

CATEGORIZAÇÃO DE OBJETOS UTILIZANDO ATENÇÃO VISUAL

MILTON ROBERTO HEINEN*, PAULO MARTINS ENGEL*

*Instituto de Informática, Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15064, 91501-970 Porto Alegre, RS, Brasil

Emails: mrheinen@inf.ufrgs.br, engel@inf.ufrgs.br

Resumo— Este artigo descreve um modelo de visão computacional baseado em técnicas de aprendizado de máquina, que tem por objetivo realizar a categorização de objetos de forma não supervisionada. Este modelo é composto de três elementos: um módulo de atenção visual bastante robusto em relação a transformações afins; um esquema de representação de informações visuais baseado em cores; e um algoritmo de categorização estatístico capaz de aprender as distribuições dos dados de entrada de forma não supervisionada. Este modelo visão computacional é validado através de diversos experimentos, que demonstram que ele consegue criar categorias que permitem a identificação de objetos de forma bastante estável.

Palavras-chave— Visão de Robôs, Atenção Visual, Categorização de Objetos, Reconhecimento de Objetos.

1 Introdução

A quantidade de informações que chega ao sistema visual dos primatas – estimada como sendo da ordem de 10^8 *bits* por segundo – excede em muito a capacidade que o cérebro tem de processá-la e assimilá-la em sua experiência consciente (Pashler, 1997). A estratégia utilizada pelos sistemas biológicos para lidar com este excesso de informações é processar de forma detalhada somente algumas partes do campo visual, chamadas de regiões de interesse, e ignorar o restante das informações (Niebur and Koch, 1998). Segundo (Desimone and Duncan, 1995), a seleção das regiões de interesse é dirigida por um mecanismo competitivo de controle de atenção, que facilita a emergência de um vencedor entre diversos alvos potenciais, permitindo ao sistema processar informações relevantes enquanto que suprime as informações irrelevantes que não podem ser processadas simultaneamente.

Inspirados nos sistemas de atenção biológicos, é possível desenvolver sistemas de atenção computacionais capazes de selecionar as regiões de interesse do campo visual, o que torna possível a análise de cenas complexas em tempo real com recursos limitados de processamento. Embora diversos modelos de atenção visual já tenham sido propostos e implementados (Koch and Ullman, 1985; Tsotsos et al., 1995; Itti et al., 1998; Frintrop, 2006), a maioria destes modelos tem como foco principal entender o funcionamento dos mecanismos de atenção dos seres vivos. Mas para que um modelo de atenção possa ser adequadamente utilizado em sistemas de visão computacional é necessário que: (i) ele seja relativamente insensível a transformações afins (rotação, translação, reflexão e escala); (ii) as escalas das fixações sejam selecionadas em conjunto com as posições das mesmas (Draper and Lionelle, 2005).

Com base nestes requisitos, um novo modelo de atenção visual, chamado de NLOOK, foi proposto e implementado (Heinen and Engel, 2008b; Heinen and Engel, 2009a; Heinen and Engel, 2009b). Este novo modelo, que possui um excelente desempenho computacional, é bem menos sensível a transformações afins que outros modelos de atenção como o NVT (Itti et al., 1998), que é o modelo de atenção visual mais conhecido e utilizado. Além disso, o NLOOK consegue selecionar tanto as posições como as escalas das fixações de forma bastante precisa e estável.

Neste artigo, a utilização do NLOOK é expandida de forma a permitir não somente a localização das regiões de interesse do campo visual (caminho *onde* do córtex visual), mas também a categorização não supervisionada de objetos (caminho *o que* do córtex visual). Este artigo está estruturado da seguinte forma: a Seção 2 descreve o modelo proposto neste artigo, bem como seus módulos; a Seção 3 descreve os experimentos realizados visando validar o modelo proposto; e a Seção 4 descreve as conclusões e perspectivas.

2 Modelo proposto

A Figura 1 mostra a arquitetura geral do modelo de categorização de objetos proposto neste artigo, que atua da seguinte forma: Inicialmente a imagem de entrada é enviada para o modelo de atenção visual NLOOK (Subseção 2.1), que extrai as regiões mais relevantes desta imagem, ou seja, os principais focos de atenção (FOAs). Estes focos de atenção são então representados utilizando a codificação angular de cores (*Color Angular Indexing* – CAI), descrita na Subseção 2.2, e a categorização das regiões de interesse é realizada utilizando o INBC, descrito na Subseção 2.3. Assim, o principal objetivo do modelo proposto não é realizar a identificação dos elementos da cena visual, mas sim categorizar os mesmos de forma não supervisionada,

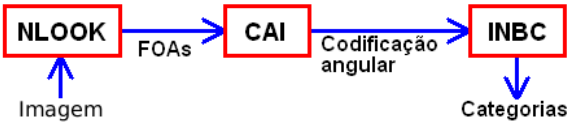


Figura 1: Arquitetura geral do modelo proposto

o que pode permitir futuramente a utilização das categorias como *landmarks* em um sistema de navegação visual. Nas próximas subseções, são descritas as características do modelo proposto em detalhes.

2.1 Modelo de atenção visual NLOOK

A Figura 2 mostra a arquitetura do NLOOK (Heinen and Engel, 2008b; Heinen and Engel, 2009a), que é o modelo de atenção visual utilizado para a extração das regiões de interesse do campo visual. O NLOOK é inspirado na conceitos de *scale-space* (Witkin, 1983).

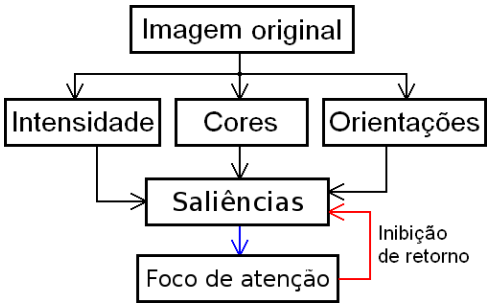


Figura 2: Arquitetura do NLOOK

Para a criação de um *scale-space*, a imagem de entrada é inicialmente sub-amostrada em diversos oitavos, e em seguida são criadas diversas escalas para cada oitavo através da convolução sucessiva das imagens iniciais com diversos *kernels* gaussianos. Por último, as diferenças de gaussianas (DoG) são obtidas através da subtração absoluta das escalas adjacentes de cada oitavo. No NLOOK, é utilizado o número máximo possível de oitavos, além do qual este seria menor que os *kernels*, e três escalas por oitavo.

Para a criação dos mapas de intensidade, a imagem original é convertida para uma imagem em tons de cinza I, e as diferenças de gaussianas são geradas para esta imagem utilizando *scale-spaces*. Em seguida cada uma destas diferenças de gaussianas são normalizadas pela subtração da média e divisão do resultado pelo desvio padrão. Diferentemente do NVT (Itti et al., 1998), os diferentes oitavos e escalas não são unidos em um único mapa, ou seja, todas as DoGs são preservadas. Para a criação dos mapas de cores, inicialmente são gerados quatro *scale-spaces* para os canais de cores R (vermelho), G (verde), B (azul) e Y (amarelo). Em seguida são geradas as diferenças de gaussianas entre os diferentes canais, ou seja, para os mapas RG as subtrações ocorrem entre os canais R e B, e para os mapas BY ocorrem entre s canais B e Y de cada oitavo/escala. Assim, são criados dois *scale-spaces* de oposição de cores: RG e BY.

Os mapas de orientação são criados de forma semelhante aos mapas de intensidade, porém antes da convolução com os *kernels* gaussianos a imagem inicial é convolucionada com filtros de Gabor (Daugman, 1988). Assim como no NVT, no NLOOK quatro orientações preferenciais $\theta \in \{0°, 45°, 90°, 135°\}$ são utilizadas, ou seja, são criados quatro *scale-spaces* de orientação. As DoGs destes quatro *scale-spaces* são então normalizadas e somadas, dando origem assim a um único *scale-space* de orientações.

Após a criação desses *scale-spaces* de características, eles são normalizados e somados em um único *scale-space* de saliências. As DoGs deste *scale-space* são então redimensionadas para a escala 0 (tamanho original da imagem) e somadas, formando assim um único mapa de saliências. Este mapa de saliências é então percorrido pelo foco de atenção da seguinte forma: inicialmente o ponto mais saliente do mapa é encontrado, e o *scale-space* de saliências é então analisado para se descobrir a escala característica deste FOA (Lindeberg, 1998), e o raio da região de interesse é calculado através da equação (Crowley et al., 2002):

$$r_{roi} = 2^{(o-1)} \times k_s \times b^{(s+\hat{s})} \quad (1)$$

onde o e s são o oitavo e a escala característica do FOA, e a constante $k_s = 1.6$ é um fator de correção empírico para a escala dado pela progressão geométrica com base $b = \sqrt{2}$. Um mecanismo de inibição de retorno, que possui o formato de uma gaussiana invertida, é então aplicado sobre o mapa de saliências único, sendo o diâmetro calculado pela Equação 1.

Através de diversos experimentos, descritos detalhadamente em (Heinen and Engel, 2008a; Heinen and Engel, 2009a; Heinen and Engel, 2009b), foi verificado que o NLOOK é capaz de selecionar as regiões mais relevantes do campo visual de forma bastante precisa, sendo bastante robusto a transformações afins. Além disso, o NLOOK consegue selecionar com bastante precisão as escalas dos FOAs, e isto o torna bastante útil em aplicações de visão computacional e robótica.

2.2 Codificação angular de cores

A sensibilidade às condições de iluminação é um problema que tem levado muitos pesquisadores a desenvolverem esquemas de representação de cores constantes, nos quais o casamento dos padrões não é afetado de forma significativa pelas condições de iluminação. A codificação angular de cores (Finlayson et al., 1996) é um esquema de representação bastante interessante, pois provê uma representação bastante compacta (apenas três valores numéricos), descritiva e robusta a mudanças de iluminação. Esta representação é baseada em estatísticas que codificam as características das distribuições de cores como ângulos entre os vetores dos canais de cores da imagem. Assim, a codificação angular de cores provê um mecanismo bastante eficiente e compacto para descrever as regiões de interesse da imagem, o que permite a categorização das mesmas de forma bastante eficiente, como será descrito a seguir.

2.3 INBC

O INBC (*Incremental Naïve Bayes Clustering*) (Engel, 2009) é um algoritmo baseado em técnicas de aprendizado não-supervisionado incremental para formação de conceitos a partir de instâncias do domínio descritas por atributos contínuos e discretos. O algoritmo INBC opera sucessivamente sobre cada dado, mantendo estimativas atualizadas dos modelos dos agrupamentos correntes. Usando o modelo corrente, o algoritmo decide se é necessário criar um novo agrupamento para o dado apresentado ao sistema. A formulação do algoritmo está baseada na hipótese da independência entre as variáveis que descrevem o domínio, equivalente à hipótese bayesiana ingênua.

Uma importante contribuição do INBC está na formulação de um procedimento incremental para a atualização dos parâmetros do modelo de mistura que representa o problema de aprendizado. A atualização dos parâmetros é vista como um processo de aproximação dos estimadores estatísticos levando em conta a hipótese da independência das variáveis. Um outro aspecto importante do INBC é a formação incremental de agrupamentos. A cada apresentação de um vetor de dados ao sistema, o algoritmo utiliza o modelo probabilístico corrente para decidir se o novo dado deve ser incorporado à configuração de agrupamentos atual, ou se este dado deve originar um novo agrupamento. A decisão é tomada em relação a um limiar de probabilidade mínima aceitável para que um vetor de dados seja considerado como pertencente a um dos componentes da mistura. Mais informações sobre o INBC podem ser encontradas em (Engel, 2009).

3 Experimentos realizados

Esta seção descreve os experimentos realizados para avaliar a performance do modelo proposto utilizando imagens reais de objetos. Para isto, foi utilizado o repositório de imagens SIVAL[1], que consiste de 1500 imagens de 1024x768 *pixels* que retratam 25 objetos diferentes (frutas, latas de refrigerante, livros, caixas, roupas, etc.) fotografados em 60 posições, condições de iluminação e cenários distintos (sobre um carpete, em cima de uma cadeira, ao ar livre, em frente à um quadro branco, etc.). Para a realização dos experimentos, o modelo proposto foi configurado para extrair os cinco focos de atenção (FOAs) mais relevantes de cada imagem. A Figura 3 mostra o espaço tridimensional da codificação angular de cores categorizado pelo INBC (cada uma das cores do gráfico da Figura 3 representa mais de uma categoria devido às limitações do software utilizado para a criação deste gráfico). No total, foram criadas 72 categorias, que abrangem entre 5 e 1321 fixações cada.

[1]SIVAL Repository – www.cs.wustl.edu/ sg/accio/SIVAL.html

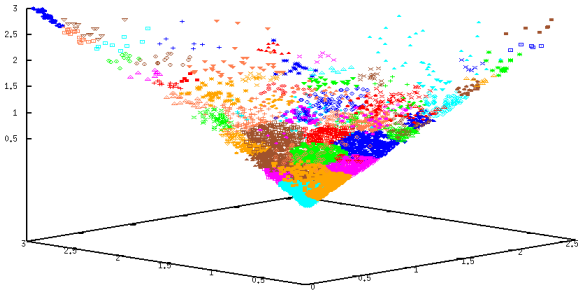


Figura 3: Categorização gerada pelo INBC

A Tabela 1 mostra as 15 categorias mais presentes, ou seja, aquelas que abrangem um maior número de FOAs. A primeira coluna descreve o número da categoria gerada pelo INBC; a segunda coluna traz o número total de fixações atribuídas à categoria; a terceira coluna descreve o objeto que teve mais fixações atribuídas a esta categoria; e a última coluna traz a quantidade de fixações atribuídas à categoria nas imagens deste objeto.

Tabela 1: Categorias mais presentes

Categoria	Total	Objeto principal	Qtde
04	1321	FeltFlower	113
28	1142	Scarf	134
13	467	WoodPot	60
70	382	Spoon	29
17	370	WoodPot	31
10	350	GoldMedal	46
03	336	Banana	57
21	323	SmileyDoll	42
35	292	GoldMedal	28
62	286	WoodPot	24
29	219	SmileyDoll	31
30	208	Notebook	74
16	116	Apple e Book	13
06	114	SmileyDoll	28
39	97	SmileyDoll	20

Percebe-se que duas das categorias (4 e 28) listadas na Tabela 1 foram atribuídas a mais de 15% dos 7500 FOAs extraídos (1500 imagens x 5 FOAs cada), mas isto se deve ao fato delas representarem elementos presentes na maioria das imagens: reflexão especular da luz (categoria 4) e objetos de mobília na cor bege/cinza (categoria 28). Assim, embora a terceira coluna da Tabela 1 traga a classe de objetos na qual a categoria possui mais fixações, esta categoria não necessariamente representa a categoria mais importante para a identificação destes objetos. As categorias mais representativas para cada objeto são apresentadas na Tabela 2. A primeira coluna descreve o objeto (foram mantidos os mesmos nomes em inglês do repositório de imagens SIVAL). As colunas 2, 4, e 6 apresentam as três categorias mais representativas de cada objeto, e as colunas 3, 5, e 7 representam a taxa de representatividade, calculada pela equação:

$$r_{co} = q_{co}/n_c \qquad (2)$$

onde r_{co} é a taxa de representatividade da categoria c para o objeto o, q_{co} é o número de ocorrências da categoria c nas imagens do objeto o, e n_c indica a quantidade de objetos nos quais a categoria c está presente. Assim, uma categoria presente nas imagens de vários objetos terá uma taxa de representatividade menor do que uma categoria presente em um único objeto.

Tabela 2: Categorias mais representativas

	Cat. A		Cat. B		Cat. C	
Objeto	C	R	C	R	C	R
Ajax	53	4.3	8	2.0	2	1.8
Apple	28	1.4	17	1.2	4	0.9
Banana	3	2.3	4	1.8	21	1.5
Scrunge	8	3.5	28	1.5	4	1.2
Candle	4	3.9	28	3.2	70	0.9
Cardboard	4	3.2	13	1.7	28	1.7
Scarf	28	5.4	4	2.6	70	1.1
CokeCan	28	1.8	34	1.5	10	1.4
Book	4	1.4	18	1.1	28	1.1
Shoe	4	3.6	28	3.0	70	1.0
Gloves	4	4.1	28	2.9	13	0.9
SoftBox	28	1.9	4	1.5	17	1.2
FeltFlower	4	4.5	13	1.6	28	1.5
WoodPot	13	2.4	4	2.0	17	1.2
GoldMedal	72	2.3	10	1.8	44	1.6
TeaBox	9	2.6	11	2.2	19	1.4
JuliesPot	52	2.5	71	2.2	12	1.5
Spoon	4	3.4	28	2.4	70	1.2
RapBook	4	2.6	28	1.8	13	1.0
SmileyDoll	21	1.7	6	1.4	4	1.3
SpriteCan	4	2.0	28	2.0	30	1.3
Notebook	30	3.5	4	2.2	28	1.5
Bowl	28	1.8	4	1.6	29	1.2
WD40Can	58	53.0	55	17.0	69	2.0
WoodPin	4	3.6	28	2.2	13	1.3

Observando os resultados da Tabela 2 percebe-se que, com exceção das categorias 58 e 55 do objeto *WD40Can*, as demais categorias possuem índices relativamente baixos. Isto se deve a forma como o modelo proposto atua, pois as fixações não representam necessariamente um objeto inteiro, mas sim partes do mesmo. Além disso, fixações de objetos diferentes podem possuir cores similares, o que fará com que a codificação angular de cores gere valores similares para ambos os objetos. Na Figura 4, por exemplo, as categorias 0, 2 e 3 representam partes distintas de um objeto de forma estável, ou seja, mesmo utilizando diferentes condições de iluminação, posição e escala as categorias são mantidas. Assim, embora individualmente estas categorias possam aparecer nas imagens de diversos objetos, a presença de todas elas em uma única imagem fortemente indica a presença deste objeto na imagem.

A Figura 5 apresenta outros objetos com as categorias selecionadas em destaque (parte das laterais das figuras foi removida somente para melhorar a visualização).

Figura 4: Categorias geradas para o objeto *Ajax*

Os círculos em cada imagem representam as fixações, os números próximos aos círculos representam as categorias, e as cores dos círculos foram calculadas através da seguinte fórmula:

$$\begin{aligned} R &= (\pi - \alpha_0) * (255/\pi) \\ G &= (\pi - \alpha_1) * (255/\pi) \\ B &= (\pi - \alpha_2) * (255/\pi) \end{aligned} \quad (3)$$

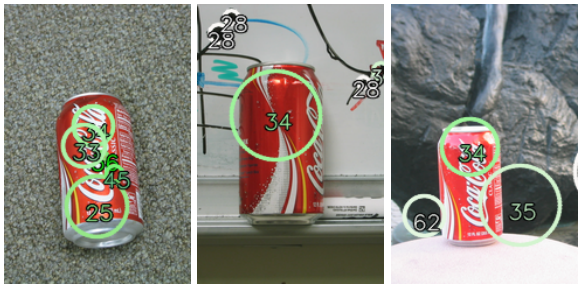
onde R, G e B são os valores dos canais RGB usados para colorir o círculo e $[\alpha_0, \alpha_1, \alpha_2]$ é um vetor contendo os valores da codificação angular de cores em radianos. Ou seja, círculos de cores próximas possuem codificações angulares similares, o que permite estimar a qualidade da categorização realizada pelo INBC de forma visual. Cabe ressaltar que as cores dos círculos não estão diretamente relacionadas com as cores dos objetos selecionados.

Percebe-se, na Figura 5, que o modelo proposto conseguiu criar categorias que permitem a identificação dos diversos objetos, apesar das diferentes condições de iluminação, posição e escala. Embora incipientes, estes resultados são bastante promissores, pois o modelo proposto conseguiu criar categorias relativamente representativas dado o número elevado de dados de treinamento (7500 vetores extraídos de 1500 imagens distintas) de forma não supervisionada.

4 Conclusões e perspectivas

Este artigo apresentou um modelo de visão computacional que tem por objetivo realizar a categorização de objetos de forma não supervisionada. Este modelo é composto de três elementos: (i) o modelo de atenção visual NLOOK; (ii) um esquema de codificação angular de cores; e (iii) o INBC, que consegue aprender as distribuições dos dados de entrada bem como categorizar os mesmos de forma não supervisionada. Através de diversos experimentos foi constatado que o modelo proposto consegue criar categorias que permitem a identificação de objetos independente das condições de iluminação, posição e escala. As perspectivas futuras incluem: (i) a utilização de informações *top-down* na elaboração dos mapas de saliências; e (ii) o uso de descritores de forma na identificação de objetos.

(a) Categorias geradas para o objeto *JuliesPot*



(b) Categorias geradas para o objeto *CokeCan*

(c) Categorias geradas para o objeto *WD40Can*

Figura 5: Exemplos de categorias geradas

Agradecimentos

Agradecemos ao apoio do CNPq que tornou possível a realização deste trabalho.

Referências

Crowley, J. L., Riff, O. and Piater, J. (2002). Fast computation of characteristic scale using a half octave pyramid, *Proc. Int. Workshop on Cognitive Vision (CogVis'2002)*, Zurich, Switzerland.

Daugman, J. G. (1988). Complete discrete 2-d gabor transforms by neural networs for image analysis and compression, *IEEE Trans. Acoustics, Speech, and Signal Processing* **36**(7): 1169–1179.

Desimone, R. and Duncan, J. (1995). Neural mechanisms of selective visual attention, *Annual Reviews Neuroscience* **18**: 193–222.

Draper, B. A. and Lionelle, A. (2005). Evaluation of selective attention under similarity transformations, *Computer Vision and Image Understanding* **100**: 152–171.

Engel, P. M. (2009). INBC: An incremental algorithm for dataflow segmantation based on a probabilistic approach, *Technical Report RP-360*, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil.

Finlayson, G. D., Chatterjee, S. S. and Funt, B. V. (1996). Color angular indexing, *Proc. 4th European Conf. in Computer Vision (ECCV'96)*, Springer-Verlag, Cambridge, UK, pp. 16–27.

Frintrop, S. (2006). *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search*, Ph.d. dissertation, Rheinische Friedrich-Wilhelms-Universitat Bonn, Bonn, Germany.

Heinen, M. R. and Engel, P. M. (2008a). Avaliação de modelos de atenção visual em relação a transformações afins, *Proc. IV Workshop de Visão Computacional (WVC)*, IEEE Press, Bauru, SP, Brazil.

Heinen, M. R. and Engel, P. M. (2008b). Visual selective attention model for robot vision, *Proc. 5th IEEE Latin American Robotics Symposium (LARS'08)*, IEEE Press, Salvador, BH, Brazil.

Heinen, M. R. and Engel, P. M. (2009a). Evaluation of visual attention models under 2d similarity transformations, *Proc. 24rd ACM Symposium on Applied Computing (SAC'09) – Special Track on Intelligent Robotic Systems*, ACM press, Honolulu, Hawaii.

Heinen, M. R. and Engel, P. M. (2009b). NLOOK: A computational attention model for robot vision, *Journal of the Brazilian Computer Society (JBCS)* p. 15. To appear.

Itti, L., Koch, C. and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis, *IEEE Trans. Pattern Analysis and Machine Intelligence* **20**(11): 1254–1259.

Koch, C. and Ullman, S. (1985). Shifts in selective visual attention: Toward the underlying neural circuitry, *Human Neurobiology* **4**(4): 219–227.

Lindeberg, T. (1998). Feature detection with automatic scale selection, *Int. Journal of Computer Vision* **30**(2): 79–116.

Niebur, E. and Koch, C. (1998). *Computational architectures for attention*, The Attentive Brain, MIT Press, Cambridge, MA, pp. 163–186.

Pashler, H. (1997). *The Psycology of Attention*, MIT Press, Cambridge, MA.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N. and Nuflo, F. (1995). Modeling visual attention via selective tuning, *Artificial Intelligence* **78**(1-2): 507–545.

Witkin, A. P. (1983). Scale-space filtering, *Proc. Int. Joint Conf. Artificial Intelligence*, Karlsruhe, Germany, pp. 1019–1022.