Detecção de padrões de legendas em imagens de ritmo visual a partir do detector de Harris

Guilherme Polo[1], **Miguel Gaiowski**[1]

[1]Instituto de Computação – Universidade Estadual de Campinas (UNICAMP)
Caixa Postal 6176 – 13083-852 – Campinas – SP – Brazil

{ggpolo,miggaiowski}@gmail.com

Resumo. *Este trabalho faz, inicialmente, um resumo do detector de cantos e bordas de Harris. Com esse detector partimos para a extração de legendas em imagens de ritmo visual. Descrevemos sete etapas complementares à Harris que fornecem resultados razoáveis para esse tipo de imagem. Em especial, aquelas onde o contraste entre regiões de legenda e demais partes da imagem são mais perceptíveis foi onde nosso trabalho apresentou melhores resultados.*

1. Introdução

A [Harris and Stephens 1988]

B

C

D

E

F

G

2. O detector de Harris

A partir de problemas existentes no detector de Moravec, definido em [Moravec 1980], surge o detector de Harris [Harris and Stephens 1988]. A ideia básica desse detector é conseguir indicar, ao mesmo tempo, pontos de canto e de borda com o cálculo de uma matriz que fornece informações locais em cada pixel da imagem.

Para definir essa matriz utilizada em Harris, primeiro consideramos uma imagem I e uma máscara w como dadas. A máscara pode ser calculada como $w_{u,v} = \exp -\frac{u^2+v^2}{2\sigma^2}$ – uma gaussiana. Com isso, essa matriz é inicialmente definida como:

$$M = \begin{bmatrix} A & C \\ C & B \end{bmatrix}$$

com

$$\begin{aligned} X &= I * [-1, 0, 1] \\ Y &= I * [-1, 0, 1]^\intercal \\ A &= (X \circ X) * w \\ B &= (Y \circ Y) * w \\ C &= (X \circ Y) * w \end{aligned}$$

sendo $f * h$ a representação da convolução de f com h e $M_1 \circ M_2$ sendo o produto de Hadamard que realiza a multiplicação ponto a ponto entre as matrizes M_1 e M_2.

As matrizes X e Y descrevem, respectivamente, uma aproximação para o gradiente em x e em y da imagem I. De acordo com [Harris and Stephens 1988], os autovalores α e β da matriz M – considerando um pixel da imagem – capturam a descrição de cantos (α e β são grandes), bordas (α é grande e β é pequeno, ou vice-versa) e regiões planas (α e β são pequenos).

Harris e Stephens [1988], então, procedem para a construção de uma matriz R:

$$\begin{aligned} Tr &= \alpha + \beta = A + B \\ Det &= \alpha\beta = A \circ B - C \circ C \\ R &= Det - k(Tr \circ Tr) \end{aligned} \tag{1}$$

A matriz formada na Equação 1 define todos os pontos de interesse que esse detector consegue capturar. O valor de k não foi estabelecido no trabalho original [Harris and Stephens 1988], mas trabalhos variados [Schmid et al. 2000, Orguner and Gustafsson 2007] mencionam valores entre $0,04$ e $0,06$. Neste trabalho fixamos $k = 0,05$. Nessa matriz, valores positivos são tomados como cantos, os negativos como bordas. Na figura 1 há uma representação visual da matriz R.

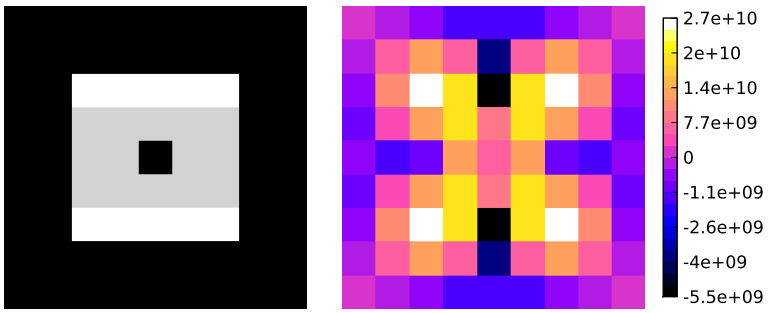

Figura 1: Uma imagem 9x9 e sua respectiva matriz R representada como um *heatmap*

É possível distinguir facilmente os pontos definidos como borda e como cantos no *heatmap* da Figura 1. Os dois pontos pretos são borda, enquanto que os quatro brancos são cantos. Apesar de haver diversos outros pontos positivos e negativos naquela figura, Harris ainda faz, conforme apresentado a seguir, duas considerações que levam a esse conjunto de apenas seis pontos.

As Equações 2 e 3 definem, respectivamente, os conjuntos de pontos de canto e de borda com base na matriz R. As coordenadas x e y que não pertencem a R são tomadas como não definidas.

$$\{(x,y) \mid \forall R[x,y] > 0 \bullet R[x,y] = \max(R[x-1:x+1, y-1:y+1)\} \quad (2)$$

$$\{(x,y) \mid \forall R[x,y] < 0 \bullet$$
$$R[x,y] < \begin{cases} min(R[x-1,y], R[x+1,y]) & \text{se } A[x,y] > B[x,y] \\ min(R[x,y-1], R[x,y+1]) & \text{c.c.} \end{cases} \quad (3)$$
$$\}$$

Na Equação 2, os pontos de canto que não são máximos em sua vizinhança direta são descartados. No caso dos pontos de borda, a Equação 3 indica que só serão bordas aqueles pontos que forem mínimos onde seu gradiente é máximo – resultando em bordas finas.

3. Implementação

Com a matriz R (Equação 1) construída e os pontos de canto (Equação 2) e borda (Equação 3) definidos, realizamos mais sete etapas na expectativa de definir regiões de legenda em imagens de ritmo visual. Na primeira, a partir de uma percentagem do maior valor em R descartamos os pontos de canto inferiores a este valor. Na segunda, definimos um limiar inferior e outro superior – também com base numa percentagem do maior valor absoluto de ponto de borda – para os pontos de borda e classificamos cada borda de acordo com estes limiares para, então, realizar a histerese [Nixon and Aguado 2008]. Estas duas primeiras etapas são realizadas em cada banda da imagem e os resultados são combinados numa única imagem. A imagem resultante até este passo é formada pelas bordas e cantos escolhidas, com os cantos sendo representados como quadrados 3x3 centrados no ponto de canto real.

A partir deste ponto, são feitas tentativas para completar retângulos quase fechados que possivelmente representam regiões de legenda nas imagens de ritmo visual. Para cada ponto ainda não preenchido, é verificado se ambos seus vizinhos imediatos na horizontal ou na vertical foram preenchidos. Caso ambos em pelo menos uma das direções tenham sido, este ponto não preenchido é marcado para preenchimento no final deste processo.

Em seguida, o algoritmo *flood fill* é aplicado para as direções horizontais e verticais partindo do ponto $(0,0)$. Nesta implementação, este ponto, garantidamente, é inicialmente preto devido a não consideração dos pontos em R, para construir as imagens até este momento, onde a máscara gaussiana escolhida não esta inteiramente contida na imagem (considerando a convolução no domínio espacial).

Na quinta etapa, os pontos que permaneceram pretos são aqueles contidos dentro de formas fechadas. Com isso, uma nova imagem é criada para apresentar aqueles pontos pretos agora como brancos. Todos os pontos brancos nesta imagem formam as possíveis regiões de legenda. Em seguida é feita uma filtragem da mediana para eliminar pequenas regiões que permaneceram devido a grande heterogeneidade das imagens de ritmo visual.

Finalmente, na sétima etapa, um processamento um tanto quanto heurístico é feito. Cada região branca da imagem é considerada uma componente conexa e é feita, então, a contagem de pixels em cada uma dessas componentes. Aquelas componentes de tama-

nho inferior a um valor estabelecido são desconsideradas na imagem final. Isso acaba limpando muitos fragmentos que sobreviveram até aqui.

4. Avaliação

Para realizar a avaliação, os parâmetros foram fixados para todos os testes da maneira seguinte: máscara gaussiana 3x3 com $\sigma = 2$, limiar para pontos de canto = $0.5(\frac{max(R)}{100})$, limiares $H = 2(\frac{|min(R)|}{100})$ e $L = 0.01(\frac{|min(R)|}{100})$ para histerese, máscara 5x5 para filtragem da mediana e descarte de componentes de tamanho inferior a 20 pixels.

De modo geral, o detector de Harris, com os demais passos aplicados, foi capaz de encontrar regiões de legenda quando estas apresentavam um contraste adequado na imagem como um todo. As Figuras 2 e 3 apresentam um resultado considerado, por nós, ruim e bom, respectivamente.

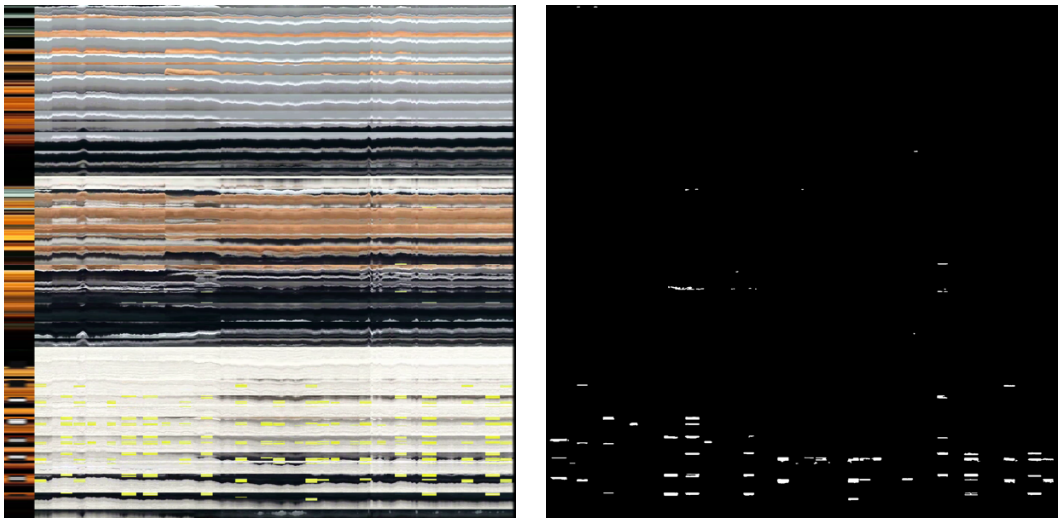

Figura 2: Muitas regiões de legenda perdidas

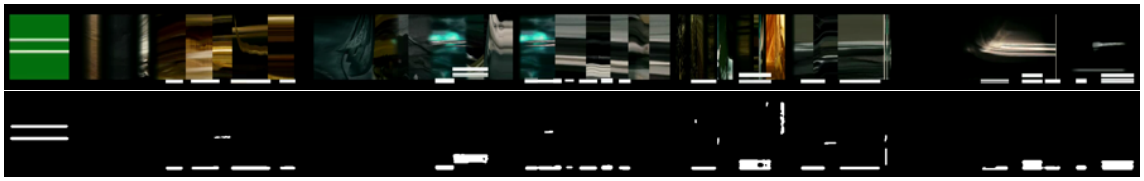

Figura 3: Grande parte das regiões de legenda encontradas; alguns erros

Na Figura 2, acreditamos que cerca de seu um terço superior não apresente legendas e nosso método detecta poucas regiões nessa área. Na última parte desta imagem é onde grande parte das legendas estão, mas a taxa de acerto foi baixa. Por outro lado, a Figura 3 apresenta o caso em que nosso método tem bons resultados. Apesar de reportar incorretamente algumas partes como sendo de legenda, todas as reais regiões de legenda foram quase completamente indicadas.

A Figura 4 apresenta uma imagem de ritmo visual significantemente maior que aquela da Figura 3. Ela suporta nossa suposição em relação ao contraste entre legendas e imagem como um todo.

Figura 4: Maioria das regiões de legendas foram reportadas

5. Conclusão

Um dos pontos a ser discutido é o fato de que, em certas imagens testadas, nem mesmo humanos (nós autores) conseguem identificar com absoluta certeza a posição das legendas. É difícil esperar de um algoritmo um trabalho que sequer há certeza sobre os resultados esperados. Mesmo assim, em vários casos foi possível guiar o algoritmo através de ajustes de parâmetros e consertos de detalhes para que chegasse em resultados aparentemente satisfatórios.

Como detector de bordas, Harris é um bom detector de cantos. Para a tarefa de encontrar legendas nas imagens de ritmo visual ele serviu apenas como um passo 0, sendo discutível se os resultados seguintes não teriam sido melhores com uso de outro detector. Uma quantidade razoável de algoritmos heurísticos foram implementados em conjunto para produzir algum resultado parcialmente adequado.

O número de linhas de código para a implementação do detector de Harris é mínima quando comparada ao resto do programa. De fato, a maior parte do trabalho foi feita por processos posteriores. Além disso, muitos parâmetros precisam ser definidos de forma quase arbitrária.

Referências

Harris, C. and Stephens, M. (1988). A Combined Corner and Edge Detector. In *Alvey Vision Conference*, pages 147 – 151.

Moravec, H. (1980). Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover. In *Technical Report CMU-RI-TR-80-03, Robotics Institute, Carnegie Mellon University & doctoral dissertaion, Stanford University*.

Nixon, M. and Aguado, A. S. (2008). *Feature Extraction & Image Processing*. Academic Press.

Orguner, U. and Gustafsson, F. (2007). Statistical Characteristics of Harris Corner Detector. In *Statistical Signal Processing*, pages 571–575.

Schmid, C., Mohr, R., and Bacukhage, C. (2000). Evaluation of Interest Point Detectors. *International Journal of Computer Vision*, 37.