基于中心宏块的视频目标跟踪算法

肖国强 康 勤 江健民 张贝贝
(西南大学计算机与信息科学学院 重庆 400715)

摘 要 目前的视频目标跟踪算法对目标的不精确分割十分敏感,从而影响目标跟踪的性能.文中提出一种新的视频目标跟踪算法,该算法对目标的过分割或欠分割有较强的鲁棒性.文中提出的跟踪算法中引入了一个中心宏块的概念,通过两个层次的相似性度量,以建立相邻帧之间目标的对应关系.同时利用 MPEG 的运动估计技术和 Kalman 滤波技术来提高目标跟踪的性能.第一个层次的相似性度量是通过 SAD 值在中心宏块之间进行局部纹理匹配;第二个层次是用描述目标内部结构的方向矢量建立目标间的对应关系.实验结果表明,文中提出的算法对于不精确分割的目标能够成功地进行跟踪,同时,对于目标的遮挡、形变、出现、消失以及光线的影响有较强的鲁棒性.

关键词 目标跟踪;视频处理;视频分割;Kalman 滤波;中心宏块
中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2011.01712

Tracking Video Object Based on Central Macroblocks

XIAO Guo-Qiang KANG Qin JIANG Jian-Min ZHANG Bei-Bei
(*College of Computer and Information Science, Southwest University, Chongqing* 400715)

Abstract Since the approaches suggested so far for video moving object tracking are sensitive to the accuracy of object segmentation, we propose a new video object tracking algorithm that provides the strength of robustness to the problems of both under-segmentation and over-segmentation. The proposed tracking algorithm introduces a concept of central macroblock, which is used to establish the correspondences between objects inside neighboring frames via two levels of similarity measurement and observations. Furthermore, MPEG motion estimation and compensation and Kalman filtering techniques are also exploited to enhance the tracking performances. While the first level of similarity measurements is limited to local texture matching via SAD (Sum of Absolute Differences) values between central macroblocks, the second level is established upon objects via directional vectors, characterizing the internal structure of the segmented objects. And both levels of similarity measurements are integrated by Kalman filtering. Experimental results carried out on PETS2001 and PETS2006 database show that the proposed algorithm achieves successful tracking performances robust to inaccuracy of object segmentation as well as other distracting factors such as occlusion, deformation, lighting effect, object disappearances and appearances etc.

Keywords object tracking; video processing; video segmentations; Kalman filtering; central macroblocks

---

收稿日期:2008-10-21;最终修改稿收到日期:2009-11-21.本课题得到重庆市自然科学基金(CSTC-2008BB2252)资助.**肖国强**,男,1965 年生,博士,教授,主要研究领域为基于内容的数字媒体处理、基于语义的视频分析、视频分割、模式识别和机器学习等. E-mail: gqxiao@swu.edu.cn. **康 勤**,女,1965 年生,高级工程师,研究兴趣为图像处理. **江健民**,男,1957 年生,博士,教授,主要研究领域为图形图像处理、视频编码、视频分析、计算机视觉和机器学习等. **张贝贝**,女,1983 年生,硕士,研究兴趣为视频处理.

1 引　言

基于内容的视频处理需要提取视频运动目标，它对视觉内容、事件和相关知识的说明和理解起到至关重要的作用. 视频目标提取主要通过目标的分割和跟踪来完成，以保证运动目标在相邻的帧中被准确地提取出来. 由于目标分割存在许多尚未解决的问题，目前还没有可靠的方法能够实现精确的视频目标分割. 因此，研究非精确分割的视频目标跟踪算法，并提高其目标跟踪的鲁棒性，这对于基于内容的视频处理、分析、诠释和应用有重要的意义.

2 相关研究

视频目标跟踪的方法可以分为 4 类：基于模型[1]、基于外观[2]、基于轮廓[3]和基于特征[4-6]的目标跟踪方法. 基于模型的跟踪方法是利用给定场景中典型物体的先验知识[1,5]，这种方法计算量大，且缺乏泛化性. 基于外观的方法依赖于视频目标的二维形状区域提供的信息[7-9]，如运动、颜色和纹理等信息，来实现目标的跟踪. 由于这些信息都是低层次的特征信息，因此，这些方法通常不能解决运动目标复杂的形变问题. 基于轮廓或网格(meshes)的方法依赖于目标的轮廓或二维网格. 这种方法利用运动信息映射出轮廓，然后将其应用到下一帧的目标分割中，从而完成目标的检测[10]. 但这种方法计算复杂，运算量大，不能处理非刚性物体的大范围运动. 对基于轮廓方法的改进是采用主动轮廓模型，如 snakes[11]和 meshes[3,12]. 基于特征的方法也有一些报道，但它们并非针对视频目标跟踪，如文献[13]提出的视频对象跟踪算法. 这些算法的主要问题是如何提取特征并确定特征与目标之间的关系. 因此，这些算法可靠性不高，容易出现跟踪错误.

另一方面，国际上已发布了一系列的视频编码标准，如从 MPEG-1 到 MPEG-4，从 H.261 到 H.264 等，来应对日益增长的视频数据. 早期的视频编码标准缺乏对高层次视频内容的描述，因此，在最近发布的标准中引入了视频对象层(VOL)的概念来支持基于内容的视频功能. 面向对象的多媒体内容的表示为用户提供了基于内容的访问和管理的灵活性，从而使基于运动目标的视频处理引起了业界的广泛关注，如基于对象的视频编码，基于对象的视频内容分析、检索和视频目标跟踪[1,2,4-5,7-11]. 因此，视频对象分割在基于内容的视频应用中，如视频对象跟踪、基于内容的视频检索、视频标注等，起到关键的作用. 然而，目标分割目前仍是一个尚未解决的问题，文献报道的目标分割算法都存在过分割或欠分割的问题[7-10,14-15]，因此，不能够提供可靠的视频目标分割.

本文提出一种在复杂背景下自动跟踪视频目标的算法，它通过目标分割、区域划分、中心点提取、中心宏块构建和方向矢量确定来实现目标跟踪. 该算法利用区域中心点和方向矢量来解决目标之间的匹配问题，从而避免了由于不精确的目标分割所带来的问题. 该算法的优势得益于在目标跟踪过程中不需要将整个目标区域投影到下一帧，而仅需处理一个 16×16 的中心宏块，因此，避免了运算复杂的运动模型.

3 算法描述

给定输入视频序列$\{I_0, I_1, \cdots, I_{i-1}, I_i, \cdots, I_m\}$，用文献[14]中提出的通过检测相邻两帧像素的变化和利用运动信息进行视频目标分割. 目标分割过程分为 3 步：第 1 步是利用 Canny 算子产生 3 个边缘图，包括当前帧的边缘图 E_n，差分帧$|I_{n-1}-I_n|$的边缘图 DE_n和背景帧的边缘图 E_b；第 2 步，通过分别比较 E_n和 DE_n以及 E_b和 DE_n之间的边缘像素产生两个运动边缘图 ME_n^{change}和 ME_n^{still}，ME_n^{change}包含了所有的当前运动边缘像素，ME_n^{still}包含了从前一帧的运动边缘像素信息中得到的所有静止边缘像素；第 3 步是通过选择 ME_n^{change}和 ME_n^{still}并集($ME_n^{\text{change}} \cup ME_n^{\text{still}}$)的所有边缘像素产生分割的视频对象(VO). 文献[15]对文献[14]的算法进行了改进，通过区域增长使 VO 的分割更加精确，同时也提高了算法的鲁棒性.

给定当前帧 I_i，其中的第 j 个目标用$O_{i,j}(j=0, 1, \cdots, N_F^i-1)$表示，这里 N_F^i表示在第 i 帧中的目标总数. 由于目标分割不能做到百分之百的精确[15]，特别是一个目标消失，其它的目标又同时进入画面，这时将产生不可避免的遮挡，出现过分割或欠分割的问题，两个重叠的目标可能被分割成一个目标或目标的某些部分被丢失. 为了克服这些因素在目标跟踪时带来的负面影响，我们进一步把一个目标分割成纹理一致的不同区域，这些区域之间相互独立，然后分别对每个区域进行跟踪，通过综合利用区域的跟踪信息来实现目标的跟踪，以弥补由于目标分

割的不精确所带来的影响.因此,给定视频帧的目标,对第 j 个目标,利用区域增长方法[16]进一步分割成 N_i^j 个互不重叠的区域,用 $R_{i,j}^k$ 表示第 i 帧中第 j 个目标的第 k 个区域,$k=0,1,2,\cdots,N_i^j-1$.

由于部分区域可能位于被分割的目标之外,我们仅利用一个区域的代表部分进行跟踪,这个代表部分应位于区域的中心.为此,需要从区域中提取一个中心点.区域 $R_{i,j}^k$ 的中心点 $\boldsymbol{C}_{i,j}^k$ 按如下的方法提取[17].首先计算

$$\mu_{i,j}^k=\frac{1}{M}\sum_{l=0}^{M-1}P_{i,j,l}^k \tag{1}$$

其中,$P_{i,j,l}^k$ 表示第 i 帧中第 j 个目标的第 k 个区域的第 l 个区域边界像素亮度值,$\mu_{i,j}^k$ 是区域边界像素 $P_{i,j,l}^k$ 的亮度平均值,M 表示区域的边界像素点总数.把 $\mu_{i,j}^k$ 作为门限,用下式产生第 k 个区域的二值图像 $g^k(x,y)$:

$$g^k(x,y)=\begin{cases}1, & P(x,y)\geq\mu_{i,j}^k\\ 0, & \text{其它}\end{cases} \tag{2}$$

其中,$P(x,y)$ 表示视频帧 $I(x,y)$ 中的像素值.最后,由下列两式得到一个区域的中心点 $\boldsymbol{C}_{i,j}^k$:

$$m_{p,q}^k=\sum_x\sum_y g^k(x,y)x^py^q,\ \ p,q\in\{0,1\} \tag{3}$$

$$\boldsymbol{C}_{i,j}^k=\left(\frac{m_{1,0}^k}{m_{0,0}^k},\frac{m_{0,1}^k}{m_{0,0}^k}\right) \tag{4}$$

以区域中心点为中心,构建一个 16×16 的中心宏块来代表该区域.宏块大小的选择是鉴于 MPEG 中运动估计和补偿技术[18],同时也便于利用它的原则来计算 SAD(Sum of Absolute Differences)值,以达到对中心宏块跟踪的目的.因此,通过中心宏块来建立相邻帧之间区域的对应关系,同时利用 Kalman 滤波技术和从 MPEG 中获得的运动矢量来提高目标的跟踪精度,这样,不仅提高了处理速度,还改善了跟踪的性能.由于目标的跟踪是建立在中心宏块之间对应关系的基础上的,为了保证视频目标的正确跟踪,我们将忽略那些不能够包含一个完整宏块的区域,这些区域可通过下式的条件进行判决:

$$\min\ d(P_{\text{boundary}},\boldsymbol{C}_{i,j}^k)>|\sqrt{128}\eta| \tag{5}$$

其中,$d(P_{\text{boundary}},\boldsymbol{C}_{i,j}^k)$ 表示区域边界点与中心点之间的 Euclidean 距离,η 为参数,取值范围 $1\leq\eta\leq2$.设置该条件的思想是使得选定区域的大小应大于中心宏块,中心宏块的边界点与中心点之间最大的 Euclidean 距离为 $\sqrt{8^2+8^2}=\sqrt{128}$.$\eta$ 用来控制跟踪目标的最小区域面积,$\eta=1$ 对应中心宏块大小的区域.

为了实现从第 $(i-1)$ 帧到第 i 帧对第 k 个目标的跟踪,我们用区域的对应关系来建立目标之间的对应关系,利用 Kalman 滤波来实现区域的跟踪[19-20].众所周知,Kalman 滤波利用状态方程和观测方程来描述动态估计和预测系统[19],我们定义每个中心宏块的状态为

$$\boldsymbol{s}(t)=(x^t,y^t,v_x^t,v_y^t) \tag{6}$$

其中,(x^t,y^t) 表示在时刻 t 中心宏块的位置,(v_x^t,v_y^t) 表示在时刻 t 分别沿 x 和 y 方向的运动速度.对于相邻帧之间的状态连续估计,有 $v_x^t=x^t-x^{t-1}=\Delta_x$ 和 $v_y^t=y^t-y^{t-1}=\Delta_y$,$(\Delta_x,\Delta_y)$ 是第 t 帧中心宏块的运动矢量.由于 MPEG 中采用从上到下,从左到右的方式扫描产生宏块,因此,中心宏块可能与 MPEG 中产生的宏块不重合,我们采用对与中心宏块所重叠的 MPEG 宏块的运动矢量进行加权来获得 (Δ_x,Δ_y).假定中心宏块与 N_{MPEG} 个 MPEG 宏块重叠,N_{MPEG} 最大值为 4.(Δ_x,Δ_y) 由下式确定:

$$(\Delta_x,\Delta_y)=\begin{cases}\Delta_x=\sum\limits_{k=1}^{N_{\text{MPEG}}}w_k\Delta_{x_k}\\ \Delta_y=\sum\limits_{k=1}^{N_{\text{MPEG}}}w_k\Delta_{y_k}\end{cases} \tag{7}$$

其中,$w_k=\frac{1}{64}\sum\limits_{P(x,y)\in B_c\cap B_M}P(x,y)$,$B_c$ 和 B_M 分别代表中心宏块和 MPEG 宏块,$(\Delta_{x_k},\Delta_{y_k})$ 为对应的第 k 个 MPEG 宏块的运动矢量.对于只有帧内编码的帧,其 (Δ_x,Δ_y) 设为 $(0,0)$.

Kalman 滤波的状态方程定义为

$$\boldsymbol{s}(t)=\boldsymbol{F}\boldsymbol{s}(t-1)+\boldsymbol{\eta}(t) \tag{8}$$

其中,$\boldsymbol{F}$ 为状态转移矩阵,表示为

$$\boldsymbol{F}=\begin{pmatrix}1&0&1&0\\0&1&0&1\\0&0&1&0\\0&0&0&1\end{pmatrix} \tag{9}$$

$\boldsymbol{\eta}(t)=(\eta_x(t),\eta_y(t),\eta_{\Delta_x}(t),\eta_{\Delta_y}(t))^{\mathrm{T}}$ 是在时刻 t 的随机噪声矢量,包括位置噪声和运动噪声.随机噪声通常为相互独立的零均值高斯白噪声,因此,其协方差矩阵为一对角矩阵,即 $E[\boldsymbol{\eta}(t)\boldsymbol{\eta}(t)^{\mathrm{T}}]=Q_t$.

根据标准的 Kalman 滤波,在时刻 $t-1$ 对时刻 t 的状态预测为

$$\hat{\boldsymbol{s}}(t|t-1)=\boldsymbol{F}\hat{\boldsymbol{s}}(t-1|t-1) \tag{10}$$

在时刻 t 的状态更新用下式:

$$\hat{\boldsymbol{s}}(t|t)=\hat{\boldsymbol{s}}(t|t-1)+K(t)[\boldsymbol{Z}(t)-\boldsymbol{H}\hat{\boldsymbol{s}}(t|t-1)] \tag{11}$$

这里,$\boldsymbol{H}=\begin{pmatrix}1&0&0&0\\0&1&0&0\end{pmatrix}$ 为测量矩阵,由观测方程

$\boldsymbol{Z}(t)=\boldsymbol{H}s(t)+\boldsymbol{W}(t)$确定，其中，$\boldsymbol{Z}(t)$为观测量，$\boldsymbol{W}(t)$为观测噪声. Kalman 增益 $\boldsymbol{K}(t)=\boldsymbol{P}(t|t-1)\cdot\boldsymbol{H}^{\mathrm{T}}[\boldsymbol{HP}(t|t-1)\boldsymbol{H}^{\mathrm{T}}-R(t)]^{-1}$，其中 $\boldsymbol{P}(t|t-1)=\boldsymbol{FP}(t-1|t-1)\boldsymbol{F}^{\mathrm{T}}+\boldsymbol{Q}(t)$是 $t-1$ 时刻的预测协方差矩阵，它的更新方程为 $\boldsymbol{P}(t|t)=\boldsymbol{P}(t|t-1)-\boldsymbol{K}(t)\boldsymbol{HP}(t|t-1)$. 这是一个基于帧的迭代过程，当 $t=0$ 时，$\boldsymbol{P}(0|0)=E[\boldsymbol{s}(0)\boldsymbol{s}(0)^{\mathrm{T}}]$.

由方程(11)可知，跟踪的状态位置由 Kalman 预测部分和校正部分确定. 状态预测利用 MPEG 运动估计技术，中心宏块的位置观测值 $\boldsymbol{Z}(t)$则利用下述方法得到.

给定第 $i-1$ 帧的第 j 个目标 $O_{i-1,j}$，它与第 i 帧的第 $\bar{j}$ 个目标 $O_{i,j}$ 的区域差分用下述矩阵描述，其中区域差分值用 SAD(Sum of Absolute Differences)表示

$$\boldsymbol{\phi}(j,\bar{j})=\begin{pmatrix} SAD_{0,0} & SAD_{0,1} & \cdots & SAD_{0,N_i^j-1} \\ SAD_{1,0} & SAD_{1,1} & \cdots & SAD_{1,N_i^j-1} \\ \vdots & \vdots & \ddots & \vdots \\ SAD_{N_{i-1}^j-1,0} & SAD_{N_{i-1}^j-1,1} & \cdots & SAD_{N_{i-1}^j-1,N_i^j-1} \end{pmatrix} \tag{12}$$

其中，$SAD=\dfrac{1}{16\times16}\sum_{x=0}^{15}\sum_{y=0}^{15}|P_i(x,y)-P_{i-1}(x,y)|$ 为两个区域的中心宏块的差分值，$P_i(x,y)$为第 i 帧中区域的中心宏块像素值.

由式(12)可知，$\boldsymbol{\phi}(j,\bar{j})$的第一行对应目标 $O_{i-1,j}$ 的第一个区域与目标 $O_{i,j}$ 的所有区域的差分值，即 $\boldsymbol{\phi}(j,\bar{j})$的每一行表示目标 $O_{i-1,j}$ 的一个区域与目标 $O_{i,j}$ 的所有区域的差分值，因此，其每一行的最小 SAD 值代表了目标 $O_{i-1,j}$ 的一个区域在目标 $O_{i,j}$ 的区域中的最佳匹配. 因此，矩阵 $\boldsymbol{\phi}(j,\bar{j})$度量了第 $i-1$ 帧的第 j 个目标 $O_{i-1,j}$ 的所有区域到第 i 帧的第 $\bar{j}$ 个目标 $O_{i,j}$ 的区域的过度过程，其区域的对应关系可表示为

$$\boldsymbol{\phi}_{\min}(j,\bar{j})=(MAC_0,MAC_1,\cdots,MAC_{N_{i-1}^j-1})^{\mathrm{T}} \tag{13}$$

其中，$MAC_k=\min\{SAD_{k,0},SAD_{k,1},\cdots,SAD_{k,N_i^j-1}\}$，$k\in[0,N_{i-1}^j-1]$，它代表了在第 i 帧的第 $\bar{j}$ 个目标中的最佳匹配宏块，对应第 $i-1$ 帧的第 j 个目标的第 k 个中心宏块与第 i 帧的第 $\bar{j}$ 个目标的所有中心宏块之间 SAD 的最小值.

因此，第 k 个中心宏块的观测值 $\boldsymbol{Z}(t)$由对应 MAC_k 宏块的位置确定. 由于第 i 帧中有 N_F^i 个目标，每个目标都将产生这样的观测量，因此，对于第 i 帧将有 N_F^i 个对应 $\{MAC_k^1,MAC_k^2,\cdots,MAC_k^{N_F^i}\}$ 的观测量.

实际上，我们的策略是通过中心宏块的对应关系建立相邻两帧区域之间的对应关系，从而建立相邻两帧目标之间的对应关系. 为了综合考虑区域间的对应关系，并把它们进一步转换为目标间的对应关系，把式(13)中对应的最小 SAD 值全部加起来，构造一个目标之间建立对应关系的代价测度，如下式所示：

$$Cost_{j,j}=\frac{1}{N_{i-1}^j}\sum_{l=0}^{N_{i-1}^j-1}MAC_l \tag{14}$$

$Cost_{j,j}$表示了第 $i-1$ 帧的第 j 个目标与第 i 帧的第 $\bar{j}$ 个目标建立对应关系的综合代价.

由于第 $i-1$ 帧中有 N_F^{i-1} 个目标，第 i 帧中有 N_F^i 个目标，因此，相邻两帧目标之间的对应关系用下面的代价矩阵来度量：

$$Correspondence_{i-1,i}=\begin{pmatrix} Cost_{0,0} & Cost_{0,1} & \cdots & Cost_{0,N_F^i-1} \\ Cost_{1,0} & Cost_{1,1} & \cdots & Cost_{1,N_F^i-1} \\ \vdots & \vdots & \ddots & \vdots \\ Cost_{N_F^{i-1}-1,0} & Cost_{N_F^{i-1}-1,1} & \cdots & Cost_{N_F^{i-1}-1,N_F^i-1} \end{pmatrix} \tag{15}$$

其中，第 1 行表示第 $i-1$ 帧中的第 0 个目标跟踪到第 i 帧中的所有目标的代价. 也就是每一行对应第 $i-1$ 帧中的目标，而每一列对应第 i 帧中的目标.

式(15)说明，第 k 行中最小的代价值 $MCost_{k,l}=\min\{Cost_{k,0},Cost_{k,1},\cdots,Cost_{k,N_F^i}\}$ 对应于第 $i-1$ 帧中的第 k 个目标最有可能跟踪到第 i 帧中的第 l 个目标. 然而，上述的目标对应关系是建立在 SAD 值基础上的，它仅仅说明了两个目标之间的纹理相似性. 为了进一步优化每个中心宏块跟踪位置的观测量，我们提出两个目标之间的结构相似性度量，具体方法如下.

为了表示每个目标的内部结构，首先对每个目标内的区域中心点按从上到下，从左到右的顺序进行标记，然后从第一个区域中心点开始直到最后一个，把区域中心点两两相连，从而构造一个方向矢量的序列. 对第 $i-1$ 帧的第 j 个目标，这个序列可用下式表示：

$$\begin{aligned} DV_{i-1}^j&=\{\boldsymbol{D}_{i-1}^0,\boldsymbol{D}_{i-1}^1,\cdots,\boldsymbol{D}_{i-1}^{N_i^j-2}\} \\ &=\{(C_0\rightarrow C_1),(C_1\rightarrow C_2),\cdots,(C_{N_{i-1}^j-2}\rightarrow C_{N_{i-1}^j-1})\} \end{aligned} \tag{16}$$

其中，$(C_m \rightarrow C_n)$表示目标中第 m 个区域中心宏块到第 n 个区域中心宏块连线的方向. 这样，任意两个目标的结构相似性可用每个矢量的方向来度量. 考虑到目标跟踪过程中可能产生旋转，我们把矢量方向仅量化成 8 个方向，即$\{[0°—45°),[45°—90°),\cdots,[315°—360°)\}$. 虽然，较少的量化方向可以为被跟踪目标的结构变化提供更大的宽容度，如旋转等，但较多的量化方向能够提供更精确的结构相似性度量. 我们通过实验确定了 45°的旋转容限.

给定两个目标的方向矢量序列，比较相对应的方向矢量，看其是否属于同一个量化的方向，据此来调整它们对应的式(15)中的代价值. 具体的调整过程如下：

$$\text{for}(\bar{k}=0,\bar{k}<N_{i-1}^{k},\bar{k}++)\left\{\right.$$

$$Cost_{k,l}=\begin{cases}Cost_{k,l}-\alpha, & \angle \boldsymbol{D}_{i-1}^{\bar{k}}=\angle \boldsymbol{D}_{i}^{\bar{k}}\\ Cost_{k,l}, & \text{其它}\end{cases} \tag{17}$$

其中，$\angle \boldsymbol{D}_{i-1}^{\bar{k}}$表示第 $\bar{k}$ 个方向矢量 $\boldsymbol{D}_{i-1}^{\bar{k}}$的量化方向，$\alpha$ 为步长，用来降低代价值，其取值范围为 1～10.

如果第 i 帧中的一个目标的方向矢量数少于或多于第 $i-1$ 帧中的任一目标的方向矢量数，则相应的代价值增加 α. 通过式(17)的结构相似性度量对式(15)的代价调整完后，找出式(15)中每一行的最小值，即 $\min\limits_{j\in[0,N_F^i-1]}(Cost_{i,j})=Cost_{h,g}$，就能够确定被跟踪的目标. 也就是第 $i-1$ 帧中的第 h 个目标跟踪到第 i 帧中的第 g 个目标. 因此，第 k 个中心宏块最后的观测量是在第 i 帧中对应 MAC_k^g的宏块的位置.

综上所述，本文提出的视频目标跟踪算法通过两个层次实现. 首先，我们建立中心宏块之间的对应关系，以度量目标内区域之间的局部相似性；第二，利用 Kalman 滤波进行基于中心宏块的目标跟踪，同时，综合纹理和结构相似性信息以保证在目标之间建立正确的目标对应关系，从而实现对视频目标的跟踪.

4 实验结果与分析

为了对本文提出的算法进行评估，我们在 Windows XP 环境下，用 VC++编程实现了本算法，并进行了大量的视频目标跟踪实验.

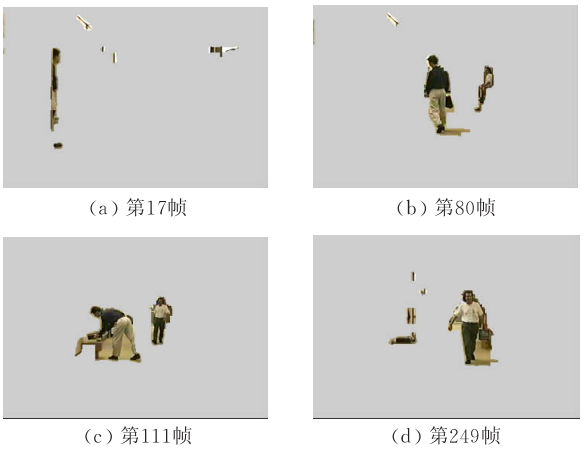
图 1 为对视频序列 Hall Monitor 进行实验的结果. 该序列共 300 帧，每帧大小 352×240. 图 1 中，(a)说明目标 A(用白色方框表示)在第 17 帧开始出现，并被跟踪到；(b)显示在第 80 帧中出现目标 B(用黑色方框表示)，并被跟踪到；从(c)中看出，在第 111 帧中，目标 A 弯下腰拿东西时，形体发生了变化，然而算法依然能够对目标 A 发生形变后实现正确的跟踪；(d)说明在第 249 帧，目标 A 消失，只剩下目标 B，而此时算法能够正确地跟踪目标 B. 图 1 所示的实验结果说明本文提出的算法在以下几种情况下都能够正确地跟踪目标的运动：(1) 从第 17 帧到第 249 帧目标 A 的形体由小到大，再由大到小，这代表了视频中的渐变过程；(2) 一个目标消失或出现，同时另一个目标仍停留在画面中；(3) 目标发生形变.

(a) 第17帧,目标A出现　(b) 第80帧,目标B出现

(c) 第111帧,目标A形变　(d) 第249帧,目标A消失

图 1　Hall Monitor 视频序列目标跟踪实验结果

图 1 中显示的视频帧的目标分割结果如图 2 所示. 从图 2 可以看出，目标的分割并不精确，同时在每一帧中都存在过分割的现象. 例如，第 17 帧中，人物目标 A 仅仅部分地出现在场景中，在目标的分割过程中产生了一些过分割的区域，但这些过分割的区域在目标跟踪的过程中被算法所抑制. 如前所述，这些过分割的区域可利用式(5)的条件进行剔出. 当



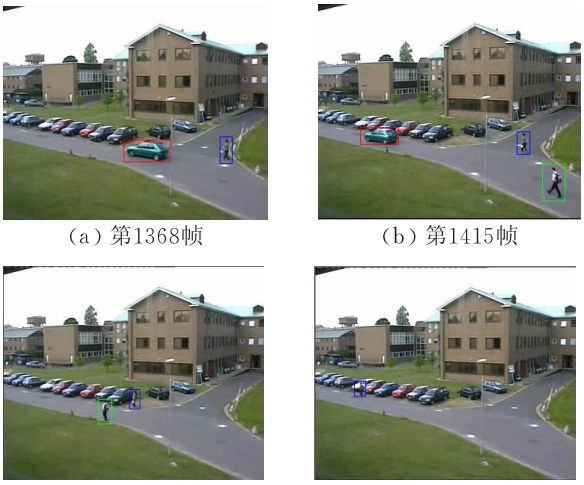
(a) 第17帧　(b) 第80帧

(c) 第111帧　(d) 第249帧

图 2　图 1 中视频帧的目标分割结果

一些背景区域作为被分割目标区域的一部分时，如图 2 中的第 80 帧和第 111 帧，中心宏块的方案将把这些背景区域对目标跟踪的影响降低到最小的程度，特别是人物目标与背景的颜色或纹理相似的情况下，如第 111 帧中人物目标的腿部与背景的颜色相似.

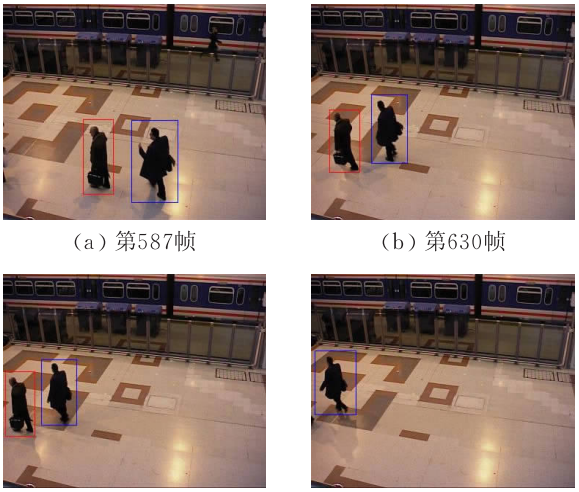
为了进一步说明本文提出的视频目标跟踪算法的鲁棒性，我们下载了大量主要用于目标跟踪的视频序列进行实验，部分实验结果如图 3 和图 4 所示. 图 3 给出了对 PETS2001 监控视频序列的实验结果，在第 1368 帧出现了汽车和行人两个目标，行人比汽车要小得多，但我们的算法仍能够对目标进行正确的跟踪. 图 3 的结果显示，本文算法在全景视图中能够正确地跟踪较小的多视频运动目标. 图 4 给出的是对 PETS2006 视频序列的实验结果，该图说明本文算法在夜晚场景并有强光影响的条件下，也能够正确地跟踪视频目标.



(a) 第1368帧　(b) 第1415帧

(c) 第1606帧　(d) 第1893帧

图 3　PETS2001 视频序列实验结果



(a) 第587帧　(b) 第630帧

(c) 第642帧　(d) 第664帧

图 4　PETS2006 视频序列实验结果

5　总　结

本文提出一种新的基于中心宏块的视频目标跟踪算法，我们的研究工作的新颖性主要体现在以下 3 方面：(1) 在区域层次上引入基于中心点的宏块来进行目标跟踪，克服了由于目标分割的不精确所带来的影响；(2) 通过在目标中引入方向矢量，在目标层次上度量目标之间的结构相似性；(3) 利用 MPEG 的运动估计技术，用 Kalman 滤波综合区域和目标对应关系，利用代价矩阵在相邻两帧的目标之间建立全局相似性度量. 大量的实验证明，本文提出的算法在各种复杂背景的情形下，都能够正确地跟踪目标，并对目标的不精确分割有很好的鲁棒性.

参　考　文　献

[1] Koller D, Danilidis K, Nagel H. Model-based object tracking in monocular image sequences of road traffic scenes. International Journal of Computer Vision, 1993, 10(3): 257-281

[2] Tao H, Sawhney H S, Kumar R. Object tracking with Bayesian estimation of dynamic layer representation. IEEE Transactions on Pattern Analysis Machine Intelligence, 2002, 24(1): 75-89

[3] Zhao J W, Wang P, Liu C Q. An object tracking algorithm based on occlusion mesh model//Proceedings of the International Conference on Machine Learning and Cybernetics. Beijing, 2002: 288-292

[4] Gevers T. Robust segmentation and tracking of colored objects in video. IEEE Transactions on Circuits and Systems for Video Technology, 2004, 14(6): 776-781

[5] Cavallaro A, Steiger O, Ebrahimi T. Tracking video objects in cluttered background. IEEE Transactions on Circuits and Systems for Video Technology, 2005, 15(4): 575-584

[6] Gnsel B, Tekalp A M, Beek P J. Content-based access to video objects: Temporal segmentation, visual summarization, and feature extraction. Signal Processing, 1998, 66(2): 261-280

[7] Meier T, Ngan K. Automatic segmentation of moving objects for video object plane generation. IEEE Transactions on Circuits and Systems for Video Technology, 1998, 8(5): 525-538

[8] Wang D. Unsupervised video segmentation based on watersheds and temporal tracking. IEEE Transactions on Circuits and Systems for Video Technology, 1998, 8(5): 539-546

[9] Marcotegui B, Zanoguera F, Correia P, Rosa R, Marques F, Mech R, Wollborn M. A video object generation tool allowing friendly user interaction//Proceedings of the International Conference on Image Processing. Kobe, Japan, 1999, 2: 391-395

[10] Gu C, Lee M C. Semiautomatic segmentation and tracking of semantic video objects. IEEE Transactions on Circuits and Systems for Video Technology, 1998, 8(5): 572-584

[11] Peterfreund N. Robust tracking of position and velocity with Kalman snakes. IEEE Transactions on Pattern Analysis Machine Intelligence, 1999, 21(6): 564-569

[12] Sun S, Haynor D R, Kim Y. Semiautomatic video object segmentation using VSnakes. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(1): 75-82

[13] Beymer D, McLauchlan P, Coifman B, Malik J. A real-time computer vision system for measuring traffic parameters// Proceedings of the International Conference on Computer Vision and Pattern Recognition. San Juan, Puerto Rico, 1997: 495-501

[14] Kim C, Hwang J N. Fast and automatic video object segmentation and tracking for content-based applications. IEEE Transactions on Circuits and Systems for Video Technology, 2002, 12(2): 122-129

[15] Gao L, Jiang J, Yang S Y. Constrained region-grow for semantic object segmentation//Lecture Notes in Computer Science 4179. Springer, 2006: 323-331

[16] Adams R, Bischof L. Seeded region growing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(6): 641-647

[17] Kirishima T, Sato K, Chihara K. Real-time gesture recognition by learning and selective control of visual interest points. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(3): 351-364

[18] Jiang J, Qiu K, Xiao G. An edge content block descriptor for MPEG compressed videos. IEEE Transactions on Circuits and Systems for Video Technology, 2008, 18(7): 994-998

[19] Bozic S M. Digital and Kalman Filtering: An Introduction to Discrete-Time Filtering and Optimum Linear Estimation. 2nd Edition. London: Edward Arnold, 1994

[20] Weng S K, Kuo C M, Tu S K. Video object tracking using adaptive Kalman filter. Journal of Visual Communication & Image Representation, 2006, 17(6): 1190-1208

XIAO Guo-Qiang, born in 1965, Ph. D., professor. His major research interests include content based digital media processing, semantic video analysis, video segmentation, pattern recognition and machine learning.

KANG Qin, born in 1965, senior engineer. Her research interests focus on image processing.

JIANG Jian-Min, born in 1957, Ph. D., professor. His major research interests include image processing, video coding, video analysis, computer vision and machine learning.

ZHANG Bei-Bei, born in 1983, M. S.. Her research interests focus on video processing.

Background

Video object tracking discussed in this paper is a topic in the field of video analysis and processing. Content based video processing requires extraction of moving video objects, which play crucial roles for interpreting visual content, events, and relevant knowledge. Such video object extraction is primarily done via object segmentation and tracking to ensure that moving objects are extracted temporarily across a number of frames. As object segmentation remains an unsolved problem and no reliable solution exists to provide accurate object segmentation, algorithms robust to possible inaccurate object segmentation should be useful for all content based video processing, analysis, interpretation, and applications.

Existing efforts on video object tracking can be classified into four categories: model-based, appearance-based, contour/mesh-based, and feature-based. In model-based object tracking, a priori knowledge for the shape of objects in a given scene is required, which is often computationally expensive and lack of generality. Appearance-based methods rely on information about the entire region of the 2D shapes for the video object. As these are mainly low-level features, such methods often fail in dealing with complex deformations. Contour and mesh based methods relies on the contour of objects or 2D meshes to do the tracking, where motion is exploited to project the contour to the object segmented in the next frame. In such contour-based tracking methods, the computational complexity is often high, and large non-rigid movements cannot be handled by these methods. Feature-based methods use features of a video object to track parts of the object, but they are not specifically designed for video object tracking. The major problem in these methods lies in the fact that it is difficult to group features and determine which of them belong to the same object.

In this paper, we propose a technique for automatically tracking video objects in cluttered background via object segmentation, regional division, central point extraction, central macroblock construction, and directional vector validation. This algorithm uses central points and their directional vectors to solve the correspondence problem, which presents advantages in its robustness to inaccurate object segmentation. The simplicity comes from the fact that instead of projecting the entire region into the next frame, only a central macroblock of 16×16 pixels needs to be processed. Therefore, there is no need for computationally expensive motion models.

This research was supported by Chongqing Natural Science Foundation under contract number CSTC-2008BB2252.