UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

Fluxo óptico para análise de calçadas

Renzo Real Machado Filho

Orientadores:
Roberto Marcondes Cesar Junior
Roberto Hirata Junior

Chamada FAPESP Futuros Cientistas

São Paulo
October 1, 2025

Resumo

Contents

1 Introdução

2 Visão Computacional: Algoritmos e Aplicações [?]

2.1 Processamento de Imagens

Nesse capítulo, trataremos do conjunto de operações e técnicas aplicadas a imagens digitais. É uma etapa de preparação para muitos algoritmos de reconhecimento de objetos, reconstrução 3D e/ou rastreamento de movimento. Nessa perspectiva, começaremos pelos chamados "operadores de ponto", que manipulam cada pixel de uma imagem, independentemente daqueles ao seu redor. Depois, passaremos pelos operadores "area-based", nos quais cada novo valor de um pixel depende de um certo número de pontos vizinhos.

Um operador genérico é uma função que toma uma ou mais imagens como inputs e produz uma outra imagem de output. Matematicamente,

$$f(x) = h(g_0(x), \ldots, g_n(x))$$

Já uma imagem digital colorida é representado como uma matriz 3D (altura $\times$ largura $\times$ 3 canais). Os 3 canais são: Blue, Green, Red (em OpenCV é BGR).

2.2 Operadores de Ponto

2.2.1 Transformações de Pixels

Ajuste de Brilho (Brightness) e Contraste (Contrast)

Tipo: Transformação linear.
Fórmula: $f(x) = a(x) \cdot g(x) + b(x)$, onde a é dito o contraste e b, o brilho.
Efeito: O brilho desloca uniformemente todos os valores de pixel na imagem. Valores positivos clareiam a imagem, valores negativos escurecem. Enquanto isso, o contraste controla a diferença entre tons claros e escuros. Valores > 1 expandem a faixa tonal (aumentam o contraste), valores entre 0 e 1 comprimem a faixa tonal (reduzem o contraste).
Quando usar: Quando a imagem está muito escura ou muito clara globalmente ou quando a imagem está "chapada", i.e, sem muita variação tonal.

Correção Gamma

Tipo: Transformação não-linear.
Fórmula: $f(x) = g(x)^{1/\gamma}$.

O que é Gamma (γ)? É um parâmetro que define a curvatura da transformação não-linear aplicada aos valores de intensidade. Ele controla como os valores intermediários (tons médios) são mapeados, enquanto preserva os extremos (preto puro e branco puro). Por padrão, usa-se $\gamma \approx 2.2$.
Efeito: Se $\gamma > 1$, escurece os tons médios enquanto preserva pretos e brancos (aumenta contraste em tons escuros). Já $\gamma < 1$, clareia os tons médios enquanto preserva pretos e brancos (aumenta contraste em tons claros).
Quando usar: Para corrigir percepção visual ou problemas de iluminação não-lineares.

Ordem das Operações

Contraste/Brilho $\rightarrow$ Gamma

O Contraste/Brilho são lineares, trabalham no domínio da intensidade enquanto o Gamma é não-linear, trabalha no domínio perceptual. Se fizéssemos gamma primeiro a transformação linear posterior distorceria a curva gamma e os valores seriam re-escalados de forma inadequada, perdendo o controle preciso sobre a correção tonal.

2.2.2 Color Balance

Esse procedimento ajusta a intensidade relativa das cores primárias. A operação é feita por canal (R, G, B). Podemos multiplicá-los por um fator que altera seu brilho ou ainda operar sobre transformações mais complexas, como o mapeamento no espaço de cores XYZ.

Espaço de Cores XYZ

É um espaço de cor matematicamente definido, usando coordenadas tridimensionais (X, Y, Z) para descrever todas as cores visíveis ao olho humano.

- **X:** Representa aproximadamente a sensibilidade ao vermelho
- **Y:** Representa o brilho luminoso (luminância)
- **Z:** Representa aproximadamente a sensibilidade ao azul

2.2.3 Composição e Mascaramento (Compositing and Matting)

Em muitos aplicativos de edição de fotos e de efeitos visuais, queremos inserir/combinar elementos em uma imagem. Esse processo é chamado de **composição** [**?**].

Paralelo a isso, também é desejável extrair objetos de imagens, processo comumente chamado de **mascaramento**. Uma máscara define a área de uma imagem que deve ser mantida ou ignorada, permitindo que apenas certas partes da imagem sejam visíveis e sejam integradas com outros elementos [**?**, **?**].

Alpha matting

É um processo que visa estimar a translucidez de um objeto em uma determinada imagem. O "alpha matting" resultante descreve, em pixels, a quantidade de cores de primeiro e segundo plano que contribuem para a cor da imagem composta. [**?**]

Alpha-Matted Color Image

É uma imagem que além dos 3 canais de cor (RGB), possui um 4º canal intermediário (Alpha - α) que representa:

- $\alpha = 1$: Pixel totalmente opaco
- $\alpha = 0$: Pixel totalmente transparente
- $0 < \alpha < 1$: Pixel parcialmente transparente

Portanto, para compor uma nova imagem sobre uma imagem antiga, o *over operator* é

$$C = (1 - \alpha)B + \alpha F$$

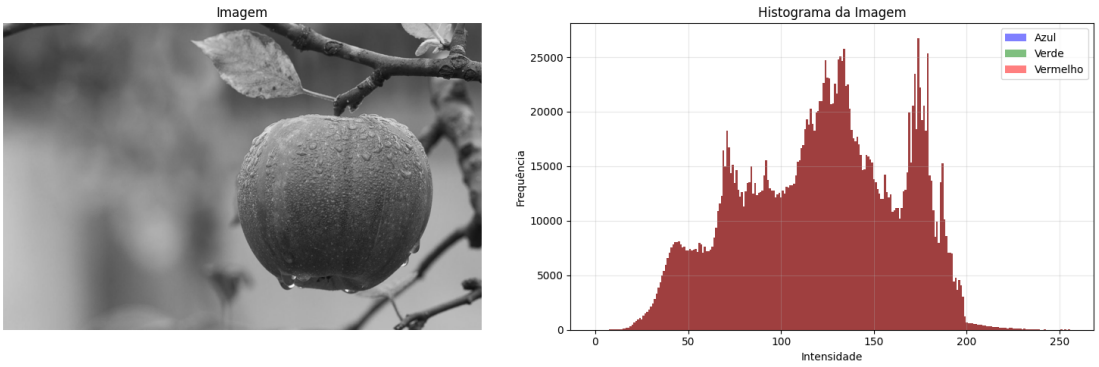
A equação acima atenua a influência da imagem de fundo B por um fator $(1-\alpha)$ e, em seguida, adiciona os valores de cor (e opacidade) correspondentes à camada de primeiro plano F.

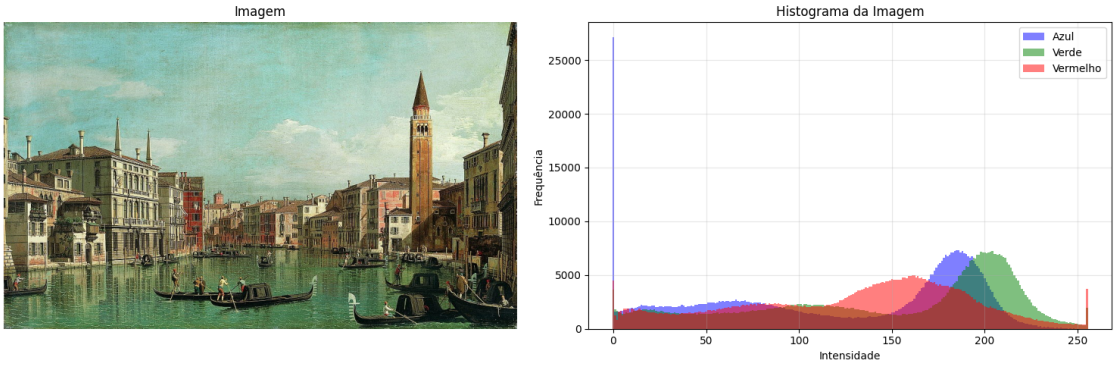
2.2.4 Equalização de Histograma

O histograma de uma imagem é um gráfico que representa a frequência de cada nível de intensidade de cinza (ou cor) presente na imagem. Em uma imagem de 8 bits em escala de cinza, por exemplo, existem 256 níveis de intensidade, que vão de 0 (preto absoluto) a 255 (branco absoluto). O eixo horizontal do histograma representa esses níveis de intensidade, enquanto o eixo vertical indica o número de pixels que possuem cada uma dessas intensidades.

Imagens com baixo contraste tendem a ter seus histogramas concentrados em uma faixa estreita de valores. Por exemplo, uma imagem escura terá a maioria de seus

pixels com valores de intensidade baixos, resultando em um histograma "amontoado" à esquerda. De forma análoga, uma imagem muito clara terá seu histograma concentrado à direita.





A equalização de histograma atua justamente nesse cenário, redistribuindo igualmente esses valores de intensidade por toda a gama possível. O objetivo é transformar o histograma original em um histograma mais próximo de uma distribuição uniforme, onde cada nível de intensidade tenha, idealmente, o mesmo número de pixels. Ao fazer isso, a diferença de intensidade entre os pixels é acentuada, o que melhora significativamente o contraste global da imagem.

Para isso, computamos a Função de Distribuição Acumulada (CDF), denotada por $c(I)$

$$c(I) = \frac{1}{N}\sum_{i=0}^{I} h(i) = c(I-1) + \frac{1}{N}h(I),$$

onde I é o nível de intensidade atual (0-255 para imagens 8-bit), $h(I)$ é a frequência do nível de intensidade I (valor do histograma) e N é quantidade de pixels na imagem.

3 Fluxo Óptico

O fluxo óptico é definido como a distribuição das velocidades aparentes do movimento dos padrões de brilho em uma imagem. [**?**].

3.1 Campo de Movimento (Motion Field)

Trata-se da velocidade de um ponto que se move na cena/imagem. Em muitos casos, esse conceito se confunde com o fluxo óptico.

3.2 *Optical Flow Constraint Equation*

Suponha que temos um ponto (x, y) numa imagem que se moveu para $(x + \delta x, y + \delta y)$. O seu deslocamento é $(\delta x, \delta y)$ e queremos medir esse movimento. Dizemos que o fluxo óptico é $(u, v) = (\frac{\delta x}{\delta t}, \frac{\delta y}{\delta t})$ num tempo δt, pequeno.

Suposição 1: O brilho de um ponto é constante ao longo do tempo. Então,

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t).$$

Suposição 2: O deslocamento $(\delta x, \delta y)$ e o passo δt são pequenos.

Com isso, por Taylor, temos

$$f(x + \delta x) = f(x) + \frac{\partial f}{\partial x}\delta x + \cdots + \frac{\partial^n f}{\partial x^n}\frac{\delta x^n}{n!}$$

Se $\delta x \to \infty$, então $f(x + \delta x) = f(x) + \frac{\partial f}{\partial x}\delta x + O(\delta x^2)$.
Para $\delta x, \delta y$ e δt suficientemente pequenos, vale que

$$f(x + \delta x, y + \delta y, t + \delta t) \approx f(x, y, t) + \frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial y}\delta y + \frac{\partial f}{\partial t}\delta t$$

Pela *suposição 2*, temos

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t) + I_x\delta x + I_y\delta y + I_t\delta t$$

Pela *suposição 1* e subtraindo ambas as equações, obtemos

$$I_x\delta x + I_y\delta y + I_t\delta t = 0$$

Dividindo por δt e com $\delta t \to 0$, temos

$$I_x \frac{\delta x}{\delta t} + I_y \frac{\delta y}{\delta t} + I_t = 0$$
$$\Leftrightarrow I_x u + I_y v + I_t = 0,$$

onde (u, v) é o fluxo óptico.

3.2.1 Problema da Abertura

3.3 O método de Lucas-Kanade

Vimos que, pelo resultado anterior, a equação de restrição de fluxo óptico é um problema mal posto. Assim, é necessário mais uma suposição.

Suposição 3: para cada pixel, assuma que o campo de movimento e o fluxo óptico são constantes em uma pequena vizinhança W.

Para quaisquer $(l, k) \in W$, vale

$$I_x(l,k)u + I_y(l,k)v + I_t(l,k) = 0$$

Isso resulta em um sistema de equações. Logo, na forma matricial para $W_{n\times n}$, temos

$$\underbrace{\begin{bmatrix} I_x(1,1) & I_y(1,1) \\ \vdots & \vdots \\ I_x(n,n) & I_y(n,n) \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} u \\ v \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} I_t(1,1) \\ \vdots \\ I_t(n,n) \end{bmatrix}}_{B}$$

Esse é um clássico problema do tipo $Ax = B$. Portanto, podemos resolvê-lo por mínimos quadrados, obtendo

$$x = (A^T A)^{-1} A^T B.$$

3.3.1 Condições

- $A^T A$ deve ser inversível
- $A^T A$ deve estar bem condicionada:
  - Sejam λ_1, λ_2 os autovalores de $A^T A$. Então, $|\frac{\lambda_1}{\lambda_2}| \approx 1$ e $\lambda_1, \lambda_2 > \epsilon$.

As seguintes situações descrevem situações comuns:

Regiões Homogêneas

Regiões sem muita textura dificultam a visualização de movimento. Matematicamente, $\lambda_1 \approx \lambda_2$, mas ambos são pequenos.

Bordas

Regiões com presença de bordas geram um problema mal condicionado. Os gradientes estão predominantemente em um direção, i.e, a variação significativa está em uma única direção. Assim, o resultado da estimação de fluxo óptico pode ser ambíguo. Matematicamente, $\lambda_1 >>> \lambda_2$.

Regiões Texturizadas

Nesse cenário, a estimativa é muito boa. Há muitas variações nos padrões de brilho. Matematicamente, $\lambda_1 \approx \lambda_2$ e $\lambda_1, \lambda_2 > \epsilon$.

3.4 Coarse-To-Fine Flow Estimation (Estimação de Fluxo do "Grosso ao Fino")

Quando temos uma sequência de imagens com grandes movimentos, a aproximação por série de Taylor não é mais válida, bem como suas suposições.

4 State-of-Art

A estimação de movimento nas cenas pode ser dividido em três métodos: knowledge-driven, data-driven e hybrid-driven.

4.1 knowledge-driven

É baseado em suposições para restringir o processo de estimação e modelar a correspondência espacial-temporal. Os primeiros métodos são Horn e Schunck (HS) e Lucas e Kanade (LK).

4.1.1 Desafios

Grandes Deslocamentos

Objetos que se movem em rapidamente producem grandes distâncias entre os pixels nas cenas, o que dificulta o cálculo. Para lidar com esse problema, é usual utilizar

uma estratégia "fino ao grosso" (coarse-to-fine).

Oclusão

A oclusão é o processo de detectar e/ou rastrear objetos quando estão ocultos ou obstruídos numa cena. Os métodos mais comuns adotam um esquema de verificação de consistência.

Variações de Iluminação

A clássica suposição de constância do brilho não se aplica em contextos de mudanças de iluminação. ???

Ruído

Filtros são aplicados em um campo para reduzir o ruído e melhorar a qualidade do mapeamento do fluxo.

4.2 Data-driven

As técnicas mais dominantes utilizam o deep learning. As CNNs tem sido muito bem sucedidas para a estimação de fluxo óptico. Por sua vez, esses métodos se dividem em U-Net e em Rede de Pirâmide Espacial.

Comumente, os métodos data-driven usam um ground truth como um sinal supervisionado, o que depende de dados rotulados. Entretanto, coletar um ground truth denso a nivel de pixel para cenas reais é um processo complicado e longo.

4.2.1 U-Net

Utiliza a estrutura de encoder-decoder. Por causa da contração e expansão do mapeamento de features nos encoders e decoders, há perda de detalhes importantes, o que é prejudicial para a estimação densa de movimento e atrapalha a acurácia e o detalhamento do campo de fluxo.

4.2.2 Rede de Pirâmide Espacial

O SPyNet foi a primeira arquitetura a utilizar essa técnica, cuja principal vantagem é o tamanho reduzido do modelo. Entretanto, sua acurácia ainda não pe capaz de bater o FlowNet2.0. O SPyNet estima grandes movimentos na camada mais grossa e deforma a segunda imagem em direção ao primeiro usando o fluxo amostrado de um nível anterior.

Esse mecanismo tem a característica de ser muito "personalizável" para a estimação de fluxo óptico. Sua construção detém diversos princípios da área, como a pirâmede espacial, deformação (warping) e pós-processamento. Isso é capaz de aumentar muito a eficiência e a acurácia da técnica.

4.3 Hybrid driven

Tais métodos usam suposições téoricas e, também, dados não-rotulados para o treino.

4.3.1 U-Net e Rede de Pirâmide Espacial

Para contornar o problema dos dados, vários trabalhos tentam aprender de maneira não supervisionada baseando-se em suposições teóricas. Outros tentam trabalhar de forma semi-supervisionada. Todavia, ainda não foram capazes de obter acurácias melhores nos mesmos datasets públicos.

4.4 Métricas

4.4.1 End-to-End Point Error (EPE) and Average Endpoint Error (AEE)

O erro de ponto final (EPE) para um determinado pixel ou ponto é calculado como a distância euclidiana entre o vetor de fluxo óptico estimado e o vetor de fluxo óptico da verdade básica naquele local.

O erro médio do ponto final (AEE), que calcula a distância euclidiana entre o campo de fluxo calculado e o campo de fluxo da verdade básica. Com isso, é nada mais que a média entre todos os pontos/pixels do cálculo anterior.

$$\frac{1}{HW}\sum \underbrace{\sqrt{(u_* - u)^2 + (v_* - v)^2}}_{EPE} \ ,$$

onde HW é número total de pixels na imagem, (u_*, v_*) o ground truth e (u, v) a estimativa do fluxo óptico.

4.4.2 Average Angular Error (AAE)

Para cada pixel, o erro angular é o ângulo entre o vetor de fluxo estimado e o vetor de fluxo do ground truth. Isso mede a precisão direcional do movimento estimado.

O AAE é a média desses erros angulares individuais em todos os pixels da imagem ou em uma região específica de interesse. Ele fornece um único valor que indica a precisão direcional geral do algoritmo de fluxo óptico.

$$\frac{1}{HW}\sum \arccos\left(\frac{u_* u + v_* v + 1}{\sqrt{(u_*^2 + v_*^2 + 1)(u^2 + v^2 + 1)}}\right)$$

4.4.3 Root-Mean-Square Error (RMSE)

Quantifica a precisão da estimativa de movimento. O valor do RMSE é expresso em pixels por quadro (px/frame). Se o RMSE for 1.5, significa que, em média, a ponta do vetor estimado está errada por 1.5 pixels em relação à ponta do vetor correto.

Quanto menor, melhor. Um RMSE baixo indica que os vetores estimados estão muito próximos dos vetores reais. Isso significa que o algoritmo é preciso. Quanto maior, pior. Um RMSE alto indica que há uma grande discrepância entre o movimento estimado e o movimento real. O algoritmo tem baixa precisão.

$$\sqrt{1/HW \sum_{(x,y)} (I_{\text{warped}}(x,y) - I(x,y))^2} \ ,$$

onde $I(x, y)$ é a intensidade do pixel na posição (x, y) do segundo quadro (o quadro para o qual estamos tentando prever o movimento) e $I_{\text{warped}}(x, y)$ é a intensidade do pixel na posição (x, y) da imagem warpeada.

Limitações: Média pode esconder detalhes: Um RMSE baixo pode ser causado por uma performance excelente em 90% da imagem, mas um desempenho terrível em 10% (e.g., em regiões de oclusão). A média pode mascarar esses erros localizados.

Não captura erros de direção: O RMSE mede principalmente a magnitude do erro. Dois vetores que apontam para direções completamente opostas, mas com magnitudes similares, podem ter um erro de magnitude pequeno, embora o erro de direção seja catastrófico.

Depende do Ground Truth: A métrica só é válida se você tiver um ground truth confiável.

5 Escala de Cinza

O Cálculo da Escala de Cinza

Seu raciocínio de usar a média (R+G+B)/3 é perfeitamente lógico e seria a forma mais simples de fazer a conversão. No entanto, o método padrão é um pouco mais complexo, e o motivo é fascinante: a biologia do olho humano!

Nossos olhos não são igualmente sensíveis a todas as cores. Somos muito mais sensíveis à luz verde, um pouco menos à vermelha e bem menos à azul.

Para criar uma representação em escala de cinza que pareça "natural" para um ser humano em termos de brilho percebido, a conversão usa uma média ponderada. Os canais de cor que percebemos melhor têm um peso maior no cálculo final.

A fórmula padrão (usada pelo OpenCV e em muitas outras aplicações) é:

$Y = 0.299R + 0.587G + 0.114B$

Onde Y é o valor do pixel de luminância (o valor de cinza).

Veja como o Verde (G) tem um peso enorme (quase 60%), enquanto o Azul (B) tem um peso bem pequeno. Isso garante que uma imagem com muito verde pareça mais clara em escala de cinza do que uma imagem com muito azul, o que corresponde à nossa percepção.