Regressão múltipla com redes neurais convolucionais para estimativa conjunta da intensidade de *Action Units*

Júlio César Batista, Olga R. P. Bellon, Luciano Silva
IMAGO-UFPR Research Group
Universidade Federal do Paraná (UFPR)
Av. Francisco H. dos Santos, 100, 81531-980, Curitiba, PR, Brazil
Email: {julio.batista,olga,luciano}@ufpr.br

***Abstract*—Este trabalho[1] apresenta uma rede neural convolucional (CNN - Convolutional Neural Network) para efetuar a estimativa conjunta da intensidade de *Action Units* (AUs) em imagens de faces. A estimativa da intensidade de AUs é essencial durante a análise de expressões faciais. Os métodos existentes não levam em consideração a possibilidade de estimativa conjunta para vários AUs em uma face e precisam de um modelo para cada AU; métodos que fazem uso dessa informação precisam reestruturar o problema como aprendizado estruturado com grafos, aumentando a complexidade. Portanto, este trabalho propõe um modelo de regressão múltipla para realizar essa estimativa conjunta e permitir a otimização de um modelo *end-to-end*. O modelo proposto foi avaliado na base BP4D (Binghamton-Pittsburgh 3D Dynamic Spontaneous Facial Expression Database), utilizada no *Facial Expression Recognition and Analysis Challenge* (FERA) 2015, que possui anotações da intensidade para cinco AUs em imagens com ambiente controlado. Os resultados obtidos, na média das cinco AUs, superam os *baselines* propostos e são similares ao estado-da-arte, superando-o em uma das AUs.**

I. Introdução

Pessoas utilizam expressões faciais diariamente para comunicação não verbal e para expressar sentimentos. Com a riqueza das informações que são transmitidas a partir de faces e expressões faciais é possível trabalhar em melhores aplicações para as áreas de computação afetiva, interação humano-computador e saúde. Essas expressões são, normalmente, categorizadas em um conjunto básico definido por: alegria, tristeza, nojo, medo, raiva, surpresa e também a face neutra. Entretanto, a análise de expressões utiliza o movimento de músculos da face, conhecidos como *Action Units* (AUs), para realizar essa categorização. Os AUs são definidos pelo *Facial Action Coding System* (FACS) [1] com base na anatomia da face. O FACS define um conjunto de 32 AUs, com os códigos e a região da face onde ocorrem listados na Tabela I, sendo que as expressões faciais são criadas a partir da combinação de determinados AUs. Por exemplo, a expressão de alegria é determinada pela ocorrência da AU12 (movimento do canto dos lábios) e, eventualmente, da AU06 (movimento da bochecha) e AU25 (separação dos lábios). A AU06 não precisa ocorrer para caracterizar um sorriso visto que muitas pessoas não demonstram esse AU ao sorrir [2].

Além da possibilidade da análise de expressões faciais a partir dos músculos da face, utilizar AUs para essa tarefa pode levar a um melhor entendimento de expressões faciais. Conforme [2], [4], as expressões faciais, e emoções, demonstradas diariamente não ocorrem em apenas seis categorias, mas podem ser compostas. Nesse sentido, determinar a expressão pela ocorrência de AUs é melhor do que elaborar todas as classes possíveis. Finalmente, [4] define a importância não só da ocorrência de AUs, mas também da intensidade com a qual estes são demonstrados nas expressões faciais.

O FACS também define uma escala ordinal de cinco níveis de intensidade para os AUs. Essa intensidade representa o nível de contração do músculo. A partir da análise dessa intensidade é possível descobrir se uma expressão é verdadeira ou não [3], se a pessoa está sentindo dor [5], ou até mesmo se possui doenças psicológicas [6]. Um exemplo da intensidade de AUs pode ser visto na Figura 1.

A partir da Figura 1 é possível perceber uma propriedade das AUs, a correlação entre AUs e intensidades. Nesse caso, quando uma AU ocorre na face, existe a chance de outra AU também ocorrer, como é o caso das AU06, AU12 e AU25 que formam um sorriso. Essa correlação ainda pode ser elevada ao nível das intensidades, conforme a Figura 1 demonstra, porque ao aumentar a intensidade da AU12, a AU25 surge naturalmente. A correlação entre as AUs também implica uma característica de aditividade nas AUs. Nesse caso, AUs na mesma região da face são considerados não aditivas, ou seja, uma AU interfere na aparência da outra. Quando os AUs são aditivos, a mudança em uma AU não interfere na outra, como

TABLE I
AUs definidas pelo FACS e a região da face onde eles ocorrem. Adaptado de: [3].

Região da face	AUs
Superior	1, 2, 4, 5, 6, 7, 43, 45, 46
Inferior	9, 10, 11, 12, 13, 14, 15, 17, 18, 20, 22, 23, 24, 25, 28, 27, 28

[1]Este artigo é baseado em uma Dissertação de Mestrado

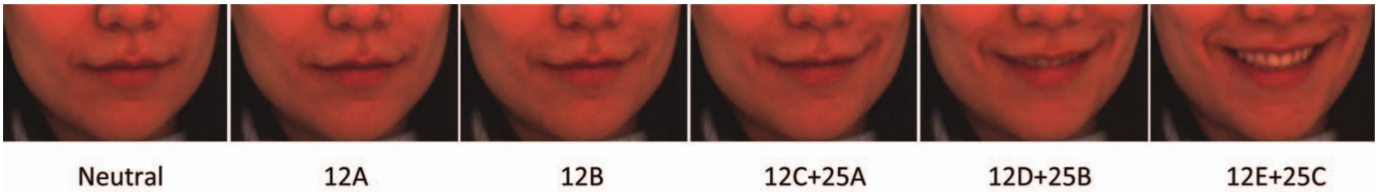


Fig. 1. Níveis de intensidade de AUs. O primeiro quadro demonstra a face neutra que muda desde um movimento leve até uma contração maior da AU12 (canto dos lábios). É possível notar também a correlação que existe com a AU25 (separação dos lábios). Fonte: [7].

no exemplo das AU12 e AU25 da Figura 1.

Considere-se que as intensidades das AUs constituem um conjunto $S = \{0, 1, 2, 3, 4, 5\}$, sendo os cinco níveis definidos pelo FACS e também a intensidade 0 indicando que a AU não está presente na imagem. Essa definição permite modelar a tarefa de estimar a intensidade de AUs como classificação multiclasse, regressão e ranqueamento. Visto que os níveis de intensidade têm uma característica ordinal, ou seja $0 < 1 < 2 < 3 < 4 < 5$, os modelos de ranqueamento e regressão são as abordagens mais sugeridas. Um ponto importante a ser analisado é que esses métodos, normalmente, requerem um modelo por AU e não fazem uso de uma representação conjunta. Com isso, não é possível levar em consideração as correlações entre os AUs e suas intensidades. Para superar essa limitação, modelos baseados em árvores latentes [8] ou em aprendizado estruturado [5], [9] podem ser utilizados.

Além dos modelos de aprendizado de máquina que são utilizados na estimativa da intensidade de AUs, também é necessário uma representação eficiente da imagem. Normalmente, a representação é dada pela textura e pela geometria da face. A parte geométrica é computada a partir de *landmarks* e envolve o cálculo de ângulos e distâncias entre esses pontos. A textura da face pode ser computada utilizando descritores como em [10], [11] que utilizaram *Histogram of Oriented Gradients* (HOG), ou em [8] que usou *Local Binary Patterns* (LBP) ou com filtros de Gabor usados em [2], [12]. Outros modelos envolvem o uso de redes neurais convolucionais, como em [13]–[16], que otimizam uma representação intermediária com base em *loss functions*.

Com base nos métodos existentes e nas propriedades das AUs, este artigo propõe:

- Uma rede neural convolucional otimizada com regressão múltipla para efetuar a estimativa conjunta da intensidade de AUs.
- A adaptação de uma *loss function* de regressão para levar em consideração as características nas intensidades de AUs.
- Uma abordagem para amostrar imagens durante a otimização para diminuir o desequilíbrio entre os níveis de intensidade das AUs.

O restante deste artigo está organizado conforme: a Seção II aborda o modelo proposto e a otimização do mesmo. Em seguida, a Seção IV descreve os protocolos utilizados nos experimentos e na avaliação dos resultados. Os resultados obtidos e as discussões estão incluídos na Seção V. Por fim, a Seção VI conclui este trabalho e sugere trabalhos futuros.

II. Modelo proposto

Esta seção descreve o modelo proposto para estimar a intensidade conjunta de AUs utilizando uma rede neural convolucional otimizada para regressão múltipla.

Dado que $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}, \mathbf{P}\}$ seja uma base de imagens com anotações de AUs utilizada para treinamento e avaliação. A matriz $\mathbf{X} \in \mathbb{R}^{M \times 224 \times 224 \times 3}$ contém M imagens RGB de faces normalizadas para $\mu = 0$ e $\sigma = 1$. $\mathbf{Y}$ é uma matriz de $M \times Q$ contendo a intensidade de Q AUs em uma imagem. Nesse caso, $\mathbf{Y}_{ik} \in \{0, 1, 2, 3, 4, 5\}$ representa a intensidade da k-ésima AU na i-ésima imagem. O vetor $\mathbf{P} \in \mathbb{R}^M$ contém as probabilidades para amostragem de uma determinada imagem durante a otimização. Portanto, $0 \leq \mathbf{P}_i \leq 1$ sendo que $\mathbf{P}_i$ é a probabilidade das anotações $\mathbf{Y}_i$ ocorrerem. O cálculo do vetor ocorre conforme a equação 1, que representa a probabilidade conjunta de um vetor de intensidades.

$$\mathbf{P}_i = \prod_{j=1}^{Q} P(i, j) \tag{1}$$

O cálculo da probabilidade $P(i, j)$ é demonstrado na equação 2. Essa probabilidade é calculada como o complemento da ocorrência da intensidade $\mathbf{Y}_{ij}$ no conjunto de treino. Assim, intensidades mais frequentes terão uma probabilidade menor do que as menos frequentes. Agora é possível usar um método de amostragem para selecionar imagens durante a otimização conforme essas probabilidades, reduzindo o desequilíbrio entre classes e gerando *batches* mais homogêneos.

$$P(i, j) = 1 - \frac{\sum_{k=1}^{M} [\mathbf{Y}_{kj} = \mathbf{Y}_{ij}]}{M} \tag{2}$$

III. Arquitetura e otimização da rede

O modelo proposto é uma adaptação da VGG16 [17]. Nesse modelo a última camada, com a probabilidade entre 1.000 classes da ImageNet [18], foi removida e foram adicionadas uma camada de *dropout* [19] e outra totalmente conectada com Q unidades. A saída do modelo é linear, sem nenhuma função de ativação para restringir as saídas no intervalo [0,5]. A saída contínua e sem restrição somente é utilizada na otimização do modelo. Para efetuar a inferência, a saída contínua é convertida em um dos níveis de intensidade utilizando a equação 3.

$$\hat{\mathbf{Z}}_{ij} = \max\{0, \min\{5, [\frac{1}{2}\hat{\mathbf{Y}}_{ij}]\}\} \quad (3)$$

A equação 3 demonstra a restrição das saídas do modelo, $\hat{\mathbf{Y}} \in \mathbb{R}^Q$, no intervalo [0,5]. Inicialmente, as estimativas do modelo são reduzidas pela metade, devido à *loss function* utilizada, demonstrada na equação 4. Após essa etapa, as estimativas são convertidas de valores contínuos para inteiros utilizando o arredondamento para o inteiro mais próximo, representado por $[\cdot]$. Finalmente, *max* restringe valores negativos para 0 e *min* limita o valor máximo como 5.

O modelo foi otimizado utilizando *Stochastic Gradient Descent*, a partir de um pré-treino na ImageNet [18], com a *loss function* apresentada na equação 4.

$$\mathcal{L}(\hat{\mathbf{Y}}, \mathbf{Y}) = \mathcal{H}(\hat{\mathbf{Y}} - 2 \cdot \mathbf{Y}) \quad (4)$$

Na *loss function* apresentada na equação 4, $\mathcal{H}(\cdot)$ representa a *Smooth L1 loss* descrita em [20]. A entrada dessa função é o erro calculado entre a saída do modelo ($\hat{\mathbf{Y}}$) e as anotações ($\mathbf{Y}$). As anotações são multiplicadas por 2 devido a utilização do arredondamento na inferência. Por exemplo, dado uma anotação y e uma saída do modelo $\hat{y}$, a inferência $\hat{y}$ estará correta se $y - 0.5 \leq \hat{y} < y + 0.5$, ou seja $[\hat{y}] = y$. Entretando, $\mathcal{H}$ assume uma margem $\delta = 1$, dessa forma $y \pm 1$ está dentro da margem de inferência correta, mas como visto no exemplo anterior essa propriedade não é válida. Assim, ao multiplicar as anotações por 2 amplia-se a margem para 1 como esperado em $\mathcal{H}$. Finalmente, durante a inferência é necessário multiplicar as saídas por $\frac{1}{2}$ para voltar para a escala de intensidades esperada conforme a equação 3.

Para a otimização do modelo são utilizadas 41.328 imagens, quantidade total de imagens da base removendo detecções erradas de face e anotações desconhecidas conforme descrito na Seção IV. Porém, essas imagens são amostradas utilizando uma distribuição multinomial utilizando as probabilidades em $\mathbf{P}$. Assim, é possível que imagens com baixa probabilidade nunca sejam utilizadas e imagens com maior probabilidade sejam utilizadas mais de uma vez. Como a probabilidade em $\mathbf{P}$ é inversa à frequência com que as anotações ocorrem na base, anotações muito frequentes na base serão pouco usadas enquanto que anotações pouco frequentes serão utilizadas mais de uma vez. Essa abordagem é utilizada para diminuir o desequilíbrio que existe entre os níveis de intensidade nas anotações, como mostra a Figura 2, fazendo a otimização mais homogênea e evitando viéses.

Por fim, antes da última camada do modelo proposto existe uma camada de *dropout* com probabilidade de 0,6 para desativar uma determinada unidade. Essa camada é utilizada para diminuir o tamanho da representação utilizado na regressão das intensidades. Ao analisar o vetor de representações com 4.096 elementos gerados pela VGG16, utilizando *Principal Component Analysis* (PCA), notou-se que era possível manter uma variância de 95% com aproximadamente 2.300 elementos. Portanto, se apenas 40% dos 4.096 elementos da representação

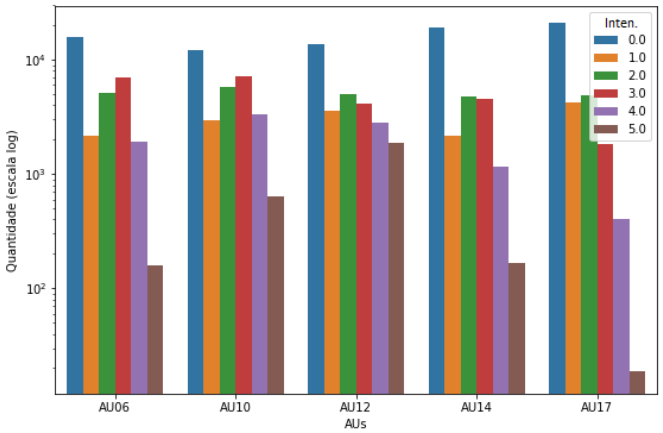


Fig. 2. Quantidade de imagens para cada nível de intensidade de cada AU no conjunto de avaliação. É possível perceber que a intensidade 0 é predominante para todas as AUs e também que níveis mais altos de intensidade são menos frequentes.

forem utilizados, a representação diminui para 1.639 elementos. Assim, é possível diminuir o tamanho da representação sem a adição de mais parâmetros com uma nova camada.

IV. Experimentos

Essa seção descreve o modelo *baseline*, a base de imagens e o protocolo utilizados na avaliação do modelo proposto para deteção conjunta de AUs . Além do protocolo padrão da base, todos os modelos foram otimizados e testados três vezes de forma independente. Assim, os resultados apresentados são computados como a média das cinco melhores épocas de cada teste. Essa metodologia visa diminuir o efeito da aleatoriedade durante a otimização e gerar resultados mais consistentes.

O modelo *baseline* consiste na CNN VGG16 [17] sem nenhum tipo de pré-treino. As modificações consistem na alteração da última camada para que a saída seja de Q unidades sem restrição para gerar as estimativas de intensidade. O outro ponto alterado foi a função utilizada na otimização, que é a mesma descrita na equação 4 sem a multiplicação por 2. Para inferência, foi utilizada a mesma abordagem apresentada na equação 3 sem o ajuste de $\frac{1}{2}$. Os demais híperparâmetros consistem em: *learning rate* (α) inicial de 0,01, *momentum* de 0,9 e *weight decay* de 0,0005. Esse modelo foi otimizado em cem épocas com α sendo atualizado para $\alpha := \frac{\alpha}{10}$ a cada vinte épocas. A única diferença dos híperparâmetros do *baseline* para o modelo proposto é o α inicial que é 0,001. Os demais valores foram mantidos iguais.

Os experimentos foram conduzidos com o subconjunto da BP4D [21], [22] utilizado no *Facial Expression Recognition and Analysis Challenge* (FERA) 2015 [23]. Essa base consiste em imagens de 21 sujeitos para treino e de 20 sujeitos para avaliação. No total, existem anotações de intensidade variando entre 0 e 5 para cinco AUs (demonstradas na Figura 3): AU06, AU10, AU12, AU14 e AU17. Entretanto, nem todas as imagens estão anotadas com todas as AUs. Em algumas imagens as AUs são anotadas com um valor 9 indicando

Fig. 3. Exemplo das AUs disponíveis para estimativa de intensidade na base BP4d. Adaptado de [3].

que a intensidade é desconhecida. As imagens que possuem alguma AU com essa anotação foram removidas do conjunto de treino visto que a abordagem proposta utiliza regressão múltipla e requer todos os AUs anotados para uma imagem. Ao remover as imagens com anotações desconhecidas, o conjunto de treino resultante ficou com 41.328 imagens. No conjunto de avaliação existe o mesmo caso com anotações desconhecidas, porém, nesse caso, as anotações desconhecidas foram desconsiderados ao calcular o resultado dos modelos. É importante notar que a base do FERA 2017 [24] não foi utilizada porque o foco dessa base é a variação nas poses da cabeça. Portanto, a base do FERA 2015 foi utilizada devido ao foco da abordagem proposta visar a estimativa de intensidade de AUs e o desequilíbrio entre as classes. Finalmente, a métrica de avaliação utilizada consiste no ICC(3, 1) [25] que é a medida utilizada no FERA 2015 para avaliar os modelos na estimativa de intensidade de AUs.

Os modelos utilizam a imagem de uma face como entrada. Para isso, as faces foram previamente detectadas utilizando uma versão modificada da Faster R-CNN [26] conforme [16]. Entretanto, os *bounding boxes* não foram transformados em quadrados, mas as imagens foram redimensionadas para que o menor lado tenha o tamanho de 224 *pixels*, demonstrado na Figura 4. Com isso, durante a otimização é feito um recorte aleatório de 224×224, como uma forma de *augmentation*; durante a avaliação é feito um corte central de 224×224.

Fig. 4. Exemplo do recorte de 224 *pixels* do menor lado do *bounding box* detectado.

V. Resultados

Esta seção descreve os resultados obtidos com o modelo proposto e a comparação com o *baseline* e com métodos estado-da-arte. A Tabela II demonstra a comparação do ICC(3, 1) obtido entre os modelos.

A partir da Tabela II é possível perceber que o modelo proposto superou o *baseline* em três de cinco AUs e também no resultado médio. Além disso, o modelo proposto superou o estado-da-arte para a AU10 e obteve um resultado similar para a AU17. Em relação ao *baseline*, o resultado para a AU06 é superior ao do modelo proposto, e também ao estado-da-arte, e a AU14 obteve um resultado similar. Para comparativo entre *baselines*, foi incluído o resultado da VGG16 demonstrado em [9]. Como é possível notar, o *baseline* proposto, utilizando regressão, superou o *baseline* em [9]. Esse comparativo indica a melhora resultante no uso de regressão ao invés de classificação na estimativa de intensidade. Devido à natureza ordinal da intensidade das AUs, faz sentido que a regressão obtenha um resultado melhor que a classificação visto que as saídas são contínuas e estabelecem essa ordem.

O resultado médio reportado na Tabela II é em relação à melhor iteração em três experimentos independentes que foram executados. Ao considerar as cinco melhores iterações de cada experimento, o *baseline* tem um resultado médio de 0,588 e o modelo proposto tem um resultado médio de 0,603. Portanto, o modelo proposto é mais robusto na variação entre treinamentos gerando um resultado mais homogêneo entre várias iterações e sem sofrer muito com o efeito da aleatoriedade.

A Figura 5 apresenta a quantidade de imagens estimadas para cada intensidade de cada AU utilizando o modelo proposto. A partir dela é possível perceber que o modelo proposto obteve uma distribuição similar à do conjunto de avaliação demonstrado na Figura 2. Entretanto, o modelo falhou ao estimar os níveis de intensidade 5 para as AU10, AU14 e AU17. Ao observar a matriz de confusão para a AU12, Figura 6, é possível verificar que, mesmo com erros, as estimativas são próximas dos níveis esperados. Um resultado qualitativo que fica evidente ao observar a matriz de confusão é que as estimativas estão próximas dos níveis esperados. Por exemplo, ao observar a intensidade 1, é possível perceber que os erros foram para os níveis 0 e 2. Conforme pode ser visto na Figura 1 a diferença entre esses níveis de intensidade é muito sutil e, portanto, requer uma representação robusta para efetuar a inferência correta. De forma geral, o resultado está próximo

TABLE II
COMPARATIVO PARA ESTIMATIVA DE INTENSIDADE ENTRE AUS UTILIZANDO ICC(3, 1).

Modelo	*Action units*					
	06	10	12	14	17	Média
Baseline VGG16 [17]	**0,770**	0,679	0,783	0,385	0.433	0,610
Proposto	0,734	**0,704**	0,824	0,385	**0,449**	0,619
Aprendizado estruturado [9]	0,750	0,690	**0,860**	**0,400**	**0,450**	**0,630**
VGG16 em [9]	0,630	0,610	0,730	0,250	0,310	0,510

do esperado com alguns casos excedem esse padrão, como o de algumas imagens que deveriam ter a AU12 com intensidade 0, mas foram estimadas com intensidade 5. A divergência nos níveis de intensidade pode estar relacionada às correlações entre as AUs que o modelo aprendeu durante a otimização que são apresentadas na Figura 7.

Um dos objetivos de realizar a regressão múltipla para a intensidade das AUs é explorar a correlação existente em uma representação intermediária do modelo. A correlação existente entre as AUs no conjunto de avaliação pode ser vista na Figura 7a e nas estimativas pode ser vista na Figura 7b. Como é possível perceber, o modelo conseguiu capturar a tendência entre os AUs. Por exemplo, existe uma correlação mais forte entre as AU06, AU10, AU12 e AU14 enquanto que a AU17 não tem correlação com as demais. Entretanto, apesar das estimativas seguirem os mesmos padrões, elas têm correlação maior que as do conjunto de treino. Portanto, essas correlações fortes entre as AUs que foram geradas pelo modelo podem influenciar os casos da intensidade esperada ser 0, mas o modelo estimar 5. O mesmo vale para a falha do modelo em estimar a intensidade 5 para as AU10, AU14 e AU17.

VI. CONCLUSÃO

Este trabalho demonstrou uma rede neural convolucional com regressão múltipla para estimar a intensidade de AUs. Além do modelo, também foi apresentada uma técnica para amostragem de imagens para treino considerando o contexto de múltiplas anotações por imagem. O modelo proposto foi

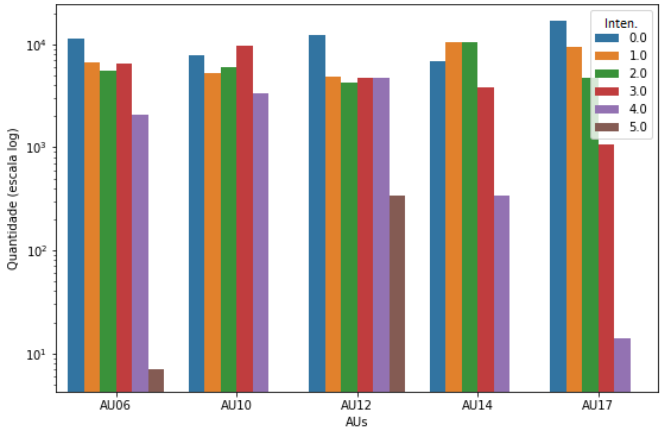


Fig. 5. Quantidade de imagens para cada nível de intensidade de cada AU estimados pelo modelo proposto.

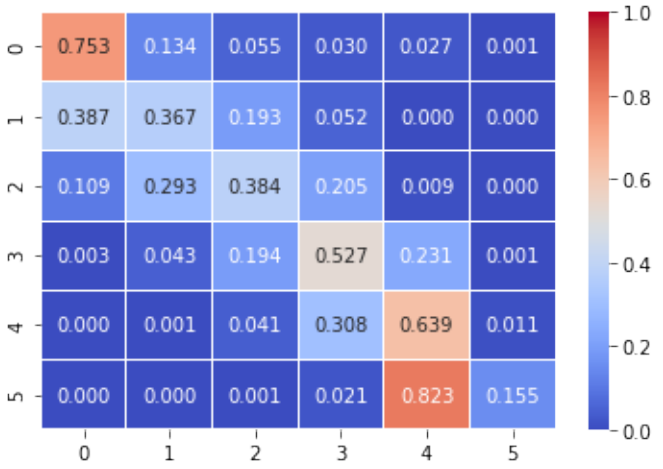


Fig. 6. Matriz de confusão com as estimativas de intensidade do modelo proposto para a AU12.

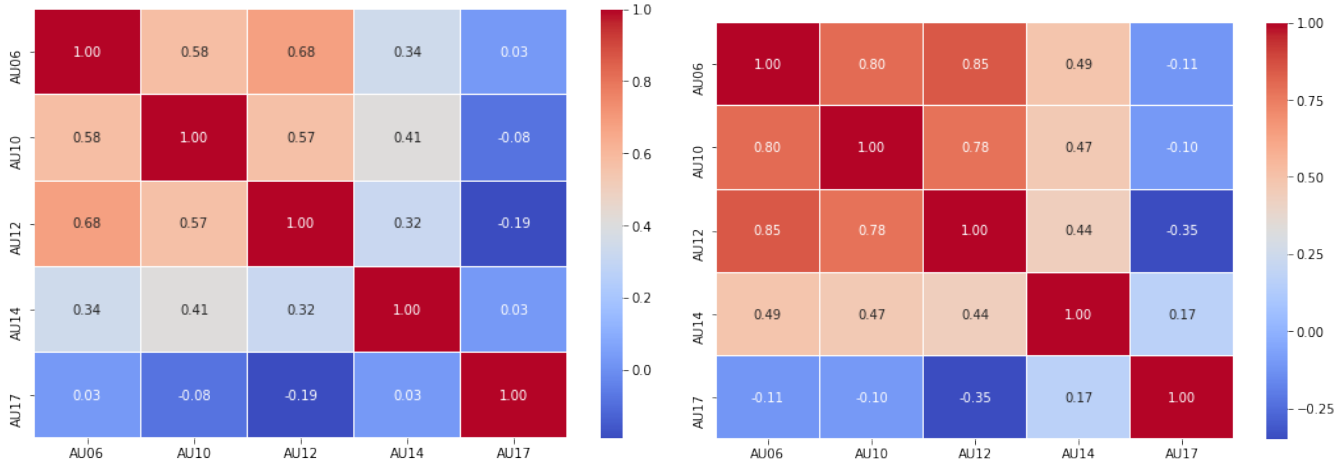
capaz de aprender correlações entre os níveis de intensidade a partir de uma representação intermediária. Entretanto, algumas correlações levaram o modelo a gerar estimativas fora do esperado. Mesmo assim, o modelo foi capaz de gerar resultados médios compatíveis ao estado-da-arte. Ao analisar os resultados para cada AU, o modelo conseguiu melhorar o estado da arte para a AU10 e obteve um resultado similar para a AU17. O uso de regressão para estimar a intensidade de AUs, que tem natureza ordinal, demonstrou-se muito bom visto que ele melhorou o *baseline* utilizado. Finalmente, como sugestões de trabalhos futuros ficam: a necessidade de estabilizar as correlações que o modelo aprendeu; melhorar a estratégia de amostragem de imagens durante a otimização; e utilizar um modelo com menos parâmetros visto que quase 50% da representação pode ser eliminada com *dropout*.

AGRADECIMENTOS

Os autores gostariam de agradecer à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio a esta pesquisa.

VII. PUBLICAÇÕES

- **Batista, J. C.**, Bellon, O. R. P., & Silva, L. (2016). Landmark-free smile intensity estimation. In 29th Conference on Graphics, Patterns and Images (SIBGRAPI), Workshop on Face Processing.



(a) Correlação entre as AUs no conjunto de avaliação.

(b) Correlação entre as AUs com o resultado do modelo proposto.

Fig. 7. Coeficiente de correlação de Pearson calculado entre as AUs.

- **Batista, J. C.**, Bellon, Albiero, V., O. R. P., & Silva, L. (2017, June). Aumpnet: Simultaneous action units detection and intensity estimation on multipose facial images using a single convolutional neural network. In IEEE International Conference on Automatic Face and Gesture Recognition (FG), pp. 866-871, Facial Expression Recognition and Analysis challenge (FERA 2017). IEEE.
- de Bittencourt Zavan, F. H., Gasparin, N., **Batista, J. C.**, e Silva, L. P., Albiero, V., Bellon, O. R. P., & Silva, L. (2017). Face Analysis in the Wild. In 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T), pp. 9-16. IEEE.

REFERENCES

[1] P. Ekman, W. Friesen, and J. Hager, *Facial Action Coding System (FACS): Manual*. A Human Face, 2002.

[2] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, 2014.

[3] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Transaction on Affective Computing*.

[4] S. Du and A. M. Martinez, "Compound facial expressions of emotion: from basic research to clinical applications," *Dialogues in Clinical Neuroscience*, vol. 17, 2015.

[5] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Copula ordinal regression for joint estimation of facial action unit intensity," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[6] J. M. Girard, J. F. Cohn, and F. D. la Torre", "Estimating smile intensity: A better way," *Pattern Recognition Letters*, 2015.

[7] S. M. Mavadati, S. Member, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, 2013.

[8] S. Kaltwang, S. Todorovic, and M. Pantic, "Latent trees for estimating intensity of facial action units," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[9] R. Walecki, O. Rudovic, V. Pavlovic, B. Schuller, and M. Pantic, "Deep structured learning for facial action unit intensity estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[10] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *IEEE Conference on Automatic Face and Gesture Recognition*, 2015.

[11] J. Nicolle, K. Bailly, and M. Chetouani, "Real-time facial action unit intensity prediction with regularized metric learning," *Image and Vision Computing*, 2016.

[12] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[13] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[14] Z. Tősér, L. A. Jeni, A. Lőrincz, and J. F. Cohn, "Deep learning for facial action unit detection under large head poses," in *Springer European Conference on Computer Vision*, 2016.

[15] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis, "Deep learning based facs action unit occurrence and intensity estimation," in *IEEE Conference on Automatic Face and Gesture Recognition*, 2015.

[16] J. C. Batista, V. Albiero, O. R. P. Bellon, and L. Silva, "Aumpnet: Simultaneous action units detection and intensity estimation on multipose facial images using a single convolutional neural network," in *IEEE Conference on Automatic Face and Gesture Recognition*, May 2017.

[17] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *International Conference on Learning Representations*, 2015.

[18] M. Simon, E. Rodner, and J. Denzler, "Imagenet pre-trained models with batch normalization," *arXiv preprint arXiv:1612.01452*, 2016.

[19] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, 2014.

[20] R. Girshick, "Fast R-CNN," in *IEEE International Conference on Computer Vision*, 2015.

[21] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, 2014.

[22] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin, "Multimodal spontaneous emotion corpus for human behavior analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[23] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn, "Fera 2015 - second facial expression recognition and analysis challenge," in *IEEE Conference on Automatic Face and Gesture Recognition*, 2015.

[24] M. Valstar, E. S. Lozano, J. F. Cohn, L. A. Jeni, J. M. Girard, L. Yin, Z. Zhang, and M. Pantic, "Fera 2017 - addressing head pose in the third facial expression recognition and analysis challenge," *arXiv preprint arXiv:1702.04174*, 2017.

[25] P. Shrout and J. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin*, 1979.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Conference on Neural Information Processing Systems*, 2015.