Detecção de Rosto Articular e Reconhecimento de Expressão Facial com MTCNN

Empresa: IEEE

2 Autor (es) Jia Xiang ; Gengming Zhu **Ver todos os autores**

Abstract: The Multi-task Cascaded Convolutional Networks (MTCNN) has recently demonstrated impressive results on jointly face detection and alignment. By using the hard sample ming... **View more**

Metadata

Abstrato:
As redes convolucionais em cascata multitarefas (MTCNN) demonstraram recentemente resultados impressionantes na detecção e alinhamento de faces em conjunto. Usando a amostra difícil ming e treinando um modelo nos conjuntos de dados FER2013, exploramos a correlação inerente entre a detecção de rosto e o reconhecimento de íons de expressão facial e relatamos os resultados do reconhecimento de expressões faciais com base no MTCNN.

Publicado em: 2017 4ª Conferência Internacional de Ciência da Informação e Engenharia de Controle (ICISCE)

Data da conferência: 21-23 de julho de 2017

Data em que foi adicionado ao IEEE *Xplore* : 16 de novembro de 2017

Informação ISBN:

Número de acesso INSPEC **:** 17376097

DOI: 10.1109 / ICISCE.2017.95

Empresa: IEEE

Local da Conferência: Changsha, China

SEÇÃO I.
Introdução

Os seres humanos interagem entre si principalmente através da fala, mas também através de gestos corporais, para enfatizar certas partes da fala e mostrar emoções. Uma das maneiras importantes pelas quais os seres humanos demonstram emoções é através de expressões faciais, que são essenciais para a comunicação não verbal entre os seres humanos. Computadores e outros dispositivos eletrônicos em nossas vidas diárias se tornarão mais amigáveis ao usuário, se puderem interpretar adequadamente as expressões faciais de uma pessoa, melhorando assim as interfaces homem-máquina. O reconhecimento de expressão facial pode ser implementado em todas as interfaces de computador.

Consequentemente, o reconhecimento da expressão facial (FER) tem sido amplamente estudado e um progresso significativo foi feito nesse campo. Em um artigo de 1971 intitulado “Constantes entre culturas no rosto e na emoção”, Ekman et al. identificamos seis expressões faciais que são universais em todas as culturas: raiva, nojo, medo, felicidade, tristeza e surpresa [1] . Como visto nos artigos descritos anteriormente, a maioria das abordagens desenvolvidas para resolver o FER usa recursos personalizados para sequências curtas de expressões faciais, e houve vários avanços nos últimos anos em termos de detecção de rosto, mecanismos de extração de recursos e técnicas utilizadas para FER .

As Redes Neurais Convolucionais (CNNs) ganharam muita popularidade nos últimos anos para aplicações relacionadas à visão e têm o potencial de obter algumas das maiores precisões no FER [2] . No entanto, a maioria dos métodos disponíveis de detecção de rosto e reconhecimento de expressão facial ignora a correlação inerente entre essas duas tarefas., Mas existem vários trabalhos existentes que tentam resolver em conjunto a detecção e o alinhamento de faces [3] , [4] . O famoso método, representado pelas redes de convolução em cascata de múltiplas tarefas (MTCNN) de Kiapeng Zhang, et al. Demonstra um resultado impressionante na detecção e alinhamento de faces (JDA). Ele propôs CNNs em cascata para integrar essas duas tarefas através do aprendizado de múltiplas tarefas.

Neste artigo, exploramos a correlação inerente entre MTCNN baseado em detecção de rosto e reconhecimento de expressão facial e implementamos as duas tarefas. A entrada no nosso sistema é uma imagem; então, usamos essa estrutura para prever a localização do rosto e o rótulo da expressão facial, que devem ser um desses rótulos: raiva, felicidade, medo, tristeza, nojo e neutro, e comparar com uma variedade de outros detectores recentes de alto desempenho.

SEÇÃO II
Trabalho relatado

A detecção de rosto é um dos problemas mais estudados em visão. Viola e Jones [5] propuseram um detector de rosto em cascata, que é a primeira vez que aplica os recursos do tipo Haar no AdaBoost para o treinamento de classificadores em cascata. Ele tem um desempenho prático em tempo real com maior precisão do que antes, mas não é capaz de lidar efetivamente com rostos não frontais e selvagens. Para superar esse problema, métodos de recursos mais robustos foram propostos, como HOG, Gabor, SIFT e SURF. Além disso, uma estratégia simples é combinar vários recursos para aprimorar a robustez da detecção. Zhu et al. [6]propuseram vários modelos de peças deformáveis para detectar faces com diferentes vistas e expressões. No entanto, eles custam muito tempo para treinar e são limitados na variedade de cenas.

Recentemente, a CNN geralmente obtém um desempenho melhor do que os métodos tradicionais de artesanato em tarefas de visão computacional. [7] usaram CNNs em cascata para detecção de face, é

uma estrutura em cascata reforçada. [8] aplicaram o Faster R-CNN [9] , um dos mais modernos na detecção de objetos genéricos, realizaram otimização de ponta a ponta e obtiveram resultados promissores. [10] construíram um modelo para realizar a detecção de faces em paralelo com o alinhamento de faces e alcançaram alto desempenho em termos de precisão e velocidade.

O problema de classificar emoções a partir de expressões faciais em imagens é amplamente estudado. Semelhante a alguns dos métodos de detecção de faces, o FER também pode ser implementado pelos métodos de recursos de artesanato. Um dos primeiros trabalhos a aplicar redes neurais para esse fim é o trabalho EMPATH de 2002 [11], prosseguindo com a filtragem Gabor nas imagens brutas, seguida de várias transformações e PCA antes de aplicar uma rede neural de 3 camadas. Vários trabalhos recentes sobre o FER utilizam com sucesso CNNs para extração e inferência de recursos. Yu e Zhang [12] alcançaram resultados de ponta no EmotiW em 2015 usando CNNs para realizar FER. Eles usaram um conjunto de CNNs com cinco camadas convolucionais. [13]também alcançaram resultados avançados no FER. Sua rede consistia em duas camadas convolucionais, max-pooling e 4 Inception, como introduzido pelo GoogLeNet.

A CNN alcançou algumas das mais altas precisões na detecção de rosto e no reconhecimento de expressão facial. Neste trabalho, principalmente a Referência [4] , inerente à detecção de faces e reconhecimento de expressões faciais com base no MTCNN.

SEÇÃO III
Visão geral do MTCNN

A. Visão geral

O MTCNN, que herdou a correlação entre a detecção e o alinhamento de faces para aumentar seu desempenho. Consiste essencialmente em três partes: (1) uma rede de propostas (P-Net) para gerar uma lista das janelas candidatas. Em seguida, usamos esse método para classificar os vetores de regressão de caixa delimitadora e sem face e estimar a localização da face e a mesclagem de candidatos à supressão não máxima (NMS). (2) uma Rede Refine (R-Net), ao contrário da P-Net, ela não obtém as propostas da região, mas rejeita muitos candidatos falsos. (3) É semelhante ao R-Net, chamado O-Net. Nesta rede, ele produzirá as posições de cinco marcos faciais.

B. Treinamento

Assim como o MTCNN, alavancamos três tarefas: classificação de rosto, regressão de caixa delimitadora e classificação de emoções faciais. As funções de custo da classificação de face e regressão da caixa delimitadora são as mesmas do MTCNN. Usando a perda de entropia cruzada para resolver a classificação de face e a perda euclidiana para cada amostra.

$$e_{Eu}^{det} = -\left(y_{Eu}^{det}\,\text{registro}\left(p_{Eu}\right) + \left(1 - y_{Eu}^{det}\right)\left(1 - \log\left(p_{Eu}(1)\right)\right)$$
$$e_{Eu}^{box} = \left\|\hat{y}_{Eu}^{box} - y_{Eu}^{box}\right\| \qquad 2)$$

Exibir fonte

Equação (1) é função de custo de classificação cara, onde p_i é a probabilidade produzido pela rede que indica uma amostra sendo uma cara. A notação $y_{Eu}^{det} \in \{0, 1\}$ denota o rótulo de verdade da terra. A equação (2) é formulada como um problema de regressão, onde $\hat{y}_{Eu}^{box}$ y_{Eu}^{box} verdade. Existem quatro coordenadas, incluindo a parte superior esquerda, altura e largura. Semelhante à tarefa de detecção de rosto, a classificação das emoções faciais é formulada como um problema de sete classificações, também usamos a perda de entropia cruzada:

$$e_{Eu}^{\text{emoção}} = -\left(y_{Eu}^{emotion} \text{registro}\left(p_{Eu} \right) + \left(1 - y_{Eu}^{emotion} \right) \left(1 - \log\left(p_{Eu} \right) \right) \right) \quad (3)$$

Exibir fonte

Onde $y_{Eu}^{\text{emoção}} \in \{ 0 , 1 , 2 , 3 , 4 , 5 , 6 \}$ denota o rótulo de verdade da terra. p_i é a probabilidade produzida pela rede que indica que uma amostra é uma emoção das expressões faciais. rótulo. p_i é a probabilidade produzida pela rede que indica que uma amostra é uma emoção das expressões faciais. A equação (3) é o formulário de múltiplas fontes.

$$\min \sum_{i=1}^{N} \sum_{j \in \{ det , box , emotion \}} \alpha_{Eu} \beta_j e_{Eu}^j \quad (4)$$

Exibir fonte

onde N é o número de trama smaples. α_{Eu} denota a importância da tarefa.

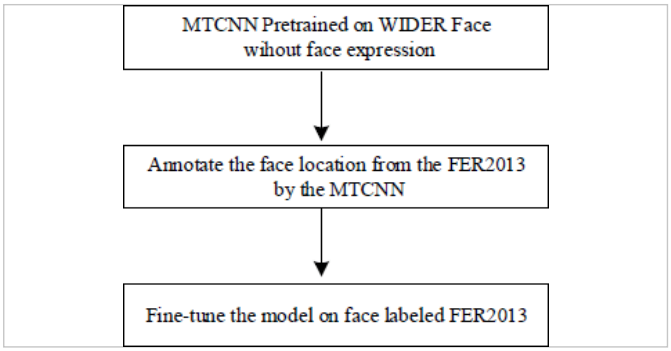

Figura 1.
Fluxograma do procedimento de treinamento

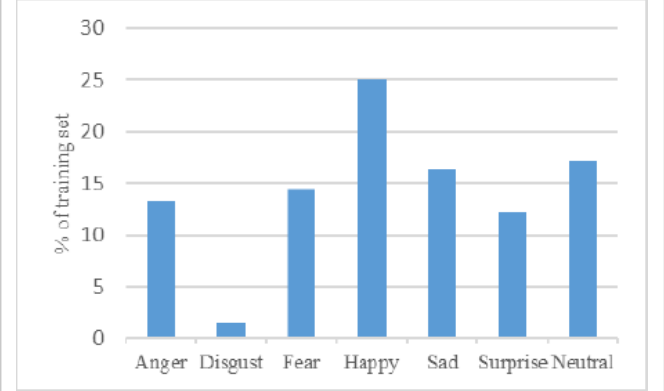

Figura 2.
Distribuição das emoções

O treinamento do MTCNN pode ser feito de uma maneira que utiliza a descida do gradiente estocástico (SGD) para os ramos de classificação e regressão. Nós treinamos a P-Net com os conjuntos de dados coletados em primeiro lugar e a R-Net com os resultados previstos da P-Net em segundo lugar. Por fim, para o treinamento da O-Net, os conjuntos de dados são coletados dos resultados previstos da P-Net e da R-Net.

C. Treinamento

Nosso método segue a estrutura semelhante de aprendizado profundo do MTCNN. É semelhante ao R-Net, chamado O-Net. Nesta rede, nosso objetivo é classificar a emoção das expressões faciais. De acordo com a proposta, modificamos a arquitetura do MTCNN para reconhecimento de expressão facial e treinamos nosso modelo de reconhecimento de expressão facial seguindo o procedimento proposto, como mostra a Figura 1 .

SEÇÃO IV

Experiências

Nesta seção, treinamos e testamos que as redes foram feitas usando a estrutura de código-fonte aberto Caffe para redes neurais profundas convolucionais e relatamos experimentos sobre comparações de reconhecimento de expressão facial e também sobre o desempenho dos principais detectores de face.

A. Conjunto de dados de treinamento

Os dados que usamos são compostos por 48×48 imagens em escala de cinza em pixel de rostos da competição Kaggle Desafios no aprendizado de representação: Desafio de reconhecimento de expressão facial [14] . É um grande conjunto de dados FER publicamente disponível que consiste em 35.887 culturas de face quase centralizadas e cada face ocupa aproximadamente a mesma quantidade de espaço em cada imagem. O conjunto de dados é dividido em conjuntos de treinamento, validação e teste com 28.709, 3.589 e 3.589 imagens, rótulos de sete categorias de expressões faciais: raiva, nojo, medo, feliz, triste, surpresa e neutro. A precisão humana neste conjunto de dados é de cerca de 65,5%. A distribuição de classe do conjunto de dados pode ser encontrada na Fig. 2 .

Para compensar o tamanho relativamente pequeno do conjunto de dados, usamos a popular técnica de aumento de dados que consiste em inverter imagens horizontalmente. Como as emoções intuitivamente não devem mudar com base no fato de as expressões faciais serem ou não espelhadas, parecia uma escolha sensata. Os dados de treinamento para a rede são descritos a seguir:

1. *Pré-treinado* : De acordo com a *Equação* (4) , usamos ($\alpha_{\text{det}} = 1$, $_{\text{caixa}} = 0{,}5$, $_{\text{emoção}} = 0$) no MTCNN para treinar a única tarefa de detecção de rosto.

2. *Anotação* : para tornar os conjuntos de dados FER2013 rotulados com a anotação de face, usamos o MTCNN pré-treinado para finalizá-lo.

3. *Ajuste fino* : semelhante ao estágio pré-treinado, usamos ($\alpha_{\text{det}} = 1$, $_{\text{caixa}} = 0{,}5$, $_{\text{emoção}} = 1$) para impulsionar o acompanhamento da tarefa de reconhecimento de expressão facial.

B. Parâmetros

Para escolher os parâmetros (força de regularização, taxa de aprendizado e decaimento), amostramos aleatoriamente no espaço de log e mantivemos aqueles que produziram a melhor precisão de validação. Os parâmetros que acabamos usando o seguinte:

- Tamanho do lote: 100.
- Taxa de aprendizagem: 0,0001.
- Decaimento da taxa de aprendizagem: 0,85.
- Momento: 0,9.
- Número de épocas: 40.

C. Mineração negativa negativa

A mineração dura negativa foi mostrada como uma estratégia eficaz para aumentar o desempenho do aprendizado profundo, especialmente para tarefas de detecção de objetos, incluindo detecção de faces. A ideia por trás desse método é que negativos negativos são as regiões em que a rede falhou ao fazer a previsão correta. Assim, os negativos negativos são alimentados novamente na rede como um reforço para melhorar nosso modelo treinado. O processo de treinamento resultante poderá melhorar

nosso modelo em direção a menos falsos positivos e melhor desempenho de classificação.

Em nossa abordagem, negativos negativos foram coletados do modelo pré-treinado da P-Net do nosso processo de treinamento. Consideramos uma região como negativa negativa se sua interseção sobre união (IOU) sobre a região de verdade do solo for menor que 0,5. Durante o processo de treinamento com negativo negativo, explicitamente adicionamos esses negativos negativos ao ROIs para ajustar o modelo e equilibramos a proporção de primeiro e segundo plano em cerca de 1: 3.

D. Resultados

Avaliamos os resultados com relação à validação e conjuntos de testes do FER2013 [14] . A precisão final de validação obtida é de 60,7%. Um gráfico de exemplo de nossa evolução da precisão pode ser encontrado na Fig. 3 . E o histórico de perdas é mostrado na Fig. 4 . Como pode ser visto, a precisão do treinamento aumenta enquanto a precisão do teste permanece quase constante após 15 épocas. Isso significa que estamos ajustando levemente nossos dados. A matriz de confusão no conjunto de validação é mostrada na Tabela 1 . Nojo é a classe em que nossa rede se sai pior e Feliz em que é mais bem-sucedida.

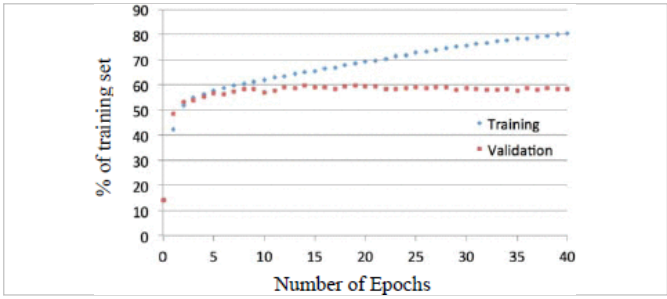

Figura 3.
Precisão de treinamento e validação em mais de 40 épocas para uma rede típica que treinamos.

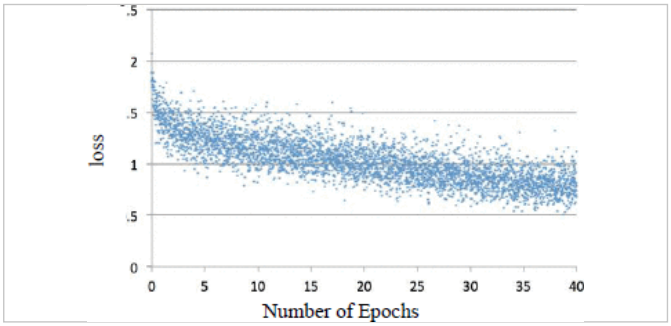

Figura 4.
Histórico de perdas

Tabela I. Matriz de confusão do conjunto de validação

	Anger	Disgust	Fear	Happy	Sad	Surprise	Neutral
Anger	60	0	9.	5	14.	2	9.
Disgust	40.	21	10	6	19	0	4
Fear	14	0	33	7	28	7	12
Happy	5.	0	3	81	4	2	6
Sad	14	0	8	7	52	3	17
Surprise	3	0	9	6	4	74	5
Neutral	11.	0	7	8	19	0	54

Algumas das dificuldades para melhorar isso é que as imagens são muito pequenas e, em alguns casos, é muito difícil distinguir qual emoção existe em cada imagem, mesmo para humanos. Para entender como a rede neural classificou diferentes imagens, usamos mapas de saliência, para detectar regiões importantes nas imagens de acordo com a rede neural. Embora a maioria dos resultados seja bastante barulhenta, algumas imagens mostraram resultados convincentes.

SEÇÃO V.
Conclusão

Neste trabalho, propusemos um novo método para detecção de rosto e reconhecimento de expressão facial usando técnicas de aprendizado profundo. Especificamente, nós os correlacionamos inerentemente. Realizamos um extenso conjunto de experimentos no conhecido FER2013 testado para o trabalho do FER. Embora nossa precisão de validação seja baixa, acreditamos que adicionar mais camadas e mais filtros melhoraria ainda mais a rede. Melhoraremos ainda mais a precisão e abordaremos a eficiência do método proposto para outros dispositivos.

Autores

Figuras

Referências

Palavras-chave

Métricas

IEEE Account

Profile Information

Purchase Details

Need Help?

Other

US & Canada: +1 800 678 4333
Worldwide: +1 732 981 0060

Conta IEEE
» Alterar nome de usuário / senha
» Atualizar endereço

Detalhes da compra
» Opções de pagamento
» Histórico de pedidos
» Exibir documentos comprados

Informação do Perfil
» Preferências de comunicação
» Profissão e Educação
» Interesses Técnicos

Preciso de ajuda?
» **EUA e Canadá:** +1 800 678 4333
» **Em todo o mundo:** +1 732 981 0060
» Contato e Suporte

Sobre o IEEE *Xplore* | Contate-Nos | Socorro | Acessibilidade | Termos de uso | Política de Não Discriminação | Mapa do site | Privacidade e exclusão de cookies