

# 视觉语音参数的自动估计

王志明<sup>1</sup> 蔡莲红<sup>2</sup> 艾海舟<sup>2</sup>

<sup>1</sup>(北京科技大学计算机科学与技术系 北京 100083)

<sup>2</sup>(清华大学计算机科学与技术系 北京 100084)

(wangzhiming@tsinghua.org.cn)

## Automatic Estimation of Visual Speech Parameters

Wang Zhiming<sup>1</sup>, Cai Lianhong<sup>2</sup>, and Ai Haizhou<sup>2</sup>

<sup>1</sup>(Department of Computer Science and Technology, University of Science and Technology Beijing, Beijing 100083)

<sup>2</sup>(Department of Computer Science and Technology, Tsinghua University, Beijing 100084)

**Abstract** Visual speech parameter estimation has an important role in the study of visual speech. In this paper, 24 speech correlating parameters are selected from MPEG-4 defined facial animation parameter (FAP) to describe visual speech. Combining the statistic learning method and rule based method, precise tracking results are obtained for mouth contour and facial feature points based on facial color probability distribution and priori knowledge on shape and edge. High frequency noise in reference points tracking is eliminated by low-pass filter, and main face pose is estimated from the four most evident reference points to remove the overall movements of the face. Finally, precise visual speech parameters are computed from the movement of these facial feature points, and these parameters have already been used in some related applications.

**Key words** visual speech; facial animation parameter (FAP); Gaussian mixture model (GMM); deformable template

**摘要** 视觉语音参数估计在视觉语音的研究中占有重要的地位。从 MPEG-4 定义的人脸动画参数 FAP 中选择 24 个与发音有直接关系的参数来描述视觉语音,将统计学习方法和基于规则的方法结合起来,利用人脸颜色概率分布信息和先验形状及边缘知识跟踪嘴唇轮廓线和人脸特征点,取得了较为精确的跟踪效果。在滤除参考点跟踪中的高频噪声后,利用人脸上最为突出的 4 个参考点估计出主要的人脸运动姿态,从而消除了全局运动的影响,最后根据这些人脸特征点的运动计算出准确的视觉语音参数,并得到了实际应用。

**关键词** 视觉语音;人脸动画参数(FAP);混合高斯模型(GMM);变形模板

中图法分类号 TP391

## 1 引言

视觉语音是指与语音相伴的可视发音器官的运

动,由于视觉语音和听觉语音都是由人的发音器官的运动产生的,它们之间存在着内在的必然联系。许多科学实验证明,视觉语音可以帮助人们理解听觉语音,提高在噪声环境下的语音可懂度,帮助有听

力障碍的儿童学习发音. 视觉语音合成在计算机人机交互、动画制作等方面也有着广泛的应用前景.

为了达到好的合成效果,首先要研究真实发音过程中视觉语音参数的变化规律,这就需要获取大量的视觉语音参数. 为了获取这些参数,目前主要有两种方法,一种是仅获取二维数据<sup>[1,2]</sup>,另一种是采用基于三维人脸模型的跟踪方法获取三维数据<sup>[3,4]</sup>. 前一种方法不能得到三维的信息,而后一种方法需要事先针对特定人建立三维人脸模型. 而目前利用多个摄像机实现三维重建的技术存在着摄像机之间难以精确同步和三维重建精度难以保证的问题,三维扫描设备的速度较慢,无法获取实时的三维数据. 因此,我们采用在人脸侧面安装一个与人脸平面成 45°角的镜子,同步地获取两个视角的人脸图像,如图 1 所示:



Fig. 1 Acquire two views face image simultaneously.

图 1 人脸双视图角同步获取

在所有人脸器官的跟踪中,嘴唇的跟踪是最困难的. 因为嘴唇是一个弹性体,在说话过程中嘴唇轮廓变化较大,加上牙齿、舌头等器官有时可见、有时不可见,造成唇区的颜色变化也较多. 本文将统计学习方法和基于规则的方法结合起来,先利用唇区颜色统计知识建立唇区和肤色区颜色概率分布模型,再利用包括先验知识的变形模板进行最佳匹配

搜索以获得准确的外唇轮廓参数. 另外,利用前景背景颜色差异、亮度差异及局部颜色特征等信息跟踪人脸其他特征点,并通过滤波和姿态校正获得各个特征点的相对运动信息,最终根据这些相对运动信息计算出视觉语音参数. 实验结果表明,我们从视频流中估计得到的参数准确可靠,可以用来进行视觉语音分析或合成.

本文的后续内容安排如下:第 2 节介绍与语音有关的视觉语音参数的选取;第 3 节给出嘴唇轮廓及其他人脸特征点的跟踪方法;第 4 节介绍如何根据这些特征点计算 FAP 参数;第 5 节是实验结果及结束语.

## 2 视觉语音参数的选择

以往语音学对发音口形的描述只是定性的,如圆唇、扁唇、开口的大小,舌位的高低等等. 而现在许多应用领域需要对视觉语音进行客观上的定量度量,如虚拟人脸合成、机器自动唇读等等. 在描述参数上,许多人使用过各种各样的自定义参数<sup>[1,2]</sup>. 但这些参数缺乏通用性,且表达能力有限. 我们在仔细研究了发音过程中可视发音器官的运动后,从 MPEG-4<sup>[5]</sup>定义的人脸动画参数 (FAP: facial animation parameter) 中选择 24 个与发音有直接关系的参数来描述视觉语音,如表 1 所示(由于条件限制,没有考虑舌头运动和表情因素). FAP 参数的优点是它已成为人脸动画的国际标准,且它通过定义人脸动画参数单位 FAPUC (facial animation parameter unit) 规范了不同人脸差异,使得同样的参数可以在不同的人脸模型上做出相似的人脸表情.

Table 1 FAP Parameters Related with Speech

表 1 与发音有关的 FAP 参数

FAP #	Name	FAP #	Name	FAP #	Name
3	<i>open_jaw</i>	11	<i>raise_b_lip_rm</i>	53	<i>stretch_l_cornerlip_o</i>
4	<i>lower_l_lip</i>	12	<i>raise_l_cornerlip</i>	54	<i>stretch_r_cornerlip_o</i>
5	<i>raise_b_midlip</i>	13	<i>raise_r_cornerlip</i>	55	<i>lower_l_lip_lm_o</i>
6	<i>stretch_l_cornerlip</i>	14	<i>thrust_jaw</i>	56	<i>lower_l_lip_rm_o</i>
7	<i>stretch_r_cornerlip</i>	16	<i>push_b_lip</i>	57	<i>raise_b_lip_lm_o</i>
8	<i>lower_l_lip_lm</i>	17	<i>push_l_lip</i>	58	<i>raise_b_lip_rm_o</i>
9	<i>lower_l_lip_rm</i>	51	<i>lower_l_lip_o</i>	59	<i>raise_l_cornerlip_o</i>
10	<i>raise_b_lip_lm</i>	52	<i>raise_b_midlip_o</i>	60	<i>raise_r_cornerlip_o</i>

在表 1 所列出的这些参数中 ,FAP3 和 FAP14 定义了下唇的上下和前后移动量 ,FAP16 和 FAP17 分别定义了下唇和上唇的突出度 ,它们可由图 1 中侧面人脸中的上下唇、下唇等特征点位置计算得出 ; FAP4 ,FAP5 和 FAP8~13 定义了内唇边缘 8 个点的上下移动量 ;FAP6 ,FAP7 和 FAP53 ,FAP54 分别定义了内外唇角的水平方向的位移 ;FAP51 ,FAP52 和 FAP55~60 定义了外唇边缘 8 个点的上下移动量 ,它们可由图 1 中正面人脸中内外唇轮廓线上的特征点位置计算得出 . 而本文正是要在不需要任何额外标记的情况下跟踪这些人脸特征点的运动 .

### 3 人脸特征点跟踪

#### 3.1 嘴唇轮廓跟踪

为了将嘴唇同周围的皮肤分开 ,Wang 等人<sup>[6]</sup>采用了 Fisher 变换做最佳二分类的方法 ,Liew 等人<sup>[7]</sup>采用了模糊聚类的方法 ,Chen<sup>[1]</sup>利用混合高斯模型 GMM(Gauss mixture model)计算各自概率的方法 . 由于唇区的颜色变化较为丰富 ,包括嘴唇、口内黑区、牙齿、舌头等 ,难以简单地将它们与肤色分开 ,我们采用 GMM 描述唇区及其周围肤色区的颜色概率分布 ,考虑到说话过程中唇区的可见部分 ,包括嘴唇、口内黑区、牙齿、舌头等 ,颜色变化较多 ,对唇区的颜色分布采用由 4 个高斯模型组成的 GMM 来描述 ;嘴唇周围的肤色变化较少 ,采用 2 个高斯模型来描述 .

在描述嘴唇形状方面 ,变形模板得到了广泛的应用 ,因为它可以简单有效地模拟嘴唇的弹形变化 . 但在搜索变形参数的过程中 ,确定好的能量函数是关键问题 . Rabi 等人<sup>[8]</sup>利用像素的上下边缘强度作为能量函数 ,但这不能有效地利用唇区颜色特点信息 ;Chen<sup>[1]</sup>利用嘴唇轮廓线内唇区/非唇区概率比作为能量函数的信息 ,但没有利用边缘信息 ,且不考虑嘴唇轮廓线以外颜色概率分布会使轮廓线趋于较小面积的形状 .

我们在利用变形模板表示先验知识的过程中 ,构造了一个有效利用颜色概率分布信息和边缘知识搜索能量函数 . 首先把唇区和肤色区像素由 RGB 颜色空间转换到 HSV 颜色空间 ,根据预先手动标出的训练数据 ,利用 EM(expectation maximization)算法 ,分别估计出相应的嘴唇区和皮肤区 GMM 参数 . 在搜索过程中 ,根据前一帧图像中特征点的位置确定感兴趣区域 ROI(region of interest) ,由唇区和肤

色区的概率模型计算 ROI 中每个像素属于这两个模型的概率 ,求其差值并做二值化 ,便得到一个 ROI 的嘴唇概率分布图 . 对得到的概率分布图利用  $3 \times 3$  的 Sobel 算子检测出水平和垂直边缘 ,并利用一个适当的阈值将其二值化 ,得到一个 ROI 的边缘分布图 .

在搜索过程中 ,我们采用一个由 3 条四次曲线(外唇)和 2 条二次曲线(内唇)组成的变形模板描述嘴唇形状<sup>[7]</sup> . 其内外唇曲线可由式(1)和式(2)表示 :

$$\text{内唇} \quad y = h \left( 1 - \frac{x^2}{w^2} \right); \quad (1)$$

$$\text{外唇} \quad y = h \left( 1 - \frac{x^2}{w^2} \right) + 4q \left( \frac{x^4}{w^4} - \frac{x^2}{w^2} \right). \quad (2)$$

由于内唇轮廓线不明显 ,在很多时候人眼也难以区分 . 在此 ,我们仅对外唇轮廓进行搜索 ,内唇参数由相应的外唇参数线性拟合估计出来 . 能量函数定义如下 :

$$E = -E_{\text{prob}} - E_{\text{edge}} = -\frac{\sum_{\omega_1} P_m(i) + \sum_{\omega_2} P_s(i)}{\omega_1 + \omega_2} - \frac{\sum_{\omega_3} P_e(i)}{\omega_3}, \quad (3)$$

其中 , $E_{\text{prob}}$  为概率能量 , $\omega_1$  , $\omega_2$  分别表示 ROI 内嘴唇轮廓线内部和外部像素集 , $p_m(i)$  表示像素点  $i$  是否属于唇区 , $p_s(i)$  表示像素点  $i$  是否属于肤色区 , $E_{\text{prob}}$  是 ROI 中轮廓线之内属于嘴唇的像素点数加上轮廓线之外不属于嘴唇的像素点数除以总的像素点数的值 ,在  $0 \sim 1$  之间 ,其值的增大使得唇形轮廓线之内嘴唇的概率尽可能的大 ,唇形轮廓线之外皮肤的概率尽可能的大 ; $\omega_3$  表示外唇轮廓线上的像素集合 , $p_e(i)$  表示像素点  $i$  是否是边缘 . 它的值是唇形轮廓线上边缘像素点的个数除以轮廓线上所有像素点的个数 ,其值也在  $0 \sim 1$  之间 ,其值的增大使得唇形轮廓线上边缘概率尽可能的大 . 根据这一能量函数 ,利用梯度下降法可找到最优的外唇轮廓线 .

在得到外唇参数后 ,我们假设外唇参数内唇参数之间存在着近似线性对应关系 ,可以用式(4)由外唇轮廓线估计内唇参数 :

$$P_{\text{in}} = A \cdot P_{\text{out}}, \quad (4)$$

其中

$$P_{\text{in}} = [W_i \ h_3 \ h_4]^T,$$

$$P_{\text{out}} = [W_o \ h_1 \ h_2 \ q_1 \ q_2 \ 1]^T$$

分别为内、外唇曲线参数 , $W_i$  , $W_o$  分别表示内外唇宽度 ; $h_1$  , $h_2$  分别表示式(2)中上下外唇  $h$  参数 ; $h_3$  , $h_4$  分别表示式(1)中上下内唇  $h$  参数 ; $q_1$  , $q_2$  分

别表示式(2)中上下外唇  $q$  参数;  $A$  为  $3 \times 6$  的线性变换系数矩阵, 可根据手动标出的点由最小二乘法估计.

### 3.2 参考点的跟踪

对于鼻孔点的定位, 首先利用前一帧图像中鼻孔点的位置和两个鼻孔点间的距离, 确定搜索鼻孔点的矩形框. 将框内彩色图像转换为灰度图像, 并根据亮度选择一个适当的阈值做二分类, 会得出两个明显的暗区, 分别计算它们的重心即为两个鼻孔点的位置.

另外, 为了进行姿态估计, 我们还根据局部图像的自身特点, 利用图像编码中运动补偿矢量的搜索方法, 在正面图中跟踪了眼镜边框两点(对于不戴眼镜的录像者, 可以用两个外眼角点来代替).

### 3.3 侧面特征点的跟踪

在录像时选择与人脸肤色反差较大的颜色作为录像背景, 这样在跟踪过程中可以很容易地根据颜色差别将侧面图区分为人脸肤色区和背景区, 进而提取出人脸的侧面轮廓线. 在这条轮廓线上, 从鼻子到下腭区有 3 个明显的突出点, 分别对应于鼻尖点和上下唇突出点. 下腭点的张开点位置可由下唇以下曲率最大的一点得出. 下腭突出点的位置可以从下唇突出点向下腭张开点搜索, 找到斜率与水平线垂直的一点. 当没有这样的点存在时, 取垂直方向上下腭张开点与下唇突出点的中点作为下腭突出点.

## 4 参数计算

### 4.1 滤波

由于光照变化、跟踪算法的精度等因素的影响, 跟踪得到每一帧图像中 4 个参数点的位置会包括一些噪声, 形成高频抖动. 需要利用头部运动姿态在时间上的连续性对跟踪到的参考点位置做平滑处理. 实际处理中我们采用以下低通滤波器对参考点的坐标进行平滑:

$$\begin{aligned} x'(t) &= (x(t-1) + 2x(t) + x(t+1))/4, \\ y'(t) &= (y(t-1) + 2y(t) + y(t+1))/4. \end{aligned} \quad (5)$$

其中  $x(t)$  和  $x'(t)$  分别表示平滑前后的  $t$  时刻特征点  $x$  坐标,  $y(t)$  和  $y'(t)$  则为  $y$  坐标. 对于个别帧的个别点出现的较大误差, 平滑之前在一个交互式界面下利用鼠标手动进行修正.

### 4.2 姿态校正

由于人说话过程中往往会不由自主的带有一些头部的运动, 而我们在计算视觉语音参数时需要去除这些头部全局运动的影响, 因此需要估计头部的全局运动姿态. 在限定的录像条件下, 说话者的头部的全局运动主要是  $x$  方向和  $y$  方向,  $z$  方向运动较小, 所造成的图像伸缩变化很小, 可忽略不计. 它对于侧面图的影响由作为参考的鼻尖点平移得到纠正. 假设眼镜边框上的两点及两个鼻孔点处在同一平面上并与图像平面平行, 头部旋转运动主要是绕  $x$  轴旋转(左右倾斜(绕  $z$  轴旋转)和旋转(绕  $y$  轴旋转)较小, 可忽略不计. 因此, 对第  $n$  帧图像, 总的姿态校正需要计算 3 个变量: 平移量  $\Delta x(n)$  和  $\Delta y(n)$ , 绕  $x$  轴旋转角度  $\theta(n)$ , 如图 2 所示:

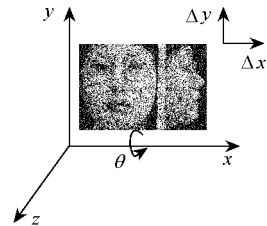


Fig. 2 Global face movement.

图 2 人脸的全局运动

设  $y_h(n) = (R_1(n).y + R_2(n).y)/2$  表示第  $n$  帧图像中眼镜框上的两个参考点的  $y$  坐标均值,  $y(n) = (R_3(n).y + R_4(n).y)/2$  表示第  $n$  帧图像中两个鼻孔点的  $y$  坐标均值, 则有:

$$\Delta x(n) = \frac{1}{4} \sum_{i=1}^4 (R_i(0).x - R_i(n).x), \quad (6)$$

$$\Delta y(n) = \frac{y(0)y_h(n) - y(n)y_h(0)}{y(0) - y_h(0)}, \quad (7)$$

$$\theta(n) = \arccos\left(\frac{y(n) - y_h(n)}{y(0) - y_h(0)}\right), \quad (8)$$

其中  $R_i(0).x$  和  $R_i(n).x$  分别表示初始帧和第  $n$  帧图像中第  $i$  个参考点的  $x$  坐标,  $i=1, 2$  对应于眼镜框上的两个参数点,  $i=3, 4$  对应于两个鼻孔点. 侧面图中的平移量由鼻尖点的  $x$  方向平移与人脸绕  $x$  轴旋转角度  $\theta(n)$  计算得到.

### 4.3 FAP 计算

在得到人脸姿态后, 我们可以计算出在新的姿态下各个人脸特征点在自然状态下的位置, 而后计算出相应点当前位置坐标与这一位置的差值. 并利用根据自然状态下的唇宽  $MW_0$  和口鼻距离  $MNS_0$  求得 FAPU:  $MW$  和  $MNS$  (具体定义可参见 MPEG-4 国际标准<sup>[5]</sup>), 再根据以下两式计算出相应

的 FAP 参数:

FAP#6,7,53 和 54:

$$Fap = (P'_i \cdot x - P_i \cdot x) / MW. \quad (9)$$

另外 20 个 FAP 参数:

$$Fap = (P'_i \cdot y - P_i \cdot y) / MNS. \quad (10)$$

$P'_i \cdot x, P_i \cdot x$  表示相应特征点的当前帧  $x$  坐标和根据姿态计算出的自然状态下  $x$  坐标;  $P'_i \cdot y, P_i \cdot y$  表示相应特征点的当前帧  $y$  坐标和根据姿态计算出的自然状态下  $y$  坐标.

### 5 实验结果及结束语

图 3 从左到右分别是 ROI 用嘴唇 GMM 计算得到的概率;用皮肤 GMM 计算得到的概率;二者之

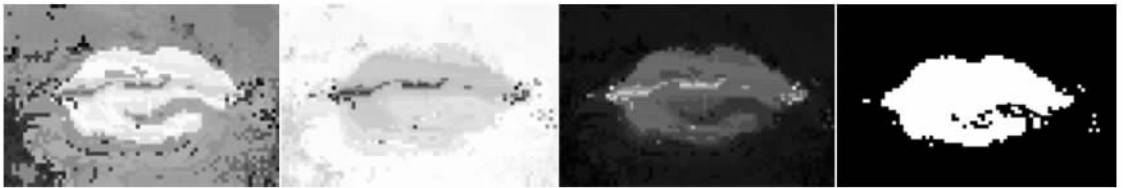


Fig. 3 GMM probability of mouth region in HSV space, skin color space, their difference and binary result.

图 3 HSV 空间嘴唇 GMM 和肤色 GMM 概率值、其差值及二值化结果



Fig. 4 Binary edge.

图 4 边缘二值化结果

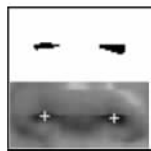


Fig. 5 Nostrils.

图 5 鼻孔点

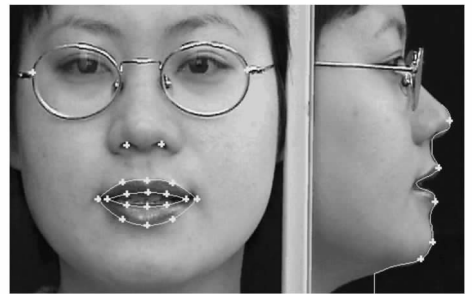


Fig. 6 Face feature points tracking result.

图 6 人脸特征点跟踪的结果



Fig. 7 Synthesis result based on estimated parameters.

图 7 利用估计参数合成的图像

本文给出了一种简便有效的视觉语音参数估计方法,在不需要多摄像机和预先建立三维人脸模型

差及二值化结果.从图 3 中可以看出,经过以上处理后,嘴唇区变得较为明显.图 4 是计算出的边缘强度经二值化后的结果,显示出明显的嘴唇轮廓线.我们在 300 帧图像上对内唇参数线性估计结果与真实数据进行了比较,平均幅度误差为 7.95%,线性相关性为 0.87,说明线性映射可以得到较好的估计.

图 5 是鼻孔搜索结果,上面是亮度二值化后的结果,下面是确定的鼻孔点位置.典型的整体跟踪结果如图 6 所示.

我们已采用本文所述的方法成功地从视频流中获取了大量的视觉语音参数,并基于这些数据实现了一个由 FAP 参数控制的视觉语音合成系统,图 7 是利用估计出的 FAP 参数所合成的正面和侧面图.

的情况下得到较为准确的三维 FAP 参数.在最为困难的外唇轮廓线跟踪中,我们将统计学习方法和基于规则的方法结合起来,有效地利用了颜色概率统计信息和形状、边缘等先验知识,取得了较好的跟踪效果.在计算 FAP 参数的过程中,我们利用低通滤波消除了高频噪声的影响,巧妙利用人脸上最为突出的 4 个参考点估计出主要的人脸姿态运动,从而消除了人脸全局运动的影响.

下一步的工作包括采用更为复杂的姿态估计算法,以便减小对录像者说话过程中头部运动的限制.另外,用更多的参考点进行姿态估计,使姿态估计结果变得更为准确.

## 参 考 文 献

- 1 T. Chen. Audiovisual speech processing. *IEEE Signal Processing Magazine*, 2001, 18(1): 9~21
- 2 P. Y. Hong, Z. Wen, T. S. Huang. Real-time speech-driven face animation with expressions using neural networks. *IEEE Trans. Neural Networks*, 2002, 13(4): 916~927
- 3 J. W. Kim, M. Song, I. J. Kim, *et al.* Automatic FDP/FAP generation from an image sequence. The 2000 IEEE Int'l Symposium on Circuits and Systems, ISCAS 2000, Geneva Switzerland, 2000
- 4 N. Sarris, N. Grammalidis, M. G. Strintzis. FAP extraction using three-dimensional motion estimation. *IEEE Trans. Circuits and Systems for Video Technology*, 2002, 12(10): 865~876
- 5 International standard, Information technology—Coding of audiovisual objects—Part 2: Visual; Amendment 1: Visual extensions, ISO/IEC 14496-2:1999/Amd.1:2000(E)
- 6 R. Wang, W. Gao, J. Y. Ma. An approach to robust and fast locating lip motion. The 3rd Int'l Conf. Multimodal Interfaces, Heidelberg, 2000
- 7 A. W. C. Liew, S. H. Leung, W. H. Lau. Region-based approach to robust lip contour extraction. *Electronics Letters*, 2000, 36(15): 1272~1274
- 8 G. Rabi, S. W. Lu. Energy minimization for extracting mouth curves in a facial image. The Int'l Conf. Intelligent Information Systems, Bahamas, 1997



**Wang Zhiming**, born in 1968. Doctor and engineer of University of Science and Technology Beijing. His main research interests include image processing, pattern recognition, interaction between speech and facial image.

王志明, 1968年生, 博士, 工程师, 主要研究方向为图像处理、模式识别、语音与人脸图像的交互作用等。



**Cai Lianhong**, born in 1945. Professor and Ph. D. supervisor of Tsinghua University. Her main research interests include speech processing and synthesis, multimedia technology and biometric recognition.

蔡莲红, 1945年生, 教授, 博士生导师, 主要研究方向为语音处理与合成、多媒体技术、生物特征识别等。



**Ai Haizhou**, born in 1964. Doctor, professor, and Ph. D. supervisor of Tsinghua University. His main research interests include computer vision and pattern recognition.

艾海舟, 1964年生, 博士, 教授, 博士生导师, 主要研究方向为计算机视觉、模式识别。

## Research Background

This study is supported by the National Research Foundation for the Doctoral Program of Higher Education of China under grant No. 20010003049, and Scholastic Science Foundation of University of Science and Technology, Beijing under grant No. 20040509190.

Human speech is bimodal in nature. Both audio speech and visual speech are produced by articulatory organs. Visual speech means the movements of visible articulatory organs of the speaker, such as lips, tongues, jaws, facial muscles, etc. Study of the inherent relationship between audio speech and visual speech gains more and more attention in recent years. There are many applications in this domain, such as lip reading, visual speech synthesis, audio-visual speech recognition, etc. The fundamental problem of these applications is to describe the visual speech qualitatively and quantitatively. Among many parametric descriptions, FAP (facial animation parameter) defined by MPEG-4 is the most popular one. How to acquire huge amounts of visual parameters (FAP) from video automatically, fast and precisely is a difficult problem so far.

We study the techniques of tracking facial feature points and lip contour precisely and robustly, and face global movement and visual parameters are estimated automatically. Visual parameters acquired are used in visual speech synthesis, and experimental results are attractive. Algorithms proposed in this paper could also be used in face animation, facial expression recognition, etc.