Projeto de pesquisa

Atenção visual para sistemas robóticos com *Deep Learning*

Aluno: Erik de Godoy Perillo
Orientadora: Profa. Dra. Esther Luna Colombini

Resumo

A informação visual é um elemento-chave no processo de entender o ambiente em que um ser se encontra. Para sistemas robóticos isso também é verdade. No entanto, a alta dimensionalidade dos dados captados por câmeras usadas para este fim é em geral problemática, muitas vezes havendo redundância e irrelevância de informação. Nos seres humanos este filtro sensorial é realizado pela Atenção. Neste contexto, este projeto propõe a aplicação de *Deep Learning* para a obtenção de sistemas de saliência visual para sistemas robóticos. Este trabalho dá continuidade ao trabalho anterior, onde as técnicas mais recentes para abordar o problema foram avaliadas, focando-se em um novo modelo competitivo com os atuais e otimizado para o domínio de robótica. Neste trabalho, visamos a extensão do modelo para fluxos contínuos de imagem, representados por vídeo.

Universidade Estadual de Campinas

26 de abril de 2017

1 Introdução

Um desafio ainda em aberto na robótica é a concepção de robôs que lidam com o imprevisível, reagindo de forma apropriada às mais diversas situações do mundo real. Sistemas que interagem com o ambiente, objetos e pessoas com uma variedade de maneiras semelhante à nossa têm o potencial de ser usados em diversar aplicações domésticas, industriais e em agricultura, sendo assim muito benéficos para a sociedade.

Um dos componentes fundamentais para tais sistemas de navegação autônoma é seu sistema visual. Dentre os diversos sub-problemas relacionados a se obter tal sistema, está o da saliência: dada uma imagem, qual região é mais relevante e como tal região varia no tempo?

O problema da saliência visual tem sido atacado de diversas maneiras há anos. Recentemente, com o progresso do *Deep Learning*, diversos novos modelos e técnicas apareceram com desempenho consideravelmente superior [9]. Modelos atuais, entretanto, focam apenas no problema da saliência visual para imagens estáticas. Em fluxos contínuos de imagens, ou seja, vídeos, há efeitos que não existem para imagens fixas que mudam o foco da atenção.

1.1 Motivação

Um problema de sensores usados para tarefas mais complexas de navegação em geral é que o volume de dados a ser processado pode ser demasiadamente grande. Para um robô que interage continuamente com o ambiente, é improvável que em todos os instantes toda informação proveniente de seus sensores seja processada ou mesmo necessária. Isso é especialmente crítico para sistemas robóticos que precisam de respostas rápidas para interagir com o ambiente em que estão inseridos.

1.2 Saliência visual

Saliência visual pode ser definida como uma região visual em que se dá foco para um maior trabalho cognitivo em um certo momento. [16] Com base no comportamento de humanos, modelos computacionais que simulam a atenção o fazem geralmente por meio de mapas de saliência, que são computados para uma certa imagem.

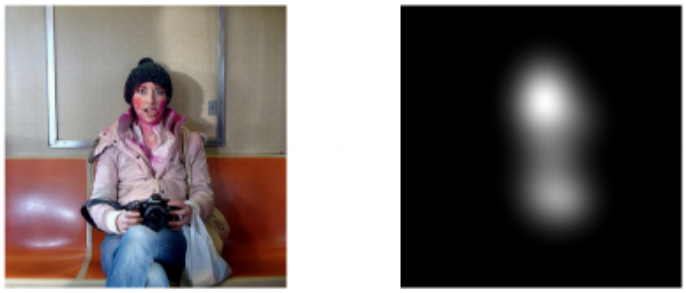

Figura 1: Exemplo de mapa de saliência para uma certa imagem [6].

Diversos métodos para geração de tais mapas foram propostos ao longo dos anos. Vocus [5], por exemplo, é um modelo que leva em consideração aspectos globais da imagem e contraste em relação a fatores como cor, direção, luminância em diversas dimensões.

A introdução de *Deep Learning* à área de saliência visual possibilitou mapas com desempenho superior em geral a modelos antigos, com a vantagem intrínseca da técnica que é a geração automática de *features* a serem extraídas. Modelos como *Salicon* [6], *DeepFix* [8] e *Salnet* [11] têm desempenho superior aos modelos atencionais tradicionais em todas as métricas estabelecidas pelo *MIT Saliency Benchmark* [9].

Considerando que nosso trabalho tem como foco usar modelos atencionais para agentes robóticos exploratórios [12], os modelos atuais apresentam dois problemas fundamentais: a) eles são demasiadamente pesados computacionalmente, dificultando a implementação em um sistema embarcado, e b) não são feitos para fluxos de imagens, ou seja, vídeos, não levando em consideração efeitos que acontecem neste cenário como inibição de retorno [3].

2 Objetivos

Neste contexto, este projeto tem por objetivo construir um modelo de saliência visual que seja mais eficiente computacionalmente (sem muita perda de desempenho) e que simule o comportamento de humanos com relação à variação espacial do foco atencional no tempo em fluxos contínuos de imagens. Mais especificamente, objetivamos:

- Obtenção de uma arquitetura de rede neural artificial otimizada para a extração de mapas de saliência da imagem e que seja relativamente eficiente computacionalmente;
- Adaptação da rede obtida para vídeo, com mecanismos extras que simulem o comportamento humano padrão;
- Implementação do sistema em uma plataforma de testes robótica para sua avaliação.

3 Materiais e Métodos

3.1 Modelo atual

As arquiteturas com melhor desempenho médio nas métricas do *MIT Saliency Benchmark* têm um grande número de camadas de convolução. Dentre *Salnet, DeepFix, Salicon*, as três usam transferência de aprendizado da já treinada rede *VGG-16* [15]. Tal rede chega a ter camadas de convolução com 512 filtros de dimensões 3x3. Os espaços de cor dos três modelos analisados são RGB e a normalização dos dados nos modelos é feita por todo o dataset.

O modelo desenvolvido por nós usa um número consideravelmente menor de parâmetros: são 6 camadas de convolução, mas o total dos filtros de todas as camadas é de 209. O espaço de cor utilizado foi o LAB, pois há indícios que este seja um espaço de cor que melhor representa o sistema visual humano [5]. A normalização dos dados foi feita por imagem individualmente, pois supõe-se que no contexto da saliência, o que mais importa é a relação dos pixels em um contexto global da imagem. O modelo é treinado com um conjunto de dados obtido pela mesclagem do *Salicon* [14], *MIT-1003* [7] *CAT-2000* [1].

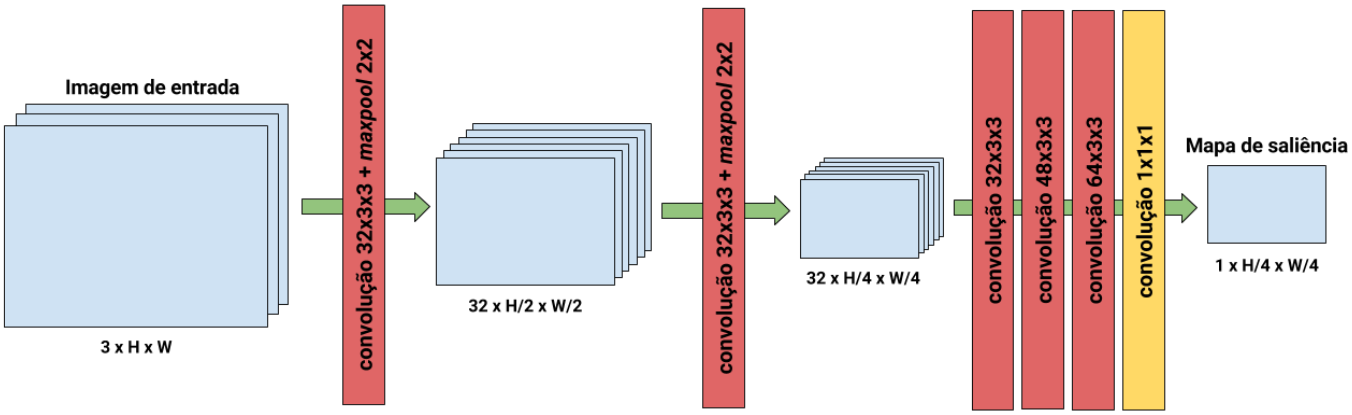

Figura 2: Arquitetura do modelo proposto atualmente [13].

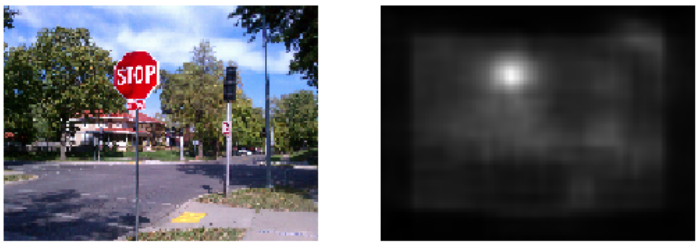

Figura 3: Predição de saliência do modelo atual para uma imagem.

O modelo ainda encontra-se em estágio de desenvolvimento, mas já consegue-se obter 0.6 na métrica *Coefficient of Correlation* (CC), que é considerada uma boa métrica que penaliza tanto falsos positivos quanto negativos [2]. Tal número foi obtido em um conjunto separado de testes que não foi visto pelo modelo durante o treinamento. Ainda não há dados para o conjunto utilizado no *MIT Saliency Benchmark*, mas o número obtido fica comparável a modelos como o *Salnet*.

Tabela 1: Resultados de modelos no *MIT Saliency benchmark* [9].

Modelo/Métrica	**Similarity**	**CC**	**AUC Judd**	**NSS**	**EMD**
DeepFix [8]	**0.67**	**0.78**	**0.87**	**2.26**	**2.04**
SALICON [6]	0.60	0.74	**0.87**	2.12	2.62
ML-Net [4]	0.59	0.67	0.85	2.05	2.63
Deep Convnet [11]	0.52	0.58	0.83	1.51	3.31
Shallow Convnet [11]	0.46	0.53	0.80	1.47	3.99

3.2 Extensão para vídeos

O modelo atualmente sendo trabalhado ainda precisa de refinamento e melhor treinamento, que será possível graças à obtenção recente de uma plataforma específica para treinamento da rede, isso será possível em breve. Após um refinamento, uma extensão para vídeos será feita. O foco atencional depende de diversos outros fatores além do *frame* atual [3] e espera-se incorporar tais fatores no modelo novo. Para treinamento, há conjuntos de dados como *Coutrot Database 1* e *SAVAM* [10].

3.3 Plataforma de testes

Como o foco do modelo é o uso por sistemas robóticos, planeja-se usar um robô móvel com uma GPU *NVIDIA Jetson TX1* embarcada. O objetivo é avaliar o desempenho da obtenção das regiões de saliência com o passar do tempo e se as escolhas de foco de região

são adequadas para auxiliar na tarefa de navegação. Objetiva-se também verificar se o modelo é eficiente computacionalmente o suficiente para que o processamento seja feito em uma janela de tempo adequada.

4 Cronograma

Para atendimento dos objetivos propostos, serão realizadas as seguintes etapas:

- FASE 1: Revisão bibliográfica.
- FASE 2: Preparação do ambiente computacional.
- FASE 3: Modelagem e otimização da rede para imagens fixas.
- FASE 4: Extensão da rede para vídeos.
- FASE 5: Implementação e testes em plataforma robótica.

Tarefa/mês	Jul	Ago	Set	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun
FASE 1	x											
FASE 2	x											
FASE 3		x	x	x	x							
FASE 4					x	x	x	x	x	x		
FASE 5										x	x	x

5 Referências

Referências

[1] Borji et al. "CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research". Em: *CVPR 2015 workshop on "Future of Datasets"* (2015). arXiv preprint arXiv:1505.03581.

[2] Zoya Bylinskii et al. "What do different evaluation metrics tell us about saliency models?" Em: *arXiv preprint arXiv:1604.03605* (2016).

[3] E.L. Colombini, A. da Silva Simoes e C.H. Costa Ribeiro. "An Attentional Model for Autonomous Mobile Robots". Em: *IEEE Systems* 99 (2016), pp. 1–12.

[4] Marcella Cornia et al. "A Deep Multi-Level Network for Saliency Prediction". Em: *arXiv preprint arXiv:1609.01064* (2016).

[5] Simone Frintrop. "VOCUS: a visual attention system for object detection and goal-directed search". Em: *IN LECTURE NOTES IN ARTIFICIAL INTELLIGENCE (LNAI)*. Springer, 2005.

[6] Ming Jiang et al. "Salicon: saliency in context". Em: *CVPR* (2015).

[7] T. et al Judd. "Learning to Predict Where Humans Look". Em: *Computer Vision, IEEE 12th International Conference* (2009).

[8] Srinivas S S Kruthiventi, Kumar Ayush e R. Venkatesh Babu. "DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations". Em: *arXiv preprint arXiv:1510.02927* (2015).

[9] *MIT saliency benchmark*. 2016. URL: `http://saliency.mit.edu/index.html` (acesso em 27/09/2016).

[10] *MIT saliency benchmark: datasets*. 2016. URL: `http://saliency.mit.edu/datasets.html` (acesso em 26/04/2017).

[11] Junting Pan et al. "Shallow and Deep Convolutional Networks for Saliency Prediction". Em: *arXiv preprint arXiv:1603.00845* (2016).

[12] Erik Perillo e Esther Colombini. *Processos Atencionais e Aprendizado de Máquina para Sistemas Robóticos*. 2016.

[13] *Repositório att*. 2017. URL: `https://github.com/erikperillo/att/tree/dev/att/deep` (acesso em 26/04/2017).

[14] *Salicon dataset*. 2016. URL: `www.example.com` (acesso em 27/09/2016).

[15] Karen Simonyan e Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". Em: *CoRR* abs/1409.1556 (2014). URL: `http://arxiv.org/abs/1409.1556`.

[16] Anne M. Treisman e Garry Gelade. "A Feature-Integration theory of Attention". Em: *Cognit Psychol* (1980).