRAME 随机场模型相较于稀疏 FRAME 随机场模型在鹿这个类中低了 10 个百分点，但其他几类却完全没有问题，这个结果的内部深刻原因，目前我们暂时不太清楚，需要进一步的探索。

4.5.2 实验设计分析

在本次毕业设计的实验设计部分，也遇到了一些困难。

首先是对比模型的选择，生成式模型相对于判别式模型，是一个大类，最初定下毕设题目时，我认为有那么多的模型，总有合适的可以来对比，但事与愿违的是，大部分的模型原理上和我们的层次化稀疏 FRAME 随机场模型大相径庭，使得对比起来无法非常公平。例如深度学习相关领域的模型包括 RCNN，自编码机等等需要的数据量上百万级，一方面我们的模型无法适应如此大规模的数据量，另一方面让这些模型来跑一百个左右的数据量，它们根本无法完成最基本的训练，毫无意义。

其次是实验的完整性，必须承认的是，在一开始我们希望设计更多更全面的实验，例如曾经设想过的在生成图片上的应用，在视频检索上的应用，甚至是对一维音频的建模等等。但因为时间原因，很多设想并未实现，最终只实现了上述三个实验。

第五章　总结

5.1 毕业设计研究总结

FRAME 随机场模型是一个很优秀的生成式概率模型。在一开始接触它时，我就被它的统计思想所感染，它全然不同于领域内风生水起的神经网络模型，学习它使用它和实现它必须拥有一定的统计基础。对它进行非均一化，稀疏化和层次化后，更使得模型的适用范围大幅提高。他的可解释性和鲁棒性保证了模型的精度和调试友好，他的无监督学习和极小数据需求是的应用它变得非常的简单。

本次毕业设计我们充分透彻的分析了 FRAME 随机场模型及其拓展，使我对机器学习生成式模型这一领域有了深入的了解，虽然在实现过程中遇到了一些难题，最终也没有完成当初所有的设想，但可以说也是达成了预期的目标。

5.2 后期规划

FRAME 随机场模型是一个很优秀生成式概率模型，在对它进行稀疏化层次化的优化后，大大拓展了它的适用范围。作为一个生成式模型，它的可解释性，鲁棒性和无监督特性非常优秀。然而，虽然模型本身理论能力出色，例如但受制于代码实现的水平，目前的实际性能依旧远远不足。所以在后期，这个模型的进步范围很大。

另外，在第三章节最后我们提到，层次化的模型可以进一步的继续拓展，增加模型的层数，复杂度。目前受制于模型原理的训练过程，并无法增加更多的层，也无法使模型变得更加复杂，但我相信我们能够找到一个方法，来克服它。

与此同时，在撰写这篇论文时，我也发现自己的统计水平并不足以支撑这个模型的理论研究，只能对一些原理上的说明点到为止。在毕业以后，我将进入 UCLA 统计系进行博士学位的深造。我将在接下来的数年大幅提高自己的统计水平，特别是统计学习领域的知识。同时，继续从事生成学习相关方面的研究。我相信自己能够在这个领域做出卓越的成绩。

参考文献

[1] Felzenszwalb, Pedro F., et al. "Object detection with discriminatively trained part-based models." IEEE transactions on pattern analysis and machine intelligence 32.9 (2010): 1627-1645.

[2] Zhu, Long, et al. "Latent hierarchical structural learning for object detection." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[3] Schnitzspan, Paul, et al. "Discriminative structure learning of hierarchical representations for object detection." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[4] Robert, Christian. "Machine Learning, a Probabilistic Perspective." (2014): 62-63.

[5] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

[6] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In Advances in neural information processing systems, pp. 2672-2680. 2014.

[7] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

[8] Fidler, Sanja, and Ales Leonardis. "Towards scalable representations of object categories: Learning a hierarchy of parts." Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.

[9] Zhu, Long, et al. "Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion." Computer vision–eccv 2008 (2008): 759-773.

[10] Si, Zhangzhang, and Song-Chun Zhu. "Learning and-or templates for object recognition and detection." IEEE transactions on pattern analysis and machine intelligence 35, no. 9 (2013): 2189-2205.

[11] Dai, Jifeng, et al. "Unsupervised learning of dictionaries of hierarchical compositional models." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[12] Mehrotra, Rajiv, Kameswara Rao Namuduri, and Nagarajan Ranganathan. "Gabor filter-based edge detection." Pattern recognition 25.12 (1992): 1479-1494.

[13] Xie, Jianwen, Yifei Xu, Ying Nian Wu, and Song-Chun Zhu. "Generative Hierarchical Structure Learning of Sparse FRAME Models." In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference.

[14] Xie, Jianwen, Wenze Hu, Song-Chun Zhu, and Ying Nian Wu. "Learning sparse FRAME models for natural image patterns." International Journal of Computer Vision 114, no. 2-3 (2015): 91-112.

[15] Xie, Jianwen, Wenze Hu, Song-Chun Zhu, and Ying Nian Wu. "Learning Inhomogeneous FRAME models for object patterns." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1035-1042. 2014.

[16] Zhu, Song Chun, Yingnian Wu, and David Mumford. "Filters, random fields and

maximum entropy (FRAME): Towards a unified theory for texture modeling." International Journal of Computer Vision 27, no. 2 (1998): 107-126.

[17] Zhu, Long, Yuanhao Chen, Alan Yuille, and William Freeman. "Latent hierarchical structural learning for object detection." In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 1062-1069. IEEE, 2010.

[18] Wu, Ying Nian, Zhangzhang Si, Haifeng Gong, and Song-Chun Zhu. "Learning active basis model for object detection and recognition." International journal of computer vision 90, no. 2 (2010): 198-235.

[19] Barbu, Adrian, Tianfu Wu, and Ying Nian Wu. "Learning mixtures of Bernoulli templates by two-round EM with performance guarantee." Electronic Journal of Statistics 8.2 (2014): 3004-3030.

[20] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.

[21] Felzenszwalb, Pedro, David McAllester, and Deva Ramanan. "A discriminatively trained, multiscale, deformable part model." In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1-8. IEEE, 2008.

谢辞

我的毕业设计主要基于我2016年暑假在加州大学洛杉矶分校UCLA统计系进行的科研实践的内容。暑假回国后，我依旧与 UCLA 统计系的导师和学长保持着高频率的交流，继续 FRAME 随机场模型的研究工作。工作内容在去年九月投稿人工智能领域国际顶会 AAAI 被拒稿。经过两个多月的实验优化和调整，十一月底投稿了计算机视觉领域国际顶会 CVPR，并在今年三月被成功收录。在大四一年中，对此模型的改进研究从未停止，我和 UCLA 统计系学长的交流几乎每天都会远程进行。

所以在这里我想要感谢 UCLA 统计系的吴英年老师和 UCLA 统计系的博士后谢建文学长，他们是我去年暑期科研实践的导师和学长，吴英年老师同样也将会是今年下半年开始我在 UCLA 博士生涯的导师。他是一位非常帮助学生的老师，每次邮件他他都会秒回，几乎所有的博士生都非常喜欢他。谢建文学长则是一位非常 push 我的学长，在我编写代码实验遇到问题时，他也会很快解决我的问题，即使是洛杉矶时间的凌晨他也会在工作。

然后要感谢的是上海交通大学电子信息与电气工程学院智能人机交互实验室的张丽清老师以及实验室的周正中学长，在大二我对科研还茫然无知的时候，是张老师和周学长帮助我从头开始学习如何搞科研，如何学习机器学习的知识。张老师给我推荐过许多有用的书籍和论文，在组会时讲过许多有用的报告，在跟张老师交流时也能获得很多的建议和帮助。周学长则能在我有问题时给予我帮助，同时作为图像处理组的组长 push 我干活。

最后感谢交大 ACM 班中的每一位同学，在我的大学生涯中帮助我成长，解答我的各种问题，特别是他们的优秀给予了我无穷的压力与动力，促使我能够认真学习潜心研究。感谢致远学院的每一位老师，特别是俞勇老师，在寻找科研实习，联系导师，申请和进行毕业设计时提供了帮助。感谢我的父母，为我提供良好的条件得以使我完成大学生涯和这篇毕业论文。

一路走来，谢谢身边的每一个人，谢谢！

Generative Hierarchical Learning on Computer Vision

Big Abstract

Nowadays, there is a rapid growth in the area of computer vision, especially the deep learning algorithms including Convolution Neural Network. They are capable to achieve dozens of problems in vision. However, most of them are discriminative models, which is unexplainable, high computation complexity and large data requirement. On the contrary, the generative models are interpretable and low data requirement.

An explainable model is highly desirable, if users are to understand, interpret and effectively manage the behaviors of the models. Models with hierarchical and compositional representations have been shown to be a powerful basis for achieving both prediction accuracy and Interpretability. They are capable of learning reconfigurable representations to deal with both structural and appearance variations of objects. These models can be paired with either discriminative learning method or generative learning method. Discriminative learning seeks to identify and weigh the most discriminant features and structures for explaining the object categories, while generative learning enables us to learn the parameters and interpretable patterns for explaining the image data instead of predicting the image categories. Moreover, generative learning is not only important for making the model explainable, it can also be used for unsupervised learning from unlabeled images.

FRAME MODEL

This paper focus on FRAME model. FRAME model, which is Filter, Random Field and Maximum Entropy, is a generative probabilistic model. The model considers the image as a Markov random field, with the Gabor filter as the under layer expression. It can be learned by a method similar to the maximum entropy principle. The originally FRAME model is proposed for stochastic texture patterns, whose Markov random field is called spatially stationary. Furthermore, it is the maximum entropy distribution that reproduces the observed marginal histograms of responses from a bank of filters, where for each filter tuned to a specific scale and orientation, the marginal histogram is spatially pooled over all the pixels in the image domain. After that, Xie proposed sparse FRAME model as an inhomogeneous and sparse generalization of the original FRAME model. The sparse FRAME is a non-stationary Markov random field model that reproduces the observed statistical properties of filter responses at a subset of selected location, scales and orientations. It is a generative model with a well-defined probability distribution on the image intensities. Unlike the original FRAME model for texture patterns, each spares FRAME model is intended to model an object pattern, and can be considered a deformable template.

As a sparse FRAME model, given a batch of training images, the trained template of sparse FRAME model is a subset of Gabor filters, which are different shapes, locations, scales and orientations of Gabor filters. Normally, we choose 200 Gabor filters as a template for animal face

data. Using this template, we can inference the deformable template of other image, which the chosen Gabor filters can be shift on its location, scale and orientations.

HIERARCHICAL SPARSE FRAME MODEL

After study on FRAME model and its generalization. We propose a method for generative learning of hierarchical random field models, which we call the hierarchical sparse FRAME model. The original sparse FRAME models can only deal with small deformations (e.g., edge perturbations), and may fail when there exist large geometric changes (e.g., part deformations). To address this limitation, we propose to extend the original sparse FRAME model to a hierarchical version, is a generalization of the original sparse FRAME model by decomposing it into multiple parts that are allowed to shift their locations and rotations relative to each other. Each part template is in turn composed of a group of Gabor wavelets that are allowed to shift their locations and orientations relative to each other. The hierarchical sparse FRAME model is a hierarchical compositional deformable template.

As a hierarchical sparse FRAME model, given a batch of training images, the trained template of hierarchical sparse FRAME model is the similar to the sparse FRAME model. The difference is that our model split these Gabor filters into a few groups, represented as parts of the object. These parts are often split by locations. On our paper, we simply split the parts by 3*3 bounding boxes. Being regarded as a part, Gabor filters in the same part can additionally shift locations, scales and orientations simultaneously. It will provide more robustness when the parts of the object move significantly.

As the model is a fully generative model, it can be learned in an unsupervised manner, where the locations, scales and orientations of the object, parts, and edges (Gabor wavelets) are unknown, by an EM-type algorithm that alternates inference and re-learning steps. The model is learned in a generative manner in the sense that the learning is carried out by maximum likelihood estimation and also it involves synthesizing image patterns via MCMC sampling.

(1) Inference: Given the current model, we match it to each training image by inferring the unknown locations, scales, and rotations of the object and its parts by recursive sum-max maps. In other word, it is to determine a certain geometric configuration of the template for a given object such that the log-likelihood is maximized. This can be efficiently achieved by a bottom-up/top-down dynamic programming, which is implemented by recursive sum-max maps.

(2) Re-learning: Given the inferred geometric configurations of the objects and their parts, we re-learn the model parameters by maximum likelihood estimation via stochastic gradient algorithm. we first align the objects and parts by morphing the corresponding image patches. We then learn an original sparse FRAME model on the aligned training images, and then divide the object template into parts templates.

The hierarchical sparse FRAME model is essentially a 3-layer sparsely connected convolutional

neural network with sophisticated max-pooling, except that the first layer of Gabor filters is already given and the weighting parameters are learned in a generative manner. As a result, we are possible to extend it by providing more complexity on number of layer and method of mapping and pooling in future work.

EXPERMENTS

We implement the hierarchical sparse FRAME model into three well-designed experiments to show the interpretation, robustness and unsupervised advantage.

The first experiment is object detection and key point alignment. We choose 8 categories of animal face data set. Each category has 40 images, in which 10 of them are training images, 30 of them are testing images. All of them are labeled with bounding box and key points position. After training the h-FRAME model, the model will map each key point to one of the Gabor filter. By model interface and deformable template, our method can predict the position of the object and its key points' positions. In this experiment, we compare our hierarchical sparse FRAME model with sparse FRAME model, and-or graph model and deformable part model.

The second one is clustering problem. This experiment tests the unsupervised manner of our model. We select 12 different datasets, which each of them have 2-5 subset with 15 images each. The model will learn 2-5 different templates represent 2-5 subset by inputting the unlabeled mixed training images. To avoid the unstable result by different random seeds. We use multi random seeds and select the highest interface score as the final result. In this experiment, we compare our hierarchical sparse FRAME model with sparse FRAME model, active basis model, two-step EM algorithm, HoG features and and-or graph model.

The third one is classification problem. This experiment shows the meaningfulness of template we trained of hierarchical sparse FRAME model. We regard the template as a feature extractor and use SVM to evaluate the precision of the classification problem. We select thousands of animal images with 20 different categories as the data set. We train 5-11 templates for each categories which represent different sub-categories in the same category. Combining with 5*20 = 100 templates for every categories and its templates, we have a 100 templates codebooks and can be transferred into a vector through SPM. In this experiment, we compare our hierarchical sparse FRAME model with and-or graph model.

Three experiments, along with the comparison between different versions of FRAME model and the comparison with different generative and hierarchical model, show that our proposed model is capable of learning meaningful and interpretable templates.