

一种鲁棒高精度的人脸三维运动跟踪算法

於 俊^{1,2} 汪增福^{1,2,3}

¹(语音及语言信息处理国家工程实验室(中国科学技术大学) 合肥 230027)

²(中国科学技术大学自动化系 合肥 230027)

³(中国科学院智能机械研究所 合肥 230031)

(harryjun@ustc.edu.cn)

A Robust and High Accurate 3D Facial Motion Tracking Algorithm

Yu Jun^{1,2} and Wang Zengfu^{1,2,3}

¹(National Laboratory of Speech and Language Information Processing (University of Science & Technology of China), Hefei 230027)

²(Department of Automation, University of Science & Technology of China, Hefei 230027)

³(Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031)

Abstract A 3D facial motion tracking approach is proposed based on the incorporation of online appearance model (OAM) and cylinder head model (CHM) in framework of particle filter. It includes: 1) For the construction of OAM, multi-measurements are infused to reduce the influence of lighting and person dependence. 2) OAM can provide the detailed descriptive parameters used for facial motion tracking, however, it is not suitable for robust facial motion tracking across large pose variation. To alleviate these problems, OAM and CHM which is suitable for robust global head motion tracking across large pose variation, are combined, where the global head motion parameters obtained from the CHM are used as the cues of the OAM parameters for a good fitting. And the good fitting result is set as the initial of CHM in next frame. 3) Motion filtering is applied by particle filter combined with local optimization and improved resampling. Experiments of tracking real video sequences demonstrate that accurate tracking is obtained even in the presence of perturbing factors including significant head pose and facial expression variations, occlusions, and illumination changes. And it is also shown that facial motion tracking combining OAM and CHM is more pose robust than that of OAM in terms of 24% higher tracking rate and 11% wider pose coverage. The between subjective experiment indicates the suitability of subjective face identification on synthesized video with tracked facial motion parameters by it.

Key words facial motion tracking; online appearance model (OAM); cylinder head model (CHM); particle filter; information fusion

摘 要 提出了一种在粒子滤波框架下的结合在线外观模型(online appearance model, OAM)和柱状人头模型(cylinder head model, CHM)的人脸三维运动跟踪方案,具体包括:1)融合多种观测信息来降低 OAM 的光照敏感性和个体相关性;2)针对 OAM 适合跟踪局部运动但在大姿态下会跟踪失败的问题,将 OAM 与适合于大姿态下全局运动跟踪的 CHM 结合起来,在当前帧将 CHM 匹配得到的全局运

动参数作为 OAM 匹配的初始值,将 OAM 匹配得到的人脸运动参数作为下一帧 CHM 匹配的初始值; 3) 基于局部优化和改进重采样来改进粒子运动滤波策略. 实验表明:该系统在大姿态、表情剧烈变化、遮挡和强光照下能得到较好的跟踪效果,且 OAM+CHM 的跟踪正确率高于 OAM 的 24%,OAM+CHM 的姿态跟踪范围大于 OAM 的 11%. 主观实验表明:由跟踪得到的人脸运动参数合成的虚拟人脸具有较高的辨识度.

关键词 人脸运动跟踪;在线外观模型;柱状人头模型;粒子滤波;信息融合

中图法分类号 TP391.4

鉴于人脸运动(描述姿态/缩放/平移的全局运动和描述表情的局部运动)能够表达人类情感和说话等方面的信息,人脸运动跟踪在人机交互、视频编码以及人脸合成等领域有着广泛的应用.

单视频下的人脸三维运动跟踪可以通过基于特征的方法或基于外观模型的方法来完成.前者^[1-2]基于由光流或块匹配获得的二维图像运动信息来估计三维模型的运动.由于人脸没有足够稳定的特征以及当前帧的运动估计依赖于以前各帧的结果,从而随着估计误差的积累存在运动漂移问题.为了克服该问题,需要引入关键帧等措施^[3-5]或者融合多种特征信息来提高跟踪的鲁棒性^[6].后者通过建立和更新外观模型来匹配当前帧,它没有运动漂移问题,但对模型的健壮性和灵活性有较高要求.基于外观模型的方法有确定性和统计性之分.前者^[7-8]在视频开始捕捉一个参考纹理作为跟踪对象,然后根据估计得到的运动参数不断对参考纹理进行几何变形以更新跟踪对象.因为该模型是面向特定对象、确定性的,所以对于光照和表情变化下的人脸对象时效果有限.后者进一步有离线和在线之分.离线方法^[9-12]利用目标的统计及先验知识来构造模型,抗干扰能力强,但模型训练复杂且无法适应新情况.在线方法(online appearance model, OAM)^[13-14]通过在线不断学习目标的变化情况,能够学习到离线模型没有学习到的新情况,故较离线方法有更大优势,但它只能跟踪一定姿态内的人脸运动,并且如何更新模型以适应目标的变化以及应对光照和个体相关性的影响是难点.针对于此,一种途径^[15-16]是挖掘 OAM 的潜力;另一种途径^[17]是融合形状等多种信息.文献^[18]提出一个用于三维非刚体物体跟踪的 G-模型,将上述方法纳入到一个通用框架来解释.

特别地,全局人脸运动估计可以通过匹配简单几何人头模型,比如柱状人头模型(cylinder head model, CHM)^[8,19]、椭球^[20]和人头相似形状体^[5,21]等来完成.该类方法假设几何人头模型在跟踪中不

会发生局部形状改变,并通过最小化观测量与模型上纹理的差值来完成运动量的提取,或者融入边缘等信息来增加跟踪的鲁棒性^[22].因为全局运动只需 6 个参数即可描述,所以较于上述 OAM 的高维参数空间,该低维参数空间增加了跟踪的鲁棒性.文献^[23]比较了不同几何模型的跟踪性能,发现最简单形状的椭球跟踪性能最好,原因在于越复杂形状的模型越容易施加错误的先验信息,进而产生错误的跟踪结果.另外,该类方法不需要学习环节,从而具有个体无关性.进一步,由于该类方法采用整个头部的纹理来匹配,所以十分适合于大姿态下的人脸运动跟踪.当然,它们不能提供人脸局部运动的描述,因此如果将其与 OAM 结合起来将能得到更好的结果.

人脸运动是非线性非高斯分布的,故需要合适的方法来提取出真实状态,而滤波算法是其中的一个有效途径.在优化问题中,粒子滤波(particle filter, PF)^[16]因其具有的全局优化特性及对模型的弱依赖性,得到了广泛应用.然而,在人脸三维运动跟踪中采用 PF 面临如下问题:1) 计算盲目. PF 用观测似然度来更新粒子权重,没有考虑状态的最新值,由此产生的粒子与真实状态分布产生的粒子偏差较大.2) 计算量大. 粒子数如果太小,产生的粒子群会覆盖不了搜索空间,从而导致搜索结果易陷入局部极值.因此,为了达到较高的状态估计精度,需要大量的粒子来模拟状态的后验分布.3) 粒子退化. PF 采用以权值为基础的重采样方法来控制粒子退化.但实际中,因为选择的观测似然度难以反映真实情况,所以根据它得到的粒子权值和重采样难以有效控制粒子退化.

1 创新点

本文提出了一种在粒子滤波框架下结合 OAM 和 CHM 的人脸三维运动跟踪方案.

相对于现有的 OAM 方案^[15-17], 本方案具有以下特性: 1) 融合多种外观观测量来减少光照和个体相关性的影响; 2) 基于局部优化和改进重采样来改进粒子运动滤波策略; 3) 根据人脸三维模型和 OAM 来处理遮挡; 4) 加入眨眼幅度检测措施来获取细微的人脸运动。

进一步, 本方案将 OAM 和 CHM 紧密地结合在一起。一方面, 从 OAM 来看, 将 CHM 匹配得到的全局运动值作为 OAM 匹配的良好初始值, 当 OAM 匹配失败时, 被用来初始化人脸跟踪, 从而可以有效地扩大姿态跟踪范围; 另一方面, 从 CHM 来看, 从 OAM 匹配得到详细的局部人脸特征运动, 使得 CHM 匹配在获取全局运动时能够有效地克服人脸表情等局部运动的影响, 从而使得 CHM 匹配更加鲁棒。

2 系统框架

图 1 是系统框架。对于人脸视频的当前帧, 在粒子滤波框架下进行人脸运动跟踪, 具体包括: 1) 基于 OAM 的人脸运动跟踪; 2) 基于 CHM 的人脸运动跟踪; 3) OAM 和 CHM 互为补充, 紧密结合。

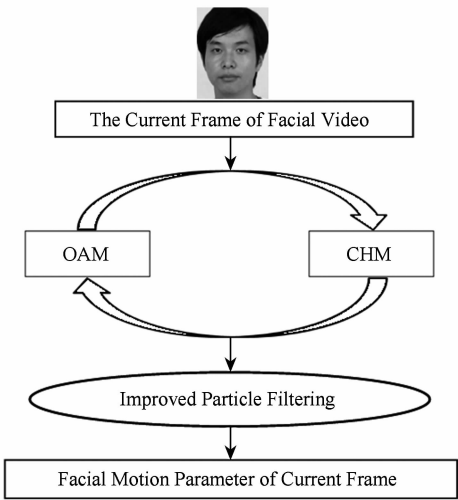


Fig. 1 System framework.

图 1 系统框架

3 基于 OAM 的人脸三维运动跟踪

因为 OAM 在不需要额外离线信息等方面的独特优势, 在此, 我们深入挖掘它的潜力。

3.1 人脸通用三维模型

基于实时性的考虑, 采用简单有效的 CANDIDE3^[24]

人脸三维模型(如图 2 所示)。对于该模型中所有顶点坐标的级联向量 \mathbf{g} , 若假定形状单元和表情单元相互独立, 可由式(1)表示^[24]:

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{S}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\alpha}, \tag{1}$$

其中, $\bar{\mathbf{g}}$ 表示模型的标准形状, \mathbf{S} 和 \mathbf{A} 的列分别表示 13 个形状单元和 7 个表情单元, $\boldsymbol{\beta}$ 和 $\boldsymbol{\alpha}$ 分别表示形状和表情运动参数, 即局部非刚体运动参数, $\mathbf{S}\boldsymbol{\beta}$ 和 $\mathbf{A}\boldsymbol{\alpha}$ 分别表示特定人的形状变化和表情变化。

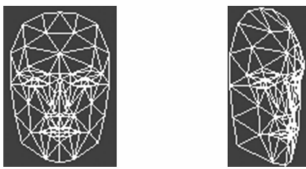


Fig. 2 CANDIDE3 3D facial model.

图 2 CANDIDE3 人脸三维模型

根据文献[9], 定义全局刚体运动参数为

$$\mathbf{h} = [\theta_x, \theta_y, \theta_z, t_x, t_y, t_z]^T. \tag{2}$$

综合以上两类参数定义人脸运动参数为

$$\mathbf{b} = [\mathbf{h}^T, \boldsymbol{\beta}^T, \boldsymbol{\alpha}^T]^T = [\theta_x, \theta_y, \theta_z, t_x, t_y, t_z, \boldsymbol{\sigma}^T, \boldsymbol{\alpha}^T]^T. \tag{3}$$

考虑到形状参数 $\boldsymbol{\beta}$ 在人脸模型特定化后就是固定的, 因此在后续的人脸运动跟踪中, \mathbf{b} 可简化为

$$\mathbf{b} = [\theta_x, \theta_y, \theta_z, t_x, t_y, t_z, \boldsymbol{\alpha}^T]^T = [\mathbf{h}^T, \boldsymbol{\alpha}^T]^T. \tag{4}$$

3.2 提取多种观测量

文献[6]融合多种特征信息来提高跟踪的鲁棒性。受此启发, 本文提取多种外观观测量并进行融合。

观测量从对应于输入人脸图像的几何归一化人脸图像(geometrical normalized image, GNI)^[9]中提取。GNI 由以下过程(如图 3 所示)获得: 1) 三维模型根据人脸运动参数投影到像平面上得到投影坐标; 2) 令投影坐标为纹理坐标, 将输入人脸图像纹理映射到三维模型上, 三维模型投影到像平面上得到投影图像; 3) 标准形状以垂直正面姿态投影到像平面上得到二维三角网格; 4) 将投影图像分段仿射变换到该二维三角网格的内部。另外, 为了避免头发的干扰, 把 GNI 的额头部分去掉; 为了减少光照的影响, 对 GNI 中的颜色值进行去均值、方差规范化等

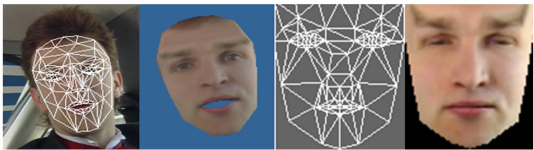


Fig. 3 Process of getting GNI.

图 3 GNI 的获取过程

操作;为了检测眨眼幅度,使 GNI 中的眼睛闭合.

本文从 GNI 中提取两部分外观观测量:颜色值和对光照比图像进行 Gabor 小波变换得到的系数.

第 2 部分观测量的获取依据和过程如下:光照比图像^[17]与表面反射性质无关,对于不同人脸(不同反射率)的面部运动,它刻画的是人脸的共性部分,从而消除了个体相关性,而且文献[17]证明了它的高频分量受光照影响较小;另外,Gabor 小波能够捕捉对应于空间尺度、位置及方向的局部结构信息^[25].鉴于此,本文将当前帧对应的 GNI(相对于首帧中人脸发生了运动)和首帧对应的 GNI 之间的光照比图像的高频(人脸运动)易发区域(如图 4 所示)上的 Gabor 小波变换系数作为第 2 部分的观测量.采用 8 个方向、5 个尺度的 Gabor 小波,由于系数中的相位对噪声敏感,只采用系数的幅值作为观测量.特别地,为了克服噪声的影响,选取每个区域中所有点的观测量的平均值作为该区域的观测量.

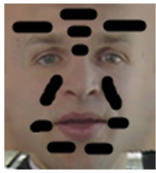


Fig. 4 Facial areas where motions with high frequencies are occur more likely.

图 4 人脸上高频运动易发区域

3.3 构建 OAMs 与融合多种观测量

3.3.1 构建第 1 个 OAM

设向量大小为 d 的 $\mathbf{y}(\mathbf{b}_t)$, 缩写为 \mathbf{y}_t , 为第 1 部分观测量的级联向量表示形式. 如文献[16]中的 s, w, l 分量, 我们将 \mathbf{y}_t 建模为 3 个分量的混合高斯模型(观测模型), 且令 $\{\boldsymbol{\mu}_{i,t}; i=s, w, l\}$ 为高斯均值向量, $\{\boldsymbol{\Sigma}_{i,t}; i=s, w, l\}$ 为方差矩阵, $\{\mathbf{m}_{i,t}; i=s, w, l\}$ 为权值, $\{\boldsymbol{\sigma}_{i,t}; i=s, w, l\}$ 为 $\{\boldsymbol{\Sigma}_{i,t}; i=s, w, l\}$ 对角元素的平方根所组成的向量. 观测似然度 $p(\mathbf{y}_t|\mathbf{b}_t) = \prod_{j=1}^d \left\{ \sum_{i=s,w,l} m_{i,t} N(\mathbf{y}_t(j); \boldsymbol{\mu}_{i,t}(j), \boldsymbol{\sigma}_{i,t}^2(j)) \right\}$ 为在 \mathbf{b}_t 下得到的 \mathbf{y}_t 与混合高斯模型的相似度.

定义在时刻 t 的 OAM: A_t 代表了到时刻 $t-1$ 为止所有观测量 $\mathbf{y}_{1:t-1}$ 的变化过程. 为了使它能够跟踪目标的变化, 当得到 $\mathbf{b}_t, \mathbf{y}_t$ 后, 定义 $\{\mathbf{m}_{i,t}; i=s, w, l\}$ 和稳定分量(s 分量) $\boldsymbol{\mu}_{s,t}, \boldsymbol{\sigma}_{s,t}$ 的更新方程, 其中 $\beta=0.2$ 是遗忘因子:

$$\begin{aligned} \mathbf{m}_{i,t} &= (1-\beta)\mathbf{m}_{i,t-1} + \beta\mathbf{m}_{i,t-1}N(\mathbf{y}_{t-1}; \boldsymbol{\mu}_{i,t-1}, \boldsymbol{\sigma}_{i,t-1}^2), \\ \boldsymbol{\mu}_{s,t} &= (1-\beta)\boldsymbol{\mu}_{s,t-1}/m_{s,t} + \beta\mathbf{y}_{t-1}m_{s,t-1}/m_{s,t}, \end{aligned}$$

$$\boldsymbol{\sigma}_{s,t}^2 = (1-\beta)\boldsymbol{\sigma}_{s,t-1}^2/m_{s,t} + \beta\mathbf{y}_{t-1}^2m_{s,t-1}/m_{s,t} - \boldsymbol{\mu}_{s,t-1}^2. \quad (5)$$

3.3.2 构建第 2 个 OAM

设向量大小为 n 的 $\mathbf{G}(\mathbf{b}_t)$, 缩写为 \mathbf{G}_t , 为第 2 部分观测量的级联向量表示形式. 与 \mathbf{y}_t 相似, 同样将它建模为 3 个分量的混合高斯模型, $p(\mathbf{G}_t|\mathbf{b}_t)$ 和 $\{\boldsymbol{\mu}'_{i,t}, \boldsymbol{\Sigma}'_{i,t}, \mathbf{m}'_{i,t}, \boldsymbol{\sigma}'_{i,t}; i=s, w, l\}$ 的定义与 3.3.1 节类似.

定义在时刻 t 的 OAM: A'_t 代表了到时刻 $t-1$ 为止的所有观测量 $\mathbf{G}_{1:t-1}$ 的变化过程. 同样定义 $\{\mathbf{m}'_{i,t}; i=s, w, l\}$ 和稳定分量(s 分量) $\boldsymbol{\mu}'_{s,t}, \boldsymbol{\sigma}'_{s,t}$ 的更新方程, 其中 $\beta'=0.2$ 是遗忘因子:

$$\begin{aligned} \mathbf{m}'_{i,t} &= (1-\beta')\mathbf{m}'_{i,t-1} + \beta'\mathbf{m}'_{i,t-1}N(\mathbf{G}_{t-1}; \boldsymbol{\mu}'_{i,t-1}, \boldsymbol{\sigma}'_{i,t-1}^2), \\ \boldsymbol{\mu}'_{s,t} &= (1-\beta')\boldsymbol{\mu}'_{s,t-1}/m'_{s,t} + \beta'\mathbf{G}_{t-1}m'_{s,t-1}/m'_{s,t}, \\ \boldsymbol{\sigma}'_{s,t}^2 &= (1-\beta')\boldsymbol{\sigma}'_{s,t-1}^2/m'_{s,t} + \beta'\mathbf{G}_{t-1}^2m'_{s,t-1}/m'_{s,t} - \boldsymbol{\mu}'_{s,t-1}^2. \end{aligned} \quad (6)$$

3.3.3 基于 OAM 融合多种观测量

采用乘法方式进行融合: $p(\mathbf{y}_t|\mathbf{b}_t) \times p(\mathbf{G}_t|\mathbf{b}_t)$, 该融合结果将在粒子滤波中被设为粒子的权重.

3.4 应对遮挡的措施

首先设置人脸三维模型中每个三维面片与其他三维面片不同的颜色, 然后根据人脸运动参数投影到图像平面上, 最后在投影后的三维面片的中心坐标处读取颜色值, 如果读取的值与设置的值相同则该三维面片没有被遮挡, 否则被遮挡.

对于每个三维面片, GNI 有对应的二维面片(如图 3 所示). 在这里, 被遮挡三维面片对应的二维面片是遮挡处理的对象. 对于在每一次迭代搜索 \mathbf{b}_t 后得到的 GNI 中第 k 个二维面片内的第 j 个像素位置, 令 $Tr_t^{(k,j)}$ 为 GNI 在此处的颜色值, $Tr\mu_{s,t-1}^{(k,j)}$, $Tr\sigma_{s,t-1}^{(k,j)}$ 为 $\boldsymbol{\mu}_{s,t-1}, \boldsymbol{\sigma}_{s,t-1}$ 在此处的值. 当第 k 个三维面片被遮挡时, $Tr_t^{(k,j)}$ 由前一时刻的值和 OAM 来估计:

$$\begin{cases} Tr_t^{(k,j)} = Tr_{t-1}^{(k,j)} + Tr\mu_{s,t-1}^{(k,j)} + Tr\sigma_{s,t-1}^{(k,j)}, \\ Tr_{t-1}^{(k,j)} \geq Tr\mu_{s,t-1}^{(k,j)}; \\ Tr_t^{(k,j)} = Tr_{t-1}^{(k,j)} + Tr\mu_{s,t-1}^{(k,j)} - Tr\sigma_{s,t-1}^{(k,j)}, \\ Tr_{t-1}^{(k,j)} < Tr\mu_{s,t-1}^{(k,j)}. \end{cases} \quad (7)$$

当 $Tr_t^{(k,j)}$ 为 Gabor 小波系数时与上述方程类似.

3.5 眨眼幅度检测

鉴于眨眼动作等细微的人脸运动对人脸表情起着重要的作用, 我们在 GNI 对应的二维三角网格中, 令上眼皮下端顶点的坐标和离它最近的下眼皮上端顶点的坐标等于这两个坐标的平均值, 从而使得 GNI 中的眼睛处于闭合状态, 那么学习到的 OAM 的均值也是闭眼状态. 如果当前帧中的眼睛是张开

的,那么根据从当前帧中提取的观测量与 OAM 之间在眼部的差异,匹配检测出眨眼幅度。

4 基于 CHM 的全局人脸三维运动跟踪

如前所述,简单几何人头模型十分适合于大姿态下的人脸运动跟踪。在现有几何模型中,考虑到与人脸的相似性和跟踪的复杂度等诸多因素,我们采用 CHM(如图 5 所示)来跟踪人脸运动。文献[8, 19]在这方面取得了较好的效果。文献[23]证明了几何模型的姿态跟踪范围与跟踪性能成正比,而 CHM 的姿态跟踪范围要比 OAM 的大,从而也验证了采用 CHM 的可行性。

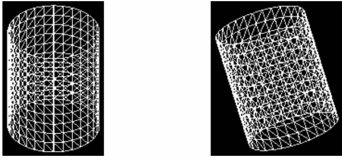


Fig. 5 Cylinder head model.

图 5 柱状人头模型

CHM 的运动是全局刚体运动,可用式(2)中的 \mathbf{h} 来表示。圆柱表面上的点 $\mathbf{p}_t = [x, y, z]$ 在 \mathbf{h} 作用下的新坐标可用式(8)来表示:

$$\mathbf{M}(\mathbf{p}_t, \mathbf{h}) = \mathbf{R} \cdot \mathbf{p}_t + \mathbf{T}, \quad (8)$$

这里 \mathbf{M} 是以 \mathbf{h} 表示的刚体变换矩阵, $\mathbf{R} \in \mathbb{R}^{3 \times 3}$ 和 $\mathbf{T} \in \mathbb{R}^{3 \times 1}$ 是与 \mathbf{h} 对应的旋转矩阵和平移向量。

对应于 $\mathbf{p}_t, \boldsymbol{\mu}_t = [\mu, \nu]$ 是它在图像平面上的投影点坐标。假设 $\Delta \mathbf{h}$ 是时刻 t 和时刻 $t+1$ 的全局刚体运动,则 $\boldsymbol{\mu}_t$ 在时刻 $t+1$ 坐标为

$$\boldsymbol{\mu}_{t+1} = F[\mathbf{p}_t, \Delta \mathbf{h}], \quad (9)$$

$F = \begin{bmatrix} x - y\theta_z + z\theta_y + t_x \\ x\theta_z + y - z\theta_x + t_y \end{bmatrix} \cdot \frac{f}{-x\theta_y + y\theta_x + z + t_z}$, f 是焦距。假设相邻两帧所处的光照环境是不变的,那么对于其中的像素 $I_t(\cdot)$ 有式(10)成立:

$$I_{t+1}(F[\mathbf{p}_t, \Delta \mathbf{h}]) = I_t(\boldsymbol{\mu}_t). \quad (10)$$

进而可通过最小化过程来解 $\Delta \mathbf{h}$:

$$\min E(\Delta \mathbf{h}) = \sum_{\boldsymbol{\mu} \in \Omega} \{I_{t+1}(F[\mathbf{p}_t, \Delta \mathbf{h}]) - I_t(\boldsymbol{\mu}_t)\}^2, \quad (11)$$

其中, Ω 是在 $I_t(\cdot)$ 和 $I_{t+1}(\cdot)$ 中均可见的区域。式(11)可通过 Lucas-Kanade 方法来解:

$$\Delta \mathbf{h} = - \left(\sum_{\Omega} (I_{\text{sg}} F_{\mathbf{h}})^T (I_{\text{sg}} F_{\mathbf{h}}) \right)^{-1} \sum_{\Omega} (I_{\text{ig}} (I_{\text{sg}} F_{\mathbf{h}})^T), \quad (12)$$

其中, I_{sg} 和 I_{ig} 为空域和时域梯度, $F_{\mathbf{h}} = \partial F / \partial \mathbf{h} |_{\Delta \mathbf{h} = 0}$ 。

5 改进的粒子滤波算法

5.1 结合局部优化来降低计算盲目性和计算量

局部优化可以结合当前时刻的最新观测值来生成与真实状态分布产生的粒子偏差较小的粒子,从而降低计算盲目性,而且粒子数目可以根据局部优化的结果进行调整来降低计算量。鉴于此,本文在更新粒子权重之前增加局部优化:

$$\mathbf{b}_t = \mathbf{b}_{t-1} + \Delta \mathbf{b}_t. \quad (13)$$

$\Delta \mathbf{b}_t$ 的求解通过对 \mathbf{y}_t 与 A_t 进行配准来进行:

$$\min_{\mathbf{b}_t} e(\mathbf{b}_t) = \min_{\mathbf{b}_t} D(\mathbf{y}_t, \boldsymbol{\mu}_{s,t}) = \sum_{i=1}^d ((y_{i(t)} - \mu_{s,i(t)}) / \sigma_{s,i(t)})^2. \quad (14)$$

从 \mathbf{b} 的初始值(前一帧的 \mathbf{b}_{t-1})开始计算 $\mathbf{r}(\mathbf{b}) = (\mathbf{y}_{t-1} - \boldsymbol{\mu}_t)$ 和 $e(\mathbf{b})$, 得到 \mathbf{b} 的更新值:

$$\Delta \mathbf{b} = -(GD_t^T GD_t)^{-1} GD_t^T \mathbf{r}, \quad (15)$$

$GD_t = \partial \mathbf{r} / \partial \mathbf{b}$ 为梯度矩阵。进而得到新的 \mathbf{b}, e :

$$\mathbf{b}' = \mathbf{b} + \lambda \Delta \mathbf{b}, \quad e' = e(\mathbf{b}'), \quad (16)$$

$\lambda = 0.5$ 为更新步长。如果 $e' < e$, 则 \mathbf{b} 根据式(16)更新直至本次匹配误差与上一次匹配误差的差值在小阈值之内即收敛,否则更换更小的步长直至收敛,如果超过一定匹配次数不收敛则终止更新。

这里的自适应性体现在所使用的梯度对于每一帧是不断变化的:在得到 \mathbf{b}_t 后,对当前帧摄动计算得到 GD_t , 然后它被送往下一帧来计算 \mathbf{b}_{t+1} 。

接着设置粒子的数量正比于上述局部优化后得到的误差 e , 从而自适应地降低了计算量。

5.2 改进的重采样

首先对粒子进行 PERM(pruned-enriched rose-nbluth method)采样^[26]:权值 π_t^i 小于 π^- 的粒子按照概率 a 被接受,若被接受,新权值为 π_t^i / a ; π_t^i 大于 π^+ 的粒子被 K 次重采样,新权值为 π_t^i / K ; 否则粒子保留不变。然后对粒子进行标准重采样。这样做的优点是,少数权值很大而有可能不是真实人脸运动的粒子也被复制多次,但权值自适应地减小,少数权值较小而有可能是真实人脸运动的粒子以一定的概率保留,且权值被自适应地增大,从而弥补了只单独使用标准重采样的不足。

另外,PERM 采样的参数进行在线调整:首先根据当前帧的人脸运动参数和人脸三维模型合成虚拟图像,然后计算它与当前帧像素值差的绝对值和。如果该值小则表示当前保留的粒子是好的,就弱化

PERM 采样:增大 a, π^+ , 减小 K, π^- ; 否则强化 PERM 采样:减小 a, π^+ , 增大 K, π^- .

6 OAM+CHM 的人脸三维运动跟踪

如引言所述, OAM 适合于一定姿态内的人脸全局和局部运动跟踪, 而 CHM 适合于大姿态的人脸全局运动跟踪. 鉴于此, 我们将它们各自的优势结合起来.

1) 以 CHM 匹配的结果初始化 OAM

如 4.2 节所述, 当 CHM 匹配得到全局运动参数后, 将它们作为后续 OAM 匹配的初始值.

2) 以 OAM 匹配的结果初始化 CHM

CHM 匹配的性能受初始值准确性的影响很大. 尽管其初始值可通过人脸检测的结果来设定, 但人脸检测的结果只是粗略地标记出图像中人脸的大致位置和大小. 鉴于此, 考虑到 OAM 不仅能够提供人脸全局位置和大小, 而且能够提供人脸局部特征的精确位置, 因此我们用 OAM 匹配的结果来初始化 CHM, 即将第 3 节所述的 OAM 匹配得到的全局运动参数作为后续 CHM 匹配的初始值, 并且将 OAM 在图像上的投影区域作为 CHM 的模板图像.

7 实验结果与分析

实验配置: AMD Athlon™ II X4 640 3.01 GHz; 内存 2 GB; NVIDIA GT200; GPU 开发包 CUDA1.3.

7.1 人脸运动跟踪

7.1.1 人脸运动跟踪结果

Carphone 是受光照影响和人脸姿态变化较大的视频. 从图 6 可以看出, 即使在大姿态、夸张表情、强光照、出现遮挡的情况下, 人脸运动也能被准确地跟踪到.

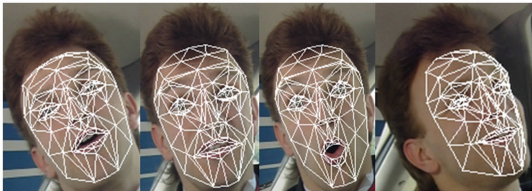


Fig. 6 Facial motion tracking results of Carphone.
图 6 Carphone 的人脸运动跟踪结果

7.1.2 眨眼幅度检测

为了验证检测眨眼幅度的能力, 我们采集了相

应的视频. 从图 7 可以看出, 眨眼幅度被准确地检测出来, 特别是图 7(a) 的闭眼状态:

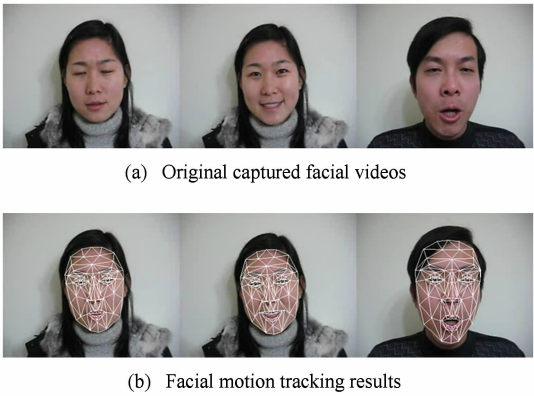


Fig. 7 Facial motion tracking results of captured videos.
图 7 采集视频中人脸运动的跟踪结果

7.1.3 单个观测量对比多个观测量

Foreman 是受光照影响较大的视频. 我们以此来验证本文提出的融合多个观测量措施的有效性. 由图 8 可见, 采用单个观测量的匹配结果会变坏, 而采取多个观测量的匹配结果会更好. 7.2 节将给出采用单个、多个观测量的量化比较.

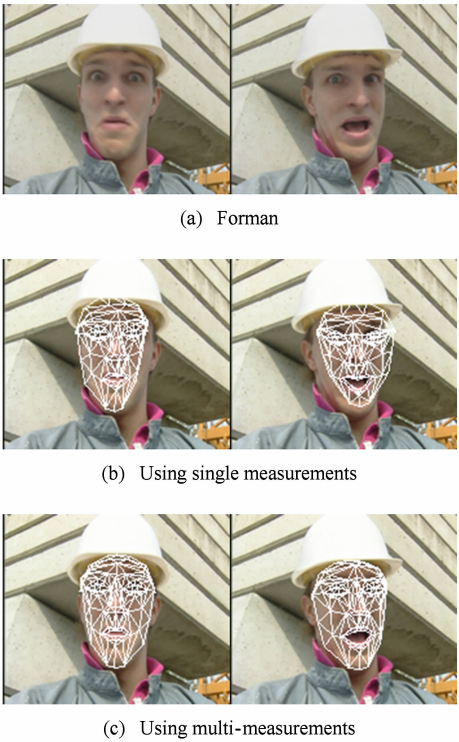


Fig. 8 Facial motion tracking results using and not using multi-measurements.
图 8 采用多个、单个观测量的人脸运动跟踪结果

7.1.4 OAM+CHM 对比 OAM

本文首先在文献[8]提供的数据集上验证 CHM

的大姿态人脸跟踪能力(如图 9 所示). 该数据集中的人脸姿态变化范围大,而且通过传感器提供了人脸姿态的真实值.

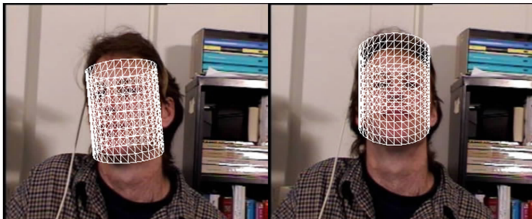


Fig. 9 Facial motion tracking results by CHM.
图 9 CHM 的跟踪结果

表 1 给出了在数据集上跟踪值与真实值的平均误差,由此可见,CHM 在大姿态下能得到较好的跟踪结果.

Table 1 Tracking Performances Using CHM
表 1 CHM 的跟踪性能

Average Angle Error of Pose: X	Average Angle Error of Pose: Y	Average Angle Error of Pose: Z
4.3	4.5	2.2

为了比较分别在 OAM+CHM 和 OAM 下人脸运动跟踪的性能,定义 2 个指标:

跟踪正确率是在测试视频中被跟踪正确的帧数与总帧数的比值,而被跟踪正确的判断依据是 7.2 节的式(20)定义的指标小于一定的阈值.

跟踪姿态范围定义为 $(\min(\theta_k), \max(\theta_k))$, $k \in \phi$, ϕ 是被跟踪正确的帧的集合, θ_k 是 ϕ 中第 k 帧对应的姿态.

由图 10 可见,在大姿态下只采用 OAM 跟踪会失败,而 OAM+CHM 的结果会更好.

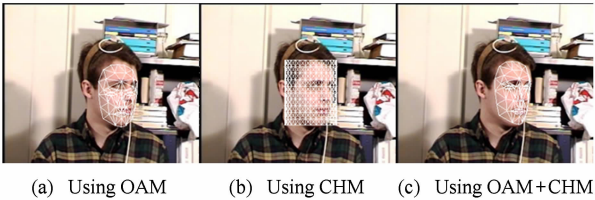


Fig. 10 Facial motion tracking results by OAM, CHM and OAM+CHM.

图 10 OAM,CHM,OAM+CHM 的跟踪结果

表 2 是在文献[8]提供的数据集上的比较结果,由此可见,OAM+CHM 的跟踪正确率高于 OAM 的 24%,OAM+CHM 的姿态跟踪范围高于 OAM 的 11%.

Table 2 Comparison of The Tracking Performances in the Test Video

表 2 在测试视频上的跟踪性能比较

Tracking Performance	OAM	OAM+CHM
Tracking Rate	0.67	0.83
Pose Coverage(X)	(-18.1,3.4)	(-19.3,4.9)
Pose Coverage(Y)	(-35.9,35.8)	(-40.5,41.7)
Pose Coverage(Z)	(-18.7,16.3)	(-18.7,16.3)

7.1.5 改进的粒子滤波

将本文的基于局部优化+PERM 采样+自适应粒子数的粒子滤波方法称为 LOPSAPF(localized optimization PERM sampling adaptive particle filter).

假设非线性系统的状态空间模型为

$$x_k = 1 + \sin[0.06\pi(k-1)] + 0.8x_{k-1} + v_{k-1}; \tag{18}$$

$$y_k = \begin{cases} 0.4x_k^2 + u_k, & k \leq 40; \\ 0.7x_k - 3 + u_k, & k > 40; \end{cases} \tag{19}$$

其中, v_k 服从 $\zeta_a(3,2)$, u_k 服从 $N(0,0.0001)$. 实验中其他算法的粒子数目为 200,LOPSAPF 算法的粒子数目是变化的,初始值为 200.

表 3 给出了经过 100 次独立实验,不同算法的平均耗时、均方误差的均值以及方差.可以看出:LOPSAPF 算法的估计精度优于其他算法.

Table 3 The Performance Comparison Using Several Non-linear Filter Algorithms

表 3 不同非线性滤波算法的比较

Algorithm	Mean Square Error		Average Time after 100 Iterations/s
	Mean	Variance	
EKF ^[27]	0.36846	0.0142716	0.054
UKF ^[28]	0.25912	0.0113	0.047
SIR ^[29]	0.18189	0.041127	0.085
EKPF ^[30]	0.28128	0.014819	0.076
UPF ^[31]	0.043593	0.0043239	0.064
LOPSAPF	0.016754	0.0004159	0.069

7.1.6 应对遮挡

从图 11 的遮挡处理过程看出,图 11(b)中右眉毛的右边比图 11(a)中右眉毛的右边更加合理,从而可从中提取出更加合理的观测量,进而得到更准确的跟踪结果.

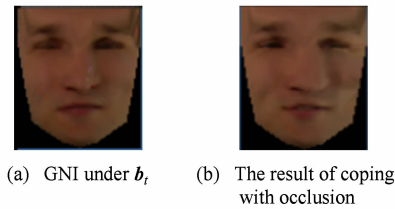


Fig. 11 Process of coping occlusion.
图 11 遮挡处理过程

7.2 跟踪质量的客观评测与比较

客观评测和比较按照如下 3 方面进行:

1) 定义一个评价人脸三维运动跟踪质量的量化指标-跟踪平均像素值差异(如图 12 所示):

$$Q_T = \sum_{i=1}^N \sum_{j=1}^{M_i} \text{abs}(y_{\text{syn}}^{(i,j)} - y_{\text{org}}^{(i,j)}) / (N \times M_i), \quad (20)$$

N 为测试视频的帧数; M_i 为测试视频第 i 帧图像中人脸区域的像素个数; $y_{\text{org}}^{(i,j)}$ 为该区域第 j 个像素的颜色值; $y_{\text{syn}}^{(i,j)}$ 为由跟踪结果将测试视频第 i 帧图像纹理映射到 CANDIDE3 模型上所合成虚拟图像中人脸区域第 j 个像素的颜色值; $\text{abs}()$ 为绝对值函数.

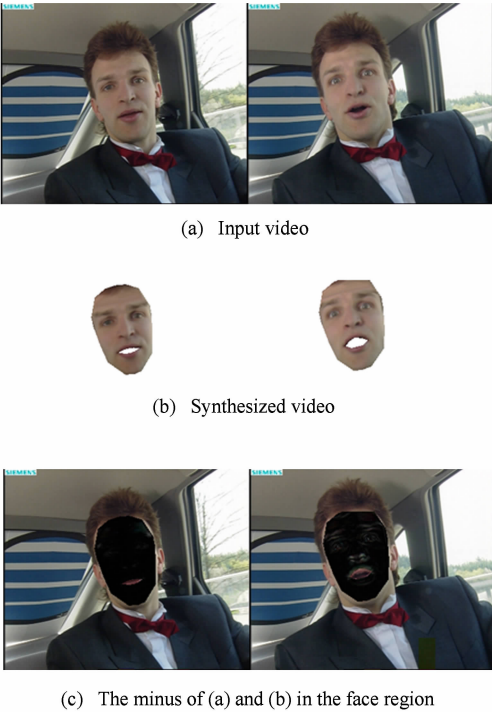


Fig. 12 The process of calculating tracking average pixel value difference.
图 12 跟踪平均像素值差异的计算过程

鉴于文献[15]是 OAM 中最重要的方法之一,我们以此为基准进行比较.对文献[8]提供的数

集、MPEG-4 中的 13 个人脸运动视频、CMU Cohn-Kanade DataBase^[32] 中的 110 个视频和我们捕捉的 78 个人脸视频进行了对比实验,从表 4 可以看出,本文算法在跟踪精度上优于文献[15]算法.

Table 4 The Average Time Each Frame Takes and The Mean of Q_T Using Our Algorithm and The Algorithm in Ref[15]

Algorithm	Average Time Each Frame Takes/s	Mean of Q_T
Our Algorithm	0.039	3.79
Algorithm in Ref[15]	0.036	5.4

2) 1)中的 Q_T 是作用到整体人脸上,但我们发现,小的 Q_T 并不意味着跟踪结果总是好的.因此需要寻找更好的评价指标.鉴于此,考虑到人们在直观上判断跟踪结果好坏的依据是三维模型在像平面上的投影与二维图像上的特征吻合得好坏与否.因此我们只利用 Q_T 在某些特征区域内的值来评价跟踪结果的好坏.这里选取的特征是嘴唇和眉毛,这是因为它们相比脸部其他区域的特征更为显著.

实验类似于 1),但 Q_T 中的 $y_{\text{org}}^{(i,j)}$ 和 $y_{\text{syn}}^{(i,j)}$ 是在嘴唇和眉毛区域的颜色值.从表 5 可以看出,本文算法在跟踪精度上优于文献[15]算法.

Table 5 The Mean of Q_T Using Our Algorithm and The Algorithm in Ref[15]

Algorithm	Mean of Q_T
Our Algorithm	4.54
Algorithm in Ref[15]	6.1

3) 考虑到虽然文献[8]的数据集通过传感器提供了所谓的真实值,但传感器也是有误差的,因此我们用计算机图形学的技术来间接获得真实值,即在给定人脸运动(真实值)和光照下绘制人脸图像,然后对绘制的人脸图像进行人脸运动跟踪,最后将估计出来的运动值与真实值进行比较.这里量化指标改为估计出来的运动值与真实值之间误差的平均值.进一步也对采用单个或多个观测量进行量化比较.

前续工作^[33-34] 中合成的在不同姿态和光照下的人脸图像如图 13 所示,实验结果如表 6 所示.从表 6 中可以看出,本文算法在跟踪精度上优于文献[15]算法,且采用多个观测量要优于采用单个观测量.

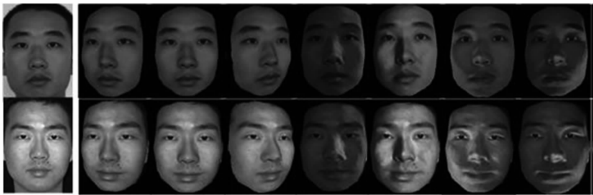


Fig. 13 Face images under various poses and lighting.
图 13 在不同姿态和光照下绘制的人脸图像

Table 6 Evaluation of Tracking Accuracy
表 6 跟踪精度评价

Angle	Roll	Pitch	Row
Our Algorithm with Multi-Measurements	2.38	1.75	1.69
Our Algorithm with Single Measurement	2.73	1.94	1.86
Algorithm in Ref[15]	2.71	1.98	1.85

7.3 主观评测与比较

我们构建了如图 14 所示的人脸运动跟踪系统. 该系统的左上窗口是输入视频, 右上窗口是根据跟踪得到的人脸运动参数所合成的虚拟视频. 下面利用人群对该系统进行主观评测.

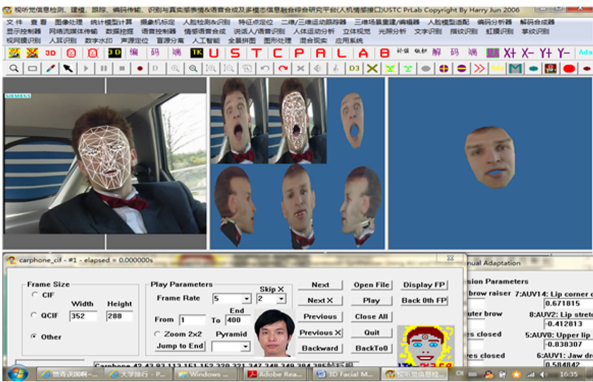


Fig. 14 Facial motion tracking system.
图 14 人脸运动跟踪系统

该主观评测的调查对象是 34 个人, 他们的年龄、性别以及出生地的情况如表 7 所示. 从表 7 可以看出, 调查对象的分布具有广泛的代表性.

Table 9 Mean Score of Developed System and Ref[15]
表 9 该系统和文献[15]的平均得分

Construct	Question	Mean Score of Our System	Mean Score of Ref[15]
Expressiveness	① I recognize avatar's expressions displayed in the synthesized video.	7.87	6.73
	② Avatar's movements look natural in the synthesized video.	8.06	7.01
	③ I like avatar's movements in the synthesized video.	8.23	7.13
Appearance	① I like avatar's appearance in the synthesized video.	7.83	6.48
	② The avatar looks like a human being in the synthesized video.	8.12	7.47

Table 7 Distributions of Participants
表 7 调查对象的分布情况

Construct	Possible Answers	Frequency
Age	① Younger than 20 ② Between 20 and 30 ③ Older than 30	8/18/8
Gender	① Male ② Female	20/14
Birth Place	① East Region ② Middle Region ③ West Region	7/14/13

该主观评测的目的是由跟踪得到的人脸运动参数合成的虚拟人脸是否具有较高的辨识度. 该主观评测分为两步:

1) 建立问卷. 在表 8 中展示了问卷中的问题. 对于其中问题的回答从“绝对不同意”到“完全同意”分为 10 个级别. 为了检验该问卷的内在可信度, 采用 Cronbach's alpha 测试来验证, 研究表明, 该测试结果在 0.7 以上表示该问卷是有效的.

Table 8 Cronbach's alpha Results of Questionnaire
表 8 问卷的 Cronbach's alpha 测试结果

Construct	Question	Cronbach's alpha
Expressiveness	① I recognize avatar's expressions displayed in the synthesized video.	0.743
	② Avatar's movements look natural in the synthesized video.	
	③ I like avatar's movements in the synthesized video.	
Appearance	① I like avatar's appearance in the synthesized video.	0.811
	② The avatar looks like a human being in the synthesized video.	

由表 8 可见, Cronbach's alpha 的测试结果均在 0.7 以上, 从而验证了该问卷的有效性.

2) 比较由跟踪得到的人脸运动参数合成的虚拟人脸的辨识度. 首先该系统和文献[15]分别跟踪数个人脸视频, 然后让调查者观察输入视频中人脸和虚拟人脸的差异, 最后让他们填写问卷, 设定评测的最高分是 10 分, 最低分是 0 分, 以平均得分作为比较的依据, 如表 9 所示:

由表 9 可见,本系统的得分在 7.5 以上,表明它由跟踪得到的人脸运动参数合成的虚拟人脸有较高的辨识度;文献[15]的得分在 7.5 以下,且比本系统低,表明它的结果辨识度比本系统低。

8 结论与展望

本文提出了一种在粒子滤波框架下的结合 OAM 和 CHM 的人脸三维运动跟踪方案. 自适应的 OAM 能够在线快速地学习到人脸的运动情况. 多种观测量的融合有效地减少了光照和个体相关性的影响. 结合局部优化和 PERM 采样来改进粒子滤波,较好地解决了计算盲目、计算量大和粒子退化等问题. 结合三维模型和 OAM 有效地处理了遮挡. 增加了眨眼幅度检测操作,从而能够有效地获取人脸的细微运动. 特别地,为了解决 OAM 只适合于一定姿态内人脸运动跟踪的限制,将 OAM 和 CHM 紧密地结合在一起,从而有效地扩大了姿态跟踪范围和提高了跟踪正确率.

客观实验表明,该系统在大姿态、表情剧烈变化、遮挡和强光照下能得到较好的跟踪效果,且 OAM+CHM 的跟踪正确率高于 OAM 的 24%, OAM+CHM 的姿态跟踪范围高于 OAM 的 11%. 主观实验表明,由跟踪得到的人脸运动参数合成的虚拟人脸具有较高的辨识度.

针对本文的实验结果提出进一步研究方向:利用跟踪到的人脸运动参数进行表情识别.

参 考 文 献

- [1] Black M J, et al. Recognizing facial expressions in image sequences using local parameterized models of image motion [J]. International Journal of Computer Vision, 1997, 25 (1): 23-28
- [2] Gokturk S, et al. A data-driven model for monocular face tracking [C] //Proc of the 7th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2001: 701-708
- [3] Zhang Wei, Wang Qiang, et al. Real time feature based 3D deformable face tracking [C] //Proc of the 10th European Conf on Computer Vision. Piscataway, NJ: IEEE, 2008: 720-732
- [4] Wang Q, Zhang W, Tang X, et al. Real-time bayesian 3D pose tracking [J]. IEEE Trans on Circuits and Systems for Video Technology, 2006, 16(12): 1533-1541
- [5] Vacchetti L, Lepetit V, Fua P. Stable real-time 3D tracking using online and offline information [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2004, 26(10): 1385-1391
- [6] Wei-Kai L, Douglas F, G'erard M. Integrating multiple visual cues for robust real-time 3d face tracking [C] //Proc of the 3rd Int Workshop on Analysis and Modeling of Faces and Gestures. Piscataway, NJ: IEEE, 2007: 109-123
- [7] Strom J. Model based head tracking and coding [D]. Linköping, Sweden: Linköping University, 2002
- [8] Cascia M L, et al. Fast, reliable head tracking under varying illumination: An approach based on registration of texture mapped 3D models [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(4): 322-336
- [9] Ahlberg J. Model-based coding: Extraction, coding, and evaluation of face model parameters [D]. Linköping, Sweden: Linköping University, 2002
- [10] Matthews I, Xiao J, Baker S. 2D vs. 3D deformable face models: Representational power, construction, and real-time fitting [J]. International Journal of Computer Vision, 2007, 75(1): 93-113
- [11] Wei-Kai L, Gerard M. 3D face tracking and expression inference from a 2D sequence using manifold learning [C] //Proc of the 24th Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2008: 3597-3604
- [12] Sung J, Kanade T, Kim D. Pose robust face tracking by combining active appearance models and cylinder head models [J]. International Journal of Computer Vision, 2008, 80 (2): 260-274
- [13] Jepson A D, Fleet D J, El-Maraghi T F. Robust online appearance models for visual tracking [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2003, 25(10): 1296-1311
- [14] Lui Y M, Beveridge J R, Whitley L D. Adaptive appearance model and condensation algorithm for robust face tracking [J]. IEEE Trans on Systems Man and Cybernetics, Part A—Systems and Humans, 2010, 40(3): 437-448
- [15] Dornaika F, Davoine F. Simultaneous facial action tracking and expression recognition in the presence of head motion [J]. Journal of Computer Vision, 2008, 76(3): 257-281
- [16] Zhou S, Chellappa R, Moghaddam B. Visual tracking and recognition using appearance-adaptive models in particle filters [J]. IEEE Trans on Image Process, 2004, 13(11): 1491-1506
- [17] Wen Z, Huang T. Capturing subtle facial motions in 3D face tracking [C] //Proc of the 8th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2003: 1343-1350
- [18] Marks T K, Hershey J R, Movellan J R. Tracking motion, deformation, and texture using conditionally gaussian processes [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(2): 348-363
- [19] Xiao J, Moriyama T, Kanade T. Robust full-motion recovery of head by dynamic templates and registration techniques [J]. International Journal of Imaging Systems and Technology, 2003, 13(2): 85-94

- [20] Basu S, Essa I, Pentland A. Motion regularization for model based head tracking [C] //Proc of the 4th Int Conf on Pattern Recognition. Piscataway, NJ: IEEE, 1996: 611-616
- [21] Malciu M, Preteux F. A robust model-based approach for 3D head tracking in video sequences [C] //Proc of the 4th Int Conf on Automatic Face and Gesture Recognition. Piscataway, NJ: IEEE, 2000: 169-174
- [22] DeCarlo D, Metaxas D. The integration of optical flow and deformable models with applications to human face shape and motion estimation [C] //Proc of the 12th Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 1996: 231-238
- [23] Fidaleo D, Medioni G, Fua P, et al. An investigation of model bias in 3D face tracking [C] //Proc of the 2nd Int Workshop Analysis and Modeling Faces and Gestures. Piscataway, NJ: IEEE, 2005: 125-129
- [24] Ahlberg J. CANDIDE3—An updated parameterized face, LiTH-ISY-R-2326 [R]. Linköping, Sweden: Department of Electrical Engineering, Linköping University, 2001: 1-16
- [25] Zhang Gang, Ma Zongmin. An approach of using Gabor wavelets for texture feature extraction [J]. Journal of Image and Graphic, 2010, 15(2): 247-254 (in Chinese)
(张刚, 马宗民. 一种采用 Gabor 小波的纹理特征提取方法 [J]. 中国图象图形学报, 2010, 15(2): 247-254)
- [26] Grassberger P. The pruned-enriched rosenbluth method: simulations of theta polymers of chain length up to 1 000 000 [J]. Physical Review E, 1997, 56(3): 3682-3693
- [27] Shi K L, Chan T F, et al. Speed estimation of an induction motor drive using an optimized extended Kalman filter [J]. IEEE Trans on Industrial Electronics, 2002, 49(1): 124-133
- [28] Julier S, Uhlmann J, et al. A new method for the nonlinear transformation of means and covariances in filters and estimators [J]. IEEE Trans on Automatic Control, 2000, 45(3): 477-482
- [29] Gordon N J, Salmond D J. Novel approach to nonlinear/non-Gaussian bayesian state estimation [J]. IEE Proceedings F, 1993, 140(2): 107-113
- [30] Merwe R V, Wan E. The square-root unscented Kalman filter for state and parameter estimation [C] //Proc of the 27th Int Conf on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE, 2001: 234-240
- [31] Vander M R, Doucet A, et al. The unscented particle filter [EB/OL]. Cambridge: Cambridge University, 2000 [2013-07-23]. <http://citeseer.ist.psu.edu/325754.html>
- [32] Kanade T, Cohn J, et al. Comprehensive database for facial expression analysis [C] //Proc of the 4th Int Conf on Automatic Face and Gesture Recognition. Piscataway, NJ: IEEE, 2000: 46-53
- [33] Hu Yuankui, Zheng Ying, Wang Zengfu. Reconstruction of 3D face from a single 2D image for face recognition [C] //Proc of the 2nd Joint IEEE IWVSPETS 2005. Piscataway, NJ: IEEE, 2005: 217-222
- [34] Hu Yuankui, Wang Zengfu. A low-dimensional illumination space representation of human faces for arbitrary lighting conditions [C] //Proc of the 14th Int Conf on Pattern Recognition. Piscataway, NJ: IEEE, 2006: 1147-1150



Yu Jun, born in 1985. PhD and associate professor. His main research interests include human machine interaction, intelligent robot.



Wang Zengfu, born in 1960. PhD and professor. His main research interests include human machine interaction, intelligent robot (zfwang@ustc.edu.cn).