Making an image worth a thousand visual words

LAVI

EESC-USP

Laboratório de Visão Computacional

USP

XI Workshop de Visão Computacional
WVC´2015

October 05th – 07th, 2015

São Carlos – SP - Brazil

Proceedings

University of São Paulo
São Carlos School of Engineering

Making an Image Worth a Thousand Visual Words

Glauco Vitor Pedrosa, Agma Juci Machado Traina
Instituto de Ciências Matemáticas e de Computação - ICMC
Universidade de São Paulo - USP
{gpedrosa,agma}@icmc.usp.br

***Abstract*—The automatic dissimilarity analysis between images depends heavily on the use of descriptors to characterize the images' content in compact and discriminative features. This work investigates the use of visual dictionaries to represent and retrieve the local image features using the popular Bag-of-Visual-Words modeling approach. We evaluated the impact of different parameters in the construction of this modeling approach, showing that an image can be effectively described using less than a thousand words.**

I. Introdução

Assim como diz o antigo provérbio "Uma imagem vale mais que mil palavras...", na área da Computação pode-se afirmar que uma imagem possui uma variedade enorme de dados a serem interpretados tanto quantitativa quanto qualitativamente. A representação das características visuais de uma imagem é uma tarefa importante no processo de análise, classificação e recuperação. A extração dos atributos visuais sintetiza a imagem através de um *vetor-de-características*, que passa a representa-la nos processos de mineração e recuperação por conteúdo. Esse vetor-de-características deve ser combinado com uma *função de distância*, cujo objetivo é calcular o grau de similaridade entre dois vetores, retornando um valor que quantifica o quão dissimilar eles são. Desse modo, um descritor é um par (vetor-de-características e função de distância) utilizado para representar e recuperar imagens .

A questão é que o desenvolvimento de um descritor é uma tarefa muito desafiadora, pois existem muitos fatores que devem se considerados na descrição de uma imagem, tais como: resolução, as variações de iluminação, objetos oclusos, etc. Descritores podem ser *globais* ou *locais*, dependendo de como analisam o conteúdo visual da imagem. Os descritores globais são mais populares na literatura, mas são conhecidos por serem limitados, pois descrevem uma imagem de uma forma holística. Já descritores locais, como o SIFT (*Scale-Invariant Feature Transform*) [1], fornece uma maneira de descrever várias regiões ao redor de pontos-de-interesse dentro das imagens, o que demonstra grande poder descritivo, uma vez que é possível caracterizar apenas regiões específicas da imagem que sejam de interesse do usuário, excluindo informações irrelevantes/desnecessárias.

Analisar a similaridade de imagens através apenas de suas características locais não é uma tarefa fácil e computacionalmente barata. Descritores locais exigem mais memória e funções de distâncias "especiais" para medir a similaridade na tarefa de recuperação, já que a cardinalidade das características locais podem variar de imagem para imagem. Pesquisas tem sido realizadas para contornar tal problema e permitir sumarizar os descritores locais em apenas um único vetor-de-características e, assim, reduzir o custo computacional da análise da similaridade, tal como a abordagem *Bag-of-Visual-Words* que é baseada na descrição da imagem em *palavras-visuais*.

Neste artigo é apresentada a modelagem *Bag-of-Visual-Words* (BoVW) com definições, discussões e avaliações experimentais, mostrando as técnicas adotadas na literatura e os desafios computacionais do modelo. Aqui será abordada em detalhes a análise da influência de quatro fatores necessários na implementação da abordagem BoVW:

- o tamanho do dicionário visual;
- o detector dos pontos-de-interesse;
- o tipo de codificação;
- a função de distância usada para medir a similaridade.

A hipótese considerada é que esses quatro fatores não são generalizáveis para todos os tipos de imagens, influenciando diretamente na qualidade da caracterização e, por conseguinte, na precisão da recuperação. Para isso, foram realizados experimentos para investigar o efeito da variação conjunta desses fatores na precisão da recuperação de imagens por conteúdo usando diferentes bases de imagens. O objetivo do trabalho é mostrar que, escolhendo uma combinação ideal de detector e função de similaridade, é possível descrever, de uma maneira eficiente, as características locais de uma imagem usando um dicionário-visual pequeno, com menos de mil palavras.

Na Seção 2 será apresentada toda a metodologia para descrever uma imagem em palavras-visuais; a Seção 3 mostra as avaliações experimentais realizadas usando diferentes bases de dados, e por fim, na Seção 4 será apresentada as considerações finais.

II. A Abordagem *Bag-of-Visual-Words* (BoVW)

Nas últimas décadas a abordagem *Bag-of-Visual-Words* (BoVW), também denominada de *Bag-of-Features*, *Bag-of-Visual-Features* ou *Bag-of-Keypoints*, se tornou uma abordagem bastante popular em várias áreas de visão computacional, como classificação de imagens, busca de vídeos, reconhecimento de textura, entre outras tarefas [2], [3], [4], [5], [6], [7], [8]. Parte da sua popularidade é devido à sua simplicidade: a metodologia BoVW é baseada na representação não-ordenada de descritores locais aplicados em uma imagem e são, portanto, conceitualmente e computacionalmente mais simples do que muitos métodos alternativos. Apesar disso, ou talvez por causa disso, sistemas de recuperação que utilizam a abordagem BoVW tem apresentado um desempenho superior em vários *benchmarks* e têm conseguido avanços significativos em escalabilidade.

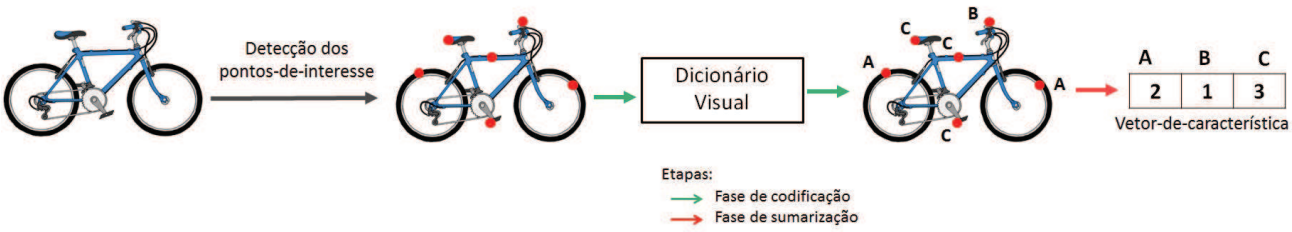


Fig. 1. Metodologia clássica para representar uma imagem em *Bag-of-Visual-Words*.

A. *Metodologia*

O processo necessário para representar uma imagem em BoVW está esquematizado pela Figura 1 e pode ser definido pelas seguintes fases: (i) detecção e representação das regiões-de-interesse; (ii) atribuição de cada região-de-interesse à uma palavra-visual de acordo com um dicionário (vocabulário) de palavras-visuais pré-definido; (iii) contagem da ocorrência (frequência) de cada palavra-visual contida na imagem.

Em suma, a modelagem pode ser encapsulada em duas etapas: codificação (*coding*) e sumarização (*pooling*). A etapa de codificação discretiza as características locais em palavras-visuais e a etapa de sumarização utiliza alguma abordagem que sintetiza as palavras-visuais da imagem em um único vetor-de-característica que será usado posteriormente para classificação e/ou recuperação de imagens. A representação BoVW é flexível, uma vez que cada uma de suas fases podem ser determinadas por diferentes técnicas de acordo com o domínio da aplicação [9], tal como será avaliada na seção de resultados experimentais desse trabalho.

Uma questão fundamental da abordagem BoVW se refere à decisão de qual detector de pontos-de-interesse usar, ou se deve ou não usar um detector de pontos-de-interesse, tal como avaliado por [10]. Geralmente os métodos de detecção de pontos-de-interesse mais utilizados na literatura são Harris-Affine ou Diferenças-de-Gaussianas, que é o detector utilizado pelo descritor SIFT. Porém, diversos trabalhos tem utilizado técnicas mais simples, como a detecção de pontos-de-interesse densos e aleatórios.

B. *Construção do Dicionário Visual*

A determinação do dicionário de palavras-visuais é uma tarefa fundamental para a abordagem BoVW, pois ele é o responsável por determinar quais são as características e padrões que representam a estrutura de uma imagem. Apesar da diversidade de aplicações, quase todos os trabalhos na literatura apresentam a mesma estratégia para a geração do dicionário de palavras-visuais. Essa estratégia está esquematizada na Figura 2 e é definida pelos seguintes passos: primeiramente um subconjunto de imagens do banco de dados é escolhido; para cada imagem, suas regiões-de-interesse são detectadas e descritas utilizando algum descritor gerando vetores-de-características; por fim, é realizado um agrupamento dos dados desse espaço de características utilizando algum algoritmo de agrupamento. O centróide de cada grupo é considerado uma palavra-visual do dicionário.

O desafio na construção de um Dicionário Visual depende de duas escolhas: o algoritmo de agrupamento e a quantidade de palavras-visuais. O algoritmo de agrupamento é um fator que interfere diretamente na qualidade e no desempenho da geração de um dicionário visual [11], [12]. Além disso, a *quantidade* de palavras-visuais do dicionário é outro fator que influencia diretamente na qualidade da representação da imagem. Um dicionário pequeno tem pouco poder discriminativo, uma vez que dois agrupamentos podem ser atribuídos à mesma palavra visual. Por outro lado, um vocabulário grande é pouco generalista e afeta a eficiência da abordagem. Na literatura, a quantidade de palavras-visuais é uma informação definida empiricamente e que pode variar entre diferentes base de dados.

O agrupamento pode ser realizado, por exemplo, através do algoritmo *k-means*, que é um método de agrupamento simples e bastante utilizado na literatura. O *k-means* particiona um conjunto de pontos entre k subconjuntos disjuntos visando a minimizar a distância intra-grupos e maximizar a distância inter-grupos. Os trabalhos na literatura tendem a escolher empiricamente a quantidade de palavras visuais do dicionário representadas pelo *k-clusters* obtidos através da execução do *k-means*. Entretanto, [13] observaram que o uso de *k-means* para a geração das palavras visuais funciona bem em imagem com texturas homogêneas, mas para imagens com objetos naturais e com regiões densas esse tipo de agrupamento não gera bons resultados para a criação de um dicionário visual. Além disso, [14] indicam que é possível substituir a abordagem de agrupamento usando o algoritmo *k-means* por uma simples seleção aleatória, sem perda estatisticamente significativa de informações, mas com uma redução radical no custo computacional. O uso de uma técnica variante, conhecida como *bisecting k-means* [11], tem demonstrado desempenho superior à técnica original no domínio da recuperação de informações textuais, mas pouco explorada na geração de palavras-visuais no domínio da recuperação de imagens.

Vários algoritmos de agrupamentos estão disponíveis na literatura, porém um problema compartilhado entre eles é a necessidade de se fornecer a quantidade de grupos *a priori*. Geralmente, esse é um valor definido empiricamente pelo usuário. Um fator importante, que interfere diretamente no desempenho da abordagem *Bag-of-Visual-Words*, é o tamanho do dicionário de palavras. Um vocabulário pequeno tem pouco poder discriminativo, uma vez que dois agrupamentos podem ser atribuídos à mesma palavra visual. Por outro lado, um vocabulário grande é pouco generalista e afeta a eficiência da abordagem. Por exemplo, [5] utiliza um vocabulário formado por 200-400 palavras visuais; [15] adota 1.000 palavras e [2] um vocabulário com 6.000-10.000 palavras.

Existem alguns métodos que permitem determinar

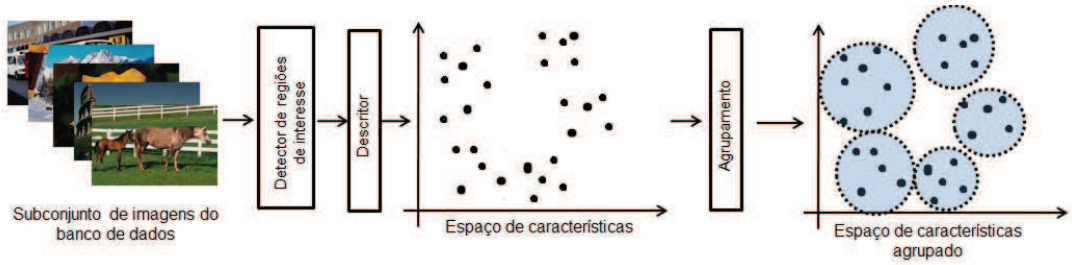


Fig. 2. Processo utilizado para a geração de um dicionário de palavras-visuais: dado um subconjunto de imagens de treinamento, suas características são extraídas e descritas. As palavras-visuais que irão compor o vocabulário do dicionário são definidas como sendo os centróides dos grupos a partir de um agrupamento realizado nesse espaço de características.

automaticamente a quantidade de grupos. Por exemplo, o algoritmo X-means, proposto por [16] utiliza um critério de agrupamento particional utilizando uma medida conhecida como BIC (*Bayesian Information Criterion*). Entretanto, essa técnica necessita saber de antemão sobre a densidade e compacidade dos grupos, pois o algoritmo assume que todos os grupos são esféricos e com o mesmo tamanho. Para uma classe específica de algoritmos de agrupamento, em particular os algoritmos de agrupamento hierárquicos, não é necessário especificar a quantidade de grupos *a priori* [17], [18]. O resultado de um algoritmo de agrupamento hierárquico pode ser graficamente visualizado como uma árvore, chamada de *dendrograma*, que mostra o processo de fusão (*merge*) dos elementos e os *clusters* intermediários. Cortando um dendrograma em um certo nível, pode-se obter um conjunto de agrupamentos (*clusters*).

C. Fase de Codificação

Uma vez definido o dicionário de palavras-visuais, cada região-de-interesse da imagem detectada e representada por um vetor-de-característica será associada à uma palavra-visual. Em inglês, essa fase é denominada de *assignment* ou *coding*. A Figura 3 ajuda a ilustrar a fase de *assignment* da abordagem *Bag-of-Visual-Words*: dado que uma palavra-visual é o resultado de um agrupamento do espaço de característica, tal como discutido anteriormente, qual palavra-visual será atribuída ao triângulo amarelo?

Na literatura, a fase de codificação pode ser realizada por três diferentes abordagens:

1) *hard assignment* ou também chamada de *nearest neighbor*;
2) *multiple assignment* ou k-*nearest neighbor*;
3) *soft assignment*.

A abordagem *hard* é a mais tradicional na literatura, em que a palavra-visual que possui a menor distância ao vetor-de-característica é determinada, e adiciona-se uma unidade ao *bin* correspondente à essa palavra-visual no histograma. No exemplo da Figura 3, o triângulo amarelo seria atribuído à palavra-visual w_4, que é a mais próxima.

Muitas vezes, pode ocorrer de um dado vetor-de-característica ter uma distância bem próxima entre duas palavras. No exemplo da Figura 3 o triângulo amarelo está bem próximo das palavras w_4 e w_5, porém na abordagem *hard* somente a palavra-visual mais próxima w_4

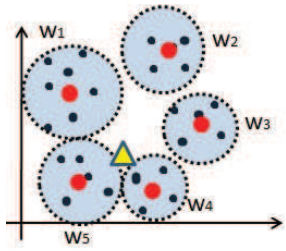


Fig. 3. Qual palavra-visual será atribuída ao triângulo amarelo?

seria selecionada para representar essa característica. Assim, aquelas características perto dos limites de Voronoi não são bem representadas. Para tentar solucionar este problema, os pesquisadores têm explorado as abordagens *multiple assignment* e *soft assignment*. Essas duas estratégias foram concebidas para mitigar o impacto negativo quando um grande número de vetores-de-características de uma imagem estão perto de uma fronteira de Voronoi de dois ou mais agrupamentos.

A abordagem *multiple assignment* [19], atribui ao vetor-de-características todas as k palavras-visuais mais próximas. Nesse caso, adiciona-se uma unidade ao *bin* correspondente à cada k palavra próxima. Supondo $k = 3$, o triângulo da Figura 3 seria atribuído às palavras w_3, w_4 e w_5.

A abordagem *soft assignment*, presente nos trabalhos de [9] e [20], também atribui mais de uma palavras-visual à um vetor-de-característica, porém ela atribui um peso de tal forma que uma palavra-visual próxima tem maior relevância do que uma palavra-visual mais distante. A técnica de codificação *soft* surgiu como um alternativa à codificação *hard* que despreza o fato de que um vetor-de-característica pode estar à mesma distância entre palavra-visuais diferentes. A codificação *soft* permite uma caracterização mais robusta evitando ambiguidade.

D. Fase de Sumarização

A fase de sumarização (*pooling*), é a etapa responsável por sintetizar em um único vetor-de-característica a representação final da imagem. As três técnicas de *pooling* mais tradicionais na literatura são: *sum-pooling*, *average-pooling* e *max-pooling*. Todas elas fornecem uma representação de tamanho fixo e são baseadas na contagem da ocorrência não-ordenada das palavras-visuais no espaço da imagem.

As técnicas *sum-pooling* e *average-pooling* são semelhantes: cada uma gera um histograma de palavras-visuais, porém a diferença entre elas está no fato de que *sum-pooling* representa a soma das ocorrências de cada palavra-visual na imagem, enquanto *average-pooling* se refere à media, ou seja, um histograma normalizado. Já a técnica de *max-pooling*, quando usada com a técnica de codificação *hard*, gera um vetor binário, indicando a presença ou ausência de uma palavra-visual na imagem. Quando *max-pooling* é usado com a codificação *soft*, gera um histograma indicando a máxima ocorrência de cada palavra-visual na imagem.

Outras técnicas de *pooling* mais sofisticadas codificam informação de distância na representação final. Por exemplo, a ideia da técnica BossaNova [21] é computar um histograma de distâncias entre cada característica local extraída da imagem e suas respectivas palavras-visuais. A técnica VLAD [22] acumula, para cada palavra-visual, a diferença entre a distância de cada ponto à sua respectiva palavra-visual. Esses métodos de sumarização permitem caracterizar a distribuição dos vetores-de-características locais aos seus respectivos centros, que correspondem às palavras-visuais.

O grande problema com essas técnicas é a perda da informação espacial das palavras-visuais na imagem. Alguns trabalhos mostram que a inclusão da localização espacial das palavras-visuais é uma características bastante importante em uma técnica de *pooling* para aumentar o poder discriminativo da abordagem [23], [24], [25].

III. Avaliações Experimentais

Para implementar a abordagem BoVW em um sistema CBIR (*Content-Based Image Retrieval*) quatro grandes questões devem ser consideradas para se obter uma caracterização semântica mais próxima à expectativa do usuário:

1) Qual a melhor maneira de detectar e representar as regiões-de-interesse de uma imagem?
2) Como gerar um dicionário eficiente e eficaz?
3) Como codificar e sumarizar de maneira eficiente cada região-de-interesse?
4) Como medir a similaridade?

Diversas técnicas tem sido propostas na literatura para atender esses requisitos, porém continua sendo um campo de pesquisa com oportunidades para avanços e diversas lacunas que podem e devem ser exploradas para aumentar ainda mais a eficiência da abordagem em um sistema CBIR.

Nesta seção são discutidos e avaliados quatro importantes aspectos da abordagem BoVW: o tamanho do dicionário visual, o detector das regiões-de-interesse, a técnica de codificação e a função de distância utilizada para medir a similaridade entre imagens. A proposta é investigar esses quatro parâmetros em diferentes bases e como eles afetam em conjunto a precisão dos resultados na recuperação de imagens por conteúdo.

A. Bases Avaliadas

Os experimentos foram realizados usando cinco diferentes bases de imagens, todas elas de domínio público e disponíveis na literatura:

- Base 1: essa é a base Corel1000 [1], que consiste em uma base de dados formada por cenas naturais bem complexas e cheias de detalhes, composta por 1.000 imagens divididas em 10 classes.
- Base 2: é uma base formada por imagens de texturas utilizada no trabalho de [5], composta por imagens de superfícies de materiais, tais como madeira, mármore e pele sob diferentes pontos de vista, escalas e condições de iluminação. O conjunto consiste em 1.000 imagens de tamanho 640x480 pixels, divididos em 40 amostras de 25 diferentes classes.
- Base 3: essa é uma base composta por 5.042 imagens biomédicas de 32 categorias diferentes, tais como Raio-X, CT, MRI, etc. Ela é um subconjunto da base do ImageCLEFmed 2007.
- Base 4: é a base 15-Scenes, composta por 4.485 imagens de cenas naturais separadas em 15 categorias. Cada categoria é composta por 210 a 410 imagens, e o tamanho médio das imagens é de 300x250 pixels.
- Base 5: é a base de imagens da Oxford Flowers, composta por 1.360 imagens de 17 espécies diferentes de flores(80 imagens por categorias), com variações de escala, iluminação e oclusões parciais.

Em conjunto, todas essas bases abrangem uma grande diversidade de tipos de imagens, cada uma com características específicas e aplicações variadas. O objetivo aqui é explorar as situações em que cada técnica possui um desempenho superior e analisar como os resultados podem mudar de um domínio para outro.

B. Abordagens Avaliadas

Quatro diferentes detectores de pontos-de-interesse foram avaliados nos experimentos:

- Harris-Affine (Harris);
- Diferença-de-Gaussianas (DoG);
- Pontos Aleatórios (*Random Sampling*);
- Denso (*Dense Sampling*).

A Figura 4 mostra o resultado da aplicação desses detectores em algumas imagens.

Para descrever os pontos foi utilizado a proposta do descritor SIFT, que é baseada na magnitude e gradiente de uma região 16x16 ao redor de cada ponto-de-interesse detectado. O agrupamento utilizado para gerar o dicionário visual foi obtido usando a estratégia do algoritmo *bisecting k-means*. As codificações utilizadas foram *hard* e *soft* e a técnica de sumarização *average-pooling*.

C. Resultados Obtidos

Os gráficos da Figura 5 mostram os valores de mAP (*mean Average Precision*) obtidos em cada base avaliada e usando diferentes tamanhos de dicionário, entre 310 e 2500. Em geral, na literatura os melhores resultados são obtidos usando

[1] base disponível em: http://wang.ist.psu.edu/docs/related/

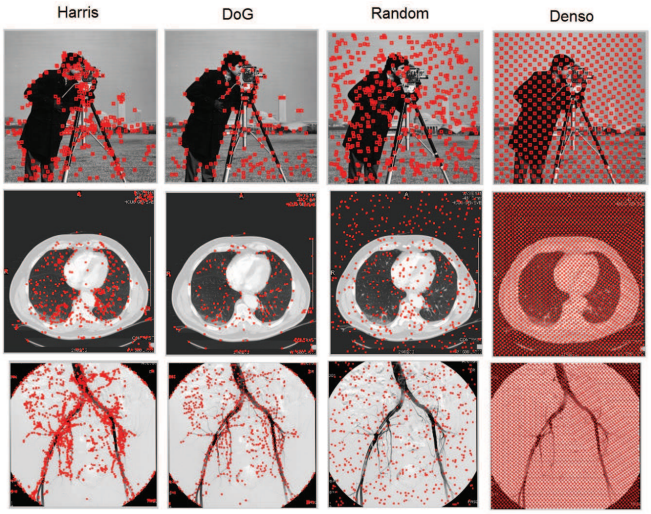


Fig. 4. Resultado da aplicação dos detectores de pontos-de-interesse utilizados nos experimentos.

dicionários grandes, acima de 1.000 palavras-visuais. Porém, pelos experimentos realizados nota-se que, para a maioria das bases, aumentando o tamanho do dicionário os valores de mAP não aumentam expressivamente, em média apenas 0.05% . Em alguns casos, por exemplo para a base 4 (15-scenes), o aumento do tamanho do dicionário fez reduzir a precisão da recuperação. Isso mostra que dicionários muito grandes não são muito vantajosos, o pouco ganho na precisão (e até a perda) não compensa o custo computacional exigido por um vetor-de-característica de alta dimensionalidade.

Além disso, a grande maioria dos trabalhos propostos na literatura defendem a codificação *soft* sobre a *hard*, por ser menos ambígua e gerar uma caracterização com menor erro de quantização. Porém, outro fato importante observado nos experimentos foi que, em todas as bases, os melhores resultados foram obtidos através da codificação *hard*, porém a diferença de precisão entre as codificações *hard* e *soft* são pequenas em algumas bases (2,4,5) e pode chegar até 0,2% em outras (bases 1,3). Apenas para a base 3, a codificação *soft* utilizando o detector DoG obteve resultado superior à codificação *hard*.

O detector de pontos-de-interesse tem um forte impacto na precisão da recuperação. Os melhores resultados foram obtidos usando os detectores Harris e *Dense*. Esse último, por sua vez, obteve os melhores resultados em três bases (1,3,4). O detector DoG por sua vez, que é a técnica utilizada pelo descritor SIFT, gerou resultados semelhantes ao detector baseado em pontos-aleatórios (*Random*). Inclusive para duas bases (1 e 3), o detector *Random* obteve uma acurácia quase semelhante ao detector *Dense*. Isso mostra que, para algumas bases, não é necessário utilizar detectores de pontos-de-interesse "sofisticados", ou seja, técnicas que possuam alguma semântica envolvida na detecção dos pontos. Pelos experimentos notou-se que as bases que não necessitam de detectores sofisticados são aquelas formadas por images de cenas naturais e médicas. Para bases complexas, como texturas e flores, o detector Harris se apresentou como o mais adequado.

Foi analisada também a influência da função distância na precisão final da recuperação de imagens. Para isso, foram realizados testes usando seis diferentes funções de distâncias, L_1, L_2, χ^2, Canberra, Cosseno e Intersecção de Histogramas. A Tabela I apresenta os valores de mAP obtidos para as bases avaliadas usando um dicionário formado por 310 palavras-visuais, codificação *hard* e usando o detector de pontos-de-interesse *Dense*.

A função de distância escolhida também exerce influência na acurácia da recuperação de imagens. Para a maioria das bases (1,4,5) as funções de distâncias L_1 e Intersecção de Histogramas apresentaram os melhores resultados. Apenas para as bases 2 e 3, a função de distância Canberra apresentou resultado superior, porém não muito melhor que as funções L_1 e Intersecção de Histogramas. Desse modo, a distância L_1 se apresentou a mais vantajosa, pois é a distância que tem menor custo computacional dentre as avaliadas.

TABLE I. MAP VALORES OBTIDOS USANDO DIFERENTES FUNÇÕES DE DISTÂNCIAS.

Base	**Função de Distância**					
	L_1	L_2	χ^2	**Canberra**	**Cosseno**	**Intersecção**
Base 1	***0,503***	0,436	0,419	0,465	0,472	***0,503***
Base 2	0,268	0,230	0,245	***0,321***	0,243	0,266
Base 3	0,712	0,641	0,672	***0,715***	0,680	0,712
Base 4	***0,568***	0,504	0,385	0,493	0,519	***0,568***
Base 5	***0,118***	0,113	0,114	0,114	0,113	***0,118***

IV. CONSIDERAÇÕES FINAIS

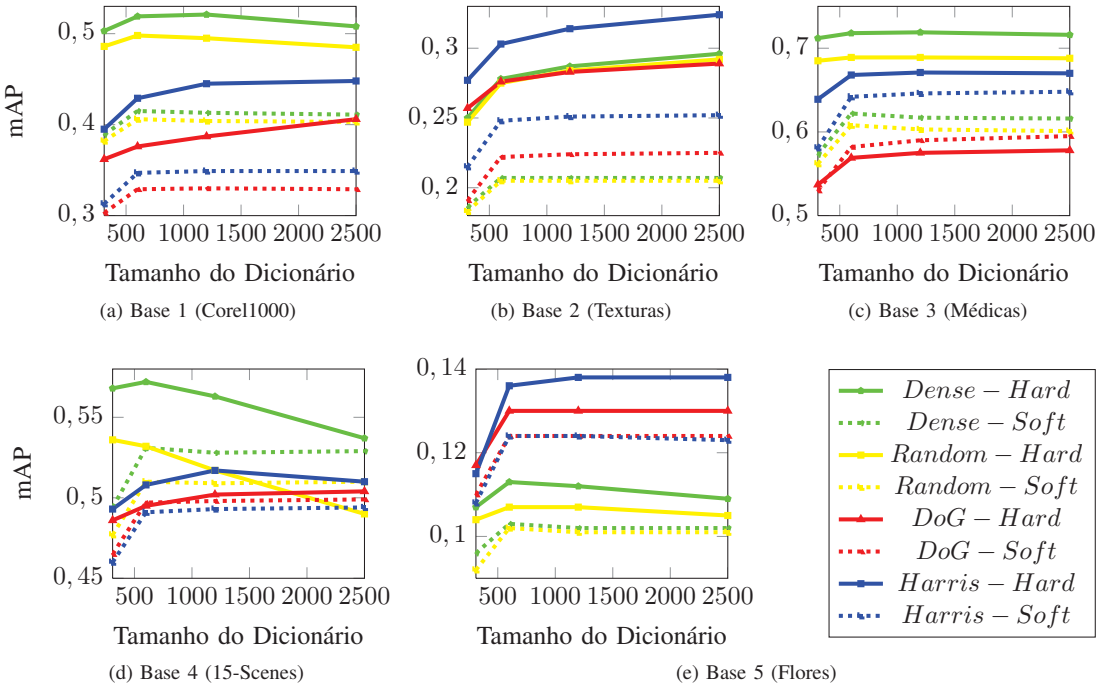
Neste artigo foi apresentado a abordagem *Bag-of-Visual-Words* (BoVW). Tal abordagem, apesar de sua popularidade na área, tem muitos parâmetros que afetam o seu desempenho, por exemplo o método de geração do dicionário, o tamanho do dicionário, o tipo de detector utilizado etc. O bom desempenho final depende fortemente dessa escolha, o que não é uma tarefa fácil, pois como avaliado pelos experimentos, não se pode generalizar um conjunto de técnicas que atenda todos os diferentes conjuntos de imagens. A motivação desse trabalho foi investigar/mostrar como a escolha de diferentes técnicas na metodologia BoVW pode influenciar o desempenho na tarefa de recuperação de imagens, usando combinações que possam quantificar o melhor possível a similaridade perceptual entre as imagens. Os resultados obtidos mostram que, escolhendo a melhor combinação entre detector e função de similaridade, é possível ter um bom desempenho na tarefa de recuperação descrevendo imagens com menos de mil palavras-visuais.

AGRADECIMENTOS

Este trabalho é apoiado, em parte, pela FAPESP, CAPES, SticAMSUD, o projeto RESCUER, financiado pela Comissão Européia (Grant 614154) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico CNPq/MCTI

REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.



(a) Base 1 (Corel1000) (b) Base 2 (Texturas) (c) Base 3 (Médicas)

(d) Base 4 (15-Scenes) (e) Base 5 (Flores)

Fig. 5. Valores de mAP (mean Average Precision) obtidos usando diferentes: tamanho de dicionários, detectores de pontos-de-interesse e técnicas de codificação.

[2] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision (ICCV)*, vol. 2, 2003, pp. 1470–1477.

[3] N. C. Batista, A. P. B. Lopes, and A. d. A. Araújo, "Detecting buildings in historical photographs using bag-of-keypoints," in *XXII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, 2009.

[4] A. P. B. Lopes, S. E. F. d. Avila, A. N. A. Peixoto, R. S. Oliveira, M. d. M. Coelho, and A. d. A. Araújo, "Nude detection in video using bag-of-visual-features," in *XXII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, 2009.

[5] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1265–1278, 2005.

[6] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[7] G. V. Pedrosa, A. J. M. Traina, and C. A. Z. Barcelos, "Shape description based on bag of salience points," in *ACM Symposium on Applied Computing (SAC)*, 2015, pp. 74–79.

[8] Q. Chaudry, S. H. Raza, A. N. Young, and M. D. Wang, "Automated renal cell carcinoma subtype classification using morphological, textural and wavelets based features," *Signal Processing Systems*, vol. 55, no. 1-3, pp. 15–23, 2009.

[9] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *International Conference on Image and Video Retrieval (CIVR)*, 2007, pp. 494–501.

[10] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *European Conference on Computer Vision (ECCV)*, 2006.

[11] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," 2000, pp. 1–2.

[12] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[13] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2005, pp. 604–610.

[14] E. R. S. Santos, A. P. B. Lopes, E. A. Valle, J. M. Almeida, and A. A. Araújo, "Vocabulários visuais para recuperação de informação multimídia," in *Simpósio Brasileiro em Sistemas Multimídia e Web (WebMedia)*, vol. 2, 2010, pp. 21–24.

[15] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, Jun. 2007.

[16] D. Pelleg and A. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," in *International Conference on Machine Learning*. Morgan Kaufmann, 2000, pp. 727–734.

[17] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *KDD Workshop on Text Mining*, 2000, pp. 1–20.

[18] P. Willett and A. El-Hamdouchi, "Comparison of hierarchic agglomerative clustering methods for document retrieval," *Computer Journal*, vol. 32, no. 3, pp. 220–227, Jun. 1989.

[19] H. Jegou, H. Harzallah, and C. Schmid, "A contextual dissimilarity measure for accurate and efficient image search." in *CVPR*, 2007.

[20] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *In CVPR*, 2008.

[21] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araujo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, vol. 117, pp. 453 – 465, 2013.

[22] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

[23] O. A. Penatti, F. B. Silva, E. Valle, V. Gouet-Brunet, and R. S. Torres, "Visual word spatial arrangement for image retrieval and classification," *Pattern Recognition*, vol. 47, no. 2, pp. 705 – 720, 2014.

[24] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR, 2006, pp. 2169–2178.

[25] G. Pedrosa and A. Traina, "From bag-of-visual-words to bag-of-visual-phrases using n-grams," in *26th Conference on Graphics, Patterns and Images (SIBGRAPI)*, Aug 2013, pp. 304–311.