Grammatical Inference and SIFT for Scene Recognition

Lucas C. Ribas*[1], Marcelo Borth‡, Amaury A. Castro Jr.*, Wesley N. Gonçalves* e Hemerson Pistori†
*Universidade Federal de Mato Grosso do Sul, UFMS - CPPP
LaRPP Laboratório de Robótica de Ponta Porã, Ponta Porã, MS 79907-414
Email: lucascorreiaribas@gmail.com, {amaury.junior, wesley.goncalves}@ufms.br
†Universidade Católica Dom Bosco, UCDB
INOVISAO Laboratório de Visão Computacional, Campo Grande, MS
Email: pistori@ucdb.br
‡Instituto Federal de Mato Grosso do Sul, IFMS
Ponta Porã, MS 79909-000
Email: marcelo.borth@ifms.edu.br

***Resumo*—Grammatical inference in computer vision regained attention in recent years due to the emergence of new local feature techniques, such as Scale Invariant Feature Transform (SIFT) e Speeded Up Robust Features (SURF). This paper presents a methodology that converts an image into a string based on the SIFT and bag-of-visual-words (BOW). Given the strings, grammar induction techniques are used for scene recognition. In the BOW, the vocabulary is usually learned using an unsupervised approach. To improve image description, this paper proposes a supervised vocabulary learning. This approach improves the string obtained from the image and the experimental results have demonstrated its robustness in a challenging scene database.**

I. Introdução

Nos últimos anos, o reconhecimento de cenas e ambientes (e.g. corredor, escritório, sala, entre outras) é um problema que tem recebido muita atenção na comunidade de visão computacional. Entre as principais aplicações, o reconhecimento de cenas auxilia a localização de robôs móveis, aumentando a compreensão semântica do ambiente e a complexidade das tarefas que eles podem realizar.

Para a descrição das imagens, a extração de características locais tem sido extensivamente empregada no reconhecimento de cenas. Os extratores *Scale Invariant Feature Transform* (SIFT) [1] e *Speeded Up Robust Features* (SURF) [2] são dois dos principais extratores desta categoria. Em conjunto com a metodologia de histograma de palavras visuais (*Bag-of-visual-words − BOW*) [3], estes extratores têm apresentado resultados promissores na tarefa descrita acima. O BOW utiliza as características extraídas pelo SIFT ou SURF para construção de um vocabulário. Para isso, são utilizadas as características locais extraídas de imagens de todas as categorias/classes. Embora grandes avanços foram reportados nos últimos anos, os histogramas (ocorrência de palavras do vocabulário) desconsideram a relação espacial das características locais.

Para superar esse problema, trabalhos foram propostos para incluir informação espacial na metodologia BOW [4], [5], [6], [7], [8]. Entre os trabalhos atuais, Pistori et al. [5] e Tu et al. [9] propõem o uso de inferência gramatical. Estes trabalhos resgatam as técnicas de reconhecimento sintático de padrões visuais muito usadas nas décadas de 60 e 70 [10]. De acordo com [5], a principal questão no uso de gramáticas em imagens está na representação da imagem por uma cadeia que represente as características relevantes da mesma e não ignorem relações estruturais entre palavras visuais.

Este artigo tem como proposta utilizar inferência gramatical para o reconhecimento de cenas. No método, características locais são extraídas das imagens e utilizadas para construção do vocabulário. Em seguida, as características são rotuladas e uma cadeia é obtida para cada imagem. A cadeia impõe a ordem e, consequentemente, a informação espacial entre as características locais. Por fim, uma gramática é construída para representar cada categoria. Dada uma gramática e uma cadeia obtida de uma imagem, é possível responder se a cadeia pertence à gramática, implicitamente, respondendo se a imagem pertence à categoria que a gramática representa. Além do método acima, este artigo apresenta uma comparação entre a construção do vocabulário de forma supervisionada [11] e não-supervisionada.

Os resultados obtidos em uma base de imagens complexa mostraram que o uso da inferência gramatical e características locais é adequado para o reconhecimento de cenas. Além disso, a construção do vocabulário supervisionado forneceu resultados superiores em comparação com a construção do vocabulário não-supervisionado. Esse resultado demonstra que a obtenção de palavras individuais para cada categoria é importante para o processo de descrição de imagens.

O restante do artigo é descrito como segue. A Seção 2 apresenta os conceitos e métodos necessários para a compreensão do artigo. O método para utilizar inferência gramatical no reconhecimento de cenas é descrito na Seção 3. A Seção 4 e 5 apresentam a configuração dos experimentos e os resultados obtidos, respectivamente. Por fim, as conclusões e os trabalhos futuros são discutidos na Seção 6.

[1]Bolsista do Programa de Educação Tutorial PET/MEC/SESu (PET/Fronteira)

II. TEORIAS

A. Scale Invariant Feature Transform (SIFT)

O *Scale Invariant Feature Transform* (SIFT) [1] é um dos mais importantes métodos para extração de características de uma imagem. O SIFT extrai da imagem uma coleção de vetores de características locais, chamados de *keypoints* ou pontos-chave [1]. Cada um dos pontos-chave é invariante a escala, rotação, e parcialmente invariante a mudança de iluminação. Essas propriedades são de suma importância para tarefas como reconhecimento de cenas e objetos.

O método para extração de características pode ser descrito como segue. O principal passo do método é a identificação de pontos-chave ρ_i que sejam invariantes a mudança de escala da imagem. Este passo é implementado usando a função de diferença de Gaussianas [12]. Em seguida, cada ponto-chave ρ_i é representado por quatro elementos:

1) (x_i, y_i), localização espacial na imagem.
2) σ_i, escala em que ele foi detectado.
3) θ_i, orientação predominante do gradiente.
4) $\varphi_i \in \Re^{128}$, vetor de características contendo 128 valores que descrevem a região ao redor do ponto.

Segundo Lowe [1] um aspecto importante deste algoritmo é a capacidade de gerar um grande número de características, que pode cobrir toda imagem e em diferentes variações e escalas. A Figura 1 mostra um exemplo de uma imagem com os pontos-chave detectados pelo SIFT. A localização (x_i, y_i) do ponto-chave na imagem é representada pelos pontos brancos. Enquanto que os círculos brancos envolta dos pontos-chave estão relacionados à escala e os traços a orientação.

B. Histograma de Palavras Visuais

O histograma de palavras visuais (*Bag-of-Visual-Words* − BOW) [3] é uma das técnicas mais populares usadas para reconhecimento de cenas nos últimos anos. De acordo com Csurka et al. [3], as principais etapas dessa técnica, visualizadas na Figura 2, são:

a) Detecção e descrição da imagem (SIFT).
b) Agrupamento dos pontos-chave em grupos de acordo com seus descritores, ou seja, criar um vocabulário (K-Means).
c) Rotulação de cada ponto-chave com uma palavra do vocabulário.
d) Construção de um histograma de ocorrência das palavras na imagem.

Uma etapa importante nessa técnica é a criação do vocabulário, nesta, o tamanho do vocabulário k é um parâmetro importante na descrição das imagens. Apesar dos avanços recentes e dos resultados promissores, o poder descritivo dessa técnica acaba sendo limitado, pois esta descarta informações espaciais das palavras na imagem. Quando apenas contamos a ocorrência de uma palavra em uma imagem, não estamos considerando onde esta palavra está localizada na imagem e nem a relação de uma palavra com as demais. Essas informações podem ser características importantes para a classificação da imagem.

C. K-Testable

O K-Testable é uma técnica que dado um tamanho de memória k_t, tenta-se encontrar um autômato para reconhecer uma linguagem representada por um número de cadeias passadas como parâmetro. Uma linguagem k-testável é uma subclasse de uma linguagem regular que encontra prefixos, sufixos e sub-cadeias nos dados de treinamento [5]. A principal característica é que cada caractere é dependente apenas dos k_t-1 caracteres anteriores e segundo Pistori et al. [5] a análise de uma cadeia de caracteres pode ser feita usando uma memória de tamanho fixo k_t.

III. MÉTODO PROPOSTO

Este trabalho propõe uma melhoria no método que utiliza inferência gramatical no problema de classificação de imagens proposto por Pistori et al. [5]. A primeira é em relação ao algoritmo de extração de características locais, neste trabalho foi usado o SIFT enquanto que o trabalho anterior utilizou o SURF. Outro ponto relevante é quanto ao tamanho k do vocabulário, que neste trabalho foi experimentado em uma ampla faixa de valores. Os valores de k baixos acabam limitando a descrição das características de cada categoria/classe, consequentemente, o desempenho acaba sendo limitado. Por fim, os vocabulários são construídos de duas formas: não-supervisionada (trabalho anterior) e supervisionada.

O método proposto pode ser descrito em 5 etapas: a) extração de pontos-chave, b) construção do vocabulário, c) rotulação dos pontos-chave, d) construção da cadeia, e) aprendizagem da gramática. As seções abaixo descrevem cada etapa em detalhes.

A. Extração de Pontos-chave

Nesta etapa, os pontos-chaves de cada imagem são encontrados por meio do SIFT. Considere ρ_i^j como sendo o ponto-chave i extraído da imagem j. Cada ponto é representado por 5 elementos descritos nas seções anteriores: $\rho_i^j = \{(x_i^j, y_i^j), \sigma_i^j, \theta_i^j, \varphi_i^j\},\ 1 \leq i \leq M$. Em geral, o número de pontos-chave M variam de imagem para imagem, pois estes dependem diretamente do gradiente.

B. Construção do Vocabulário

Dados os pontos-chave extraídos das imagens, esta etapa constrói um vocabulário com base no conjunto de descritores D (Equação 1). O conjunto de descritores é composto por todos os descritores de todas as imagens. Devido às restrições de memória, em alguns casos, utiliza-se um subconjunto de D escolhido de forma aleatória.

$$D = [\varphi_i^j],\ 1 \leq i \leq M_j, 1 \leq j \leq N \tag{1}$$

onde N é o número de imagens de treinamento e M_j é o número de pontos extraídos da imagem j.

Dado o conjunto de descritores, estes são agrupados por meio do algoritmo k-means e o conjunto de k centróides C é obtido:

$$C = \text{k-means}(D) \tag{2}$$

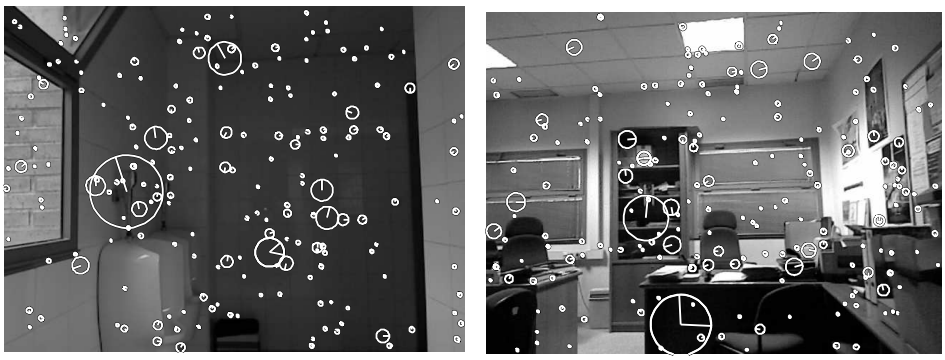


Figura 1. Imagens de cenas da base de imagens com pontos-chave detectados pelo SIFT. Cada ponto-chave é representado por uma posição espacial na imagem (x_i, y_i), uma escala σ_i, uma orientação θ_i e um vetor de características que descreve a região ao redor.

O conjunto C é descrito como o vocabulário aprendido para as imagens de treinamento. Cada centróide $C_l \in C,\ 1 \leq l \leq k$ possui a mesma dimensão dos descritores, isto é, $C_l \in \Re^{128}$.

O processo descrito acima é dito ser não-supervisionado, pois descritores de todas as classes são utilizados para construir o vocabulário. Para tornar esse processo supervisionado, um vocabulário é construído para cada classe das imagens de treinamento. Para isso, os centróides C_p são obtidos por meio do k-means aplicado em um conjunto D_p contendo somente os descritores das imagens pertencentes à classe p:

$$D_p = [\varphi_i^j],\ \text{se } j \text{ pertence à classe } p \quad (3)$$
$$C_p = \text{k-means}(D_p) \quad (4)$$

Considerando n_c como o número de classes, n_c conjuntos de centróides serão obtidos. A vantagem desta construção é que palavras serão obtidas para cada classe, o que não é garantido na construção não-supervisionada. Por fim, para a construção do vocabulário supervisionado, os n_c conjuntos de centróides são concatenados.

$$C = [C_p],\ 1 \leq p \leq n_c \quad (5)$$

C. Rotulação dos Pontos-chave

A próxima etapa é a rotulação de cada ponto-chave das imagens de treinamento. Dados os pontos-chave ρ_i^j para a imagem j, cada ponto-chave é rotulado com o índice da palavra mais próxima do vocabulário, isto é, o centróide mais próximo conforme apresentado na Equação 6.

$$r_i^j = \arg\min_{l=1}^{k} |\varphi_i^j, C_l| \quad (6)$$

onde r_i^j corresponde ao rótulo do ponto-chave i da imagem j e $|.|$ é a distância Euclidiana.

Após essa etapa, cada ponto-chave possui um rótulo de forma que pontos-chaves com descritores parecidos possuem os mesmos rótulos. Um exemplo pode ser visto na Figura 3, onde os rótulos são indicados por letras do alfabeto.

D. Construção da Cadeia

Nesta etapa os pontos-chave na imagem são percorridos gerando uma cadeia com as palavras visuais correspondentes. Para percorrer os pontos-chave na imagem, a ordem radial foi definida por ser invariante à rotação. Nesta ordem, as palavras são extraídas da imagem de acordo com a distância Euclidiana da sua posição espacial para o centro da imagem, onde as palavras mais próximas do centro da imagem serão extraídas primeiro. Um exemplo é ilustrado na Figura 3 onde podem ser visto os círculos concêntricos e a cadeia formada é CACFEBCACBFCBEFDEFA.

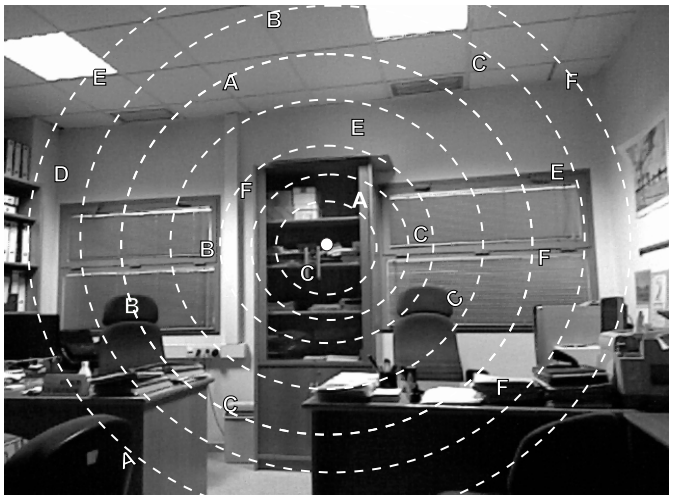


Figura 3. Extração das palavras da imagem e formação das cadeias. Na imagem acima a cadeia gerada é CACFEBCACBFCBEFDEFA.

E. Aprendizagem da Gramática

A última etapa do método é inferir uma gramática para cada classe através das cadeias de palavras. Para a inferência, o algoritmo K-Testable é utilizado.

Após as etapas acima, as cadeias que representam as imagens de teste são validadas nas gramáticas geradas. Para validar uma cadeia, é contado o número de erros da cadeia para cada gramática que representa uma classe. Erros ocorrem quando há um caractere na cadeia que não pertence a linguagem ou não existe a transição na linguagem. A gramática que retornar o menor número de erros é a qual pertence à cadeia.

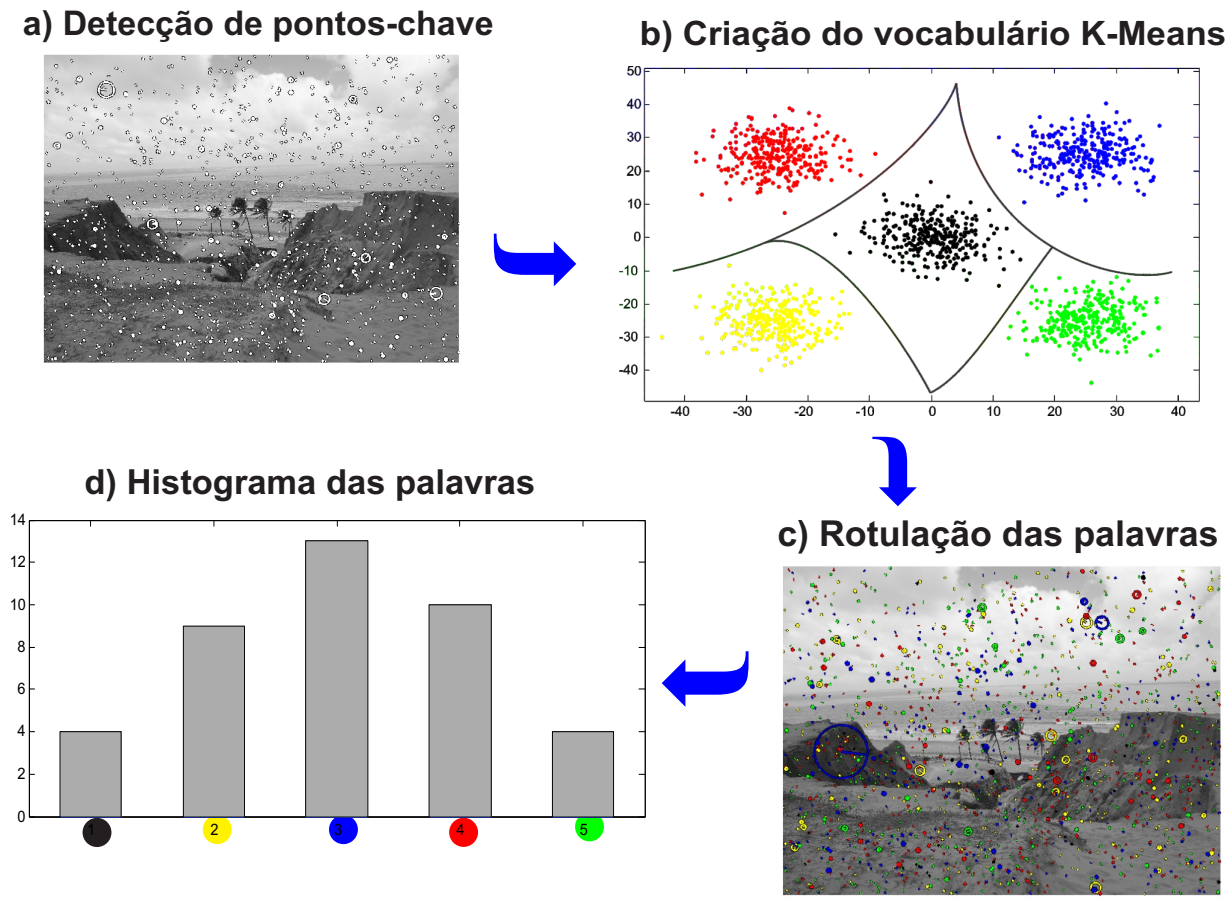


Figura 2. Etapas para extração de características utilizando o BOW. a) Detecção dos pontos-chaves utilizando o algoritmo SIFT. b) Criação do vocabulário utilizando o algoritmo K-Means. c) Rotulação das palavras de acordo com o vocabulário gerado pelo K-Means. d) Contagem de ocorrência das palavras em uma imagem.

IV. Experimentos

Os experimentos foram realizados utilizando imagens de cenas, a Figura 4 mostra exemplos de imagens da base. A base de imagens usada é composta por um total de 1105 imagens coloridas, divididas em 6 classes de cenas classificadas por humanos [13]. A Tabela I resume o número de imagens de cada classe. O tamanho de cada imagem é 640×480.

A construção do vocabulário foi realizada de duas formas, supervisionada e não-supervisionada com k variando entre 60 e 2700. O parâmetro k_t do K-Testable utilizado nos experimentos foi de k_t = 2. Para a divisão do conjunto de dados de treinamento e teste foi usado o modelo de validação cruzada, com um número de pastas igual a 10.

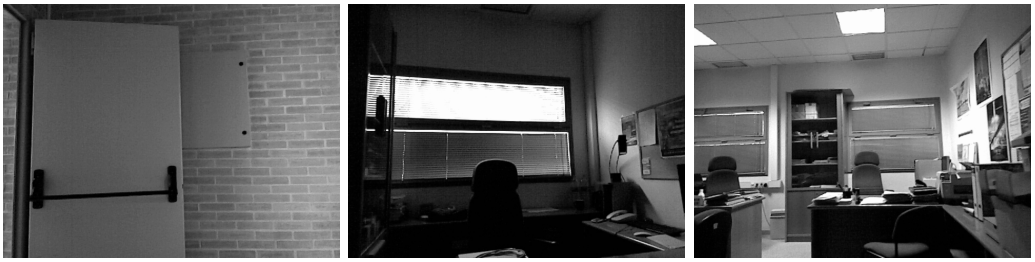
Tabela I. Número de imagens de cada classe utilizado nos experimentos.

Classe	Quantidade de imagens	Porcentagem do Total
Elevador	40	3.61
Banheiro	188	17.01
Sala de professor	200	18.09
Secretaria	181	16.38
Laboratório	282	25.52
Sala de TI	214	19.36

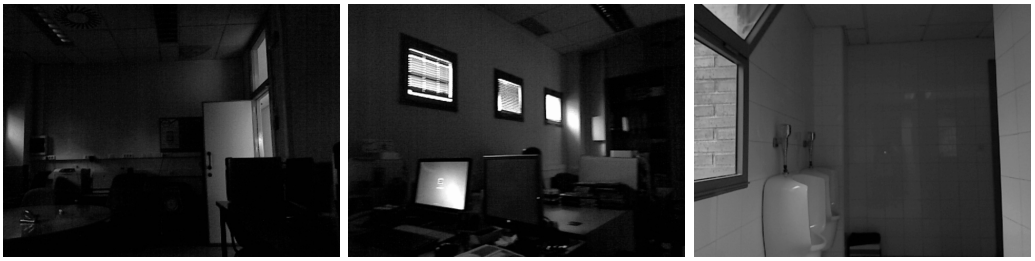
V. Resultados e Discussões

A Tabela II apresenta os resultados dos experimentos na forma supervisionada e não-supervisionada e com diferentes valores de k. A maior taxa de classificação correta (TCC) foi de 77.33 % usando construção de vocabulário supervisionada e um tamanho de vocabulário igual 2.400. Na Figura 5 temos um gráfico comparando os dois processos de construção de vocabulário para os diferentes tamanhos de k. Nesta, pode-se observar que os vocabulários construídos de forma supervisionada e com valores de k altos levam vantagem em relação ao método comparado. Os vocabulários construídos de forma supervisionada garantem palavras para cada classe, enquanto, na construção não-supervisionada não há essa garantia, isso pode explicar o desempenho superior da construção supervisionada. Outro fator relevante é quanto ao desempenho superior dos experimentos usando tamanhos de vocabulário maiores, o que pode ser explicado pela maior descrição das características das imagens ao utilizar valores altos de vocabulário.

A Figura 6 compara as matrizes de confusão para a construção supervisionada e não-supervisionada do vocabulário. Na construção supervisionada do vocabulário a maior TCC foi de 77.33 % com k = 2400 e na não-supervisionada a TCC foi de 72.70 % com k = 1440. Nas duas matrizes, pode-se



(a) Área de Elevador (b) Sala de Professor (c) Secretaria



(d) Sala de Estudante (e) Sala de TI (f) Banheiro

Figura 4. A base é composta por 6 classes com imagens de 640×480.

Tabela II. RESULTADOS DOS EXPERIMENTOS (TCC E DESVIO PADRÃO) PARA DIFERENTES TAMANHOS DE VOCABULÁRIO. AS MELHORES TCC PARA CADA TAMANHO DE VOCABULÁRIO ESTÁ EM NEGRITO.

Tamanho de Vocabulário	TCC (%)	
	Não-Supervisionado	Supervisionado
60	28.10 ($\pm$0.01)	**29.15** ($\pm$0.02)
120	**33.80** ($\pm$0.02)	33.70 ($\pm$0.03)
240	52.30 ($\pm$0.02)	**52.43** ($\pm$0.02)
480	63.80 ($\pm$0.02)	**70.05** ($\pm$0.03)
960	69.70($\pm$0.02)	**71.22** ($\pm$0.03)
1200	69.80 ($\pm$0.04)	**73.79** ($\pm$0.02)
1440	72.70 ($\pm$0.03)	**74.32** ($\pm$0.02)
1800	71.15 ($\pm$0.05)	**76.03** ($\pm$0.02)
2100	72.00 ($\pm$0.02)	**77.18** ($\pm$0.03)
2400	69.42 ($\pm$0.02)	**77.33** ($\pm$0.02)
2700	68.63 ($\pm$0.04)	**76.77** ($\pm$0.02)

observar que em algumas classes possuem uma considerável taxa de classificação incorreta. Como exemplo, a classe Sala de Professor é classificada incorretamente como Laboratório em 39.50 % na construção não-supervisionada e em 36.50 % na supervisionada. Um dos prováveis motivos dessa confusão na classificação está na complexidade da base de imagens. Nesta, há imagens de diferentes classes, mas com bastantes características similares, como computadores, cadeiras, mesas e etc.

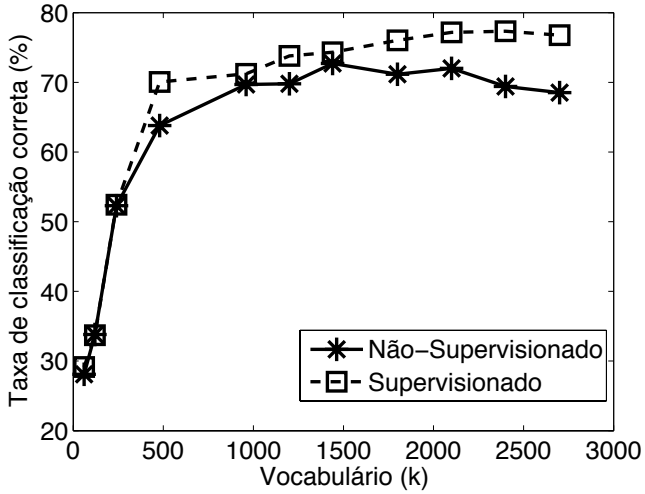
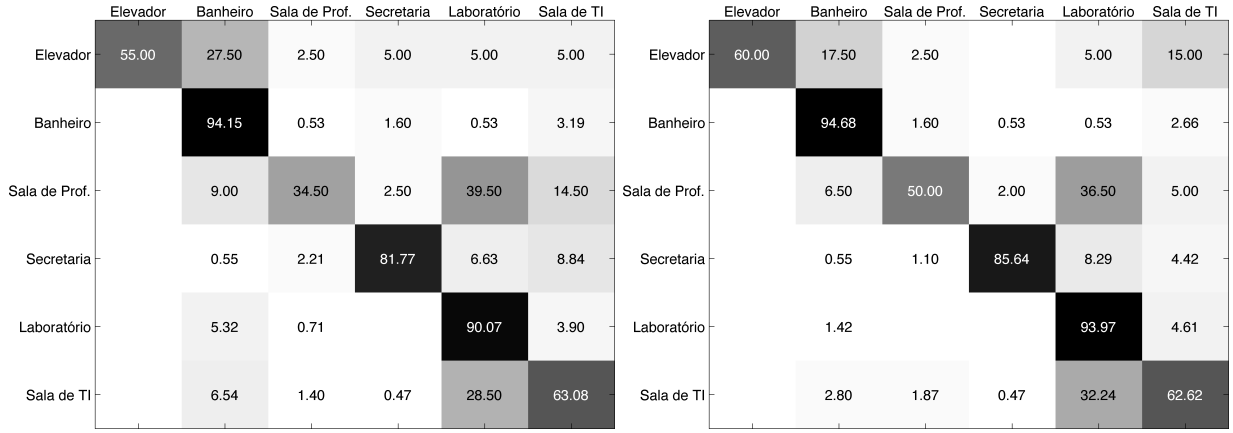


Figura 5. Gráfico com a TCC (eixo y) para diferentes tamanhos de vocabulário k (eixo x) e com modo de treinamento supervisionado e não-supervisionado.

VI. CONCLUSÃO E TRABALHOS FUTUROS

O problema de classificação de imagens (e.g. reconhecimento de cenas) em visão computacional tem sido um campo de estudo que vem recebendo grande atenção na área nos últimos anos. Uma das linhas de pesquisa neste campo é a conversão de imagens em cadeias e o uso de gramáticas para a classificação de imagens. Esta ganhou maior atenção nos últimos anos devido ao surgimento de novas técnicas para extração de características de imagem (e.g. SIFT). Neste trabalho a proposta é a conversão de imagens em cadeias com base na metodologia de BOW, como já foi proposto em [5]. No entanto, a principal contribuição deste artigo está na construção supervisionado do vocabulário e no uso de uma ampla faixa de valores de tamanho de vocabulário. Como pode se visto nos resultados há um considerável aumento de desempenho ao usar valores altos de k, combinado com vocabulários construídos de forma supervisionada.

Para trabalhos futuros, sugere-se o uso de novas estratégias para percorrer as palavras na imagem e gramáticas mais poderosas. A estratégia de percorrer as palavras na imagem é determinante para adicionar informação espacial e características da imagem nas cadeias. Portanto, se as características da imagem estiverem implícita na cadeia, o desempenho na classificação será melhor. O uso de gramáticas mais poderosas, também é outro fator a ser levado em consideração em trabalhos futuros. Gramáticas que aceitam exemplos positivos e negativos ou gramáticas não-regulares, podem ser alternativas para melho-



(a) Não-Supervisionado (b) Supervisionado

Figura 6. Matriz de confusão com a TCC para cada classe de cena referente a maior TCC total do modo não-supervisionado 72.70 % e supervisionado 77.33 %.

rar o desempenho. Para outros trabalhos pretende-se avaliar imagens com ruídos e diferentes condições de iluminação.

AGRADECIMENTOS

A Pró-Reitoria de Extensão, Cultura e Assuntos Estudantis (PREAE) da Universidade Federal de Mato Grosso do Sul (UFMS) e ao Programa NERDS da Fronteira pelo apoio financeiro na locomoção e estadia. Ao MEC/SESu, através do Programa de Educação Tutorial (PET/Fronteira), no qual o primeiro autor deste artigo é bolsista.

REFERÊNCIAS

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

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, Jun. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2007.09.014

[3] G. C. e Christopher R. Dança e Lixin Fan e Jutta Willamowski e Cédric Bray, "categorização visual com sacos de pontos-chave," in *Em Workshop sobre Estatística Aprendizagem em Visão Computacional, ECCV*, 2004, pp. 1–22.

[4] E. Zhang and M. Mayo, "Improving bag-of-words model with spatial information," in *Image and Vision Computing New Zealand (IVCNZ), 2010 25th International Conference of*, Nov 2010, pp. 1–8.

[5] H. Pistori, A. Calway, and P. Flach, "A new strategy for applying grammatical inference to image classification problems," in *Industrial Technology (ICIT), 2013 IEEE International Conference on*, 2013, pp. 1032–1037.

[6] C. Zhang, S. Wang, Q. Huang, J. Liu, C. Liang, and Q. Tian, "Image classification using spatial pyramid robust sparse coding," *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1046 – 1052, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865513000573

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

[8] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: from visual words to visual phrases," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, June 2007, pp. 1–8.

[9] Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu, "Image parsing: Unifying segmentation, detection, and recognition," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 113–140, 2005.

[10] K. S. Fu and K. S. Fu, *Syntactic methods in pattern recognition*. Academic press New York, 1974, vol. 197, no. 1.

[11] M. Jiu, C. Wolf, C. Garcia, and A. Baskurt, "Supervised learning and codebook optimization for bag-of-words models," *Cognitive Computation*, vol. 4, no. 4, pp. 409–419, 2012. [Online]. Available: http://dx.doi.org/10.1007/s12559-012-9137-4

[12] G. L. G. Gonzales, "Aplicação da técnica sift para determinação de campos de deformações de materiais usando visão computacional," Mestrado, PONTIFíCIA UNIVERSIDADE CATóLICA DO RIO DE JANEIRO - PUC-RIO, 2011.

[13] J. Martinez-Gomez, I. Garcia-Varea, M. Cazorla, and B. Caputo, "Overview of the imageclef 2013 robot vision task," in *Working Notes, CLEF 2013*, 2013.