第 36 卷 第 9 期 计 算 机 工 程 2010 年 5 月
Vol.36 No.9 Computer Engineering May 2010
·多媒体技术及应用· 文章编号：1000—3428(2010)09—0232—03 文献标识码：A 中图分类号：N945

结合识别信息的多目标视频分割

黄叶珏[1]，褚一平[2]
(1. 浙江工业职业技术学院计算机学院，绍兴 312000；2. 杭州电子科技大学计算机学院，杭州 310018)

摘 要：针对实际应用中待分割目标类型已知的情况，提出一种结合识别信息的多目标视频分割算法，使用训练数据集构建目标以及背景的特征字典，计算视频帧的超像素，构造一个分层条件随机场模型，用于约束视频帧的局部邻域和全局邻域，通过求解分层条件随机场模型，获得最终分割结果。实验结果表明，该算法能够对视频中相互遮挡及残缺不全的多个目标进行有效分割。

关键词：多目标视频分割；特征字典；分层条件随机场

Multiple Objects Video Segmentation with Recognition Information

HUANG Ye-jue[1], CHU Yi-ping[2]
(1. School of Computer, Zhejiang Industry Polytechnic College, Shaoxing 312000;
2. College of Computer, Hangzhou Dianzi University, Hangzhou 310018)

【Abstract】The classes of objects are specific in most practical applications, an algorithm of multiple objects video segmentation with recognition information is proposed. The algorithm learns feature dictionary of object and background from training data, and constructs a hierarchical conditional random field model via computing super pixel for video frames, by which the local and global neighboring constraints in video frames are modeled. The final segmentation results are obtained by solving the hierarchical conditional random field model. Experimental results illustrate the algorithm can segment the objects of both occluded each other and the partial objects.

【Key words】multiple objects video segmentation; feature dictionary; hierarchical conditional random fields

1 概述

在视频监控等应用中，视频分割算法[1]一般采用基于背景建模的方法。它主要由 2 个部分组成：(1)视频背景建模算法。典型的有背景累积算法、高斯混合模型方法、非参数估计方法以及基于隐马尔可夫模型的方法。(2)视频帧与当前背景模型对比运算以此获得前景目标。这类运算有简单的阈值化方法、马尔可夫随机场阈值化方法以及条件随机场方法。基于背景建模的分割算法的优点是算法无需拥有前景目标的先验知识，它可以分割出多种类型的前景目标。它们的缺点是：(1)算法要求背景必须是相对固定的。如摄像机必须固定或者晃动的程度较小；前景目标反光导致摄像机自动增益可能会引起算法失效。(2)活动阴影的消除问题。基于背景建模的方法容易把活动阴影误分类为前景目标，所以，需要作专门的处理。由于应用环境的复杂多样性，目前处理活动阴影的算法很难达到对各种环境都保持健壮性。(3)多目标分割问题。由于算法不拥有前景目标的信息，当多个目标相互遮挡或者距离比较近的，寻找连通区域的方法很难把各个目标分割开来。

在实际应用中，需处理的对象一般是已知的。如在视频交通流量统计系统中，需要分割的目标为汽车；在客流量统计应用中，需处理的目标对象为旅客。所以，在这些应用中可以通过学习目标的先验知识来解决上述问题。本文提出一种融合识别信息的多目标视频分割算法。通过对目标特征的学习，建立目标特征字典，结合分层条件随机场模型对局部和全局的邻域进行约束，以此实现多个目标的检测与分割。算法可以对视频中多个目标进行分割，包括相互遮挡的目标以及残缺不全的目标。

2 多目标视频分割算法

文献[2]提出使用 SIFT 不变量特征建立特征字典的方法，实现视频中文本的检索。文献[3]通过对背景和阴影进行建模，利用隐条件随机场对视频的邻域进行约束，实现对视频的前景和阴影的分割。本文提出一种融合识别信息的多目标视频分割算法，利用 SIFT 特征建立特征字典，对文献[2]中的特征字典表示方法进行扩展，增加块掩码、目标特征几何分布以及背景等信息，实现对多个目标的识别和分割。同时，对文献[3]中的隐条件随机场模型进行推广，使之构建在图像的超像素之上，形成一个分层条件随机场模型，实现对局部邻域和全局邻域的约束。

对于视频中的每一帧，通过特征提取算法提取帧图像的特征值，利用这些特征值与特征字典进行匹配。如果该特征匹配为前景目标的特征，则根据字典的目标中心点分布函数估计目标的中心位置。凝聚聚类(agglomerative clustering)算法把中心位置相近的特征聚为同一类。同一类的目标特征根据各自的块掩码，把以特征位置为中心的 32×32 像素块标签为该目标标号，然后构建一个分层随机场模型对这些结果以及相应的局部和全局邻域进行约束，求解出最优的分割结果。

在分层模型中，$X=\{X_1,X_2,\cdots,X_m\}$ 表示视频中一帧视频

基金项目：浙江省教育厅基金资助项目(Y200805048)；杭州电子科技大学校科学研究基金资助项目(KYS055609006)；现代通信国家重点实验室基金资助项目(9140c110206070c11)

作者简介：黄叶珏(1978－)，女，讲师、硕士，主研方向：计算机图像，信息安全；褚一平，讲师、博士

收稿日期：2009-12-20 **E-mail：**yejuehuang@163.com

帧图像的观察值集合，其中，m为帧图像的像素数目。L表示标签集合，称为标签层。在观察集合与标签层之间存在一个隐状态层$H=\{H_1,H_2,\cdots,H_m\}$，它与X的元素一一对应。L与H具有相同的取值范围$\{0,1,2,\cdots\}$，其中，0 表示背景；2 表示第 1 个前景目标；3 表示第 2 个前景目标，以此类推。每个$H_i\in H$的空域邻域表示为$H_j(j\in N_i)$，N_i表示节点i在同层中的邻域。H_i与标签层的邻域表示为$L_j(j\in M_i)$，其中，M_i表示节点i在不同层之间的邻域。与此类似，$L_j(j\in N_i)$表示$L_i\in H$的空域邻域，$H_j(j\in M_i)$表示L_i与隐状态层之间的邻域。H，L以及它们的邻域形成了一个分层的无向图，它们具有马尔可夫属性。

帧图像中颜色相同或者相近的相邻像素可以聚集在一起形成一个超像素，超像素可以看作是更高一级别的图像表示形式。把视频帧图像中颜色相近的连通区域聚在一起形成超像素，当这些超像素足够地小时，超像素的边界与目标的边界是一致的。为了避免求解特征向量，采用多级图划分方法[4]来计算超像素。超像素之内的各个像素之间的邻域形成了局部邻域关系，超像素与超像素之间形成了全局邻域关系。每个超像素由L_i来标签，该超像素内各个像素的标签由H_i表示，它们之间也形成邻域关系。本文定义一个复制函数实现标签层与观察集对应，它实际上是把标签层的标签作为对应超像素中的每一个像素的标签。

2.1 特征字典

使用 SIFT 不变量特征提取算法来提取视频图像的特征值，形成一个 128 维的特征向量，通过 k-mean 聚类算法来构建特征字典。本文的特征字典由目标特征和背景特征 2 种类型的特征构成。特征属于目标特征还是背景特征是根据训练数据集中以特征位置为中心的 8×8 块中的像素标签为目标的数目来确定，如果块中标签为目标的数目大于 30%，则该特征认为是目标特征。特征字典中背景特征条目只保存 128 维的特征向量，而目标特征条目由 128 维的特征向量、32×32 的块掩码以及目标中心点的几何分布函数$C(\cdot)$3g 个部分组成。32×32 的块掩码是根据已标签的训练数据集以特征点位置为中心提取的。$C(\cdot)$由距离参数d和夹角参数θ组成，d表示特征的位置到目标中心点位置的欧氏距离，θ表示d与 SIFT 特征的主方向之间的夹角。在已标签的训练数据集中，根据目标的中心位置可以计算出每个目标特征的距离参数d和夹角参数θ。

2.2 分层随机场模型

若给定一帧视频X，则分层条件随机场模型的概率可以表示为

$$P(L\mid X)=\frac{1}{Z(X)}\sum_{H}\exp\left(-E(L,H,X)\right) \tag{1}$$

其中，$Z(X)$为分配函数，它使得条件概率是一个归一化的值，它的计算公式为$Z(X)=\sum_{L,H}\exp\left(-E(L,H,X)\right)$，$E(L,H,X)$为能量函数，它定义为

$$\begin{aligned}E(L,H,X)=&\sum_k\lambda_k f_1(H_k,X_k)+\\&\nu\sum_i\sum_{j\in N_i}f_2(H_i,H_j,L_i,L_j)+\\&\mu\sum_i\sum_{j\in M_i}f_3(H_i,H_j,L_i,L_j)\end{aligned} \tag{2}$$

其中，λ_k，ν和μ为模型的参数；特征强度函数$f_1(H_k,X_k)$表示单个节点的强度，定义为$f_1(H_k,X_k)=\delta(H_k,H_{k,m})$；$\delta(\cdot)$为 Kronecker delta 函数；$H_{k,m}$为特征匹配字典后根据块掩码标签的值。同层节点邻域之间的强度函数定义为

$$\begin{aligned}f_2(H_i,H_j,L_i,L_j)=&w(H_i,H_j)\times\delta(H_i,H_j)+\\&w(L_i,L_j)\times\delta(L_i,L_j)\end{aligned} \tag{3}$$

其中，像素强度权重为$w(H_i,H_j)=\exp\left(-(X_i-X_j)^2/\sigma\right)$，而$w(L_i,L_j)$是根据$L_i$和$L_j$对应的超像素的平均强度计算而得。不同分层之间的邻域强度为

$$\begin{aligned}\sum_i\sum_{j\in M_i}f_3(H_i,H_j,L_i,L_j)=&\sum_i\sum_{j\in M_i}\delta(L_i,H_j)+\\&\sum_i\sum_{j\in M_i}\delta(H_i,L_j)\end{aligned} \tag{4}$$

给定观察集X以及模型参数λ_k，ν以及μ，可以根据下式推断出分割结果：

$$\hat{L}=\arg\max P(L\mid X;\lambda_k,\nu,\mu) \tag{5}$$

上式可以使用 Gibbs 采样算法或者置信度传播算法进行求解。使用随机梯度下降算法[5]学习模型参数。给定训练数据集$(X^{(i)},L^{(i)})\ i=1,2,\cdots$，参数由以下列公式更新：

$$\begin{aligned}\lambda_k^{(n+1)}&=\lambda_k^{(n)}+\eta^{(n)}\nabla_{\lambda_k}\log P(L^{(i)}\mid X^{(i)})\\\nu^{(n+1)}&=\nu^{(n)}+\eta^{(n)}\nabla_{\nu}\log P(L^{(i)}\mid X^{(i)})\\\mu^{(n+1)}&=\mu^{(n)}+\eta^{(n)}\nabla_{\mu}\log P(L^{(i)}\mid X^{(i)})\end{aligned} \tag{6}$$

其中，$\eta^{(n)}$为学习率，取常数。梯度公式分别为

$$\begin{aligned}\nabla_{\lambda_k}\log P(L^{(i)}\mid X^{(i)})=&\sum_H P(H\mid L^{(i)},X^{(i)})f_1(H,X^{(i)})-\\&\sum_{H,L}P(H,L\mid X^{(i)})f_1(H,X^{(i)})\\\nabla_{\nu}\log P(L^{(i)}\mid X^{(i)})=&\sum_H P(H\mid L^{(i)},X^{(i)})f_2(H_i,H_j,L_i,L_j)-\\&\sum_{H,L}P(H,L\mid X^{(i)})f_2(H_i,H_j,L_i,L_j)\\\nabla_{\mu}\log P(L^{(i)}\mid X^{(i)})=&\sum_H P(H\mid L^{(i)},X^{(i)})f_3(H_i,H_j,L_i,L_j)-\\&\sum_{H,L}P(H,L\mid X^{(i)})f_3(H_i,H_j,L_i,L_j)\end{aligned} \tag{7}$$

其中，$P(H\mid L^{(i)},X^{(i)})$和$P(H,L\mid X^{(i)})$可以由置信度传播算法计算。

2.3 算法流程

算法的整个流程分为训练和分割 2 个部分。训练部分主要是对训练数据提取特征构建特征字典，通过训练数据学习分层随机场模型的参数。分割部分主要是对测试的视频帧提取特征，把提取的特征与字典中的特征条目进行匹配，确定提取的特征是否为目标特征。再通过凝聚聚类算法计算出这些特征属于哪一个目标，使用分层随机场模型对这些特征以及局部和全局邻域进行约束，求解出最终的分割结果。

给定训练数据集$(X^{(i)},L^{(i)})\ i=1,2,\cdots$，$X^{(i)}$表示一帧图像，图像中包含一个目标或者只有背景图像，$L^{(i)}$表示与$X^{(i)}$对应的标签，它对$X^{(i)}$中的每个像素进行标签。训练算法首先确定每帧$X^{(i)}$中是否包含目标，如果包含一个目标，则计算目标的中心点坐标。然后对$X^{(i)}$提取 128 维归一化的 SIFT 特征值，根据特征的位置以及$L^{(i)}$确定该特征是否属于目标特征。如果该特征是目标特征，则计算与目标中心点的距离参数d以及夹角参数θ，提取以特征位置为中心的 32×32 的块掩码。处理完所有的训练数据后，对所有提取的 SIFT 特征值使用 k-means 算法聚类，剔除重复或者相近的特征值，组成特征字典。分层随机场模型的参数通过式(6)进行增量式学习获得。首先对训练数据集$(X^{(i)},L^{(i)})\ i=1,2,\cdots$中的$X^{(i)}$使用多级图划分方法计算超像素构建分层随机场模型，再通过式(6)估

计模型参数，直至训练数据集中的每一个数据训练完成。

训练过程如下：

(1)根据训练数据提取特征，计算 128 维的 SIFT 特征向量，根据特征的位置以及相应的标签确定该特征是属于前景目标还是背景。如果是前景目标，则提取掩码块，计算距离参数 d 和夹角参数 θ。

(2)对所有的 SIFT 特征向量使用 k-means 聚类算法生成特征字典。

(3)根据式(6)学习分层随机场模型的参数。

分割过程如下：

(1)对每一帧测试图像提取所有的 SIFT 特征，把这些特征与特征字典进行匹配，找到字典中距离最近的特征。如果距离在阈值允许的范围，则这个匹配是有效的，如果它属于前景目标，则估计所属目标的中心点坐标。

(2)根据估计的目标中心点坐标对所有的匹配为前景的特征应用凝聚聚类算法。剔除元素个数少于阈值的类别，可以得到实际的目标个数。根据聚类的结果以及相应的掩码，对对应的像素进行标签。

(3)计算超像素，构建分层随机场模型，根据第(2)步的标签计算模型的特征函数，根据式(5)解出最终的分割结果。

经过训练获得特征字典和分层随机场模型的参数之后，就可以根据这些数据对每帧的视频图像进行分割。对每一帧视频图像应用SIFT特征提取算法计算得128维归一化的特征向量。使用这些特征向量与特征字典的特征进行匹配。特征的匹配通过欧氏距离作为度量，搜索出特征字典中与之距离最近的条目。如果它们的距离在给定的阈值范围，则认为它们是匹配的特征。由此可以确定视频帧中的各个匹配后的特征是属于前景目标还是背景。

对所有属于前景目标的特征，根据与特征字典的匹配条目估计出所属目标的中心点坐标。采用凝聚聚类算法对这些估计出的坐标进行聚类，剔除元素少于给定阈值(一般取 5~6)的类别，就可以确定帧中目标的数目以及这些目标对应的特征，再根据特征字典中的 32×32 的块掩码对特征位置为中心的 32×32 像素块标签为对应的目标号。标签完所有的对应特征后，剩余的像素都标签为背景。

采用多级图划分方法计算超像素构建分层随机场模型。每个超像素中的各个像素点的四邻域组成局部邻域关系，这些邻域关系共同组成分层型的隐状态层。超像素与超像素之间的邻域关系为全局邻域关系，它们共同组成标签层。根据式(2)对上述标签的结果计算各个像素的特征函数以及它们之间的局部和全局的特征函数。这些特征函数通过分层随机场模型进行约束，根据式(5)使用 Gibbs 采样算法或者置信度传播算法对模型进行推断，可以求出最优的分割结果。使用复制函数把分层随机场模型中标签层的标签映射到视频帧图像的每个像素中。

3 实验结果

程序由 Visual C++2005 编写，测试视频的尺寸为 320×240。分层条件随机场模型的隐状态层采用图像的四邻域关系。在分割结果中不采用任何的滤波器进行滤波，其中，黑色表示背景，其他不同的颜色表示不同的目标。

训练数据集是由正反两面以及不同视角拍摄的车辆图片组成，本文的训练集包括正面车辆图片 99 张和反面车辆图片 164 张，这些图片有不同视角拍摄的一个目标，另外 9 张不包含目标的背景图片。训练集中，每一张原始图片对应一张二值的标签图片，这些标签图片由手工完成标签工作。测试数据集是通过摄像机在不同场景下拍摄的视频，大约测试了 12 800 帧左右不同场景下的视频。

图 1 为不同场景下各个目标分离情况下的分割结果，图中视频为连续 2 帧~3 帧的对应结果。在分离的情况下，分割的结果比较理想。图 2 为视频目标之间存在相互遮挡或者目标之间距离很近情况下的分割结果。可以看出，本文的算法可以正确地处理该情况下的目标分割。图 3 为视频目标存在残缺情况的分割结果，由于算法在中心点估计时会剔除凝聚聚类结果中元素数目少于给定阈值的类别，因此当残缺目标过小时，它们会被剔除掉。

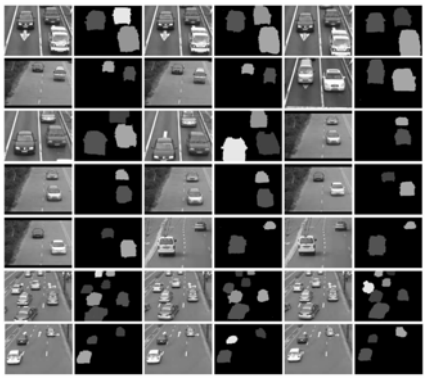

图 1　视频目标间不存在相互遮挡情况下的分割结果

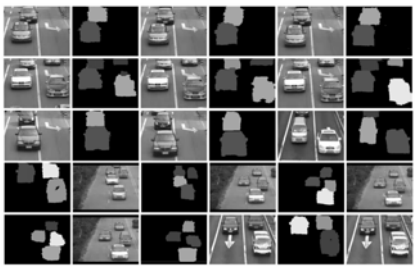

图 2　视频目标间相互遮挡或目标间距离很近情况下的分割结果

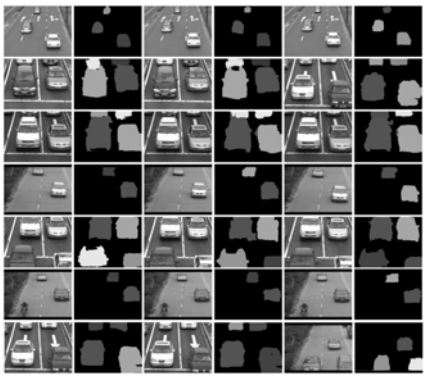

图 3　视频目标存在残缺情况的分割结果

4 结束语

本文提出一种融合识别信息的多目标视频分割算法，算法可以完成对视频中包括相互遮挡以及残缺等多种特殊的目标进行分割。通过特征字典匹配来确定目标特征，利用凝聚聚类算法确定特征所属目标的类别，不必专门处理活动阴影。通过构建基于超像素的分层条件随机场模型来约束局部邻域关系和全局邻域关系，进一步提高分割的质量。

（下转第 237 页）