Kaio: Detector de Expressões Faciais e Classificador de Emoções

1. Introdução

Vivemos em um mundo cada vez mais conectado e imerso em ambientes repletos de dispositivos computadorizados. Tal fenômeno proporcionou o surgimento de estudos em temas como Computação Ubíqua e Pervasiva [Satyanarayanan 2001] e Internet das Coisas [Gubbi et al. 2013]. Desenvolvimentos nestes temas trazem como consequência a necessidade cada vez maior de interação entre o ser humano com máquinas e dispositivos eletrônicos. Novas formas de interação como uso de gestos [Mitra and Acharya 2007] e comandos de voz [Cohen et al. 2004] podem trazer mais agilidade e conforto para que o ser humano realize suas tarefas. Mas tem-se a preocupação em evitar que tais atividades sejam muito frias e monótonas.

O uso de interfaces homem-máquina que levem em consideração as emoções humanas vem sendo alvo estudos recentes. Por exemplo, [Fragopanagos and Taylor 2005], levantam a necessidade de que interações mais naturais com dispositivos sejam baseadas em forma de reconhecimento de emoções, permitindo a esses dispositivos se adaptarem e otimizarem o processo interativo. Assim, propõem o uso de técnicas de inteligência artificial sobre imagens e vozes para a detecção das emoções do usuário. No entanto, não discutem métodos objetivos de como tais técnicas poderiam compor aplicações práticas.

Neste artigo, propomos uma arquitetura baseada em visão computacional para detecção de emoção a partir de expressões faciais e apresentamos um estudo de caso focado na interação com personagens digitais. A arquitetura proposta visa na avaliação de métodos adequados de aprendizado de máquina e formas de interação com diferentes dispositivos. A aplicação em personagens digitais visa servir de modelo, embora também tenha a motivação de ser aplicado em situações de aprendizado de crianças, pessoas com necessidade especiais ou simples entretenimento.

O restante do artigo está estruturado da seguinte forma. Nesta seção apresentamos o problema e as métricas. A seção 2 apresenta uma análise detalhada sobre os dados. A seção 3 apresenta a arquitetura proposta e a seção 4 discute os resultados obtidos da experimentação. Finalmente, a seção 5 conclui o trabalho e apresenta trabalhos futuros.

1.1. Problema

O uso de ferramentas que levem em consideração as emoções humanas ainda são alvo de estudo em muitas pesquisas, pois problemas como este focam em encontrar uma forma de tornar mais atraente e natural a interação homem-máquina.

O objetivo deste trabalho é construir um sistema inteligente que leve em consideração as emoções humanas por meio da detecção de expressões faciais para interagir com um personagem digital.

Para tal será necessário construir um modelo de aprendizagem que classifique emoções baseado em métodos automáticos de aprendizagem de máquina. Portanto, será necessário o uso de técnicas destinadas a classificação multiclasse. E para simplificar esta tarefa, utilizaremos entradas de texto produzidas por meio de um detector facial a

parte, feito com a biblioteca *Android Mobile Vision - AMV* [Google 2017], em um módulo separado do classificador, para fornecer as características necessárias da detecção.

Uma vez definido o problema, para cada sequência de detecção produzida teremos 4 (quatro) características e 1 (uma) classe alvo sobre cada usuário. Esta detecção será feita pela biblioteca AMV em um aplicativo Android.

Para ficar claro o funcionamento desta detecção, vejamos uma sequência de expressões faciais detectada pela biblioteca AMV para um usuário com aparência de tristeza, conforme a Tabela 1:

Table 1. Sequência de expressões de tristeza

user	rate blink left	rate blink right	rate smile or not	feel
1	0.99	0.99	0.01	0
1	0.99	0.99	0.01	0
1	0.92	0.87	0.02	0
1	0.92	0.87	0.02	0
1	0.92	0.87	0.02	0
1	0.99	0.99	0.08	0
1	0.99	0.99	0.08	0
1	0.99	0.99	0.08	0
1	0.99	0.99	0.08	0
1	0.99	0.99	0.02	0

Neste exemplo a pessoa teve a liberdade de piscar qualquer um dos olhos, sorrir ou apenas abrir a boca, e todas estas ações serão interpretadas pelo detector como uma sequência de expressões faciais que serão compreendidas como emoções, e consequentemente, convertida em probabilidades, significando:

- Probabilidade de uma pessoa está sorrindo (*rate smile or not*)
- Probabilidade de uma pessoa estar com olhos abertos (*rate blink left or right*).

As probabilidades são valores entre 0.0 e 1.0 (inclusive) através dos campos **rate blink left, rate blink right** e **rate smile or not**, já as emoções são fornecidas pelo campo **feel** que corresponde ao nosso alvo para o modelo de aprendizagem, esta emoção varia entre 3 tipos, tais como: 0 - Tristeza; 1 - Raiva; e 2 - Felicidade.

Portanto, o objetivo do nosso classificador dado uma sequência de expressões faciais é classificar uma emoção em tristeza, raiva ou felicidade.

1.2. Métricas

No contexto de Recuperação de Informação (RI) e Inteligência Artificial (AI), o desempenho de um sistema inteligente para classificação de emoções pode ser mensurado por meio de métricas, tais como: acurácia, precisão, matriz de confusão e validação cruzada, conforme afirmam [Ikonomakis et al. 2005], [MIKAMI et al. 2009] e [Freire et al. 2009].

Acurácia: Segundo [Sabbagh 2011], mensura o total de classificações corretas dividido pelo total de exemplos. Calculada conforme a Figura 1.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FN + FP + TN)}$$

Figure 1. Acurácia

Precisão: Segundo [Freire et al. 2009], mensura o quanto que foi exato a resposta, em uma proporção de classificações relevantes. Calculada conforme a Figura 2.

$$\text{Positive Predictive Value or Precision} = \frac{TP}{(TP + FP)}$$

Figure 2. Precisão

Matriz de confusão: Segundo [Freire et al. 2009], a medida oferece uma efetiva avaliação do modelo de classificação considerando o número de classificações corretas e as classificações estimadas para cada classe em um conjunto de dados. Calculada por meio da Figura 3.

	Predicted	
	Positive	Negative
Actual True	**TP**	**FN**
Actual False	**FP**	**TN**

Figure 3. Matriz de confusão

Validação Cruzada: Segundo [MIKAMI et al. 2009] corresponde ao processo de divisão dos dados coletados em uma parte para treinamento e outra parte para testes. Segundo o autor, é um processo importante para analisar os resultados alcançados pelo modelo de aprendizagem, testando o sistema com dados reais diferente dos dados que ele foi treinamento.

2. Análise

Nesta seção é descrito a origem dos dados, extração de informações e características importantes que serão utilizadas pelo classificador de emoções.

2.1. Dados

Para que seja possível analisar e avaliar o classificador de emoções, segundo seu desempenho, dados de usuários foram coletados para o treinamento do sistema, em uma fase anterior ao uso do sistema, que ocorreu em 2 dias, com 12 pessoas. O treinamento capturou um total de 360 expressões faciais por usuário, divididas em 60 expressões para 1 emoção (tristeza, raiva ou felicidade). Entre as pessoas envolvidas, 2 foram mulheres, 10 homens, considerando pessoas com barba, sem barba, usando óculos e sem óculos. O treinamento considerou as seguintes regras:

- Treinar com 12 pessoas voluntárias;

- Treinar 6 emoções com a mesma pessoa;
- Executar 2 treinamentos para 1 emoção, considerando as 3 (três) emoções possíveis;
- Treinar 8 segundos para cada emoção escolhida;
- Realizar o treinamento das emoções de forma aleatória com as pessoas.

Uma questão a ser respondida, por exemplo, é o por quê que o treinamento ocorreu em apenas 8 segundos para cada emoção. A resposta é simples, escolhemos este valor com o objetivo de permitir que uma pessoa utilizasse o aplicativo de detecção em um tempo curto, de maneira que, fosse o mais natural possível e 8 segundos veio a ser uma realidade próxima a realidade humana, pois normalmente, quando uma pessoa olha para outra, o julgamento sobre o estado emocional da outra acontece em poucos segundos, diante disso justifica-se a escolha do tempo.

No final do treinamento, a base de treino resultou em **4264 expressões faciais** salvas no arquivo **detect.csv**, de todos os usuários e as 3 emoções de tristeza, raiva e felicidade. Esse treinamento serviu como subsídio para o sistema ser capaz de reconhecer as emoções. Para cada tupla das expressões capturada, as características e a classe alvo retornadas pela biblioteca AMV foram:

Características:

- *user*: Identificação do usuário que executou o treino;
- r*ate_blink_left*: Probabilidade do olho esquerdo está aberto (0.0 a 1.0);
- *rate_blink_right*: Probabilidade do olho direito está aberto (0.0 a 1.0);
- *rate_smile_or_not*: Probabilidade da pessoa está sorrindo (0.0 a 1.0);

Alvo:

- *feel*: Emoção do usuário (0-tristeza , 1-raiva, 2-felicidade).

3. Kaio: Detector de Expressões Faciais para Classificar Emoções

Esta seção apresenta uma visão geral da ferramenta proposta para detectar expressões faciais e classificá-las em emoções, descrevendo a sua arquitetura, implementação e processo de classificação para possibilitar uma nova forma de interatividade com personagens digitais.

3.1. Descrição da Ferramenta

A solução proposta consiste na arquitetura de uma ferramenta capaz de possibilitar uma nova forma de interatividade com personagens digitais. Diante disso, utilizando técnicas de aprendizagem de máquina e processamento de imagens, a ferramenta deverá classificar emoções por meio da detecção de expressões faciais. Assim, será possível acessar as informações de interação classificadas em níveis de emoção do usuário, a princípio dividida em: **Tristeza**, **Raiva** e **Felicidade**.

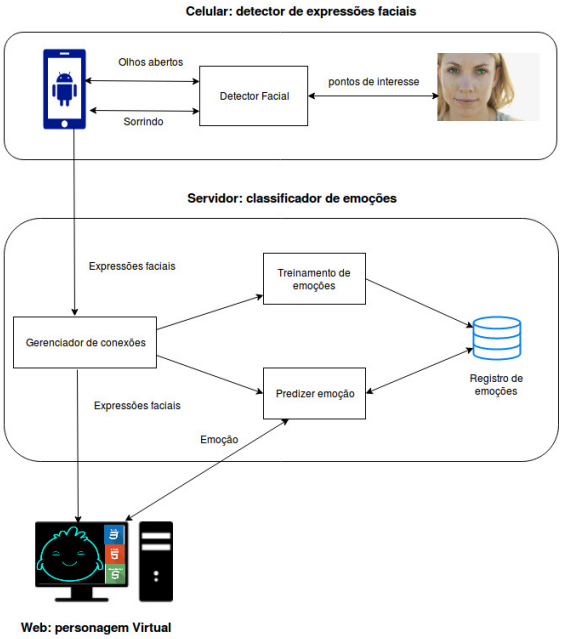

Figure 4. Processo de Funcionamento da Ferramenta

3.2. Arquitetura da Ferramenta

A arquitetura proposta foi distribuída em 3 (três) componentes:

- Celular: detector de expressões faciais;
- Servidor: classificador de emoções;
- Web: personagem virtual.

Conforme a Figura 4 apresentou temos uma arquitetura modular dividida em partes separadas com o objetivo de facilitar o desenvolvimento e garantir um baixo acoplamento de cada componente.

3.2.1. Celular: Detector de Expressões Faciais

Este componente é responsável por realizar a detecção das expressões faciais de um usuário por meio do uso de um celular com câmera e sistema operacional *Android*. O componente monitora pontos de interesses, tais como, olhos e boca por meio do uso da biblioteca AMV que repassa as informações para o dispositivo em forma de probabilidades, informando o quanto que o usuário está sorrindo (*smilling*) e as chances de estar com olhos abertos (*eyes open*).

A detecção dos pontos de interesse são transmitidas para o servidor responsável por classificar emoções, que recebe essa informação, e repassa para o componente personagem virtual.

3.2.2. Servidor: Classificador de Emoções

O principal componente neste trabalho é responsável por classificar emoções por meio da detecção de expressões faciais. Neste encontra-se os módulos para treinamento e

predição, além do gerenciador de conexões.

- **Módulo de Treinamento**: Na Figura 4 o módulo ”Treinamento de emoções” corresponde a fase que o sistema precisa aprender por um determinado tempo expressões faciais de usuários. Cada usuário ao submeter-se a detecção de expressões faciais, fornece para o sistema informações do tipo: taxa de olho esquerdo aberto; taxa de olho direito aberto; e taxa de sorriso. Estas informações são primordiais para a construção do modelo neste trabalho, e servirão como referência para classificar as emoções, entretanto, é importante frisar, que o treinamento não ocorre durante o uso do sistema, e sim em uma fase anterior, conforme será descrito na seção 4;
- **Módulo de Predição**: Na Figura 4 o módulo “Predizer emoção” corresponde a fase que o sistema já possui uma base de conhecimento com o registro de emoções, já passou pela fase de treinamento, e conhece algumas expressões faciais. Desta maneira, emprega-se um modelo de aprendizagem para ser testado e avaliado em cima da base de conhecimento, preparando para predições. No final, as emoções classificadas irão variar em três níveis, indicando a emoção do usuário em: “Tristeza”, “Raiva” e “Felicidade”.
- **Módulo Gerenciador de Conexões**: Na Figura 4 o módulo ”Gerenciador de Conexões” é o responsável por delegar as operações solicitadas ao servidor que poderá optar por ativar dois módulos: treinamento ou predição.

3.2.3. Web: Personagem Virtual

Este componente é responsável por reproduzir a nova forma de interação proposta neste trabalho, e foi desenvolvido com o objetivo de demonstrar as possibilidades de interação. Ele recebe o processamento da detecção, e executa as animações de um personagem digital, denominado ”Kaio”, que de acordo com o estado do usuário, assume os estados de piscar e sorrir, imitando o usuário. No final da interação, este componente prediz qual estado emocional o usuário se encontra, classificado nos três níveis citados anteriormente.

4. Experimentos e Resultados

Esta seção descreve os dados que foram analisados e os experimentos para avaliar o sistema de classificação das emoções.

4.1. Avaliação

A avaliação do sistema ocorreu em 3 dias, com 10 pessoas diferente das pessoas voluntárias do treinamento. Nesta fase, foram consideradas as seguintes regras:

- Avaliar com 10 pessoas voluntárias;
- Avaliar 3 vezes com a mesma pessoa;
- Permitir que a mesma pessoa escolha a ordem que desejar treinar, sem exceder as 3 tentativas por pessoa;
- Permitir que a pessoa voluntária anote a emoção que ela realizou e qual o sistema classificou;
- Ter 2 pessoas avaliadoras para julgar a emoção que a pessoa voluntária executou.

No final da avaliação, uma base de 30 observações foi adquirida, que serviu de referência para análise e os resultados alcançados neste trabalho. Para cada observação, as pessoas envolvidas e os avaliadores julgaram uma emoção, e a pessoa voluntária observou também a emoção que o sistema realmente classificou. As emoções foram classificadas em: 0 - Tristeza; 1 - Raiva; e 2 - Felicidade.

4.2. Algoritmo de aprendizagem

Para escolher o melhor modelo de aprendizagem para o sistema, foi elaborado um comparativo empírico baseado na acurácia entre as principais técnicas de aprendizagem de máquina destinadas a classificação. Segundo [Calvo and D'Mello 2010], [Forsyth and Ponce 2011] e [Pradhan 2013], as principais técnicas empregadas geralmente são Redes Neurais, Modelos Estatísticos, Modelos Genéticos e Árvores de Decisão. Para escolher o melhor algoritmo elencou aquele que obteve a maior precisão durante os experimentos, conforme a Figura 5

Algoritmo	Taxa de acerto (%)
DecisionTreeClassifier	84.54%
ExtraTrees	74.71%
KNeighborsClassifier	72.21%
AdaBoostClassifier	63.28%
RBF SVM	62.60%
OneVsOne	61.30%
Linear SVC	60.96%
OneVsRest	60.96%
QuadraticDiscriminantAnalysis	59.81%
GaussianNB	59.47%
MLPClassifier	59.34%
MultinomialNB	57.21%
SVC with Kernel Linear	53.66%
Algoritmo base	36.77%

Figure 5. Ranking dos algoritmos com maior precisão

Nesse sentido, verificou-se que a técnica de melhor desempenho dentre todas, foi a Árvore de Decisão, a qual foi adotada neste trabalho. Segundo [Du and Zhan 2002] o uso da Árvore de Decisão se destaca em relação ao fácil nível de interpretação, uma classificação explícita, além de baixo custo computacional gerado pela técnica. Outro aspecto interessante, segundo [Kołakowska 2013], são os desempenhos desta técnica, pois dependendo das configurações, os algoritmos podem melhorar sem a necessidade de mais informações.

4.3. Análise e Resultados

Nesta fase foram empregadas as métricas descritas, analisando o desempenho do sistema.

4.3.1. Validação Cruzada

Dividiu-se 90% das 4264 expressões faciais para treino e 10% para testes, então com uma matriz de confusão foi possível verificar o desempenho do sistema, conforme a Figura 6.

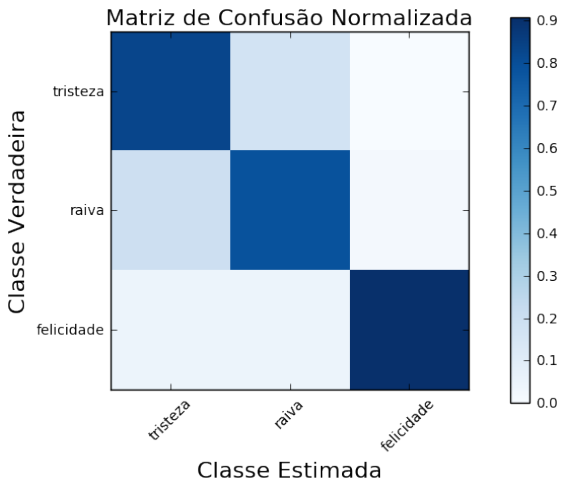

Figure 6. Matriz de Confusão do Sistema

De forma geral, observa-se que o sistema atingiu uma alta precisão para classificar as emoções, pois obteve mais acertos do que erros. A emoção denominada como felicidade, o sistema parece não confundir tanto com as outras emoções, entretanto, para tristeza e raiva o sistema parece se confundir algumas vezes entre as duas emoções.

4.3.2. Comparativo com Avaliação Humana

Nesta fase de avaliação, comparamos a precisão do sistema versus a observação humana feita por dois avaliadores, baseado nos testes feitos pelas 10 pessoas, durante a fase de avaliação. Conforme a demonstra Figura 7.

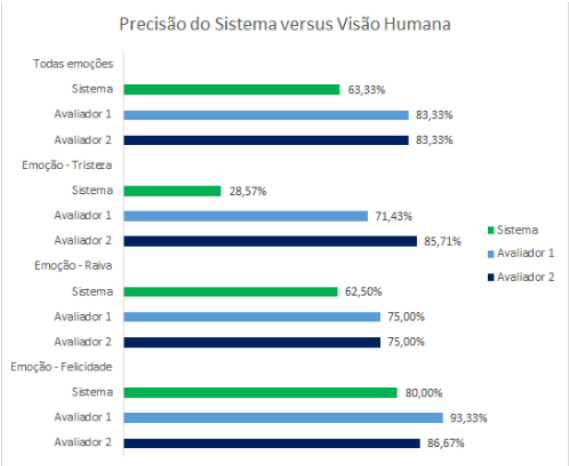

Figure 7. Precisão do Sistema versus Visão Humana

As três primeiras barras correspondem às precisões das avaliações de todas as emoções, classificadas pelo sistema, avaliador 1 e avaliador 2, em relação aos testes feitos pelas pessoas voluntárias durante a avaliação. Pode-se observar que a precisão dos avaliadores humanos em relação à emoção detectada pelos avaliadores foi de 83%, o que mostra um ruído natural na comunicação entre humanos. Já o sistema proposto acertou 63%, 20 pontos percentuais a menos que os avaliadores humanos. Isso se deve à menos características do sistema na detecção de pontos de expressão da face.

As outras barras na Figura 7, demonstram o desempenho do sistema versus a visão humana para cada emoção de forma individual. É possível observar que os acertos do sistema em relação aos avaliadores se manteve em 20 pontos percentuais a menos para a emoção de Raiva, no caso da Felicidade, os acertos ficaram entre 6 e 13 pontos percentuais a menos em relação aos avaliadores, relevando a melhor classificação do sistema. Entretanto, para Tristeza, os acertos não foram maiores que os erros, revelando que o sistema mostrou dificuldade para classificar corretamente esta emoção.

5. Conclusão

O trabalho permitiu avaliar e apresentar um estudo de caso focado na interação com personagens digitais para classificar a emoção do usuário por meio da detecção de expressões faciais. Os resultados mostram que mesmo seres humanos apresentam erros ao interpretar emoções de outros seres humanos, mas que métodos automáticos, como o proposto neste trabalho, podem ser úteis em alguns cenários.

Para algumas emoções como raiva e tristeza, o sistema não teve um desempenho tão alto quanto a classificação da felicidade, devido aos poucos atributos fornecidos pela biblioteca *Android Mobile Vision* para detecção facial. No entanto, a possibilidade de acertar a emoção felicidade com precisão de 80% já pode ser muito útil para aplicações de entretenimento infantil como proposto. A biblioteca oferece apenas 3 pontos de interesse - olhos (esquerdo e direito) e boca, talvez se fosse possível ter outros pontos de detecção na biblioteca como a sobrancelha, poderia permitir diferenciar a Raiva de Tristeza por meio de mais características, isso fica como proposta para trabalhos futuros.

Todavia, espera-se que o trabalho possa servir como modelo para interação entre diferentes dispositivos e principalmente para personagens digitais. E se for possível aplicar a ferramenta no entretenimento ou até mesmo como ferramenta educacional, será um ganho a mais para a pesquisa.

References

Calvo, R. A. and D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37.

Cohen, M. H., Cohen, M. H., Giangola, J. P., and Balogh, J. (2004). *Voice user interface design*. Addison-Wesley Professional.

Du, W. and Zhan, Z. (2002). Building decision tree classifier on private data. In *Proceedings of the IEEE International Conference on Privacy, Security and Data Mining - Volume 14*, CRPIT '14, pages 1–8, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

Forsyth, D. and Ponce, J. (2011). *Computer vision: a modern approach*. Upper Saddle River, NJ; London: Prentice Hall.

Fragopanagos, N. and Taylor, J. G. (2005). Emotion recognition in human–computer interaction. *Neural Networks*, 18(4):389–405.

Freire, P. P., de Oliveira Lombardi, L., Ciferri, R. R., Thiago Alexandre Salgueiro Pardo, C. D. d. A. C., and Vieira, M. T. P. (2009). Relatório técnico: Métricas de avaliação. projeto: Um ambiente para análise de dados da doença anemia falciforme.

Google (2017). Detect Faces. URL: https://developers.google.com/vision/ Access in 2017–03–22.

Gubbi, J., Buyya, R., Marusic, S., and Palaniswami, M. (2013). Internet of things (iot): A vision, architectural elements, and future directions. *Future generation computer systems*, 29(7):1645–1660.

Ikonomakis, M., Kotsiantis, S., and Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS transactions on computers*, 4(8):966–974.

Kołakowska, A. (2013). A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *Human System Interaction (HSI), 2013 The 6th International Conference on*, pages 548–555. IEEE.

MIKAMI, R., SANTOS, L., VENDRAMIN, A., and KAESTNER, C. (2009). Procedimentos de validação cruzada em mineração de dados para ambiente de computação paralela. *Artigo do Departamento Acadêmico de Informática Universidade Tecnológica Federal do Paraná, Curitiba, Brasil*.

Mitra, S. and Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3):311–324.

Pradhan, B. (2013). A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using gis. *Computers & Geosciences*, 51:350–365.

Sabbagh, R. (2011). Precisão e acurácia em estimativas ágeis. URL: https://www.infoq.com/br/articles/precisao-acuracia-agile/ Access in 2017–03–22.

Satyanarayanan, M. (2001). Pervasive computing: Vision and challenges. *IEEE Personal communications*, 8(4):10–17.