基于运动关注深度迁移网络的跨库微表情识别*

夏万闯　郑文明　宗源　江星洵
东南大学
{xiawanchuag,wenming_zheng,xhzongyuan,jiangxingxun}@seu.edu.cn

摘要

微表情持续时间短、强度微弱、局部发生，这使得人较难准确的识别微表情。本文针对跨库面部微表情识别这一微表情识别领域的重要问题，结合微表情的自身特点和域自适应方法，围绕微表情的显著特征表示和领域间分布差异，提出了基于光流的运动关注深度迁移网络 (MADTN)。MADTN 首先用光流信息编码微表情时空变化信息，用来表征面部微表情的肌肉运动。然后将光流运动信息与面部外观信息结合，使模型能够准确地关注面部运动区域。同时，为增强微表情帧间强度、排除冗余信息，MADTN 仅采用三帧（起始帧、关键帧、终止帧）来构建运动关注微表情特征表示图。我们在两个标准跨库微表情识别任务上对 MADTN 进行了评估。实验结果表明，MADTN 的整体准确率，较之目前最好的方法，至少提高了 2.9% 的准确度。这也证明了，MADTN 的泛化性远好于其他方法。

1. 引言

微表情是人类无意识产生的面部微动作，其反应了人类的真实情感状态 [6]。微表情在医疗诊断 [32]、商业谈判 [33]、刑事审讯 [31]等方面具有广泛的应用。因此，自动微表情识别成为了一个备受关注的课题。由于微表情持续时间短、动作微弱等自身特点 [31]，人们裸眼识别微表情十分困难 [9]。因此，借助专业知识，设计自动识别微表情是十分必要的。研究人员也针对微表情的特点提出了一系列识别方法。

微表情的特性主要由运动信息展现，光流 (OF) 算法可以有效地描绘运动信息。在之前的研究工作中，研究者们基于光流信息，提出了一系列方法 [23, 37, 22, 3]。MDMO[23]计算了每个 ROI(Region of Interest) 区域内光流的主要运动方向信息，包括局部统计信息和空间位置信息。FDM[37]通过在不同尺度上迭代地计算局部运动主成分方向的光流信息来表示微表情。BI-WOOF 方法 [22]用每个 ROI 的 OF 幅度和 OS 幅度对 HOOF[3]进行加权。这些方法都是利用基于 ROI 层级的光流信息，所以，ROI 区域的合理选择对方法性能起着至关重要的作用。当 ROI 区域划分不准确时，会存在只考虑多个主要运动方向之一信息的情况，这会降低模型的性能。另外，这些方法仅考虑运动信息而忽视了面部的人脸外观信息这一重要信息。近年来，卷积神经网络（CNN）在计算机视觉的各个领域中展现出它强大的能力。研究者们基于 CNN 提出许多自动微表情识别方法 [19, 30, 18]。Kim 等人 [19]利用 CNN 和 LSTM 编码微表情序列，设计了一个浅层网络，有效防止了微表情的过拟合问题。Peng 等人 [30]利用微表情视频序列和光流信息去训练了一个 3D-CNN。ELRCN[18]方法在通道方向堆叠微表情的视频帧和相邻帧间的光流信息。此外，目前大部分的微表情识别研究，是假定训练集和测试集的特征满足独立同分布的。然而，真实世界中微表情样本，往往由不同设备在不同光照、角度和背景下拍摄而得，这使得训练集与测试集的特征不满足独立同分布条件，也即

*本文为 ICPR'20 Workshop FBE 论文 [36]的中文翻译版。

真实世界中源域数据和目标域数据遵循不同的分布。在这种情况下，将源域数据上训练的模型直接用于目标域数据，所体现的性能较差。与此同时，标注足够的数据对于任何任务来说通常是昂贵的并且是耗时的。领域自适应方法可以缓解这个问题。它利用有标注信息的源域，并迁移源域的知识到有少量标注或没有标注的目标域。跨库微表情识别的训练集和测试集来自不同的数据库，它们之间存在着较大的分布差异，是一种典型的领域自适应问题。之前的领域自适应方法，通过在浅层学习域不变的特征的方式，来连接源域和目标域 [12, 16, 29]。最近的深度网络域自适应方法则是利用深度网络来学习可迁移的特征表示，其思路是在深度架构中嵌入域自适应模块，去匹配跨领域的特征分布 [10, 11, 24, 34, 26, 27, 35]。由于跨库微表情识别可以很好地模拟真实世界中微表情识别遇到的问题，它逐渐成为微表情识别中的重要话题。

跨库微表情识别，不仅涉及到域自适应问题，也与微表情的显著表示有关。本文就是从这两个角度出发，提出了运动关注深度迁移网络（MADTN），来提高跨库微表情识别的的准确率和泛化性。MADTN 将运动信息和人脸信息相结合，使其能够准确地反应肌肉运动出现的人脸区域。MADTN 利用光流信息生成注意力图，在像素水平上对面部信息进行加权，使模型能重点关注人脸运动区域，产生更具区分度的微表情特征表示。为缓解相邻帧间光流微弱、噪声强度与微表情强度相近，无法提取显著描绘微表情运动变化光流的困难，MADTN 挑选运动范围大、代表性强的三帧（起始帧、峰值帧和终止帧）来计算光流。MADTN 计算了微表情样本的起始帧、峰值帧之间的光流信息，以及峰值帧、终止帧之间的光流信息，并用其对人脸外观加权。深度学习领域的注意力机制 [17]通过学习到的注意力图对不同特征加权。但这种学习到的注意力图的权值具有不确定性，而光流生成的注意力图能准确地反应微表情的运动区域。而 MADTN 在像素层级上，通过光流信息加权面部信息，产生判别性更强的特征表示。另一方面，受卷积神经网络的启发，MADTN 设计了一个简单有效的卷积神经网络，用于特征的提取。它由 5 层卷积神经网络构成，参数量小但十分有效。此外，MADTN 在深度网络中引入最大均值差异 (Maximum Mean Discrepancy，MMD)[27]，用于减小域间的特征分布差异。我们遵循 Zong 等人 [41]设计的实验范式，进行了大量跨库微表情识别实验，评估了 MADTN 的模型性能。实验结果表明了本章提出的 MADTN 方法的有效性和泛化性强。

2. 运动关注特征表示

2.1. 结构概述

本文所提出的运动关注深度迁移网络 (MADTN) 模型如图1所示。MADTN 提出了一种整合光流和人脸外观信息的预处理方式，称为运动关注特征表示，它有助于高判别性微表情特征的产生，如图1(a) 和图1(b) 所示。图1(a) 是光流运动信息和人脸外观信息；图1(b) 是运动关注微表情特征表示图，由图1(a) 中的三幅灰度图组成。MADTN 设计了一个简单有效的浅层卷积神经网络，用于提取微表情特征，如图1(c) 所示，并融入了 MMD，用于跨库微表情识别。下面分别对 MADTN 模型的两个主要模块进行详细介绍。

2.2. 运动关注特征表示

光流算法可以有效地编码人脸微表情视频的运动变化。光流估计算法 [7]通过跟踪大量点在视频序列中的位移来推断物体的运动。我们用 $I(x, y, t)$ 表示 t 时刻某像素在 (x, y) 处的强度，假设在很短的时间 $\mathrm{d}t$ 内，它移动了 $(\mathrm{d}x, \mathrm{d}y)$ 的距离。由亮度恒定不变原理可知，该像素在运动前后的光强度是不变的，即：

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \tag{1}$$

将式1右端进行泰勒展开，得：

$$I(x, y, t) = I(x, y, t) + \frac{\partial I}{\partial x}\mathrm{d}x + \frac{\partial I}{\partial y}\mathrm{d}y + \frac{\partial I}{\partial t}\mathrm{d}t + \varepsilon \tag{2}$$

其中，ε 代表高阶无穷小项，可以忽略不计。进一步可有：

$$\frac{\partial I}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial I}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial I}{\partial t}\frac{\partial t}{\partial t} = 0 \tag{3}$$

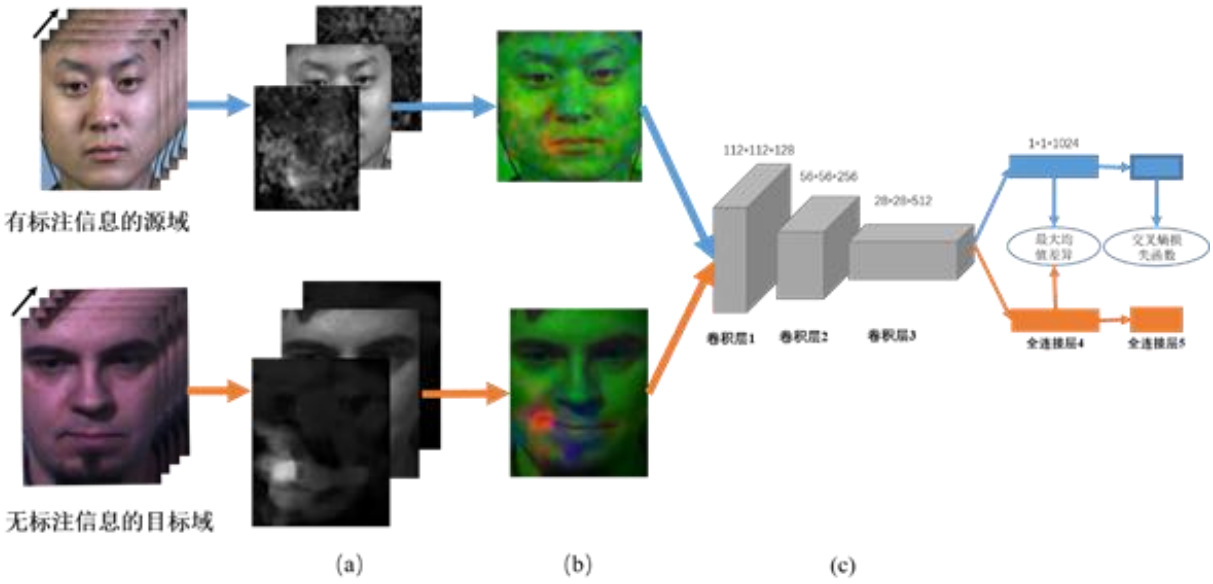


图 1. 运动关注深度迁移网络 (MADTN) 模型

设 $V_x = \frac{\mathrm{d}x}{\mathrm{d}t}$ 和 $V_y = \frac{\mathrm{d}y}{\mathrm{d}t}$ 分别为光流沿 X 轴和 Y 轴的速度矢量，得 $V = [V_x, V_y]^{\mathrm{T}}$。令 $I_x = \frac{\partial I}{\partial x}, I_y = \frac{\partial I}{\partial y}, I_t = \frac{\partial I}{\partial t}$ 分别表示图像中像素点的强度沿 X，Y，T 方向的偏导数。式3可以改写成

$$I_x V_x + I_y V_y + I_t V_t = 0 \tag{4}$$

其中，I_x、I_y、I_t 可由图像数据求得，而 $V = [V_x, V_y]$ 为所求光流矢量。

由于微表情视频的帧率一般较高，两帧之间的运动变化太微弱，故选取微表情视频的起始帧、关键帧、终止帧三帧计算它们之间的光流。我们计算起始帧到关键帧的光流，用它代表微表情肌肉放松到收缩状态之间的变化；计算关键帧到终止帧的光流，用它代表肌肉收缩到放松状态之间的变化。最后将起始帧和关键帧间的光流、关键帧的人脸灰度图、关键帧到终止帧间的光流合成运动关注微表情特征表示图，处理过程图如图2所示。其中，图2(a) 分别是起始帧、关键帧、终止帧；图2(b) 分别是起始帧和关键帧间的光流图、关键帧的灰度图、关键帧和终止帧间的光流图；图2(c) 是由图2(b) 合成的运动关注微表情特征表示图，其由图2(b) 的三幅灰度图充当此图像的每一维而组成。从图2(c) 看，可以明显发现此人的嘴角发生了明显的动作。经过这样的处理后，可以得到更具区分度的特征表示图。并且该方法仅选用三帧，有效地减少了时序上的冗余信息。

下面介绍获取两帧图片对应光流灰度图的具体过程。计算两帧图像之间光流时，会分别得到 X 方向和 Y 方向的位移分量图 V_x 和 V_y。我们进一步将笛卡尔坐标下的 $V = (V_x, V_y)$ 转变为极坐标 $V = (r, \theta)$，r 和 θ 分别表示光流的幅值和方向。在 HSV 颜色空间模型下，色调 H 的范围是 $[0°, 360°]$，饱和度 S 和明度 V 的范围是 $[0, 1]$。当合成图像时，把 V 的值设为 1，然后把 r 和 θ 分别赋值给 S 和 H。最后，将 HSV 颜色空间下的彩色图转变为一张灰度图，得到两帧图像的光流灰度图。

我们将起始帧与峰值帧的光流灰度图、峰值帧灰度图、峰值帧与结束帧光流灰度图这三幅灰度图，分别赋值给 RGB 图像的 R、G、B 维度，如图3(c) 所示，合成微表情运动关注特征。通过运动关注模块的加权，不同类别的微表情会加权不同的聚焦区域。我们从 SMIC-HS 数据库中挑选一个被试的不同微表情的样本，展示其对应的运动关注特征图，如

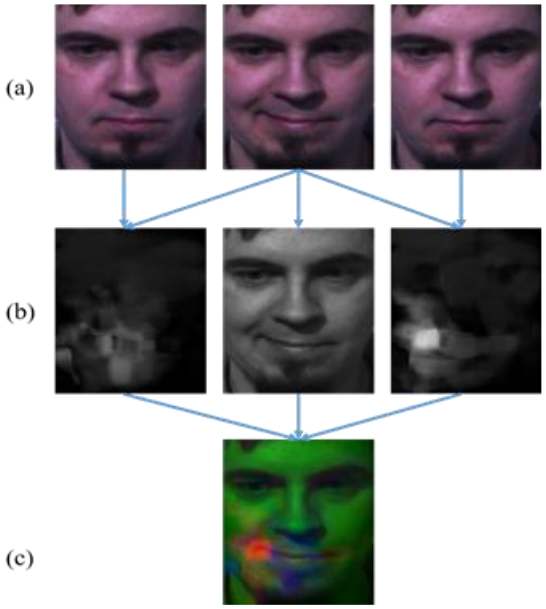


图 2. 运动关注特征合成过程图

图3所示。从图3可以很明显地看出，不同类别的微表情特征存在较大的差异，这就有利于模型对它们进行分类。

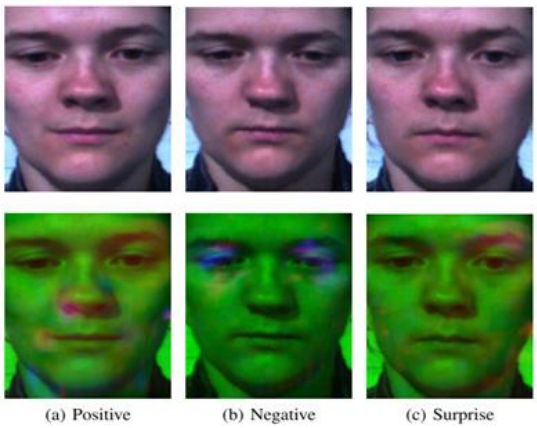


图 3. SMIC-HS 数据库不同类别的关键帧和运动关注特征表示

2.3. 深度迁移网络

迁移学习中将有标注信息的训练数据库称为源域 $S = (x_i^s, y_i^s)_{i=1}^{N_s}$，训练时没有标注信息的测试数据库称为目标域 $T = (x_i^t)_{i=1}^{N_t}$。其中，x_i^s 是第 i 个源域样本，y_i^s 是第 i 个源域样本的标注信息，N_s 是源域的样本数目，x_i^t 是第 i 个目标域样本，N_t 是目标域的样本数量。由于源域和目标域样本获取的条件不同，源域数据库的分布 $P_s(x)$ 和目标域数据库的分布 $P_t(x)$ 就不相同。跨域微表情识别，是利用源域的监督数据和目标域的无监督数据，来减少源域和目标域之间的差异，使得源域上的知识能够最大程度的迁移到目标域上，以期提高目标域的识别率。

我们设计了一个简单但有效的浅层卷积神经网络，用来提取预处理之后的样本特征。其网络结构图如图1(c) 所示。它是五层的卷积神经网络，分为三层卷积层和全连接层。每一层卷积后都跟着最大池化层进行降维，并且每层后都使用 Leaky Rectified Linear Unit(LReLU)[28]增加模型的非线性，以改进模型的拟合能力。文献 [39]表明，随着网络层数的加深，将直接迁移学到的知识运用到目标域的效果不尽人意。所以，我们将 MMD 嵌入 CNN 结构中，拉近源域和目标域的差异，更好地利用有标注的源域信息和未标注的目标域信息，使学到的特征在两个领域内都具有普遍性和高区分性。

最大均值差异（MMD）[1]被广泛用来计算两个域分布之间的距离，是迁移学习中常用的损失函数。MMD 及其变种在再生核希尔伯特空间（Reproducing Kernel Hilbert Space，RKHS）定义如下：

$$MMD[\mathcal{H}, P_s, P_t] = \left\| \frac{1}{N_s} \sum_{i=1}^{N_s} \kappa(f_i^s) - \frac{1}{N_t} \sum_{i=1}^{N_t} \kappa(f_i^t) \right\|_{\mathcal{H}} \tag{5}$$

其中，$\mathcal{H}$ 表示再生核希尔伯特空间，$\kappa(\cdot)$ 代表映射函数，N_s 和 N_t 分别代表源域和目标域的数目。f_i^s 和 f_j^t 分别代表第 i 个源域样本 x_i^s 和第 j 个目标域样本 x_j^t 对应特征。如果它们的分布相似的话，$MMD[\mathcal{H}, P_s, P_t]$ 会趋于零 [14]。根据上述假设，

Gretton 等人 [14]提出 $MMD^2[\mathcal{H}, S, T]$ 的无偏估计：

$$\begin{aligned}\mathrm{MMD}^2\left[\mathcal{H}, S, T\right] = & \frac{1}{N_s(N_s-1)}\sum_{i\neq j}^{N_s}\kappa(x_i^s, x_j^s) \\ & + \frac{1}{N_t(N_t-1)}\sum_{i\neq j}^{N_t}\kappa(f_i^t, f_j^t) \\ & - \frac{2}{N_sN_t}\sum_{i=1}^{N_s}\sum_{j=1}^{N_t}\kappa(f_i^s, f_j^t)\end{aligned} \quad (6)$$

此处映射函数 κ 选择高斯核，即 $\kappa(f^s, f^t) = e^{-\frac{\|f^s - f^t\|}{2\delta^2}}$，$\delta$ 是高斯函数的参数。嵌入 MMD 损失函数后，MADTN 的目标函数如下：

$$L = L_s(\phi((f^s), y^s)) + \lambda MMD^2[\mathcal{H}, S, T] \quad (7)$$

其中，$L_s(\phi(f^s), y^s)$ 代表源域的交叉熵函数，λ 是衡量两个损失函数的超参数。通过联合训练这两个损失函数，不仅能学习到有区分度的微表情特征表示，而且也能减少源域和目标域之间的差异。

3. 实验

现有的跨库微表情识别的实验协议各不相同。为统一协议，Zong 等人 [41]设计了统一的评估方式，来评估跨库微表情识别这一任务。所以为了公平的对比，本文遵循其提出的实验协议。我们在 CASME II 和 SMIC 数据库中进行两类实验，评估所提出的运动关注深度迁移网络 MADTN 在跨库微表情识别任务上的性能。

3.1. 数据库及协议

SMIC. SMIC 数据库的包含 SMIC-HS、SMIC-VIS、SMIC-NIR 三个子集。SMIC-HS 收集自每秒 100 帧的高速相机；SMIC-VIS 收集自每秒 25 帧的正常相机；SMIC-NIR 收集自每秒 25 帧的红外相机。SMIC-HS 包含了来自 16 个被试者共 164 个微表情样本；SMIC-VIS 和 SMIC-NIR 数据库包含了 16 个被试中后 8 个被试共 71 个样本。SMIC 数据库包含三类微表明：正性 (Positive)、负性 (Negative)、惊讶 (Surprise)。

CASME II. CASME II 包含了来自 26 个被试者的 257 个微表情样本。CASME II 的微表情样本分为 7 类：开心 (Happy)、厌恶 (Disgust)、压抑 (Repression)、伤心 (Sad)、害怕 (Fear)、惊奇 (Surprise) 和其他 (Others)。CASME II 数据库样本的帧率为 200 帧每秒。

协议. 本文进行的跨库微表情识别实验，需要源数据库和目标数据库的标签信息一致。但是 CASME II 数据库和 SMIC 数据库的标签不一致，所以需要对原数据库进行一定的处理。SMIC 数据库有三个子库 SMIC-HS、SMIC-VIS、SMIC-NIR，所有样本分为三类：正性（Positive）、负性（Negative）、惊讶（Surprise）。CASME II 数据库分为七类：高兴（Happy）、厌恶（Disgust）、压抑（Repression）、伤心（Sad）、害怕（Fear）、惊讶（Surprise）、其他（Others）。在跨库微表情识别实验中，不同的数据库应有统一的标签。所以，根据 SMIC 标签的定义，把 CAMSE II 中的标签重新标注：高兴被重新标注为正性，厌恶、伤心、害怕被标注为负性，惊讶不变。统一标签的微表情数据库情况如表1所示。Zong 等人 [41]建立了两种类型的标准实验：TYPE-I 和 TYPE-II，实验协议如表2所示。其中，H 代表 SMIC-HS、V 代表 SMIC-VIS、N 代表 SMIC-NIR、C 代表 CASME II，H→V 表示 SMIC-HS 是源数据库，V 是 SMIC-VIS 是目标数据库，其他类似。TYPE-I 是 SMIC 内部各个子集之间的跨库实验，这个任务中的差异主要是不同摄像头导致的成像差异。TYPE-II 是 CAMSE II 和 SMIC 各个子库之间的跨库协议，这个任务中的差异是样本收集环境不一样，故它比 TPYE-I 存在着更大的差异。

表 1. 标签统一的微表情数据库样本分布情况

微表情数据库	正性	负性	惊讶
CASME II	32	73	25
SMIC-HS	51	70	43
SMIC-VIS	23	28	20
SMIC-NIR	23	28	20

表 2. 跨库微表情识别实验任务

类型	实验	源数据库	目标数据库
TYPE-I	H→V	SMIC-HS	SMIC-VIS
	V→H	SMIC-VIS	SMIC-HS
	H→N	SMIC-HS	SMIC-NIR
	N→H	SMIC-NIR	SMIC-HS
	V→N	SMIC-VIR	SMIC-NIR
	N→V	SMIC-NIR	SMIC-VIS
TYPE-II	C→H	CASME II	SMIC-HS
	H→C	SMIC-HS	CASME II
	C→V	CASME II	SMIC-VIS
	V→C	SMIC-VIS	CASME II
	C→N	CASME II	SMIC-NIR
	N→C	SMIC-NIR	CASME II

3.2. 实现细节

在样本送入微表情运动加权模块之前，需要先获取对应的人脸区域。首先，我们将视频离散成图像序列，然后通过 MTCNN[40]检测微表情起始帧的人脸。为了尽可能地保存微表情相关信息，只保留人脸区域大小如图4所示，其上至额头上部，下至下巴，左右至耳朵根部。由于光流估计算法计算两帧之间的运动变化信息，需要保证两帧内的人脸位置不能被人为改变。故根据起始帧的人脸区域去裁剪微表情序列后面的所有帧。微表情的持续时间非常短暂，不到 0.5 秒，人脸位置不会产生太大的变化，不会离开第一帧的人脸框，能够完整保存之后帧的人脸信息。

训练模型时，为了避免过拟合问题，在训练时，从整个视频序列中选取最中间的 20% 的帧作为关键帧扩充样本，而测试时，仅选择最中间的帧作为关键帧。由于微表情整个变化过程中的肌肉运动变化程度是不同的，如此处理可以提高样本的多样性。输入图像在 $[-30°, 30°]$ 随机旋转做数据增强，并将其分辨率变化为 112×112。模型中，我们设置高斯核化参数的取值范围为 $\delta = [1, 2, 4, 8, 16]$。我们通过 Adam[20]优化算法训练模型，学习率为 2×10^{-4}，并设置 λ 的值为 2。实验结果用 mean F1-score 和准确率 (Accuracy) 来评估。mean F1-score 是每一类 F1-score 值的平均值（不考虑每一类的样本数），特别适用于样本不平衡问题。

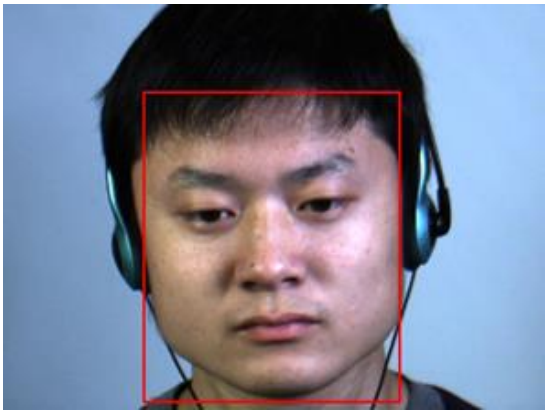


图 4. 人脸裁切示意图

4. 结果

4.1. 实验结果

表3和表4分别展现了本文所提的 MADTN 在 TYPE-I 和 TPYE-II 上的的实验结果。为了公平比较，对比方法的结果直接来自文献 [41]。每个实验中最好的结果用加粗字体表示。从表3和表4的结果可以看出，在 mean F1-score 和 Accuracy 这两个指标上，本文提出的方法 MADTN 在整体上达到当前最优水平，并且在大部分的跨库实验上，性能都优于其他方法。特别地，在 TYPE-II 任务上，MADTN 整体准确率较之当前最好方法 RSTR 高出 14.4%。在各个跨库实验上，MADTN 也远优于其他方法，在 C→N 跨库实验上的准确率甚至能提高 22%。这有效地说明了 MADTN 的优越性。从表3和表4的结果可以看出，只有在 TYPE-I 任务的 H→V 跨库实验上，MADTN 没有超过其他方法。此外，H→V 跨库实验的结果比 V→H 跨库实验结果高的原因，可能是由数据集本身的特点决定的。由于 SMIC-VIS 由常规速度的摄像头捕获而得，SMIC-HS 由高速摄像头捕获而得，且 SMIC-VIS 只包含了 SMIC-HS 的 16 个被试中后 8 个被试的样本，所以，SMIC-VIS

可视为 SMIC-HS 的低时间分辨率子集。我们可以认为，这相当于在 H→V 跨库实验中，测试集中的信息被包含在训练集中，而 V→H 中，测试集中存在许多不在训练集上的信息。

其他方法在两类（TYPE-I 和 TYPE-II）任务的结果上，可以看到 TYPE-II 的实验结果明显低于 TYPE-I 的，这表明 TYPE-II 跨库任务中数据库之间的差异是大于 TYPE-I 中的。从数据库的文献[21]和 [38]中，SMIC 子库之间的差异主要是记录样本的拍摄设备不同，以及 SMIC-HS 拥有更多的样本。而 CASME II 和 SMIC 的采集设备和环境、被试的人种、诱发材料等各个方面都是不同的。这就导致它们之间存在较大的分布差异，而且大多是与微表情无关的差异，但也是不可避免的。正是存在这种差异，导致识别准确率下降，泛化性不够。也正是本文所提的 MADTN 尝试去解决的问题。从实验结果可以看出来，其他方法都出现了大幅度的下降，而本文的方法 MADTN 仅出现了小幅度的下降。这说明本文所提出的方法 MADTN 具有强泛化性。

从实验中，我们可以知道：构建具有显著差异的特征表示是微表情识别的关键。增加类间差异、增强与微表情相关的信息，都能降低域分布差异对识别效果的影响，有利于跨领域微表情识别。微表情识别任务首先从微表情本身的特点出发，提高特征的区分度，与任务相关的特征在特征分布中占主导地位，自然就会抑制无关噪声，不同数据库的域分布差异也会尽可能小。

实验结果的混淆矩阵如图5所示，可以看到在大部分跨库实验（H→V, V→N, N→V, H→C, C→N, N→C）中，MADTN 模型识别每类微表情情感的准确率都相差不大，这说明 MADTN 在一定程度上能抑制类别不平衡对实验结果的影响。

4.2. 消融实验

为了更深入地探究 MADTN 模型，我们在运动关注特征表示的结构上进行消融实验，来研究 MADTN 不同的结构对实验结果的影响。在 C→H, C→V, H→N 跨库实验上评估运动关注特征表示图的不同组成成分的识别效果。我们在 MADTN 模型

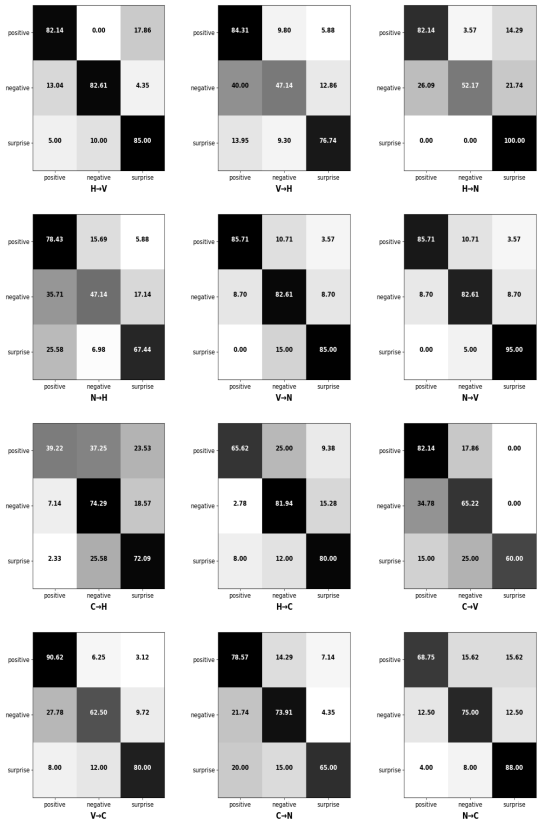


图 5. MADTN 模型在跨库实验上的混淆矩阵

的运动关注图的处理上进行消融实验，并在 C→H, C→V, H→N 三个跨库实验上进行了评估。本次实验设置与之前的设置保持相同，为了保证网络结构不变，如果某一组成成分不存在，将对应成分置零。运动关注特征表示图由起始帧与关键帧间的光流图、关键帧的人脸灰度图和关键帧与终止帧间的光流图组成，如图2(b) 所示。不同组成成分的实验结果如表5所示。其中，Face 代表关键帧的人脸灰度图，onOF 代表起始帧和关键帧之间的光流，endOF 代表关键帧和终止帧之间的光流。onOF-Face 代表输入的成分包含 onOF 和 Face，其他类似。对比 Face 和 onOF-Face 实验，可以发现加入光流信息 onOF 后，实验结果有了明显提升，这表明了光流信息能有效地反应微表情的运动变化信息，光流信息和人脸信息结合能有效增加微表情样本的区分度，提升模型的性能。对比 onOF-Face 和 onOF-Face-endOF 实验结果，可以发现，加入关键帧和终止帧的光流信

表 3. 第一类 TYPE-I 跨库微表情识别实验结果 (mean F1-score/Accuracy)

方法	H→V	V→H	H→N	N→H	V→N	N→V	平均
SVM[2]	0.8002/80.28	0.5421/54.27	0.5455/53.52	0.4878/54.88	0.6186/63.38	0.6078/63.38	0.6003/61.62
IW-SVM[15]	0.8868/88.73	0.5852/58.54	0.7469/74.65	0.5427/54.27	0.6620/69.01	0.7228/73.24	0.6911/68.07
TCA[29]	0.8269/83.10	0.5477/54.88	0.5828/59.15	0.5443/57.32	0.5810/61.97	0.6598/67.61	0.6238/64.01
GFK[13]	0.8448/84.51	0.5957/59.15	0.6977/70.42	0.6197/62.80	0.7619/76.06	0.8142/81.69	0.7223/72.44
SA[8]	0.8037/80.28	0.5955/59.15	0.7465/74.65	0.5644/56.10	0.7004/71.83	0.7394/74.65	0.6917/69.44
STM[4, 5]	0.8253/83.10	0.5059/51.22	0.6628/66.20	0.5351/56.10	0.6427/67.61	0.6922/70.42	0.6440/65.78
TKL[25]	0.7742/77.46	0.5738/57.32	0.7051/70.42	0.6116/62.60	0.7558/76.06	0.7579/76.06	0.6964/69.92
TSRG[42]	0.8869/88.73	0.5652/56.71	0.6484/64.79	0.5770/57.93	0.7056/70.42	0.8116/81.69	0.6991/70.05
DRFS-T[43]	0.8643/85.92	0.5767/57.32	0.7179/71.83	0.6163/61.59	0.7286/73.24	0.7732/77.46	0.7128/71.23
DRLS[43]	0.8604/85.92	0.6120/60.98	0.6599/66.20	0.5599/55.49	0.6620/69.01	0.5771/61.97	0.6552/66.60
RSTR[41]	0.8721/87.32	0.6401/64.02	0.7466/74.65	0.5765/57.32	0.7506/76.06	0.8428/84.51	0.7381/73.98
MADTN	0.8302/83.11	0.6704/66.46	0.7641/77.44	0.6252/62.21	0.8435/84.52	0.8732/87.30	0.7678/76.84

表 4. 第二类 TYPE-II 跨库微表情识别实验结果 (mean F1-score/Accuracy)

方法	C→H	H→C	C→V	V→C	C→N	N→C	平均
SVM[2]	0.3697/45.12	0.3245/48.46	0.4701/50.70	0.5367/53.08	0.5295/52.11	0.2368/23.85	0.4112/45.55
IW-SVM[15]	0.3541/41.46	0.5829/62.31	0.5778/59.15	0.5537/54.62	0.5117/50.70	0.3456/36.15	0.4876/50.73
TCA[29]	0.4637/46.34	0.4870/53.08	0.6834/69.01	0.5789/59.23	0.4992/50.70	0.3937/42.31	0.5177/53.45
GFK[13]	0.4126/46.95	0.4776/50.77	0.6361/66.20	0.6056/61.50	0.5180/53.52	0.4469/46.92	0.5161/54.31
SA[8]	0.4302/47.56	0.5447/62.31	0.5939/59.15	0.5243/51.54	0.4738/47.89	0.3592/36.92	0.4877/50.90
STM[4, 5]	0.3604/43.90	0.6115/63.85	0.4015/52.11	0.2715/30.00	0.3523/42.25	0.3850/41.54	0.3982/45.61
TKL[25]	0.3829/44.51	0.4661/54.62	0.6042/60.56	0.5378/53.08	0.5392/54.93	0.4248/43.85	0.4925/51.93
TSRG[42]	0.5042/51.83	0.5171/60.77	0.5935/59.15	0.6208/63.08	0.5624/56.34	0.4105/46.15	0.5348/56.22
DRFS-T[43]	0.4524/46.95	0.5460/60.00	0.6217/63.38	0.6762/68.46	0.5369/56.34	0.4653/50.77	0.5498/57.65
DRLS[43]	0.4924/53.05	0.5267/59.23	0.5757/57.75	0.5942/60.00	0.4885/49.83	0.3838/42.37	0.5102/53.71
RSTR[41]	0.5297/54.27	0.5622/60.77	0.5882/59.15	0.7021/70.77	0.5009/50.70	0.4693/50.77	0.5587/57.74
MADTN	0.6100/62.79	0.7486/77.54	0.7056/70.41	0.7304/72.85	0.7305/73.24	0.7403/75.98	0.7109/72.14

息，也能大幅提高识别性能，这反应从关键帧到终止帧之间的微表情的肌肉放松过程也能有效地反应微表情，运动的变化信息可能在某些微表情样本的后半段更加显著。通过对比 onOF-endOF 和 onOF-Face-endOF 的实验结果，人脸的外观信息也能有效帮助提高识别性能，这是因为不同实验样本可能由于被试人不同，人脸外观差异巨大，导致微表情运动发生的区域和强度存在差异，联合人脸外观信息和光流运动信息，可以清晰地表明微表情运动发生的人脸区域。通过这种方式，卷积神经网络学习到相关信息，降低此差异对结果的影响。对比 Face 和 onOF-endOF 的结果，可以发现，光流运动信息比人脸外观信息更具区分度，大幅超过了只有人脸信息成分的识别结果。这说明了微表情是一个肌肉运动的微动作，在微表情识别中，时序信息比空间信息更重要。通过对比不同成分的实验，可以发现这些信息都对提高微表情识别准确率有所帮助。时间、空间特征的结合有助于生成具有区分度的特征，也表明了本文提出的运动关注特征表示的有效性。

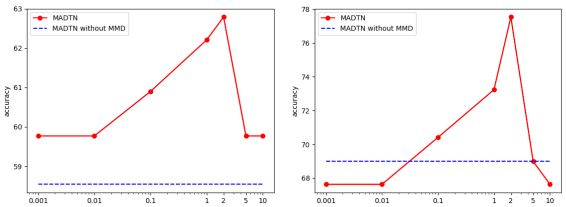
4.3. 超参讨论实验

我们进一步讨论了最大均值差异 MMD 对实验结果的影响。我们挑选 C→H、C→V 跨库实验，对 MMD 设置不同的权重因子 λ，以准确度为评价指标进行实验。实验结果如图6所示。蓝虚线表明 λ 等于 0 的实验结果，红线为不同 λ 值的实验结果。从图

表 5. MADTN 的不同输入成分的实验结果 (mean F1-score/Accuracy)

输入成分	C→H	C→V	H→N
Face	0.4115/42.07	0.4715/52.10	0.4673/52.10
onOF-Face	0.5598/56.69	0.6207/61.96	0.5713/57.76
onOF-endOF	0.5858/59.13	0.6520/64.79	0.7352/73.24
onOF-Face-endOF(MADTN)	0.6100/62.79	0.7056/70.41	0.7641/77.44

中，可以看到 MMD 能有效地缓解源域和目标域之间的分布差异，提高识别性能，但从图6(b) 红蓝线，可以看到不合适的 λ 值，不能有效地平衡分类损失函数和 MMD 损失函数，也会降低识别性能。如果不能找到两类损失函数的平衡点，那会大大降低模型在实验上的性能。如果偏向分类函数，模型虽然能在源域很好地识别，但泛化性不足，无法准确识别目标域数据；如果偏向 MMD 损失函数，虽然领域间特征相似，但特征无法有效表示微表情，因而产生迁移的负面效果，影响识别性能。所以，选择合适的平衡因子是十分重要的，它能充分利用它们的优势，降低负面影响，提高模型识别性能。



(a) C→H (b) C→V

图 6. 不同权重因子 λ 对 MADTN 模型影响的实验结果

5. 总结与展望

本文提出了运动关注深度迁移网络（MADTN）来解决跨库微表情识别这个任务。本方法只选取三帧来表示微表情，能有效地减少冗余信息；光流运动信息和人脸外观信息地结合，能有效地表明微表情运动发生的区域，使网络能重点关注该区域；可视化结果显示其是有区分度的特征表示。深度迁移网络能有效的缓解源域和目标域之间的分布差异。在两个标准跨库微表情识别任务上，MADTN 取得了显著的结果，超过了之前的方法，验证了本文提出的方法的优越性和强泛化性。

致 谢. 本项目受到了国家重点研发计划 (2018YFB1305200)，国家自然科学基金 (61921004, 61902064, 81971282, U2003207, 62076064) 和中央高校基本科研业务费专项资金 (2242018K3DN01) 的资助。

参考文献

[1] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics, 22(14):e49–e57, 2006. 4

[2] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3):1–27, 2011. 8

[3] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 1932–1939. IEEE, 2009. 1

[4] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3515–3522, 2013. 8

[5] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial expression analysis. IEEE transactions on pattern analysis and machine intelligence, 39(3):529–545, 2016. 8

[6] P. Ekman and W. V. Friesen. Nonverbal leakage and clues to deception. Psychiatry, 32(1):88–106, 1969. 1

[7] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In Scandinavian conference on Image analysis, pages 363–370. Springer, 2003. 2

[8] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In Proceedings of the IEEE international conference on computer vision, pages 2960–2967, 2013. 8

[9] M. Frank, M. Herbasz, K. Sinuk, A. Keller, and C. Nolan. I see how you feel: Training laypeople and professionals to recognize fleeting emotions. In The Annual Meeting of the International Communication Association. Sheraton New York, New York City, 2009. 1

[10] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In International conference on machine learning, pages 1180–1189. PMLR, 2015. 2

[11] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. The journal of machine learning research, 17(1):2096–2030, 2016. 2

[12] B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In International Conference on Machine Learning, pages 222–230. PMLR, 2013. 2

[13] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 2066–2073. IEEE, 2012. 8

[14] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. The Journal of Machine Learning Research, 13(1):723–773, 2012. 4, 5

[15] A. Hassan, R. Damper, and M. Niranjan. On acoustic emotion recognition: compensating for covariate shift. IEEE Transactions on Audio, Speech, and Language Processing, 21(7):1458–1468, 2013. 8

[16] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. Smola. Correcting sample selection bias by unlabeled data. Advances in neural information processing systems, 19:601–608, 2006. 2

[17] L. Itti and C. Koch. Computational modelling of visual attention. Nature reviews neuroscience, 2(3):194–203, 2001. 2

[18] H.-Q. Khor, J. See, R. C. W. Phan, and W. Lin. Enriched long-term recurrent convolutional network for facial micro-expression recognition. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pages 667–674. IEEE, 2018. 1

[19] D. H. Kim, W. J. Baddar, and Y. M. Ro. Micro-expression recognition with expression-state constrained spatio-temporal feature representations. In Proceedings of the 24th ACM international conference on Multimedia, pages 382–386, 2016. 1

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 6

[21] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In 2013 10th IEEE International Conference and Workshops

on Automatic Face and Gesture Recognition (FG), pages 1–6. IEEE, 2013. 7

[22] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan. Less is more: Micro-expression recognition from video using apex frame. Signal Processing: Image Communication, 62:82–92, 2018. 1

[23] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu. A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Transactions on Affective Computing, 7(4):299–310, 2015. 1

[24] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In International conference on machine learning, pages 97–105. PMLR, 2015. 2

[25] M. Long, J. Wang, J. Sun, and S. Y. Philip. Domain invariant transfer kernel learning. IEEE Transactions on Knowledge and Data Engineering, 27(6):1519–1532, 2014. 8

[26] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. arXiv preprint arXiv:1602.04433, 2016. 2

[27] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. In International conference on machine learning, pages 2208–2217. PMLR, 2017. 2

[28] A. L. Maas, A. Y. Hannun, A. Y. Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In Proc. icml, volume 30, page 3. Citeseer, 2013. 4

[29] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks, 22(2):199–210, 2010. 2, 8

[30] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu. Dual temporal scale convolutional neural network for micro-expression recognition. Frontiers in psychology, 8:1745, 2017. 1

[31] S. Porter and L. Ten Brinke. Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. Psychological science, 19(5):508–514, 2008. 1

[32] T. A. Russell, E. Chu, and M. L. Phillips. A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool. British journal of clinical psychology, 45(4):579–583, 2006. 1

[33] F. Salter, K. Grammer, and A. Rikowski. Sex differences in negotiating with powerful males. Human Nature, 16(3):306–321, 2005. 1

[34] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In Proceedings of the IEEE international conference on computer vision, pages 4068–4076, 2015. 2

[35] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7167–7176, 2017. 2

[36] W. Xia, W. Zheng, Y. Zong, and X. Jiang. Motion attention deep transfer network for cross-database micro-expression recognition. In Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III, pages 679–693. Springer, 2021. 1

[37] F. Xu, J. Zhang, and J. Z. Wang. Microexpression identification and categorization using a facial dynamics map. IEEE Transactions on Affective Computing, 8(2):254–267, 2017. 1

[38] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. PloS one, 9(1), 2014. 7

[39] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? arXiv preprint arXiv:1411.1792, 2014. 4

[40] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503, 2016. 6

[41] T. Zhang, Y. Zong, W. Zheng, C. P. Chen, X. Hong, C. Tang, Z. Cui, and G. Zhao. Cross-database micro-expression recognition: A benchmark. IEEE Transactions on Knowledge and Data Engineering, 2020. 2, 5, 6, 8

[42] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao. Learning a target sample re-generator for cross-database micro-expression recognition. In Proceedings of the 25th ACM international conference on Multimedia, pages 872–880, 2017. 8

[43] Y. Zong, W. Zheng, X. Huang, J. Shi, Z. Cui, and G. Zhao. Domain regeneration for cross-database micro-expression recognition. IEEE Transactions on Image Processing, 27(5):2484–2498, 2018. 8