Seleção Não-Supervisionada de Métodos de Binarização para Documentos Históricos

MAC 499 – Trabalho de Formatura Supervisionado
Instituto de Matemática e Estatística da Universidade de São Paulo (IME-USP)

Aluno: Denis T. Ikeda
Orientador: Ronaldo Fumio Hashimoto

1. Apresentação

Este trabalho tem como proposta elaborar uma seleção não-supervisionada de métodos de binarização para documentos históricos, levando em consideração os avanços recentes apresentados nas competições DIBCO [1][2][3].

Ela seleciona um dos resultados de um conjunto de métodos de binarização utilizando uma estimativa dos valores de *Precision* e *Recall* [4], descartando os resultados discrepantes e ordenando os restantes pelo valor de *F-Measure*. Esta estimativa é feita pois assume-se que a binarização ideal *(ground truth)* não está disponível.

2. Binarização

Binarização é o ato de transformar uma imagem em níveis de cinza em uma imagem binária, que utiliza somente preto e branco. Seu objetivo é separar todo o texto da imagem de entrada dos ruídos e do fundo.

Esta é uma importante etapa de pré-processamento de documentos para obtenção de seu texto através do *OCR* (do inglês, reconhecimento ótico de caracteres). Utilizada isoladamente, diminui radicalmente o tamanho dos arquivos destes documentos e melhora sua legibilidade ao eliminar os ruídos.

3. O problema

Métodos de binarização para documentos em bom estado existem há décadas, mas como documentos históricos costumam possuir defeitos como manchas, envelhecimento, dobras e o verso aparecendo na frente, o desempenho deles torna-se insatisfatório.

Para averiguar o interesse nesta área, e obter uma noção do estado da arte atual, as competições DIBCO [1][2][3] foram realizadas, e apresentam vários método interessantes de desempenho surpreendente.

No entanto, mesmo métodos excelentes, como o primeiro lugar do DIBCO 2011, não produzem resultados aceitáveis para certas imagens. Logo, a binarização precisa ser supervisionada para obter qualidade consistente.

Figuras. 3 imagens de teste dos conjuntos de dados estão dispostas em cada coluna. Ao rotular de cima para baixo, as linhas são: (a) imagem original e (b) *ground truth* associado; o restante são resultados dos métodos de (c) Otsu, (d) Niblack, (e) Sauvola, (f) White, (g) Su, e (h) a seleção de métodos proposta.

Os métodos locais adaptativos utilizam tamanho de janela 15 por 15, incluindo o método de Su, que originalmente estima seu tamanho de janela. Os parâmetros de Niblack e Sauvola são os recomendados (k = -0,2 para o primeiro, k = 0,5 e R = 128 para o segundo), o parâmetro de White é bias = 1.2, pois o recomendado bias = 2 não produzia bons resultados, e o parâmetro de Su é Nmin = 8, também originalmente estimado.

As configurações da seleção proposta estão inteiramente descritas na monografia deste poster.

4. O método proposto

Ao invés de propor um novo método de binarização, este estudo tenta aumentar a consistência da binarização através de uma seleção de métodos, reduzindo ou até eliminando a necessidade de supervisioná-la.

Para isto, um conjunto de métodos de binarização é executado, e seus resultados, guardados. A partir deles, estima-se o *Precision*, a probabilidade de um pixel da resposta estar correto, e *Recall*, a probabilidade de um pixel correto estar na resposta, como em [4]. Os resultados são então ordenados por *F-Measure*, média harmônica destes valores.

O resultado de maior *F-Measure* é selecionado. Entretanto, como valores díspares de *Recall* influenciam negativamente na seleção, o resultado com a maior distância do intervalo considerado aceitável é filtrado, e os valores de *Precision* e *Recall* do conjunto reduzido são utilizados para uma nova filtragem até que todos resultados estejam no intervalo aceitável.

5. Resultados e conclusões

Para comparar os resultados, foi implementado um software de avaliação que utiliza as métricas apresentadas nas competições DIBCO. Aqui, serão apenas mostradas as de 2009: F-Measure, PSNR, NRM e MPM. Valores altos das duas primeiras significam melhores resultados; das duas últimas, piores. Além das métricas, foram implementados alguns métodos clássicos (Otsu, Niblack, Sauvola e White) e o vencedor do H-DIBCO 2010 simplificado (Su), um pouco pior que o original.

Como os resultados nas tabelas mostram, a seleção, cujo conjunto para escolha são todos os seus concorrentes, consegue resultados superiores para este conjunto de imagens nas duas formas de posicionamento (*ranking*). Embora não escolha sempre o resultado ótimo, ele supera os métodos de seu conjunto para escolha ao evitar seus piores casos.

	posição	placar
Proposta	1	5
Su	2	8
White	3	14
Otsu	4	15
Sauvola	5	19
Niblack	6	23

	posição	placar
Proposta	1	60
Su	2	69
Otsu	3	93
White	4	111
Sauvola	5	137
Niblack	6	190

	FM (%)	PSNR	NRM ($\times 10^{-2}$)	MPM ($\times 10^{-3}$)
Otsu	78,53	15,26	5,54	13,86
Niblack	38,85	5,76	19,75	183,08
Sauvola	61,66	13,84	25,46	1,35
White	80,49	15,40	13,41	2,85
Su	89,97	18,06	6,93	0,75
Proposta	91,43	18,68	5,33	0,84

Tabela 1 mostra as posições dos métodos avaliados para o conjunto de dados do DIBCO 2009 considerando a média das avaliações das métricas, ordenando cada avaliação de acordo com seu significado, e somando todas posições, formando o placar. As posições finais levam em conta a ordem crescente de placar.

Tabela 2 mostra as posições considerando as avaliações de cada imagem ao invés de ordenar pela média.

Tabela 3 mostra a média da avaliação de cada métrica para todos os métodos.

Bibliografia

[1] GATOS, Basilis; NTIROGIANNIS, Konstantinos; PRATIKAKIS, Ioannis. ICDAR 2009 Document Image Binarization Contest (DIBCO 2009). In: INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION, 10., 2009, Barcelona, Spain. *Proceedings*... [Washington DC]: IEEE Computer Society, 2009. p. 1375–1382.

[2] _____. H-DIBCO 2010 – Handwritten Document Image Binarization Competition. In: INTERNATIONAL CONFERENCE ON FRONTIERS IN HANDWRITING RECOGNITION, 12., 2010, Kolkata, India. *Proceedings of the*... Los Alamitos: IEEE Computer Society, 2010. p. 727–732.

[3] _____. ICDAR 2011 Document Image Binarization Contest (DIBCO 2011). In: INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION, 11., 2011, Beijing, China. *Proceedings*... [S.I.]: IEEE Computer Society, 2011. p. 1506–1510.

[4] LAMIROY, Bart; SUN, Tao. Precision and Recall Without Ground Truth. In: IAPR INTERNATIONAL WORKSHOP ON GRAPHICS RECOGNITION, 9., 2011, Seoul, Korea. *Proceedings*... [S.I.: s.n.], 2011. p. [?].