Aprendizado Multi-Instâncias e Redes Convolucionais aplicados a Recuperação de Imagens baseada em Conteúdo e Rascunhos

Leonardo S. F. Ribeiro, Tu Bui, John Collomosse, Moacir Ponti
ICMC / Universidade de São Paulo — São Carlos - SP, Brasil
CVSSP / University of Surrey — Guildford, United Kingdom

***Resumo*—A recuperação de imagens baseada em conteúdo (CBIR) permite a busca e indexação de imagens com base diretamente no seu conteúdo visual. A recuperação de imagens baseada em rascunhos (SBIR) é uma variação que utiliza como consulta um rascunho de um objeto, o que adiciona diversos desafios. Nesses cenários propomos duas abordagens: o aprendizado multi-instâncias ao representar imagens para CBIR e redes convolucionais com função de custo triplet na obtenção de descritores para SBIR, em particular versões compactas desses descritores. Os resultados apontam melhorias com relação aos métodos base e do estado da arte, demonstrando o potencial das abordagens propostas.**

***Abstract*—Content based image rerieval (CBIR) allows for querying and indexing of image datasets based on the visual content of those images. Sketch based image retrieval (SBIR) is a variation that uses object sketches as queries, adding complexity to the retrieval challenge. To solve this problem we came up with two proposed approaches: Multiple Instance Learning (MIL) methods to represent images for CBIR and Convolutional Neural Networks with a triplet loss function to generate descriptors for SBIR, particularly, compact versions of such descriptors. The results show improvements when compared with base methods and the state of the art, demonstrating the potential of the shown approaches.**

***Keywords*-CBIR, SBIR, Deep Learning, Aprendizado multi-instâncias**

I. Introdução

A Internet é uma importante rede de distribuição de conteúdo visual. Em 2018 espera-se que 80% do tráfego será composto de conteúdo deste tipo, dois terços dos quais em dispositivos móveis [1]. Este tipo de demanda desperta a necessidade do desenvolvimento de métodos para categorização e recuperação de conteúdo visual. Nesse contexto, apresentamos nesse artigo o aprendizado multi instância (*Multiple Instance Learning* – MIL) para classificação de imagens naturais, e as redes neurais convolucionais para recuperação de imagens baseada em rascunhos (*Sketch Based Image Retrieval* – SBIR), com vistas à melhoria de acurácia e ao mesmo tempo mantendo viável a escalabilidade para dispositivos móveis.

Aprendizado Multi Instância: a formulação tradicional de aprendizado supervisionado determina que um objeto é representado por uma instância, associada a um rótulo de classe [2]. Formalmente, sendo $\mathcal{X}$ o espaço de instâncias (ou espaço de características) e $\mathcal{Y}$ o conjunto de rótulos de classe, a tarefa é aprender uma função $f : \mathcal{X} \rightarrow \mathcal{Y}$ a partir de um determinado conjunto de dados $\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$, onde $x_i \in \mathcal{X}$ e $y_i \in \mathcal{Y}$ ao rótulo conhecido de x_i.

Existem problemas do mundo real que não se encaixam bem nesse modelo, um desses casos é quando um objeto é representado por múltiplas instâncias. Por exemplo, uma imagem geralmente contém várias regiões de interesse e cada uma pode ser uma instância e o conjunto de essas instâncias representa a imagem [3].

A abordagem *Multiple Instance Learning* (MIL) [4] descreve problemas em que um objeto é descrito por um conjunto de instâncias (bag), este associado a um rótulo de classe. A tarefa em MIL é aprender um conceito a partir de dados que consistem de $bags$ de instâncias. Cada bag é rotulado como positivo ou negativo (em problemas de duas classes), e cada um descrito como um conjunto de vetores. Um bag é positivo se pelo menos um dos vetores em seu conjunto se encontra dentro do conceito pretendido e negativo se nenhum dos vetores reside no conceito.

Cada instância que compõe o *bag* pode possuir ou não rótulo conhecido durante o treinamento, no entanto na fase de classificação deve-se classificar o *bag* como um todo e não os exemplos individualmente. Um exemplo de aplicação do MIL para a classificação de imagens de câncer de pulmão é apresentada por [5].

Formalmente, a tarefa do MIL é aprender uma função [2]: $f_{MIL} : 2^{\mathcal{X}} \rightarrow \{-1, +1\}$ a partir de um determinado conjunto de dados $\{(X_1, y_1), (X_2, y_2), \cdots, (X_m, y_m)\}$, em que $X_i \subseteq \mathcal{X}$ é um conjunto de instâncias $\{x_1^{(i)}, x_2^{(i)}, \cdots, x_{n_i}^{(i)}\}, x_j^{(i)} \in \mathcal{X}(j = 1, 2, \cdots, n_i), y_i \in \{-1, +1\}$ é o rótulo de X_i [1].

Recuperação de Imagens Baseada em Rascunhos: bases de dados de imagens e vídeos incluem comumente anotações manuais indicando a presença de conceitos ou objetos. Essas anotações permitem codificar de forma concisa a semântica dos dados, mas são menos eficientes para representar aparência visual. Assim, a recuperação de imagens baseada em conteúdo permite utilizar dados visuais diretamente no processo de recuperação, sem a necessidade de metadados. A área de recuperação de imagens baseada em conteúdo é importante em diversos cenários. No entanto, uma área pouco explorada é a recuperação de imagens baseada em rascunhos ou esboços

[1]De acordo com a sintaxe comumente utilizada no aprendizado multi-instância, (xi, yi) é um $bag(saco)$ rotulado enquanto X_i um bag sem rótulo.

(*sketch based image retrieval* – SBIR).

Um rascunho é uma representação intuitiva da aparência do objeto, o que é interessante para uso em dispositivos móveis dada a facilidade para produção de rascunhos nestes aparelhos, além da conveniência de interações curtas com o usuário. Dado o apelo para o uso em dispositivos modos, é fundamental em SBIR otimizar o uso de processamento e memória.

Uma das abordagens possíveis é o uso de *bag of visual words* em conjunto com descritores visuais, dos quais os descritores HOG (*Histogram of Gradients*) mostraram bons resultados em conjunto com métodos mais complexos para descrever os rascunhos [6]. Mais recentemente, o advento das redes neurais convolucionais modernas também atingiu a área de aprendizado com rascunhos.

As redes neurais convolucionais (CNNs) são técnicas que podem ser vistas como um *framework* unificado para o aprendizado de extratores de características (descritores) e classificadores de grande desempenho quando comparados aos métodos baseados em desenvolvidos manualmente [7]. Yu et al. [7] propôs a CNN *Sketch-A-Net*, que superou Eitz et al. [8] na classificação da base de dados Tu-Berlin (maior base de rascunhos categorizados disponível atualmente).

A. Contribuições

Apresentamos nesse artigo contribuições em duas frentes: primeiramente, a melhoria da recuperação de imagens baseada em conteúdo usando abordagens multi-instância, e em segundo o uso de CNNs para a extração de características na recuperação de imagens baseada em rascunhos. Com relação à segunda contribuição, reportamos uma extensão dos resultados de Bui et al. [9], que adaptou a rede em uma arquitetura siamesa utilizando função de perda *triplet loss* pra uso em SBIRs capaz de generalizar a recuperação entre diferentes categorias de objetos. Em especial, a contribuição está na compactação do descritor por meio de quantização e análise de componentes principais.

II. Aprendizado Multi Instância para Classificação de Imagens

Os estudos realizados nesta área podem ser divididos em passos: (1) Investigação de métodos para segmentação, descritores de imagem baseados no espaço de cores Luv e coeficientes Wavelet e o algoritmo de agrupamento k-Médias, (2) análise dos resultados da segmentação, visual e estatisticamente, com o apoio do software Weka e (3) experimentos com classificação de imagens e rascunhos, utilizando as bases de dados: Flickr15K [6], composta de imagens e rascunhos; e Corel1000 [10], cujos rascunhos foram criados por voluntários durante o período de pesquisa.

Segmentação: Inicialmente, cada imagem é dividida em blocos de 4×4 pixels e um vetor de características é extraído para cada bloco. Obtidos os vetores de características, o algoritmo de agrupamento *k-medias*, já implementado na biblioteca, é aplicado. No entanto ressaltamos que o algoritmo não define o número de grupos (k) em que será feita a divisão; Para esta tarefa foi empregado um método semelhante a [10], incrementando k em passos restringindo-o a thresholds baseados no desempenho da segmentação.

Extração de Características: Foram implementados métodos de extração de três grupos de características visuais de forma que estes pudessem ser utilizados (compondo vetores de características) para segmentação e/ou recuperação das imagens a partir de rascunhos (SBIR). O primeiro grupo, composto de três características, é a média dos componentes de cor dos blocos 4×4 utilizando o espaço de cores LUV, onde L representa a luminância, U e V representam as informações de cor (crominância).

O segundo grupo de características também é composto de três valores que representam a energia nas bandas de alta frequência de tranformadas *Wavelet* [11]; A motivação para o uso dos coeficientes como descritores visuais vem da propriedade dos mesmos em representar de forma distintiva características de textura, a intuição por traz desta ideia vem do fato de que diferentes banda de frequência representam variação em diferentes direções; a banda HL, por exemplo, representa atividades na direção horizontal.

O terceiro grupo de características são parte do *HOG* (Histograma de Gradientes Orientados), um descritor conhecido por sua eficácia para a identificação de formas humanas em imagens [12] e, de especial interesse a nossa aplicação final, na classificação de imagens digitais, rascunhos e recuperação baseada em rascunhos [6].

Com estes descritores implementados, os vetores de características, quando utilizando de todos os grupos, possuem o formato apresentado pela Tabela I

Cor			Wavelets			HoG
L	U	V	HL	LH	HH	8 bins (ângulos)

Tabela I
Formato do vetor de características

A. Resultados

Tabela II
Coeficientes Kappa

Base	Segment.	Classific.	KNN	B. C-KNN
corel1000	L e W	L, W e H	0.42 ± 0.05	0.47 ± 0.04
corel1000	L e W	L e W	0.35 ± 0.04	0.43 ± 0.04
corel1000	L	L, W e H	0.44 ± 0.04	0.51 ± 0.04
corel1000	L	L	0.31 ± 0.05	0.34 ± 0.05
flickr15K	L e W	L, W e H	0.22 ± 0.03	0.24 ± 0.04
flickr15K	L e W	L e W	0.17 ± 0.03	0.20 ± 0.03
flickr15K	L	L, W e H	0.21 ± 0.03	0.26 ± 0.04
flickr15K	L	L	0.11 ± 0.03	0.12 ± 0.03

Foram feitos testes usando os classificadores *KNN* (k-Nearest Neighbors) e *Bayesian Citation KNN* (uma versão modificada do mesmo algoritmo com vistas a aplicação em cenários multi-instância) em diferentes configurações de segmentação e quantidade de características utilizadas. Os resultados da *cross-validation*, usada para comparar os descritores, podem ser contemplados na Tabela II com a ajuda do coeficiente Kappa. Note que, na tabela, representamos

os descritores com L para LUV, W para Wavelet e H para HOG. É possível notar que, em quase todas as configurações apresentadas, a abordagem multi-instância (B.C. KNN) supera o classificador comum, demonstrando a vantagem do método de extração.

Testes semelhantes foram realizados para a classificação de rascunhos, mas os resultados não foram satisfatórios; a classificação de rascunhos apresentou resultados muito próximos àqueles que seriam obtidos caso um classificador aleatório fosse utilizado (ou seja, $kappa \approx 0$). Para o tratamento de rascunhos, um novo problema (Recuperação de Imagens Baseada em Rascunhos) e uma nova abordagem (Redes Neurais Convolucionais) foi utilizada, detalhada na seção III.

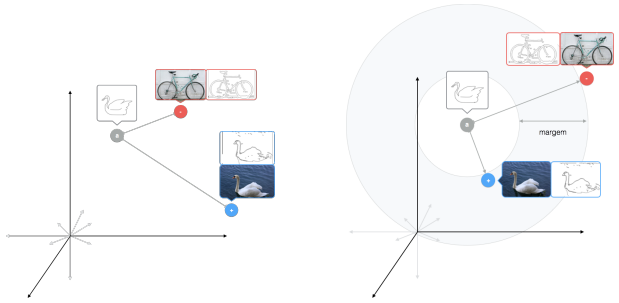
III. Redes Convolucionais para Recuperação de Imagens

A proposta é tratar a recuperação de imagens como um problema de regressão, onde um descritor multidimensional e aprendido através de uma rede de arquitetura siamesa tripla (três redes neurais unidas) com pesos parcialmente compartilhados. Utilizamos a CNN para aprender o mapeamento inter domínio entre rascunhos e mapas de bordas obtidos de imagens naturais [13] e exploramos representações eficientes dos descritores produzidos. Após o treinamento da rede realizamos experimentos com a compactação do descritor usando quantização e projeções PCA (*Principal Component Analysis*), visto que a descritores compactos são imporantes em SBIR para o uso eficiente em dispositivos restritos como celulares de tablets.

Rede Convolucional Siamesa: Redes convolucionais clássicas (sem ramificações) trouxeram avanços revolucionários para a detecção de objetos e categorização de imagens. Esses sucessos foram refletidos nas áreas relacionadas a rascunhos, como no caso da arquitetura Sketch-a-Net que atingiu performance de classificação semelhante a de humanos. Pode-se acreditar, de forma ingênua, que descritores extraídos diretamente destes classificadores seriam suficientes para uso em recuperação por conteúdo, o que não é verdade já que estes descritores não apresentam boa generalização para buscas fora do conjunto de treinamento. Utilizamos então uma rede siamesa que resringe apenas a distância relativa entre um objeto âncora e exemplos positivos e negativos (Figura 1). Em nosso caso um triplet é composto de um rascunho âncora e exemplos positivos e negativos escolhidos entre os mapas de borda gerados a partir das imagens naturais.

Para os experimentos foram utilizados duas bases de dados públicas de rascunhos: a *Tu-berlin*, desenvolvida para classificação [14] e a *Flickr15K*, de *SBIR* [6]:

- A base de dados de classificação Tu-Berlin, utilizada apenas na fase de **treino**, é composta de 250 categorias de rascunhos com 80 destes por categoria. Além dos rascunhos, o método requere imagens naturais para o treino da rede siamesa; essas imagens foram obtidas através de *APIs* públicas dos serviços de recuperação de imagens *Google, Bing e Flickr*.



(a) Antes do treinamento (b) Após treinamento, com margem

Figura 1. A função de perda *Triplet loss* minimiza a distância entre um âncora e um exemplo positivo (mesma categoria) e maximiza a distância entre um âncora e um exemplo negativo (qualquer outra categoria).

- A base Flickr15K, usada para **teste**, é composta de 33 categorias de rascunhos, com 20 rascunhos por categoria, além de diferentes quantidades de imagens naturais por categoria, totalizando 15024 imagens deste tipo.

Compactação: Em níveis de aplicação é de suma importância o desenvolvimento de uma representação compacta dos dados de tarefas de classificação e recuperação, permitindo transmissão e armazenamento eficientes, fatores relevantes para uso em dispositivos móveis [15], [16].

Investigamos um método de compressão por projeção baseada em PCA como em [17], seguido da quantização de cada dimensão em inteiros de n-bits. Especificamente, para a representação dos dados para SBIR, a combinação de projeções baseadas em PCA e quantização independente dos elementos do descritor promovem excelente redução de dimensionalidade sem perdas de performance. Também fizemos experimentos com técnicas de binarização via PCA ou LSH, mas estas causaram degradação na performance.

A. Resultados

Com um treinamento utilizando as 250 categorias disponíveis para treino e as 100 dimensões dos descritores produzidos sem modificações, a performance da recuperação tem *mean average precision* (MAP) de 22.04% conforme resultados em [9]. Com nosso método de compactação do descritor, conforme Tabela III a projeção baseada em PCA diminui 90% do armazenamento em sua versão mais compacta sem grandes perdas de performance, com uma projeção de 8 a 12 dimensões.

Tabela III
Uso de diferentes números de componentes principais (PCs) em busca da diminuição da dimensionalidade do descritor

#PC	2	4	6	8	10	12	14
mAP	11.60%	17.88%	19.82%	22.19%	22.07%	22.32%	22.06%

Reduzimos ainda mais o armazenamento necessário aplicando quantização, de pontos flutuantes (32-64 bits) para inteiros com n bits. O resultado dos experimentos com quantização podem ser contemplados na Figura 2; observamos

que quantizações com menos de $n = 8$ bits são inadequadas, sugerindo que quantização para 1 byte-por-dimensão seja a solução ótima. Nessa configuração a performance flutua ao redor de $\sim$ 1%, mas a redução no armazenamento da representação é de 98%, de 3200 para 64 bits (8 bits por dimensão, 8 dimensões); cada imagem na base de dados pode ser representada por um único descritor de 64 bits, fazendo com que a base Flickr15K, em sua totalidade, ocupe apenas ≈120 kilobytes de armazenamento.

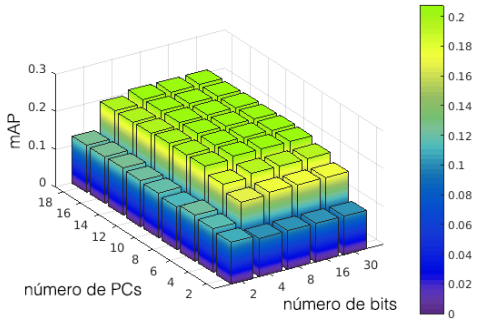


Figura 2. Comparação de diferentes configurações de quantização, destacando que nossa escolha de 8 componentes principais quantizados para 8 bits é a opção mais compacta sem grandes perdas na performance

Além da redução no armazenamento, as representações compactas (em números inteiros) são mais rapidamente comparadas do que valores reais. A Tabela IV mostra o tempo para realização de uma busca em algumas configurações de compatibilização. Enquanto a busca com o descritor original toma **29.7976ms**, recuperar imagens usando o descritor quantizado com nossa configuração de escolha reduz o tempo de processamento em aproximadamente **41%**.

Tabela IV
TEMPO DE UMA BUSCA NA NA BASE DE DADOS FLICKR15K

PCs \nBits	4	8	16
4	14.5762ms	17.3261ms	17.9398ms
8	14.3429ms	17.4888ms	17.9492ms
12	14.0348ms	17.7112ms	18.7156ms
16	15.9036ms	17.8845ms	19.0868ms

IV. CONCLUSÃO

Descrevemos uma série de métodos para melhoria dos resultados de recuperação de imagens. Primeiramente, para obter melhoria na recuperação de imagens naturais baseada em conteúdo, fizemos uso do aprendizado multi-instância. Esse no entanto, não foi suficiente para lidar com a recuperação baseada em rascunhos, necessitando de técnicas de aprendizado profundo, em particular redes convolucionais com função de custo triplet.

A abordagem multi-instância permite desacoplar regiões da imagem, e considerar cada imagem como um conjunto de suas regiões ao invés de globalmente, o que ajuda a representar imagens com poluição visual, ou objetos provenientes de outras classes que não a de interesse. As redes convolucionais com funções de custo triplet, por sua vez, permitem relacionar domínios diferentes (no nosso caso rascunhos e imagens), e produzem descritores representativos que, conforme demonstramos podem ser compactados para maior eficiência.

Trabalhos futuros podem combinar as duas abordagens, permitindo por exemplo lidar com poluição visual nos cenários de SBIR e CBIR, além de explorar outras abordagens de aprendizado profundo como as redes geradoras adversariais.

ACKNOWLEDGMENT

Os autores agradecem ao patrocínio da FAPESP (processos #2014/14557-0, #2015/26050-0 e #15/13504-2).

REFERÊNCIAS

[1] Cicso, *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019 White Paper*, 2014.

[2] Z.-h. Zhou and M.-l. Zhang, "Multi-instance multi-label learning with application to scene classification," in *Advances in Neural Information Processing Systems 19*, 2007.

[3] Y. Wang, C. Zhang, and Z. Wang, "Rate distortion multiple instance learning for image classification," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, Sept 2013, pp. 3235–3238.

[4] Y. Xiao, B. Liu, Z. Hao, and L. Cao, "A similarity-based classification framework for multiple-instance learning," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–1, 2013.

[5] L. Zhu, B. Zhao, and Y. Gao, "Multi-class multi-instance learning for lung cancer image classification based on bag feature selection," in *Int. Conf. on Fuzzy Systems and Knowledge Discovery*, vol. 2, 2008, pp. 487–492.

[6] R. Hu and J. Collomosse, "A performance evaluation of gradient field HOG descriptor for sketch based image retrieval," *Computer Vision and Image Understanding*, vol. 117, no. 7, pp. 790–806, 2013.

[7] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Sketch -a-net that beats humans," in *Proc. BMVC*. IEEE, 2015.

[8] M. Eitz, R. Richter, T. Boubekeur, K. Hildebrand, and M. Alexa, "Sketch-based shape retrieval," *ACM Trans. Graph. (Proc. SIGGRAPH)*, vol. 31, no. 4, pp. 31:1–31:10, 2012.

[9] T. Bui, L. Ribeiro, M. Ponti, and J. Collomosse, "Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network," *(paper submitted)*, 2016.

[10] D. C. J. Wang, J. Li and G. Wiederhold, "Simplicity: semantics-sensitive integrated matching for picture libraries," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, pp. 947 – 963, 2001.

[11] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1992.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[13] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 898–916, 2011.

[14] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors," *IEEE Trans. Visualization and Computer Graphics*, vol. 17, no. 11, pp. 1624–1636, 2011.

[15] M. Ponti and L. Escobar, "Compact color features with bitwise quantization and reduced resolution for mobile processing," in *IEEE Global Conference on Signal and Information Processing*, 2013, pp. 751–754.

[16] M. Ponti, T. S. Nazaré, and G. S. Thumé, "Image quantization as a dimensionality reduction procedure in color and texture feature extraction," *Neurocomputing*, vol. 173, pp. 385–396, 2016.

[17] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in Neural Information Processing Systems*, 2014, pp. 1988–1996.