1.专利 基于鉴别局部保持投影的多模态情感特征融合方法.pdf

基于鉴别局部保持投影的多模态情感特征融合方法，该方法首先对多模态情感数据库中每种模态的样本数据提取情感特征，如语音特征、表情特征、姿态特征等，然后采用鉴别局部保持投影方法将各种模态的情感特征映射到统一的鉴别子空间中，最后将映射后的多组特征进行串联融合，得到融合的多模态情感特征。以融合的多模态情感特征作为输入的分类器能够有效地识别出生气、反感、害怕、高兴、悲伤和惊讶等基本情感，为开发人类情感分类识别系统以及人机交互的实现提供了一种新的方法和途径。

# 徐嵚嵛, 卢官明, 闫静杰. 基于鉴别局部保持投影的多模态情感特征融合方法:, CN 106096642 A[P]. 2016.

==============================================================================

2.专利 一种基于遗传算法的多模态情感特征融合方法.pdf

本发明公开了一种基于遗传算法的多模态情感特征融合方法:建立一个多模态情感数据库，然后对于所述数据库中的每个样本，提取多模态情感数据库样本的各模态情感特征，构造多模态情感特征矩阵，将遗传算法用于多个模态的特征融合，包括基于遗传算法的特征选择、交叉以及重组，最后采用遗传算法对多模态情感特征进行F次迭代的特征选择和融合。

# 程晓, 卢官明, 闫静杰. 一种基于遗传算法的多模态情感特征融合方法:, CN 106096641 A[P]. 2016.

==============================================================================

3.对所提取视频镜头的多模态底层特征,根据共生数据嵌入和相似度融合进行多模态子空间相关性传递而得到镜头之间的相似度关系,接着通过局部不变投影对原始数据进行降维以获得低维语义空间内的坐标,再利用标注信息训练分类模型,从而可对训练集外的测试数据进行语义概念检测,实现视频语义信息挖掘。实验表明这一方法有较高的准确率。

# 刘亚楠. 多模态特征融合和变量选择的视频语义理解[D]. 浙江大学, 2010.

==============================================================================

4.一种基于典型关联分析(Canonical Correlation Analysis，CCA)的特征融合方法，并将该算法融入到标准的Corr-AE模型结构中，提出了一种改进的Corr-AE模型。实验表明，改进模型比标准Corr-AE模型具有更好的跨模态检索性能:

# 臧虎. 基于深层网络的多模态特征融合问题的研究与应用[D]. 北京邮电大学, 2014.

==============================================================================

5.为了提高情感识别的分类准确率,提出一种将栈式自编码神经网络(SAE)和长短周期记忆单元循环神经网络(LSTM RNN)融合的多模态融合特征情感识别方法.该方法通过 SAE 对不同模态的生理特征进行信息融合和压缩,随后用LSTM RNN对长时间周期的融合进行情感分类识别.通过将该方法用到开源数据集中进行验证,得到情感分类准确率达到0.7926.实验结果表明,SAE对多模态生理特征进行了有效融合,LSTM RNN能够有效地对长时间周期中的关键特征进行识别.

# 李幼军, 黄佳进, 王海渊,等. 基于SAE和LSTM RNN的多模态生理信号融合和情感识别研究[J]. 通信学报, 2017(12).

==============================================================================

6.本文提出一种基于基准点与特征点连线的手背静脉识别算法。该方法在对手背静脉图像进行预处理后，提取静脉的交叉点和端点作为特征点，用特征点之间的相对距离确定图像匹配的基准点，采用基准点和特征点之间的相对距离以及基准点和特征点相邻连线的夹角作为识别特征，并将这两种特征融合，进行手背静脉识别。该算法具有位移旋转不变性,对手背摆放位置不敏感。实验结果证明了该方法的有效性。

# 李飞. 基于手部生物特征多模态识别算法研究[D]. 天津理工大学, 2013.

7. 重点介绍了多模态医学图像融合方法,包括加权平均法、多分辨金字塔法、小波变换法等常用技术,并分析了当前该领域存在的技术难点以及发展前景。

# 谈泉, 林家瑞. 多模态医学图像融合技术的研究与进展[J]. 国际生物医学工程杂志, 2006, 29(3):158-160.

**8.** 本文针对基于利用单一特征对图像检索的片面性,进行了基于综合颜色和纹理特征的图像检索。仅基于一种特征的方法只能表达图像的部分属性,对图像内容的描述比较片面。颜色特征充分利用了图像的色彩信息,侧重于图像整体信息的描述,纹理特征只利用了图像的灰度信息,更偏重于局部。在本文中,统计颜色直方图的方法不同于通常的颜色直方图的统计方法,纹理特征的表达与一般的表达方法也有差别。

# 邓金杰, 肖诗斌, 吕学强,等. 基于多特征融合的图像检索研究[C]// 图像图形技术与应用学术会议. 2009.

**9.** 提出一种多模态融合的情感识别算法，从面部图像序列和语音信号中提取表情和语音特征，基于隐马尔可夫模型和多层感知器设计融合表情和语音模态的情感分类器。建立面部表情图像的主动外观模型，实现面部特征点的定位和跟踪；根据面部特征点的位移，计算面部动画参数作为表情特征。对语音信号作时域、和频域分析，提取各帧的短时平均能量、基音频率和共振峰作为语音特征。利用提取的表情和语音特征，采用Viterbi算法训练各种表情和语音情感的隐马尔可夫模型；利用特征向量关于各隐马尔可夫模型的条件概率，采用反向传播学习算法训练多层感知器。实验结果表明，融合表情和语音的情感识别算法在识别样本中的高兴、悲伤、愤怒、厌恶等情感状态时具有较高的准确率。 提出的多模态识别算法较好地利用了视频和音频中的情感信息，相比于仅利用语音模态的识别结果有较大的提升，相比于表情模态的识别结果也有一定改进，是一种可以采用的情感识别算法。

# 曹田熠. 多模态融合的情感识别研究[D]. 天津大学, 2012.

**10.** 本文研究了多模态生物特征识别的特征层、分数层融合算法，在开集测试集上测试各种算法的性能，涉及到的生物特征有人脸、掌纹、虹膜三种。 本文的主要研究成果总结如下： 1 为了克服典型相关分析(CCA)在信息融合应用中的不足，本文提出了一种有监督的学习方法。增强相关分析(ECA)，进而利用核技巧提出了核化的增强相关分析(KECA)，并把ECA和KECA用于多模态生物特征的特征层融合。与CCA相比，ECA有效的利用了类别信息，也适用于有样本缺失的情形。开集测试证明，采用ECA、KECA进行特征层融合时，系统的性能(等错误率和正确接收率)较CCA有所提高，且KECA的性能高于ECA。 2 设计了一种新颖的近红外人脸、双眼虹膜图像采集设备，构建了国内首个包含近红外人脸、左眼虹膜、右眼虹膜这三种模态的包含噪声样本的多模态数据库。在该库基础上，研究了上述三种模态的融合方法。首先，首次提出了基于联合虹膜编码的双眼虹膜特征层融合算法，其性能高于双眼虹膜分数层融合的性能。其次，采用混合融合结构将三种模态进行融合，即双眼虹膜进行特征层融合后再与近红外人脸进行分数层融合，实验表明其性能高于把三种模态进行分数层融合的方法。 3 提出了基于最小二乘法的分数层融合算法(LSMSF)。该算法用最小二乘法估计融合函数的参数，融合函数有三种形式：幂级数函数、多变量多项式函数、简化的多变量多项式函数。采用交叉验证的方法全面评估了LSMSF和传统分数层融合算法的性能，评估包括：单模态性能、分数归一化方法、数据库训练集和测试集的不同划分对算法性能的影响。实验表明，LSMSF的性能均高于传统分数层融合算法，采用多变量多项式函数的LSMSF的性能最好。

# 10.张志坚. 多模态生物特征识别融合算法的研究[D]. 中国科学技术大学, 2008.

**11.** 提出了一种融合音频、字幕以及视觉等多模态信息的新闻故事单元分割算法。创新性地将字幕变化、音频类型以及镜头类型等高层次内容特征联系起来共同处理，巧妙地将新闻镜头序列转换成为多个关键词序列，使新闻故事单元分割问题转换成为文本序列分割的问题。该算法采用条件随机场进行建模，充分利用了每个序列内以及序列之间的上下文信息，得到了较好的分割性能。 此外，论文还综述了视频内容分析技术，构造了一个基于规则和隐马尔可夫模型的分层音频分类方法，实现了一个较完整的新闻视频中字幕提取框架，最终设计并实现了一个基于COM架构的视频内容分析与摘要系统。 综上所述，本论文分别从音频、字幕、视觉以及它们之间的有效融合等方面对新闻视频进行了基于内容的分析，实验结果证明了这些算法的有效性。

# 冀中. 基于多模态信息的新闻视频内容分析技术研究[D]. 天津大学, 2007.

**12.** 设计了一个应用多模态数据融合的车载防撞预警系统，并在eclipse环境下进行了编程实现与测试,在多模态数据融合方面做了有益的研究工作。

# 戎翔. 多模态数据融合的研究[D]. 南京邮电大学, 2012.

**13.** 提出一种有效地融合多模态信息来检测足球视频精彩事件的系统算法和框架.首先从视频中抽取音频流,然后基于CHMM进行音频分类.接着根据时间对应关系在包含激昂解说音和欢呼声的相邻镜头里结合球门和慢镜头检测射门事件,其中慢镜头检测是基于徽标的.对射门事件进一步根据激昂解说音和欢呼声的长短、慢镜头的长短及比分字幕的出现检测进球事件.在哨音出现的相邻镜头中结合是否有慢镜头回放及回放长度来检测犯规事件.实验表明,提出的系统算法及框架是高效率的.

# 张玉珍, 魏带娣, 王建宇,等. 基于多模态融合的足球视频语义分析[J]. 计算机科学, 2010, 37(7):273-276.

**14.** 如何跨越低层特征描述到高层语义知识的"语义鸿沟"已成为跨媒体检索(CMR)问题的关键,提出一个基于多模态融合描述和时空上下文语义的跨媒体检索模型,对多模态融合的特征采用主成分分析(PCA)和独立成分分析(ICA)相结合的降维算法、采用基于支持向量机(SVM)和隐马尔可夫模型(HMM)的混合分类器进行语义映射,同时给出了时空模糊聚类分析方法和基于相关反馈的跨媒体检索算法。并在此基础上开发出基于该模型的原型系统,成功验证了该模型的可行性和正确性,可为相关系统的设计者提供思路。

# 刘扬, 郑逢斌, 姜保庆,等. 基于多模态融合和时空上下文语义的跨媒体检索模型的研究[J]. 计算机应用, 2009, 29(4):1182-1187.

15. 研究和改进了基于小波变换的图像融合算法,提出了一种改进小波变换金字塔融合算法.该算法对经过多层小波变换后的高低频小波系数,分别使用局部均值加权和分层线性加权的小波系数融合规则进行融合.通过编程试验,将该算法与空域和频域的各种融合算法进行对比,并使用主观和客观两种评价方法进行综合评价.实验结果表明,针对灰度图像特别是医学图像的融合,提出的融合算法与文中的对比算法相比较,在主观和客观的综合评价上表现出更好的融合效果.

# 戴峻峰. 基于小波变换的多模态图像融合算法的改进[J]. 计算机工程与设计, 2010, 31(19):4250-4252.

16. A novel hierarchical image fusion scheme based on wavelet multiscale decomposition is presented. The basic idea is to perform a wavelet multiscale decomposition of each source image first, then the wavelet coefficients of the fused image are constructed using region-based selection and weighted operators according to different fusion rules, finally the fused image is obtained by taking inverse wavelet transform. This approach is successfully used in image fusion. In addition, with the use of the parameters such as entropy, cross entropy, mutual information, root mean square error, peak-to-peak signal-to-noise ratio, the performance of the fusion scheme is evaluated and analyzed. The experimental results show that the fusion scheme is effective.

# Liu G. A WAVELET-DECOMPOSITION-BASED IMAGE FUSION SCHEME AND ITS PERFORMANCE EVALUATION[J]. 2002, 28(6):927-934.

17. This article presents a wavelet transformation based multi-mode medical image fusion algorithm which combined with the edge characteristics of sub-image, making wavelet transformation on multi-source medical image to be integrated firstly, and then set up appropriate fusion operator to make integration according to edge feature of sub-images transformed and human eyes' different sensitivity on images in HVS, and reconstruct fusion image through inverse transformation at last. Tested by the integration experiment on brain MRI-PET images, it is proved that this method can combine anatomical information and functional information together more effectively, and retain the edge characteristics of original image better.

# Zhang X, Zheng Y, Peng Y, et al. Research on Multi-Mode Medical Image Fusion Algorithm Based on Wavelet Transform and the Edge Characteristics of Images[C]// International Congress on Image and Signal Processing. IEEE, 2009:1-4.

18.提出多模态特征层融合的情感识别方法。在多模态特征融合方法中使用了如上介绍的特征多样性特性计算方法对多模态特征进行选择和融合。我们在特征层融合中测试了多种细粒度方法的组合方式，同时，我们也将提出的特征层融合方法与其它特征选择方法进行了对比。在特征层融合方法中，当我们使用细粒度方法的组合分别对情感两个维度（即，效价维和愉悦度）进行二元分类时，发现方法组合Fisher Criterion Score（特征排序方法）+在等级-评分函数上计算的多样性特性+多层感知器（分类器）获得了最好的识别效果。通过将多模态特征层融合方法与其它特征选择方法进行比较，我们发现特征层融合方法在选择了合适的细粒度方法组合时能够获得比这些对比方法更好的识别结果。

提出决策层融合的情感识别方法。在决策层融合方法中，首先使用分类器对单模态数据进行分类，然后将单模态数据上获得的所有分类输出进行融合得到最后的情感决策结果。我们不仅在决策层融合方法中测试了多种细粒度方法组合方式，也探索了多种决策层融合算法。实验结果表明，决策层融合方法能够获得比单模态更好的情感识别效果。

# 陈婧. 情感识别中的多模态信息融合方法研究[D]. 兰州大学, 2016.

19. 一种基于图像和生理信号的多模态特征融合情感识别方法。利用奇异值分解(singular value decomposition,SVD)方法和小波分解法分别对图像信息和生理信号进行特征提取,然后采用主成分分析(principal component analysis,PCA)方法对多模态特征进行降维融合,将反向传播(back propagation,BP)神经网络作为分类器,对不同情感进行分类识别。情感诱发试验结果表明,该方法能有效提高情感识别的正确率。

# 杨亚龙, 汪明月, 方潜生. 基于图像与生理参数特征融合的情感识别方法研究[J]. 合肥工业大学学报:自然科学版, 2017, 40(6):747-751.

20. 本文基于数据融合理论,利用非负矩阵分解理论和小波分析方法对MRI和MRSI数据融合问题进行研究,提出一种非监督的多模态数据融合方法

# 20.刘鑫. 基于矩阵分解的MRI和MRSI数据融合方法研究[D]. 电子科技大学, 2015.

21. 首先对预处理后的语音信号提取梅尔倒谱系数特征,并用隐马尔科夫构建语音情感识别模型,然后计算脉搏信号K值和小波包系数能量值,并输入到最小二乘支持向量机识别模型中进行判别,最后对两个模型的判别结果进行决策级的融合.实验结果表明:对于哀伤和平静两种情感,语音识别率较高,融合后识别率达到100%;对于高兴和愤怒两种情感,语音识别率为75%和80%,融合后提高到95%和90%.

# 周红标. 融合语音和脉搏的多模态情感识别研究[J]. 微电子学与计算机, 2015(6):5-9.

22. 在决策层，本文将时序建模与多模态决策层融合相结合，提出了一种多模态时序融合方法。该方法通过同时融合来自多组特征以及各组特征不同时刻的预测结果，实现了更长跨度的时序建模，并与特征层的时序建模实现相互补充。本文所提出的多尺度时序建模方法在2014年举办的音视频融合的情感识别公开评测（AVEC2014）中取得了评测第二名的成绩。

# 巢林林. 音视频融合的情感识别技术研究[D]. 中国科学院大学, 2016.

23. 视频分类算法通过融合多模态特征提高分类性能.大多数多模态特征融合算法采用统一模式处理不同模态特征,忽略文本与声音/图像特征在视频分类能力和可靠性方面的差异.该文提出一种新闻视频分类算法,采用文本为主,声音/图像为辅的特征融合模式.该算法充分考虑文本特征在内容分类方面的高可靠性,并在适当情况下,辅以声音/图像信息为补充.对10类新闻视频进行分类实验,并与其他两种分类算法对比,表明在提取相同的底层特征情况下,该文算法的平均分类性能最佳.

# 王鹏, 蔡锐, 杨士强. "文本为主"的多模态特征融合的新闻视频分类算法[J]. 清华大学学报(自然科学版), 2005, 45(4):475-478.

24. 针对非平稳时间序列信号,提出一种基于经验模态分解（EMD）的特征提取和多模态特征融合支持向量机的故障诊断方法.首先对原始信号进行EMD分解,选择能量最大的几个基本模式分量（IMF）并提取其小波包特征,获得对每个IMF独立的特征子集;然后在每个IMF特征子集中训练SVM弱分类器,并根据各特征子集对应的IMF能量权重进行加权融合,获得故障状态的强分类器.将该方法应用于6135型柴油机振动信号故障诊断中,实验结果表明了其可行性和有效性.

# 沈志熙, 黄席樾, 马笑潇. 基于EMD多模态特征融合支持向量机的故障诊断[J]. 控制与决策, 2009, 24(6):889-893.

25. 利用加权融合和特征空间变换的方法分别对判决层和特征层进行融合,并比较了这2种融合算法在语音信号与心电信号融合情感识别中的性能.实验结果表明:在相同测试条件下,基于心电信号和基于语音信号的单模态情感分类器获得的平均识别率分别为71%和80%;通过特征层融合,多模态分类器的识别率则达到90%以上;特征层融合算法的平均识别率高于判决层融合算法。因此，依据语音信号、心电信号等不同来源的情感特征可以构建出可靠的情感识别系统

# 黄程韦，金赟，王青云，赵力， 邹采荣. 基于语音信号与心电信号的多模态情感识别[J].東南大學學報（自然科學版）,2010,40(5):895 – 900

26. 比较研究了多模态人脸识别中的5种匹配得分级融合方法。首先用局部二值模式(Local Binary Pattern,LBP)算子分别提取人脸灰度图像和深度图像的区域LBP直方图序列(LBP Histogram Sequence,LBPHS),采用Fisherfaces分别构建相应的线性子空间,用余弦相似度计算投影向量的匹配得分,再采用5种方法对匹配得分进行融合。在FRGC数据库上的实验结果表明,除最小匹配得分外,其他融合方法的识别性能都要优于单一模态的方法。

# 叶剑华, 刘正光. 多模态人脸识别融合方法比较研究[J]. 计算机工程与应用, 2009, 45(19):153-156.

27.本文是一个调研型论文，现有的多模态融合的研究主要通过几种基于融合方法和融合层（特征层，决策层，融合层）次的分类方法来呈现

# Atrey P K, Hossain M A, Saddik A E, et al. Multimodal fusion for multimedia analysis: a survey[J]. Multimedia Systems, 2010, 16(6):345-379.

28. Multimodal streams of sensory information are naturally parsed and integrated by humans using signal-level feature extraction and higher level cognitive processes. Detection of attention-invoking audiovisual segments is formulated in this work on the basis of saliency models for the audio, visual, and textual information conveyed in a video stream.

# Evangelopoulos G, Zlatintsi A, Potamianos A, et al. Multimodal Saliency and Fusion for Movie Summarization Based on Aural, Visual, and Textual Attention[J]. IEEE Transactions on Multimedia, 2013, 15(7):1553-1568.

**29.** In this paper, we propose a two-step approach. The first step finds statistically independent modalities from raw features (e.g., extracted from the visual, audio, and caption track of videos). In the second step, we use super-kernel fusion to determine the optimal combination of individual modalities. We carefully analyze the tradeoffs between three design factors that affect fusion performance: modality independence, curse of dimensionality, and fusion-model complexity. Through analytical and empirical studies, we demonstrate that our two-step approach, which achieves a careful balance of the three design factors, can improve class-prediction accuracy over traditional techniques.

# Wu Y, Chang E Y, Chang C C, et al. Optimal multimodal fusion for multimedia data analysis[C]// ACM International Conference on Multimedia, New York, Ny, Usa, October. DBLP, 2004:572-579.

30.Hierarchical patch-level information fusion via an ensemble classifier。Maximal diagnostic accuracies of 93.52% (AD vs. NC), 85.19% (MCI vs. NC), and 74.58% (MCI converter vs. MCI non-converter)

# 30.Suk H I, Lee S W, Shen D. Hierarchical Feature Representation and Multimodal Fusion with Deep Learning for AD/MCI Diagnosis[J]. Neuroimage, 2014, 101:569.

31.This work investigated several statistical approaches built upon global/local visual features, audio features, and audio-visual combinations. Three multi-modal fusion frameworks (ensemble, context fusion, and joint boosting) are also evaluated. Experiment results show that visual and audio models perform best for different sets of concepts. Both provide significant contributions to multimodal fusion, via expansion of the classifier pool for context fusion and the feature bases for feature sharing. The fused multimodal models are shown to significantly reduce the detection errors (compared to single modality models), resulting in a promising accuracy of 83% over diverse concepts. To the best of our knowledge, this is the first work on systematic investigation of multimodal classification using a large-scale ontology and realistic video corpus.

# Chang S F, Dan E, Jiang W, et al. Large-scale multimodal semantic concept detection for consumer video[C]// International Workshop on Workshop on Multimedia Information Retrieval. ACM, 2007:255-264.

32. Why we need data fusion" and "how we perform it". The first question is motivated by numerous examples in science and technology, followed by a mathematical framework that showcases some of the benefits that data fusion provides. In order to address the second question, "diversity" is introduced as a key concept, and a number of data-driven solutions based on matrix and tensor decompositions are discussed, emphasizing how they account for diversity across the datasets. The aim of this paper is to provide the reader, regardless of his or her community of origin, with a taste of the vastness of the field, the prospects and opportunities that it holds.

# Lahat D, Adali T, Jutten C. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects[J]. Proceedings of the IEEE, 2015, 103(9):1449-1477..

33. We rigorously analyze and combine a large set of low-level features that capture appearance, color, motion, audio and audio-visual co-occurrence patterns in videos. We also evaluate the utility of high-level (i.e., semantic) visual information obtained from detecting scene, object, and action concepts. Further, we exploit multimodal information by analyzing available spoken and videotext content using state-of-the-art automatic speech recognition (ASR) and videotext recognition systems. We combine these diverse features using a two-step strategy employing multiple kernel learning (MKL) and late score level fusion methods. Based on the TRECVID MED 2011 evaluations for detecting 10 events in a large benchmark set of ∼45000 videos, our system showed the best performance among the 19 international teams.

# Natarajan P, Wu S, Vitaladevuni S, et al. Multimodal feature fusion for robust event detection in web videos[C]// Computer Vision and Pattern Recognition. IEEE, 2012:1298-1305.

34. Given an online video document, which usually consists of video content and related information (such as query, title, tags, and surroundings), video recommendation is formulated as finding a list of the most relevant videos in terms of **multimodal** relevance. We express the **multimodal** relevance between two video documents as the combination of textual, visual, and aural relevance. Furthermore, since different video documents have different weights of the relevance for three modalities, we adopt relevance feedback to automatically adjust intra-weights within each modality and inter-weights among different modalities by users' click-though data, as well as attention **fusion** function to fuse **multimodal** relevance together. Unlike traditional recommenders in which a sufficient collection of users' profiles is assumed available, this proposed system is able to recommend videos without users' profiles. We conducted an extensive experiment on 20 videos searched by top 10 representative queries from more than 13k online videos, reported the effectiveness of our video recommendation system.

# Yang B, Mei T, Hua X S, et al. Online video recommendation based on multimodal fusion and relevance feedback[C]// ACM International Conference on Image and Video Retrieval. ACM, 2007:73-80.

35.In order to overcome the limitation of single mode emotion recognition,a novel multimodal emotion recognition algorithm is proposed,taking speech signal and facial expression signal as the research subjects. First,the speech signal feature and facial expression signal feature is fused, and sample sets by putting back sampling are obtained, and then sub-classifiers are acquired by Adaboost algorithm. Second, the difference is measured between two classifiers by double error difference selection strategy. Finally, the recognition result is obtained by the majority voting rule. Experiments show the method improves the accuracy of emotion recognition by giving full play to the advantages of decision level fusion and feature level fusion, and makes the whole fusion process close to human emotion recognition more, with a recognition rate 91. 2%.

# Han Z Y, Wang J, University B. Research on Feature Fusion Algorithm for Multimodal Emotion Recognition[J]. Computer Technology & Development, 2016.

37.专利

A computer-based system and method to improve the multimodal fusion output at the decision level is disclosed. The method proposes computation of a confidence weighted measure for the individual score values obtained for each modality and fuse these new updated scores to get the final decision. These confidence weights are the performance parameters (measured in terms of F-measure) during the offline training step. The process significantly increases the accuracy of the multimodal system.

# Gupta R K, Gattani S, Sinha A, et al. Method and system for association and decision fusion of multimodal inputs: US, US8700557[P]. 2014.

38. Towards this goal, a multi-step approach is followed: initially, automated audio and visual analysis is performed to extract audio and visual cues. Then, two different fusion approaches are deployed: (i) a multimodal one that provides binary decisions on the existence of violence or not, employing machine learning techniques, (ii) an ontological and reasoning one, that combines the audio-visual cues with violence and multimedia ontologies. The latter reasons out not only the existence of violence or not in a video scene, but also the type of violence (fight, screams, gunshots). Both approaches are experimentally tested, validated and compared for the binary decision problem of violence detection. Finally, results for the violence type identification are presented for the ontological fusion approach. For evaluation purposes, a large dataset of real movie data has been populated.

# Perperis T, Giannakopoulos T, Makris A, et al. Multimodal and ontology-based fusion approaches of audio and visual processing for violence detection in movies[J]. Expert Systems with Applications, 2011, 38(11):14102-14116.

39. In this paper, we investigate fusion methods for multimodal identification using several unimodal identification results. One fingerprint identification system and two face identification systems are used as fusion sources. We discuss rank level and score level fusion methods. Whereas the latter combines similarity scores, the other one combines the orders of the magnitudes of the similarity scores. For rank level methods, Borda Count and Bayes Fuse are considered and, for score level methods, Sum Rule and Binary Classification Approach are considered. Especially, we take a more detailed look at Binary Classification Approach, which simplifies a multiple class problem into a binary class problem. Finally, we compare experimental results using the fusion methods in different combinations of the sources.

# Lee Y, Lee K, Jee H, et al. Fusion for multimodal biometric identification[C]// International Conference on Audio- and Video-Based Biometric Person Authentication. Springer-Verlag, 2005:1071-1079.

40.State of the art classification methods are employed to extract audio and visual midlevel semantics. The segmentation and modality specific analysis algorithms instantiate the corresponding video structure and modality specific ontologies developed in the context of the knowledge engineering framework. A set of consecutive and interleaved ontological and SWRL rule reasoning steps map sets and sequences of extracted low to midlevel semantics into higher level concepts represented in the harmful content domain ontology. We present the involved ontologies, the corresponding SWRL rule sets and the reasoning mechanism in detail. Finally we present the evaluation of the proposed approach in a preanotated movie dataset, compare its results with the single modality approaches and a kNN late fusion meta classifier. We comment on the higher level semantics extraction ability and evaluate a set of extensions employed in the basic structure of the framework. The extensions concern the development of a scene detection module that combines markov clustering with SQWRL queries, the incorporation of existing rating and movie genre metadata in the violence identification procedure and the detection of pornography.

# 40. Thanassis Perperis , High Level Multimodal Fusion and Semantics Extraction

41. Here, we proposed a fusion with reference model called “multi-site canonical correlation analysis with reference + joint-independent component analysis” (MCCAR+jICA), which can precisely identify co-varying multimodal imaging patterns closely related to the reference, such as cognitive scores. In a three-way fusion simulation, the proposed method was compared with its alternatives on multiple facets; MCCAR+jICA outperforms others with higher estimation precision and high accuracy on identifying a target component with the right correspondence. In human imaging data, working memory performance was utilized as a reference to investigate the co-varying working memory-associated brain patterns among three modalities and how they are impaired in schizophrenia. Two independent cohorts (294 and 83 subjects respectively) were used. Similar brain maps were identified between the two cohorts along with substantial overlaps in the central executive network in fMRI, salience network in sMRI, and major white matter tracts in dMRI. These regions have been linked with working memory deficits in schizophrenia in multiple reports and MCCAR+jICA further verified them in a repeatable, joint manner, demonstrating the ability of the proposed method to identify potential neuromarkers for mental disorders.

# Qi S, Calhoun V D, Tgm V E, et al. Multimodal Fusion with Reference: Searching for Joint Neuromarkers of Working Memory Deficits in Schizophrenia.[J]. IEEE Transactions on Medical Imaging, 2018, PP(99):1-1.

42. This paper presents a multimodal fusion approach using kernel-based Extreme Learning Machine (ELM) for video emotion recognition by combing video content and electroencephalogram (EEG) signals. Firstly, several audio-based features and visual-based features are extracted from video clips and EEG features are obtained by using Wavelet Packet Decomposition (WPD). Secondly, video features are selected using Double Input Symmetrical Relevance (DISR) and EEG features are selected by Decision Tree (DT). Thirdly, multimodal fusion using kernel-based ELM is adopted for classification by combing video and EEG features at decision-level. In order to test the validity of the proposed method, we design and conduct the EEG experiment to collect data that consisted of video clips and EEG signals of subjects. We compare our method separately with single mode methods of using video content only and EEG signals only on classification accuracy. The experimental results show that the proposed fusion method produces better classification performance than those of the video emotion recognition methods which use either video content or EEG signals alone.

# Duan L, Ge H, Yang Z, et al. Multimodal Fusion Using Kernel-Based ELM for Video Emotion Recognition[J]. 2016.

43. Data generated from real world events are usually temporal and contain multimodal information such as audio, visual, depth, sensor etc. which are required to be intelligently combined for classification tasks. In this paper, we propose a novel generalized deep neural network architecture where temporal streams from multiple modalities are combined. There are total M+1 (M is the number of modalities) components in the proposed network. The first component is a novel temporally hybrid Recurrent Neural Network (RNN) that exploits the complimentary nature of the multimodal temporal information by allowing the network to learn both modality specific temporal dynamics as well as the dynamics in a multimodal feature space. M additional components are added to the network which extract discriminative but non-temporal cues from each modality. Finally, the predictions from all of these components are linearly combined using a set of automatically learned weights. We perform exhaustive experiments on three different datasets spanning four modalities. The proposed network is relatively 3.5%, 5.7% and 2% better than the best performing temporal multimodal baseline for UCF-101, CCV and Multimodal Gesture datasets respectively.

# Gandhi A, Sharma A, Biswas A, et al. GeThR-Net: A Generalized Temporally Hybrid Recurrent Neural Network for Multimodal Information Fusion[C]// European Conference on Computer Vision. Springer International Publishing, 2016:883-899.

44. This paper focuses on the application of novel information theoretic tools in the area of information fusion. Feature transformation and fusion is critical for the performance of information fusion, however the majority of the existing works depend on the second order statistics, which is only optimal for Gaussian-like distribution. In this paper, the integration of information fusion techniques and kernel entropy component analysis provides a new information theoretic tool. The fusion of features is realized using descriptor of information entropy and optimized by entropy estimation. A novel multimodal information fusion strategy of audio emotion recognition based on kernel entropy component analysis (KECA) has been presented. The effectiveness of the proposed solution is evaluated though experimentation on two audiovisual emotion databases. Experimental results show that the proposed solution outperforms the existing methods, especially when the dimension of feature space is substantially reduced. The proposed method offers general theoretical analysis which gives us an approach to implement information theory into multimedia research.

# Xie Z, Guan L. Multimodal Information Fusion of Audio Emotion Recognition Based on Kernel Entropy Component Analysis[C]// IEEE International Symposium on Multimedia. IEEE Computer Society, 2012:1-8.

45. We present a novel audiovisual emotion recognition solution using multimodal information fusion based on entropy estimation. Considering the limitations of existing methods, we propose a new dual-level fusion framework which consists of feature level fusion module based on kernel entropy component analysis and score level fusion module based on maximum correntropy criterion. In our system, audio and visual channels are utilized to detect and classify emotional states for intelligent human computer interfaces. Our extensive experimental study on eNTERFACE database and RML database demonstrates the feasibility of the proposed multimodal emotion recognition framework based on integrated analysis of speech and facial expression. The experimental results show that the proposed methods are capable of providing improved performance. The comparison with other methods shows that the proposed two-stage fusion platform outperforms the traditional algorithms in terms of both accuracy and reliability.

# Xie Z, Yun T, Guan L. A new audiovisual emotion recognition system using entropy-estimation-based multimodal information fusion[C]// IEEE International Symposium on Circuits and Systems. IEEE, 2015:726-729.

46. This paper presents an architecture for fusion of multimodal input streams for natural interaction with a humanoid robot as well as results from a user study with our system. The presented fusion architecture consists of an application independent parser of input events, and application specific rules. In the presented user study, people could interact with a robot in a kitchen scenario, using speech and gesture input. In the study, we could observe that our fusion approach is very tolerant against falsely detected pointing gestures. This is because we use speech as the main modality and pointing gestures mainly for disambiguation of objects. In the paper we also report about the temporal correlation of speech and gesture events as observed in the user study.

# Holzapfel H, Nickel K, Stiefelhagen R. Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3D pointing gestures[J]. 2004:175-182.

47. In biometric systems, the threat of “spoofing”, where an imposter will fake a biometric trait, has leaded to the increased use of multimodal biometric systems. It is assumed that an imposter must spoof all modalities in the system to be accepted. This paper looks at the cases where some but not all modalities are spoofed. The contribution of this paper is to outline a method for assessment of multimodal systems and underlying fusion algorithms. The framework for this method is described and experiments are conducted on a multimodal database of face, iris, and fingerprint match scores.

# Johnson P A, Tan B, Schuckers S. Multimodal fusion vulnerability to non-zero effort (spoof) imposters[C]// IEEE International Workshop on Information Forensics and Security. IEEE, 2010:1-5.

48. Deep networks have been successfully applied to unsupervised feature learning for single modalities (e.g., text, images or audio). In this work, we propose a novel application of deep networks to learn features over multiple modalities. We present a series of tasks for multimodal learning and show how to train deep networks that learn features to address these tasks. In particular, we demonstrate cross modality feature learning, where better features for one modality (e.g., video) can be learned if multiple modalities (e.g., audio and video) are present at feature learning time. Furthermore, we show how to learn a shared representation between modalities and evaluate it on a unique task, where the classifier is trained with audio-only data but tested with video-only data and vice-versa. Our models are validated on the CUAVE and AVLetters datasets on audio-visual speech classification, demonstrating best published visual speech classification on AVLetters and effective shared representation learning. 1.

# Ngiam, Jiquan & Khosla, Aditya & Kim, Mingyu & Nam, Juhan & Lee, Honglak & Y. Ng, Andrew. (2011). Multimodal Deep Learning. Proceedings of the 28th International Conference on Machine Learning, ICML 2011. 689-696.

49. Practical Brain-Computer Interfaces (BCIs) for disabled people should allow them to use all their remaining functionalities as control possibilities. Sometimes these people have residual activity of their muscles, most likely in the morning when they are not exhausted. In this work we fuse electromyographic (EMG) with electroencephalographic (EEG) activity in the framework of a so called “Hybrid-BCI” (hBCI) approach. Thereby, subjects could achieve a good control of their hBCI independently of their level of muscular fatigue. Furthermore, although EMG alone yields good performance, it is outperformed by the hybrid fusing of EEG and EMG. Two different fusion techniques are explored showing graceful performance degradation in the case of signal attenuation. Such a system allows a very reliable control and a smooth handover if the subjects get exhausted or fatigued during the day.

# Leeb R, Sagha H, Chavarriaga R, et al. Multimodal fusion of muscle and brain signals for a hybrid-BCI[C]// Engineering in Medicine & Biology Society. Conf Proc IEEE Eng Med Biol Soc, 2010:4343.

50.In this paper we propose a multimodal fusion framework based on novel face-voice fusion techniques for biometric person authentication and liveness verification. Checking liveness guards the system against spoof/replay attacks by ensuring that the biometric data is captured from an authorised live person. The proposed framework based on bi-modal feature fusion, cross-modal fusion as well as 3D shape and texture fusion techniques, allow a significant improvement in system performance against impostor attacks, type-1 replay attacks (still photo and pre-recorded audio), and challenging type-2 replay attacks(CG animated video from a still photo and pre-recorded audio) and robustness to pose and illumination variations.

# 50. Chetty G, Wagner M. Audio-visual multimodal fusion for biometric person authentication and liveness verification[C]// Nicta-Hcsnet Multimodal User Interaction Workshop. Australian Computer Society, Inc. 2006:17-24.

51. Video media carries multimodal information including visual, audio, textual data. Considerable research has been focused on utilizing multimodal features for better understanding of video content. However, many problems remain such as how to combine multimodal features and what are the effects of different combinations. In this paper, we propose to find the optimal combination of multimodal information in order to improve the performance of video concept detection using two methods, one is gradient-descent-optimization linear fusion and the other is super-kernel nonlinear fusion. Gradient-descent-optimization linear fusion learns an optimal weighted linear combination of single modalities based on fusing individual kernel matrices with gradient descent techniques. Super-kernel nonlinear fusion trains separate classifiers for single modalities as the first step. Once individual models have been designed, super-kernel nonlinear fusion learns an optimal nonlinear combination of individual models by fusing single-modality classifiers. Our experiments show that both methods improve performance significantly on TREC-Video 2003 benchmarks.

# Wu Y, Lin C Y, Chang E Y, et al. Multimodal information fusion for video concept detection[C]// International Conference on Image Processing. IEEE, 2004:2391-2394 Vol. 4.

52. This is a new hybrid fusion strategy based primarily on the implementation of two former and differentiated approaches to multimodal fusion [11] in multimodal dialogue systems. Both approaches, their predecessors and their respective advantages and disadvantages will be described in order to illustrate how the new strategy merges them into a more solid and coherent solution. The first strategy was largely based on Johnston's approach [5] and implies the inclusion of multimodal grammar entries and temporal constraints. The second approach implied the fusion of information coming from different channels at dialogue level. The new hybrid strategy hereby described requires the inclusion of multimodal grammar entries and temporal constraints plus the additional information at dialogue level utilized in the second strategy. Within this new approach therefore, the fusion process will be initiated at grammar level and will be culminated at dialogue level.

# Carredano G A. Multimodal fusion: a new hybrid strategy for dialogue systems[C]// International Conference on Multimodal Interfaces. ACM, 2006:357-363.

53. In this work we explicitly take into account feature measurement uncertainty and we show how classification rules should be adjusted to compensate for its effects. Our approach is particularly fruitful in multimodal fusion scenarios, such as audio-visual speech recognition, where multiple streams of complementary time-evolving features are integrated. For such applications, provided that the measurement noise uncertainty for each feature stream can be estimated, the proposed framework leads to highly adaptive multimodal fusion rules which are widely applicable and easy to implement. We further show that previous multimodal fusion methods relying on stream weights fall under our scheme under certain assumptions; this provides novel insights into their applicability for various tasks and suggests new practical ways for estimating the stream weights adaptively. The potential of our approach is demonstrated in audio-visual speech recognition using either synchronous or asynchronous models.

# Pitsikalis V, Katsamanis A, Papandreou G, et al. Adaptive multimodal fusion by uncertainty compensation[C]// INTERSPEECH 2006 - Icslp, Ninth International Conference on Spoken Language Processing, Pittsburgh, Pa, Usa, September. DBLP, 2006:423 - 435.

54. In this paper, we explicitly take feature measurement uncertainty into account and show how multimodal classification and learning rules should be adjusted to compensate for its effects. Our approach is particularly fruitful in multimodal fusion scenarios, such as audiovisual speech recognition, where multiple streams of complementary time-evolving features are integrated. For such applications, provided that the measurement noise uncertainty for each feature stream can be estimated, the proposed framework leads to highly adaptive multimodal fusion rules which are easy and efficient to implement. Our technique is widely applicable and can be transparently integrated with either synchronous or asynchronous multimodal sequence integration architectures. We further show that multimodal fusion methods relying on stream weights can naturally emerge from our scheme under certain assumptions; this connection provides valuable insights into the adaptivity properties of our multimodal uncertainty compensation approach. We show how these ideas can be practically applied for audiovisual speech recognition. In this context, we propose improved techniques for person-independent visual feature extraction and uncertainty estimation with active appearance models, and also discuss how enhanced audio features along with their uncertainty estimates can be effectively computed. We demonstrate the efficacy of our approach in audiovisual speech recognition experiments on the CUAVE database using either synchronous or asynchronous multimodal integration models.

# Papandreou G, Katsamanis A, Pitsikalis V, et al. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition[J]. IEEE Transactions on Audio Speech & Language Processing, 2009, 17(3):423-435.

55. The study of multimodality is comparatively less developed for affective interfaces than for their traditional counterparts. However, one condition for the successful development of affective interface technologies is the development of frameworks for the real-time multimodal fusion. In this paper, we describe an approach to multimodal affective fusion, which relies on a dimensional model, Pleasure-Arousal-Dominance (PAD) to support the fusion of affective modalities, each input modality being represented as a PAD vector. We describe how this model supports both affective content fusion and temporal fusion within a unified approach. We report results from early user studies which confirm the existence of a correlation between measured affective input and user temperament scores.

# Gilroy S W, Cavazza M, Niiranen M, et al. PAD-based multimodal affective fusion[C]// International Conference on Affective Computing and Intelligent Interaction and Workshops. IEEE, 2009:1-8.

56. We propose effective multimodal fusion strategies for video search. Multimodal search is a widely applicable information-retrieval problem, and fusion strategies are essential to the system in order to utilize all available retrieval experts and to boost the performance. Prior work has focused on hard-and soft-modeling of query classes and learning weights for each class, while the class partition is either manually defined or learned from data but still insensitive to the testing query. We propose a query-dependent fusion strategy that dynamically generates a class among the training queries that are closest to the testing query, based on light-weight query features defined on the outcome of semantic analysis on the query text. A set of optimal weights are then learned on the dynamic class, which aims to model both the co-occurring query features and unusual test queries. Used in conjunction with the rest of our multimodal retrieval system, dynamic query classes performs favorably with hard and soft query classes, and the system performance improves upon the best automatic search run of TRECVID05 and TRECVID06 by 34% and 8%, respectively.

# Xie L X L, Natsev A, Tesic J. Dynamic Multimodal Fusion in Video Search[J]. 2007:1499-1502.

57. This paper investigates a novel approach to model the relation between such modality types bypassing the need for summarising each modality independently of each other. For that purpose, we introduce a deep learning model based on convolutional neural networks that is adapted to process multiple modalities at different time resolutions we name deep multimodal fusion. Furthermore, we introduce and compare three alternative methods (convolution, training and pooling fusion) to integrate sequences of events with continuous signals within this model. We evaluate deep multimodal fusion using a game user dataset where player physiological signals are recorded in parallel with game events. Results suggest that the proposed architecture can appropriately capture multimodal information as it yields higher prediction accuracies compared to single-modality models. In addition, it appears that pooling fusion, based on a novel filter-pooling method provides the more effective fusion approach for the investigated types of data

# Yannakakis G N. Deep Multimodal Fusion: Combining Discrete Events and Continuous Signals[C]// International Conference on Multimodal Interaction. ACM, 2014:34-41.

58. The capability of interpreting the conceptual and affective information associated with natural language through different modalities is a key issue for the enhancement of human-agent interaction. The proposed methodology, termed sentic blending, enables the continuous interpretation of semantics and sentics (i.e., the conceptual and affective information associated with natural language) based on the integration of an affective common-sense knowledge base with any multimodal signal-processing module. In this work, in particular, sentic blending is interfaced with a facial emotional classifier and an opinion mining engine. One of the main distinguishing features of the proposed technique is that it does not simply perform cognitive and affective classification in terms of discrete labels, but it operates in a multidimensional space that enables the generation of a continuous stream characterising user's semantic and sentic progress over time, despite the outputs of the unimodal categorical modules have very different time-scales and output labels.

# Cambria E, Howard N, Hsu J, et al. Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics[C]// IEEE, 2013:108-117.

59. Because a people detection system that considers only a single feature tends to be unstable, many people detection systems have been proposed to extract multiple features simultaneously. These detection systems usually integrate features using a heuristic method based on the designers' observations and induction. Whenever the number of features to be considered is changed, the designer must change and adjust the integration mechanism accordingly. To avoid this tedious process, we propose a multimodal fusion system that can detect and track people in a scalable, accurate, robust, and flexible manner. Each module considers a single feature and all modules operate independently at the same time. A depth module is constructed to detect people based on the depth-from-stereo method, and a novel approach is proposed to extract people by analyzing the vertical projection in each layer. A color module that detects the human face, and a motion module that detects human movement are also developed. The outputs from these individual modules are fused together and tracked over time, using a Kalman filter.

# Yang, Mau‐Tsuen, Wang, Shih‐Chun, Lin, Yong‐Yuan. A multimodal fusion system for people detection and tracking[J]. International Journal of Imaging Systems & Technology, 2005, 15(2):131-142.

60. It is well-known that early integration (also called data fusion) is effective when the modalities are correlated, and late integration (also called decision or opinion fusion) is optimal when modalities are uncorrelated. In this paper, we propose a new multimodal fusion strategy for open-set speaker identification using a combination of early and late integration following canonical correlation analysis (CCA) of speech and lip texture features. We also propose a method for high precision synchronization of the speech and lip features using CCA prior to the proposed fusion. Experimental results show that i) the proposed fusion strategy yields the best equal error rates (EER), which are used to quantify the performance of the fusion strategy for open-set speaker identification, and ii) precise synchronization prior to fusion improves the EER; hence, the best EER is obtained when the proposed synchronization scheme is employed together with the proposed fusion strategy. We note that the proposed fusion strategy outperforms others because the features used in the late integration are truly uncorrelated, since they are output of the CCA analysis.

# 60.Sargin M E, Yemez Y, Erzin E, et al. Audiovisual Synchronization and Fusion Using Canonical Correlation Analysis[J]. IEEE Transactions on Multimedia, 2007, 9(7):1396-1403.

61. Image collections on the internet and other sources of information can naturally include attached text descriptions. This work considers the problem of fusing two data modalities: visual content and text keywords, to allow a flexible image indexing scheme. The proposed strategy learns multimodal relationships using matrix reconstruction principles and factorization algorithms, allowing one data modality to be represented in another modality space. We further exploit this exchangeability property, to fuse the modalities in any of the representation spaces by backprojecting predicted data to the input space. An experimental evaluation was carried out on the Corel 5K and MIRFlickr data sets using example images without text as query paradigm. Experimental results demonstrate the ability of the proposed strategy to find multimodal links between data and make them useful to improve the image retrieval performance.

# Caicedo J C. Multimodal fusion for image retrieval using matrix factorization[C]// Proceedings of the 2nd ACM International Conference on Multimedia Retrieval. ACM, 2012:1-8.

62.

To improve classification accuracy, most of the previous work focuses on extracting more features and building algorithms for a specific task. This paper proposes a framework based on the novel and robust Collateral Representative Subspace Projection Modeling (C-RSPM) supervised classification model for general histology image classification. In the proposed framework, a cell image is first divided into 25 blocks to reduce the spatial complexity of computation, and one C-RSPM model is built on each block set which contains blocks in the same location from different images. For each testing image, our proposed framework first classifies each of its blocks using the C-RSPM classification model built for that block set, and then applies a multimodal late fusion algorithm with a weighted majority voting strategy to decide the final class label of the whole image. Experimenting using three-fold cross validation with three benchmark histology data sets shows that the proposed framework outperforms other well-known classifiers in the comparison and gives better results than the highest accuracy reported previously.

# Meng T, Lin L, Shyu M L, et al. Histology Image Classification Using Supervised Classification and Multimodal Fusion[C]// IEEE International Symposium on Multimedia. IEEE, 2011:145-152.

63.匹配分数是传统的融合分数指标,但是其不能很好地区分类内和类间数据,分类置信度虽然可以较好地将类内类间数据分开,但对于匹配分数仅次于分类阈值的数据,其分类效果不是很理想.因此,首先提出了一种基于分类距离分数的融合分数指标,其不仅携带一级分类信息,也含有匹配分数与分类阈值之间的距离信息,可增大融合后类内类间分数之间的距离,为融合算法提供了一个具有有效判别信息的特征融合集,提高了融合指标的利用率;进一步,利用信息熵表示信息价值多少的这一特性,定义特征关联系数和特征权重系数,并将加权融合和传统SUM规则统一在一个自适应算法框架中,提高了融合识别率.实验结果验证了所提出方法的有效性.

# 张露, 陶亮. 基于分类距离分数的自适应多模态生物特征融合[J]. 计算机研究与发展, 2018(1):151-162.

64. 本文针对单一模式识别的局限性,提出基于人耳、人脸的多模态识别,并应用PCA改进的SIFT算法对人脸、人耳进行特征提取,并在匹配层进行融合,有效提高了识别率。相较于传统的SIFT算法,改进的SIFT算法具有更强的稳定性和鲁棒性。随着信息技术的发展,基于生物特征的模式识别成为研究热点,现有的识别技术包括人脸识别、指纹识别、虹膜识别等。

# 耿子烨, 高靖翔, 鲍佳宁. 基于改进的SIFT算法的多模态识别[J]. 中国科技信息, 2018(1):81-82.

65. 针对视频分割中底层特征与高层语义之间的“语义鸿沟”问题，提出了一种基于多模态融合和镜头间竞争力的场景分割算法，对视频帧的图像、文本、音频等模态进行特征提取，用欧式距离、余弦距离计算出同种模态数据的相似性，用典型相关分析法计算出不同模态数据的相关度，分别对各模态数据的相似性和相关度进行融合得到镜头之间的相似度和相关度，采用镜头间竞争力的方法分别对相似镜头和相关镜头进行场景分割并对分割出的两个场景边界集合取交集得到最终的场景边界，从而实现对视频的场景分割。实验结果表明，该方法在场景分割中具有较高的性能，查全率和查准率分别达到82．1％和86．7％。

# 杨亭, 丰洪才, 金凯,等. 基于多模态融合和竞争力的视频场景分割算法[J]. 武汉理工大学学报(信息与管理工程版), 2014(6):759-763.

66. 近年来,采用脑电信号(EEG)与功能磁共振成像(fMRI)多模态融合方法确定脑激活源位置、分析激活区域间连通性的方法逐步成为认知科学研究的主流。EEG/fMRI多模态融合,可以有效克服单模态分析的缺陷,形成对脑神经活动的高空间分辨率高时间分辨率分析。 论文首先分别介绍了EEG和fMRI单模态分析研究主要的问题、算法和应用,评价了各自的不足,然后介绍了EEG/fMRI融合分析的整个过程,包括数据采集,噪声滤除,脑模型重构,常用的血液动力学响应函数,多模态数据配准方法,最后对多模态融合算法进行分类,总结了当前主流的算法和思路,举例说明EEG/fMRI融合方法在具体脑功能算法中的广泛应用。论文重点突出了脑模型重构和融合方法分析。在脑模型重构中,分别介绍了球脑模型,真实脑模型和复杂球模型构造原理,给出了皮层激活源电压和头皮测量电压间关系。在融合方法分析中,论文将目前主流的融合方法分为两类,约束融合方法和直接融合方法。分别列举了卡尔曼滤波算法,Twomey规范化算法,线性逆估计算法和时空最优算法,多路偏最小二乘算法,贝叶斯融合法。利用多路偏最小二乘方法分析了一组静息数据,并获得有意义的结果。论文最后讨论了EEG/fMRI多模态融合分析面临的一些问题,展望了未来研究的方向和目标。

# 何继军. 脑电与功能磁共振成像(EEG/fMRI)多模态融合分析[D]. 国防科学技术大学, 2006.

67. 本文分别从文本模态，图像模态，多模态融合三个方面对微博分类问题进行了研究，主要贡献如下: 　　(1)实现了基于文本模态的微博分类方法 ... 展开 微博分类作为处理和组织大量微博数据的关键技术，可以很大程度上解决微博信息爆炸的现象。但由于微博文本具有长度短、口语化等特点，传统的文本分类技术对微博文本不太适用。实现了基于多模态融合的微博分类方法 　　在文本和图像两个单模态模型的基础上，本文实现了两种基于多模态融合的微博分类方法。第一种是基于语言模型的单模态微博分类结果融合，即在两个单模态模型输出的概率向量基础上，结合语言模型、文本长度、图像个数等特征进行单模态结果的融合。第二种是基于多模态特征融合的微博分类，该方法把两个模态的特征使用卷积等方法进行规格统一，构成一个数据流，使用LSTM模型对两个模态进行融合。实验表明，基于多模态融合的微博分类效果要显著优于单模态的模型。对比基于Word2Vec语义扩充的微博分类结果，特征融合的方法F1值提升了4.4个百分点，达到了84.94％。

# 张水源. 基于多模态融合的微博分类[D]. 中国科学院大学, 2016.

68. 图像和文本分类是模式识别领域的一个重要研究方向,而毒品网页过滤则可以是看作特殊的图像和文本分类技术。通用的图像和文本分类技术对建立快速而有效地信息检索和管理系统有着重要的意义,而毒品图像和文本过滤技术则可以自动的过滤互联网上的毒品信息,净化互联网环境,保证互联网的信息安全。研究了利用多种模态多层次信息进行融合的方法。本文利用多模态分类器来进行分类,第一个模态是多模态多示例图像分类器,另一是正文文本分类器。然后,利用D-S证据理论将二个分类器的结果进行融合。无论是单独的文本分类器还是图像分类器,其召回率,即把毒品网页分对的比率都不高。但是当两者采用D-S证据理论进行融合后,其召回率大大提高,并且准确率较之文本分类器也没有改变太多。实验证明,本文提出的方法不但考虑实用性,速度,而且还考虑到了分类器的性能以及对于毒品分类这个特殊领域的特殊要求。

# 汪银娟. 基于多模态融合的毒品网页过滤技术[D]. 中国民航大学, 2012.

69. ,本文提出一种新的网络视频多模态融合检索方案。分别从视频视觉内容,视频标题和标签的文本信息,以及视频上传时间、类别、作者三种人与视频交互产生的社会特征等异构信息出发进行多模态融合研究,并将此方法应用到大规模图像视频检索任务中。Youtube数据集上的实验结果显示:相对于传统单一文本特征、单一视觉特征的检索方案以及两模态融合的检索方案,我们的文本、视觉和用户社会特征多模态融合融合方案表现出更好的性能。其次,本文提出了一种主动式的内部参数调优算法和模态间参数的自学习算法。上文提到的多模态融合检索方案面临着两个问题,各模态内部和模态间参数过多的问题。对于第一个问题,如果在模态内部做简单加权,模态内部权重系数一般只能由个人经验给出,受主观因素影响大。针对多种社会特征,本文通过各模态内参数调整并自动迭代调优,实现了内部参数的自适应学习。针对第二个问题,在单纯应用一个或两个模态的情况下,可通过试参方式对参数进行选取,模态特性过多时调参过程复杂,本文通过考查多模态类结构的分类有效性,利用分类方法达到模态间参数自学习的目的。最后,本文利用给出的多模融合方案对视频主题进行分类,实验结果表明:在此方案下的主题间分类可以取得较好的效果。

# 温有福. 面向网络视频检索的多模态融合方法研究[D]. 北京交通大学, 2017.

70. 音乐数据通常由音频信号与歌词文本两种模态所组成,传统的音乐情感分类方法多数侧重于对单一模态数据的分析,由于单模态数据中所包含语义信息的局限性,使得其往往无法完全表达音乐中蕴含的情感信息,因此研究有效的方法挖掘和利用音乐中多种模态数据间相关性与互补性,对提高现有音乐情感分类方法的性能具有十分重要的意义。本文以融合多种模态数据的音乐情感分类方法为具体研究对象,探索综合利用多种模态音乐数据中包含的相关情感信息以有效提高情感分类的准确率。不同于以往基于文档级音乐表示的多模态音乐情感分类方法,本文提出了句子层次的音乐特征表示机制以从更细粒度精确刻画其情感属性,并且提出了基于区分度排序与同义词扩展的音乐文本预处理方法以提高音乐文本数据的情感类别区分能力。另一方面,为消除不同模态数据间的异构性以实现更加有效的融合,本文提出了针对音乐情感分类的多模态Locality Preserving Projection算法,将音频与歌词模态数据映射到更具情感分类能力的公共隐式特征空间中。在其基础上,本文进一步提出了基于多模态投票的Hough森林音乐情感分类方法,利用不同模态的音乐数据在时间上的相关性提升了音乐情感分类的准确率。另一方面,本文提出了基于多模态K近邻与图学习的音乐情感分类方法,基于不同模态特征在隐式空间中的相似度实现情感标签的有效传播,充分利用了不同模态音乐数据之间的相关性和互补性,提高了音乐情感分类的准确率。实验结果表明,本文提出的方法有效融合了不同模态音乐数据中的情感信息,有效提升了音乐情感分类的精度。

# 70.薛昊. 基于多模态融合的音乐情感分类方法研究[D]. 南京大学, 2016.

71.考虑到多模态数据分析的上述困难，本文采用了主题概率模型，通过主题这一语义层次概念建立异构数据间的关联。本文在实现PLSA(Probabilistic Latentsemantic Analysis)算法模型的基础上，利用该模型探索了建立视、听觉模态特征之间相关性，并通过给出了实验比较与分析来验证了这一思路的可行性。在此基础上，我们首先对PLSA模型进行扩展，提出了一个新的基于多模态融合和传播的跨模态信息检索算法，该方法首先对能够处理连续特征的多层概率统计模型建模，通过共同的主题变量关联起来，接着采用不对称学习的方法来融合视觉特征和听觉特征这两种不同模态的语义信息。最后，通过本文提出的多模态相似度传播模型，利用不同模态的相关性、互补性，迭代传播模态间的相似度值，充分发掘了多模态间的相关信息，避免了单纯倚重一类特征造成的检索偏差。实验结果证明，本文方法可以有效实现多模态数据间的关联分析与检索。

# 林婉霞. 基于多模态融合和传播的跨模态信息检索算法[D]. 南京大学, 2012.

72.专利- 本发明公开了一种基于深度学习的多模态耦合的歌曲情感识别方法，其特征是按如下步骤进行：1获取歌曲歌词文本数据和音频语音数据；2对歌词文本内容进行文本特征提取，获得歌词文本信息特征；3提取歌曲语音数据的第一语音特征和第二语音特征并进行第一次融合，获得歌曲语音信息特征；4对歌词文本信息特征和歌曲语音信息特征进行第二次融合，获得歌曲的综合信息特征；5利用深度分类器对综合信息特征进行训练，获得歌曲情感识别模型，以歌曲情感识别模型实现对歌曲的多模态融合的情感识别。本发明能全面结合歌曲的歌词文本信息和歌曲音频信息两个方面的数据信息，从而提高人机交互中的对歌曲情感状态判断的准确度。

# 孙晓, 陈炜亮, 任福继. 基于深度学习的多模态融合的歌曲情感识别方法, Emotion recognition method based on the song depth study of multimodal fusion:, CN 106228977 A[P]. 2016.

73. 本发明公开一种多模态融合手功能康复训练与智能评估系统，该系统包括：康复训练机构、传感器数据采集装置、康复评估系统以及情景互动游戏系统，康复训练机构用于对手指和手腕进行康复训练，传感器数据采集装置用于实时采集表面肌电信号、力信号以及姿态信号中的一种或多种信号，康复评估系统根据传感器数据采集装置采集的多模态信号对患者的康复训练情况进行量化评估，情景互动游戏系统用于为患者提供游戏康复训练、效果评估和视觉等信息反馈。本发明公开多模态融合手功能康复训练与智能评估系统，针对脑卒中患者康复训练，可以对康复训练过程进行实时反馈调控、量化评估和游戏反馈，有效调动其训练积极性，增强康复训练效果。

# 郭立泉, 王计平, 郁磊,等. 多模态融合手功能康复训练与智能评估系统:, CN 105963926 A[P]. 2016.

74. 本文主要研究了手指静脉识别中的关键技术,以及基于指纹和手指静脉的多模态生物识别系统。首先简要的介绍了生物识别技术的相关问题以及研究现状。之后在指纹合成技术的启发下,提出了一种合成手指静脉图像的方法并将其实现,用于手指静脉识别算法的研究和测试。对于手指静脉识别的关键技术,主要研究了小波域下的手指静脉特征提取和匹配算法,采用二进小波分析与纹理特征相结合的方法实现了手指静脉识别。同时进一步的对静脉模式的提取算法进行研究,在分析了手指静脉红外图像的特点后,提出了位置灰度剖面图的概念,在该图中计算每个点的曲率,通过寻找局部曲率最大点的集合获得静脉模式,实验证明,该方法能够有效的获得更加清晰准确的静脉模式。最后在研究了指纹和指静脉图像质量的评测后,实现了基于这两种生物特征的匹配层融合的多模态生物识别系统。设计指纹和指静脉的多模态生物识别系统,可以同时获得手指静脉和指纹图像,与采用多种设备的融合系统相比,降低成本,大大提高了生物认证的效率,弥补了单种生物特征认证不稳定、错误率较高的缺点,是对人类自身认证方式的模拟,具有很好的研究和应用的前景。

# 李雪妍. 融合指纹和指静脉的多模态生物识别技术的研究[D]. 吉林大学, 2008.

75. 目的探讨一种基于小波变换的医学图像融合的新算法,为充分发挥现有医学成像系统的功能提供一种可靠手段。方法采用一种基于小波变换的医学图像融合新算法。首先对已配准的医学图像进行小波分解,把图像分解成低频和高频子图像,采用加权因子的融合规则对低频系数进行融合,采用基于区域能量的融合规则对高频系数进行融合,最后通过重构得到融合图像。结果该算法有效地将多模态医学图像所提供的信息融合在一起,为临床诊断提供了更充分、更可靠的信息。结论该算法是一种可行有效的医学图像融合算法。

# 张彬, 郑永果, 马芳,等. 基于小波变换的多模态医学图像融合算法[J]. 重庆医学, 2011, 40(7):652-654.

76. 事件检测是视频语义分析中的一大难题。在视频中有多种语义丰富的模态信息，融合多模态信息可以帮助准确的检索出所需的事件。提出了一种基于HMM有效融合多模态对象的足球视频语义分析方法。首先从视濒中抽取音频流，然后基于CHMM将音频分类。接着根据时间对应关系将音频对象与视频流融合，然后在相应的视频流镜头中基于DHMM融合多种模态对象实现精彩事件如射门、犯规及一般事件的检测，对射门事件进一步结合比分字幕的出现检测进球事件。另外，对DHMM模型的结构、参数初始值尤其是参数约束条件进行了详细地描述。实验证实提出的算法具有较好的效率。

# 张玉珍[1, 丁思捷, 王建宇,等. 基于HMM的融合多模态的事件检测[J]. 系统仿真学报, 2012, 24(8):80-84.

77. 以微博为代表的社会媒体的蓬勃发展在加速信息交流的同时,也促使虚假谣言信息迅速在社会网络上传播,造成严重的后果.自动谣言检测问题受到了国内外学术界、产业界的广泛关注.围绕社会多媒体谣言检测这一问题,本文总结了融合多模态特征的谣言检测相关技术.首先从基本概念出发,阐述了谣言的定义和社会多媒体的特点,给出了社会多媒体谣言检测问题的定义.针对谣言检测面临的多模态特征抽取和模型构建两大难点,分别总结和归纳了各种类型的特征及其提取方法和不同的机器学习检测模型.这些特征和算法是检测谣言的基本手段,也是接下来研究的基础,可为进一步谣言检测的研究提供参考.

78.Speech is the most natural form of communication. Very less work has been carried out by fusion of speech parameters. This paper attempts to review the fusion of speech parameters to recognize emotions in human speech. Emotion recognition and verification is the process of determination of the psychological state of the speaker. Every emotion comprises different vocal parameters exhibiting diverse characteristics of speech. These features can be extracted using efficient parameters like MFCC coefficients, energy, pitch etc., if the fusion results of these parameters is the input to the trained neural net, it will be possible to analyze correct emotion of speaker. Harnessing the approaches of signal processing and pattern recognition algorithms, a smart and emotions specific man machine interaction can be achieved. Neural networks can be explored to model the prosodic parameters of the syllables from their positional, contextual and phonological features. It may also be possible to search effective emotion parameters and recognize emotions accurately. In future, it will lead to creation machine capable to work considering human emotions.

# R. V. Darekar, A. P. Dhande. EMOTION DETECTION WITH MULTIMODAL FUSION USING SPEECH - A REVIEW[J]. International Journal of Computer Science and Communication Engineering,2014,3(1):40-45.

78. 研究了主持人镜头检测，提出了一种基于音视频特征融合的主持人镜头提取方法。主持人模板是通过综合考虑静音片段的特征之后自动提取的。然后利用主持人镜头背景不变的特性，通过色矩计算和颜色模板匹配来进行主持人镜头的检测。该方法实现了主持人镜头的自动化聚焦，具有准确度高和适应性强等优点。

研究了视频摘要的生成及表现形式。综合考虑新闻视频中的多种模态信息，提出了基于多模态特征融合的新闻视频摘要生成方法，视频摘要表现形式包括新闻故事板形式的静态摘要、基于比例压缩的视频缩略、基于主持人镜头的视频缩略和基于新闻标题条的视频缩略。  
最后，设计实现了一个多特征融合的新闻视频摘要原型系统，系统集成了本文研究的主要成果，实验结果表明，用户对摘要结果的满意度较高。

# 孟文婷. 基于多模态特征融合的新闻视频摘要技术研究[D]. 华中师范大学, 2012.

79.专利-本发明公开了一种基于多模态序列融合的动作识别方法，包括以下步骤：获得原始视频多种模态下的信息，并进行预处理获取初始RGB图像序列及初始深度图像序列，以及获取骨架特征序列；对初始RGB图像序列进行背景建模，通过得到的背景建模结果来提取第一人体区域外接矩形，同时提取初始深度图像序列上相应位置的第二人体区域外接矩形；在第一、第二人体区域外接矩形上分别提取人体动作特征，得到RGB模态和深度模态下的特征向量；根据得到的RGB-LBP、D-LBP特征向量序列，以及骨架特征序列，通过多视角判别模型来进行动作识别。本发明能够将多种模态的序列信息进行互补融合，能够有效提高动作识别准确率。

# 刘安安, 苏育挺, 马莉. 一种基于多模态序列融合的动作识别方法:, CN104156693A[P]. 2014.

80.身份鉴别所需要的高度准确性和鲁棒性往往很难通过使用单模态的生物证人方法来达到.本文提出了一种基于动态贝叶斯网络的融合框架,将声纹信息与人脸信息在特征层进行融合。

# 80.李冬冬, 吴朝晖, 杨莹春. 动态贝叶斯网络在多模态说话人鉴别上的研究[C]// 中国科协优秀博士生学术年会. 2004.

81. The study at hand aims at the development of a multimodal, ensemble-based system for emotion recognition. Special attention is given to a problem often neglected: missing data in one or more modalities. In offline evaluation the issue can be easily solved by excluding those parts of the corpus where one or more channels are corrupted or not suitable for evaluation. In real applications, however, we cannot neglect the challenge of missing data and have to find adequate ways to handle it. To address this, we do not expect examined data to be completely available at all time in our experiments. The presented system solves the problem at the multimodal fusion stage, so various ensemble techniques-covering established ones as well as rather novel emotion specific approaches-will be explained and enriched with strategies on how to compensate for temporarily unavailable modalities. We will compare and discuss advantages and drawbacks of fusion categories and extensive evaluation of mentioned techniques is carried out on the CALLAS Expressivity Corpus, featuring facial, vocal, and gestural modalities.

# Wagner J, Andre E, Lingenfelser F, et al. Exploring Fusion Methods for Multimodal Emotion Recognition with Missing Data[J]. IEEE Transactions on Affective Computing, 2011, 2(4):206-218.

82. Novel fusion algorithm, ISR, improves results from existing rank fusion algorithms. We got the best result on multimodal case-based retrieval in 2013 ImageCLEFMedical.

# Mourão A, Martins F, Magalhães J. Multimodal medical information retrieval with unsupervised rank fusion.[J]. Computerized Medical Imaging & Graphics, 2015, 39:35-45.

83. The use of clustering algorithms for decision-level data fusion is proposed. Person authentication results coming from several modalities (e.g., still image, speech), are combined by using fuzzy k-means (FKM) and fuzzy vector quantization (FVQ) algorithms, and a median radial basis function (MRBF) network. The quality measure of the modalities data is used for fuzzification. Two modifications of the FKM and FVQ algorithms, based on a fuzzy vector distance definition, are proposed to handle the fuzzy data and utilize the quality measure. Simulations show that fuzzy clustering algorithms have better performance compared to the classical clustering algorithms and other known fusion algorithms. MRBF has better performance especially when two modalities are combined. Moreover, the use of the quality via the proposed modified algorithms increases the performance of the fusion system.

# Chatzis V, Bors A G, Pitas I. Multimodal decision-level fusion for person authentication[J]. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 1999, 29(6):674-680.

84. We conduct a broad survey of query-adaptive search strategies in a variety of application domains, where the internal retrieval mechanisms used for search are adapted in response to the anticipated needs for each individual query experienced by the system. While these query-adaptive approaches can range from meta-search over text collections to multimodal search over video databases, we propose that all such systems can be framed and discussed in the context of a single, unified framework. In our paper, we keep an eye towards the domain of video search, where search cues are available from a rich set of modalities, including textual speech transcripts, low-level visual features, and high-level semantic concept detectors. The relative efficacy of each of the modalities is highly variant between many types of queries. We observe that the state of the art in query-adaptive retrieval frameworks for video collections is highly dependent upon the definition of classes of queries, which are groups of queries that share similar optimal search strategies, while many applications in text and Web retrieval have included many advanced strategies, such as direct prediction of search method performance and inclusion of contextual cues from the searcher. We conclude that such advanced strategies previously developed for text retrieval have a broad range of possible applications in future research in multimodal video search.

# Kennedy L, Chang S F, Natsev A. Query-Adaptive Fusion for Multimodal Search[J]. Proceedings of the IEEE, 2008, 96(4):567-588.

85. A multimodal biometric identification system aims to fuse two or more physical or behavioral traits. Multimodal biometric system is used in order to improve the accuracy. Multimodal biometric identification system based on palmprint fingerprint trait is proposed. Typically in a multimodal biometric system each biometric trait processes its information independently. The processed information is combined using an appropriate fusion scheme. Successively, the comparison of data base template and the input data is done with the help of Euclidean-distance matching algorithm. If the templates are matched we can allow the person to access the system. The experimental results demonstrated that the proposed multimodal biometric system achieves a recognition accuracy of 87% Multimodal biometric system provides optimal False Acceptance Rate (FAR) False Rejection Rate (FRR), thus improving system accuracy reliability.

# Chaudhari J. A multimodal biometric recognition system based on fusion of palmprint and fingerprint[J]. International Journal of Engineering Trends & Technology, 2013, 4(5):1908-1911.

86. In this paper, we present our new results in news video story segmentation and classification in the context of TRECVID video retrieval benchmarking event 2003. We applied and extended the Maximum Entropy statistical model to effectively fuse diverse features from multiple levels and modalities, including visual, audio, and text. We have included various features such as motion, face, music/speech types, prosody, and high-level text segmentation information. The statistical fusion model is used to automatically discover relevant features contributing to the detection of story boundaries. One novel aspect of our method is the use of a feature wrapper to address different types of features -- asynchronous, discrete, continuous and delta ones. We also developed several novel features related to prosody. Using the large news video set from the TRECVID 2003 benchmark, we demonstrate satisfactory performance (F1 measures up to 0.76 in ABC news and 0.73 in CNN news), present how these multi-level multi-modal features construct the probabilistic framework, and more importantly observe an interesting opportunity for further improvement.

# Kennedy L. Discovery and fusion of salient multimodal features toward news story segmentation[J]. 2003, 5307(s 1–2):244-258.

87. A new application, Fusion Viewer, available for free, has been designed and implemented with a modular object-oriented design. The viewer provides both traditional and novel tools to fuse 3D data sets such as CT (computed tomography), MRI (magnetic resonance imaging), PET (positron emission tomography), and SPECT (single photon emission tomography) of the same subject, to create maximum intensity projections (MIP) and to adjust dynamic range. In many situations, it is desirable and advantageous to acquire biomedical images in more than one modality. For example, PET can be used to acquire functional data, whereas MRI can be used to acquire morphological data. In some situations, a side-by-side comparison of the images provides enough information, but in most of the cases it may be necessary to have the exact spatial relationship between the modalities presented to the observer. To accomplish this task, the images need to first be registered and then combined (fused) to create a single image. In this paper, we discuss the options for performing such fusion in the context of multimodal breast imaging. Additionally, a novel spline-based dynamic range technique is presented in detail. It has the advantage of obtaining a high level of contrast in the intensity range of interest without discarding the intensity information outside of this range while maintaining a user interface similar to the standard window/level windowing procedure.

# Baum K G, Helguera M, Krol A. Fusion viewer: a new tool for fusion and visualization of multimodal medical data sets[J]. Journal of Digital Imaging, 2008, 21(1):59-68.

88. In this paper, we present a flexible and effective reranking method, called CR-Reranking, to improve the retrieval effectiveness. To offer high accuracy on the top-ranked results, CR-Reranking employs a cross-reference (CR) strategy to fuse multimodal cues. Specifically, multimodal features are first utilized separately to rerank the initial returned results at the cluster level, and then all the rdyanked clusters from different modalities are cooperatively used to infer the shots with high relevance. Experimental results show that the search quality, especially on the top-ranked results, is improved significantly.

# Wei S, Zhao Y, Zhu Z, et al. Multimodal Fusion for Video Search Reranking[J]. IEEE Transactions on Knowledge & Data Engineering, 2010, 22(8):1191-1199.

89.

This is a new hybrid fusion strategy based primarily on the implementation of two former and differentiated approaches to multimodal fusion [11] in multimodal dialogue systems. Both approaches, their predecessors and their respective advantages and disadvantages will be described in order to illustrate how the new strategy merges them into a more solid and coherent solution. The first strategy was largely based on Johnston's approach [5] and implies the inclusion of multimodal grammar entries and temporal constraints. The second approach implied the fusion of information coming from different channels at dialogue level. The new hybrid strategy hereby described requires the inclusion of multimodal grammar entries and temporal constraints plus the additional information at dialogue level utilized in the second strategy. Within this new approach therefore, the fusion process will be initiated at grammar level and will be culminated at dialogue level.

# Carredano G A. Multimodal fusion: a new hybrid strategy for dialogue systems[C]// International Conference on Multimodal Interfaces. ACM, 2006:357-363.

90. The multimodal physiological signals are: Electroencephalogram (EEG) (32 channels) and peripheral (8 channels: Galvanic skin response (GSR), blood volume pressure, respiration pattern, skin temperature, electromyogram (EMG) and electrooculogram (EOG)) as given in the DEAP database.

A novel approach for multimodal fusion of information from a large number of channels to classify and predict emotions has also been proposed. Discrete Wavelet Transform, a classical transform for multiresolution analysis of signal has been used in this study. The experiments are performed to classify different emotions from four classifiers. The average accuracies are 81.45%, 74.37%, 57.74% and 75.94% for SVM, MLP, KNN and MMC classifiers respectively. The best accuracy is for 'Depressing' with 85.46% using SVM. The 32 EEG channels are considered as independent modes and features from each channel are considered with equal importance. May be some of the channel data are correlated but they may contain supplementary information. In comparison with the results given by others, the high accuracy of 85% with 13 emotions and 32 subjects from our proposed method clearly proves the potential of our multimodal fusion approach.

# 90.Verma G K, Tiwary U S. Multimodal fusion framework: a multiresolution approach for emotion classification and recognition from physiological signals.[J]. Neuroimage, 2014, 102 Pt 1:162.

91. During face to face communication, it has been suggested that as much as 70% of what people communicate when talking directly with others is through paralanguage involving multiple modalities combined together (e.g. voice tone and volume, body language). In an attempt to render humancomputer interaction more similar to human-human communication and enhance its naturalness, research on sensory acquisition and interpretation of single modalities of human expressions have seen ongoing progress over the last decade. These progresses are rendering current research on artificial sensor fusion of multiple modalities an increasingly important research domain in order to reach better accuracy of congruent messages on the one hand, and possibly to be able to detect incongruent messages across multiple modalities (incongruency being itself a message about the nature of the information being conveyed). Accurate interpretation of emotional signals - quintessentially multimodal - would hence particularly benefit from multimodal sensor fusion and interpretation algorithms. In this paper we provide a state of the art multimodal fusion and describe one way to implement a generic framework for multimodal emotion recognition. The system is developed within the MAUI framework [31] and Scherer's Component Process Theory (CPT) [49, 50, 51, 24, 52], with the goal to be modular and adaptive. We want the designed framework to be able to accept different single and multi modality recognition systems and to automatically adapt the fusion algorithm to find optimal solutions. The system also aims to be adaptive to channel (and system) reliability.

# Paleari M, Lisetti C L. Toward multimodal fusion of affective cues[C]// ACM International Workshop on Human-Centered Multimedia. ACM, 2006:99-108.

92. We propose a novel hybrid model that exploits the strength of discriminative classifiers along with the representational power of generative models. Our focus is on detecting multimodal events in time varying sequences. Discriminative classifiers have been shown to achieve higher performances than the corresponding generative likelihood-based classifiers. On the other hand, generative models learn a rich informative space which allows for data generation and joint feature representation that discriminative models lack. We employ a deep temporal generative model for unsupervised learning of a shared representation across multiple modalities with time varying data. The temporal generative model takes into account short term temporal phenomena and allows for filling in missing data by generating data within or across modalities. The hybrid model involves augmenting the temporal generative model with a temporal discriminative model for event detection, and classification, which enables modeling long range temporal dynamics. We evaluate our approach on audio-visual datasets (AVEC, AVLetters, and CUAVE) and demonstrate its superiority compared to the state-of-the-art.

# Amer M R, Siddiquie B, Khan S, et al. Multimodal fusion using dynamic hybrid models[C]// Applications of Computer Vision. IEEE, 2014:556-563.

93.This paper describes our transition from a speech-only dialogue system to a multimodal one. Our description focuses on the fusion of input modalities coming from different channels. Two strategies have been implemented for comparison purposes: the first solution is largely based on Johnston’s work [Johnston et al. 1997, Johnston 1998], and involves modifying our parser to cope with simultaneous multimodal inputs, and to include temporal constraints at unification level. The second implementation proposes an original solution to the problem, and involves combining inputs coming from different multimodal channels at dialogue level. This solution is based on an implementation of the ISU approach [Traum et al. 1999, Amores et al. 2001]. These two strategies have been implemented in an InformationState-Update-based system, combining both speech and graphical inputs. A multimodal “Smart House” scenario where the user interacts with the system using a microphone and a touch-screen has been chosen. The paper includes a high-level description of the algorithms implemented and concludes with a theoretical analysis of the advantages and drawbacks of both approaches.

# Pérez G, Amores G, Manchón P. Two strategies for multimodal fusion[J]. 2005.

94. Persuasiveness is a high-level personality trait that quantifies the influence a speaker has on the beliefs, attitudes, intentions , motivations, and behavior of the audience. With social multimedia becoming an important channel in propagating ideas and opinions, analyzing persuasiveness is very important. In this work, we use the publicly available Persuasive Opinion Multimedia (POM) dataset to study persuasion. One of the challenges associated with this problem is the limited amount of annotated data. To tackle this challenge, we present a deep multimodal fusion architecture which is able to leverage complementary information from individual modalities for predicting persuasiveness. Our methods show significant improvement in performance over previous approaches.

# Nojavanasghari B, Gopinath D, Koushik J, et al. Deep multimodal fusion for persuasiveness prediction[C]// ACM International Conference on Multimodal Interaction. ACM, 2016:284-288.

95. This paper deals with two main research focuses on Affective Computing: facial emotion recognition and multimodal fusion of affective information coming from different channels. The facial sensing system developed implements an emotional classification mechanism that combines, in a novel and robust manner, the five most commonly used classifiers in the field of affect sensing, obtaining at the output an associated weight of the facial expression to each of the six Ekman’s universal emotional categories plus the neutral. The system is able to analyze any subject, male or female, of any age, and ethnicity and has been validated by means of statistical evaluation strategies, such as cross-validation, classification accuracy ratios and confusion matrices. The categorical facial sensing system has been subsequently expanded to a continuous 2D affective space which has made it also possible to face the problem of multimodal human affect recognition. A novel fusion methodology able to fuse any number of affective modules, with very different time-scales and output labels, is proposed. It relies on the 2D Whissell affective space and is able to output a continuous emotional path characterizing the user’s affective progress over time. A Kalman filtering technique controls this path in real-time to ensure temporal consistency and robustness to the system. Moreover, the methodology is adaptive to eventual temporal changes in the reliability of the different inputs’ quality. The potential of the multimodal fusion methodology is demonstrated by fusing dynamic affective information extracted from different channels (video, typed-in text and emoticons) of an Instant Messaging tool.

# Cerezo E, Hupont I, Baldassarri S, et al. Emotional facial sensing and multimodal fusion in a continuous 2D affective space[J]. Journal of Ambient Intelligence & Humanized Computing, 2012, 3(1):31-46.

96. Person identification is of great interest for various kinds of applications and interactive systems. In our system we use face recognition and voice recognition from data recorded in an interactive dialogue system. In such a system, sequential images and sequential utterances can be used to improve recognition accuracy over single hypotheses. The presented approach uses confidence-based fusion for sequence hypotheses, for multimodal fusion, and to provide a reliability measure of the classification quality that can be used to decide when to trust and when to ignore classification results.

# Holzapfel H, Waibel A. Confidence based multimodal fusion for person identification[C]// ACM International Conference on Multimedia. ACM, 2008:885-888.

97. The retrieving method proposed in this paper utilizes the fusion of the images' multimodal information (textual and visual) which is a recent trend in image retrieval researches. It combines two different data mining techniques to retrieve semantically related images: clustering and association rules mining algorithm. The semantic association rules mining is constructed at the offline phase where the association rules are discovered between the text semantic clusters and the visual clusters of the images to use it later at the online phase. The experiment was conducted on more than 54,500 images of ImageCLEF 2011 Wikipedia collection. It was compared to an online image retrieving system called MMRetrieval and to the proposed system but without using association rules. The obtained results show that the proposed method achieved the best precision score among different query categories.

# Alghamdi R A, Taileb M, Ameen M. A new multimodal fusion method based on association rules mining for image retrieval[C]// Electrotechnical Conference. IEEE, 2014:493-499.

98. In this work, the authors aim to detect social events from Web images by devising a semi-supervised multimodal fusion model, denoted as SMF. With a multimodal feature fusion layer and a feature reinforcement layer, SMF learns feature histograms to represent the images, fusing multiple heterogeneous features seamlessly and efficiently. Particularly, a self-tuning approach is proposed to tune the parameters in the process of feature reinforcement automatically. Furthermore, to deal with missing values in raw features, prior knowledge is utilized to estimate the missing ones as a preprocessing step, and SMF will further extend an extra attribute to indicate if the values in the fused feature are missing. Based on the fused expression achieved by SMF, a series of algorithms are designed by adopting clustering and classification strategies separately. Extensive experiments conducted on the MediaEval social event detection challenge reveal that SMF-based approaches outperform the baselines.

# Yang Z, Li Q, Lu Z, et al. Semi-Supervised Multimodal Fusion Model for Social Event Detection on Web Image Collections[J]. International Journal of Multimedia Data Engineering & Management, 2015, 6(4):1-22.

99.

Content-based video copy detection algorithms (CBCD) focus on detecting video segments that are identical or transformed versions of segments in a known video. In recent years some systems have proposed the combination of orthogonal modalities (e.g. derived from audio and video) to improve detection performance, although not always achieving consistent results. In this paper we propose a fusion algorithm that is able to combine as many modalities as available at the decision level. The algorithm is based on the weighted sum of the normalized scores, which are modified depending on how well they rank in each modality. This leads to a virtually parameter-free fusion algorithm. We performed several tests using 2010 TRECVID VCD datasets and obtain up to 46% relative improvement in min-NDCR while also improving the F1 metric on the fused results in comparison to just using the best single modality.

# Anguera X, Barrios J M, Adamek T, et al. Multimodal fusion for video copy detection[C]// International Conference on Multimedea 2011, Scottsdale, Az, Usa, November 28 - December. DBLP, 2011:1221-1224.

100.

A robot that interacts with a human has to be able to interpret information from various input channels: it needs to understand and analyse the utterances by the human, it has to keep track of its own environment using sensors, and it needs to incorporate background knowledge about the task it was built for. Typically, a human-robot interaction system has various specialised system components that implement these abilities. Thus, the robot also needs to merge the information from its input channels so that it is able to complete its assigned task. This integration of information from input channels is called multimodal fusion.This thesis presents two approaches for multimodal fusion for a robot that jointly cooperates with a human partner. The first approach, which is called classical multimodal fusion, focusses on processing human utterances. For that, the robot processes speech and gestures of its human partner using methods from classical artificial intelligence to yield logical representations of the utterances. Following that, these representations are enhanced with further information from other input modalities of the robot. In contrast to that, in the second approach the robot generates representations for its own actions in relation to objects in its environment, so-called embodied multimodal fusion. Here, the system uses the data from its input channels to evaluate the relevance of its own actions for a given context. After a literature review, this thesis discusses the theoretical basis of both multimodal fusion approaches and presents how these methods can be implemented on a robot that is able to work together with a human on a common construction task, for which it processes multimodal input. These implementations were used in three human-robot interaction studies, in which naive subjects worked together with the robot. The experiments were executed to study different aspects of joint action between human and robot. The results of the experiments reveal several interesting facts: the first experiment studies how the robot can explain building plans to the human. The results of the study show that the users preferred a plan explanation strategy in which the robot first names the target object and after that explains the single building steps. The first as well as the second experiment study the generation of referring expression in two different contexts. The results of the studies suggest that experiment participants rate the robot as a better dialogue partner when the robot makes full use of context information to generate referring expressions. Finally, the third experiment studies how humans perceive different roles of the robot in the interaction. The study shows that the users equally accept the robot as an instructor or as an equal partner and simply adjust their own behaviour to the robot's role.

# Giuliani M. Comparing Classical and Embodied Multimodal Fusion for Human-Robot Interaction[J]. Bone, 2011, 44(12):2067-2074.