



Fundação Universidade Federal do ABC

Pró reitoria de pesquisa

Av. dos Estados, 5001, Santa Terezinha, Santo André/SP, CEP 09210-580

Bloco L, 3ºAndar, Fone (11) 3356-7617

iniciacao@ufabc.edu.br

Relatório Final de Iniciação Científica
referente ao Edital: 01/2019

Nome do aluno: Erick Hotta Orsi

Assinatura do aluno:

Nome do orientador: João Henrique Ranhel Ribeiro

Assinatura do orientador:

Título do projeto: Câmeras de Profundidade na Criação de Interface Homem-Máquina

Palavras-chave do projeto: Interface homem-máquina, linguística, multimodalidade, modelos de classificação, expressões faciais.

Área do conhecimento do projeto: Multidisciplinar, Engenharia Elétrica e Eletrônica, Inteligência Artificial.

Bolsista: Não.

Santo André

24 de agosto de 2020

Sumário

1 Resumo	1
2 Introdução.....	1
3 Fundamentação teórica	2
3.1 Interação Multimodal.....	2
3.2 Classificação de Emoções Humanas	2
4 Metodologia	4
4.1 Materiais e Métodos	4
4.2 Etapas da pesquisa	6
5 Resultados e discussão dos resultados.....	7
6 Conclusões e perspectivas de trabalhos futuros	8
Referências.....	9

1 Resumo

O uso simultâneo de diferentes modalidades de dados está cada vez mais comum em aplicações de interação humano-computador, como por exemplo, o uso de dados visuais em conjunto com dados de áudio. O uso de ferramentas que permitem a multimodalidade de dados torna a interação entre computadores e humanos mais natural, já que humanos percebem a realidade por meio de múltiplos canais (som, visão, tato, etc) de seus sentidos. Em especial, a interação entre seres humanos é muito influenciada pela percepção de expressões faciais de outras pessoas e, conseqüentemente, de suas emoções. Atualmente, existem muitas pesquisas em que são desenvolvidas formas de utilizar dados audiovisuais concomitantes, incluindo as emoções, para melhorar a interação entre pessoas e avatares virtuais, por exemplo. Para isso, é importante que exista interdisciplinaridade entre linguistas e cientistas da computação, dentre outros, que possam realizar e analisar a classificação de expressões faciais. Entretanto, um grande problema em classificação de emoção consiste na necessidade de dados rotulados para realizar o treinamento dos modelos de classificação. No contexto de multimodalidade audiovisual, é preciso que um linguista especializado rotule cada evento linguístico para posterior treinamento de máquinas. Dessa forma é preciso detectar mudança de expressão ao longo do *streaming* de vídeo, o que é um trabalho muito demorado.

O objetivo deste projeto foi desenvolver uma forma para automatizar a detecção e quando possível fazer a classificação de emoções em vídeos e aplicá-los a uma ferramenta de análise de dados multimodais em vídeo chamado *ELAN* [ELAN, 2020] para que a modalidade de expressões faciais possa ser juntada a outros eventos linguísticos identificados na plataforma.

2 Introdução

Este projeto foi realizado em paralelo com outro projeto, Pesquisa PIE633-2020 da UFABC: *Análise de requisitos linguístico-computacionais em interfaces presenciais homem-máquina*, projeto este aprovado no comitê de ética da UFABC (CAAE: 08069119.6.0000.5594). Nesse

projeto foram gravadas entrevistas humanos e um avatar humanoide utilizando-se três câmeras: uma com tomada no rosto (close-up) do humano, outra com tomada mais ampla que capta movimentos e gestos do humano, e outra com foco na tela do computador onde está reproduzido o avatar. As tomadas são sincronizadas e analisadas na ferramenta *ELAN*, um programa de computador que permite a anotação de múltiplas informações textuais no decorrer de um vídeo. Nesta ferramenta foram marcados eventos modais (gestos, palavras, prosódia, posturas, etc.) junto com os vídeos gravados de pessoas que participaram de uma entrevista, como parte do projeto. Contudo, é bastante oneroso detectar cada um dos eventos e marca-los no *ELAN*; por exemplo, um linguista deve marcar todos os eventos de expressões faciais para analisar sua função nas falas. Os vídeos das gravações que foram passados para esta pesquisa de IC não apresentam a modalidade de expressões faciais dos entrevistados. Um dos problemas para treinar máquinas é que há necessidade de uma rotulação dos dados para que o computador entenda o que deve ser lido e como deve ser classificado cada gesto ou cada expressão facial.

O objetivo deste projeto foi analisar ferramentas que consigam automatizar a marcação da modalidade de *expressões faciais*, para que pudesse ser aplicado no *ELAN* em conjunto com os outros eventos modais marcados e comparar com as marcações dos linguistas. Para isso, foi necessário o estudo de ferramentas e de técnicas de classificação de emoção em vídeos, além do desenvolvimento de um método para gerar uma trilha das emoções classificadas que possa ser aplicada no *ELAN*.

3 Fundamentação teórica

3.1 Interação Multimodal

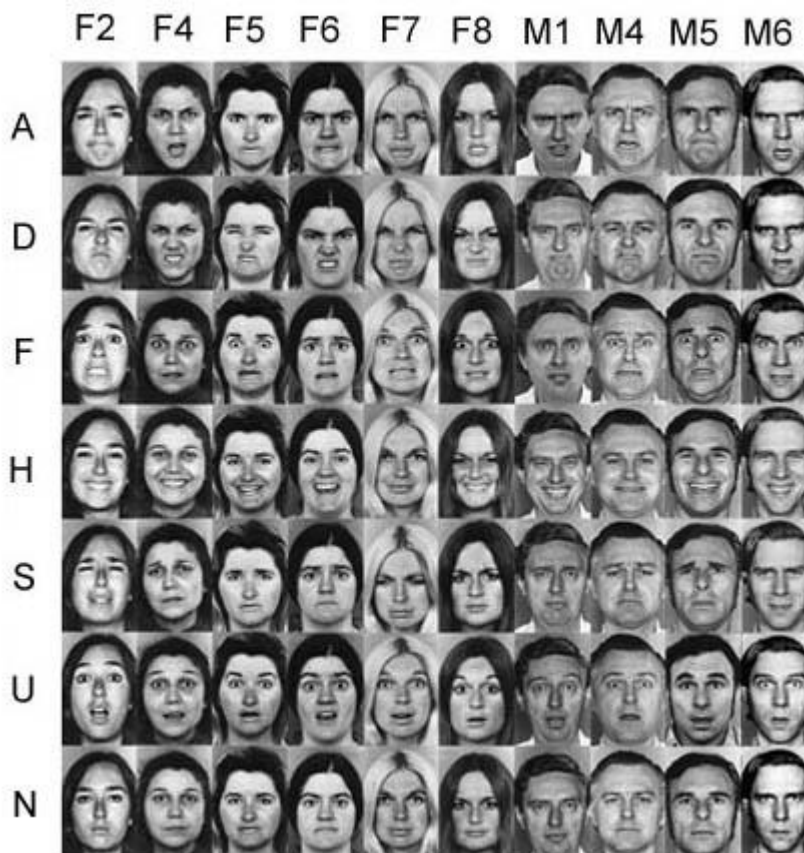
Historicamente, interação humano-computador possui mais foco em entradas de dados unimodais, utilizando apenas dados em forma ou de texto ou de vídeo, por exemplo. Entretanto, seres humanos interagem com o mundo por meio de múltiplas modalidades de percepção, utilizando todos os sentidos em paralelo e em sequência, como visão, olfato e propriocepção, o que se chama multimodalidade [[TURK, 2014](#)]. Interfaces multimodais são estruturas de interação que incluem simultaneamente a entrada de dados de múltiplos módulos, como por exemplo fala e vídeo, o que resulta em padrões de reconhecimento e métodos de classificação mais sofisticados, mais significativos e mais próximos à classificação humana [[TURK, 2014](#); [LAHAT et al, 2015](#); [VIELZEUF et al, 2017](#)]. A ferramenta *ELAN* foi desenvolvida com objetivo de ser usada para pesquisa de multimodalidade [[WITTENBERG et al, 2008](#)]. Dentre as possibilidades de interações multimodais, a multimodalidade audiovisual é uma das mais comuns e envolve fala humana e visão simultaneamente. Atualmente, é muito usada, dentre outros, em processamento de fala, reconhecimento de fala e interação humano-computador, no qual o objetivo é tornar a interação mais natural e mais intuitiva [[SHIVAPPA et al, 2010](#); [LAHAT et al, 2015](#)].

3.2 Classificação de Emoções Humanas

Expressão facial é um dos sinais mais influentes de comunicação entre seres humanos para demonstrar estados emocionais e intenções. Em sistemas de interação humano-computador, o reconhecimento automatizado de emoções humanas é bem explorado para facilitar a comunicação entre humano e computador [[LI e DENG, 2020](#)]. Historicamente foram realizados diversos estudos de emoções humanas. Em particular, Paul Ekman definiu várias

classes de emoções, incluindo emoções compostas, microexpressões e emoções básicas universais [EKMAN, 1999; EKMAN 1993]. Essas emoções básicas universais são mostradas na Imagem 1: “A” (raiva), “N” (desdém), “D” (desgosto), “H” (prazer), “F” (medo), “S” (tristeza) e “U” (surpresa); e são atualmente comumente usadas como classes para treinamento de modelos de aprendizado de máquina para classificação de emoções humanas [KO, 2018; VIELZEUF et al, 2017].

Imagem 1 – Fotografias das expressões faciais



Fonte: Artigo “Facial Expressions and Emotion – Stimuli and Tests (FEEST)” [YOUNG; EKMAN et al, 2002]

Nesse contexto, existem diversas abordagens e métodos para reconhecimento de emoção em imagem. O uso de redes neurais convolucionais (CNN) [KUBAT, 1999] é comum na identificação de rostos humanos e classificação de emoções por meio da identificação de marcos faciais ou “Facial Landmarks” (FL) [KO, 2018; WU et al, 2018; JOHNSTON e CHAZAL, 2018]. Outras abordagens possíveis são por “Local Binary Patterns” (LBP; padrões binários locais) [AHONEN et al, 2014; PIETIKAINEN et al, 2011]; por “Histogram of Oriented Gradients” (HOG, histograma de gradientes orientados) [CARCAGNI et al, 2015; DONIA et al, 2014] ou por “You Only Look Once” (YOLO, você só olha uma vez) [REDMON et al, 2016; REDMON e FARHADI, 2017]; dentre outros, cada um com aplicações em diferentes situações específicas. Por meio de métodos de classificação de emoção atuais em conjunto com uma implementação de interação multimodal, pode-se gerar modelos de classificação que consigam classificar emoções com menos erros. Em [SEBE et al, 2005], são apresentados estudos de reconhecimento visual de expressões e de reconhecimento vocal de emoções humanas, nos quais é identificado que uma grande dificuldade no contexto de treinamento de modelos para reconhecimento de emoção multimodal consiste na dificuldade da rotulação dos dados. Tal

treinamento é complexo e interdisciplinar, sendo necessário que um linguista experiente marque precisamente quando ocorrem fenômenos de variação de emoção em vídeo a partir dos dados da fala, para que um modelo seja treinado corretamente utilizando os dois módulos simultaneamente. Entretanto, esse trabalho é demorado e, em muitas situações, inviável.

4 Metodologia

4.1 Materiais e Métodos

Com o objetivo de desenvolver uma forma de marcar emoções básicas de forma correta para ajudar linguistas no reconhecimento de emoções em vídeo, foi necessário um estudo de diferentes abordagens para reconhecimento de rosto humano em vídeo e classificação de emoção, simultaneamente. A aplicação principal em vista foi o melhoramento da interação entre humanos e avatares digitais que simulam comportamento. Para isso seria necessário que a classificação fosse possível em tempo real, o que implica na consideração das limitações computacionais na escolha da forma de realizar as inferências. A partir disso, foram estudadas técnicas de reconhecimento facial e classificação de emoção usadas atualmente, como o *YOLO*, *HOG*, *LBP* e *CNN*, com o intuito de entender quais técnicas seriam mais adequadas para este projeto.

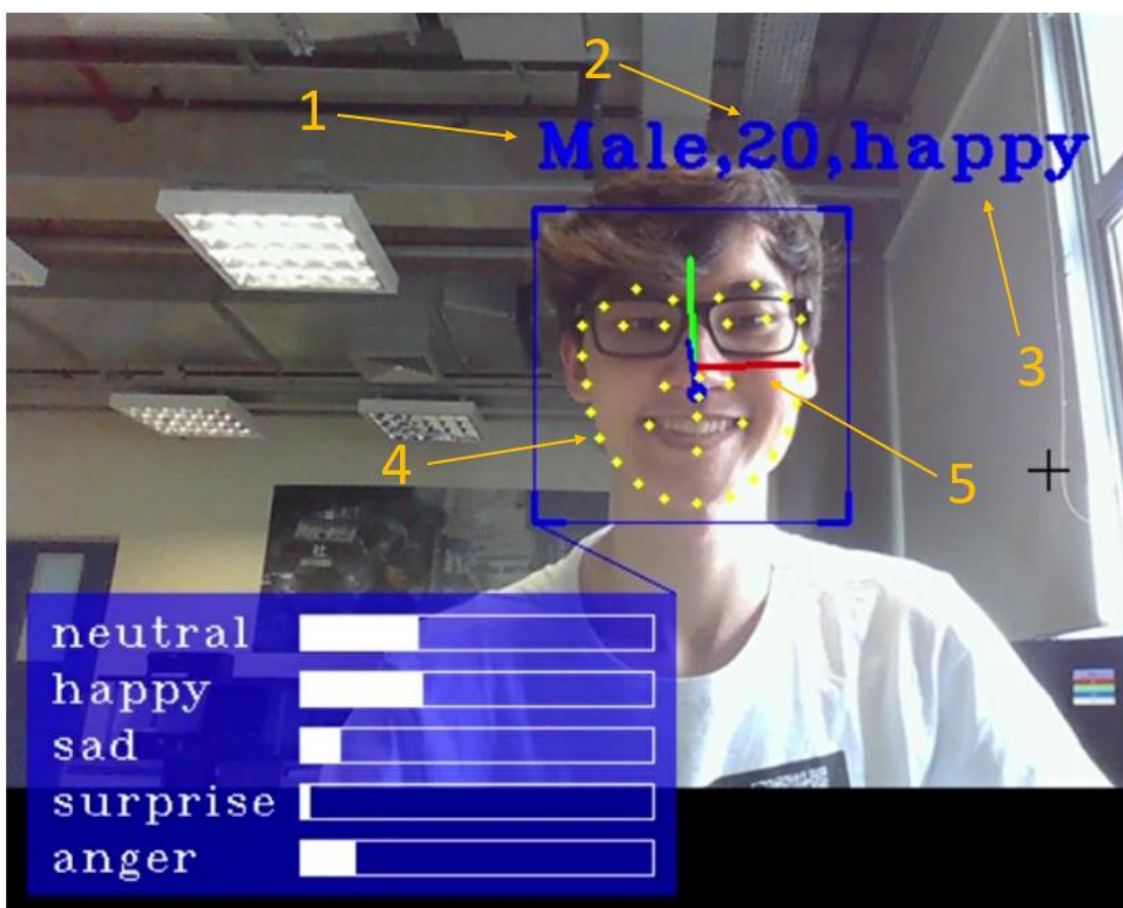
Inicialmente, foi prevista a colaboração com um projeto da Universidade de São Paulo (USP) em que as classificações de emoção seriam feitas na placa *Labrador*, do *Programa Caninos Loucos* [[CANINOS LOUCOS, 2020](#)]. Durante alguns meses no início do projeto, foram realizados estudos sobre o uso dessa placa e feitos testes de classificação para verificar a possibilidade de identificar emoções em tempo real. Porém, devido a uma incompatibilidade de hardware da câmera, não foi possível capturar imagens e realizar a classificação em tempo real, então a classificação foi feita em computador pessoal com processador *Intel(R) Core(TM) i7-1065G7 CPU @1.30GHz 1.50 GHz*, com 8.00 GB de RAM.

Para a classificação de emoções, foi escolhida a ferramenta chamada *OpenVINO™* [[OPENVINO, 2020](#)], especificamente o *Intel(R) Distribution of OpenVINO toolkit™*, que apresenta diversos modelos de inferência prontos e redes neurais otimizadas para processamento de imagens em CPU ou GPU. O *OpenVINO* realiza classificação por meio de redes neurais em conjunto com a biblioteca *OpenCV* [[OPENCV, 2020](#)], que utiliza classificadores do tipo *Haar Feature-Based Cascade* [[SHARIFARA et al, 2014](#)] e *HOG*. Em particular, no *OpenVINO* existe uma aplicação de reconhecimento de objeto chamada *Interactive Face Detection C++ Demo* que identifica rostos humanos, assim como a orientação da cabeça, os marcos faciais, o sexo biológico e a emoção de cada rosto. O modelo de inferência para a classificação de emoções é pré-treinada a partir do dataset *Affecnet* [[MOLLAHOUSSEINI et al, 2017](#)]. Este modelo reconhece apenas cinco emoções ('neutra', 'feliz', 'triste', 'surpresa', 'brava'), em vez das sete emoções classificadas pelo Paul Ekman, o que é uma diferença relevante no contexto de classificação de emoções. Apesar disso, essa foi a aplicação usada neste projeto, devido à otimização do modelo e à facilidade em aplicar e modificar o código.

Depois de estudar o funcionamento do *OpenVINO* e do modelo de inferência, foi possível aplicá-lo aos vídeos gravados no projeto de Análise de Requisitos para realizar a classificação de emoção. Em seguida, com o objetivo de desenvolver uma trilha de emoções para ser aplicada no *ELAN*, o código da aplicação do *OpenVINO* foi alterado para gerar um arquivo de texto em vez de um vídeo com as emoções em cada frame. Para isso, foram substituídas as linhas de código que desenham as emoções sobre os rostos e gravam cada frame em um vídeo novo por linhas de código que escrevem o número e a emoção classificada de cada frame em

um arquivo de texto, utilizando bibliotecas do *OpenCV*. Além disso, foram removidas as linhas de código que realizam classificação de idade, sexo biológico, marcos faciais e orientação da cabeça, já que não eram necessárias para esse projeto. Em seguida, o arquivo gerado foi reescrito, de forma a eliminar as linhas cuja emoção se repete, para que o arquivo apenas contenha os eventos de mudança de emoção. Em cada arquivo de texto, cada linha possui o número do frame em que há mudança de emoção, seguido de uma letra representando a emoção nova, separadas por vírgula. A Imagem 2 mostra uma captura de um vídeo gravado utilizando todas as funcionalidades do *Interactive Face Detection C++ Demo*, com classificações de sexo biológico (1), idade (2), expressão (3), marcos faciais (4) e orientação espacial (5).

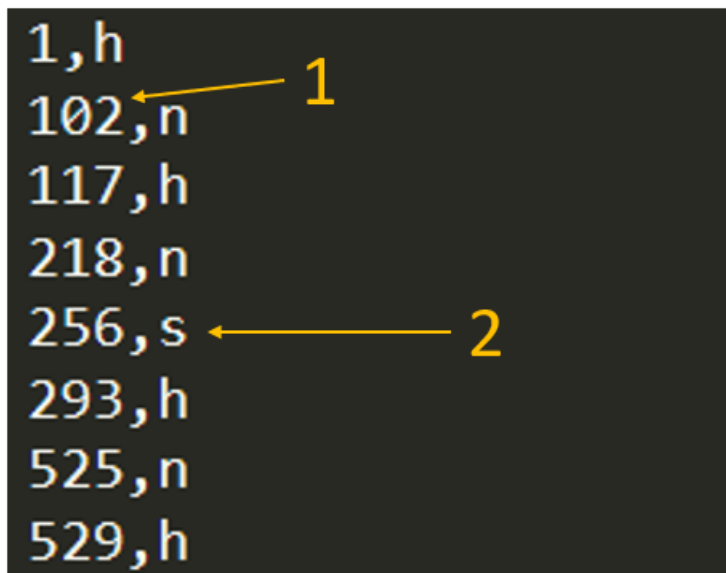
Imagem 2 – Classificações pelo *Interactive Face Detection C++ Demo* original



Fonte: Captura de um frame do vídeo gerado (2020)

A Imagem 3 mostra o arquivo de texto final gerado de forma automática, após as modificações do código, com o número do frame (1) e emoção detectada (2). Neste arquivo, cada linha possui o número do frame em que há mudança de emoção, seguido de uma letra representando a emoção nova, dentre os caracteres “n” (neutro), “h” (felicidade), “t” (tristeza), “f” (medo), “a” (raiva) e “0” (sem classificação).

Imagem 3 – Arquivo de texto gerado após modificação do código

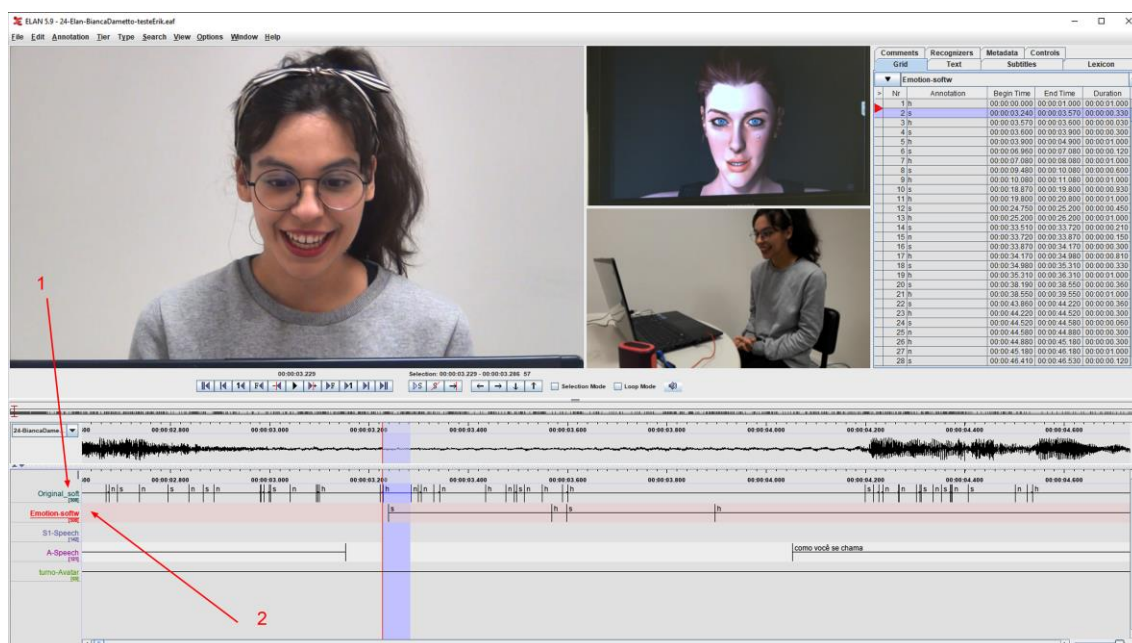


1,h
102,n
117,h
218,n
256,s
293,h
525,n
529,h

Fonte: Captura de parte do arquivo (2020)

A plataforma *ELAN* é capaz de criar uma trilha a partir desse arquivo de texto e, com isso, é possível comparar e juntar as classificações visuais, feitas com *OpenVINO*, com as classificações linguísticas, feitas com *ELAN*, como mostrado na Imagem 4, em que são mostradas a trilha original feita pelo *ELAN* (1) e a trilha corrigida a partir do arquivo de texto da classificação de emoções (2).

Imagem 4 – Trilhas *ELAN*



Fonte: Captura de tela da ferramenta *ELAN* com uma das entrevistadas no projeto de análise de requisitos (2020)

4.2 Etapas da pesquisa

Na Tabela 1 são ilustradas as etapas da pesquisa e os meses em que foram realizadas ao longo da duração da pesquisa, de setembro, 2019, a agosto, 2020.

Tabela 1 – Cronograma de etapas da pesquisa

Mês	1	2	3	4	5	6	7	8	9	10	11	12
Estudo de ferramentas												
Teste na placa												
Aplicação OpenVINO												
Criação de arquivos												
Relatório												

Fonte: Autoria própria (2020)

- Estudo de ferramentas:

Estudo de técnicas de classificação em imagem, com foco em classificação de rostos humanos e classificação de emoções humanas. Busca por ferramentas de inferência que realizam essas classificações, no qual foi escolhido o *OpenVINO*.

- Teste na placa:

Estudo e familiarização com a placa *Labrador*, do *Caninos Loucos*. Testes de classificação com redes neurais na placa. Testes de inferência da localização de rostos humanos e classificação de emoções. Posteriormente inutilizado devido a incompatibilidade de hardware da câmera e dificuldade de aplicação.

- Aplicação *OpenVino*:

Estudo e familiarização da ferramenta *OpenVINO*. Testes das inferências em computador pessoal, com foco na aplicação *Interactive Face Detection C++ Demo*.

- Criação de arquivos:

Alteração do código da *Interactive Face Detection C++ Demo* para geração automatizada de arquivos de texto durante as classificações, que foram analisados no projeto de Análise de Requisitos para criação de interfaces multimodais face-a-face.

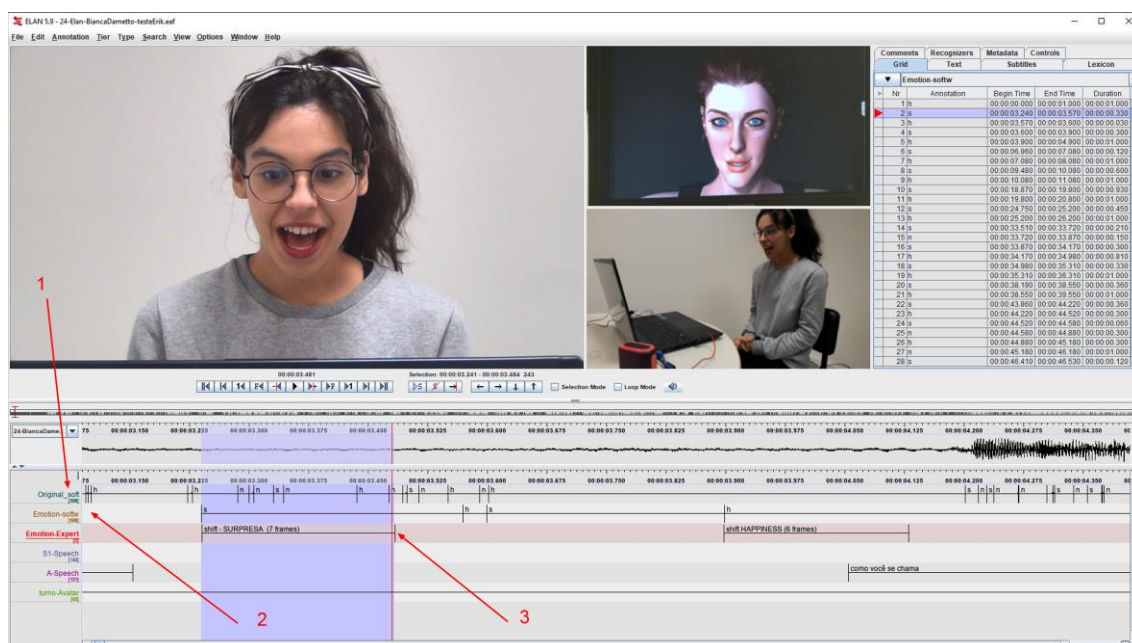
- Relatório:

Desenvolvimento do relatório final e do pôster para apresentação do projeto.

5 Resultados e discussão dos resultados

Foi possível desenvolver a automatização do processo de classificação de emoções humanas em vídeo de forma que seja aplicada à ferramenta *ELAN*. Assim, pôde-se juntar a trilha de emoções classificadas visualmente com os eventos modais linguísticos. Apesar do arquivo de texto gerar uma trilha confiável para o *ELAN*, na Imagem 5 é visível que existe algum erro temporal de alguns frames entre o que o programa marca (2) e a detecção manual por um humano especialista (3).

Imagem 5 – Trilha *ELAN* com anotações de especialista



Fonte: Captura de tela da ferramenta *ELAN* com uma das entrevistadas no projeto de análise de requisitos (2020)

Isso se deve ao erro do próprio classificador ao identificar a expressão facial aparente em cada frame, o que por sua vez se deve ao modelo de inferência utilizado. Neste projeto, foi utilizada ferramenta *OpenVINO*, para realizar a classificação de emoções, porém é possível realizar a classificação utilizando outros modelos de inferência ou modelos de treinamento, que possivelmente resultem em classificações mais assertivas.

6 Conclusões e perspectivas de trabalhos futuros

O objetivo de automatizar a geração de trilha de emoções classificadas para o *ELAN* foi alcançado, com a análise dos dados multimodais no *ELAN* realizado no projeto de Análise de requisitos linguístico-computacionais para criação de interfaces multimodais face-a-face. Como perspectivas para trabalhos futuros, um possível desdobramento é realizar as inferências com outros modelos de classificação de emoções ou realizar o treinamento de um modelo novo, com as sete emoções definidas pelo Ekman, para verificar se resultados seriam mais assertivos e se correspondem melhor às expectativas dos linguistas.

Outra possível aplicação consiste em fazer um banco de dados demarcados automaticamente para correlacionar os erros de marcação de uma dada emoção identificada pelo linguista, para calcular a taxa de acerto do programa e gerar uma trilha com mais acurácia, quando possível. Um possível desdobramento também pode ser obter por processo de regressão linear ou similar uma forma do software autocorrigir as distorções entre suas marcações e as marcações dos linguistas. Isso requer um trabalho minucioso de um linguista especialista e, por consumir muito tempo, não foi possível dentro desse projeto de pesquisa.

Além disso, a classificação automatizada desenvolvida neste projeto pode ser aplicada a capturas de vídeo em tempo real e ser adicionada como uma ferramenta de multimodalidade em interações humano-computador que possam se beneficiar de reconhecimento de

expressões faciais, como por exemplo avatares virtuais de conversa, cansaço dos usuários, estados psicológicos em entrevistas, e muitas outras aplicações.

Referências

1. ELAN (versão 5.9) [Software de computador]. *Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive*, 2020. Disponível em <<https://archive.mpi.nl/tla/elan>>
2. TURK, M.; Multimodal interaction: A review. *Pattern Recognition Letters*, 36, p. 189-195, 2014.
3. LAHAT, D.; ADALI, T.; JUTTEN, C.; Multimodal data fusion: An overview of methods, Challenges, and Prospects. *Proceedings of the IEEE*, vol. 103, no. 9, p. 1449-1477, 2015.
4. VIELZEUF, V.; PATEUX, S.; JURIE, F.; Temporal multimodal fusion for video emotion classification in the wild. *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI '17)*, p. 569-576. Association for Computing Machinery, New York, USA, 2017.
5. WITTENBERG, P. et al.; ELAN: A professional framework for multimodality research. *Proceedings of LREC, Fifth International Conference on Language Resources and Evaluation*, 2006.
6. SHIVAPPA, S. T.; TRIVENDI, M.; RAO, B. D.; Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey. *Proceedings of the IEEE*, vol. 98, no. 10, p. 1692-1715, 2010.
7. LI, S.; DENG, W.; Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, p. 1-1, 2020.
8. EKMAN, P.; Basic emotions. *Handbook of Cognition and Emotion*, Chapter 3, p. 45-60, 1999.
9. EKMAN, P.; Facial expression and emotion. *American Psychologist*, 48(4), p. 384-392, 1993.
10. KO, B. C.; A brief review of facial emotion recognition based on visual information. *Sensors*, 18(2), p. 401, 2018.
11. KUBAT, M.; Neural networks: A comprehensive foundation, de Simon Haykin Macmillan, 1994. *The Knowledge Engineering Review*, 13(4), p. 409-412, 1999.
12. WU, Y. et al.; Facial landmark detection with tweaked convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no.20, p.3067-3074, 2018.
13. JOHNSTON, B.; CHAZAL, P.; A review of image-based automatic facial landmark identification techniques. *J Image Video Proc.*, p. 86, 2018.
14. AHONEN, T.; HADID, A.; PIETIKAINEN, M.; Face recognition with local binary patterns. *Pajdla T., Matas J. (eds) Computer Vision – ECCV 2004, Lecture Notes in Computer Science*, vol. 3021, 2004.
15. PIETIKAINEN, M. et al.; Computer vision using local binary patterns. *Springer Science & Business Media*, 2011.
16. CARCAGNI, P. et al.; Facial expression recognition and histograms of oriented gradients: a comprehensive study. *SpringerPlus*, vol. 4, no.645, 2015.

17. DONIA, M. M.; YOUSSEF, A. A. A.; HASHAD, A.; Spontaneous facial expression recognition based on histogram of oriented gradients descriptor. *Computer and Information Science*, vol. 7, no. 3, 2014.
18. REDMON, J. et al.; You only look once: Unified, real-time object detection. *University of Washington, Allen Institute for AI, Facebook AI Research*, 2016.
19. REDMON, J; FARHADI, A.; YOLO9000: Better, faster, stronger. *University of Washington, Allen Institute for AI*, 2017.
20. SEBE, N. et al.; Multimodal approaches for emotion recognition: A survey. *Proceedings SPIE 5670, Internet Imaging VI*, 2005.
21. CANINOS LOUCOS [Programa de desenvolvimento de placas]. *Laboratório de Sistemas Integráveis Tecnológico (LSI-TEC)*, 2020. Disponível em <<https://caninosloucos.org/pt/>>
22. OPENVINO (versão 2020.4) [Ferramenta de Deep Learning]. *Intel®*, 2020. Disponível em <<https://software.intel.com/content/www/us/en/develop/tools/openvino-toolkit.html>>
23. OPENCV (versão 4.4.0) [Software de computador]. *OpenCV Team*, 2020. Disponível em <<https://opencv.org/about/>>
24. SHARIFARA, A.; RAHIM, M. S. M.; ANISI, Y.; A general review of human face detection including a study of neural networks and Haar feature-based cascade classifier in face detection. *2014 International Symposium on Biometrics and Security Technologies (ISBAST)*, p. 73-78, 2014.
25. MOLLAHOUSSEINI A.; HASANI, B.; MAHOOR, M. H.; AffecNet: A new database for facial expression, valence, and arousal computation in the wild. *IEEE Transactions on Affective Computing*, 2017.