Fusão de duas modalidades para reconhecimento de emoções na natureza

Documento de Conferência · Outubro 2019

DOI: 10.1145/3340555.3355719

8 autores, incluindo:

Yuan Zong
Universidade do Sudeste (China)
85 PUBLICAÇÕES 2.164 CITAÇÕES
VER PERFIL

Cheng Lu
Universidade do Sudeste (China)
23 PUBLICAÇÕES 306 CITAÇÕES
VER PERFIL

Chuangao Tang
Universidade do Sudeste (China)
32 PUBLICAÇÕES 375 CITAÇÕES
VER PERFIL

Xingxun Jiang
Universidade do Sudeste (China)
15 PUBLICAÇÕES 73 CITAÇÕES
VER PERFIL

Fusão Bimodal para Reconhecimento de Emoções no Selvagem

Sunan Li
School of Information Science and Engineering, Southeast University Nanjing, China
230189473@seu.edu.cn

Wenming Zhengÿ
Laboratório Chave de Desenvolvimento Infantil e Ciência da Aprendizagem do Ministério da Educação Universidade do Sudeste Nanjing, China
wenming_zheng@seu.edu.cn

Yuan Zong
School of Biological Science and Medical Engineering, Southeast University Nanjing, China
xhzongyuan@seu.edu.cn

Cheng Lu
Escola de Ciência da Informação e Engenharia, Universidade do Sudeste Nanjing, China
cheng.lu@seu.edu.cn

Chuangao Tang
School of Biological Science and Medical Engineering, Southeast University Nanjing, China tcg2016@seu.edu.cn

Xingxun Jiang
School of Biological Science and Medical Engineering, Southeast University Nanjing, China
jiangxingxun@seu.edu.cn

Jiateng Liu
School of Biological Science and Medical Engineering, Southeast University Nanjing, China
jiangxingxun@seu.edu.cn

Wanchuang Xia
School of Cyber Science and Engineering, Southeast University Nanjing, China
220184474@seu.edu.cn

RESUMO

O reconhecimento de emoções na natureza tem sido um tópico de pesquisa quente no campo da computação afetiva. Embora alguns progressos tenham sido alcançados, o reconhecimento de emoções na natureza ainda é um problema não resolvido devido ao desafio do movimento da cabeça, deformação facial, variação de iluminação, etc.

Para lidar com esses desafios irrestritos, propomos um método de fusão bimodal para reconhecimento de emoções baseado em vídeo na natureza. Os modelos de reconhecimento de objetos baseados em CNN de última geração são empregados para facilitar o desempenho do reconhecimento de expressões faciais . Uma memória bidirecional de longo prazo (Bi-LSTM) é empregada para capturar informações dinâmicas dos recursos aprendidos. o melhor experimental

O resultado mostra que a precisão geral do nosso algoritmo no conjunto de dados de teste do desafio EmotiW é de 62,78%, o que supera o melhor resultado do EmotiW2018 e ocupa o 2º lugar no desafio EmotiW2019.

CONCEITOS CCS

• **Organização de sistemas computacionais** ÿ **Sistemas embarcados;** Redundância; Robótica; • **Redes** ÿ Confiabilidade da rede .

PALAVRAS-CHAVE

Reconhecimento de Emoções; Aprendizado Profundo; Neural Convolucional Redes

Formato de referência ACM:

Sunan Li, Wenming Zheng, Yuan Zong, Cheng Lu, Chuangao Tang, Xingxun Jiang, Jiateng Liu e Wanchuang Xia. 2019. Fusão de duas modalidades para reconhecimento de emoções na natureza. Em 2019 Conferência Internacional sobre Interação Multimodal (ICMI'19), 14 a 18 de outubro de 2019, Suzhou, China. ACM, Nova York, NY, EUA, 5 páginas. https://doi.org/ 10.1145/3340555.3355719

ÿO autor correspondente.

ICMI'19, 14 a 18 de outubro de 2019, Suzhou,
China
ACM ISBN 978-1-4503-6860-5/19/10. . . $ 15,00
https://doi.org/10.1145/3340555.3355719

1. INTRODUÇÃO

O reconhecimento de emoções desempenha um papel fundamental na interação humano-computador e tem sido investigado por muitos anos. Existem vários tipos de modalidade de dados emocionais, incluindo expressão facial, emoção de áudio, EEG, EMG e etc.

Nessas modalidades de dados, a gravação de vídeo tem a vantagem de não ter contato, o que leva à sua ampla aplicação. Portanto , o reconhecimento de emoções baseado em áudio e vídeo atraiu o interesse de muitos pesquisadores.

O subdesafio de reconhecimento de emoções baseado em áudio e vídeo na natureza (EmotiW) foi realizado sete vezes desde 2013. Os videoclipes usados neste desafio foram extraídos de filmes para simular emoções no ambiente do mundo real [1, 2], e muitas pesquisas foram feitas no desafio EmotiW [4, 18]. Para treinar um modelo robusto de reconhecimento de emoções , os pesquisadores propuseram vários tipos de métodos, incluindo recursos artesanais e recursos baseados em aprendizado profundo . Kaya et al. [8] empregou recursos tradicionais e aprendizes baseados em mínimos quadrados para reconhecimento de emoções. Wu e outros. [19] usou vários modelos de recursos para codificar videoclipes.
Existem também algumas literaturas com foco em recursos artesanais (por exemplo, filtros de Gabor, padrões binários locais) [10, 21] para resolver o problema do reconhecimento de emoções.

Nos últimos anos, pesquisadores têm trabalhado em redes neurais de convolução (CNN) para tarefas de visão computacional e reconhecimento de padrões. Alexnet[9], GoogLeNet[17], VGG[16] e ResNet[5] foram propostos para lidar com problemas em tarefas visuais. Para capturar informações dinâmicas da expressão facial nas sequências, redes neurais recorrentes foram desenvolvidas para resolver esse problema. No desafio EmotiW 2018, Liu et al. [11] empregou uma estrutura baseada em vários recursos usando recursos acústicos e faciais que alcançaram uma precisão de 61,87% no conjunto de teste. Motivados por seu trabalho, neste artigo propomos um framework contendo modelo de imagem facial e modelo de áudio.

O modelo de imagem consiste em quatro tipos diferentes de redes neurais para extrair recursos e, em seguida, os recursos são alimentados em Bi-direção LSTM para capturar informações temporais dinâmicas para uma maior precisão [3]. Além disso, extraímos as informações complementares do áudio com dois métodos diferentes. No estágio de fusão subsequente, a estratégia de busca em grade é empregada para otimização de desempenho no conjunto de validação, e o método de fusão de informações melhora a precisão para 62,78% no conjunto de teste e o desempenho ocupa o 2º lugar no subdesafio baseado em áudio e vídeo no EmotiW2019.

O restante deste artigo está organizado da seguinte forma. O modelo de imagem facial e áudio é apresentado na seção 2. Na seção 3, expomos nossos resultados experimentais no conjunto de dados do desafio para avaliar nossa estrutura proposta. Finalmente , na seção 4, concluímos este artigo.

2 O MÉTODO PROPOSTO

Nossa estrutura proposta consiste em duas partes, ou seja, fluxo de imagem e fluxo de áudio. Quanto ao fluxo de imagens, existem três redes baseadas em CNN para aprendizado de recursos espaço-temporais exclusivos, enquanto o restante se concentra na extração de informações espaciais. As pontuações do fluxo de imagem e do fluxo de áudio são ponderadas com uma otimização de pesquisa em grade. A pontuação final ponderada é utilizada para a classificação. A estrutura é mostrada na Fig. 1 e os detalhes são descritos a seguir.

O modelo de imagem facial As sequências faciais cortadas são alimentadas nas redes neurais convolucionais, ou seja, VGG-Face, Restnet18, Densenet121 e VGG16, para extração de recursos espaciais. Nas três primeiras redes , as redes neurais convolucionais empilhadas são conectadas com um Bi-LSTM [6, 12] de duas camadas, que pode capturar informações dinâmicas. Já para o modelo CNN estático, o rótulo previsto de cada quadro é usado para o cálculo da frequência e a frequência normalizada é usada como uma pontuação de cada tipo de emoção.

Rede de extração de características espaciais baseada em CNN. No ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2014, o Visual Geometry Group propôs o VGGNet. O modelo VGG-Face é um tipo especial de VGG com 16 camadas treinadas em um banco de dados especial desenvolvido pelo grupo de geometria visual também [14]. Usamos o VGG-Face aqui para capturar as características emocionais das imagens faciais. Para obter um melhor resultado de reconhecimento, ajustamos o modelo pré-treinado em imagens de expressões faciais cortadas do conjunto de dados de treinamento. Como pesquisas anteriores foram feitas por Yosinski et al na Cornell University, os recursos extraídos de camadas mais profundas de redes tendem a ser mais específicos em tarefas do que camadas rasas[22]. Portanto, congelamos os pesos de todas as camadas convolucionais e atualizamos apenas nas camadas totalmente conectadas na rede VGG-Face pelo processamento de imagens faciais.

Embora com o aumento da profundidade da rede, dois problemas são inevitáveis. Os gradientes podem desaparecer/explodir e a precisão pode ficar saturada e depois degradar rapidamente. Para resolver esses dois problemas, He et al. propôs a estrutura residual profunda chamada 'ResNet'[5]. A concepção de 'conexões de atalho ' foi proposta pela primeira vez no ResNet e pode mudar a regra de ajuste tradicional. O principal papel das 'conexões de atalho ' era fazer com que os mapas de entrada fossem ping de mapeamento de identidade. Além disso, usamos o ResNet-18 para construir nossa estrutura devido à falta de dados de treinamento para evitar o ajuste excessivo.

O número de parâmetros do ResNet é substancialmente maior porque cada camada tem seus próprios pesos e o ResNet usa apenas recursos de saída da camada inferior adjacente como entrada, o que limita a reutilização de mapas de recursos. Para resolver esse problema, Huang et al propuseram o DenseNet, que ganhou o prêmio de melhor artigo no CVPR-2017[7]. O nome DenseNet significa que cada camada obtém entradas adicionais de todas as camadas anteriores e passa seus próprios mapas de recursos para todas as camadas subsequentes, o que significa que elas são mais densamente conectadas do que outras redes. Comparado com o ResNet, o DenseNet incentiva a reutilização de recursos e alivia o problema do gradiente de desaparecimento.

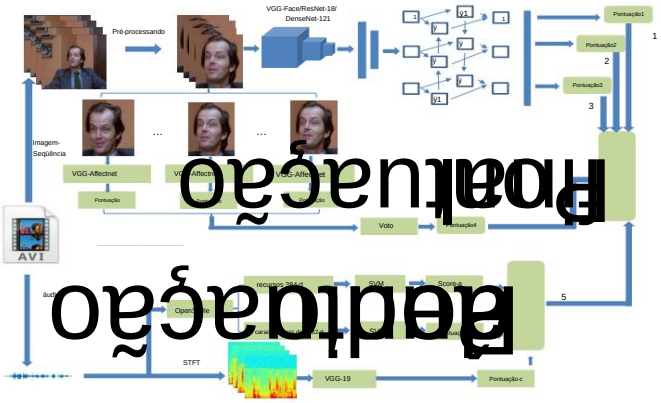

Figura 1: Visão geral da estrutura proposta.

As três redes mencionadas acima melhoram o desempenho da estrutura proposta com base na estrutura de rede. Para resolver o problema do desequilíbrio da amostra, introduzimos um novo banco de dados de emoções chamado AffectNet. O Affect da Internet é o maior banco de dados de modelos categóricos e dimensionais de afeto na natureza[13]. Nós o usamos para treinar um modelo VGG-16 para melhorar a robustez de nossa estrutura . E o modelo VGG-16 treinado produz um vetor de pontuação correspondente à probabilidade da imagem de entrada pertencer a cada emoção. Por fim, para os quadros selecionados de um videoclipe, usando o voto de maioria simples para determinar a qual emoção o vídeo pertence.

Rede de extração de recursos dinâmicos. As quatro redes mencionadas acima extraem o recurso de emoção tanto na dimensão espacial , agora precisamos de uma estrutura para combinar recursos extraídos de diferentes imagens que recortam de um vídeo para capturar informações dinâmicas. Para resolver esse problema, usamos o LSTM para extrair informações dinâmicas, o núcleo do LSTM é o estado da célula, no qual existem três portas (porta de entrada, porta de esquecimento e porta de saída) utilizadas no LSTM para atualizar o estado da célula para resolver o problema problema de problema de memória de longo prazo surgiu em RNN [6, 12]. A saída dessas três portas é mostrada na fórmula (4), onde ft ,it ,ot representa a saída da porta de entrada, porta de esquecimento e porta de saída no tempo t, respectivamente.

$$ft = ÿ\ Wf\ [htÿ1, xt\ ] + bf\ it = ÿ\ (Wi\ [htÿ1, xt\ ] + bi)\ ot = ÿ\ (Wo\ [htÿ1, xt\ ] + bo\ ) \quad (1)$$

Finalmente, o estado oculto pode ser formulado como formula(2), onde Ct representa o estado celular no tempo t.

$$ht = ot\ ÿ\ tanh\ (Ct\ ) \quad (2)$$

Para extrair as informações dinâmicas do vídeo de forma mais robusta, usamos LSTM bidirecionais que usam a sequência de quadros de uma direção e seu reverso como entrada. Então a saída de Bi-LSTM pode ser formulado como:

$$= h + h^{fb}_{t}\ ht_{T-t} \quad (3)$$

Esses recursos que constituem as sequências são obtidos a partir da última camada totalmente conectada na CNN, e as dimensões dos recursos são todas 4096.

O modelo de áudio

O áudio é uma parte significativamente importante no desafio EmotiW, devido à falta de amostras de algumas emoções específicas no banco de dados de treinamento e à semelhança de entrada, diferentes modelos de imagem funcionam de maneira tão semelhante que é difícil classificar algumas amostras e emoções corretamente. A aprendizagem de sinais de áudio desempenha um papel importante na melhoria do desempenho do nosso modelo de classificação de emoções em vídeo. Nesta estrutura, dois métodos foram usados como extração e classificação de recursos, o primeiro é um modelo de aprendizado profundo no qual extraímos a fala útil do vídeo e a convertemos em espectrograma por transformada de Fourier de tempo curto. Em seguida, processe o espectrograma como uma imagem e alimente-o em uma rede VGG-19[15] para extrair informações de profundidade e classificá-las por softmax. Em segundo lugar, para melhorar a robustez do modelo de áudio, também usamos a caixa de ferramentas openSMILE para extrair o recurso LL

Tabela 1: Precisão de reconhecimento de cada modelo no banco de dados de validação e teste.

modelos	Validação(%)	Teste(%)
VGG-Face+BLSTM	53,91%	-
ResNet-18+BLSTM	43,34%	-
DenseNet-121+BLSTM	49,35%	-
VGG-AffectNet	41,78%	-
Fusão de áudio	25,59%	-
Fusão	54,30%	62,78%

comumente utilizados no processamento de áudio. Extraímos dois tipos diferentes de features do openSMILE cujas dimensões são 384 e 1582 respectivamente. Esses dois recursos são dois recursos padrão que contêm Mel Frequency Ceptral Coefficien (MFCC), frequência fundamental, etc. Em seguida, use o Support Vector Machine (SVM) para processar esses dois recursos e gerar duas pontuações correspondentes.

Por fim, fundimos três pontuações diferentes como pontuação final da fala. Portanto, não apenas aproveitamos a forte vantagem do recurso profundo baseado na representação de frequência de tempo do espectrograma e CNN, mas também usamos os recursos LLD classificados pelo SVM para melhorar a robustez do nosso modelo de áudio.

A estratégia de fusão

Nas seções anteriores, apresentamos diferentes redes que usamos em nosso framework, cada uma delas desempenhando um papel importante e complementar na classificação da emoção. Portanto, temos que usar uma estratégia para fundir a saída de cada rede. Em nossa estrutura, cada rede produzirá um vetor de pontuação para cada amostra que doa a probabilidade da amostra pertencer à emoção específica. Neste artigo, usamos a soma ponderada para fundir todas as pontuações e obter a pontuação final conforme mostrado na fórmula (4) .

$$S\hat{} = ÿi \cdot Si \quad (4)$$

Onde Sˆ é o vetor de pontuação final e ÿi e S são os vetores de peso e pontuação da rede i, respectivamente. Para um melhor desempenho, usamos o método de pesquisa de grade na base de dados de validação para encontrar os pesos adequados, e o peso final é 0,88,0,34,0,22,0,06,0,3 para VGG-Face, ResNet-18, DenseNet 121, VGG-Affectnet e áudio respectivamente.

3 EXPERIMENTO

Pré-processamento

O conjunto de dados do subdesafio de áudio e vídeo do EmotiW2019 é afetado por condições anormais, como variação de iluminação, oclusão facial e assim por diante. Portanto, o pré-processamento é necessário em nossa estrutura. Primeiramente, alinhamos e normalizamos a face através do bloco de identificação do Seetaface usando cinco pontos de referência no MTCNN e redimensionamos a imagem facial em 256×256[20, 23]. Em seguida, cortamos a imagem em 224 × 224 aleatoriamente. Finalmente, para cada vídeo, selecionamos 16 quadros como entrada para ResNet18 VGG-Face e VGG-AffectNet, 4 quadros para DenseNet-121. Da mesma forma, para o áudio, primeiro removemos o ruído irrelevante e de fundo. Em seguida, a fala é transformada em imagens de espectrograma por transformada de Fourier de tempo curto e redimensionada para 224×224 para o modal VGG-19.

Resultado e discussão

No subdesafio de áudio e vídeo do EmotiW2019, adotamos diferentes redes neurais para extrair recursos complementares para um desempenho melhor e robusto. Os resultados das diferentes redes que propusemos na seção 2 foram mostrados na Tabela 1 , respectivamente, cujo melhor modelo: VGG-Face alcançou precisão geral de 53,91%. Vale a pena notar que, embora a precisão do modelo de fusão de áudio seja menor do que outras redes, o modelo de fusão de áudio ainda desempenha um papel importante devido às suas informações complementares. Por fim, o melhor quadro de fusão alcançou 62,78% no banco de dados de teste, 0,91% a mais que o campeão do EmotiW2018.

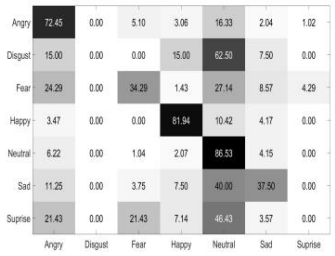

Figura 2: Matriz de confusão da 4ª submissão.

A matriz de confusão das redes de fusão é mostrada na Fig. 2. De acordo com a matriz de confusão apresentada, existem três emoções (nojo, medo e surpresa) difíceis de classificar corretamente, talvez devido à falta de dados de treinamento e à implicidade das características dessas três emoções.

4. CONCLUSÃO

Neste artigo, apresentamos um trabalho de estrutura multifuncional de bimodalidade para reconhecer emoções em estado selvagem no EmotiW2019. As informações sobre emoções são divididas em dois aspectos complementares: áudio e vídeo. Para informações de vídeo, usamos quatro redes diferentes para extrair o recurso de emoção. E, para áudio, usamos STFT e openSMILE separadamente. Em seguida, combinamos diferentes pontuações por soma ponderada, onde o peso de cada rede dependia de sua contribuição para o melhor resultado. O resultado do experimento do desafio mostrou que o framework proposto é mais robusto na tarefa de reconhecimento de emoções na natureza.

5 RECONHECER

Este trabalho foi financiado em parte pelo Programa Nacional de Pesquisa e Desenvolvimento Chave da China sob Grant 2018YFB1305200, em parte pelo Programa Nacional de Pesquisa Básica da China sob Grant 2015CB351704, em parte pela Fundação Nacional de Ciências Naturais da China sob Grant 61572009, e em parte pelos Fundos de Pesquisa Fundamental para as Universidades Centrais sob Grant 2242018K3DN01 e Grant 2242019K40047.

REFERÊNCIAS

[1] Abhinav Dhall, Roland Goecke, Shreya Ghosh e Tom Gedeon. 2019. EmotiW 2019: Tarefas automáticas de previsão de emoção, engajamento e coesão . Na Conferência Internacional da ACM sobre Interação Multimodal 2019.

[2] Abhinav Dhall, Roland Goecke, Simon Lucey e Tom Gedeon. 2012. Coletando bancos de dados de expressões faciais grandes e ricamente anotados de filmes. IEEE MultiMedia 19, 3 (2012), 34–41.

[3] Alex Graves e Jürgen Schmidhuber. 2005. Classificação de fonemas Framewise com LSTM bidirecional e outras arquiteturas de redes neurais . Redes Neurais 18, 5 (2005), 602 – 610. IJCNN 2005.

[4] Da Guo, Kai Wang, Jianfei Yang, Kaipeng Zhang, Xiaojiang Peng e Yu Qiao. 2019. Explorando regularizações com dicas de rosto, corpo e imagem para previsão de coesão de grupo. In Proceedings of the 21th ACM International Conference on Multimodal Interaction (no prelo). ACM.

[5] K. He, X. Zhang, S. Ren e J. Sun. 2016. Deep Residual Learning for Image Recognition. Em 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 770-778.

[6] Sepp Hochreiter e Jürgen Schmidhuber. 1997. Memória de longo prazo . Computação neural 9, 8 (1997), 1735–1780.

[7] G. Huang, Z. Liu, L. vd Maaten e KQ Weinberger. 2017. Redes Convolucionais Densamente Conectadas. Em 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[8] Heysem Kaya, Furkan Gürpinar, Sadaf Afshar e Albert Ali Salah. 2015. Contrastando e Combinando Alunos Baseados em Mínimos Quadrados para Reconhecimento de Emoções na Natureza. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15). ACM, Nova York, NY, EUA, 459-466.

[9] Alex Krizhevsky, Ilya Sutskever e Geoffrey E. Hinton. 2017. Classificação Ima geNet com Redes Neurais Convolucionais Profundas. Comun . ACM 60, 6 (maio de 2017), 84–90.

[10] Markus Kächele, Martin Schels, Sascha Meudt, Günther Palm e Friedhelm Schwenker. 2016. Revisitando o desafio EmotiW: quão selvagem é realmente? Journal on Multimodal User Interfaces 10, 2 (2016), 1–12.

[11] Chuanhe Liu, Tianhao Tang, Kui Lv e Minghao Wang. 2018. Reconhecimento de emoções baseado em vários recursos para videoclipes. In Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18). ACM, Nova York, NY, EUA, 630–634.

[12] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký e Sanjeev Khudanpur. 2015. Modelo de linguagem baseado em rede neural recorrente . In Interspeech, Conference of the International Speech Communication Association, Makuhari, Chiba, Japão, setembro.

[13] A. Mollahosseini, B. Hasani e MH Mahoor. 2019. AffectNet: um banco de dados para expressão facial, valência e computação de excitação na natureza. IEEE Transactions on Affective Computing 10, 1 (janeiro de 2019), 18–31.

[14] Omkar M. Parkhi, Andrea Vedaldi e Andrew Zisserman. 2015. Reconhecimento facial profundo. Na Conferência Britânica de Visão de Máquina.

[15] RV Shannon, FG Zeng, . Kamath, V., . Wygonski, J., e . Ekelid, M. 1995. Reconhecimento de fala com pistas principalmente temporais. Science 270, 5234 (1995), 303–304.

[16] Karen Simonyan e Andrew Zisserman. 2014. Redes Convolucionais Muito Profundas para Reconhecimento de Imagens em Grande Escala. Ciência da Computação (2014).

[17] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke e A. Rabinovich. 2015. Indo mais fundo com convoluções. Em 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 1–9.

[18] Kai Wang, Jianfei Yang, Da Guo, Kaipeng Zhang, Xiaojiang Peng e Yu Qiao. 2019. Bootstrap Model Ensemble e Rank Loss for Engagement Intensity Regression. In Proceedings of the 21th ACM International Conference on Multimodal Interaction (no prelo). ACM.

[19] Jianlong Wu, Zhouchen Lin e Hongbin Zha. 2015. Fusão de vários modelos para reconhecimento de emoções na natureza. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15). ACM, Nova York, NY, EUA, 475–481.

[20] Shuzhe Wu, Meina Kan, Zhenliang He, Shiguang Shan e Xilin Chen. 2017. Cascata estruturada em funil para detecção facial multivisualização com reconhecimento de alinhamento. Neurocomputação 221 (2017), 138 – 145.

[21] Anbang Yao, Junchao Shao, Ningning Ma e Yurong Chen. 2015. Capturando características faciais com reconhecimento de AU e suas relações latentes para reconhecimento de emoções na natureza. In ACM na Conferência Internacional sobre Interação Multimodal.

[22] Jason Yosinski, Jeff Clune, Yoshua Bengio e Hod Lipson. 2014. Quão transferíveis são os recursos em redes neurais profundas? Eprint Arxiv 27 (2014), 3320–3328.

[23] K. Zhang, Z. Zhang, Z. Li e Y. Qiao. 2016. Detecção de face conjunta e alinhamento usando redes convolucionais em cascata multitarefa. IEEE Signal Processing Letters 23, 10 (outubro de 2016), 1499–1503.