融合时空特征的视频序列表情识别

王晓华*①② 夏 晨① 胡 敏① 任福继①③
①(合肥工业大学计算机与信息学院 合肥 230009)
②(江苏省物联网移动互联技术工程实验室 淮安 223001)
③(德岛大学先端技术科学教育部 日本 7708502)

摘 要：针对视频表情识别，静态特征不能有效描述人脸区域沿时间轴动态变化信息的局限，该文提出一种融合动态纹理信息和运动信息的表情识别方法，借鉴 LBP-TOP 原理，提出具有时空域描述能力的时空韦伯局部描述子(STWLD)来提取动态纹理信息，同时采用分块光流直方图(BHOF)描述运动信息，最后利用 SVM 对融合后的纹理和运动信息完成表情分类。在 CK+和 MMI 表情数据库上的交叉实验结果表明，相比基于单一特征的识别方法，所提方法取得了更好的效果；与其他相关方法的对比实验也验证了该方法的优越性。

关键词：视频序列；表情识别；时空韦伯局部描述子；分块光流直方图特征
中图分类号： TP391.43 **文献标识码**： A **文章编号**：1009-5896(2018)03-0626-07
DOI: 10.11999/JEIT170592

Facial Expression Recognition Based on the Fusion of Spatio-temporal Features in Video Sequences

WANG Xiaohua①② XIA Chen① HU Min① REN Fuji①③
①(*School of Computer and Information of Hefei University of Technology*, *Hefei* 230009, *China*)
②(*The Laboratory for Internet of Things and Mobile Internet Technology of Jiangsu Province*, *Huaian*, 223001 *China*)
③(*Graduate School of Advanced Technology & Science*, *University of Tokushima*, *Tokushima* 7708502, *Japan*)

Abstract: For facial expression recognition based on video sequences, the changing information of facial regions along the time axis can be described by dynamic descriptors more effectively than static descriptors. This paper proposes an expression recognition method based on the dynamic texture and motion information, learning from the principle of Local Binary Pattern on Three Orthogonal Planes (LBP-TOP), Spatio-Temporal Weber Local Descriptor (STWLD) is proposed to describe the dynamic texture feature information of the facial expression sequence. Moreover, using Block-based Histogram of Optical Flow features (BHOF), the motion information can be described. Through the combination of the dynamic texture and motion information, and finally SVM is applied to complete the expression classification. The results of the cross experiments on the CK + and MMI expression database show that the method achieves better performance than methods using the single descriptors. The comparison experiments with other related methods also prove the superiority of the method.

Key words: Video sequences; Expression recognition; Spatio-Temporal Weber Local Descriptor (STWLD); Block-based Histogram of Optical Flow (BHOF)

1 引言

表情是人类情感表达的重要方式。人脸表情识别技术是计算机理解人类的根基，是实现人机交互智能化的有效手段。近年来，人脸表情识别技术已成为人工智能和计算机视觉领域的研究热点，同时也是最具有挑战性的研究课题之一，在人机交互、教育、通信、医疗等领域有着广泛的应用前景。

以往的表情识别方法大多是以静态图像为对象分析表情。通常可以分为两大类方法，一类基于几何特征，该类方法利用人脸局部特征点、区域的形状以及位置信息描述面部表情，优点是所需的数据量小，能够反映面部表情变化的宏观结构信息；主要有活动轮廓模型(AAM)[1]，受限局部模型(CLM)[2]，可变形部件模型(DPM)[3]。另一类基于外观特征，这类方法利用图像的全部像素进行数据统

收稿日期：2017-06-20；改回日期：2017-11-28；网络出版：2017-12-27
*通信作者：王晓华 xh_wang@hfut.edu.cn
基金项目：国家自然科学基金(61672202, 61432004, 61300119)，国家自然科学基金深圳联合基金重点项目(U1613217)，江苏省物联网移动互联技术工程实验室开放课题(JSWLW-2017-017)
Foundation Items: The National Natural Science Foundation of China (61672202, 61432004, 61300119), The National Natural Science Foundation of China -Shenzhen Joint Foundation (Key Project) (U1613217), Open foundation of The Laboratory for Internet of Things and Mobile Internet Technology of Jiangsu Province (JSWLW-2017-017)

计。与几何特征方法相比，该类方法不需要定义和追踪人脸特征点，因此更易于实现。常用的特征算子有局部二值模式 (LBP)[4], Gabor 滤波[5]，以及韦伯局部特征(WLD)[6]等。

然而对于视频表情识别，上述方法很难反映面部表情的变化信息。众所周知，面部表情是一个连续的变化过程，相邻的视频序列帧之间存在着相关性，若仅考虑视频序列的空间信息，忽略视频序列的时域信息，则会丢失有效特征信息，影响最终的识别效果。视频序列时域上的动态信息则是视频表情识别区别于静态图片表情识别的关键。文献[7]提出了将 LBP 邻域从 2 维平面扩展到 3 维空间得到 VLBP(Volume Local Binary Pattern)，该特征算子与 LBP 相比，考虑到了时域信息，因此更适合分析视频的面部表情，但其特征维数过大。为了简化 VLBP，文献[7]提出了用 LBP-TOP(Local Binary Pattern on Three Orthogonal Planes)描述视频情感序列，从时空域 3 个平面提取视频序列的 LBP 特征。文献[8]提出了利用多尺度时空局部方向角方法表征视频人脸表情，通过在 3 个正交平面上(*X*-*Y*, *X*-*T*, *Y*-*T*)上统计局部方向角特征，充分利用了图像序列时空特征信息。文献[9]提出了一种时空纹理图特征，该方法利用 3 维 Harris 角点来获取视频的时空域信息。文献[7–9]都考虑了视频序列的时空域信息，亦取得了不错的识别效果。受文献[7–9]的启发，考虑到 WLD 特征[6]是一种计算简单易于实现、鲁棒、高效的局部纹理描述子，但只能反映某个具体时间点图像的表情状态信息，忽略了时域上连续信息的情况。本文将 WLD 特征扩展到时空域，提出了时空韦伯局部描述子(STWLD)，从 3 个正交平面获取视频序列的纹理信息。相对于原始 WLD 特征算法：(1)考虑到了时域上的特征信息，表征更全面。(2)保留了对光照和噪声的鲁棒性。

除了纹理变化信息，运动信息也是视频表情识别的重要依据。光流 [10]是提取视频序列运动信息重要的方法，它能够反映前后帧图像间的相互关系。文献[11]提出了一种全局光流特征描述人脸表情的方法，首先采用多分辨率策略对图像分层，然后分块求解光流特征，最后在局部区域内统计各点光流运动情况。文献[12]提出了利用光流追踪相邻帧图像之间特征点运动信息，然后利用特征点运动信息分析表情。文献[11,12]将光流应用到人脸表情识别，但特征点追踪易丢失和运动特征统计粗略，识别效果不佳。针对上述问题，本文提出了一种分块光流直方图特征方法，该方法考虑到表情动作微弱，相邻图像帧之间的变化较小，通过将视频序列相邻帧的光流特征叠加统计得到更明显的光流特征，解决了光流微弱的问题，兼顾了局部细节信息。但因光流的效果易受到光线变化以及非刚性运动的影响，单独采用光流进行表情识别通常效果不理想。

综上所述，无论是纹理信息还是运动信息都只能从单一方面描述表情序列，鉴别能力有限。纹理特征的优势在于描述局部像素点分布，但不擅长描述像素点变化信息，而运动特征恰恰相反。因此，两者具有很强的互补性，融合使用应该能获得更好的识别效果。基于以上分析，本文提出了一种融合 STWLD 和分块光流直方图(BHOF)特征的表情识别方法。首先，对视频序列进行预处理，将预处理后的表情序列视为沿着时间轴堆叠而成的 3 维时空立方体；然后利用融合的时空特征算法提取 3 维时空立方体得到的特征向量作为视频序列最终的特征向量；最后，通过 SVM 获得分类结果。与现有的特征方法相比，融合过后的时空特征有机地结合了表情序列的动态纹理信息以及运动信息，充分考虑了表情序列图像的像素点分布与像素点变化，相较于单一形式的特征，更能够包含视频序列的有效信息，得到更为精确、可靠的结果。

2　STWLD 和 BHOF 特征

2.1　韦伯局部特征(WLD)

WLD 由差分激励 $\xi(x_c)$ 和方向 $\theta(x_c)$ 构成。由于模拟人类的感知需要找到图像中较为显著的变化，所以 WLD 利用图像中每个像素 x_c 和它周围像素之间的强度差分的比值作为像素 x_c 的变化。具体来说，图像中每一像素 x_c 的差分激励为

$$\xi(x_c)=\arctan\left[\sum_{i=0}^{n-1}\left(\frac{x_i-x_c}{x_c}\right)\right] \tag{1}$$

式中，x_i 表示像素点 x_c 的第 i 个邻域像素，n 为邻域像素数目。方向 $\theta(x_c)$ 表示像素点 x_c 的梯度方向，即

$$\theta(x_c)=\arctan(v_a/v_b) \tag{2}$$

式中，v_a 和 v_b 分别由滤波窗口 $\boldsymbol{f}_0$ 和 $\boldsymbol{f}_1$ 计算出，滤波窗口 $\boldsymbol{f}_0$ 和 $\boldsymbol{f}_1$ 分别为

$$\boldsymbol{f}_0=\begin{bmatrix}0&-1&0\\0&0&0\\0&+1&0\end{bmatrix},\ \boldsymbol{f}_1=\begin{bmatrix}0&0&0\\+1&0&-1\\0&0&0\end{bmatrix} \tag{3}$$

为了达到简化的目的，通常将梯度方向 θ 量化到 T 个主要方向，具体的量化方程为

$$\phi_t=f_q(\theta')=\frac{2t}{T}\pi,\ t=\mathrm{mod}\left(\left\lfloor\frac{\theta'}{2\pi/T}\right\rfloor,T\right) \tag{4}$$

$$\theta'=\arctan2(v_a,v_b)+\pi,$$

$$\arctan2(v_a,v_b)=\begin{cases}\theta, & v_a>0\,\&\,v_b>0\\ \theta+\pi, & v_a>0\,\&\,v_b<0\\ \theta-\pi, & v_a<0\,\&\,v_b<0\\ \theta, & v_a<0\,\&\,v_b>0\end{cases} \tag{5}$$

由式(1)~式(5)得出差分激励 $\xi(x_c)$ 和方向 $\theta(x_c)$，得到 2 维 WLD 直方图 $\{\mathrm{WLD}(\xi_j,\phi_t), j=0,1,\cdots,N-1, t=0,1,\cdots,T-1\}$，其中 N 是图像维度。为了增强 WLD 特征的区分能力，2 维 WLD 直方图特征被转化为 1 维直方图特征。

2.2 时空韦伯局部描述子(STWLD)

人脸表情通常是一个动态的连续变化过程，因此，动态信息更能够有效地描述人脸表情。但是传统的韦伯局部描述子只提取了空间域上的纹理信息，未考虑时域上的纹理变化信息，从而导致视频表情识别效果不佳。为了有效地提取视频序列在时空域上的纹理变化信息，本文将 WLD 扩展到 3 维空间，提出时空韦伯局部描述子。该描述子利用相交于中心像素点的 3 个正交平面(X-Y, X-T, Y-T)，将在这 3 个正交平面上提取的特征串联起来得到 STWLD 特征。X-Y 平面表征视频序列的局部空间信息，X-T 和 Y-T 平面则表征视频序列的时域信息。视频序列中的每一帧图像上的任一像素点都可以看作是这 3 个正交平面的交点，对任意一中心像素点，利用式(1)~式(5)可从 3 个正交平面分别计算得到 XY-WLD, XT-WLD, YT-WLD。如图 1 所示，分别提取图像序列在 3 个平面上的特征之后，将 3 个平面上的 WLD 直方图级联构成 STWLD 特征，所得最终特征向量包含了面部表情在垂直和水平方向上的纹理信息。STWLD 特征定义如式(6)：

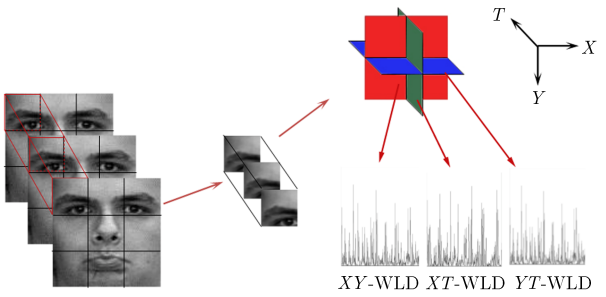


图 1 视频序列中的 STWLD 特征

$$\boldsymbol{H}_{i,j}=\sum_{x,y,t} p\{f_i(x,y,t)=j\};\quad i=0,1,2;\ j=0,1,\cdots,L-1 \tag{6}$$

式中，$i=0,1,2$ 分别代表 X-Y,X-T,Y-T 平面；$j=0,1,\cdots,L-1$, L 表示直方图总标签数。$f_i(x,y,t)$ 表示中心像素点 (x,y,t) 在 i 平面的 WLD 编码。函数

$$p(b)=\begin{cases}0, & b\text{为假}\\ 1, & b\text{为真}\end{cases} \tag{7}$$

2.3 分块光流直方图特征(BHOF)

为了更好地描述表情的变化信息，根据 Horn-Schunck[10]方法提取表情序列的全局光流特征。具体如下，假设在 t 时刻时，图像中某像素点 (x,y) 处的像素点值为 $I(x,y,t)$，光流约束基本方程为

$$I_x u_x + I_y v_y + I_t = 0 \tag{8}$$

式中，I_x,I_y,I_t 分别表示 t 时刻像素点 (x,y) 的像素值在水平、垂直、时间 3 个方向上的偏导数，u_x, v_y 分别表示该像素点在水平和垂直方向上的光流矢量分量。因为光流基本方程中有两个未知变量，而只有一个光流约束方程是解不出来的。为了求解光流的两个未知变量，Horn 和 Schunck[10]假设全局的光流场是平滑的，即光流场既满足式(8)的约束也满足全局平滑性的约束。定义目标函数式(9)，使目标函数式(9)最小化，即可获得图像中每个像素点的速度矢量场。

$$E=\iint\left(I_x u_x+I_y v_y+I_t\right)^2 \mathrm{d}x\mathrm{d}y + \lambda\iint\left\{\left(\frac{\partial u_x}{\partial x}\right)^2+\left(\frac{\partial u_x}{\partial y}\right)^2+\left(\frac{\partial v_y}{\partial x}\right)^2+\left(\frac{\partial v_y}{\partial y}\right)^2\right\}\mathrm{d}x\mathrm{d}y \tag{9}$$

其中，$\frac{\partial u_x}{\partial x},\frac{\partial u_x}{\partial y},\frac{\partial v_y}{\partial x},\frac{\partial v_y}{\partial y}$ 分别代表着光流分量 u_x,v_y 在 x 方向上和 y 方向上的偏导数。参数 λ 表示平滑度的约束参数。为了得到图像中每一像素点的光流分量 u_x,v_y。式(9)式可最终化为

$$u_{x,y}^{k+1}=\overline{u}_{x,y}^{k}-\frac{I_x\left(I_x\overline{u}_{x,y}^{k}+I_y\overline{v}_{x,y}^{k}+I_t\right)}{I_x^2+I_y^2+\lambda^2} \tag{10}$$

$$v_{x,y}^{k+1}=\overline{v}_{x,y}^{k}-\frac{I_x\left(I_x\overline{u}_{x,y}^{k}+I_y\overline{v}_{x,y}^{k}+I_t\right)}{I_x{}^2+I_y{}^2+\lambda^2} \tag{11}$$

式中，$\left(u_{x,y}^{k},v_{x,y}^{k}\right)$表示在 k 次迭代中，图像中 (x,y) 处的光流。$k=0$，光流初始值为 0。当迭代前后的差值小于设定的阈值时，迭代结束。由式(10)，式(11)可得到的光流 $\left(u_{x,y},v_{x,y}\right)$，通过式(12)可得到像素点 (x,y) 的光流向量角为

$$\theta_{x,y}=\arctan\left(\frac{v_{x,y}}{u_{x,y}}\right) \tag{12}$$

其中，$0\le\theta_{x,y}\le 2\pi$。像素点 (x,y) 的光流幅值为

$$F_{x,y}=\sqrt{u_{x,y}^2+v_{x,y}^2} \tag{13}$$

把 $0\le\theta_{x,y}\le 2\pi$ 的光流方向分成 $n(n=8)$ 个方向，根据像素点 (x,y) 的光流向量角 $\theta_{x,y}$ 的范围划分区间，以像素点 (x,y) 的光流幅值 $F_{x,y}$ 作为像素点 (x,y) 在划分区间内的权值。对光流图像的每个像素点按 n 个方向统计光流特征信息，得到光流直方图。

针对相邻帧之间表情变化较小，导致光流特征微弱。本文提出分块光流直方图特征算法，通过叠

加视频序列的光流得到更明显的光流特征。根据式(12)，式(13)得视频序列的光流图像序列，将光流图像序列视为 3 维时空体，具体如图 2 所示，3 维时空体由时间轴上连续光流图像排列组成，然后将 3 维时空体分块，统计每一子块的光流直方图，最后将所有子块的光流直方图结合成最终的 BHOF 特征向量。

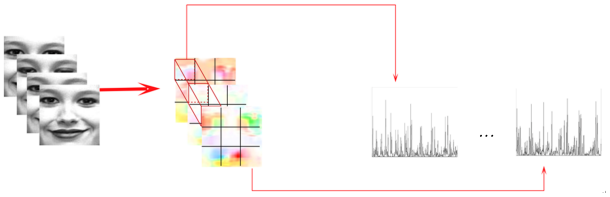


图 2 视频序列中的 BHOF 特征

2.4 特征融合

将 2.2 节和 2.3 节计算得到的 STWLD 和 BHOF 特征归一化到[0,1]，然后根据式(14)整合为一个向量。

$$\boldsymbol{F}=\gamma\boldsymbol{P}+(1-\gamma)\boldsymbol{Q},\quad 0\le\gamma\le 1 \tag{14}$$

式中，$\boldsymbol{P}$ 表示 STWLD 特征，$\boldsymbol{Q}$ 表示 BHOF 特征，$\boldsymbol{F}$ 表示加权融合过后的最终特征向量。参数 γ 的取值依赖于每个特征的表现，依据 4.2 节的分析，本文设定 $\gamma=0.54$。对特征融合以及权重 γ 的分析请见 4.2 节。

3 基于复合情感时空特征的表情识别的方法

本文算法包含 3 个阶段：视频预处理、表情特征提取、表情分类识别。具体步骤如下：

(1)首先利用 Viola-Jones 人脸检测器检测人脸，然后检测两眼位置，计算两眼的中心点，进行人脸对齐。最后，对表情序列图像进行尺度归一化处理。本文所有的表情序列图像都归一化到 96×96 像素；

(2)对预处理后的表情序列进行分块处理，将表情序列视为 3 维时空体，按照互不重叠，大小均匀的策略划分为 3 维时空矩形块。对每一子块分别提取 STWLD 和 BHOF 特征，然后分别将所有子块的 STWLD 以及 BHOF 特征级联，最后按 2.4 节方法加权结合两种特征。本文系统流程框架如图 3 所示；

(3)SVM 分类器因具有如下特性：(a)高泛化性能；(b)能够处理高维度特征；(c)基于统计学习理论，被广泛应用于人脸表情识别。本文选取 SVM 作为分类识别阶段的分类器。为了获得更好的识别结果，实验采用 RBF 核函数作为 SVM 核函数，并采用

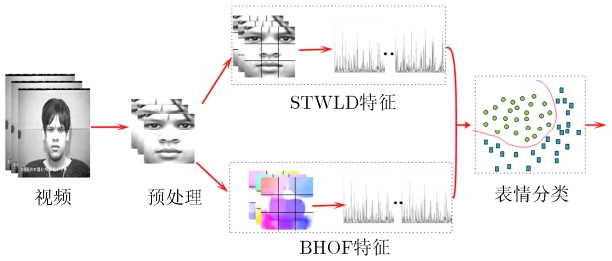


图 3 面部表情识别系统流程框架图

“one-versus-all”的策略实现多分类。在最优核参数的选择上，本文采用网格搜索和 10 次交叉验证方法确定核参数。

4 实验结果与分析

4.1 实验数据库

为了验证本文算法的有效性，实验采用目前广泛应用的表情数据库：CK+数据库[13]和 MMI 数据库[14]。CK+库是当前用来评估面部表情识别效果最广泛的数据库，该数据库包含 123 名对象的 593 个视频序列。鉴于该数据库情感类别标签未完全标注，本文选取其中已标注的 309 个视频序列进行相关实验，所选 6 类基本人脸表情(生气，厌恶，害怕，高兴，悲伤，惊讶)样本数分别为(45, 59, 25, 69, 28, 83)。MMI 数据库包含 213 个带表情标签的视频序列。与 CK+数据库不同，视频序列中人物的情感变化分为中性，开始，高峰，回落 4 个阶段。本文选取从中性到表情高峰阶段之间的视频序列图像。考虑到标签和正面人脸等因素，选择 203 个视频序列作为样本进行相关实验。所选 6 类基本人脸表情(生气，厌恶，害怕，高兴，悲伤，惊讶)样本数分别为(32, 28, 28, 42, 32, 41)。本文实验是在 Windows 7 系统下，使用 VS2013+Opencv 2.4.9 实现。本文采用 10 组交叉实验方案。

4.2 实验结果分析

在特征提取过程中，表情序列分块数会影响后续的识别效果，分块过少会导致局部区域表情特征提取不充分。分块数过多则会增加特征维数以及时间复杂度。图 4 给出了分块数与平均识别率之间的对应关系。从图 4 可以看出，随着分块数增加，人脸各区域表征的精细程度逐渐加深，识别率呈上升趋势。分块数达 64 时，在 CK+和 MMI 库上的识别效果最优。分块数超过 64 以后，识别效果虽有所提升，但增幅趋于平缓，且此时计算量加大。考虑到分块数与识别性能之间的有效性和简便性，本文选取的分块数目为 64。

相比 MMI 数据库，考虑到 CK+库图像质量高，无眼镜、围巾等物体遮挡以及头部姿态变化幅度小

等原因，本文在 CK+库上作了一系列详细的对比实验，来验证本文时空融合特征的有效性。为了验证结合两种特征的方法优于单个特征的识别效果，本文比较了 STWLD 特征、BHOF 特征、级联和加权结合两种特征的识别性能。图 5 所示为相关方法的识别结果。从图 5 中可以看出，由于 STWLD 综合考虑了时空域信息，包含更多有效的信息，STWLD 比 BHOF 特征更具有辨别能力。结合两种特征的方法识别效果要优于单个特征的识别结果。此外加权融合后的特征识别效果要优于级联方法的识别结果。

以往的大多数研究方法对于特征向量融合往往只是简单的级联是在一起。这种级联的策略是假设两种特征对于识别结果有着相同的贡献为前提的。但是，从前面的实验分析来看，并不是每个特征都具有同等重要性，某些特征对最终的识别结果影响更大。因此，如果两种特征按照同等重要性融合，就不能充分发挥两种特征结合的优势，达到最佳的识别效果。基于以上分析，本文提出一个加权策略来提高最终的识别结果。详细的加权策略如下，首先，我们获得两种特征识别率 $R=\{R_1,R_2\}$，权值计算公式为

$$w_i=\frac{\|R_i\|}{\|R\|} \tag{15}$$

这里 $\|\cdot\|$ 是 $L1$ 范式，R_i 表示单个特征的识别结果。每个特征权重向量如式(16)：

$$\boldsymbol{f}_w=w_i\boldsymbol{f}_i \tag{16}$$

这里 w_i 表示特征的权值，$\boldsymbol{f}_i$ 表示特征向量，$i=1$ 表示 STWLD 特征，$i=2$ 表示 BHOF 特征。依据 STWLD 和 BHOF 特征的识别结果确定它们的权值，识别效果越好分配越大的权重。

为了研究权值分配对识别准确率的影响，本文对融合方法在 γ 不同取值下的识别性能进行了实验。图 6 给出了参数 γ 在 CK+以及 MMI 库上不同取值下的识别结果。分析图 6 可知，随着参数 γ 值不断增加，赋予的 STWLD 特征权重越高，识别性能不断提升，这是由于 STWLD 特征比 BHOF 特征包含更多的纹理信息。当 $\gamma=0.54, 0.56$，在 CK+库上的识别效果最佳，且随着权值继续增加，识别效果会有所降低，即 BHOF 权重变小也会影响整体的识别准确率，这也证明了两种特征的互补性。当 $\gamma=0.52,0.54$ 时，在 MMI 库上识别效果最佳。综合 CK+和 MMI 库上的识别结果，本文选取 $\gamma=0.54$ 作为 STWLD 特征的权重。

为了详细分析各种特征算法的识别效果，表 1–表 3 分别给出了 STWLD, BHOF 以及时空融合特征算法的混淆矩阵。实验结果表明，如果只利用单一的 STWLD 或 BHOF 特征，分别只达到 89.3%和 76.1%的准确率，而融合两类特征后的准确率提升到 91.6%。这一方面验证了本文提出的 STWLD 特征对视频序列具有较强的描述能力，同时也充分验证了 STWLD 和 BHOF 特征具有良好的互补性。

为了验证本文方法的有效性，将本文方法与文献[7,13,15–18]在 CK+库的识别结果进行了比较。文

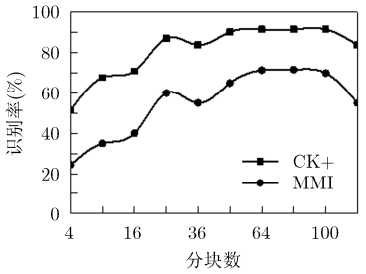


图 4 分块数和识别率之间的关系

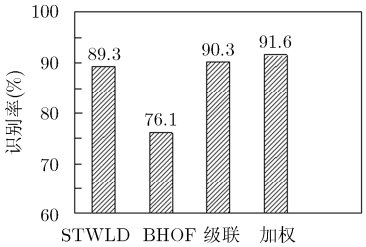
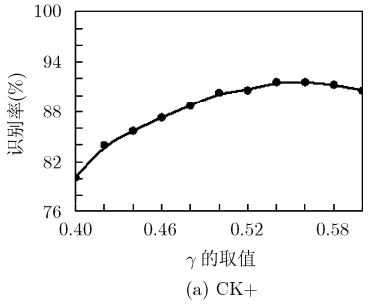
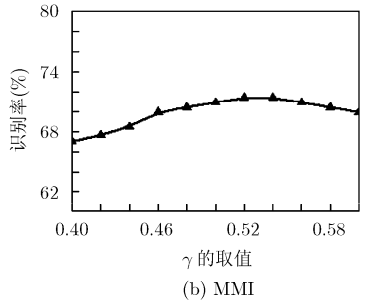


图 5 单个特征、级联以及加权融合方法在 CK+库上的识别结果



(a) CK+



(b) MMI

图 6 参数 γ 在 CK+, MMI 数据库上不同取值下的识别性能

表 1　CK+数据库上 STWLD 特征的表情识别结果

表情	生气	厌恶	害怕	高兴	悲伤	惊讶	平均(%)
生气	**40**	1	0	0	4	0	88.9
厌恶	3	**52**	2	2	0	0	88.1
害怕	3	0	**15**	2	3	2	60.0
高兴	0	0	0	**69**	0	0	100.0
悲伤	5	2	2	0	**17**	2	60.7
惊讶	0	0	0	0	0	**83**	100.0
总计							89.3

表 2　CK+数据库上 BHOF 特征的表情识别结果

表情	生气	厌恶	害怕	高兴	悲伤	惊讶	平均(%)
生气	**31**	3	2	2	6	1	68.9
厌恶	5	**48**	2	3	1	0	81.3
害怕	5	1	**11**	3	2	3	44.0
高兴	3	0	2	**60**	3	1	86.9
悲伤	5	0	2	3	**15**	3	53.6
惊讶	6	1	2	1	3	**70**	84.3
总计							76.1

表 3　CK+数据库上融合 STWLD 和 BHOF 特征的识别结果

表情	生气	厌恶	害怕	高兴	悲伤	惊讶	平均(%)
生气	**42**	0	0	0	2	1	93.3
厌恶	2	**55**	1	1	0	0	93.2
害怕	3	0	**17**	2	1	2	68.0
高兴	1	0	0	**68**	0	0	98.6
悲伤	5	1	3	0	**18**	1	64.3
惊讶	0	0	0	0	0	**83**	100.0
总计							91.6

献[7,13]仅考虑单一形式的信息，文献[15,16]复杂度过高。文献[17]考虑的是时空信息，利用的是金字塔稀疏编码特征。文献[18]采用自学习多速率编码神经网络。比较结果如表 4 所示，由于结合了纹理信息与运动信息，本文融合算法在 CK+库上达到了91.6%的识别率，优于文献[7,13,15–18]中的方法。

表 5 给出了不同方法在 MMI 库上的平均识别率。分析表 5 数据可知，各种方法在 MMI 库上的识别效果远低于 CK+库上的识别结果。这是因为 MMI 库头部姿态变化幅度大，存在帽子、眼镜、围巾等遮挡物。文献[7,19]未考虑运动信息。文献[15,20]利用的是运动单元模型，仅考虑运动信息，且实现起来较为复杂。文献[18,21]采用神经网络，需要大量的样本数据，但是 MMI 库样本数量较小限制了它们的识别效果。从实验结果来看，本文时空融合方法在 MMI 库上的识别率达到了 71.43%，优于文献[7,15,18-21]的方法。

表 4 不同方法在 CK+数据库上平均识别率比较

算法	特征	识别率(%)
文献[7]	VLBP	86.1
	LBP-TOP	88.3
文献[13]	AAM shape and AAM	88.7
文献[15]	IABN	88.1
文献[16]	Phog-top and optical flow	90.9
文献[17]	learned spatiotemporal	81.4
文献[18]	Multi-Velocity Encoder	90.6
本文算法	STWLD and BHOF	91.6

表 5 不同方法在 MMI 数据库上平均识别率比较

算法	特征	识别率(%)
文献[7]	VLBP	62.07
	LBP-TOP	59.60
文献[15]	IABN	62.50
文献[18]	Multi-Velocity Encoder based	66.15
文献[19]	Gabor and Geometry	70.67
文献[20]	ITBN	59.70
文献[21]	DTAGN	70.24
本文算法	STWLD and BHOF	71.43

5　结束语

本文提出的复合时空特征融合了 STWLD 和 BHOF 特征，与其他现有算法相比较,具有如下特点：(1)将 2 维 WLD 扩展到 3 维时空域，能够描述图像像素在空间域和时间域上的分布，且保留了原始 WLD 特征对于光照和噪声变化鲁棒性的优点。(2)将光流图像分成若干区域进行特征统计，增强特征的描述能力。(3)从纹理信息和运动信息方面综合描述视频序列，避免了单一特征的局限性，获得的特征描述更加全面。

本文方法是以正面人脸视频为对象分析表情，主要考虑表情序列在时空域的信息，方法简单易于实现，具有较好的识别效果，且计算量适中，适用于动态人脸表情识别。然而当视频序列存在遮挡、头部姿态变化情况时，会影响本文方法的识别效果；另外，人类情感的表达往往通过多种形式表达，除了面部表情，姿态动作、语言、生理信号等都是常用的表达方式，如何有效地融合多模态信息提高情感识别的准确率将会是下一步所要研究的工作。

参 考 文 献

[1] CHEON Y and KIM D. Natural facial expression recognition using differential-AAM and manifold learning[J]. *Pattern*

Recognition, 2009, 42(7): 1340–1350. doi: 10.1016/j.patcog.2008.10.010.

[2] PAN Z, POLCEANU M, and LISETTI C. On constrained local model feature normalization for facial expression recognition[C]. International Conference on Intelligent Virtual Agents. Los Angeles, CA, USA, 2016: 369–372. doi: 10.1007/978-3-319-47665-0_35.

[3] ZHU X and RAMANAN D. Face detection, pose estimation, and landmark localization in the wild[C]. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 2012: 2879–2886. doi: 10.1109/CVPR.2012.6248014.

[4] ZHAO L, WANG Z, and ZHANG G. Facial expression recognition from video sequences based on spatial-temporal motion local binary pattern and Gabor multiorientation fusion histogram[J]. *Mathematical Problems in Engineering*, 2017, (1): 1–12. doi: 10.1155/2017/7206041.

[5] ZHOU J, ZHANG S, MEI H, *et al*. A method of facial expression recognition based on Gabor and NMF[J]. *Pattern Recognition and Image Analysis*, 2016, 26(1): 119–124. doi: 10.1134/S1054661815040070.

[6] CHEN J, SHAN S, HE C, *et al*. WLD: A robust local image descriptor[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 32(9): 1705–1720. doi: 10.1109/TPAMI.2009.155.

[7] ZHAO G and PIETIKAINEN M. Dynamic texture recognition using local binary patterns with an application to facial expressions[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(6): 915–928. doi: 10.1109/TPAMI.2007.111.

[8] 付晓峰, 付晓鹃, 李建军, 等. 视频序列中基于多尺度时空局部方向角模式直方图映射的表情识别[J]. 计算机辅助设计与图形学学报, 2015, 27(6): 1060–1066.
FU Xiaofeng, FU Xiaojuan, LI Jianjun, *et al*. Facial expression recognition using multi-scale spatiotemporal local orienta-tional pattern histogram projection in video sequences[J]. *Journal of Computer Aided Design & Computer Graphics*, 2015, 27(6): 1060–1066.

[9] KAMAROL S K A, JAWARD M H, PARKKINEN J, *et al*. Spatiotemporal feature extraction for facial expression recognition[J]. *IET Image Processing*, 2016, 10(7): 534–541. doi: 10.1049/iet-ipr.2015.0519.

[10] MEINHARDT-Llopis E, PÉREZ J S, and KONDERMANN D. Horn-schunck optical flow with a multi-scale strategy[J]. *Image Processing on Line*, 2013, 20: 151–172. doi: 10.5201/ipol.2013.20.

[11] 张轩阁, 田彦涛, 颜飞, 等. 基于全局光流特征的微表情识别[J]. 模式识别与人工智能, 2016, 29(8): 760–768. doi: 10.16451/j.cnki.issn1003-6059.201608011.
ZHANG Xuange, TIAN Yantao, YAN Fei, *et al*. Micro-expression recognition based on global optical flow feature[J]. *Pattern Recognition and Artificial Intelligence*. 2016, 29(8): 760–768. doi: 10.16451/j.cnki.issn1003-6059.201608011.

[12] YACOOB Y and DAVIS L S. Recognizing human facial expressions from long image sequences using optical flow[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1996, 18(6): 636–642. doi: 10.1109/34.506414.

[13] LUCEY P, COHN J F, KANADE T, *et al*. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression[C]. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), California, USA, 2010: 94–101. doi: 10.1.1.182.3759.

[14] PANTIC M, VALSTAR M, RADEMAKER R, *et al*. Web-based database for facial expression analysis[C]. IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 2005: 317–321. doi: 10.1109/ICME.2005.1521424.

[15] 邱玉, 赵杰煜, 汪燕芳. 结合运动时序性的人脸表情识别方法[J]. 电子学报, 2016, 44(6): 1307–1313. doi: 10.3969/j.issn.0372-2112.2016.06.007.
QIU Yu, ZHAO Jieyu, and WANG Yanfang. Facial expression recognition using temporal relations among facial movements[J]. *Acta Electronica Sinica*, 2016, 44(6): 1307–1313. doi: 10.3969/j.issn.0372-2112.2016.06.007.

[16] FAN X and TJAHJADI T. A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences[J]. *Pattern Recognition*, 2015, 48(11): 3407–3416. doi: 10.1016/j.patcog.2015.04.025.

[17] LONG F and BARTLETT M S. Video-based facial expression recognition using learned spatiotemporal pyramid sparse coding features[J]. *Neurocomputing*, 2016, 173: 2049–2054. doi: 10.1016/j.neucom.2015.09.049

[18] GUPTA O, RAVIV D, and RASKAR R. Multi-velocity neural networks for facial expression recognition in videos[J]. *IEEE Transactions on Affective Computing*, 1949, 99: 1.

[19] FANG H, MAC Parthaláin N, AUBREY A J, *et al*. Facial expression recognition in dynamic sequences: An integrated approach[J]. *Pattern Recognition*, 2014, 47(3): 1271–1281. doi: 10.1016/j.patcog.2013.09.023.

[20] WANG Z, WANG S, and Ji Q. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA. 2013: 3422–3429. doi: 10.1109/CVPR.2013.439.

[21] JUNG H, LEE S, YIM J, *et al*. Joint fine-tuning in deep neural networks for facial expression recognition[C]. Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2983–2991. doi: 10.1109/ICCV.2015.341.

王晓华： 女，1976 年生，副教授，研究方向为数字图像处理、情感计算、计算机视觉.

夏 晨： 男，1992 年生，硕士生，研究方向为计算机视觉、模式识别.

胡 敏： 女，1967 年生，教授，研究方向为数字图像处理、计算机视觉、模式识别.

任福继： 男，1959 年生，教授，研究方向为信号与信息处理、情感计算、计算 机视觉、模式识别.