基于局部视觉感知及语义联想的图像理解模型

周海英[1,2] 穆志纯[1]

(北京科技大学自动化学院 北京 100083)[1] (中北大学电子与计算机科学技术学院 太原 030051)[2]

摘 要 在视觉感知获得图像视觉信息的基础上,提出了一种自底向上的视觉搜索与自顶向下的语义判定相联系的模型,该模型使图像区域和图像块与图像语义和图像类别之间建立感知联系,模拟注意焦点的移动,通过动眼扫描和搜索图像区域,对感兴趣的内容进行索引和存储,并在任务事件的驱动下进行联想,使图像理解融入了视觉感知和语义解释两个方面,符合人类的认知规律。

关键词 视觉感知,图像理解,语义联想

中图法分类号 TP391.4 **文献标识码** A

Image Understanding Model Based on Local Visual Perception and Semantic Association

ZHOU Hai-ying[1,2] MU Zhi-chun[1]

(School of Automation and Electrical Engineering, University of Science & Technology Beijing, Beijing 100083, China)[1]

(School of Electronics and Computer Science and Technology, North University of China, Taiyuan 030051, China)[2]

Abstract On the basis of image information from visual perception, a model integrated with bottom up visual searching and top down semantic determining was proposed, establishing a perception contact between image areas or image blocks and image semantic or image categories. Through the scanning and searching for image areas, the moving eyes simulate the movement of the focus of attention to index and memory the interested content so as to produce association under the driving of task events. The model blends two aspects of visual perception and semantic explanation in image understanding, which conforms to the human cognitive law.

Keywords Visual perception, Image understanding, Semantic association

人类对外部世界的认识是一个从感知到理解的复杂过程,这个过程的初级阶段是人眼视网膜对外界对象的感知过程[1]。基于感知的图像搜索建立在模拟人眼视网膜对外界对象的感知和认知的基础上,包括对图像内容的感知搜索与存储、图像语义的关联和分类。在基于用户任务要求的图像搜索中还包括视觉搜索路径的控制、搜索的局部内容的保存、潜在概念语义结点的形成与关联等。图像库中图像的信息融合与反馈可实现由视觉到类别、特征到结点的查询,提高了图像检索的柔性,更能符合用户检索的真实意图。

为了模拟人眼对外界事物的观察过程,本文提出了一种"动眼搜索模型",它是对图像中的一些对比显著的位置进行特征采样的搜索方法,因为人眼的预注意焦点大都是变化剧烈的区域。

1 图像的感知

人类视觉系统由于视觉搜索的导向作用[2],观察外部图像时注意焦点的移动具有一定的规律,对比度较大的区域优先获得焦点,构成了焦点移动的主序列,利用注意焦点处的底层特征鉴别对象并建立对象间的联系,为图像的模式分类和识别提供信息基础。

人眼的视觉观察是高度结构化的,统计分析能够揭示自然图像的鲁棒的不变性[3,4]。由于邻近的像素有较强的关联性,利用这种不变性可以对视觉刺激有效地编码[5,6],视觉感知模型中注意窗中心的基本边缘信息结合环境边缘信息构成了图像的局部信息的基础[7]。

通过感知模型模拟人的注意焦点的移动(这里称为"动眼"),在图像上进行扫描和搜索,把动眼感知到的感兴趣的内容进行索引和存储,以便在任务事件的驱动下进行联想,在大量训练样本的基础上建立起图像(或图像块)与图像类别或概念之间的索引联系[8,9]。任务事件是具有固定特征的一组索引块。在探测阶段可建立不同的感知线索,包括像素级线索、区域级线索(以区域为索引单位)和高层语义级线索。当"动眼"对图像进行基于任务的扫描时,各个索引块之间形成一定的交互联系,以便在索引阶段对事件进行记忆。在搜索阶段能够快速地根据任务特征激活相关索引事件,事件的联想通过任务事件的索引项的匹配实现。基于语义联想的图像搜索通过装入能够表达任务事件的索引特征作为"索引结点",在"动眼"对图像进行区域搜索时,阈值水平较高的位置与搜索任

到稿日期:2012-09-15 返修日期:2012-12-11 本文受国家自然科学基金(61170116)资助。

周海英(1962-),男,博士生,教授,主要研究方向为模式识别、视觉感知与图像处理,E-mail:zhydsrzxy@yahoo.com.cn;**穆志纯**(1952-),男,教授,博士生导师,主要研究方向为生物特征识别、认知心理、过程建模与过程控制。

务形成“同步关联联想”。利用视觉索引可实现有关图像的情景联想，通过指针链接建立“场景语义关联表”(见表1)。

表1 场景语义关联表

场景标号	关联图像集	语义群	联想链表指针	索引指针
ID01	ImSet1	Keywords1	AL-pointer	DB_pointer
…	…	…	…	…

例如，由“boat”、“sand beach”、“people”、“water”可以联想到“海滨风光”等图像类别：“boat”、“sand beach”、“people”、“water”→::“Seaside scenery”。

不同类别的图像的内容均由语义对象组成。语义对象具有视觉特征并与其它对象以潜在的联系而存在。视觉目标对于图像语义对象的频繁联系模式构成了分类模型，形成“语义群”。语义群内的语义关键词以一定的概率联系到某个分类模型，关联结点组成的视觉索引形成动眼探测的基本路径。

2 图像分割与语义提取

2.1 图像分割

图像的自动分割是视觉信息获取的基本手段。图像在像素级进行图像语义处理并不方便，以图像区域为对象来进行图像检索和图像理解的高层次处理成为较自然的选择。大多数的图像分割都是建立在底层视觉特征分析的基础上，利用图像在区域中的像素具有某种同质性的特点通过聚类分析实现。基本过程包括：(1)探测图像的区域的基本边缘；(2)对图像区域在同质性探测的基础上发现闭合的边界；(3)对多个同质性区域进行融合增长，以获得语义区域的精确范围。图1给出了图像中分割出的主要目标对象。

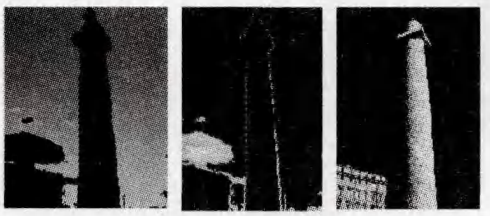


图1 分割图像中的主要对象

2.2 图像片关联

为了对动眼凝视点区域的感知图像进行计算，每个图像被分割成 $M\times N$ 个重叠的图像块。重叠的图像块的好处在于允许一定的模糊性融入空间分布。

例如，在大小为 23×23 的 N 个图像片中，任意两个中心点 (i_1,j_1) 和 (i_2,j_2) 的关联度定义为：

$$r(i_1,j_1;i_2,j_2)=N^{-1}\left[\sum_k (I_{i_1j_1k}-\bar{I}_{i_1j_1})(I_{i_2j_2k}-\bar{I}_{i_2j_2})\right]^{1/2} \tag{1}$$

式中，$\bar{I}_{i_1j_1}=N^{-1}\sum_k I_{i_1j_1k}$，$I_{i_1j_1k}$ 是第 k 个图像片区域中的视觉注意中心 (i_1,j_1) 的强度值。

为探测变化显著的位置，引入对比度(C)的计算(式(2))，即图像片内部的局部像素灰度标准差对于整个图像的平均灰度归一化后的结果，它有助于选择对比度相对较高的区域作为视觉预注意中心。

$$C=\bar{I}^{-1}N^{-1}\sum_k\left[\sum_{(i,j)\in G_k}(I_{ij}-\bar{I}_k)^2\right]^{1/2} \tag{2}$$

设 $G_1,G_2,\cdots,G_N$ 是 N 个视觉注意中心 (i,j) 所在的图像片，I_{ij} 是 (i,j) 处的像素灰度，$\bar{I}$ 是整个图像的灰度均值，$\bar{I}_k$ 是第 k 个片的灰度均值，规范化的关联函数为：$\rho(i,j,i',j')=r(i,j,i',j')/r(0,0,0,0)$，其中 $r(0,0,0,0)$ 为中心像素的自关联度。关联性反映了图像区域像素点之间的同质性，利用图像片间关联性可以对多个关联位置的采集结果进行投票决策，图2显示了图像片之间关联性较强的区域。

图2 图像片关联性较强的区域

2.3 图像区域的色彩同质性判定

对色彩近似的区域进行视觉语义判断是通过计算区域色彩相似度(RCL)来衡量的。

假设两个区域的色彩分别为 DC_1 和 DC_2：

$$\begin{aligned} DC_1&=\{(c_{1i},p_{1i})\},i=1,2,\cdots,N_1\\ DC_2&=\{(c_{2i},p_{2i})\},i=1,2,\cdots,N_2 \end{aligned} \tag{3}$$

式中，c_i 为色彩空间中对应的色彩三维向量，在RGB色彩空间中取值；p_i 为对应于色彩 i 的像素百分比，N 为色彩种数。两个彩色区域之间的距离(D)计算如下：

$$D(DC_1,DC_2)=\sum_{i=1}^{N_1}p_{1i}^2+\sum_{j=1}^{N_2}p_{2j}^2-\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}2a_{1i,2j}p_{1i}p_{2j} \tag{4}$$

式中，$a_{k,l}=\begin{cases}1-d_{k,l}/d_{\max}, & d_{k,l}\leqslant T_d\\ 0, & d_{k,l}>T_d\end{cases}$，为两种色彩 c_k 和 c_l 之间的“相似系数”。其中 $d_{k,l}=\|c_k-c_l\|$ 是两种色彩的欧式距离；T_d 是两种色彩的主观认定相似的最大阈值，由实验给定。

2.4 像素级到区域级的视觉语义联系

设一个图像片在 (x,y) 处像素取 c 值时针对某语义图像的“像素-语义”判定概率定义为：

$$RCL(x,y)=p(c)$$

式中，$p(c)=\dfrac{m(c)}{M(c)}$。其中，$m(c)$ 表示在标准语义训练图像集中，某语义图像像素值取 c 的像素总数，$M(c)$ 为所有语义训练图像集中取像素值 c 的总数。

在“像素-语义”判断的基础上，区域 R_i 相对于某个潜在图像语义的相似度判定(α_i)定义为：

$$\alpha_i=\frac{\sum_{(x,y)\in R_i}RCL(x,y)}{S} \tag{5}$$

式中，S 为区域 R_i 中的像素总数。

$$\begin{cases}\alpha_i\geqslant\tau, & \text{激活 } R_i \text{ 的语义预测}\\ \alpha_i<\tau, & \text{忽略 } R_i \text{ 的判定}\end{cases}\quad(\tau\text{ 为判定阈值})$$

3 图像搜索与图像分析

3.1 自顶向下的动眼搜索

根据已有的知识进行动眼的目标搜索策略中可能会有多个

搜索任务和搜索目标。在自顶向下的知识导引下的理想搜索中，假如目标位置具有相等的先验概率，则目标相对于背景的对比度是平滑的，在一个潜在位置处目标出现的后验概率为：

$$p_i=\frac{p_{prior}(i)\exp(v_i^2w_i)}{\sum_j^n p_{prior}(j)\exp(v_j^2w_j)} \tag{6}$$

式中，$p_{prior}(i)$是潜在位置 i 处目标出现的先验概率，这样的潜在位置设为 n 个；v_i 是该位置的对比度；w_i 是第 i 个潜在位置的预测响应：

$$w_i=\begin{cases}0.5, & \text{目标出现在 } i \text{ 位置}\\ -0.5 & \text{否则}\end{cases}$$

对于动眼移动的控制，其下一个位置以各个图像片的分布熵(H)最大为依据，同时兼顾自顶向下的知识的导引：

$$H=0.5[\sum_{i=1}^{n}\log(1+\lambda_i/\tau)+n\log(2\pi e)] \tag{7}$$

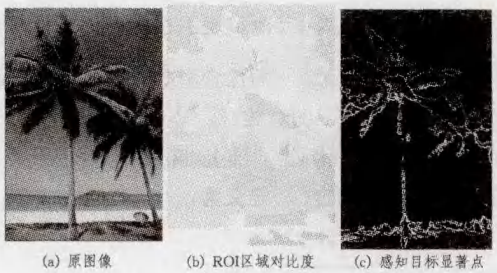
式中，λ_i 为每个图像片的中间部分的像素协方差矩阵的特征值；τ 为与像素灰度量化值有关的噪声水平(一般取 $\tau=0.05$ 或 0.2 两种水平)。

在目标搜索中由关联系数测量对比变量之间的相似程度(ce)，诱导预测兴趣目标，利用式(8)计算“视觉-语义”关联系数(ce)来对比两个图像片。

$$ce(d_f,d_p)=\frac{\sum_{p\in Sc}\dfrac{\mathrm{cov}(d_f,d_p)}{\sigma_f\sigma_p}}{|Sc|} \tag{8}$$

式中，d_f 表示动眼在凝视位置中心的图像片密度，d_p 是预测概念群(Sc)中某图像片的密度，$|Sc|$为该群组中的图像的数量，σ_f，σ_p 分别表示两种密度的方差，$\mathrm{cov}(d_f,d_p)$是两种密度的协方差。

为了模仿人对图像的理解，在感知发生后进行视觉解释是非常重要的，它相当于对感知内容进行结构性简化，突出了兴趣点集中的地方而略去了其它对理解不重要的地方[10]。在图 3 中，兴趣区域(ROI)主要集中在两个部分，一个是上半部分的树枝及叶子部分，中间部分是树干部分，另一个部分是底部的地面，其它部分作为图像背景来处理。存储了缩略信息实际就存储了针对图像的理解信息：“树枝区域-树干-地面”。



(a) 原图像　(b) ROI区域对比度　(c) 感知目标显著点

图 3

3.2 新概念的发现

当与动眼所在位置大致相同的视觉区域被分割出来时，事实上就产生出一个潜在的兴趣区域，该区域相对于整个图像的面积比决定是否值得关注或定义，低于阈值的区域看成是噪声，面积超过阈值的区域返回探测结果并与存储的图像块数据库特征进行相似性对比(式(9))，以判断其是否为已有的语义概念：

$$Sim(c_i,c_j)=S(I(c_i,c_j)) \tag{9}$$

式中，$I(c_i,c_j)=\log(\dfrac{p(c_i\cap c_j)}{p(c_i)p(c_j)})$，$S(x)=\dfrac{1}{1+e^{-x}}$。其中，$c_i$，$c_j$ 分别是未知概念和存储的预测潜在概念，$p(c_i)$，$p(c_j)$是它们单个发生的概率，$p(c_i\cap c_j)$是 c_i，c_j 的共生概率，$S(\cdot)$是 Sigmoid 函数。

若为已有的概念，则更新该概念区域对于被分类别之间的关联权值(即视觉概念区域特征相对某类别图像的后验概率)；若为新发现的图像块，则其特征保存到语义图像块数据库中。

3.3 图像概念区的位置关系分析

将新的语义概念区与原有的概念区之间进行比对并确定它们的位置关系，包括上下、左右、前后、分离、嵌入、交叉等关系。由位置关系还可以获得概念关联关系，如概念包含关系、概念平行关系等。在图像中的概念元素位置关系分析的基础上结合图像元素的关键词可以生成一系列的图像“位置-概念”关联关系，它们完善了自顶向下的知识模型。

例如，“sky”与“sea”是“上下关系”，“sky－>grass land”以及“tree－>ground”等是“上下关系”，“sun”与“sky”是嵌入关系(见图 4)，而“sky－>grass land”与“sky－>road”构成概念平行关系；“grass land”和“road”可以被概念“ground”所包含，在有监督的条件下构成概念包含关系。

在关系判断和推理规则抽取的基础上可以建立基于关联的自顶向下的图像知识库模型。

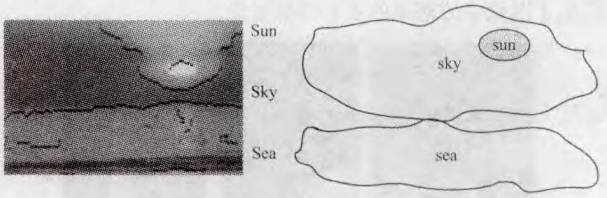


图 4　语义对象区域及位置关系

由于自底向上的图像分析主要借助的是底层的视觉特征如色彩、纹理、朝向、形状等进行判断，而自顶向下的图像分析主要借助的是预测概念之间的各种关系分析，两种模型相结合能够极大地提高图像分析理解的准确性。

3.4 图像内容的框架联系

在给定的图像训练集上，可以通过有监督的学习和反馈归纳出图像内容的理解结构，采用图像语义关键词加内容框架的表达方式：图像内容表达＝框架(Frame)＋关键词(Keywords)(见图 5)。

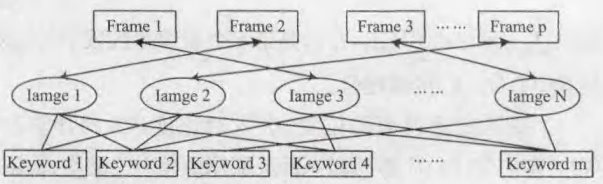


图 5　图像内容表达框架

图像基础框架定义为：

〈图像框架(K)〉::＝Frame〈框架名〉

{〈图像类别 Imclass〉:〈槽值〉
〈元素个数(Elem_num)〉:〈槽值〉
〈主要元素位置关系类型(Inter_pos)〉:〈槽值〉
〈关键词集指针(Keyword_piont)〉:〈槽值〉
〈视觉语义索引项指针(VisIndex_piont)〉:〈槽值〉
〈语义类别联系(Semclass_AKO)〉:〈槽值〉
〈下层子框架的关系指针(Sclass_Instance)〉:〈槽值〉
}

在上述总体框架结构的基础上可以根据不同的处理目的将其进一步拆分成若干子框架，相互关联的多个框架联合构成图像框架总体结构。在建立框架总体结构之后就形成了图像理解的框架知识模型。对图像框架知识库的主要运算包括两种：(a)填槽：对框架中的槽填写内容；(b)匹配：对已知的图像内容寻找合适的框架并将内容填入槽中。

建立了上述运算后引入图像理解框架推理，包括：(1)默认推理，建立父框架与子框架之间的继承关系，填写框架内容的过程中，子框架槽值将自动继承父框架相应的槽值；(2)匹配推理，根据已知的信息通过知识库中存储的框架进行预筛选，通过评价准则找出最匹配的框架。

设一个关键字集合 $Keyword_i$ 对应的聚类原型为 $Cluster_i$：$Keyword_i \rightarrow Cluster_i$，$Cluster_i$ 由一系列样本图像组成：$Cluster_i = \{image_j \mid image_j \in K_i\}$，$K_i$ 为某个类别图像框架。图像类别框架 K_i 的激活值(A_i)由式(10)计算，若超出阈值，则框架与当前查询图像(Q_i)匹配成功。

$$A_i = \frac{1}{1+\frac{p}{M_1}\exp\left(\frac{1}{(n-1)}\sum_{\substack{j=1\\ j\neq i}}^{n} d(Q_i, c_j)\right) - \frac{q}{M_2}\exp\left(\sum_{k=1}^{m}\lambda_k f_k\right)} \tag{10}$$

式中，p 为自底向上的查询权重；q 为自顶向下的查询权重；n 为图像显示的主要对象的个数；f_k 为模型特征二值函数，λ_k 为对应的权重；M_1，M_2 为为归一化常数。其中，$d(Q_i, c_j) = \frac{d_{vis}(Q_i, c_j)}{|keywords(Q_i) \cap R|}(keywords(Q_i) \cap R \neq \phi)$，这里的 $d_{vis}(Q_i, c_j)$ 是 Q_i 与候选框架 R 中的语义词 c_j 之间的视觉距离，采用低层视觉特征向量的欧式距离进行计算；$keywords(Q_i)$ 为查询图像中包含的预测语义关键字。

4 实验

图像理解框架模拟了人的视觉搜索、概念联系和假设形成的过程。实验对一组训练图像在监督条件下进行学习，对视觉目标进行索引标定并建立图像分类框架，对图像中的视觉目标给出不同比率的语义概念索引覆盖，测试同类图像的框架匹配准确率(Matching Accuracy)。实验表明，在各轮反馈训练中，对图像目标所建立的索引覆盖越大，其框架匹配的准确率越高(见图 6)，这种表达机制能够有助于克服图像理解过程中由语义鸿沟造成的障碍。

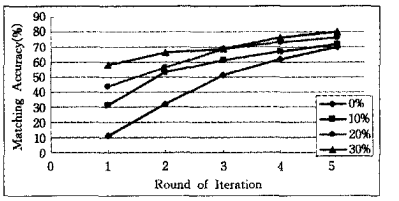


图 6 框架匹配性能

结束语 在训练样本的基础上建立起图像或图像块与类别或概念之间的索引联想。图像的索引过程分为图像探测和索引生成两个阶段。在探测阶段引入不同的感知线索，如像素级线索、区域级线索和高层语义线索。通过学习形成分类和局部模式并与基于视觉特征的参考框架相结合，通过扫描路径上的多个凝视点的耦合使图像识别从部分图像中开始，识别的稳定性随着更多的凝视点的增加而增强，从而部分地克服了“语义鸿沟”造成的障碍。在反馈的基础上强化了相关图像(或图像组)之间的感知联系。这种方法有利于概念和语义的组织，提高了图像理解的准确性，并能够较好地满足用户对某类图像查询细节的要求。

参 考 文 献

[1] Rensink R A. Visual search for change: A probe into the nature of attentional processing[J]. Visual Cognition, 2000(7): 345-376

[2] Chun M M, Jiang Yu-hong. Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention [J]. Cognitive Psychology, 1998(36): 28-71

[3] Eero P. Natural image statistics and neural representation[J]. Neuroscience, 2001, 24(1): 193-216

[4] Gray W D, Fu W T. Soft constraints in interactive behavior: The case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head[J]. Cognitive Science, 2004(28): 359-382

[5] 梁天一，宋国新，虞慧群. 基于稀疏编码的图像语义分类器模型[J]. 华东理工大学学报，2007，33(6)：827-830

[6] Olshausen B A. Sparse coding with an overcomplete basic set: a strategy employed by V1? [J]. Vision Research, 1997, 37(23): 3311-3325

[7] Rybak I A. A model of attention-guided visual perception and recognition [J]. Vision Research, 1998(38): 2387-2400

[8] Gray W D, Fu W T. Soft constraints in interactive behavior: The case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head [J]. Cognitive Science, 2004 (28): 359-382

[9] Smallwood J. The consequences of encoding information on the maintenance of internally generated images and thoughts: The role of meaning complexes[J]. Consciousness and Cognition, 2004 (13): 789-820

[10] Guyonneau R. A key to understanding rapid processing in the visual system [J]. Physiology, 2004(98): 487-497