Reconhecimento de Expressões Faciais Aplicada à Análise de Vídeos com Reações Espontâneas usando SVM com resposta Probabilística

Wdnei R. Paixão, Flávio G. Pereira, Karin S. Komati
Programa de Pós-Graduação em Engenharia de Controle e Automação (ProPECAut)
Instituto Federal do Espirito Santo - Campus Serra
Rodovia ES-010 - Km 6,5 - Manguinhos, Serra, ES, Brasil
Email: wdneipaixao@gmail.com, flavio.garcia@ifes.edu.br, kkomati@ifes.edu.br

***Abstract*—This work defines an initial proposal of a system for automated facial emotion classification applied to video that contains recordings of spontaneous reactions of spectators. The proposed approach uses a variation of the classifier Support Vector Machines (SVM) with outputs in posteriori probability values and the Histogram of Oriented Gradients (HoG) as a feature descriptor. For the training, the Radboud Face Database (RaFD) was used. The results presented show the viability of the use in the mass media to assess the mood of the audience in quantitative terms of probability with respect to time.**

***Resumo*—Este trabalho define uma proposta inicial de um sistema para o reconhecimento automatizado de emoções faciais em vídeo que contém gravações das reações de espectadores de forma espontânea. A abordagem proposta utiliza uma variação do classificador Maquinas de Vetores de Suporte (SVM) com saídas em valores de probabilidade a posteriori e o método Histograma de Gradientes Orientados (HoG) como extrator de características. Para o treinamento, utilizou-se a base de imagens Radboud Face Database (RaFD). Os resultados apresentados mostram a viabilidade da utilização em meios de comunicação para avaliação de humor da audiência em termos quantitativos de probabilidade em relação ao tempo.**

I. Introdução

Detecção de emoção através de expressões faciais é um problema cada vez mais estudado, pois reconhecer as reações do espectador em tempo real é uma ferramenta para avaliar se o vídeo alcançou os objetivos emocionais planejados pela indústria de marketing [1]. A exemplo disso, no período dos debates das eleições dos EUA, em 2015 e 2016, durante o debate do atual presidente Donald Trump, utilizou-se um sistema automatizado para verificar as reações dos espectadores [2].

Há vários trabalhos para reconhecer emoções em faces, destacando-se o trabalho de Santos-Paiva et al. [3] que apresentou um bom desempenho (94% de exatidão) utilizando o descritor Padrões Binários Locais (LBP, do inglês *Local Binary Pattern*), somente nas regiões dos olhos e boca, e o classificador Máquina de Vetores de Suporte (SVM, do inglês *Support Vector Machines*). No trabalho de Batista et al. [4] é realizada uma comparação do comportamento do classificador SVM em sua adaptação multi-classe, aplicado à intensidade de sorriso, mostrando que há diferenças de exatidão das abordagens Um-Contra-Todos e Um-Contra-Um.

Há também vários trabalhos que reconhecem emoções em vídeos, um deles é a proposta de Michel e El Kaliouby [5], que utiliza uma técnica de rastreamento dos pontos de interesse, na qual o ponto inicial e o quadro de maior movimento representam a informação da expressão. O trabalho de Hsu et al. [6] realiza a extração de características com filtros de Gabor e fusão de classificadores. O trabalho de Da Silva et al. [7] apresenta uma revisão de vários trabalhos, e afirma que a combinação mais consistente é a associação de HoG (*Histogram of Oriented Gradients*) e SVM.

Em todos os trabalhos citados, o resultado é a indicação de uma única emoção. Diferente dos trabalhos mencionados, este trabalho propõe oferecer como resposta uma probabilidade (porcentagem) de cada emoção contida na expressão. Assim, no presente trabalho apresenta-se um sistema de classificação automatizada de emoções faciais em vídeo, aplicadas a reação de espectadores, empregando o extrator de características HoG e o classificador SVM com saídas em valores de probabilidade a *posteriori*.

II. Abordagem Proposta

A arquitetura da abordagem proposta para uma única imagem é apresentada na Figura 1. Para o processamento de vídeo, o fluxo de processamento será repetido para cada quadro. O único processamento feito na etapa da *Imagem de Entrada* foi a conversão da imagem em escala de cinza. A próxima etapa (Encontrar Faces com algoritmo Viola-Jones) é a detecção das faces na imagem em escala de cinza usando o algoritmo Viola-Jones [8], que já indica a localização dos olhos. Esta informação é aproveitada para a extração das regiões de interesse: regiões dos olhos e região da boca, assim como aplicado no trabalho de Santos-Paiva et al. [3], pois são as regiões que mais apresentam características emocionais. No entanto, o restante da face é desprezado por ocasionar ruído na classificação, tais como cabelo e orelhas.

Figura 1. Fluxo de processamento proposto para identificação automática de emoções faciais

A extração de características é feita utilizando-se HoG [9]. Sua funcionalidade é baseada no princípio de que um objeto pode ser descrito a partir da distribuição da intensidade de gradientes ou vetores direcionais. O gradiente é a medida da variação de uma grandeza, numa determinada direção, tendo em conta sua dimensão espacial. Assim, para atingir precisamente o formato do objeto, a imagem de entrada é dividida em pequenas regiões chamadas de células (*cells*). Para cada célula, um histograma com vetores direcionais é gerado, e o resultado final consiste no conjunto de todos os histogramas. No entanto, outras parametrizações podem ser realizadas, por exemplo o tamanho das células e blocos (conjunto de células a serem normalizadas), assim como, o número de *bins* dos gradientes.

Neste trabalho, é realizada a divisão de cada região de interesse em 9 blocos verticais e horizontais com células com tamanho $N/9$, sendo MxN a dimensão da imagem. Sendo cada bloco de HoG composto por histogramas direcionais com 9 *bins*. Assim, temos que cada região de interesse é representada por 9x9x9 = 729 atributos. Portanto, cada face é representada por 2x729 = 1.458 atributos.

Por fim, tem-se o classificador SVM com *kernel* polinomial quadrático utilizando abordagem multi-classe *Um-Contra-Todos* [10], além de sua adaptação para a geração de probabilidades a *posteriori* [11]. Assim, na última etapa temos a saída do classificador com as classes/emoções e suas respectivas probabilidades.

O classificador SVM separa as classes linearmente de tal modo que seja possível definir uma superfície de decisão ou hiperplano [12]. Em outras palavras, uma região é demarcada usando-se alguns valores como suporte, em que cada classe torna-se uma região [13]. Porém, SVM produz valores que não são probabilísticos. Para que fosse possível transformar a saída do SVM original em probabilidade, utilizou-se o trabalho de Platt et al. [11] que realiza uma alteração em seu *kernel*, no qual, ao invés de estimar a densidade condicional de uma classe $p(entrada|classe)$, foi utilizada uma modelagem paramétrica para uma estimação a *posteriori* $P(classe = 1|entrada)$ direta, empregando uma aproximação por sigmóide. Deste modo, a combinação SVM com sigmóide preserva a qualidade dos valores do SVM, enquanto produz probabilidades.

III. Experimentos, Resultados e Discussão

Os experimentos foram feitos de duas formas: uma utilizando uma base de fotos com expressões anotadas e outra feita com vídeo.

O primeiro experimento visa avaliar a performance do sistema proposto para imagens estáticas. Para tanto, foi utilizada a base de imagens Radboud Face Database (RaFD) [14]. Esta base apresenta um conjunto de fotos de 67 modelos, incluindo homens, mulheres e crianças caucasianos, de ambos os sexos, e homens holandeses de etnia marroquina, demonstrando emoções faciais em vários ângulos. Foram usadas apenas as fotos frontais desta base de 7 emoções, logo foram utilizadas 469 fotos (67 x 7).

Foi utilizada a técnica de *Hold-Out* sendo 70% base como treinamento e 30% como testes (140 fotos), para validar a base. Foram realizas 100 repetições para obter a média. A Figura 2 mostra a matriz confusão do sistema aplicada à base RaFD, alcançando exatidão média de 85,7%. Como o resultado do sistema é probabilístico, considerou-se como a emoção reconhecida aquela que apresentou a maior probabilidade.

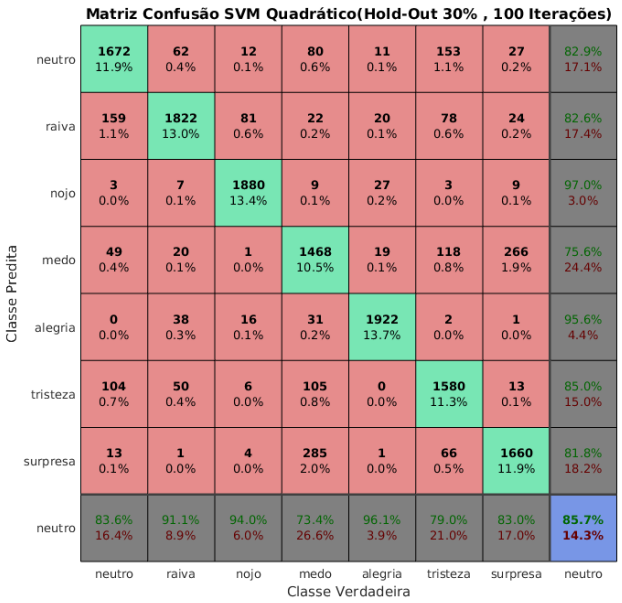

Figura 2. Matriz confusão do Sistema (Hold-Out 30%; 100 repetições)

No segundo experimento, o treinamento do sistema foi feito utilizando todas as 469 imagens da base RaFD. Para analisar as emoções e reações de espectadores foram utilizados vídeos do YouTube, nos quais há a gravação

das reações de pessoas ao assistir *trailers* de filmes. Infelizmente, não foram encontradas base de imagens que atendessem aos requisitos do trabalho. Todos os vídeos utilizados continham a licença padrão do YouTube, a qual permite o direito do uso e reprodução pública.

A Figura 3 mostra um quadro do vídeo utilizado como exemplo do protótipo criado neste trabalho, a face de cada ator foi reconhecida e cada emoção também foi discriminada abaixo contendo a probabilidade a *posteriori* de cada emoção encontrada pelo classificador. Nesta figura, a ordem das faces foi numerada de acordo com o resultado do algoritmo Viola-Jones, assim a mulher mais à esquerda com retângulo azul escuro é a "face 1", a segunda mulher com o retângulo verde é a "face 2", o rapaz é a "face 3" e a mulher mais a direita com retângulo azul claro é a "face 4". O quadro original não contém os retângulos coloridos sobre os rostos, nem as faixas coloridas na parte inferior.

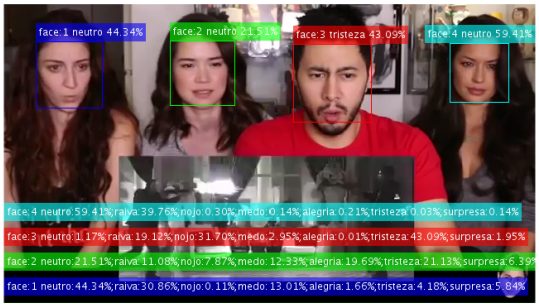

Figura 3. Análise de 4 pessoas assistindo um *trailer* de um filme indiano. Fonte: retirado de [15]

Na Figura 3, mostra-se a emoção reconhecida apenas naquele quadro. Já os gráficos das Figuras 4 e 5 mostram o resultado da análise das emoções decorridas no tempo, nos vários quadros, para a "face 1" e "face 4", respectivamente. As emoções são mostradas na seguinte ordem: *neutro* em azul escuro, *raiva* em verde, *nojo* em vermelho, *medo* em azul claro, *alegria* em lilás, *tristeza* em amarelo e *surpresa* em laranja. Vale ressaltar que a análise foi realizada somente durante a visualização do *trailer*, para que não fosse avaliado o momento das falas. Algumas amostras das faces são apresentadas ao lado do gráfico para exemplificar a expressão das atrizes. Os gráficos das faces 2 e 3 foram omitidas por questão de espaço do artigo.

Nesta análise, percebe-se que há variações no tempo, pois uma pessoa muda a expressão facial durante o tempo. No gráfico da "face 1" (Figura 4), apresentou uma boa variação entre as várias emoções: varia entre raiva e alegria entre os 40 e 60 segundos; após os 105 segundos apresenta raiva variando com tristeza. Mostra-se bem alegre entre os 80 e 103 segundos, variando no meio com surpresa. Esta atriz é bem expressiva, com movimentos de sobrancelha, boca e olhos bem visíveis.

Na análise dos vídeos, foi verificada, que muitas reações dos atores condiziam com o que estava sendo apresentado, principalmente as emoções *alegria* e *surpresa*, no entanto, foi percebido erros no reconhecimento de emoções. Principalmente no gráfico da "face 4" (Figura 5) que apresenta muita variação entre raiva e neutro, e poucas amostras de outras emoções. Conjectura-se que a pouca variação de emoções pode ser explicada pela curvatura bem arqueada das sobrancelhas, expressão caracterizada como raiva. A amostra da face de raiva da atriz pode não ser considerada como raiva e sim, como um sorriso sedutor, assim como a amostra da face de surpresa que está mais para alegria. Além disso, esta atriz por muitas vezes coloca as mãos de forma a cobrir parte do rosto, vira o rosto de lado e mexe no cabelo.

IV. Considerações Finais

No presente trabalho foi apresentado a proposta de um sistema automatizado para avaliar as emoções em faces em vídeo, associados à reações espontâneas, respondendo com valores probabilísticos. O sistema atual, que ainda se encontra em um estágio inicial de desenvolvimento, apresentou resultados satisfatórios na classificação de diferentes emoções.

Classificar as emoções em valores de probabilidade facilita a análise e oferece a vantagem de avaliar em quais momentos aconteceram quais emoções em cada face. Assim, a emoção expressa pode não estar na classe predita com maior probabilidade, mas em uma segunda ou terceira instância de probabilidade, fatores que podem ser usados por um especialista.

Como trabalho futuro, pretende-se melhorar o treinamento e torná-lo dinâmico, bem como avaliar outros métodos de extração de caraterísticas e classificadores.

Referências

[1] H. Moon, R. Sharma, and N. Jung, "Method and system for measuring emotional and attentional response to dynamic digital media content," Mar. 19 2013, uS Patent 8,401,248. [Online]. Available: https://www.google.com/patents/US8401248

[2] Emotient, "Emotion-reading technology first and only to analyze audience reactions to republican presidential," http://prn.to/1W5letd, Aug 2015, (Accessed on 07/14/2017).

[3] F. A. Santos Paiva, P. D. P. Costa, and J. M. De Martino, "Supervised methods for classifying facial emotions," in *Electronic Proceedings of the 29th Conference on Graphics, Patterns and Images (SIBGRAPI'16)*, october 2016.

[4] J. C. Batista, O. R. Bellon, and L. Silva, "Landmark-free smile intensity estimation," in *Electronic Proceedings of the 29th Conference on Graphics, Patterns and Images (SIBGRAPI'16)*, october 2016.

[5] P. Michel and R. El Kaliouby, "Real time facial expression recognition in video using support vector machines," in *Proceedings of the 5th international conference on Multimodal interfaces*. ACM, 2003, pp. 258–264.

[6] S.-C. Hsu, H.-H. Huang, and C.-L. Huang, "Facial expression recognition for human-robot interaction," in *Robotic Computing (IRC), IEEE Int. Conference on*. IEEE, 2017, pp. 1–7.

[7] F. A. M. Da Silva and H. Pedrini, "Effects of cultural characteristics on building an emotion classifier through facial expression analysis," *Journal of Electronic Imaging*, vol. 24, no. 2, pp. 023 015–023 015, 2015.

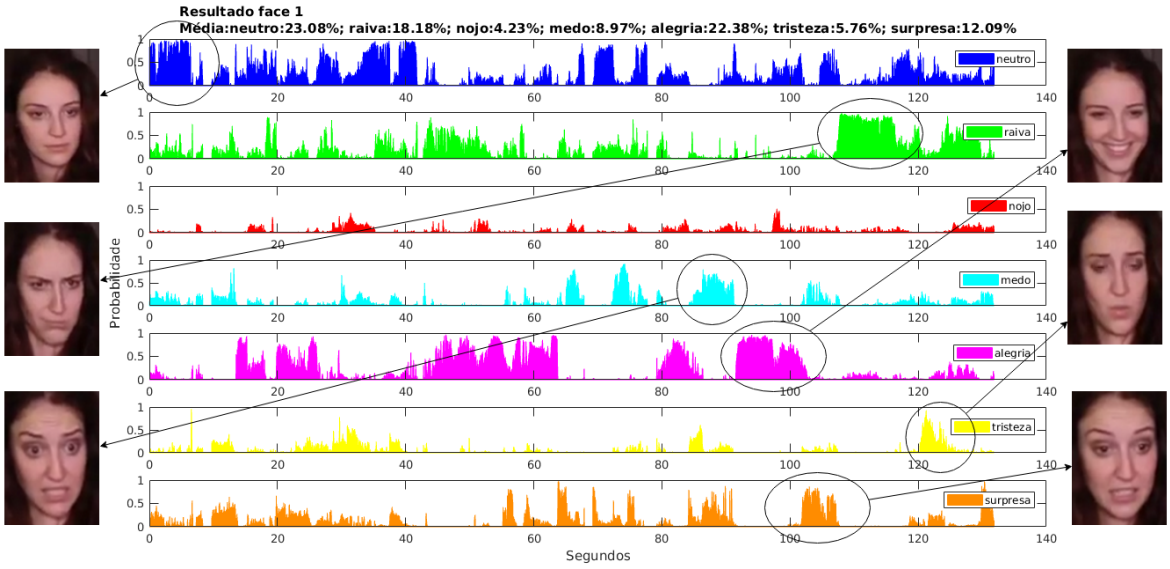

Figura 4. Análise de emoção face 1

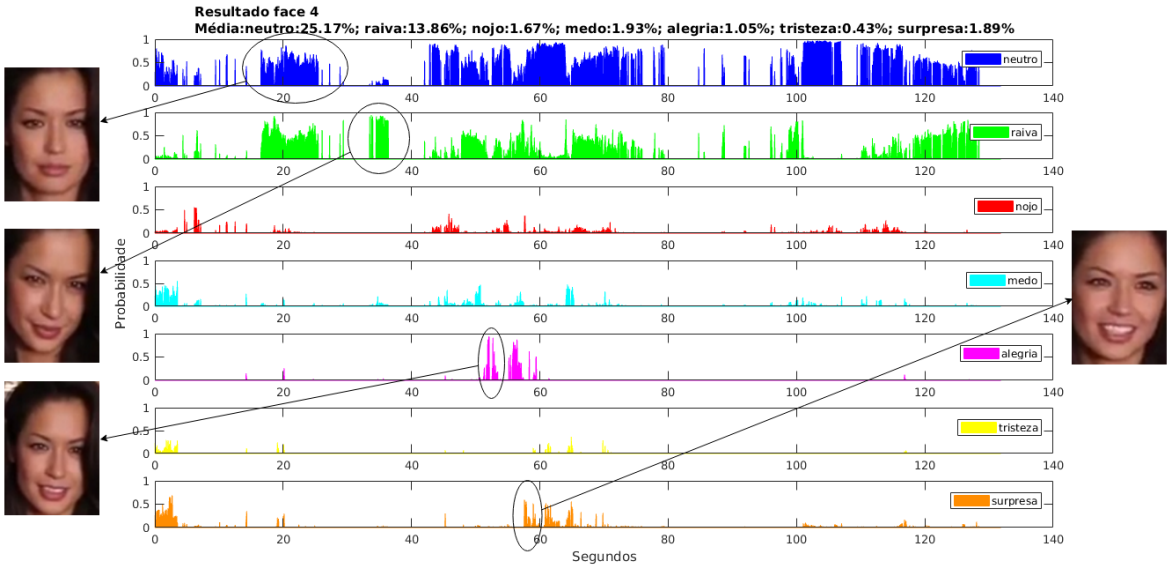

Figura 5. Análise de emoção face 4

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[10] R. K. Eichelberger and V. S. Sheng, "Does one-against-all or one-against-one improve the performance of multiclass classifications?" in *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, 2013.

[11] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[12] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[13] R. Baeza-Yates and B. Ribeiro-Neto, *Recuperação de Informação - 2ed: Conceitos e Tecnologia das Máquinas de Busca*. Bookman Editora, 2013. [Online]. Available: https://books.google.com.br/books?id=YWk3AgAAQBAJ

[14] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[15] J. KOAY, J. ROBINSON, A. KIRK, and H. JAYMES, "Dilwale | shah rukh khan | trailer reaction discussion 4way," https://www.youtube.com/watch?v=2mutNo-vUi8, Nov 2016, (Accessed on 07/14/2017).