# Uso de Técnicas de IA para Navegação Socialmente Aceitável

**Victor G. de Carvalho, Murillo R. Batista, Josué J. G. Ramos**

victorgdc@usp.br, (mbatista, jgramos)@cti.gov.br

**Divisão de Sistemas Ciberfísicos - DISCF**
**CTI/MCTI Renato Archer – Campinas/SP**

**Instituto de Ciências Matemáticas e de Computação**
**Universidade de São Paulo – São Carlos/SP**

***Abstract.*** *Over the past years, the interest in developing independent systems for emotional reflection has grown rapidly, involving environments where machines need to interact with or monitor humans. The focus of the project is related to the use of multimodal information to collect data about the emotional state of the target person to meet socially acceptable navigation objectives, through deep learning techniques. A base model with a convolutional neural network divided into three modules with specific functions was used. The network was trained on the EMOTIC dataset, containing images of people in a series of everyday scenarios.*

***Resumo.*** *Nos últimos anos, o interesse em desenvolver sistemas autônomos para reconhecer emoções cresceu rapidamente, envolvendo ambientes onde as máquinas precisam interagir ou monitorar humanos. O foco do projeto está relacionado ao uso de informações multimodais para coletar dados sobre o estado emocional da pessoa alvo para atender objetivos de navegação socialmente aceitável, através de técnicas de aprendizado profundo. Utilizou-se um modelo baseline com uma rede neural convolucional dividido em três módulos com funções específicas. A rede foi treinada no dataset EMOTIC, contendo imagens de pessoas em uma série de cenários cotidianos.*

## 1. Contexto do Problema

A navegação robótica em espaços públicos é uma tarefa complexa que requer o enfrentamento de uma variedade de desafios de engenharia e de fatores humanos. Esses desafios motivaram uma grande quantidade de pesquisas, resultando em desenvolvimentos importantes para os campos da robótica e da interação humano-robô nas últimas décadas [13], [14]. Ao mesmo tempo, abordar a navegação de robôs conscientes da atividade humana com uma arquitetura cognitiva levanta diversas dificuldades na integração dos componentes, assim como na orquestração de comportamentos e habilidades para realizar tarefas sociais.

Num cenário do mundo real, o sistema de navegação não deve considerar os indivíduos apenas como obstáculos. É necessário oferecer uma representação de pessoas particular e dinâmica para aprimorar a experiência da interação humano-robô [11], [12]. Os comportamentos do robô devem ser modificados por humanos, direta ou indiretamente. Em nossas vidas cotidianas e interações sociais, muitas vezes tentamos perceber os estados emocionais das pessoas ao nosso redor, mesmo que

involuntariamente. Por isso, há uma grande mobilização em pesquisas para dotar máquinas com capacidade semelhante de reconhecer emoções.

Do ponto de vista da visão computacional, a maioria dos esforços anteriores têm se concentrado em analisar as expressões faciais [5], [6] e, em alguns casos, também a postura corporal [7], [8]. Alguns desses métodos funcionam notavelmente bem em ambientes específicos. No entanto, seu desempenho é limitado em ambientes naturais e irrestritos. Estudos de psicólogos mostram que o contexto da cena, além da expressão facial e postura corporal, fornece informações importantes para nossa percepção das emoções das pessoas [9], [10]. No entanto, o processamento do contexto para o reconhecimento automático de emoções não foi explorado em profundidade, em parte devido à falta de dados adequados.

Nesse ínterim, é introduzida a dinâmica proposta por este artigo: utilizar o reconhecimento da diversidade de expressões humanas como elo de sustentação para a tomada de decisões mais socialmente aceitáveis durante o processo de navegação de robôs.

## 2. Dataset Utilizado

### 2.1 Apresentação da base de dados

Inicialmente, foram listados os datasets atrelados ao reconhecimento de expressões/emoções mais promissoras para a construção do modelo. Após período de análise e comparações, foi escolhido o conjunto de dados EMOTIC (2020). Esta versão, utilizada no treinamento da rede neural, é uma extensão daquela destacada no artigo [1].

O EMOTIC é uma coleção de imagens de pessoas em ambientes irrestritos anotados de acordo com seus estados emocionais aparentes. O conjunto de dados contém 23.571 imagens e 34.320 pessoas anotadas. Algumas das imagens foram coletadas manualmente da Internet pela pesquisa do motor de busca Google. O resto das imagens pertence a 2 conjuntos de dados de referência públicos: COCO [2] e Ade20k [3]. No geral, as imagens mostram uma ampla diversidade de contextos, contendo pessoas em lugares diferentes, ambientes sociais e realizando atividades diferentes.


**Figura 1. Três exemplos de imagens rotuladas do dataset e seus respectivos bounding boxes[18]**

O dataset combina dois tipos diferentes de representação de emoções:

- **Dimensões Contínuas:** as imagens são anotadas de acordo com o modelo VAD [4], que representa as emoções por meio de uma combinação de 3 dimensões contínuas: Valência, Excitação e Dominância. Na representação utilizada cada dimensão leva um valor inteiro que está no intervalo [1 - 10].
- **Categorias de Emoção:** Além do modelo VAD também, foi estabelecida uma lista de 26 categorias de emoções que representam diversos estados. A lista de categorias de emoções foi criada inicialmente coletando de forma manual um vocabulário afetivo de dicionários e livros sobre psicologia. Este vocabulário consiste em uma lista de aproximadamente 400 palavras, que foram reunidas em clusters de significado similar, gerando as 26 categorias que fundamentam o dataset.

### 2.2 Rotulagem e Divisão

Para coletar as anotações do conjunto de dados EMOTIC, foi utilizado crowdsourcing através da plataforma Amazon Mechanical Turk (AMT). Duas tarefas foram atribuídas, uma para cada um dos 2 formatos de representação de emoção. Os anotadores foram solicitados a rotular cada imagem de acordo com o que pensam que as pessoas nas imagens estão sentindo. Os cientistas responsáveis pelo processo confiaram na capacidade humana de fazer suposições razoáveis sobre o estado emocional de outras pessoas e no seu conhecimento sensorial e raciocínio sobre informações visuais

Após a primeira rodada de rotulagens (1 anotador por imagem), as imagens foram divididas em três conjuntos: Treinamento (70%), Validação (10%) e Teste (20%), mantendo uma distribuição de categorias afetivas semelhantes entre os diferentes conjuntos. Depois disso, Validação e Teste foram rotulados por 4 e 2 anotadores extras, respectivamente. Como consequência, imagens no conjunto de validação são rotuladas por um total de 5 anotadores, enquanto as imagens no conjunto de testes são rotuladas por 3 anotadores - apesar de que esses números podem variar devido à remoção de rotulagens ruidosas.

## 3. Arquitetura do Modelo

Inicialmente, o treinamento foi realizado a partir de uma configuração própria de camadas e neurônios, utilizando os conhecimentos obtidos no livro-texto da pesquisa [15]. Entretanto, os resultados não foram satisfatórios, e em seguida novas tentativas de treino foram realizadas com configurações neurais já consolidadas como AlexNet [16] e ResNet [17]. Apesar de serem arquiteturas prestigiadas, também não foram capazes de gerar um modelo robusto o suficiente para abordar o problema, com uma acurácia de menos de 30%.

Desse modo, diante da dificuldade de encontrar uma opção viável para realizar a tarefa, foi escolhido o modelo utilizado no artigo [18]. Se trata de uma rede neural convolucional dividida em 3 módulos: extração de características corporais; extração de características da imagem (contexto); rede de fusão;

O primeiro módulo captura o corpo visível da pessoa e gera features relacionadas a ele. Para capturar esses aspectos, o módulo foi pré-treinado utilizando a categoria “pessoa” do dataset ImageNet [16], que é centrado em objetos. O segundo módulo recebe a imagem inteira como entrada e gera features relacionadas à cena. Este foi pré-treinado com o dataset centrado em cenas Places [19].

Ambos os módulos de extração de features foram baseados na rede convolucional proposta em [20]. Finalmente, o terceiro módulo combina estes recursos para fazer uma regressão refinada dos dois tipos de representações emocionais (como descrito na seção 2.1).

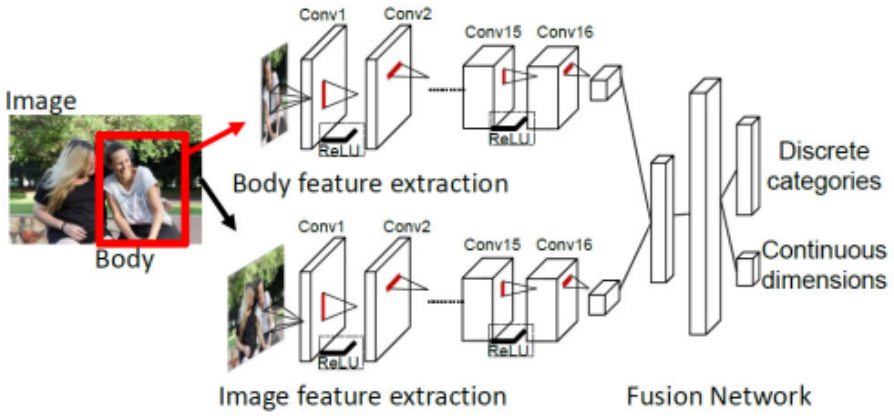


**Figura 2. Diagrama representativo do pipeline da rede neural[18]**

O módulo de fusão consiste em duas camadas totalmente conectadas. A primeira camada é usada para reduzir o dimensionalidade das features para 256 e, em seguida, uma segunda camada totalmente conectada é usada para aprender representações independentes para cada tarefa [21]. A saída desta segunda camada se ramifica em 2 grupos de saídas separadas: uma com 26 unidades representando as categorias de emoções discretas; e o segundo com 3 unidades representando as 3 dimensões contínuas.

## 4. Resultados

Para alimentar a rede neural e ao mesmo tempo satisfazer a tarefa de navegação, foi utilizado o reconhecimento de indivíduos em tempo real através do modelo YOLO [23], especializado em detecção de objetos.

A rede obteve um desempenho satisfatório no conjunto de treino, com uma acurácia de cerca de 88% e uma curva de perda progressivamente decrescente para os dois tipos de saídas: categóricas e contínuas.

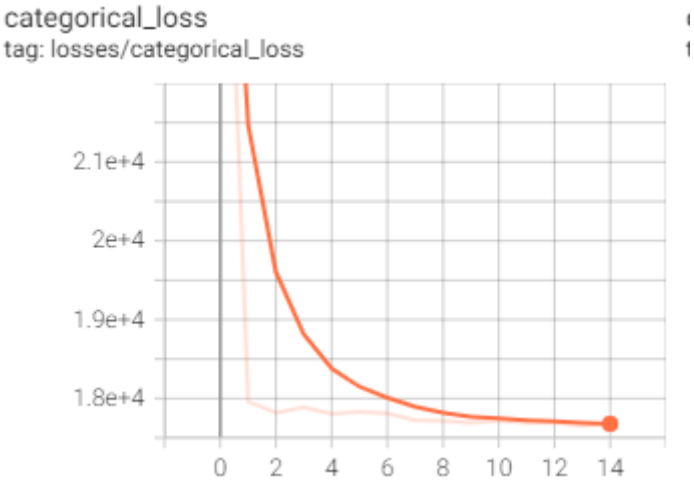


**Figura 3. Curva de Loss x Epochs dos atributos categóricos durante o treino**


**Figura 4. Curva de Loss x Epochs dos atributos contínuos durante o treino**

Já durante a fase de validação, a rede neural apresentou uma acurácia de quase 83%. Apesar das curvas de perda ruidosas, utilizou-se apenas os parâmetros nos pontos de mínimo.

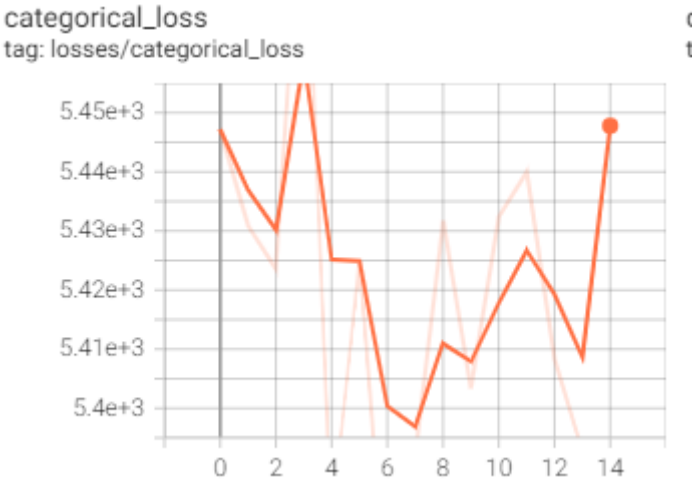


**Figura 5. Curva de Loss x Epochs dos atributos categóricos durante validação**

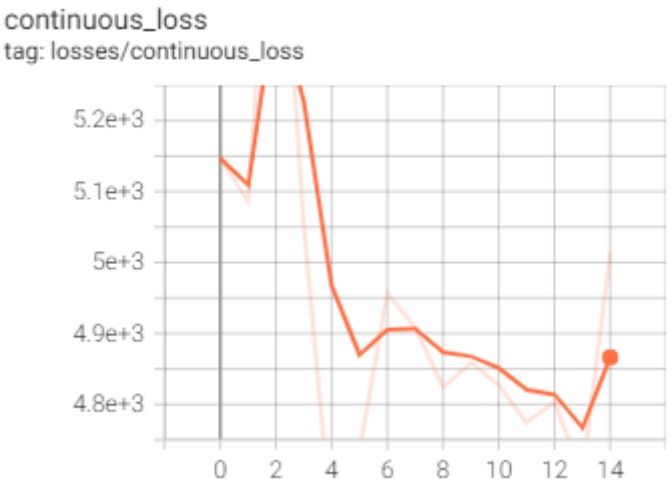


**Figura 6. Curva de Loss x Epochs dos atributos contínuos durante validação.**

As curvas de cor laranja acentuada representam a média móvel exponencial com suavização de 0.6, enquanto que as de cor opaca representam valores absolutos.

A fim de simplificar o problema, foi decidido que seriam utilizados apenas os valores contínuos para construir o sistema de navegação sociável, ou seja, somente o modelo VAD [4]. Entretanto, após testes utilizando imagens de câmera, percebeu-se que a variação destes valores era mínima quando se alterava a posição corporal ou expressões faciais. A partir daí, criou-se um impedimento na transferência de aprendizado sustentado no dataset para o sistema de navegação, o que acabou frustrando os planos iniciais.

## 5. Conclusão

Apesar de não ter alcançado o objetivo planejado, este estudo proporcionou aprendizados valiosos sobre a complexidade do reconhecimento de emoções e a natureza imprevisível das interações entre máquina e ser humano. Através das dificuldades enfrentadas e dos resultados não conformes, emergiram insights que podem ser usados como trampolim para pesquisas futuras. A ideia é que, após ajustar os parâmetros da rede neural ao contexto proposto, o modelo seja implementado na plataforma ROSANA [22] para realizar testes concretos a respeito da viabilidade dessa abordagem ao problema da navegação social de robôs.

## 6. Referências

[1] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, “Emotion recognition in context,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[2] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: common objects in context,” CoRR, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[3] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, “Semantic understanding of scenes through ade20k dataset,” 2016. [Online]. Available: https://arxiv.org/pdf/1608.05442

[4] A. Mehrabian, “Framework for a comprehensive description and measurement of emotional states.” Genetic, social, and general psychology monographs, 1995.

[5] M. Pantic and L. J. Rothkrantz, “Expert system for automatic analysis of facial expressions,” Image and Vision Computing, vol. 18, no. 11, pp. 881–905, 2000.

[6] Z. Li, J.-i. Imai, and M. Kaneko, “Facial-component-based bag of words and phog descriptor for facial expression recognition.” in SMC, 2009, pp. 1353–1358.

[7] A. Kleinsmith and N. Bianchi-Berthouze, “Recognizing affective dimensions from body posture,” in Proceedings of the 2Nd International Conference on Affective Computing and Intelligent Interaction, ser. ACII ’07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 48–58. [Online]. Available: http://dx.doi.org/10.1007/978-3-540- 74889-2 5

[8] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed, “Automatic recognition of non-acted affective postures,” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 41, no. 4, pp. 1027–1038, Aug 2011.

[9] L. F. Barrett, B. Mesquita, and M. Gendron, “Context in emotion perception,” Current Directions in Psychological Science, vol. 20, no. 5, pp. 286–290, 2011.

[10] L. F. Barrett, How emotions are made: The secret life of the brain. Houghton Mifflin Harcourt, 2017.

[11] Ginés Clavero, J., Martín Rico, F., Rodríguez-Lera, F.J. et al. Impact of decision-making system in social navigation. Multimed Tools Appl 81, 3459–3481 (2022).

[12] MAVROGIANNIS, Christoforos et al. Core challenges of social robot navigation: A survey. arXiv preprint arXiv:2103.05668, 2021.

[13] NING, Chen et al. Survey of pedestrian detection with occlusion. Complex & Intelligent Systems, v. 7, n. 1, p. 577-587, 2021.

[14] Y. Che, A. M. Okamura and D. Sadigh, "Efficient and Trustworthy Social Navigation via Explicit and Implicit Robot–Human Communication," in IEEE Transactions on Robotics, vol. 36, no. 3, pp. 692-707, June 2020, doi: 10.1109/TRO.2020.2964824.

[15] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O’Reilly.

[16] Krizhevsky Alex, Sutskever Ilya, and Hinton Geoffrey E.. 2012. ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25, F. Pereira, Burges C. J. C., L. Bottou, and Weinberger K. Q. (Eds.). Curran Associates, Inc., 1097–1105.

[17] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[18] Kosti, R., Alvarez, J. M., Recasens, A., & Lapedriza, A. (2019). Context based emotion recognition using emotic dataset. IEEE transactions on pattern analysis and machine intelligence, 42(11), 2755-2766.

[19] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva, "Places: An image database for deep scene understanding," CoRR, vol. abs/1610.02055, 2015

[20] J. Alvarez and L. Petersson, "Decomposeme: Simplifying convnets for end-to-end learning," CoRR, vol. abs/1606.05426, 2016

[21] R. Caruana, A Dozen Tricks with Multitask Learning, 2012, pp. 163–189

[22] PAIVA, Pedro VV et al. ROSANA: Robot for Social Interaction in Unstructured Dynamic Environments. In: 2020 Latin American Robotics Symposium (LARS), 2020 Brazilian Symposium on Robotics (SBR) and 2020 Workshop on Robotics in Education (WRE). IEEE, 2020. p. 1-6.

[23] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).