Criação de base de manuscritos em língua japonesa

Luiz Fellipe Machi Pereira (PIC/UEM), Yandre Maldonado e Gomes da Costa (Orientador) e Diego Bertolini (Co-orientador),
e-mail: ra103491@uem.br, yandre@din.uem.br e diegobertolini@gmail.com.

UEM/CTC/DIN.

Área: Ciência da Computação.
Subárea: Metodologia e Técnicas da Computação.

Palavras-chave: reconhecimento de padrões, manuscrito.

Resumo

Neste trabalho é apresentada a base de dados Japanese Kana and Brazilian Portuguese (JKBP), disponibilizada gratuitamente com o objetivo de apoiar o desenvolvimento de tarefas de reconhecimento baseadas em manuscritos. Pelo que sabemos, esta é a primeira base de dados disponibilizada à comunidade científica composta por manuscritos brasileiros e japoneses. São apresentados experimentos básicos realizados na JKBP, explorando algumas de suas possíveis de tarefas de identificação, como de silabário e de escritor. Duas abordagens foram testadas, a primeira usa o documento inteiro para a extração de características. A segunda abordagem leva em consideração a divisão da amostra do manuscrito em blocos e faz a fusão dos resultados obtidos em cada bloco. Os resultados alcançados mostraram que o processo de zoneamento e o uso da fusão tardia proporcionam um aumento significativo do desempenho em alguns cenários. As melhores taxas de acerto alcançadas foram de 97,98% e 83,77% nas tarefas de identificação do escritor usando manuscritos em português e japonês, respectivamente, e 100% na classificação de silabário.

Introdução

A identificação automática do escritor é um tópico amplamente discutido pela comunidade de pesquisa de reconhecimento de padrões. Com o aumento dos esforços de pesquisa sobre esse problema, surge a necessidade de bancos de dados de manuscritos, especialmente se levarmos em conta bancos de dados disponibilizados publicamente, considerando diversos idiomas e scripts, e criados em condições que emulam um cenário real.
Neste trabalho, apresentamos uma nova base de dados de manuscritos offline, multi-script e texto-dependente. A criação desta base de dados justifica-se pela escassez de bases de dados de manuscritos disponíveis para o Kana japonês, e também pela escassez de bases de dados que possam apoiar investigações sobre identificação de escritores no cenário multi-script. A base de dados aqui apresentada foi especialmente curada

para investigações sobre tarefas de classificação de manuscritos em português do Brasil e Japonês. Além da tarefa de identificação do escritor, a base também permite classificação de script, faixa etária e gênero.

Materiais e métodos

Extração de features - Neste trabalho usamos Speed Up Robust Features (SURF), proposto por (BAY *et al*, 2006), como o descritor de textura a ser aplicado nas imagens do manuscrito. Este descritor de textura foi escolhido para realizar os experimentos devido aos bons resultados obtidos anteriormente na tarefa de identificação do escritor, conforme mostrado em (PINHELLI *et al*, 2020) e (SOUZA *et al*, 2019).
Experimentos que dividem as imagens em blocos sem sobreposição foram feitos inspirados em (M. Roberto e Souza *et al*). Em nossos testes, limitamos o número de blocos em quatro. Após dividir a imagem em blocos, o descritor SURF foi aplicado a cada bloco para extrair características e, possivelmente, encontrar bons pontos de interesse locais que não seriam considerados em toda a imagem.

Classificadores - A classificação foi realizada utilizando Support Vector Machine (SVM), pois já foi utilizada com sucesso para tarefas de classificação neste cenário, conforme mostrado nos trabalhos (PINHELLI *et al*, 2020) e (SOUZA *et al*, 2019). Para os experimentos, a classificação foi realizada utilizando C-SVM presente no framework LIBSVM (CHANG;LIN,2011) com kernel Radial Basis Function (RBF) e a otimização da configuração de hiperparâmetros feita por grid search.
A fim de evitar resultados tendenciosos, o protocolo Document Filter (DF) foi aplicado aos experimentos que usaram divisões de blocos. Após as predições de SVM serem obtidas para cada bloco individualmente, elas foram combinadas para obter uma única decisão final para o documento como um todo. Para isso, em todos os experimentos realizamos testes usando as regras de fusão Soma, Produto e Máximo.

Resultados e Discussão

Identificação de autor - Para os experimentos de identificação do escritor, dois conjuntos diferentes de experimentos foram realizados. O primeiro conjunto refere-se a testes com manuscritos apenas em português e o segundo conjunto refere-se a testes que utilizaram apenas manuscritos japoneses. Nestes conjuntos, foram feitos testes com extração de características usando todo o documento, e usando zoneamento e fusão, os os conjuntos de treino e teste foram criados utilizando cross validation (CV).
Utilizando a imagem do documento inteira, o descritor SURF foi aplicado em cada amostra de cada autor e foi feito um 5-fold stratified CV. Essa configuração alcançou resultados satisfatórios na tarefa de reconhecimento de autores utilizando amostras em português, com precisão de 0,9509 e F-

Measure média de 0,9531, enquanto os resultados para amostras em japonês inferiores, com precisão de 0,5895 e 0,5989 de F-Measure.
Para melhorar os resultados, foi aplicado o processo de classificação por divisão em blocos, e continuou-se a empregar 5-fold CV. Em testes com ambas as linguagens, a precisão produzida pela regra do produto superou a produzida pela regra do Máxima e pela regra Soma. A precisão para reconhecimento do escritor usando manuscritos em português não mudou muito, aumentando para 0,978947, mas melhorou a precisão para testes com japonês, passando a ser igual a 0,7754.
Observando os resultados obtidos anteriormente, foi possível perceber que quando a pasta selecionada para teste era a pasta com apenas amostras de katakana, os índices apresentados foram muito piores. Este fato pode estar relacionado ao desequilíbrio no número de amostras de cada silabário, enquanto quatro textos eram escritos maioritariamente em Hiragana apenas um texto era totalmente em Katakana. Para validar essa suspeita, uma variação experimental foi proposta. Neste novo experimento, todas as amostras de Katakana foram removidas do teste e o número de folds utilizados diminuiu para quatro. A melhor taxa de acerto obtida nesta variação foi de 0,8837 e 0,043860 de desvio padrão, utilizando a divisão de quatro blocos e a regra de fusão tardia do produto, além da F-Measure média igual a 0,6380. Esses resultados mostraram que a suspeita levantada anteriormente era consistente.

Identificação de silabário - Nesta Subseção discutimos os procedimentos adotados para classificar os manuscritos em três classes de acordo com seu silabário, ou seja, Hiragana, Katakana ou Romano (Português). Para a realização desses testes é importante notar que a classe referente aos Hiragana foi composta pelos manuscritos em Hiragana e pelos manuscritos que usaram Hiragana com Katakana. Para os experimentos com quatro e cinco fold o resultado apresentado não foi satisfatório, pois o classificador sempre previu a mesma classe (Romano), para todas as amostras, esse resultado provavelmente está ligado ao desequilíbrio da quantidade de amostras por classe. Com base nessa hipótese, a ideia foi reduzir a quantidade de amostras para a classe romana, para isto, três das cinco amostras dessa classe foram selecionados aleatoriamente, tendo o cuidado de selecionar as mesmas amostras para todos os escritores, de modo a não participarem desta etapa. Ao aplicar este método, os 570 manuscritos foram reduzidos a 399, estes por sua vez foram divididos em 3 folds, ou seja, cada fold possui 133 amostras, sendo este conjunto composto por todos os 7 manuscritos de cada um dos 19 participantes por fold.
Usando três folds e dividindo cada imagem em quatro blocos, o teste com melhores valores de acurácia e F-Measure foi aquele que não utilizou os manuscritos 6, 7 e 8 de cada escritor. O melhor valor de precisão, igual a 1 e com desvio padrão igual a 0, foi obtido pela regra da Soma e a F-Measure média obtida nesta configuração foi de 0,973496.

Conclusões

Neste trabalho, apresentamos a JKBP, uma base de dados disponibilizada para apoiar o desenvolvimento de tarefas de reconhecimento baseadas em manuscritos. A base é composta manuscritos copiados de dez textos, cinco em português brasileiro e cinco em japonês, obtidos de 57 voluntários.
Em trabalhos futuros, pretendemos expandir a JKBP, aumentando o número de colaboradores e introduzindo amostras com kanji e mais amostras com Katakana. Também pretendemos realizar novos experimentos, desta vez fazendo uso da dissimilaridade para identificar o autor no cenário multi-script e fazendo testes com outros classificadores.

Agradecimentos

Agradecemos à professora Kiyomi Kimura Fugie, do Instituto de Estudos Japoneses (IEJ) da UEM, pelo auxílio na elaboração e revisão dos textos, ao IEJ e à Escola de Língua Japonesa do Maringá, vinculada à Associação Cultural e Esportiva de Maringá (ACEMA), por ceder espaço e alunos para a coleta de manuscritos, ao PET-Informática da UEM, pelo apoio financeiro.

Referências

CHANG, Chih-Chung; LIN, Chih-Jen. **LIBSVM: A library for support vector machines**. ACM transactions on intelligent systems and technology (TIST), v. 2, n. 3, p. 1-27, 2011.

PINHELLI, Fabio et al. **Single-sample writers-" Document Filter" and their impacts on writer identification**. arXiv preprint arXiv:2005.08424, 2020.

BAY, Herbert; TUYTELAARS, Tinne; VAN GOOL, Luc. **Surf: Speeded up robust features**. In: European conference on computer vision. Springer, Berlin, Heidelberg, 2006. p. 404-417.

SOUZA, Marcos Roberto et al. **Offline Handwritten Script Recognition Based on Texture Descriptors**. In: 2019 International Conference on Systems, Signals and Image Processing (IWSSIP). IEEE, 2019. p. 57-62.