指导教师：杨涛　　提交时间：2016/3/17

CVPR2015 Paper Translation

No: 1

姓名：潘妍

学号：2013302467

班号：10011301

基于 3D 模型的连续情感识别

Hui Chen[1], Jiangdong Li[2], Fengjun Zhang[3], Yang Li[1], Hongan Wang[2,3]
Beijing Key Lab of Human-computer Interaction, Institute of Software, Chinese Academy of Sciences[1]
University of Chinese Academy of Sciences[2]
State Key Lab of Computer Science, Institute of Software, Chinese Academy of Sciences[3]
Beijing, China, 100190

摘要

我们建立了一个基于 3D 模型的实时访法，能够持续地识别出空间中，人们在自然交流时面部表情所传达的情感。在我们的方法中，3D 面部模型重建于 2D 图像，这为健壮性的提高提供了重要的线索，从而能够克服包括头部过度旋转、头部快速移动和局部面部阻挡等的巨大改变。为了准确地识别出情感构建了一种新奇的基于森林的随机算法，它能够同时归一化两个 3D 面部追踪和连续情感估计的回归式。并且，通过重构的 3D 面部模型，暂时的信息和用户独立的情感表现也可以通过我们图像融合过程被记录下来。实验结果表明了我们的算法可以获得最高水平实时连续情感识别的皮尔逊相关系数。

1. 简介

连续情感分析要求得到并处理较长且未分割的自然输入，并且预测二维空间中代表的情感[15]。计算机能够理解自然交互中的情感就有能力做出更智慧的决定并提供更好的交互体验，这是已被承认的[17, 27]。通过将情感分为不同的类别，一些以人为中心的系统[12, 20]已经被设计成对于不同的用户情感类别可以做出不同的反应，这为用户提供了更好的交互体验，说明了在人机交互方面情感评估的重要性和必要性。在日常交流中，人们持续地谈话和思考，情感便自然而然地流露出来。因此为了更高质量的人机交互，日常交流中的情感应该以不同的情感维度被评估出来。

视觉信号已被证明是最有效且最重要的情感识别的线索[1, 18, 22]。自然这样的观点就被提出，在日常生活中的情感往往比被表现出来的改变得更慢，从而持续的表情变化更加微妙。在日常交流中的重要表达往往不会被充分地表现出来，这就导致不同的情感状态之间的模糊差异。另外，人们以多变的方式表达自己的情感，相似的情感传达的信息令人更加迷惑，这就为确切的用户表情与更普遍的表现连接起来带来更大的挑战。除此之外，从日常交流中不活的情感总是伴随大的变化，例如更自由的头部旋转，快速的头部移动，部分的面部遮挡等等。这些特征增加了连续情感的复杂度，使准确可靠地评估日常交流中的情感变得困难。

为了解决这些挑战，我们提出了一个实时的基于 3D 模型的方法，能够识别日常交流中三维空间的人类情感。我们的方法将 3D 面部模型引入连续情感识别，使处理这些变化，大角度头部旋转，快速头部移动和部分面部遮挡，有了更高的健壮性。用户特有的暂时特征和用户独立的情感表现也被构造出来，能够更加准确地描述情感。一个新奇的随机基于森林的框架评估出情感，以回归的方式同时呈现出 3D 面部记录和连续的情感评估。

2. 相关工作

人类情感通过以两种方式表示，绝对分类和维度。依据面部表情编码系统（FACS），P.Ekman[9]建议情感被分为六类：愉快，伤心，生气，惊喜，害怕和厌恶。日常人类情感是复杂的，表达时有着模糊的界限，因此独立的分类无法反映出微妙的情感转变和情感差异。因此，许多工作使用维度表示来说明不同情感维度下的人类情感。PAD 情感空间是典型的一个，它以愉悦度、激活度和优势度三个维度来描述连续的情感。Fontaine 等人 [13]以唤起程度、评价值、能量值和期望值四个维度来描述连续情感。维度表示法能够以众多连续的测量值来分析情感和更好地描述情感变化，因此更适于日常人机交互中的情感表达。

大多数已有的情感识别算法使用从图像中获取的 2D 特征来预测情感，可以用外貌特征和几何特征[11]再细分。例如，Wu 等人 [36]用 Gabor 滤波器的运动能量后的强度来分类情感。Kapoor 等人[18]用嘴部位的像素差异来评估情感。这样的算法基于外貌特征并且当面部姿势不改变时得到了正确的结果。一些工作用 2D 面部几何特征。Valstar 和 Pantic 用 20 个 2D 面部点的几何特征来预测情感。Kobayashi 和 Hara 用面部几何模型来识别情感。也有一些类似情感识别的工作也使用表情和几何的 2D 混合特征，例如主动外观模型（AAM）。2D 特征可以直接获取，但是正如 Sandbach 等人在他们的调查中指出的那样，这样的算法对于在交流中的巨大改变是不足够稳定的，当使用 2D 特征时必须一个固定不变的面部姿势，这表明基于 2D 特征的算法还不具有足够的健壮性以识别持续地表情。

在一些算法中结合了 3D 特征。与使用 2D 图像特征的方法相比，基于 3D 特征的方法对于情感识别更加健壮有力。用 3D 特征的工作可分为基于图形和基于深度两类。基于图形的算法使用 3D 曲线图形的参数，3D 坐标的位置或 3D 坐标的改变来分类情感。 例如，Huang 等人[31， 39]用贝塞尔曲线来描述面部表情，把多方面的参数变化作为情感变化的标志。他们的实验结果表明基于贝塞尔曲线的方法在分类自然表情时运行地很哈。一些其他的工作用面部深度特征来识别情感。Fanelli 等人[10]使用深度信息将情感分为独立的类别。现有的基于 3D 特征模型已足够健壮但

他们仍很少被用于连续情感识别。在这篇论文中，我们展示了一个有效的回归方法用 3D 面部信息来评估立体空间中的持续情感。

持续地情感展示是顺序化的行为。为了包含进动态的暂时信息，消除依赖用户的信息并且克服交流环境的巨大变化，混合的情感展示已经被设计出来。Yang 和 Bhanu[37, 38]展示他们的图像融合方法，该方法用 SIFT 流算法将来自一个视频的图像与另一个图像相结合。SIFT 流[23]对于 2D 图像队列是一个健壮的算法，在脸部识别方法也表现得很好，但是它相对耗时。在 3D 模型的帮助下，我们提出一个实时图像混合方法来分别表现持续的情感和用户独立的情感。

已经有许多方法被设计去识别连续的情感[16]。一些经典的方案有：支持向量机回归（SVR）[30]，相关向量机（RVM）[25]，条件随机场（CRF）[2]等等。随机森林[4]作为一个流行的方法，已经被广泛使用在分类和回归任务中。随机森林包含众多分类和回归树（CART）。它能够在不过度拟合的情况处理大数据的测试样本，并且有这样的特征，健壮性，高效性和强大的拟合能力。由于二叉树的结构，它能够在短时间消耗中获得结果。Fanelli 等人[10]提出一个基于框架的随机森林通过从深度相机中采集的数据回归式来评估头部姿势。输出表明随机森林可以高质量地解决面部回归问题。在我们的工作中，我们进一步研究了随机森林的回归能力，它在 3D 面部轨迹和持续情感识别中都起到了作用。

3. 提出方法

我们提出一个基于森林的随机算法，能够在日常交流中识别出立体空间中的情感。与现有算法不同的是[12, 32]，我们的方法使用从2D图像重建的3D面部模型，这维持了面部坐标的位置关系，提供了更具健壮性的规则来克服改变的环境。通过基于3D头部模型的融合图像持续的情感表现和用户独立的情感表现也被记录下来，从而更准确地描述情感。

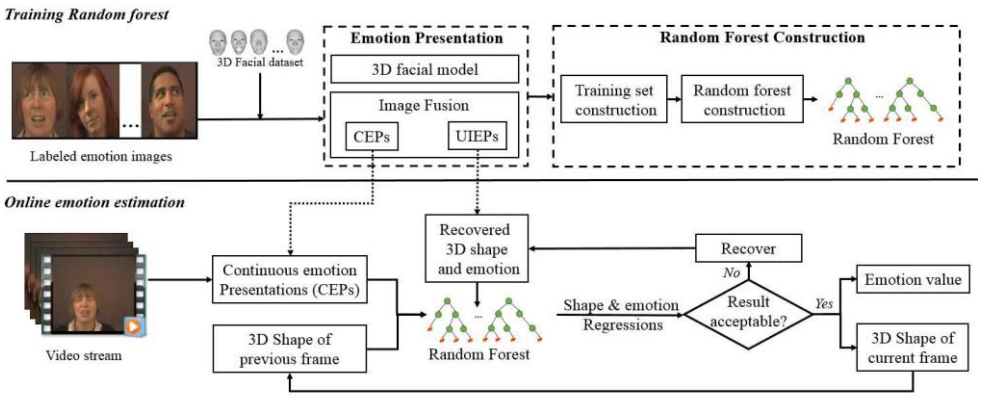

Figure 1. 我们基于 3D 模型连续情感识别和记录方法的框架

我们的工作框架展示在图 1 中。在测试阶段，输入图像的 3D 面部模型首先被存储。然后连续情感表现（CEP）和不受用户约束情感表现（UIEP）通过图像融合构建。3D 面部图形，CEP 图像与其情感值构成了一个增强的训练体系，随之随机森林被构建。在情感评估阶段，两个回归式被同时放入随机森林中：一个为记录 3D 面部表情，另一个为识别当前的情感。当前时间步长的 CEP 图像和先前时间步长的 3D 面部图形被当做输入，然后情感值和当前步长的 3D 面部图形被计算作输出。当随机森林没有合理的输出时，会在 UIEP 图像的帮助下执行恢复操作来获取重新获得的 3D 面部图形和情感。

3.1. 数据准备

连续情感数据集总是以包含许多图像的视频剪辑的方式呈现，这些图像的大部分在情感值和表现方面是相似的。为了降低数据冗余并且提高测试数据的代表性，被缩减的测试图像首先从所有帧中被提取出来。在图像提取步骤中，我们确保被选择的图像有被平均分配的情感值，覆盖整个情感范围并且保留了日常交流中不同的头部姿势。为了每一个有影响的维度，在我们的方法中大约 160 个相关的小测试图像首先被提取。然后，每一个被选择图像的面部坐标通过 Baltrusaitis 等人[3]的算法被自动检测到。考虑到情感信息大多表现在嘴部，眼睛和眉毛

Figure 2 一些被选择图像的面部标记.

这个事实，我们的方法仅仅选择了 42 个内部标志，包括 8 个眉毛标志，12 个眼角标志，4 个鼻子标志和 18 个嘴部标志。Figure2 展示了一些被选择的图像中已标记处的标志。

3.2. 还原 3D 面部模型

FaceWarehouse，一个 3D 面部数据库[6]，它包含 150 个不同种族的人的 3D 面部模型并且每一个对象有包含 11k 顶点的 47 个 FACS 融合图形，在其帮助下我们的方法还原了每一个标记图像的 3D 面部图形。这可以用一个三元式描述出来：

$$F = Cr \times w^{T}_{id} \times w^{T}_{exp} \quad (1)$$

C_r是一个包含11K顶点的3D面部图形融合，w^{T}_{id}，w^{T}_{exp}在式中分别是身份权值和表情权值的列向量。

根据Cao Chen等人[5]所做工作，从2D图像中构建3D模型可被分为两步：第一步是计算最佳w^{T}_{id}。用w^{T}_{id}最佳图形融合，每一个被选择图像的3D面部图形在第二步中被构建。这两步以迭代的方式进行。

与Cao Chen等人工作不同的是，他们关注于特定的用户，我们想以统一的方式代表不同人的输入图像，考虑到所有的输入图像应该被

FaceWarehouse中相同w^T_{id}的混合变形构建。所以当计算最适宜w^T_{id}时，考虑这个约束的能量公式定义如下：

$$E_{id}=\sum_{i=1}^{N}\sum_{b=1}^{42}\left\|P(M^i(C_r\times w_{id}^T\times w_{exp,i}^T)^b-u_i^b)\right\|^2 \quad (2)$$

其中N是被选择图像的数量；P指预测矩阵；M^i指通过EpnP算法[21]计算得到的相机外在参数矩阵；$w_{exp,i}^T$代表第i个凸显过得最相似表达；u_i^b是图像中第b个标记。有最少能量E_{id}的恒等式w^T_{id}被认为是最佳等式，最佳w^T_{id}的47个融合变形作为基础融合变形置入3D面部模型重构中。

一旦基础融合变形被请求，每一幅图像的3D面部模型可以被恢复，通过基础融合变形[21]的线性插入，被选择2D图像的3D面部模型可以全部被重新构建。

3.3. 图像融合

在 3D 情感表现的帮助下，一幅图像融合方法可以被实施。图 3 展示了我们的图像融合方法的流水线。首先，我们用算法标注输入图像的标点，重构 3D 面部模型。然后 3D 面部图形被转换成空间二维直角坐标再用以下公式投影成 2D 面部坐标系：

$$u_i^{OP_b}=P(M_{R|t}*V^b) \quad (3)$$

其中 P 代表相机的投影矩阵，$M_{R|t}$ 代表 3D 图形从原始位置到当前空间坐标系的直角坐标转换矩阵。V^b 代表 3D 面部图形的第 b 个标记。

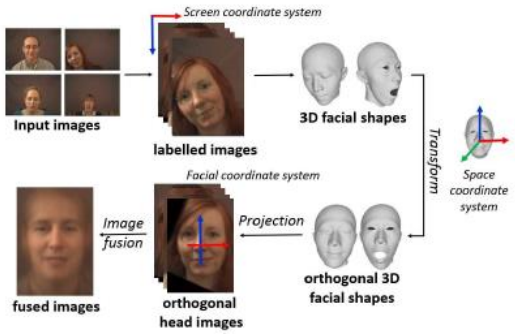

Figure 3 图像融合流水线

用原始标记和插入标记$u_i^{OP_b}$，可以获取从原始屏幕空间到面部坐标空间的对应转换矩阵。原始图像的面部部分被统一为 2D 面部坐标系。在所有的原始图像面部部分转换成统一的面部坐标系后，这些图像被添加在上面成为一个融合式表现。

对于不同目标，在我们的工作中图像融合方法被用作产生用户特定的连续情感表现形式和不受用户约束的情感表现。连续情感表现（CEP）将一个视频剪辑中邻近帧拼接，可将情感的动态特征和暂态情况包含其中。不受用户约束的情感表现（UIEP）将在不同视频有着相同情感的图片融合成一个图片表现形式，来保留同一情感状态的显著特征并评估不同人之间的差异。

3.4. 测试

测试集结构。随机森林由许多分类和回归树（CARTs）构成。如在 3.1 部分表述的，选择的情感样本数量相对较小时，不足以确保 CART 的健壮性和精确性。所以我们首先扩大了情感样本使其足够大以测试。

假定{CEP_i，M_i，S_i，A_i}是第 i 个测试图像的情感样本，其中 CEP_i 是第 i 个图像的被融合持续表情表现；S_i 是重新构建的 3D 情感图形；A_i 是被标记的情感值；M_i 是恒等矩阵。我们首先分别沿着三个坐标轴作 3D 情感图形 Si 的变换得到 M-1 额外的 3D 情感图形，这将{CEP_i，M_i，S_i，A_i}增大至 N*M 个测试样本，其中 M_{ij} 是 S_i 到 S_{ij} 的转换矩阵。从而算出 M_{ij} 相对应的单应矩阵 M_{HOMO}。用 M_{HOMO} ，CEP_i 可被转换成 CEP_{ij}，被用作 S_{ij} 的连续情感表达式。

每个被转换的情感样本{CEP_i，M_i，S_i，A_i}的一些相似的情感样本随之可被找到。假定{CEP_{ij} ，M^l_{ij}，S_l ，A_l}代表另一个情感样本，两个情感样本的不同点如下计算：

$$E_l = \sum_{b=1}^{42} \left\| S_{ij}^b - S_l^b \right\|^2 + w_a \left\| A_i - A_l \right\| \quad (4)$$

$$S_l = M_{ij}^l S_{ij} \quad (5)$$

上标b代表在3D面部图形S_i和S_{ij}的第b个标记；A_i和A_l分别是相应图形的情感值；M^l_{ij}是3D图形间的转换矩阵；ωa是衡量图形多样性和情感多样性的影响，这里是350。通过最小化以上能量E_l可以得到最优情感样本。从而情感样本可被扩展为{CEP^1_{ij}，M^1_{ij}，S^1_{ij}，A_l}。最终分别沿着三条坐标轴转换S^1_{ij}，并从被转化的图形中随机选择K个图形，我们将得到增大的情感图形{CEP^{1k}_{ij}，M^{1k}_{ij}，S^{1k}_{ij}，A_l}。增大之后，测试情感样本的数量从N增大到N*M*L*K。这里，我们置N=160，M=9,L=3,K=7。

用增强情感样本，为了测试随机森林而构建测试点。置于情感样本{CEP^{1k}_{ij}，M^{1k}_{ij}，S^{1k}_{ij}，A_l}，形成了一些测试点反映了3D情感图形的替换，情感值的差异和图像的表现。来自原始图形Si的情感图形S^{1k}_{ij}中的每一个面部标记的替换被记做$Dis_s(S^{1k}_{ij}, S_i)$。情感值间的不同以$Dis_a(A_l, A_i)$表示。为了代表2D图像的表示，我们从CEP^{1k}_{ij}中面部区域随机地选择了Q个点，将这强度值浓缩为强度向量$Int(CEP^{1k}_{ij})$，在我们的实验中将Q固定为400。因此，一个点向量以

P={ $Int(CEP^{1k}_{ij}), Dis_s(S^{1k}_{ij}, S_i), Dis_a(A_l, A_i)$}被建立。图4展示了为一个情感样本产生测试点的例子。我们在每一个CEP中随机地选择Z强度向量，在每一个情感样本中得到Z个点{$P_z|1\leqslant z\leqslant Z$}，Z为100.最终，一个包括N*M*L*K*Z个测试点的测试集被构建。

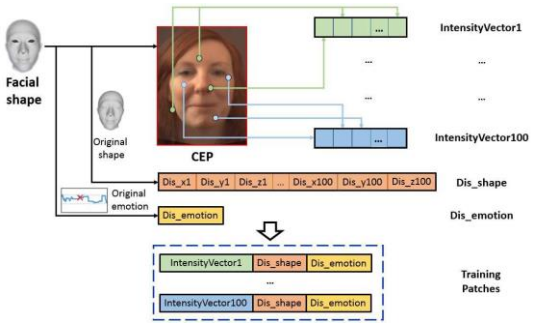

Figure 4 测试点结构

随机数结构。利用形成的点，包括许多 CARTs 的随机森林可以被构建。当测试每一个 CART 时，仅 70%的点被用来避免过度拟合。

在每一个非叶子节点中，执行一个

二分测试来分离测试点，如下定义：

$$|F_1|^{-1}\sum_{q1\in F_1} Int_{q1} - |F_2|^{-1}\sum_{q2\in F_2} Int_{q2} > \tau \quad (6)$$

F1 和 F2 是当前测试点的两个帧，Int 代表强度向量，τ 是随机的临界值。在我们的测试中，F1 和 F2 的长度被置为 60，二分测试临界值的范围是[-30,30].

对于每一个非叶结点通过随机地选择 F1，F2 和 τ 的参数产生了 2000 个二分点$\{t^x\}$。

通过回归不确定度 U_R 评定每个二值化检验的质量，这包括两部分：图形回归不确定度 URS 和影响回归不确定 URa。这两个回归不确定被定义如下：

$$U_{R_s}(P\mid t^x) = H(P)_s - w_L H(P_L)_s - w_R H(P_R)_s \quad (7)$$

$$U_{R_a}(P\mid t^x) = H(P)_a - w_L H(P_L)_a - w_R H(P_R)_a \quad (8)$$

H（P）代表点集的微分熵，ωL，ωR 分别是左分支和右分支中点的比例。假定测试集的分布是正常的。所以回归不确定度可以按一下公式计算：

$$U_{R_s}(P\mid t^x) = \log(|\Sigma^s|) - \sum_{i=\{L,R\}} w_i \log(|\Sigma^s|) \quad (9)$$

$$U_{R_a}(P\mid t^x) = \log(|\Sigma^a|) - \sum_{i=\{L,R\}} w_i \log(|\Sigma^a|) \quad (10)$$

Σ^S和Σ^a是图形标记和情感值该变量的协方差。总不确定度 UR 表示如下：

$$U_R(P\mid t^x) = U_{R_s}(P\mid t^x) + \lambda U_{R_a}(P\mid t^x) \quad (11)$$

λ 是值为 1 的经验权重。通过最小化 UR 可以最小化这些协方差矩阵的决定性因素从而找到当前节点的最优二值化，记作 t^{opt}。

一旦找到 t^{opt}，我们将最优二值化的参数记作随机回归森林的一部分并将当前节点的测试点分别它的左孩子节点和右孩子节点。如果一个节点到达最深层 Lmax 或它所含点的数量少于最小临界值 Pmin，则视其为叶子节点。这里 Lmax 是 15，Pmin 是 20.叶子节点不再分割并保留关于它所含点的信息，包括图形位移{Aves，$|\Sigma^S|$}的均值和协方差与情感值位移{Avea，$|\Sigma^a|$}的均值和协方差。

3.5. 在线情感估算

准备工作。在线情感估算之前需要做一些准备工作。在情感识别过程中，被忽视的情况将出现。在这种情况下，3D 面部模型和情感值需要恢复。所以我们提前准备了图形恢复集和情感恢复集。为了准备图形恢复集，在固定的时间间隔中从输入视频中选择一些帧。这些被选择图像的 3D 面部图形被重新构建成一个图形恢复集 Rshape。对于情感恢复集，有着相似情感值的图像产生并存为一些组别。这些组中不受用户约束的情感表现（UIEPs）随之被计算。每一个 UIEP 的面部标记被自动侦测，每一个标记区域（标记点周围 10*10 的点集）的 LBP 特征被存为 LBP 情感表现形式。LBP 的情续表现和所有 UIEPs 相应的情感值被收集成情感恢复集 Remotion。

另一个准备工作是产生第一帧的 3D 情感图形和情感值。我们用在 3.1 部分中提到的方法来恢复第一帧的 3D 图形。当计算第一帧的情感值时，我们通过比较第一帧 LBP 情感表现和 Remotion 中 UIEPs 的 LBP 情感表现的相似性来计算其 LBP 情感表现并发现

其情感值。

情感估算。把当前时间步中的 CEPt，先前时间步的 3D 情感图形和情感值{St-1,At-1}作为输入，当前时间步 t 的 3D 情感图形和情感值{St,At}可以回归方式被估算出来。

给出前一时间步的输入情感图形 St-1 和情感值 At-1，测试数据库中一些最为类似的 3D 情感图形和他们所对应的情感标记{S^{ω}，A^{ω}}被挑选出来。然后从 St-1 到 S^{ω}的仿射矩阵和其对应的 $M\omega$ 单应矩阵随之产生，这样 CEP_t 被转换为 $CEP^{\omega}{}_t$ 作为 S^{ω}的连续情感表现。

从 S^{ω}中我们随机地从面部区域选择 400 个点产生一个点集 P^{ω} = { $Int(CEP^{\omega}{}_t)$, $Dis_s(S^{\omega}, St-1)$, $Dis_a(A_{\omega}, A_{t-1})$}。每一个测试点被放入随机森林中，每一个 CART 的叶子节点通过图形位移$|\Sigma^{S}|$和情感位移$|\Sigma^{a}|$得到。为选取可用叶子设临界值 θ s=10， θ a=-1.5。如果 $\log(|\Sigma^{S}|)$比 θ s 大或 $\log(|\Sigma^{a}|)$ Ashra09 比 θ a 大，便舍弃该叶子。最终形成一个可用叶子集。然后分别通过平均图形位移和情感位移计算的图形和情感的回归值。把他们分别加入 S^{ω}和 A^{ω}中得到新的图形 $S^{\omega*}$和情感值 $A^{\omega*}$。

用以上被选择的相似 3D 图形，我们最终可以得到 3D 图形和情绪的估算值集合{$S^{\omega*}$, $A^{\omega*}$}。这些结果的中值{$S^{\omega}{}'$, $A^{\omega}{}'$}被选作当前

Algorithm 1 Emotion estimation

$\{S^w, A^w\} \leftarrow$ a set of most-like 3D shapes and emotion values from training set $\{CEP_{ij}^{lk}, M_{ij}^{lk}, S_{ij}^{lk}, A_l\}$
for each simple in $\{S^w, A^w\}$ **do**
  $M_w \leftarrow$ affine transformation matrix from S_{t-1} to S^w
  $P^w \leftarrow \{Int(CEP_t^w), Dis_s(S^w, S_{t-1}), Dis_a(A^w, A_{t-1})\}$
  for $n = 1$ to N **do**
    $Leaf_n \leftarrow$ leaf node of the n^{th} CART reached by P^w
    if $Leaf_n \rightarrow |\Sigma^s| > \theta_s$ or $Leaf_n \rightarrow |\Sigma^a| > \theta_a$ **then**
      discard $Leaf_n$
    end if
  end for
  $\{Leaf_n\} \leftarrow$ the acceptable leaves
  $Reg_s \leftarrow (\Sigma_{n=1}^{N} ave_s^n)/N$
  $Reg_a \leftarrow (\Sigma_{n=1}^{N} ave_a^n)/N$
  $S^{w*} \leftarrow S^w + Reg_s$
  $A^{w*} \leftarrow A^w + Reg_a$
end for
$\{S^{w*}, A^{w*}\} \leftarrow$ alternative results from random forests
$\{S^{w'}, A^{w'}\} \leftarrow$ median result of $\{S^{w*}, A^{w*}\}$
$S_t \leftarrow M_{w'}^{-1} S^{w'}$
$A_t \leftarrow A^{w'}$
$return(S_t, A_t)$

3D 图形和情绪的最终估算。用逆矩阵 M_{ω}^{-1}转换 $S^{\omega}{}'$可以得到 3D 图形 St。当前时间步的情感值 At 等于 $A^{\omega}{}'$。

由于连续情感变化趋势经常是细微的，我们用之前 500 个情感值来计算当前情感值，把这个结果作为当前时间步的最终情感值。

恢复。有两种情况我们需要恢复 3D 面部图形和情感值。一个是没有得到可用叶子。子啊这种情况下，图形和情感都要恢复。

在图形恢复中，我们从图形恢复集 Rshape 中找到最接近当前时间步的 3D 图形，把它作为新的输入图形。用新的输入图形和当前 CEPt，CEPt 的 LBP 情感表现 LBPt 得以计算。将上一帧的情感值 Et-1 作为约束，能量 Ea 被如下定义：

$$E_a = \|LBP_t - LBP_i\|^2 + \beta \|A_{t-1} - A_i\|^2 \quad (12)$$

LBPi 是情感恢复集 Remotion 中一个 LBP 情感表现； β 是值为 45 的经验权重。通过最小化能量 Ea 得到最接近当前帧表现的一些 UIEPs。他们情感值的

均值作为被恢复的情感值。

另一种情况是邻近帧间检测到情感值的较大变化。当连续情感变化细微时，如果邻近情感值的变化比经验临界值 θ diffA 更大时，我们认为估算的情感值错误并执行如上的情感恢复。在我们的测试中，θ diffA 被设为 0.2.算法 1 中展示了在线情感估算的伪代码。

4. 实验

为了评估这个方法的合理性，我们建立了一个样本模型并通过下面三方面内容评定我们的方法：<1>3D 面部追踪的精密性,2>情绪辨别的相互系数关系 3>我们模型的计算成果。

我们的连续情感辨别和跟踪系统在一个双核志强英特尔处理器（3.2GHz）和 4G 内存的电脑上执行。

4.1. 数据库

听觉/视觉的情感挑战数据库（AVEC 2012）[30]是一个连续不断的的公共数据库，被视频用 SEMAINE 语句[24]顺序自动记录。在这个数据库里，每一帧的情感都被人们用像愉悦度，评价值等等这样的维度标注，每个视频的长度大概 3-5 分钟 视频中每个图片的分辨率是 780*580，fps 是 50 帧。

我们用 AVEC 2012 处理我们的数据库并且根据皮尔逊相关系数评估情绪辨别的能力。

由于唤醒程度和评价值已经更加频繁的应用到情感表现，我们用那两个维度去测试我们的模型并用我们的结果跟 AVEC 2012 的底线去比较，好多个呈现出了最好的结果，这个正好测试了相同的数据集。

4.2. 实验结果

感情辨别的评估很大一部分建立在面部表情分界的位置，我们首先把我们表情跟踪准确性的算法跟许多个典型成果比较。在表格 1 里，我们为图片上界标的不同算法测量 RMSE,并跟实况调查比较。表情跟踪模型的结果包括多线模型[35]，2D 回归[35]，和用 3D 回归模型[35]跟踪界标的艺术陈述都被提到了。通过这个表格，我们能够发现我们的算法比 2D 模型能更完整精确，而且能够实现相似级别结果中的 3D 跟踪的最佳值，这就意味着我们的算法在面部表情跟踪的情感评估已经足够精确。

表格 5 展示出了一些我们在 3D 面部追踪模型的结果，红色的点代表界标的实地状况，绿色的点代表使我们模型的追踪结果。通过输出我们能够发现我们的主要的基于模型的样式能够在像错过飞机时头转动，快速头部运动和部分表情封闭这样改变的相互作

Figure 5 3D 面部追踪的结果。红点：实际情况；绿点：追踪结果

RMSE	<3 pixels	<4.5 pixels	<6 pixels
Multilinear Model [35]	20.8%	24.2%	41.7%
2D Regression [7]	50.8%	64.2%	72.5%
3D Regression [5]	73.3%	80.8%	100%
Our Method	70%	83.93%	94.91%

Table 1 面部追踪中 RMSE 帧百分比

Correlation coefficient	Arousal	Valence	Mean
SVR [30]	0.151	0.207	0.179
Multiscale Dynamic Cues [26]	0.509	0.314	0.4165
CFER [32]	0.30	0.41	0.355
CCRF [2]	0.341	0.343	0.342
Our Method	0.564	0.454	0.509

Table 2 在 AVEC2012 数据库中测试的经典情感回归方法的皮尔逊相关系数

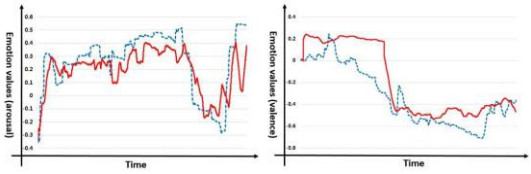

Figure 6 我们的情感评估结果与实际的比较。左：愉悦度维度，右：评价值维度

用的环境下保持好的结果。

表格 2 展示出我们的跟多个典型算法情感评估比较的结果。第一条线展示出 AVEC 2012 比赛[30]的基准线，用 SVR 作为回归者。第二条线[26]展示出了多尺寸动力模型线索的结果，第三条线是连续面部表情表现形式（CFER）[32]的结果，第四条线是连续条件随机场 CCRF [32]的结果，通过比较我们发现我们算法在连续表情的评估方面比其他四个模型性能好。

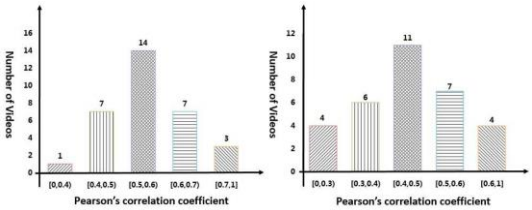

Figure 7 愉悦度维度（左）和评价值维度（右）的相关系数分布

在表格 6 展示了两个跟表情评估跟一个实况资料比较的视频。红色的实线是我们模型的结果蓝色加点的先是实况资料，左边的一个是愉悦度维度，右边的一个是估计值维度。 通过图形我们能够看到我们的算法能够准确的辨别表情的价值。

图 7 中的直方图给我们宏观展示了不同回归系数下愉悦度维度和估计值维度视频的数量，通过那两个直方图，我们能够看到我们系统的结果大概跟正常分布一致，，这就意味着我们提议的算法是可靠的。在就大多数情况下，我们的算法能达到 0.5-0.6 在愉悦度维度下。0.4-0.5 在估计值维度下，这跟表格 2 的结果一致。

我们系统的时间性能绝大多数取决于 CARTs 数。表格 3 展示出我们系统在不停 CARTs 数下的性能和皮尔逊相关系数在那些情况下。我们能够看到，随着 CARTs 数 的增加，我们系统的性能迅速下降，情感评估的相互关系更好的增长。当 CARTs 数比 6 大 评估准确

度几乎稳定在 0.51 左右。由于在真实时间工作对一个相互作用的系统是很必要的，我们用 6CARTs 来练习作为卖出的效率跟准确度。然后我们的模型能够在 20 帧左右的速度掌握情感辨别，这个很多相互作用的系统都不能做到。

Number of CARTs	Frames handled in one second	Pearson's correlation coefficient
3	25	0.413
5	23	0.472
6	21	0.511
8	18	0.512
10	17	0.507
12	15	0.514

Table 3 用不同 CARTs 的时间表现

5. 结论

在这篇论文中，我们提出了一个基于 3D 模型的连续情感识别方法。我们将 3D 面部表情模型引进我们的工作中，从 2D 图像中恢复出 3D 面部模型。用重构的 3D 面部图形，提出了图像融合方法来产生一个用户的连续情感表现（CEP）和不受用户约束的情感表现（UIEP）。用 3D 面部模型和融合图像，构建出一个随机森林，它使 3D 标记跟踪和情感估计同步地成为整体。

我们的算法在 AVEC2012 数据库中的视频段中得以测试。实验结果表明我们的实时方法有足够的能力获取连续情感估计的可取结果。进一步，我们系统的高效性使在人机交互方面提供及时智能反馈成为可能。

虽然我们的算法基于 3D 面部模型，但仅需要输入 2D 图像或 2D 视频流，摆脱了设备的束缚。由于系统的易用性，有希望使我们的系统安装在移动设备如智能机，平板电脑等上。

我们的工作中提取的图像特征时图像的维度，这是图像的最基本特征。在未来工作中，我们将尝试用更具健壮性的特征，例如从恢复的 3D 面部模型中提取的特征，这将有更好的效能。

在论文中，我们仅仅在 AVEC2012 数据库中测试了我们的方法，因为 AVEC2012 是典型且流行的，被连续情感识别的研究者们广泛接受。仍然有其他一些好的数据库可以被用来测试我们的方法。未来我们将在更多的数据库上进行测试。

除此之外，研究者们已经指出了情感维度间彼此关联[14]。在未来工作中，我们将关注于开发和模拟不同情感维度间的关系来更好地实现情感分析。

6. 感谢

The authors gratefully acknowledge SSPNET for providing the AVEC 2012 dataset and Cao Chen, et al. for provi-ding the FaceWarehouse dataset. We also thanks the source code about facial landmarks labeling provided by T. Baltrusaitis et al. This research was supported by by the National Natural Science Foundation of China: 61135003, 61232013, 61173059 and 973 Project:2013CB329305.

7. References

[1] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon. The painful facecpain expression recognition using active appearance models. *Image and Vision Computing*, 27(12):1788 - 1796, 2009. 0, 1

[2] T. Baltrusaitis, N. Banda, and P. Robinson. Dimensional affect recognition using continuous conditional random fields. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1 - 8, 2013. 1, 6

[3] T. Baltrusaitis, P. Robinson, and L. P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *IEEE International Conference on Computer Vision Workshops*, pages 354 - 361, 2013. 2, 3, 6

[4] L. Breiman. Random forests. *Machine learning*, 45(1):5 - 32, 2001. 1

[5] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics*, 32(4):41, 2013. 2, 6

[6] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse:a 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413 - 425, 2014. 2

[7] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177 - 190, 2014. 6

[8] A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests:A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2 - 3):81 - 227, 2012. 1

[9] P. Ekman. An argument for basic emotions. cognition and emotion. *IEEE Transactions on Cybernetics*, 6(3 - 4):169 - 200, 1992. 0

[10] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. V. Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101(3):437 - 458, 2013. 1

[11] B. Fasel and J. Luettin.

Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259 - 275, 2003. 1, 5

[12] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3):143 - 166, 2003. 0

[13] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth. The world of emotions is not two dimensional. *Psychological science*, 18(12):1050 - 1057, 2007. 1

[14] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, 1(1):68 - 99, 2010. 7

[15] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120 - 136, 2013. 0

[16] H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: a survey. In *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops*, pages 827 - 834, 2011. 1

[17] E. Hudlicka. To feel or not to feel: The role of affect in humanccomputer interaction. *International Journal of Human Computer Studies*, 59(1):1 - 32, 2003. 0

[18] A. Kapoor, W. Burleson, and R. W. Picard. Automatic prediction of frustration. *International Journal of Human Computer Studies*, 65(8):724 - 736, 2007. 0, 1

[19] H. Kobayashi and F. Hara. Facial interaction between animated 3d face robot and human beings. In *Proceedings of the International Conference on Systems, Man and Cybernetics*, pages 3732 - 3737, 1997. 1

[20] B. Kort and R. Reilly. Analytical models of emotions, learning and relationships: towards an affect-sensitive cognitive machine. In *Conference on Virtual Worlds and Simulation*, 2002. 0

[21] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate O(n) solution to the PNP problem. *International Journal of Computer Vision*, 81(2):155 - 166,

2009. 3

[22] G. C. Littlewort, M. S. Bartlett, and K. Lee. Faces of pain: automated measurement of spontaneous all facial expressions of genuine and posed pain. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, pages 15–21, 2007. 0

[23] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2011. 1

[24] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1):5–17, 2012. 6

[25] M. A. Nicolaou, H. Gunes, and M. Pantic. Outputassociative rvm regression for dimensional and continuous emotion prediction. *Image and Vision Computing*, 30(3):186–196, 2012. 1

[26] J. Nicolle, V. Rapp, K. Bailly, and et al. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pages 501C–508, 2012.6

[27] C. Peter and R. Beale. Affect and emotion in humancomputer interaction: From theory to applications. *Lecture Notes in Computer Science*, 4868, 2008. 0

[28] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert. Recognition of 3d facial expression dynamics. *Image and Vision Computing*, 30(10):762 – 773, 2012. 1

[29] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683 –697, 2012. 1

[30] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pages 449–456, 2012. 1, 6

[31] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S.Huang. Authentic facial expression analysis. *Image and Vision Computing*, 25(12):1856－1863, 2007. 1

[32] C. Soladi´e, H. Salam, C. Pelachaud, N. Stoiber, and R. S´eguier. A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pages 493－500, 2012. 1,6

[33] P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394,1994. 1

[34] M. F. Valstar and M. Pantic. Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. *HumanCComputer Interaction*, pages 118－127, 2007. 1

[35] D. Vlasic, M. Brand, H. Pfister, and J. Popovi´c. Face transfer with multilinear models. *ACM Transactions on Graphics*, 24(3):426－433, 2005. 6

[36] T. Wu, M. S. Bartlett, and J. R. Movellan. Facial expression recognition using gabor motion energy filters. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 42－47, 2010. 1

[37] S. Yang and B. Bhanu. Facial expression recognition using emotion avatar image. In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, pages 866－871, 2011. 1

[38] S. Yang and B. Bhanu. Understanding discrete facial expressions in video using an emotion avatar image. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(4):980－992, 2012. 1

[39] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang. Audiocvisual affective expression recognition through multistream fused hmm. *IEEE Transactions on Multimedia*, 10(4):570－577, 2008. 1