Vocabulários Visuais para Recuperação de Informação Multimídia*

Elerson R. da S. Santos
UFMG, Brazil
elerson@dcc.ufmg.br

Ana Paula B. Lopes
UFMG, Brazil
paula@dcc.ufmg.br

Eduardo A. do Valle Jr.
UNICAMP, Brazil
mail@eduardovalle.com

Jussara M. de Almeida
UFMG, Brazil
jussara@dcc.ufmg.br

Arnaldo A. de Araújo
UFMG, Brazil
arnaldo@dcc.ufmg.br

ABSTRACT

In most current commercial search engines, images and videos are represented essentially based on textual information associated with them. Such an approach can only achieve limited results, which can be enhanced by the addition of content-based information. One of the most promising content-based approaches to classify visual data in a semantic level are those based on Bags of Visual Features (BoVFs) representations. Nevertheless, extracting BoVFs have a high-computational cost, making those not scalable to large databases. In this work, we perform an experimental investigation in which we compare three strategies for visual vocabulary selection, a key step in the BoVF-building process. Our results indicate that it is possible to substitute the classical approach – based on k-means clustering algorithm – for a simple random selection, without statistically significant loss of information, but with an radical decrease in computational cost.

RESUMO

Na maioria dos sistemas de busca atuais, imagens e videos são representados essecialmente com base em informações textuais a eles associados.Tal abordagem pode apresentar resultados limitados, que podem ser melhorados com a adição de informações baseadas no conteúdo. Um dos enfoques mais promissores que se baseiam em conteúdo para classificar dados visuais em nível semântico são aqueles baseados em histogramas de características visuais(BoVFs). No entanto, os BoVFs tem um custo computacional alto, fazendo esses inviáveis para grandes bases de dados. Neste trabalho, realizamos uma investigação experimental na qual comparamos três estratégias para a seleção do vocabulário visual, que é um passo fundamental no processo de construção dos BoVFs. Nossos resultados indicam que é possível substituir a abordagem clássica - baseada no algoritmo de agrupamentos *k-means* - por uma simples seleção aleatória, sem uma perda estatísticamente significativa de informações, mas com uma redução radical no custo computacional.

*Visual Vocabularies for Multimedia Information Retrieval

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering, Retrieval models*

General Terms

Performance

Keywords

Recuperação de Informação Multimídia

1. INTRODUÇÃO

A quantidade crescente de imagens e videos – profissionais e amadores – disponibilizados na WEB coloca em evidência a dificuldade dos atuais sistemas de recuperação de informação em lidar com conteúdo multimídia. A representação utilizada pela maioria dos sistemas comerciais é feita quase exclusivamente com base em informações textuais associadas à esses elementos. Isso acontece porque a análise de conteúdos multimídia, diferentemente de análise de conteúdo textual, não pode ser feita diretamente. A associação entre os pixels de uma imagem ou vídeo com o significado semântico destes é uma tarefa complexa e um problema de pesquisa em aberto, chamado *lacuna semântica* [10]. Uma das linhas de pesquisa que têm apresentado os melhores resultados para o estreitamento da lacuna semântica são baseadas em representações por histogramas de pontos de interesse (BoVFs, do inglês, *Bag of Visual Features*) [13]. Entretanto, a extração de BoVFs requer um processamento de alto custo computacional, inviabilizando a sua aplicação em bases de dados de larga escala.

Neste trabalho, é feita uma investigação experimental que fornece subsídios para reduzir consideravelmente o tempo de processamento necessário para extração de BoVFs, sem perda significativa de informação. Essa redução no custo computacional é um passo importante para tornar computacionalmente viável a introdução de informação de conteúdo em sistemas de busca tradicionais.

2. TRABALHOS RELACIONADOS

A representação com base em histogramas de pontos de interesse (BoVFs) surgiu nos últimos anos como uma das mais promissoras para classificação de conceitos de alto nível semântico à partir de imagens e vídeos [13]. Em [1], BoVFs são usados para classificação de objetos, enquanto que em

[6], eles são usados na classificação de tipos cenas. BoVFs são usados também em [7] para classificação de imagens contendo nudez. Na revisão proposta em [8], abordagens baseadas em BoVFs para classificação de ações humanas em vídeos são enfatizadas em relação à outras propostas.

Um dos pontos-chave na construção de representações BoVF é a criação do dicionário ou vocabulário visual. Normalmente, isso é feito pelo agrupamento dos pontos de interesse no espaço de características, quase sempre com emprego do clássico algoritmo de agrupamento *k-means* [13]. Entretanto, o custo computacional do *k-means* é bastante alto, o que compromete a escalabilidade dos métodos baseados BoVFs.

Algumas alternativas ao *k-means* podem ser encontradas na literatura. Em [4], por exemplo, um algoritmo alternativo ao *k-means* é proposto no contexto de classificação de objetos. No caso do reconhecimento de ações em vídeos, [12], por exemplo, usam árvores para criação dos vocabulários visuais, enquanto que em [3] é proposta uma melhoria do vocabulário criado pelo *k-means*.

Neste trabalho, é feita uma avaliação experimental que compara os resultados obtidos com o *k-means* com os de outras duas alternativas: seleção aleatória pura (SAP) e seleção aleatória melhorada (SAA), uma estratégia gulosa para seleção dos pontos para o vocabulário visual que será proposta.

3. A CONSTRUÇÃO DO BOVF

O primeiro passo para a criação do BoVF é a deteção de pontos de interesse. Existem na literatura muitos algoritmos para selecionar pontos de interesse, sendo o mais difundido o Transformada de Características Invariante à Escala (SIFT, do inglês, *Scale-Invariant Feature Transform*), que é apontado por [11] como um divisor de águas na área, uma vez que foi utilizado com sucesso em diversas aplicações desafiadoras, incluindo reconhecimento de objetos por meio de BoVFs. Para detectar pontos de interesse em videos levando em conta as variações no tempo, o algoritmo proposto em [5] para detectar pontos de interesse espaço-temporais (STIP, do inglês, *Spatio-Temporal Interest Points*),vem assumindo nos ultimos anos um papel semelhante ao do SIFT para imagens.

Após serem detectados os pontos de interesse, é necessária alguma forma de descrevê-los. Nesse artigo, foi utilizado o detector de pontos de interesse STIP, juntamente com os descritores de histogramas de gradientes orientados (HoG, do inglês *histograms of oriented gradients*), e histogramas de fluxo óptico (HoF, do inglês *histograms of optical flow*) combinados[1].

Com os descritores extraídos dos vídeos, o passo seguinte é a redução de dimensionalidade desses descritores. A redução de dimensionalidade é utilizada para reduzir o custo computacional requerido nas etapas seguintes, e é feita tipicamente por meio do algoritmo de análise de componentes principais (PCA, do inglês, *Principal Component Analysis*).

A próxima etapa é a criação do vocabulário visual. Isso é tradicionalmente feito por meio da utilização de um algoritmo de agrupamento, que define cada grupo de descritores semelhantes como uma dessas palavras visuais. O nosso trabalho concentra-se nessa etapa da formação do BoVF, que é melhor detalhada na Seção 3.1

Com o vocabulário criado, o último passo para a criação do BoVF é o cálculo do histograma de palavras visuais, que servirá de representação do vídeo para a etapa de classificação. Para isso, cada descritor é associado ao agrupamento (i.e. palavra visual) cujo centróide está mais próximo dele. Finalmente, o histograma é normalizado.

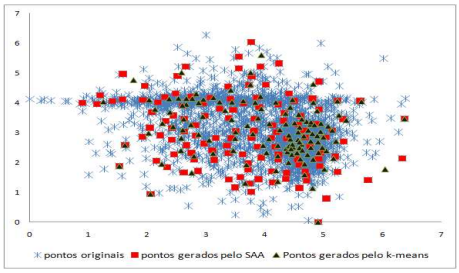

Figure 1: Comparação entre os centroides selecionados pelo k-means e o SAA.

3.1 Diferentes Métodos para Construção do Vocabulário

Apesar da larga utilização do algoritmo de agrupamento *k-means* para a construção do vocabulário, não é muito claro na literatura que essa seja a melhor forma de fazê-lo. Além disso, o custo computacional do *k-means* é consideravelmente alto, limitando a escalabilidade dos métodos baseados em BoVFs. Os resultados em [9], por exemplo, indicam uma oscilação aleatória dos resultados obtidos com o *k-means*. Isso leva à suposição de que uma seleção aleatória de descritores para compor o vocabulário visual possa obter resultados semelhantes, com um custo computacional praticamente nulo. Essa é a suposição que nosso trabalho se propõe a verificar experimentalmente. Além disso, é proposta e avaliada também uma estratégia gulosa para a seleção dos pontos, como alternativa intermediária entre o *k-means* e a seleção aleatória.

3.1.1 Seleção Aleatória Aprimorada

A Seleção Aleatória Aprimorada (SAA) é apresentada neste trabalho, e foi desenvolvida com o objetivo de avaliar uma alternativa intermediária entre o *k-means* e a seleção aleatória pura (SAP) dos centróides que irão compor o vocabulário visual. O algoritmo proposto baseia-se na observação encontrada em [4], de que o *k-means* tende a selecionar muitos pontos em regiões do espaço de características onde a distribuição é mais densa, deixando outras regiões sub-representadas no vocabulário. A solução de proposta em [4], no entando, envolve um novo algoritmo de agrupamento e amostragens densas de pontos, o que compromete o objetivo de redução do custo computacional.

A SSA começa selecionando um conjunto aleatório de pontos. Para cada ponto, todos os pontos que estão aquém de um limiar para a distância mínima são eliminados do conjunto. O algoritmo repete o passo anterior com todos os pontos da lista, em uma abordagem gulosa, até que todos os pontos tenham sido avaliados. Desta forma uma distância

[1]O código executável para esses algoritmos, fornecido pelos autores, pode ser encontrado em http://www.irisa.fr/vista/Equipe/People/Laptev/download.html.

mínima entre os pontos escolhidos para o vocabulário fica garantida, além de se obter uma melhor distribuição dos pontos, conforme mostra a Figura 1.

4. AVALIAÇÃO EXPERIMENTAL

Para a avaliar as diferentes estratégias descritas para formação do vocabulário visual, foi utilizada a base de vídeos Weizmann [2]. Essa base possui vídeos de 9 pessoas, cada uma executando 10 ações diferentes. As ações executadas são: *abaixar*, *acenar com as duas mãos*, *acenar com uma mão*, *andar de lado*, *caminhar*, *correr*, *fazer polichinelo*, *pular no lugar*, *pular para frente* e *pular em uma perna*. Essa base é considerada um padrão *de facto* para a tarefa de reconhecimento de ações, tendo sido utilizada por muitos autores para avaliar diferentes abordagens para esse fim.

As taxas de reconhecimento dos BoVFs construídos para essa base foram estimadas utilizado-se o algoritmo de classificação Máquinas de Vetores de Suporte (SVM, do inglês, *Support Vector Machines*). Para estimar o melhor valor para parâmetro de erro do SVM, foi utilizada validação cruzada com 5 dobras. O parâmetro de erro na margem C foi variado entre 10^{-8} e 10^{8} (em passos multiplicativos de 10), determinando, assim, a melhor taxa de reconhecimento em cada caso. Além disso, para amenizar os efeitos da seleção aleatória dos pontos (que também ocorre na inicialização do *k-means* e do SAA), cada experimento foi repetido 10 vezes para cada algoritmo e tamanho de vocabulário.

Os experimentos foram executados de forma que o tamanho do vocabulário variasse de aproximadamente 100 à 1500[2], com e sem a utilização do PCA – exceto no SAA com distância χ^2, já que esta não está definida para valores negativos, que aparecem nos BoVFs transformados para o espaço gerado pelo PCA. A ideia nesse caso é avaliar o impacto do PCA na qualidade da classificação e portanto, uma possível remoção de mais esse passo no processamento dos BoVFs.

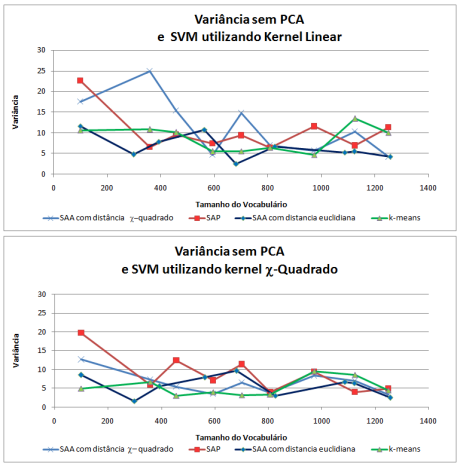

Figure 3: Variâncias apresentadas pelos métodos, com e sem aplicação do PCA.

Os gráficos da Figura 2 mostram a média das taxas de reconhecimento obtida sobre as 10 execuções para todos os casos. Neles pode-se ver que o *k-means* apresenta taxas de reconhecimento um pouco acima que os demais algoritmos na maioria dos casos. O exame visual também mostra que os resultados sem o PCA (à direita) tendem a ser um pouco melhores que os com PCA (à esquerda). Com relação aos diferentes *kernels* para o SVM (linear ou χ^2), a Figura 2 sugere uma maior instabilidade nos resultados obtidos com o *kernel* linear. Tal instabilidade fica mais evidente na Figura 3, que mostra a variância dos resultados das 10 rodadas de cada caso.

[2]Os valores exatos dependem da seleção feita pelo SAA.

Dado que os cálculos sem PCA apresentam resultados ligeiramente superiores, eles são usados para uma análise mais detalhada das diferenças entre os diferentes casos. A Tabela 1 mostra os intervalos de confiança para as taxas de reconhecimento com $k \approx 700$, para diferentes algoritmos e *kernels*. O exame dessa tabela mostra que, mesmo relaxando o intervalo para uma confiança de apenas 90%, as diferenças apresentadas não são estatisticamente significativas. Esses dados fornecem duas indicações importantes. Primeiramente, contrariando a intuição sugerida pela Figura 1 e por [4], um vocabulário melhor distribuído no espaço de características não foi capaz de melhorar a qualidade da representação do vídeo para fins de classificação. Além disso, do ponto de vista estatístico, *não há diferença entre selecionar as palavras visuais aleatoriamente ou por agrupamento com k-means*. Este resultado é um indicativo extremamente relevante para os propósitos deste trabalho, pois ele sugere que é possível substituir um passo de alto custo computacional – o agrupamento – por outro de custo constante na composição dos BoVFs.

Table 1: Intervalos de confiança para as taxas de reconhecimento com $k \approx 700$, sem PCA e com kernel linear, com 90% de confiança.

Método	Kernel	Média (%)	Interv. Conf. (90%)
SAA	linear	86.7	(79.6 – 93.7)
SAA	χ^2	86.6	(81.9 – 91.2)
SAP	linear	88.1	(82.5 – 93.7)
SAP	χ^2	86.7	(80.5 – 92.8)
k-means	linear	87.2	(82.9 – 91.5)
k-means	χ^2	86.5	(83.2 – 89.7)

Finalmente, os tempos de execução para construção do vocabulário usando 80% dos vídeos disponíveis (i.e., as bases de treino na validação cruzada) são mostrados na Figura 4. É importante notar os seguintes aspectos: a) mesmo com essa base de tamanho reduzido, o tempo de execução é bastante alto, considerando que este é apenas um passo do processo de construção dos BoVFs; b) o tempo do *k-means* sem PCA é o maior de todos, devido às dimensões do descritor original; e c) o tempo da SAP não é mostrado porque ele é insignificante (< 1s). A máquina utilizada possui um processador Intel core2 duo E6750 2.66GHz e 2GB de memória.

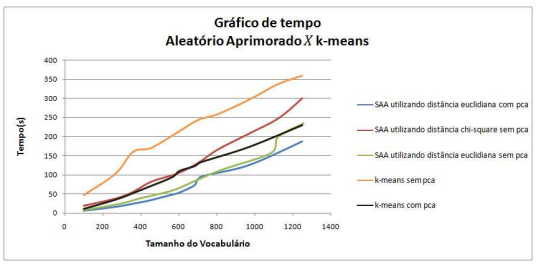

Figure 4: Tempos de execução para a construção do vocábulário usando 80% dos vídeos da base Weizmann.

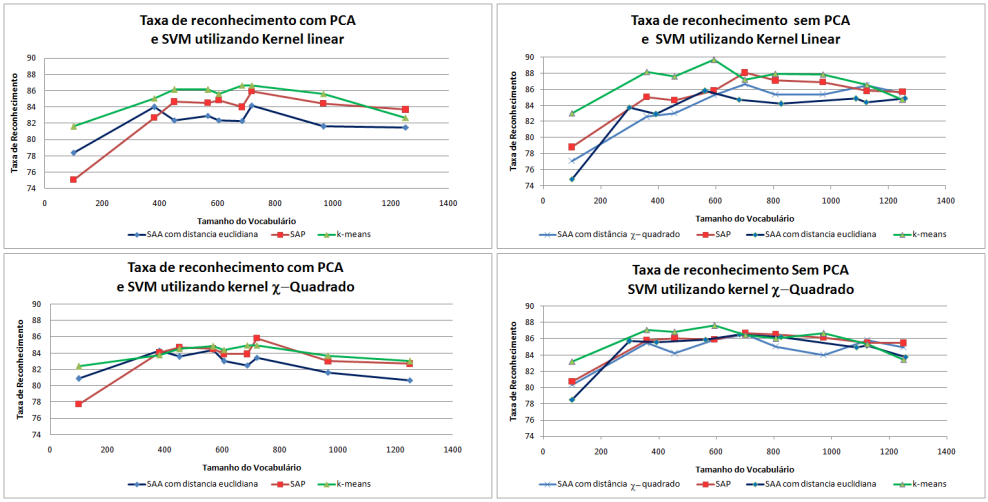

Figure 2: Taxas de reconhecimento usando a SAA, a SAP e o *k-means*, com e sem redução por PCA.

5. CONCLUSÃO

Nesse artigo foi proposta uma avaliação experimental do impacto de diferentes de algoritmos para construção de vocabulários visuais. O objetivo dessa análise é reduzir o custo computacional da criação de descritores BoVFs, tornando métodos de classificação de imagens e vídeos baseados em BoVFs mais escaláveis.

Os resultados obtidos experimentalmente com uma base padrão para reconhecimento de ações humanas indicam que: a) apesar da intuição sugerir que um vocabulário com melhor cobertura do espaço de características gere melhores descritores BoVF, essa expectativa não se verifica na prática; b) as etapas de redução de características dos descritores locais e agrupamento com *k-means* no fluxo de processamento do BoVF podem ser substituídas por uma seleção aleatória de pontos, cujo custo computacional é praticamente nulo.

Os resultados obtidos nesses experimentos são um passo importante na obtenção de descritores para imagens e vídeos que sejam baseados no seu conteúdo visual que possam escalar para ser incorporados em sistemas de recuperação de informação reais. Em trabalhos futuros, pretende-se expandir os experimentos para outras bases de dados, tanto de imagens como de vídeos. As novas bases deverão ser maiores e ser compostas de imagens mais realísticas, com objetivo de verificar a generabilidade dos resultados alcançados até o momento.

6. AGRADECIMENTOS

Este trabalho é parcialmente financiado pelo Instituto Nacional de Ciência e Tecnologia para a Web - INCTWeb(MCT/ CNPq 573871/2008-6), CNPq, FAPEMIG, CAPES e FAPESP.

7. REFERÊNCIAS

[1] S. Agarwal and A. Awan. Learning to detect objects in images via a sparse, part-based representation. *Trans. Pattern Anal. Mach. Intell.*, 26(11):1475–1490, 2004. Member-Dan Roth.

[2] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *Trans. Pattern Anal. Mach. Intell.*, 29(12):2247–2253, December 2007.

[3] Y.-G. Jiang and C.-W. Ngo. Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval. *Comput. Vis. Image Underst.*, 113(3):405–414, 2009.

[4] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 1, pages 604–610, October 2005.

[5] I. Laptev and T. Lindeberg. Space-time interest points. In *Proceedings of ICCV '03*, pages 432–439, 2003.

[6] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of IEEE CVPR '06*, volume II, pages 2169–2178, 2006.

[7] A. P. B. Lopes, S. E. F. de Avila, A. N. A. Peixoto, R. S. Oliveira, and A. de Albuquerque Araújo. A bag-of-features approach based on hue-sift descriptor for nude detection. In *Proceedings of EURASIP EUSIPCO '09*, 2009.

[8] A. P. B. Lopes, E. A. do Valle Jr., J. M. de Almeida, and A. A. de Araújo. Action recognition in videos: from motion capture labs to the web (preprint). 2010.

[9] A. P. B. Lopes, R. S. Oliveira, J. M. de Almeida, and A. de Albuquerque Araújo. Spatio-temporal frames in a bag-of-visual-features approach for human actions recognition. In *Proceedings of SBC SIBGRAPI '09*, pages 1–7, 2009.

[10] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[11] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis.*, 3(3):177–280, 2008.

[12] H. Uemura, S. Ishikawa, and K. Mikolajczyk. Feature tracking and motion compensation for action recognition. In *Proceedings of BMVA BMVC '08*, pages 1–8, 2008.

[13] E. Valle and M. Cord. Advanced techniques in CBIR local descriptors, visual dictionaries and bags of features. In *Proceedings of SBC SIBGRAPI '09*, 2009.