

彩色序列图像中的人脸检测与跟踪^{*}

李久贤, 夏思宇, 袁晓辉, 夏良正
(东南大学 自动控制系, 南京 210096)

摘 要: 提出了一种彩色视频序列图像中的人脸检测与跟踪方法。该方法将人脸检测与人脸跟踪有效地结合在一起, 采用 Condensation 滤波跟踪算法对区域进行跟踪, 在跟踪过程中提出引入基于支持向量机的人脸置信度, 样本的置信度随时间进行更新, 人脸检测的结果基于置信度的后验概率。同时, 该方法对 Condensation 滤波跟踪算法作了改进, 在跟踪过程中采用了基于 Metropolis 算法的重采样方法以及自适应的动态模型, 实现了复杂背景下的对人脸自由运动的跟踪, 且精度较高。实验结果表明, 该方法有效地解决了复杂背景中人脸姿态变化情况下的人脸检测与跟踪问题, 与静态人脸检测相比有更好的检测效果。

关键词: 模式识别; 人脸检测; 人脸跟踪; Condensation 滤波; 支持向量机

中图分类号: TP391.4 文献标志码: A 文章编号: 1001-3695(2007)08-0189-04

Face detection and tracking in color image sequences

LI Jiu-xian, XIA Si-yu, YUAN Xiao-hui, XIA Liang-zheng
(Dept. of Automatic Control Engineering, Southeast University, Nanjing 210096, China)

Abstract: Human face detection and tracking received extensive attention because of the potential applications in many fields, such as video coding, surveillance and human-computer interface. The aim of this paper was to detect and to track human face in color image sequences effectively. An approach to simultaneous detection and tracking in video images was presented. The approach was based on posterior estimation using conditional density propagation. A face identity score was obtained for each image position and at several scales and posed based on support vector machine classifier. The identity scores were propagated over time using Condensation filter. Additionally, a simple and efficient sampling method based on the Metropolis algorithm and an adaptive dynamical model were used in Condensation algorithm for robust face tracking. Experimental results show that this approach could detect and track human face's free movement in videos containing complex backgrounds, and be more effective compared with still-to-still face detection.

Key words: pattern recognition; face detection; face tracking; Condensation filter; support vector machine(SVM)

0 引言

近年来, 人脸检测与跟踪技术得到了很大的发展。在电视电话会议、远程教学、监视与监控等场合, 都需要对特定人脸目标进行实时跟踪、分析和传递。可视电话、视频会议、基于内容的压缩与检索、身份鉴别、人机智能交互等许多应用都与好的人脸检测与跟踪算法紧密相关。

目前, 已有不少的人脸检测与跟踪的算法^[1,2], 但是其中大部分都是将人脸检测与跟踪分成两个步骤, 人脸跟踪是在人脸检测完成之后的基础上进行的, 而没有将这两个相互关联的部分有效地结合起来。视觉研究表明, 人眼对目标物体的辨识是一个过程。首先, 人眼会对运动的物体更加有兴趣, 随后, 在它进行跟踪的过程中会运用人脑中存储的各种信息对其加以辨识。比如, 当一个物体出现在很远的地方时, 人眼可能无法辨识, 但是当这个物体逐渐靠近, 人眼便可能会慢慢“认出”这个物体的类别。另外, 与静态图像相比, 视频序列图像包含丰富的时域信息, 但是目前的大多数算法仅仅将它用在跟踪算

法中, 对人脸的检测仍是采用基于静态图像的方法。

1 Condensation 滤波跟踪方法

人脸跟踪通常可以看成是在给定观测量下, 求解系统隐含状态量的估计问题。贝叶斯估计方法将未知(隐含)状态量的先验知识和描述观测量与状态量关系的似然函数结合起来, 利用贝叶斯公式得到未知量的后验概率。但是迭代贝叶斯估计只在特定的模型和假设下有解析解, 其特定模型包括线性高斯状态空间模型(卡尔曼滤波)和有限状态空间隐含马尔可夫模型(隐马尔可夫滤波)。然而, 在许多实际问题中, 状态空间模型含有非线性和非高斯成分, 因此没有闭合的最优解。在近三十年来, 人们提出了许多次优方案, 如扩展卡尔曼滤波(EKF)、确定性数值积分方法等。但当状态维数增加时, 近似误差的收敛率下降。

1996 年, Isard 与 Blake 将基于贝叶斯规则的概率模型引入到计算机视觉中, 用于跟踪非刚体的、多关节表示的手掌运动, 称之为 Condensation 滤波跟踪方法^[3]。与 EKF 不同的是,

收稿日期: 2006-05-07; 修返日期: 2006-07-29 基金项目: 浙江省长三角联合攻关项目(2005E60007)

作者简介: 李久贤(1961-), 男, 教授, 主要研究方向为图像处理与模式识别; 夏思宇(1978-), 男, 博士研究生, 主要研究方向为人脸检测、跟踪与识别(xia081@gmail.com); 袁晓辉(1962-), 男, 教授, 主要研究方向为图像处理与模式识别; 夏良正(1942-), 男, 教授, 博导, 主要研究方向为模式识别与图像处理、计算机视觉。

Condensation 方法不是近似模型以使其符合某一给定类型的后验概率,它是利用加权的采样值来逼近真实后验概率值。

通过建立一个状态空间模型,可以将视频序列的跟踪问题表示为一个动态系统的状态估计问题。在状态空间模型中,状态向量 X_t 用来描述系统的信息,观测向量 Z_t 对应于状态的观测值。状态方程与观测方程分别为

$$X_t = f(X_{t-1}, W_{t-1}) \tag{1}$$

$$Z_t = h(X_t, V_t) \tag{2}$$

其中: W, V 分别对应的过程噪声与观测噪声。

以贝叶斯学派的观点,跟踪问题就是在给定观测向量 Z_t 的条件下,估算状态向量 X_t 的值,即估计后验概率密度函数 $p(X_t/Z_t)$ 。假设在时刻 $t-1$ 概率密度函数 $p(X_{t-1}/Z_{t-1})$ 已知,则利用系统模型式(1)就可预测时刻 t 的先验概率密度:

$$p(X_t/Z_{t-1}) = \int p(X_t/X_{t-1}) p(X_{t-1}/Z_{t-1}) dX_{t-1} \tag{3}$$

此时,在式(3)中,利用了式(1)所描述的一阶马尔可夫过程

$$p(X_t/X_{t-1}, Z_{t-1}) = p(X_t/X_{t-1}) \tag{4}$$

利用贝叶斯规则,可得

$$p(X_t/Z_t) = [p(Z_t/X_t) p(X_t/Z_{t-1})] / [p(Z_t/Z_{t-1})] \tag{5}$$

式(5)是最优贝叶斯估计的一般概念表达式,通常不可能对它进行精确的分析。

Condensation 滤波跟踪方法的主要思想是利用一组带有相关权值的随机样本,以及基于这些样本的估算来表示后验概率密度 $p(X_t/Z_t)$ 。当样本数非常大时,这种概率估算将等同于后验概率密度。假设在时刻 t ,系统包含 N 个样本,每个样本 s_i 均由加权样本对表示 (s_i, π_i) 。其中: s_i 为一个样本的状态量; π_i 为样本的权值。首先,对这 N 个样本以其权值 π_i 为概率进行采样,以抑制权值较小的样本,增强权值较大的样本,从而得到 N 个新的样本 s'_i ;然后,对每个样本 s'_i 进行预测,从 $p(X_{t+1}|X_t = s'_i)$ 中生成 s_{t+1} (其中 $p(X_{t+1}|X_t)$ 表示系统状态的演化,由系统动态模型决定);最后,对权值进行更新, $\pi_{t+1} = p(Z_{t+1}|X_{t+1} = s_{t+1})$ 。其中 $p(Z_{t+1}|X_{t+1})$ 由系统观测模型决定。则在时刻 $t+1$,系统状态的期望可表示为

$$E[f(x)|Z_{t+1}] \approx \sum_{n=1}^N \pi_{t+1}^{(n)} f(s_{t+1}^{(n)}) \tag{6}$$

Condensation 方法适合跟踪那些非线性非高斯的运动,即并不要求概率密度函数为高斯分布,而这是卡尔曼滤波器所无法胜任的。另外,Condensation 方法的鲁棒性很强,能适应复杂的跟踪环境。在实际跟踪过程中,由于光照变化、遮挡或其他干扰物的出现,常常会出现一些不确定的观测数据。Condensation 的鲁棒性源于对这些不确定性的描述,在跟踪过程中并不是只保留权值最大的样本,那些权值较小的样本仍然有可能保留下来,在后续的跟踪中起作用。

2 人脸检测与跟踪

2.1 基于 SVM 的置信度

支持向量机是 Vapnik 等人提出的基于结构风险最小化 (structural risk minimization, SRM) 原理的统计学习理论,用于分类与回归问题^[4]。其基本思想是通过一个映射将输入向量映射到一个高维空间中,使原先空间中的非线性问题变为线性问题,然后在高维特征空间中求解最优分类超平面。得到的判别式如下:

$$f(x) = \sum_{i=1} \alpha_i y_i K(x, x_i) + b \tag{7}$$

其中: x_i 为支持向量; α_i 为相应权重; b 为阈值; $K(x, x_i)$ 为核函数;满足 Mercer 条件,对应某一变换空间中的内积。

在训练过程中,笔者采用了 Platt 提出的 SMO 算法,用人脸/非人脸样本对 SVM 进行训练。关于“人脸样本”的选取,用手工方法对 Yale 人脸库中图像的人脸区域进行了标定。Yale 库中共有 15 个人,每人有 11 张 160 × 160 大小的人脸图像样本,包含各种表情变化和光照变化。将每一个样本调整为 20 × 20 的 256 级灰度像素点阵,由于人脸呈一个椭圆形状,在点阵的四个角部分可以不予考虑,这样可以将输入点数量减少到 340 个。同时为了增强对有少许倾斜和尺寸少许变化的人脸检测的鲁棒性,将每个样本向左右分别旋转一定角度。共得到 1 000 余个人脸样本,这些样本经过预处理后用于训练。训练非线性 SVM 的非人脸样本采用基于自举方法收集,共 950 个样本。

本文采用多项式函数作为核函数:

$$k(x, z) = ((x \cdot z) + t)^p \tag{8}$$

其中,取 $p=2, t=1$ 。决策函数为

$$y = \sum_{i=1}^{N_s} \alpha_i y_i K(x_i, x) - b \tag{9}$$

其中: x_i 为支持向量; α_i 为相应权重; y_i 为 x_i 对应的标号; N_s 为支持向量个数。

为了引入置信度,没有直接使用决策函数式(9)作为 SVM 分类器,而是引入如下的映射使得输出规范到区间 [0, 1], 以此数值反映输入样本与人脸的相似程度:

$$\lambda(x) = 1/[1 + e^{-y(x)}] \tag{10}$$

其中: $y(x)$ 由式(9)求得。 λ 为 1 时表明输入样本与人脸最相似;相反, λ 为 0 时表明输入样本与人脸的相似程度最小。

2.2 改进的 Condensation 滤波跟踪算法

采用人脸的肤色特征对区域进行跟踪,因为肤色不依赖于面部的细节信息,肤色提取具有速度快、姿态不变性等特点。在跟踪过程中,笔者将置信度引入状态向量,对每个样本的置信度进行更新,并采用基于 Metropolis 算法的重采样方法^[5, 6]与自适应的动态模型。具体步骤如下:

a) 初始化。为了标记人脸在图像中的位置,取系统的状态量为 $X = [x, y, w, h, \lambda]$ 。其中: x, y 为人脸区域中心点的坐标, w, h 分别为人脸区域的宽与高; λ (由式(10)得到) 表示置信度。首先,在给定的人脸区域附近随机地选取 $N=100$ 个样本(这里,假定初始帧的肤色区域位置已得到);然后分别计算每个样本的置信度 λ_i 与观测值 $P_i = M/(w_i \times h_i)$ 。其中 M_i 为第 i 个样本中肤色像素点的数量;最后,归一化后得到各个样本的权值 $\pi_i = P_i / \sum_{k=1}^N P_k$ 。这里的状态量与权值用来预测下一帧中样本的概率分布。

b) 采样。采用一种基于 Metropolis 算法的采样方法,依据每个样本所对应的权值来进行采样,采样后的样本将用来做下一步的预测。Metropolis 是一种有效的重点抽样法,其算法为:系统从能量一个状态变化到另一个状态时,相应的能量从 E_1 变化到 E_2 , 概率为 $p = \exp[-(E_2 - E_1)/k]$ 。如果 $E_2 < E_1$, 系统接收此状态;否则,以一个随机概率接收此或丢弃此状态。这种经过定次数的迭代,系统会逐渐趋于一个稳定的分布状

态。将这种抽样方法应用到 Condensation 滤波跟踪过程中,能更加有效地保留高权值的样本及更好地反映样本分布。

c) 预测。采用一个自适应的动态模型来预测下一帧中样本的状态量。其定义如下:

$$\begin{cases} x_{t+1} = x_t + N(\mu_x, k_x \delta_x) \\ y_{t+1} = y_t + N(\mu_y, k_y \delta_y) \end{cases} \quad (11)$$

$$\begin{cases} w_{t+1} = w_t \times N(\mu_w, k_w \delta_w) \\ h_{t+1} = h_t \times N(\mu_h, k_h \delta_h) \end{cases} \quad (12)$$

其中:

$$\begin{cases} \mu_x = x_t - x_{t-1} \\ \mu_y = y_t - y_{t-1} \end{cases} \quad (13)$$

$$\begin{cases} \mu_w = w_t - w_{t-1} \\ \mu_h = h_t - h_{t-1} \end{cases} \quad (14)$$

其中: k 为方差系数(实验中, k_x 与 k_y 取为 2, k_w 与 k_h 取为 0.5); δ 取值为其相应均值的绝对值。这是考虑到目标移动的速度越快,其机动性可能越高。式(13)与(14)为实验中所采用的估计公式。样本下一帧的状态量 x_{t+1} 、 w_{t+1} 为当前值加上一个变化量,这个变化量变化的范围由目标位置移动与形状改变的快慢来决定,这样能解决由于人脸运动突然停止,或改变方向而导致跟踪失败的问题。方差系数的选取是通过实验经验得到,保证了模型既能适应缓慢的人脸运动,又能及时跟踪目标的快速移动。同时,考虑到目标的平移变化的速度要大于前后变化的速度,所以反映跟踪区域位置的动态模型的方差系数数值要比反映跟踪区域大小的动态模型的大。

假设在这个时间间隔中样本的置信度是不变的:

$$\lambda'_{t+1} = \lambda_t \quad (15)$$

d) 更新。对预测得到的每个样本 s_{t+1} 计算其观测量:

$$P_{t+1,i} = [(N_{t+1,i}) / (w_{t+1,i} \times h_{t+1,i})] \times \theta_{t+1,i}, \theta_{t+1,i} = (N_{t+1,i}) / S_t \quad (i = 1, \dots, N) \quad (16)$$

其中: $N_{t+1,i}$ 由肤色检测算法得到; S_t 为上一帧中跟踪得到的肤色区域面积; $\theta_{t+1,i}$ 表示第 i 个样本中肤色点所反映出的人脸信息量大小。然后计算每个新样本的置信度 λ''_{t+1} 。

更新每个样本的权值:

$$\pi_{t+1,i} = (P_{t+1,i}) / (\sum_{k=1}^N P_{t+1,i}) \quad (i = 1, \dots, N) \quad (17)$$

更新每个样本的置信度。这里,假定样本的置信度只随时间的变化而增大。这样,当人脸区域被确认后,即使随后人脸姿态发生改变或着被遮挡,算法仍能检测与跟踪到人脸。

$$\lambda_{t+1,i} = \begin{cases} \lambda''_{t+1,i} & \text{if } \lambda''_{t+1,i} \geq \lambda'_{t+1} \\ \lambda'_{t+1,i} & \text{if } \lambda''_{t+1,i} < \lambda'_{t+1} \end{cases} \quad (i = 1, \dots, N) \quad (18)$$

在时刻 $t+1$, 系统状态量均值与方差分别为

$$\text{Mean}(X) = \sum_{i=1}^N \pi_{i,t+1} X'_i \quad (19)$$
$$\text{Var}(X) = \sum_{i=1}^N \pi_{i,t+1} X_i^2 - \text{Mean}(X)^2 \quad (20)$$
$$\text{Mean}(\lambda) = \sum_{i=1}^N \lambda_{t+1,i} / N$$

其中: $X' = [x, y, w, h]'$ 。设定一个阈值 $\delta=0.5$, 当置信度的均值大于这个阈值时,即判断此区域为人脸区域,否则为非人脸区域。

2.3 人脸区域的消失与新的人脸区域的出现

在对人脸进行跟踪的过程中,每隔 M 帧便会对整幅图像进行肤色检测。如果检测到新的区域,且离现在所跟踪的区域较远,则标记为新的跟踪区域并进行跟踪。这种策略同时也能

解决两个人脸区域重合后又分开的情况,因为两个跟踪系统可能会在分开之后集中到同一个人脸区域,采用此方法后,那个漏掉的人脸区域会被重新检测到,并归为新的目标区域。

在两种情况下,认为人脸区域消失或不存在。其一是每个样本的观测值都很小,甚至为 0,且一直持续 M 帧;其二是置信度曲线一直在阈值之下。在这两种情况下就不再对区域进行跟踪。

3 实验结果

对各种典型视频序列图像,如 american miss、carphone、IIT-NRC facial video database^[7] 等进行了实验,另外也从各种影片中截取了视频序列图像进行测试。这里列举三组实验来说明,其一是来自 IIT-NRC facial video database 的序列图像;另外两组是来自影片中的部分片断,图像大小均调整为 160 ×120 像素。实验平台是 P4 2.4 GHz 的微机,处理速度在 10 fps 左右。

图 1 是对来自 IIT-NRC facial video database 中的视频序列进行跟踪的部分结果。可以看出在人脸大小不断变化的情况下,算法能很好地进行了跟踪。将跟踪得到的人脸区域的中心点位置与人工提取的相应帧的中心点位置进行了比较,结果如图 2(a) 所示,可以看出平均跟踪误差在三个像素点之内。(b) 为跟踪区域的置信度曲线,可以看出一直在阈值之上,表明跟踪区域为人脸区域。



图 1 IIT-NRC facial video database 部分跟踪结果

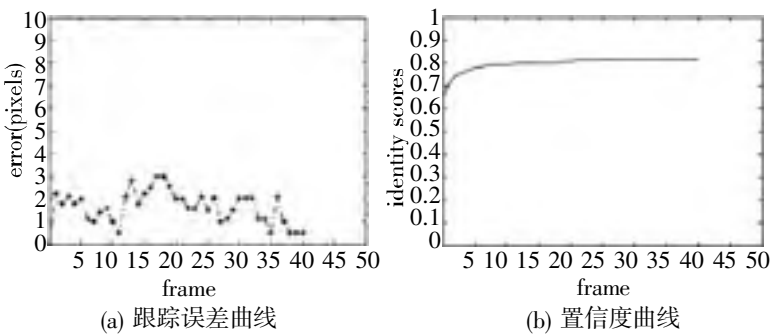


图 2 跟踪误差与置信度曲线

图 3 与 4 分别是对来自两段影片中的视频序列图像进行跟踪的部分结果。从图 3 中可以看出在复杂背景及人脸姿态发生变化的情况下,算法仍能进行成功跟踪。从置信度曲线(图 5(a))的变化可以看出,当人脸慢慢从侧面转向正面时,跟踪区域的置信度由阈值之下逐渐变化到阈值之上,表明已认出人脸区域。图 4 是算法同时对两个肤色区域进行跟踪的结果。从图 5(b)中可以看到,人脸区域的置信度曲线一直在阈值之上,而人手区域的置信度一直处于阈值之下。



图 3 来自影片 1 中的视频序列图像的部分跟踪结果

本文将算法与基于静态图像的人脸检测算法进行了比较。

针对一段视频文件中的 100 帧彩色图像(共含 100 个人脸区域), 分别用本文算法与采用肤色分割和 SVM 验证的人脸检测算法进行测试。



图 4 来自影片 2 中的视频序列图像的部分跟踪结果

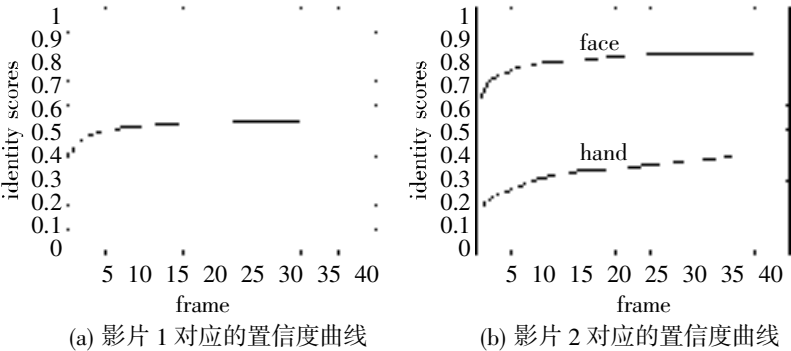


图 5 置信度曲线

表 1 实验结果统计

测试算法	正确检测	漏检	错误报警	正确率 / %
本文算法	100	0	0	100
肤色分割 + SVM	84	16	3	84

从表 1 中的实验结果可以看到, 相对于静态图像的人脸检测算法, 本文方法具有更高的检测正确率。从对结果的分析中, 可以发现, 由于充分利用了视频图像中的时域信息, 即使当人脸的姿态发生改变或者被遮挡, 本文方法仍能正确检测到人脸, 而基于静态图像的人脸检测算法只能检测到一定姿态下的人脸区域。

4 结束语

针对彩色视频序列图像, 本文提出一种基于 Condensation

(conditional density propagation) 滤波^[3] 的人脸检测与跟踪方法。在跟踪过程中, 用训练好的 SVM 对每个样本进行分类, 并赋予一个置信度。该置信度随着时间变化而更新。在任意时刻 k , 人脸检测的结果基于置信度的后验概率。并且, 对 Condensation 滤波跟踪方法进行改进, 有效的保证了复杂背景下多目标跟踪的准确性, 实现了人脸姿态变化与遮挡情况下的人脸检测与跟踪。

参考文献:

[1] ONG S, PSARROU A. Tracking and recognition of face sequences [C] //Proc of European Workshop on Combined Real and Synthetic Image Processing for Broadcast and Video Production. Hamburg: [s. n.], 1994: 97-112.

[2] YANG J, WAIBEL A. Tracking human faces in real-time, Technical Report CMU-CS-95-210 [R] . [S. l.] : Carnegie Mellon University, 1995.

[3] ISARD M, BLAKE A, Condensation-conditional density propagation for visual tracking [J] . International Journal of Computer Vision, 1998, 29(1) : 5-28.

[4] VAPNIK V. The nature of statistical learning theory[M] . New York: Springer-Veriag, 1995.

[5] METROPOLIS N, ROSENBLUTH A W, ROSENBLUTH M N, *et al.* Equation of state calculations by fast computing machines [J] . J of Chem Phys, 1953, 21(6) : 1087-1092.

[6] GUO Hong, HOW Wen-chi, YAN Feng, *et al.* A mote carlo sampling method for drawing representative samples from large databases [C] //Proc of the 16 th International Conference on Scientific and Statistical Database Management. Santorini: [s. n.], 2004: 1-2.

[7] GORODNICHY D O. IIT-NRC facial video database [EB/OL] . (2005-12-06) . <http://synapse.vit.iit.nrc.ca/db/video/faces/cv-glab>.

(上接第 188 页) 修复, 本方法适用于非纹理的图像修复。在具体操作过程中, 本文又对非纹理图像待修复区域作了进一步的划分, 对于较平坦区域和灰度跳跃比较大区域分别采用不同的修复方法。对于平坦的区域通过简单外推法效果较好, 但对于灰度跳跃比较大区域, 若用简单外推法, 则会导致图像边缘模糊, 甚至丢失边缘信息。图像修复的关键在于建立图像模型, 图像模型一旦确定, 就可以根据该模型计算待修复区域的像素值。通过具体的实验, 充分验证了本文方法效果显著, 即修复的视觉效果好, 运行效率高。决定实验效果有多个因素, 一个因素是极值差的选择问题, 根据实验效果一般选择在 30 ~50; 另一个因素是极值窗口尺寸的选择, 本文分别用窗口尺寸 3 和 5 分别进行试验, 窗口尺寸大的运行时间较长, 但准确度较高。从运行时间上看, 时间的长短与破损点的总数有关。以后将进一步探索影响运行效率的内在因素, 及其之间的关系, 以进一步提高运行时间。

参考文献:

[1] ERTALMIO M, SAPIRO G, CASELLES V, *et al.* Image inpainting [C] //The ACM SIGGRAPH Conference on Computer Graphics. New

Orleans: ACM Press, 2000: 417-424.

[2] BERTALMIO M, BERTOZZI A L, SAPIRO G. Navier-Stokes, Fluid-Dynamics, and Image and Video Inpainting[C] //Proc of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2001: 355-362.

[3] CHAN T, SHEN J. Mathematical models for local deterministic inpaintings [R] . UCLA CAM Report. 11, 2000.

[4] CHAN T, SHEN J. Non-texture inpainting by curvature-driven diffusions [R] . UCLA CAM Report. 35, 2000.

[5] SHIH T K, CHANG Rong-chi, LU Liang-chen, *et al.* Adaptive digital image inpainting[C] //Proc of the 18th International Conference on Advanced Information Networking and Applications. Washington, DC: IEEE Computer Society, 2004: 71-76.

[6] MUMFORD D, SHAH J. Optional approximations by piecewise smooth functions and associated variational problems [J] . Communications on Pure and Applied Mathematics, 1989, 42: 577-685.

[7] ADLATEINSSON D, SETHIAN J A. A fast level set method for propagation interfaces [J] . Jour Comp Phys, 1995, 118: 269-277.