基于无监督提取表情时空特征的情感识别

王金伟[1,2] 马希荣[2] 孙济洲[1]

(天津大学计算机科学与技术学院 天津 300072)[1] (天津师范大学计算机与信息工程学院 天津 300387)[2]

摘 要 情感识别是解决智能教学系统中情感缺失问题的关键技术。针对识别时如何从视频中有效提取人脸表情时空特征的问题,提出一种采用堆叠卷积独立子空间分析模型进行无监督特征提取的识别方法,来对疑惑、愉快和厌倦3种学习中最常出现的情感进行识别。该方法检测视频中的人脸区域并进行规范化处理,采用堆叠卷积独立子空间分析模型从视频块中无监督地学习表情的时空特征,采用线性支持向量机进行分类。实验结果表明,相比使用人工特征的方法,该方法能够更有效地提取视频中人脸表情的时空特征,获得更高的识别率,同时符合实时性要求。

关键词 情感识别,无监督学习,独立子空间分析,时空特征,人脸表情

中图法分类号 TP391.41 **文献标识码** A

Emotion Recognition Based on Unsupervised Extraction of Facial Expression Spatio-temporal Features

WANG Jin-wei[1,2] MA Xi-rong[2] SUN Ji-zhou[1]

(School of Computer Science and Technology, Tianjin University, Tianjin 300072, China)[1]

(College of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387, China)[2]

Abstract Emotion recognition is the key to solving the problem of the absence of emotional communication in intelligent tutoring systems. According to the problem of effective extraction of facial expression spatio-temporal features from videos for emotion recognition, a recognition method based on unsupervised feature extraction using stacked convolutional independent subspace analysis (ISA) model was proposed to recognize three emotions including puzzlement, delight and boredom that most often appear in learning. This method first detects face in video and normalizes it, then adopts stacked convolutional ISA model to learn (without supervision) facial expression spatio-temporal features from video blocks, finally uses linear SVM classifier to recognize different emotions. Experimental results indicate that this method can extract spatio-temporal expression features more effectively than the use of hand-designed features, as well as recognition rate is better, and it meets the requirement of real-time.

Keywords Emotion recognition, Unsupervised learning, Independent subspace analysis, Spatio-temporal feature, Facial expression

1 引言

情感识别目前已成为实现自然人机交互的一个重要研究课题。特别是在智能教学系统(Intelligent Tutoring System, ITS)中,情感识别可以解决计算机教学的情感缺失问题。心理学研究表明,人们在学习中最常产生的情感为疑惑、愉快和厌倦[1],其中愉快是积极情感,而疑惑和厌倦是消极情感。因此,研究对这些情感的有效识别方法有助于ITS系统像传统课堂教学中的教师一样,随时掌握学习者的情感状态,及时提供恰当的指导,激发积极情感,消除消极情感,提高学习效率。

识别人类情感的最直接渠道是面部表情,它及时地反映出人们的心理状态。表情的变化是一个动态过程,在实际中一般通过视频记录,因此需要从中提取时空特征。通常的提取方法往往依赖于人工设计的特征,例如HOG[2]、HOF[3]和HOG3D[4]。这些特征的设计需要深层的领域知识和大量时间,而且这些特征并不一定能有效地描述被识别对象的特点。近年来,无监督的特征学习逐渐成为研究热点,例如Sparse Coding[3]、Deep Belief Nets[5]和Stacked Autoencoders[6]。不同于人工特征,无监督特征学习是通过用海量无标签数据训练多层模型来自动获得能够更好地描述被识别对象的特征,最终提升识别率。但这些方法在处理视频等高维数据时,模型的训练时间会急剧增加,从而限制了它们的应用范围。针对这一问题,Quoc V. Le等人在独立子空间分析(Independent Subspace Analysis, ISA)的基础上提出了堆叠卷积ISA(Stacked convolutional ISA, SISA)模型[7],该模型通过采用分层结构、卷积和PCA降维技术,大幅缩短了训练时间,从而能够应用于视频数据。

因此,本文基于SISA模型,提出了一个无监督提取表情时空特征的情感识别方法,来对疑惑、愉快和厌倦3种情感进行识别。根据表情变化在视频中的空间局部性和时间阶段性

到稿日期:2013-07-13 返修日期:2013-09-30 本文受国家自然基金(61203259,61103074),天津市自然基金(11JCYBJC00600)资助。

王金伟(1980—),男,博士生,讲师,主要研究方向为情感识别、并行计算,E-mail:wangjinwei@tju.edu.cn;马希荣(1962—),女,博士,教授,主要研究方向为情感计算;孙济洲(1949—),男,博士,教授,主要研究方向为模式识别、并行计算。

特点，本方法对 SISA 模型的训练和识别流程做了两点改进：(1)采用预处理后的视频替代原始视频进行训练和识别，即检测视频中的人脸位置，并对其进行几何规范化和直方图均衡化。(2)采用顺序方式替代随机方式抽取样本中的局部视频块进行特征学习。通过对比实验表明，本文的方法较其他人工特征方法，能够获得更高的识别率。

2 基于 SISA 模型的无监督特征提取

2.1 ISA 模型

ISA 是无监督的图像特征学习方法，它模拟了人类视觉系统 V1 区简单细胞与复杂细胞感受野的响应模式，它假设每个复杂细胞都接收一组简单细胞的数据，由此构成子空间的概念。

如图 1 所示，ISA 模型是一个双层结构，其中第一层单元模拟简单细胞，第二层单元模拟复杂细胞，模型的输入是图像块。第一层对输入进行平方变换，第二层对第一层的输出进行开方变换。输入和第一层之间的权重 W 需要通过学习获得，第一层和第二层网络之间的权重 V 一般是固定的。可以看到图中 ISA 第二层的每个单元和第一层的两个单元相连，即一个复杂细胞连接两个简单细胞。因此该 ISA 模型的子空间大小为 2。

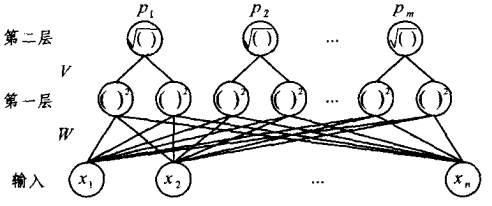

图 1 ISA 模型双层结构

更进一步，可以设 ISA 模型的输入为 x^t，则通过两层的变换，第二层的输出为：

$$p_i(x^t;W,V)=\sqrt{\sum_{k=1}^{m}V_{ik}(\sum_{j=1}^{n}W_{kj}x_j^t)^2} \tag{1}$$

其中，$\{x^t\}_{t=1}^{T}$ 是经过白化后的输入向量，即对图像块进行了去均值和协方差单位化。$W\in\mathcal{R}^{k\times n}$ 是第一层的权值矩阵，也是 ISA 的变换矩阵，$V\in\mathcal{R}^{m\times k}$ 是第一层和第二层之间的权值矩阵，n、k 和 m 分别是输入向量维度、第一层单元数和第二层单元数。

ISA 通过求解下面的最小化问题来学习权值矩阵 W。

$$\min_{W}\sum_{t=1}^{T}\sum_{i=1}^{m}p_i(x^t;W,V)\quad \text{s.t. } WW^{\mathrm{T}}=I \tag{2}$$

其中，W 的正交化限制保证了特征的多样化。

训练 ISA 模型需要使用梯度下降法，但该方法每一步都要计算正交化，计算正交化的时间复杂度是立方级。因此，如果直接把 ISA 模型应用于视频数据，输入向量的维度会剧增，模型的训练速度将变得非常慢。

2.2 SISA 模型

SISA 模型通过堆叠和卷积技术，大大加快了模型的训练速度，其主要思想是：先用大量小样本块训练一个 ISA 用于提取特征，然后再用卷积和堆叠技术构建更大样本块上的模型。图 2 是一个两层的 SISA 模型，第一层由若干 ISA 组成，第二层由一个 ISA 构成。其具体构造过程如下：

第 1 步　以非监督学习方式用大量的无标签小样本块训练一个 ISA，该 ISA 能够提取出小块的特征。

第 2 步　将训练好的 ISA 进行“复制”和“粘贴”，得到若干相同的 ISA，这些 ISA 组成 SISA 模型的第一层。

第 3 步　在已构建的模型的第一层上面再堆叠一个新的 ISA，作为模型的第二层。

第 4 步　用无标签的大样本块训练第二层的 ISA。方法是，从大样本块中抽取若干小块，每个小块作为第一层中一个 ISA 的输入。通过这些小块的重叠实现小样本块 ISA 模型和大样本块的卷积。然后将第一层的输出作为第二层的输入，对第二层的 ISA 进行无监督学习。最终得到一个两层 SISA 模型。

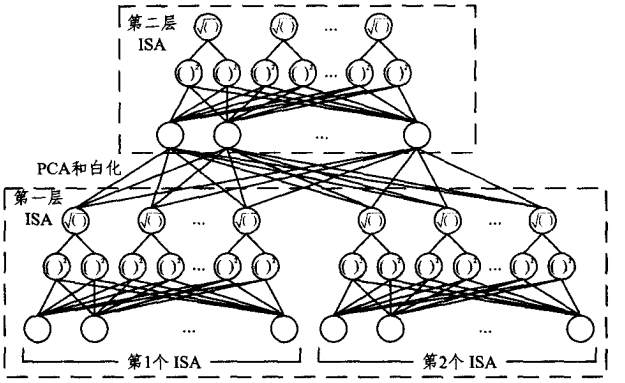

图 2 两层的 SISA 模型

在以上过程中，原始的输入数据和第一层的输出数据都要先经过白化和 PCA 降维再进入下一层网络。

SISA 模型的构造过程实际上是深度学习的逐层贪婪训练法。每一层的训练都采用了梯度下降算法，直至收敛。为了加快训练速度，SISA 模型不是一次性输入所有数据，而是截取其中的一个个小块作为训练样本，然后通过堆叠和卷积技术缓慢过渡到大样本块，再加上 PCA 降维，最终使训练时间减少到几个小时。

3 无监督提取时空特征的情感识别方法

通过对 SISA 模型的研究，本文提出了一个基于采用 SISA 模型的无监督提取表情时空特征的情感识别方法。该方法首先从视频片段中抽取时空三维样本块对一个双层 SISA 模型进行无监督训练，得到能够提取表情时空特征的特征提取器。然后再用该特征提取器组合支持向量机(SVM)对疑惑、愉快和厌倦 3 种情感的视频片段进行分类识别。

在最初的设计中，我们发现如果直接用原始视频对 SISA 模型进行训练和识别，识别率较低。原因是视频中的动态变化部分具有空间局部性，即表情的变化只集中在人的面部，头发、耳朵和身体的其他部分以及背景图像基本上是不动的，而 SISA 模型本身并不能自动识别视频中变化的局部区域并对其进行抽样。因此，本方法先对原始视频进行预处理，即检测视频每帧图像中人脸的位置，并对其进行几何规范化和直方图均衡化，然后再作为输入对 SISA 模型进行训练。这样大大提高了所抽取样本块的有效性。另外，大部分表情的变化过程具有顺序性和阶段性，即先后经过起始、发展、高潮和结束等阶段。因此，为了使学习到的特征反映出这种特点，本文的方法采用时空上的顺序方式替代原先的随机方式来抽取视

频中的局部块进行特征学习。

图 3 是本文方法的概要图，其具体步骤如下：

第 1 步　视频预处理。通过 Viola-Jones 物体检测算法，先检测视频中每一帧图像上的人脸，然后在人脸的上半区检测双眼的位置。依据双眼的坐标，对原图像进行旋转、剪切和缩放，得到人脸的几何规范化图像。对规范化的人脸图像再进行直方图均衡化，提高图像的对比度和动态范围。

第 2 步　特征提取。用双层 SISA 模型对以上预处理后的视频进行特征提取。SISA 模型第一层中每个 ISA 的输入都是按顺序抽取的小视频块扁平化后的向量，第二层的 ISA 以第一层所有 ISA 的输出作为输入，最后得到时空特征。为了提高后续的识别精度，将 SISA 第一层的输出经过 PCA 降维后与第二层的输出合并，作为最后的特征向量。

第 3 步　情感识别。将 SISA 模型输出的特征向量输入到 SVM 分类器中对视频进行最终的分类识别。

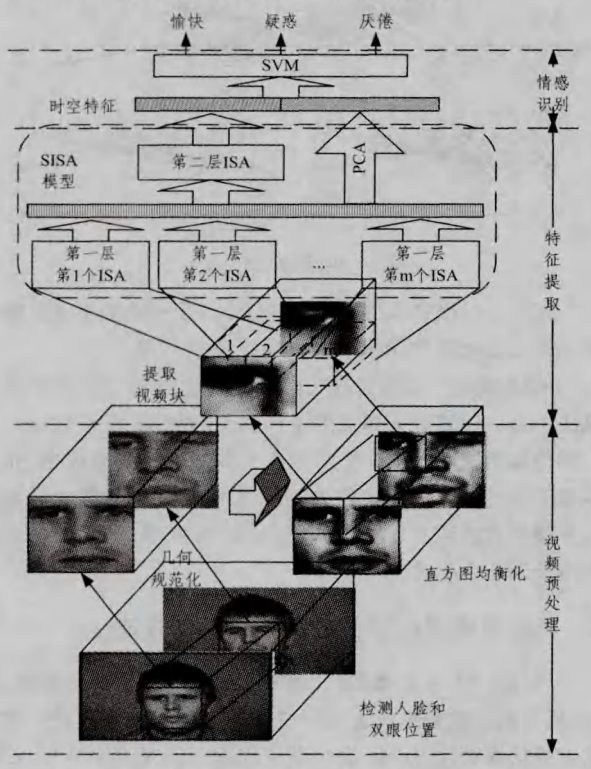

图 3　本文的情感识别方法

4　实验

4.1　实验数据集

本文采用美国德州大学的 Moving Faces and People 视频数据库[8]作为实验对象，该数据库总容量为 160GB，包含了 284 个成人（男性 76 人，女性 208 人）的面部表情、语音交谈以及步态的视频片段。这些受试者涵盖了不同的种族和年龄。不同于其他数据库中受试者表演出的 6 种基本情感[9]，该数据库中的情感是由受试者通过观看一段有声视频而自然产生，与实际学习生活中由视觉和听觉激发情感的方式相同，因此更加复杂和精细。本文从数据库里受试者表现出愉快、疑惑和厌倦 3 种情感的面部表情视频片段中，每种随机抽取了 100 段，共 300 段作为训练和测试数据集，来测试上文提出的情感识别方法的性能。每个视频片段的分辨率都为 720×480，时长约 5 秒，每秒 29.97 帧。图 4 是其中的两个示例。在本文的实验中，所有片段都经过图 3 方法中的视频预处理，最终被转换成 80×80 的规范化灰度人脸视频片段，每段 140 帧。

(a) 疑惑

(b) 厌倦

图 4　实验数据集示例

4.2　实验设置和结果分析

对于 SISA 模型，本文采用 16×16×10 的小视频块作为输入，其中时间维度为 10，设输出特征数为 300，对第一层的 ISA 进行训练；采用 20×20×14 的大视频块，以 4 作为时空卷积步长，可分成 8 个 16×16×10 的小视频块，设输出特征数为 100，对第二层进行训练。因此，最终得到一个第一层 8 个 ISA、第二层 1 个 ISA 的 SISA 模型。在提取特征时，每个视频片段可以分成 160 个大视频块，通过训练好的 SISA 模型，每个视频块对应输出 200 个特征，其中包括第二层输出的 100 个特征和第一层输出经 PCA 降维后得到的 100 个特征，最终得到一个 32000(160×200)维的特征向量。

为了便于比较，本文还采用了 HOG/HOF 和 HOG3D 两种人工特征方法对数据集进行特征提取和情感识别。这两种方法首先使用 Dense 采样，以 8 和 3.5 为时空块因子，50%为时空重叠率，从每个视频片段中获取了 1050 个兴趣点，然后对这些兴趣点区域进行计算，最终得到 170100 维的 HOG/HOF 特征向量和 60000 维的 HOG3D 特征向量。

根据文献[10]的建议，本文使用线性 SVM 作为分类器。因为 SVM 本身是二值分类器，为了多类识别，实验采用 one-versus-rest 技术，即先为愉快、疑惑和厌倦每个情感训练一个二值分类器，然后选择具有最高决策值的类作为识别结果。实验采用 5-fold 交叉验证，表 1—表 3 是 3 种方法平均识别率的混淆矩阵。

表 1　本文方法的混淆矩阵

	愉快(%)	疑惑(%)	厌倦(%)
愉快	**90**	5	5
疑惑	6	**83**	11
厌倦	8	18	**74**

表 2　Dense+HOG/HOF 方法的混淆矩阵

	愉快(%)	疑惑(%)	厌倦(%)
愉快	**87**	6	7
疑惑	6	**82**	12
厌倦	5	28	**67**

表 3　Dense+HOG3D 方法的混淆矩阵

	愉快(%)	疑惑(%)	厌倦(%)
愉快	**80**	13	7
疑惑	12	**74**	14
厌倦	5	28	**67**

对比3个混淆矩阵，可以看出，与两种人工特征方法相比，本文所采用的无监督特征学习方法对3种情感具有最高的识别率，HOG/HOF方法次之，HOG3D方法的识别率最低。其中，本文方法对厌倦的识别率较其他两种方法的提高较为显著。对疑惑的识别率，HOG/HOF方法接近于本文方法，而HOG3D方法的识别率相对较低。

对比3种情感的识别率，可以发现其中愉快的识别率最高，3种方法都达到或超过了80%，本文方法达到了90%；疑惑的识别率次之；厌倦的识别率最低，除本文方法外，其他两种方法的识别率都低于70%，容易和另外两种情感混淆。通过分析视频片段，发现愉快情感的面部表情相对单一，主要是笑脸，较容易识别；疑惑的面部动作主要是眉毛的靠近和眼睑的缩紧，部分受试者的表现非常微小，识别起来有一定难度；而厌倦的面部动作较为多样，包括眨眼、动嘴、上睑下降和打哈欠，还有部分表现也非常微小，因此给识别增加了很大难度。

表4是3种方法的平均识别率和特征向量维度，从中可以看到，本文方法的特征向量维度最低，而识别率最高。相比之下，HOG/HOF方法的特征向量维度高达17万，但没有产生较高的识别率。

表4　3种方法的平均识别率和特征向量维度

方法	平均识别率(%)	特征向量维度(万)
本文方法	82.3	3.2
Dense+HOG/HOF	79.0	17.0
Dense+HOG3D	73.7	6.0

同时还进行了算法耗时实验，在配置为AMD Athlon IIX255，3.1 GHz/8 GB的计算机上用Matlab进行编程测试。从预处理、提取特征到分类识别，本文的方法平均耗时0.004秒/帧，合250帧/秒。这表明，该方法完全能满足实时视频中情感识别的要求。

结束语　本文提出了一种基于无监督提取表情时空特征的情感识别方法，来对愉快、疑惑和厌倦3种情感进行识别。该方法首先对表情视频片段进行规范化预处理，然后采用SISA模型从中学习和提取表情的时空特征，最后用线性SVM进行情感分类。通过实验验证，本文的方法不仅识别速度快，而且相比其他两种人工特征方法，在相对低的特征空间维度下也能有效地提取出人脸表情的时空特征，从而获得较高的识别率。

参考文献

[1] McDaniel B T, D'Mello S K, King B G, et al. Facial features for affective state detection in learning environments[C]// Proceedings of the 29th Annual Cognitive Science Society Conference, 2007. Nashville, TX, USA, Cognitive Science Society, 2007: 467-472

[2] Dahmane M, Meunier J. Emotion recognition using dynamic grid-based hog features[C]// Proceedings of IEEE International Conference and Workshop on Automatic Face and Gesture Recognition, 2011. IEEE, Santa Barbara, CA, USA, 2011: 884-888

[3] Song Y, Morency L P, Davis R. Learning a sparse codebook of facial and body microexpressions for emotion recognition[C]// Proceedings of the 15th ACM on International conference on multimodal interaction, 2013. Sydney, Australia, ACM, 2013: 237-244

[4] Hayat M, Bennamoun M, El-Sallam A. Evaluation of spatiotemporal detectors and descriptors for facial expression recognition [C]// Proceedings of IEEE 5th International Conference on Human System Interactions, 2012. IEEE, Perth, West Australia, 2012: 43-47

[5] Schmidt E M, Kim Y E. Learning emotion-based acoustic features with deep belief networks [C] // Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2011. IEEE, New Paltz, NY, USA, 2011: 65-68

[6] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion [J]. The Journal of Machine Learning Research, 2010, 11: 3371-3408

[7] Le Q V, Zou W Y, Yeung S Y, et al. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011. IEEE, Colorado Springs, CO, USA, 2011: 3361-3368

[8] O'Toole A J, Harms J, Snow S L, et al. A video database of moving faces and people [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 812-816

[9] 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]// Proceedings of IEEE Workshops on Computer Vision and Pattern Recognition, 2010. IEEE, San Francisco, CA, USA, 2010: 94-101

[10] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification [J]. The Journal of Machine Learning Research, 2008, 9: 1871-1874

(上接第262页)

[13] Emerson E A, Mok A K, Sistla A P, et al. Quantitative temporal reasoning [C] // Computer-Aided Verification. Berlin Heidelberg: Springer, 1991: 136-145

[14] Alur R, Dill D L. A theory of timed automata[J]. Theoretical computer science, 1994, 126(2): 183-235

[15] Alur R. Timed automata[C]// Computer Aided Verification. Berlin Heidelberg: Springer, 1999: 8-22

[16] Alur R. Techniques for automatic verification of real-time systems[D]. Stanford University, 1991

[17] Alur R, Courcoubetis C, Dill D. Model-checking for real-time systems[C]// Logic in Computer Science, 1990. LICS'90, Proceedings. , Fifth Annual IEEE Symposium on e. IEEE, 1990: 414-425

[18] 钱俊彦，赵岭忠，古天龙. 一种基于时间自动机的时钟等价性优化方法[J]. 计算机工程，2005，9(18)：71-73

[19] Alur R, Courcoubetis C, Dill D. Model-checking in dense real-time[J]. Information and Computation, 1993, 104(1): 2-34