# 基于 what 和 where 信息的目标检测方法

田 媚,罗四维,廖灵芝
(北京交通大学计算机与信息技术学院,北京 100044)

**摘 要:** 根据视觉系统两条通路理论,提出了一种基于 what 和 where 信息的目标检测方法.采用以环境为中心的 where 信息进行自顶向下的注意控制,指导 what 信息驱动的自底向上的注意.自顶向下的注意包括预注意和集中注意两个阶段,预注意依据 where 信息为特定目标出现与否提供先验,做出是否继续搜索的判定.集中注意的结果与 what 信息相结合,将注意指向目标最有可能出现的图像区域,并得到一系列样本显著区域.应用于多幅自然图像的实验结果证明了算法的有效性.

**关键词:** 自顶向下的注意; where 信息; what 信息; 目标检测
**中图分类号:** TP391 **文献标识码:** A **文章编号:** 0372-2112 (2007) 11-2055-07

# Object Detection Method Based on "What" and "Where" Information

TIAN Mei, LUO Si-wei, LIAO Ling-zhi
(*School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044, *China*)

**Abstract:** Inspired by the theory of two visual pathways, a novel model for object detection is proposed based on "what" and "where" information. Context-centered "where" information is used to control top-down attention, and guide bottom-up attention which is driven by "what" information. The procedure of top-down attention can be divided into two stages: pre-attention and focus attention. In the stage of pre-attention, "where" information can be used to provide the prior knowledge of presence or absence of objects which decides whether search operation is followed. By integrating the result of focus attention with "what" information, attention is directed to the region that is most likely to contain the object and series of salient regions for samples are detected. Experimental results with natural images demonstrate its effectiveness.

**Key words:** top-down attention; "where" information; "what" information; object detection

## 1 引言

生物在感知外部世界时,视觉系统分为两个皮层视觉子系统,即两条视觉通路——what 通路和 where 通路[1].what 通路传输的信息与外部世界的目标对象相关;where 通路用来传输对象的空间信息.这种有关视觉系统的理论与 Marr 的观点不谋而合.在 Marr 的计算理论中,视觉被看作是一种信息处理过程,其主要目的是从图像中发现存在于外部世界中的目标以及目标所在的空间位置[2].因此,研究视觉感知系统信息处理理论的一个主要内容涉及目标对象和空间位置感知,其实质就是研究两个视觉子系统 what 通路和 where 通路的功能.根据两条通路理论,我们分别定义在 what 和 where 通路中传输的信息——what 和 where 信息.结合注意机制,what 信息可用于驱动自底向上的注意,形成感知和进行目标识别,where 信息可以用来驱动自顶向下的注意,处理空间信息.本文采用自顶向下的注意与自底向上的注意相结合的方法,同时考虑 what 信息和 where 信息,完成目标检测.

视觉注意的作用是将人类注意快速指向视觉环境中感兴趣的目标[3].通过近年来的一些研究,人们对于理解视皮层信息处理的基本原理已经取得了巨大的进步,从而使得自底向上的注意引导备受国内外研究者的关注,并成功地建立了一些基于注意的目标检测模型[4~6].这些模型的检测过程与目标不相关,当在复杂混乱的场景中寻找的特定目标不是最显著时,检测效率不高.为了说明目标相关约束,Sun 提出了基于目标的注意计算模型[7].基于目标的注意应用在检测方面比较有代表性的例子是交通标志的检测[8].但是所有这些模型都使用目标自身的特征进行目标检测和识别,忽略了高层信息的指导,因此面临着两个难题:第一,当图像质量下降使得目标自身信息不足,从而不能进行可靠的

收稿日期:2006-09-08;修回日期:2007-01-22
基金项目:国家自然科学基金(No.60373029);国家教育部博士点基金(No.20050004001)

检测和识别时,这种方法将会失效.第二,对应于相同目标的不同形态和位置,这种方法不能很好地利用先验,需要搜索图像中所有的空间.

为了解决这些难题,研究者们引入自顶向下的注意进行目标检测和识别[9~13].Rybak[10]在定义目标显著性时增加了一个"语义参数",但是这个"语义参数"实质只是在高层视觉结构缺乏注意的自顶向下控制时,预先定义以强调图像中具有重要意义的部分,并不是真正的自顶向下的注意控制.Salah[11]将可观测马尔科夫模型引入到模拟任务驱动的注意中来,并在数字识别和人脸识别的实验中取得了很好的效果.但是该模型只是将基于数据驱动的注意焦点转移图作为先验,模拟自顶向下的注意控制,并没有用到真正的高层信息.这些基于自顶向下注意的目标检测方法使用的"高层信息"包括存储在记忆中的模板、可以调节视觉感知的阈值、根据需求或动机设置的偏置或权重等,都只是简化后的近似高层信息,仅模拟了注意机制的调制作用,只适用于解释视觉感知的初级阶段.

通过对视觉的研究发现环境因素在目标检测过程中有非常重要的作用.当要寻找的特定目标与场景环境一致时,该目标更容易被注意[14].环境信息既可以为哪种目标最有可能出现提供很强的先验,也可以为图像中期望的目标出现的位置提供先验.此外,环境还可以帮助消除局部特征不充分时引起的歧义.因此,鉴于现有的目标检测方法主要依靠低层数据信息,缺乏对高层信息的有效定义,本文定义了新的以环境为中心的 where 信息,模拟 where 通路中传输的自顶向下的注意控制信息,由此获得相关目标的先验知识,用以指导与 what 信息相关的自底向上的注意,为目标检测提供有效的捷径.

本文将基于 where 信息的自顶向下的注意分为两个阶段:预注意和集中注意.这里的预注意与自底向上注意机制中的预注意[3]不同.Itti 通过预注意将颜色、朝向和亮度等初级视觉特征快速、自动地并行加工,形成多个显著图为图像中每个位置的显著性提供度量.我们的算法在预注意阶段根据 where 信息为特定目标出现与否提供先验,做出是否继续搜索的决定.集中注意阶段在预注意的基础上给出目标最有可能出现的位置信息.这样,既提高了目标检测效率,将计算资源优先分配给那些目标出现概率比较高的图像;又提高了目标检测的可靠性,将注意集中到目标可能出现的位置区域,使检测过程不受区域外其它显著目标的影响.

## 2 算法结构

当整幅图像是某个目标的特写镜头时,目标在图像中占主要部分,环境信息主要由目标决定,可以用 what 信息近似表示 where 信息.这时,基于自底向上的注意,只用 what 信息就可以完成目标检测.但是,当目标的大小与整幅图像相比较小时,环境信息主要由背景决定,而不是由目标决定.这时就需要采用自顶向下的注意与自底向上的注意相结合的方法,同时考虑 what 信息和 where 信息,完成目标检测.在本文中我们将集中研究第二种情况,这在目标检测方法中还没有引起足够的重视.

我们用基于统计的方法进行目标检测,因为环境与其所含目标之间存在很强的关系,给定图像的 where 信息 $\boldsymbol{Ve}$,目标出现在由 $\boldsymbol{Ve}$ 表示的环境中的条件概率可以写作 $P(O|\boldsymbol{Ve})$.考虑目标 $O$ 的类别属性 $s$ 和位置属性 $l=(x,y)$,连续使用贝叶斯规则

$$P(O|\boldsymbol{Ve})=P(l|s,\boldsymbol{Ve})P(s|\boldsymbol{Ve}) \quad (1)$$

似然函数 $P(s|\boldsymbol{Ve})$和 $P(l|s,\boldsymbol{Ve})$分别对应自顶向下注意引导的两个阶段:预注意和集中注意.本文将其与 what 信息相结合,提出了基于 what 和 where 信息的目标检测方法.如图 1 所示,输入图像经过 Gabor 滤波后得到 16 幅特征显著图.从那儿开始,目标检测分成两条并行分等级的通路.一条通路提取 what 信息,另一条提取 where 信息.根据图像的 where 信息进行自顶向下的注意控制.在预注意阶段,估计似然函数 $P(s|\boldsymbol{Ve})$的值.若 $P(s|\boldsymbol{Ve})<Q$ 则停止搜索,若 $P(s|\boldsymbol{Ve})\geqslant Q$ 则转入集中注意阶段,$Q$ 为阈值.集中注意阶段估计出的 $p(l|s,\boldsymbol{Ve})$可以用来指导自底向上的注意,并将注意指向目标最有可能出现的图像区域,即集中注意区域.将 $p(l|s,\boldsymbol{Ve})$与 what 信息 $\boldsymbol{Vl}$ 相结合产生综合信息 $\boldsymbol{Vi}$,并由此得出最终的目标检测结果.

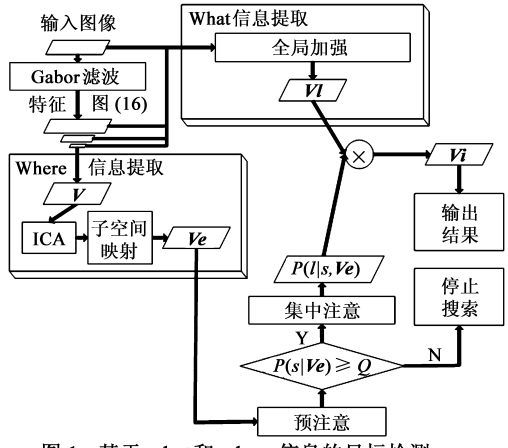



图 1 基于 what 和 where 信息的目标检测

## 3 what 和 where 信息的获取与表示

根据人类视觉感知系统的初级视觉特性,what 通路和 where 通路都是从视网膜开始,经过侧膝体(LGN)、初级视皮层(V1),然后才分开.因此,我们提出的目标

检测模型从视觉输入中提取 what 和 where 信息的初始阶段相同,都选择 Gabor 滤波器处理输入图像,模拟不同位置、不同尺度的感受野特性.由于 where 信息表示整个场景的环境信息,为全局信息,滤波后得到 where 信息的高维编码,还需要进一步从中提取相关信息得到简单有效的表示.而 what 信息表示图像中的局部信息,所以处理方法有所不同,我们将在后面的两个小节中详细地介绍.

### 3.1　以环境为中心的 where 信息提取

真实世界中的场景具有一定的规则性,属于同一类的场景具有相似稳定的空间结构,可以不经过图像分割就提取出来.这使得我们可以将整个场景看成一个单独的目标,定义与整体环境特性相关的特征,而不必定义场景中的独立目标.为了表示整体环境信息,模拟从视网膜到 LGN 的处理过程,我们采用 Gabor 滤波器对输入的整幅图像滤波,二维 Gabor 滤波器的时域公式如下

$$h_e(x,y)=g(x',y')\cos(2\pi f_0 x') \tag{2}$$

$$h_o(x,y)=g(x',y')\sin(2\pi f_0 x') \tag{3}$$

其中 $h_e(x,y)$和 $h_o(x,y)$分别表示偶对称和奇对称的 Gabor 滤波器,$g(x',y')$为高斯函数,这里的 $x'=x\cos\theta+y\sin\theta$, $y'=-x\sin\theta+y\cos\theta$, $g(x,y)=\exp\left(-\frac{x^2}{2\sigma_x^2}-\frac{y^2}{2\sigma_y^2}\right)$.$\theta$ 是 Gabor 滤波器的朝向,$f_0$ 是中心频率.参数 $\sigma_x$ 和 $\sigma_y$ 是在空域 $x'$, $y'$ 方向的高斯方差.我们使用的滤波器包括 4 个朝向 $\theta\in\{0,\pi/4,\pi/2,3\pi/4\}$ 和 4 个频率 $f_0\in\{0.1,0.2,0.3,0.4\}$.图像 $\boldsymbol{I}(x,y)$经过滤波后的输出可以表示为

$$\boldsymbol{v}(x,y)=\sqrt{[\boldsymbol{v}_e(x,y)]^2+[\boldsymbol{v}_o(x,y)]^2} \tag{4}$$

其中,$\boldsymbol{v}_e(x,y)=\boldsymbol{I}(x,y)*h_e(x-x_0,y-y_0)$,$\boldsymbol{v}_o(x,y)=I(x,y)*h_o(x-x_0,y-y_0)$,$(x_0,y_0)$为感受野中心位置,$*$ 表示卷积.每个 Gabor 滤波器相当于一个神经元,对原始输入图像 $\boldsymbol{I}(x,y)$在某个频率和朝向上给出响应,从而得出一组新的图像$\{\boldsymbol{v}_k(x,y),k=1,2,\cdots,16=4\times4\}$.变量 $k$ 表示不同空间频率和朝向的滤波器.用这组图像构成 where 信息的高维编码 $\boldsymbol{V}$, $\boldsymbol{V}_{i,j}=\boldsymbol{v}_k(x,y)$,$i$ 为不同朝向索引号,$j$ 为不同频率索引号.

基于编码 $\boldsymbol{V}$ 的高阶统计特性,我们采用独立分量分析(ICA)[15]对高维编码进行线性分解,降低编码维数,并使得到的独立分量满足稀疏特性.将每个 $\boldsymbol{V}_{i,j}$作为一个样本,提取所有样本的线性基函数.这样,对于所有的样本,基函数都是相同的,不同的是系数向量,每一个样本就可以用其统计特征系数向量表示.对于不同朝向 $\theta$、不同频率 $f_0$ 的训练样本集$\{\boldsymbol{V}_{i,j}^{(k)}\}$,采用 Hyvörinen 的快速定点算法[15]进行训练,估计出相应的转换矩阵 $\boldsymbol{W}_{i,j}$.这样每一个样本的特征系数 $\boldsymbol{S}_{i,j}^{(k)}=\langle\boldsymbol{W}_{i,j},\boldsymbol{V}_{i,j}^{(k)}\rangle$就可以计算出来.

经过前面两个阶段的处理,可以用一组特征系数 $\boldsymbol{S}_{i,j}$表示原始图像.我们将分解得到的同一朝向、同一频率的基函数定义为一个线性子空间,其输出响应则定义为到该子空间的投影距离.所有不同朝向、不同频率的响应值就构成了一个响应矩阵 $\boldsymbol{Ve}=\{ve_{i,j}\}$.矩阵中的每一个 $ve_{i,j}$的计算公式为

$$ve_{i,j}=\sqrt{\sum_{n=1}^{d}(a_{i,j}^{(n)})^2} \tag{5}$$

这里的 $a_{i,j}$是标量,表示特征系数向量 $\boldsymbol{S}_{i,j}$的一个特征值.$n$ 是特征值的索引号,$d$ 为选取的特征值的个数.经过这一步的处理,可以进一步降低编码维数.我们提出用响应矩阵 $\boldsymbol{Ve}$ 作为原始图像的 where 信息.因为原始图像中所有区域对最终的响应矩阵都有一定的贡献,$\boldsymbol{Ve}$ 不包含特定目标的独立信息,又没有抛弃任何属于目标的度量,所以 $\boldsymbol{Ve}$ 可以表示整个场景的环境信息.

### 3.2　以目标为中心的 what 信息提取

我们对输入图像提取较为敏感的朝向、频率、亮度这 3 类特征,形成各个特征维的显著图.然后用全局加强法[16]对这些显著图进行合并,形成一幅最终的显著图,并用其作为原始图像的 what 信息.

与 where 信息第一阶段处理过程相同,我们用 4 个朝向、4 个频率的 Gabor 滤波器对输入图像滤波,得到 16 幅朝向、频率特征图$\{\boldsymbol{v}_k(x,y),k=1,2,\cdots,16=4\times4\}$.直接将图像 $\boldsymbol{I}(x,y)$作为亮度特征图,记为 $\boldsymbol{v}_{17}(x,y)$.将各个特征图的特征值归一化到同一个范围内后,找出每一幅特征图的全局极大 $M$ 和除此全局极大之外的其他局部极大的平均值 $\bar{m}$,给每一幅特征图乘以加强因子$(M-\bar{m})^2$,这就是每幅特征图的权.最终的显著

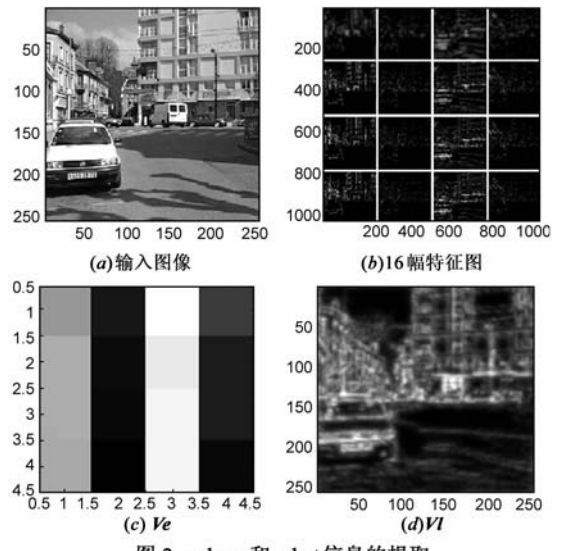



图 2　where和 what信息的提取

图是各个特征显著图的加权和,并将其作为 what 信息.

图 2 是对一幅的 8bits 灰度图提取 where 信息和 what 信息的实验结果. 图 2($a$)为原始输入图像,图 2($b$)是以朝向 $\theta$(水平)、频率 $f_0$(竖直)为索引排列的 16 幅特征图,图 2($c$)为原始图像的 where 信息 $\boldsymbol{Ve}$,其中 $d=16$. 图 2($d$)为原始图像的 what 信息 $\boldsymbol{VI}$.

## 4 目标检测算法

### 4.1 基于 where 信息的预注意

式(1)中目标似然值的计算是一个学习的过程,得到的信息是关于过去在相似环境中成功发现目标的搜索经验. 即给定 where 信息,通过学习得出什么样的目标最有可能出现和出现的位置. 在预注意阶段,似然函数 $P(s\mid \boldsymbol{Ve})$可以表示成

$$P(s\mid \boldsymbol{Ve})=\frac{p(\boldsymbol{Ve}\mid s)p(s)}{p(\boldsymbol{Ve}\mid s)p(s)+p(\boldsymbol{Ve}\mid \neg\, s)p(\neg\, s)} \tag{6}$$

只有在系统积累了足够的经验之后,where 信息对目标检测才是有效的,即用于学习 $P(s\mid \boldsymbol{Ve})$的训练集中要包含大量的图片. 根据概率论,在大样本事件中,一个样本出现与否的概率均为 1/2. 因此,可以近似定义目标的先验概率 $p(s)=p(\neg\, s)=1/2$. 用来学习似然函数 $p(\boldsymbol{Ve}\mid s)$的训练集是一组包含目标的图片,训练数据是这组图片的 where 信息 $\boldsymbol{VE}=\{\boldsymbol{Ve}_1,\boldsymbol{Ve}_2,\cdots,\boldsymbol{Ve}_t,\cdots,\boldsymbol{Ve}_N\}$,$N$ 是参加训练的图片数. 则似然函数的定义可以表示为

$$p(\boldsymbol{Ve}\mid s)=\prod_{t=1}^{N}p(\boldsymbol{Ve}_t\mid s) \tag{7}$$

模型中引入环境类别信息 $C=\{C_i\}_{i=1,K}$,$\omega_i$ 表示第 $i$ 类的先验概率,则有

$$p(\boldsymbol{Ve}_t\mid s)=\sum_{i=1}^{K}\omega_i p(\boldsymbol{Ve}_t\mid C_i,s),\ \sum_{i=1}^{K}\omega_i=1 \tag{8}$$

那么,根据 $p(\boldsymbol{Ve}_t\mid C_i,s)$符合的密度模型,可以用一个特定的混合模型来模拟目标 $s$ 的似然函数 $p(\boldsymbol{Ve}\mid s)$.

对于不能预先确认的问题,特定的函数具有一定的局限性,所以对于第 $i$ 类我们引入均值、方差分别为 $\mu_i$ 和 $\sigma_i^2$ 的广义高斯模型[17]

$$p(\boldsymbol{Ve}_t\mid \mu_i,\sigma_i,\beta)=\frac{\omega(\beta)}{\sigma_i}\exp\left[-c(\beta)\left|\frac{\boldsymbol{Ve}_t-\mu_i}{\sigma_i}\right|^{2/(1+\beta)}\right] \tag{9}$$

$$c(\beta)=\left[\frac{\Gamma\left[\frac{3}{2}(1+\beta)\right]}{\Gamma\left[\frac{1}{2}(1+\beta)\right]}\right]^{1/(1+\beta)},$$

$$\omega(\beta)=\frac{\Gamma\left[\frac{3}{2}(1+\beta)\right]^{1/2}}{(1+\beta)\Gamma\left[\frac{1}{2}(1+\beta)\right]^{3/2}},\Gamma(a)=\int_0^{\infty}e^{-t}t^{a-1}\mathrm{d}t.$$

选取不同的 $\beta$ 值,可以形成各种不同的混合模型. 本文选取 $\beta=-0.25$,$\beta=0$(正态分布),$\beta=0.495$(Tanh 分布)和 $\beta=1$(拉普拉斯分布)时的模型作为候选模型,分别对应了亚高斯、高斯和超高斯模型. 然后,根据 IGMSC 模型选择准则[18],从候选模型中选出最适合的正态分布作为 $p(\boldsymbol{Ve}_t\mid C_i,s)$符合的密度模型,形成高斯混合模型来模拟训练.

这样,在由 $K$ 个高斯分布组成的高斯混合模型中,假定第 $i$ 类符合均值为 $\mu_i$,方差为 $\sigma_i^2$ 的高斯分布. 对应于目标类的模型参数为 $\Theta=\{\theta_i,\omega_i\}_{i=1,K}$,其中第 $i$ 个高斯分布的参数为 $\theta_i=(\mu_i,\sigma_i)$,每个数据的分布可以用混合高斯表示为

$$p(\boldsymbol{Ve}_t\mid s)=p(\boldsymbol{Ve}_t\mid \Theta)=\sum_{i=1}^{K}\omega_i p(\boldsymbol{Ve}_t\mid C_i,\theta_i) \tag{10}$$

其中,

$$p(\boldsymbol{Ve}_t\mid C_i,\theta_i)=\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{(\boldsymbol{Ve}_t-\mu_i)^2}{2\sigma_i^2}\right] \tag{11}$$

定义函数 $L(\Theta\mid \boldsymbol{Ve})$为

$$L(\Theta\mid \boldsymbol{Ve})=\log p(\boldsymbol{Ve}\mid \Theta)=\sum_{t=1}^{N}\log\left[\sum_{i=1}^{K}\omega_i p(\boldsymbol{Ve}_t\mid C_i,\theta_i)\right] \tag{12}$$

目标类的模型参数 $\Theta$ 通过 EM[19]算法获得.

E 步,首先根据 Bayes 规则,在给定最新的 $\hat{\theta}_i=(\hat{\mu}_i,\hat{\sigma}_i)$和 $\hat{\omega}_i$ 估计的基础上求出后验概率

$$p(C_i\mid \boldsymbol{Ve}_t,\hat{\theta}_i)=\frac{\hat{\omega}_i p(\boldsymbol{Ve}_t\mid C_i,\hat{\theta}_i)}{\sum_{i=1}^{K}\hat{\omega}_i p(\boldsymbol{Ve}_t\mid C_i,\hat{\theta}_i)} \tag{13}$$

然后,使用当前参数 $\hat{\Theta}$ 和 where 信息 $\boldsymbol{Ve}$ 计算完整样本数据的 $L(\Theta\mid \boldsymbol{Ve},C)$的期望值[19]

$$Q(\Theta,\hat{\Theta})=\mathrm{E}[\log p(\boldsymbol{Ve},C\mid \Theta)\boldsymbol{Ve},\hat{\Theta}]=\int_C\log p(\boldsymbol{Ve},C\mid \Theta)p(C\mid \boldsymbol{Ve},\hat{\Theta})\mathrm{d}C \tag{14}$$

M 步,选择使 $Q(\Theta,\hat{\Theta})$最大时 $\Theta$ 的值

$$\hat{\omega}_i^{\mathrm{new}}=\frac{1}{N}\sum_{t=1}^{N}p(C_i\mid \boldsymbol{Ve}_t,\hat{\theta}_i) \tag{15}$$

$$\hat{\mu}_i^{\mathrm{new}}=\frac{1}{\sum_{t=1}^{N}p(C_i\mid \boldsymbol{Ve}_t,\hat{\theta}_i)}\sum_{t=1}^{N}p(C_i\mid \boldsymbol{Ve}_t,\hat{\theta}_i)\boldsymbol{Ve}_t \tag{16}$$

$$(\hat{\sigma}_i^{\mathrm{new}})^2=\frac{1}{\sum_{t=1}^{N}p(C_i\mid \boldsymbol{Ve}_t,\hat{\theta}_i)}\sum_{t=1}^{N}p(C_i\mid \boldsymbol{Ve}_t,\hat{\theta}_i)\cdot(\boldsymbol{Ve}_t-\hat{\mu}_i^{\mathrm{new}})\cdot(\boldsymbol{Ve}_t-\hat{\mu}_i^{\mathrm{new}})^T \tag{17}$$

这样 E 步和 M 步迭代进行,直到收敛为一个稳定值. 对于似然函数 $p(\boldsymbol{Ve}\mid \neg\, s)$采用相同的方法. 将学习的结果带入式(6),就可以得出预注意的结果.

### 4.2 where 信息驱动的集中注意

在集中注意阶段,似然函数 $P(l\mid s,\boldsymbol{Ve})$可以表示成

$$P(l \mid s, \boldsymbol{Ve}) = \frac{p(l, \boldsymbol{Ve} \mid s)}{p(\boldsymbol{Ve} \mid s)} \tag{18}$$

似然函数 $P(l, \boldsymbol{Ve} \mid s)$通过一个高斯混合模型来模拟训练,学习给出了 where 信息和属于某类目标的典型位置之间的关系.用来学习似然函数 $p(l, \boldsymbol{Ve} \mid s)$的训练集是一组包含目标 $s$ 的图片,训练数据是这组图片的 where 信息 $\boldsymbol{VE} = \{\boldsymbol{Ve}_1, \boldsymbol{Ve}_2, \cdots, \boldsymbol{Ve}_t, \cdots, \boldsymbol{Ve}_N\}$和目标 $s$ 在场景中的位置 $L = \{l_1, l_2, \cdots, l_t, \cdots, l_N\}$. $p(l, \boldsymbol{Ve} \mid s)$由 $K$ 个高斯族组成,每个族分解成两个高斯函数的积,分别对应 where 信息$(\theta_i = (\mu_i, \sigma_i))$和目标位置$(\delta_i = (\mu_i', \sigma_i'))$.所以,对应于目标类的模型参数为 $\Theta' = \{\theta_i, \delta_i, \omega_i\}_{i=1,\cdots,K}$,每个数据的分布可以用混合高斯表示为

$$p(l_t, \boldsymbol{Ve}_t \mid s) = \sum_{i=1}^{K} \omega_i p(\boldsymbol{Ve}_t \mid C_i, \theta_i) p(l_t \mid \delta_i) \tag{19}$$

与预注意学习过程类似,目标类 $s$ 的模型参数 $\Theta'$ 通过 EM 算法获得.

### 4.3　what 信息与 where 信息的结合

在集中注意阶段,我们通过训练估计出模型参数 $\Theta = \{\theta_i, \delta_i, \omega_i\}_{i=1,K}$,再对测试图像的每个位置计算似然函数 $p(l \mid s, \boldsymbol{Ve})$来找出集中注意区域.通过实验观察测试图像中 $p(l \mid s, \boldsymbol{Ve})$的分布情况,发现 where 信息为估计集中注意区域的竖直位置 $y$ 提供了很强的先验,但是对水平位置 $x$ 的确定几乎没有贡献.为了确定集中注意区域,我们定义 $l_{x_0, y_0} = \sum_{l_{x,y}} l_{x,y} p(l_{x,y} \mid s, \boldsymbol{Ve})$为区域中心,区域的宽度为测试图像的宽度,然后从 $y = y_0$ 开始循环计算 $\Delta_+$ 和 $\Delta_-$,以确定区域的高度.$\Delta_+$ 定义为

$$\Delta_+ = \frac{\sum_{x=0}^{255} p(l_{x,y} \mid s, \boldsymbol{Ve})}{\sum_{x=0}^{255} p(l_{x,y+1} \mid s, \boldsymbol{Ve})}, y = y + 1 \tag{20}$$

当 $\Delta_+ > 10$ 时停止计算,并定义 $y_+ = y$.对于 $\Delta_-$ 采用相同的方法,不同的是分母中的似然函数为 $p(l_{x,y-1} \mid s, \boldsymbol{Ve})$,每计算一次 $y = y - 1$,最终 $y_- = y$.此时定义 $h = y_+ + y_-$ 为集中注意区域的高度.

将测试图像每个位置的 $p(l \mid s, \boldsymbol{Ve})$与测试图像对应位置的 what 信息 $\boldsymbol{Vl}$ 相乘得出综合信息 $\boldsymbol{Vi}$.假定集中注意区域可以划分成 $b$ 个 $n \times n$ 的图像块(在我们的实验中,$n$ 取 16),我们取 $4b$ 个随机分块,可以保证覆盖重构误差最小.将每个子块对应的综合信息 $\boldsymbol{Vi}$ 作为一个样本,转换成样本向量,并用此行向量作为该样本的描述信息.然后,我们计算各样本向量与集中注意区域中其它样本向量对应项之间综合信息差异的平方和,并将这个和作为对应样本的显著性度量.那些与其它样本相比具有不同特性的样本区域更显著,目标出现概率较大.由于样本的综合信息 $\boldsymbol{Vi}$ 来自图像的局部信息,所以图像的全局旋转变换对 $\boldsymbol{Vi}$ 几乎没有影响.而且,算法将位置参数当作变量,在整幅图像中取 $4b$ 个随机分块,所以即使目标在图像中发生平移,该算法仍能通过其它子块发现该显著目标.

## 5　实验

我们进行训练的图像来自 Database of Cars and Faces in Context.数据库由 2500 多幅图像组成,涉及房间、超市、旅馆、街道等多种场景.实验中训练图像大小为 256 × 256,灰度级为 0 ~ 255.选定汽车和人两类目标进行实验.在预注意阶段,我们用包含目标 $s$ 的 200 幅图像作为学习似然函数 $p(\boldsymbol{Ve} \mid s)$的训练集,高斯混合模型参数 $K = 2$,对应两类环境(室内环境和室外环境).图 3 给出了在 8 幅图像中寻找目标类汽车和人时 $P(s \mid \boldsymbol{Ve})$的分布情况(图像来自训练集).似然函数 $P(s_{\text{cars}} \mid \boldsymbol{Ve})$和 $P(s_{\text{people}} \mid \boldsymbol{Ve})$分别表示汽车和人出现在图像中的概率.可以看出,对于 $P(s \mid \boldsymbol{Ve}) \approx 1$ 的图像,不需要仔细搜索整幅图就可以肯定目标的存在.与之相反,我们可以判定目标不会出现在 $P(s \mid \boldsymbol{Ve}) \approx 0$ 的图像中,不必继续在该图像中搜索目标.在图 3 第 6 幅场景中并没有出现汽车,但似然函数 $P(s_{\text{cars}} \mid \boldsymbol{Ve}) \sim 1$.说明在这个阶段,注意仅受 where 信息的驱动,而跟属于目标的 what 信息无关.

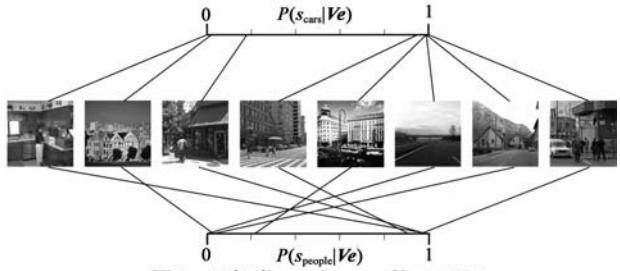



图 3　目标类 $s_{\text{cars}}$ 和 $s_{\text{people}}$ 的 $P(s|\boldsymbol{Ve})$

另外,注意到在图 3 中 $P(s \mid \boldsymbol{Ve})$主要分布在 0 和 1 附近.我们从图像库中随机选出 50 幅图像计算 $P(s_{\text{people}} \mid \boldsymbol{Ve})$,其中有 29 幅 $P(s_{\text{peolpe}} \mid \boldsymbol{Ve}) \geqslant 0.95$,有 17 幅 $P(s_{\text{people}} \mid \boldsymbol{Ve}) < 0.05$,图 4 是满足两种条件的场景集合示例.这说明预注意过程可以为大多数图像提供关于特

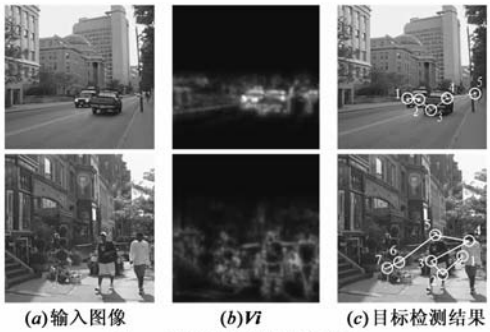
(a) $P(s_{\text{people}}|\boldsymbol{Ve})>0.95$

(b) $P(s_{\text{people}}|\boldsymbol{Ve})<0.05$

图 4　由 $P(s_{\text{people}}|\boldsymbol{Ve})>0.95$ 和 $P(s_{\text{people}}|\boldsymbol{Ve})<0.05$ 定义的场景集合示例

定目标是否出现的可靠的先验.一旦得出目标不会出现在某个图像中($P(s|\boldsymbol{Ve})<0.05$),就立即停止目标检测,从而在很大程度上节约计算资源,提高目标检测的工作效率.实验中 $Q=0.05$ 保证系统会对不能准确判定是否包含目标的图像继续进行目标检测.

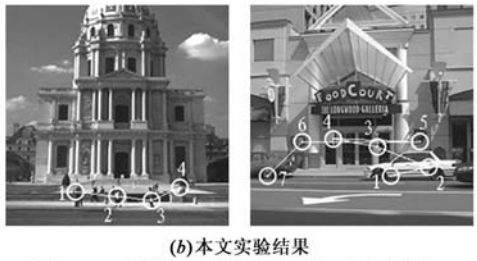
图 5 是使用本文算法在图像中寻找汽车和人时的实验结果.图 5($a$)为原始输入图像,图 5($b$)是综合信息 $\boldsymbol{Vi}$.两幅图中黑色区域对应的 $p(l|s,\boldsymbol{Ve})\sim 0$,使得该区域的 what 信息被遮蔽.也就是说目标几乎不会出现在这部分空间位置,进行目标检测时不需要搜索这部分区域,可以将计算资源分配给由 where 信息选择的集中注意区域.集中注意区域对应的 $p(l|s,\boldsymbol{Ve})\sim 1$,目标出现的概率很高.图 5($c$)为整个检测过程完成后的最终结果.按照样本显著性由强到弱的顺序,图 5($c$)的两幅图分别显示了 5 个(用 1 到 5 这五个数字标出)和 7 个(用 1 到 7 这七个数字标出)选定的图像区域(去掉了相互重叠的样本区域).当采用数据库中满足 $P(s|\boldsymbol{Ve})>0.05$ 的图像时,我们的算法可以保证 91% 的目标将会出现在集中注意区域中.



($a$)输入图像 ($b$)$\boldsymbol{Vi}$ ($c$)目标检测结果

图 5 目标检测的实验结果

($a$)Itti实验结果



($b$)本文实验结果

图 6 Itti算法和本文算法的对比实验结果

图 6 是 Itti 算法(图 6($a$))和本文算法(图 6($b$))的对比实验结果.当感兴趣的目标不是图像中最显著的目标时,Itti 算法得到的显著区域包含了一些其它目标,而本文算法的目标检测没有受到那些目标的影响.原因是 Itti 算法没有考虑高层信息,注意的集中仅受低层特征显著性的影响.本文算法用 where 信息作为高层信息指导与低层 what 信息相关的注意控制,从而减小了搜索区域,避免将计算资源花费在根据经验目标不太可能出现的空间位置,提高了目标检测的可靠性.

## 6 结论

本文提出了一种新的基于 what 和 where 信息的目标检测方法.应用于多幅自然图像的实验均取得了较为满意的实验结果.与现有算法相比,本文算法具有以下几个突出的特点:(1)采用 where 信息作为自顶向下的注意控制信息,指导自底向上的注意.现有算法出发点都是基于自底向上注意的数据信息,缺乏对高层信息的有效定义.本文定义的 where 信息不是以目标为中心的局部信息的简单叠加,而是将整个场景看成一个独立目标得出的真正意义上的环境信息,它不包含特定目标的独立信息,又没有抛弃任何属于目标的度量.从而为自底向上的注意提供可靠的先验.(2)本文算法将自顶向下的注意控制分为两个阶段,在预注意完成后根据条件就可以停止整个检测过程,从而在很大程度上节约计算资源.将集中注意的结果与 what 信息相结合,为将注意集中到与目标相关的显著区域提供了有效机制.

本文算法没有考虑根据 what 信息进行知觉编组,而这是生物视觉中普遍存在的现象.因此,我们下一步的工作将研究用知觉编组作为连接 what 信息和 where 信息的桥梁,以更好地将自顶向下的注意与自底向上的注意相结合.此外,根据视觉注意过程中资源共享,考虑将环境类别信息用于新环境的类型预测,也是我们今后需要完成的工作.

**参考文献:**


[1] Creem S H, Proffitt D R. Defing the cortical visual systems: "what", "where", and "how"[J]. Acta Psychologica, 2001, 107:43 – 68.

[2] 马尔著,姚国正,刘磊,汪云九译.视觉计算理论[M].北京:科学出版社,1988.1 – 5.
Marr D. Yao G Z, Liu L, Wang Y J. Vision[M]. Beijing: science press, 1988.1 – 5.(in Chinese)

[3] Itti L, Koch C. Computational modeling of visual attention[J]. Nature Reviews Neuroscience, 2001, 2(3): 194 – 230.

[4] Itti L. Models of bottom-up attention and saliency[A]. Neurobiology of Attention[C]. San Diego, CA: Elsevier, 2005. 576 – 582.

[5] 张鹏,王润生.基于视点转移和视区追踪的图像显著区域检测[J].软件学报,2004,15(06):891 – 898.
Zhang P,Wang R S.Detecting Salient Regions Based on Location Shift and Extent Trace[J].Journal of Software,2004,15(06):891 – 898.(in Chinese)

[6] Frintrop S, Rome E. Simulating visual attention for object recognition[A].Proceedings of the Workshop on Early Cognitive Vision[C].Isle of Skye,Scotland,2004.

[7] Sun Y,Fisher R.Object-based visual attention for computer vision[J].Artificial Intelligence,2003,146(1):77 – 123.

[8] Ouerhani N.Visual attention:from bio-inspired modeling to real-time implementation[D].Switzerland: Institute of Micro technology,2003.

[9] 龙甫荟,郑南宁.一种引入注意机制的视觉计算模型[J].中国图象图形学报,1998,3(7):592 – 595.
Long F H,Zheng N N.A visual computing model based on attention mechanism[J].Journal of Image and Graphics,1998,3(7):592 – 595.(in Chinese)

[10] Rybak I A,Gusakova V I,Golovan A V,Podladchikova L N,Shevtsova N A.A model of attention-guided visual perception and recognition[J].Vision Research,1998,38:2387 – 2400.

[11] Salah A A,Alpaydin E,Akarun L.A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition[J].IEEE Trans on Pattern Analysis and Machine Intelligence,2002,24(3):420 – 425.

[12] Itti L.Models of bottom-up and top-down visual attention[D].Pasadena:California Institute of Technology,2000.

[13] Navalpakkam V, Itti L. A goal oriented attention guidance model[J].Lecture Notes in Computer Science,2002,2525:453 – 461.

[14] Henderson J M.Human gaze control during real-world scene perception[J].Trends in Cognitive Sciences,2003,7(11):498 – 504.

[15] Hyvärinen A.Fast and robust fixed-point algorithms for independent component analysis[J].IEEE Transactions on Neural Network,1999,10(3):626 – 634.

[16] Itti L, Koch C, Feature combination strategies for saliency-based visual attention systems[J].Journal of Electronic Imaging,2001,10(1):161 – 169.

[17] Lee T W, Lewicki M S. The generalized Gaussian mixture model using ICA[J].International Workshop on Independent Component Analysis,2000,239 – 244.

[18] Liu Y H,Luo S W,Li A J,Yu H B.A new model selection criterion based on information geometry[A].7th Intern.Conf on Signal Processing[C].Beijing:IEEE press,2004.2:1562 – 1565.

[19] Bilmes J A.A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models[R].Berkeley:University of Berkeley,1998.

**作者简介:**

**田　媚**　女,1980 年生于陕西省,现为北京交通大学计算机与信息技术学院博士研究生,主要研究领域为神经计算,模式识别.
E-mail:tmlily@126.com

**罗四维**　男,1944 年生于北京,博士,博士生导师,北京交通大学计算机与信息技术学院教授,主要研究领域为神经计算,并行处理.

**廖灵芝**　女,1978 年生于四川省,现为北京交通大学计算机与信息技术学院博士研究生,主要研究领域为为神经计算.