# Descritor local baseado na transformação SIFT para segmentação de vídeos via grafos

Gustavo M. Q. de Mendonça e Ricardo L. de Queiroz

***Resumo*— Na segmentação de vídeos quadro a quadro, a manutenção da coerência temporal depende diretamente da qualidade do rastreamento de regiões ao longo do tempo. Uma vez representadas como vértices de um grafo, regiões convertidas em superpixels podem ser relacionadas de acordo com suas características. Neste trabalho, adaptamos para o domínio dos superpixels processados como grafos de regiões, princípios de um extrator de características bastante difundido, o SIFT. Um descritor é criado para cada vértice de um grafo, a partir de histogramas de orientação do gradiente de setores ao redor do vértice, calculado de forma a garantir invariância a escala, rotação e iluminação. Os resultados iniciais se mostram promissores, visto que as correspondências entre pares de imagens (intencionalmente transformadas) e quadros sequenciais de vídeos exibem razoável coerência.**

***Palavras-Chave*— Segmentação de vídeo, grafos, descritor local, SIFT.**

***Abstract*— In a frame to frame video segmentation, the temporal coherence maintenance depends directly on the quality of the regions tracking along the time. Once represented by vertices of a graph, regions converted into superpixels can be related by its features. In this work, we adapt to the domain of superpixels processed as region graphs, principles of widespread feature extractor, the SIFT. A descriptor is created to each vertex of graph, from orientation histograms of the gradient of bins around the vertex, calculated to ensure a scale, rotation and lighting invariance. The initial results show promise, as matched regions of image pairs (intentionally transformed) and sequential frames of a video display reasonable coherence.**

***Keywords*— Video segmentation, graph, local descritpor, SIFT.**

## I. INTRODUÇÃO

A segmentação de vídeos é um problema básico em visão computacional. Na restrição de regiões que se estendam espacialmente e temporalmente, em algum nível se faz necessário um agrupamento/rotulação não supervisionado. Em geral, esses agrupamentos utilizam relações de textura, cor e/ou movimento para serem construídos [1], relações entre vizinhanças que acabam se enquadrando em um problema de grafos. Os pixels de quadros sequenciais de um video podem ser analisados em unidades de volume (voxels), onde agrupamentos dão origem a supervoxels. Entretanto, a quantidade de dados gerados por um volume espaço-temporal de um vídeo demanda um grande esforço computacional [2, 3]. Uma forma de reduzir o esforço computacional nesse tipo de abordagem, é a aplicação de uma segmentação hierárquica utilizando grafos [4, 5]. Apesar de questionamentos quanto à manutenção de coerência temporal, dada a instabilidade de uma segmentação quadro a quadro [4], uma abordagem de correspondências entre superpixels (conjunto de pixels de uma imagem) é vastamente empregada [1, 6, 7, 8, 9].

O casamento quadro a quadro de superpixels tem como base o confronto entre as características de aparência (cor, textura) e posição dessas regiões ao longo do tempo. Contudo, essa base é limitada em informação. Isto é, duas regiões pertencentes a dois quadros consecutivos apresentando cor e posição muito próximas (senão iguais) não representam necessariamente um mesmo objeto. Propriedades mais discriminantes se fazem necessárias para a manutenção da coerência temporal.

A Transformação de Características Invariante à Escala (SIFT) [10] é um algoritmo bastante difundido em visão computacional, ao se mostrar eficiente na determinação de pontos correspondentes entre imagens confrontadas. A aplicação da SIFT em processamento de vídeos não se restringe apenas ao rastreamento de objetos [11], podendo ser aplicada à estabilização de vídeos [12]. A utilização do histograma associado às regiões e de um descritor produzido pela SIFT, ou semelhante, exibe uma melhora de desempenho na segmentação de vídeos [4, 13, 14]. Entretanto, o cálculo dos descritores é realizado nas imagens construídas por pixels isolados. Tal abordagem não aproveita a simplificação das imagens enquanto representadas por grafos de regiões e nem a redução do esforço computacional associado a essa simplificação.

Este trabalho tem como objetivo a construção de um descritor local que opere diretamente no domínio dos grafos de região, criando associações (pesos de ligação) entre superpixels de quadros consecutivos, as quais podem definir, com razoável grau de coerência, regiões correspondentes. Primeiramente, uma transformada *watershed*, acompanhada de uma proposta de algoritmo de agrupamento, efetua uma sobresegmentação das imagens, criando os superpixels. Em seguida, as definições de gradiente são adaptadas ao domínio dos grafos de região. Uma introdução aos conceitos básicos de Grafos e da SIFT é apresentada nos parágrafos seguintes, englobando as notações adotadas neste trabalho.

Um grafo $G = (V, E)$ consiste em um par onde $V$ e $E$ são ambos conjuntos finitos de elementos. Os elementos $v \in V$ são chamados vértices (nós) e os elementos $e \in E \subset \{\{i, j\}, i, j \in V, i \neq j\}$ são chamados de arestas. Se uma aresta $e_{i,j}$ conecta $i$ a $j$, então estes são adjacentes, ou seja, $i$ é vizinho de $j$ (vice e versa). No caso de um triplete $G = (V, E, W)$ (grafo ponderado), $w_{i,j} \in W$ determina a força de ligação de uma aresta $e_{i,j}$ [15].

Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília-DF, Brasil, E-mails: gustavo@image.unb.br, queiroz@ieee.org. Este trabalho recebeu apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Uma imagem digital $F(x,y)$ é um tipo especial de grafo (grafo de adjacência de pixels ou grafo de pixels), com vizinhança bem definida em grade retangular. Os vértices $v$ desse grafo são representados pelos pixels, que estão associados a uma posição $\vec{\mathbf{r}}_v = (x_v, y_v)$, e uma função de intensidades $F_v = F(x_v, y_v)$.

Nesse contexto, um conjunto de pixels (sub-grafo de pixels) pode ser substituído por uma região homogênea (superpixel), promovendo uma segmentação em baixo nível [16]. A cada subgrafo $V'$, substituído por um vértice $v'$, associa-se uma posição $\vec{\mathbf{r}}_{v'}$ e uma função de intensidade $L_{v'}$, calculados com base na média das posições e intensidades dos pixels de $V'$, e uma área $A_{v'}$ (número de pixels contidos em $V'$). Cria-se então um grafo de adjacência de regiões (grafo de regiões) a partir de um grafo de pixels.

A transformação SIFT [10] segue 4 passos: (1) detecção e seleção de extremos em um espaço de escalas; (2) localização de pontos-chave; (3) definição da orientação e magnitude dos pontos-chave; e (4) criação de um descritor para os pontos-chave.

No primeiro passo, as escalas representam um conjunto de imagens oriundas da convolução da imagem da $I(x,y)$ com Gaussianas $G(x,y,k\sigma)$ de diferentes desvios padrão, determinados por um fator de escala $k$:

$$L(x,y,k\sigma) = I(x,y) * G(x,y,k\sigma). \tag{1}$$

Esse procedimento simula redimensionamentos consecutivos na imagem para a seleção de pontos que se preservam a essas mudanças de escala, pontos-chave. A diferenças entre as imagens nos espaços escalas e algoritmos de seleção determinam pontos-chave e suas posições.

A definição da orientação e magnitude dos pontos-chave, terceiro passo, é realizada na respectiva imagem do espaço de escalas do ponto-chave. Mapas de gradiente são criados, dos quais as projeções horizontal e vertical, $m_x = L(x+1,y) - L(x-1,y)$ e $m_y = L(x,y+1) - L(x,y-1)$, respectivamente, geram uma magnitude $m(x,y) = \sqrt{m_x^2 + m_y^2}$ e uma orientação $\theta(x,y) = \tan^{-1}(m_y/m_x)$. Pode-se representar essas propriedades em um vetor $\vec{\mathbf{m}}_i$ para cada pixel por meio de um grafo:

$$\vec{\mathbf{m}}_i = \sum_{j \in V_i^4} L_j \vec{\mathbf{u}}_{i,j}, \tag{2}$$

onde $V_i^4$ é a vizinhança 4-conectividade do pixel/vértice $i$, $L_i$ o valor de sua intensidade e $\vec{\mathbf{u}}_{i,j} = \frac{\vec{\mathbf{r}}_j - \vec{\mathbf{r}}_i}{||\vec{\mathbf{r}}_j - \vec{\mathbf{r}}_i||}$ é o vetor unitário que define a direção entre $i$ e $j$, neste caso, o conjunto $[(\pm 1, 0), (0, \pm 1)]$.

Um histograma de orientações é criado para uma região ao redor do ponto-chave. Cada amostra adicionada ao histograma é ponderada pela magnitude do gradiente e por uma Gaussiana circular simétrica em relação à localização do ponto-chave. A partir de um limiar, até três picos de orientações e suas respectivas magnitudes são relacionados à localização do ponto.

Novos histogramas são calculado em setores (*bins*) ao redor do ponto-chave, com as orientações dos gradientes referenciadas pela orientação do ponto-chave, o que gera invariância à rotação. Os histogramas, com elementos também ponderados pelas magnitudes dos gradientes e pela janela Gaussiana, são dispostos em um único vetor, definido como descritor do ponto-chave. Esse descritor é normalizado em valores entre 0 e 1, objetivando a invariância às mudanças na iluminação.

## II. Método Proposto

### A. Determinação do grafo de regiões

A transformada de um grafo de pixels para um grafo de regiões é feita por meio da *watershed*. Aplica-se a *watershed* a imagem $U(x,y)$ (8 bits), que é o módulo do gradiente de uma imagem $F(x,y)$ (normalizada entre 0 e 1), com componentes horizontais $D_x(x,y)$ e verticais $D_y(x,y)$ aproximadas pela convolução de $F(x,y)$ com filtros Sobel detectores de bordas (horizontais e verticais) [17], o módulo então é elevado a uma potência $p$:

$$U(x,y) = ||\big(D_x(x,y), D_y(x,y)\big)||^p. \tag{3}$$

Visto que a sobre-segmentação oferecida pela *watershed* pode gerar uma quantidade indesejada de vértices, quando aplicada diretamente ao módulo das componentes, ou seja, $p = 1$ (Fig. 2(b)), a potência $p$ é ajustada de forma que se possa estipular um limiar mínimo para as represas formadas pela detecção de bordas [17]. As Fig. 1 (b) à (f) exibem o controle da quantidade de nós $N$ por meio da potência $p$.

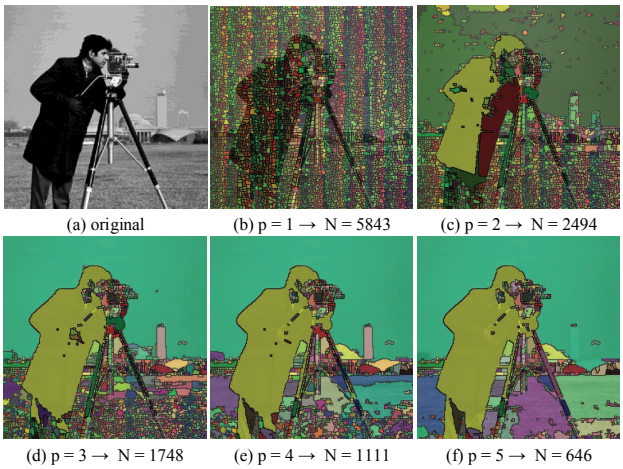



Fig. 1. *Watershed* aplicada à imagem exibida em (a), para diferentes valores da potência $p$ (Equação (3)) relacionadas de (b) a (f). Nota-se uma gradual diminuição no número de superpixels $N$ e a exclusão de algumas estruturas (torres, por exmeplo) de acordo com o aumento de $p$ .

### B. Agrupamento por escalas

Apesar de promover uma redução na quantidade de regiões (vértices) no grafo, um aumento considerável do termo $p$, equação (3), elimina regiões de baixo contraste possivelmente relevantes. Foi, então, implementado um algoritmo de agrupamento, o qual aproveita os superpixels formados pela transformada *watershed* em diferentes escalas da imagem original.

Os superpixels/vértices determinados por uma *watershed* da imagem original, processada com os devidos parâmetros, são agrupados de acordo com a posição de seu centro de

massa em relação aos superpixels de uma *watershed* da imagem original reduzida por um fator $k$ e processada com os mesmos parâmetros. Para melhor definição das bordas, esse procedimento pode ser repetido $m$ vezes por um fator de redução $k' = 2^{1/n}$ até o fator desejado $k = k'^m$, sendo $n$ o número de passos entre uma imagem e sua redução por um fator 2.

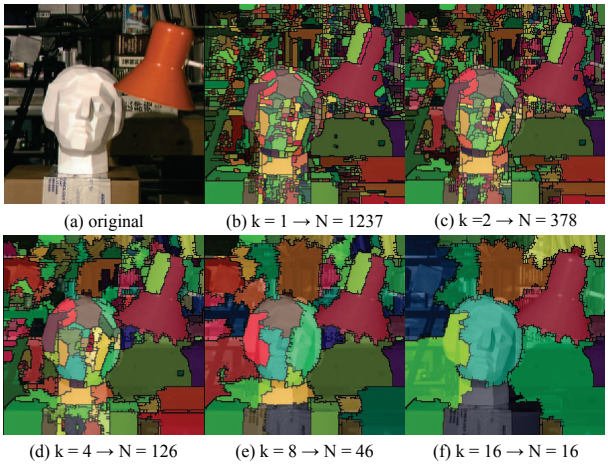



Fig. 2. Agrupamento de superpixels aplicando-se $watershed$ em diferentes escalas da imagem exibida em (a), um fator redução intermediário $k' = \sqrt{2}$ leva a $watershed$ primária em (b) a agrupamentos gerados por um fator de redução na escala com $k = 2$ em (c) e $k = 16$ em (f).

### C. Descritor local para grafos de regiões

De forma similar à equação (2), a orientação e magnitude do gradiente de um vértice $i$ no domínio dos grafos de região é determinado pelo vetor:

$$\vec{\mathbf{m}}_i = \sum_{j \in V, j \neq i} \frac{L_j - L_i}{\frac{||\vec{\mathbf{r}}_j - \vec{\mathbf{r}}_i||}{\sqrt{A_i/\pi}} + 1} \vec{\mathbf{u}}_{i,j}, \qquad (4)$$

onde $L_i - L_j$ representa a diferença de intensidade entre o superpixel $i$ e um vizinho, que dividida pela distância entre $i$ e $j$ normalizada, na direção do vetor unitário $\vec{\mathbf{u}}_{i,j}$, fornece o módulo e direção da contribuição promovida por esse vizinho. A soma da contribuição de todos os vértices do grafo em relação a $i$, retorna o gradiente associado a esse vértice.

A Fig. 3(a) exibe a distribuição das intensidades médias (tons de cinza) da imagem da Fig. 2(a) em um grafo de regiões definido com parâmetros $k = 4$ e $p = 2$ (Fig. 2(c)). Cada vértice tem associado a ele um vetor de gradiente (Fig. 3(b)), calculado pela equação (4).

O termo $\sqrt{A_i/\pi}$ da equação (4) é representado na Fig. 3(b), como o raio equivalente dos superpixels da Fig. 3(a). Esse termo normaliza a distância entre $i$ um vértice e seu vizinho, em termo da área $A_i$, de forma que o tamanho do superpixel $i$ não influencie no cálculo de sua interação com sua vizinhança imediata. A normalização também garante independência quanto a escala na qual se apresenta a imagem.

Diferentemente de [10], todos os vértices são considerados pontos-chave, com orientação e magnitude determinadas pelo próprio vetor $\vec{\mathbf{m}}_i$, não por um histograma de sua vizinhança. Ao invés da distribuição dos setores em uma região retangular,

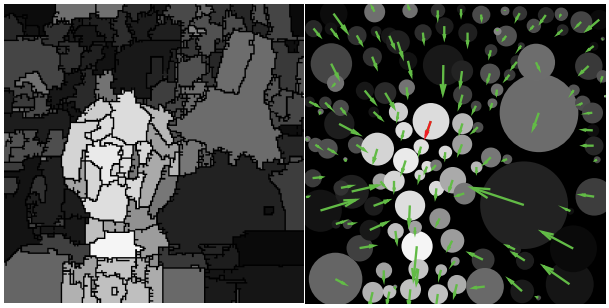



Fig. 3. (a) superpixels exibidos de acordo com os níveis de cinza associados a eles; (b) superpixels representados por círculos de raios iguais a $\sqrt{A_i/\pi}$, ostentando os vetores de gradiente calculados por meio da equação (4), estando em vermelho o vetor do vértice utilizado para exemplificar o cálculo do descritor.

esses são distribuídos em uma grade polar (Fig. 4(a)). O raio equivalente $\sqrt{A_i/\pi}$ determina também o tamanho da região circular que é dividida em 4 quadrantes e dois raios, $R$ e $2R$ (Fig. 4(a)), contabilizando $4 \times 2 = 8$ setores. Definiu-se $R = 10\sqrt{A_i/\pi}$ uma vez que tal proporção apresentou os melhores resultados em testes preliminares.

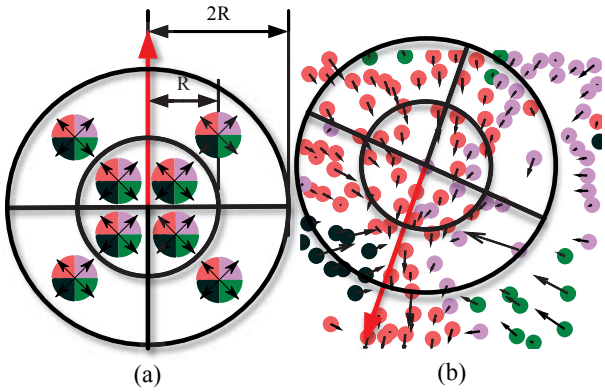


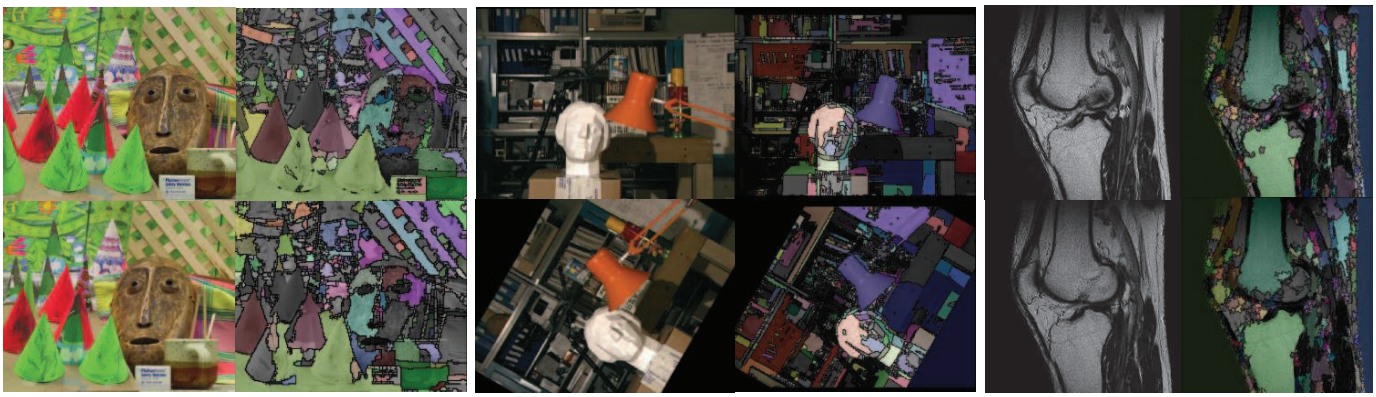

Fig. 4. (a) exibe a representação da grade polar da qual setores fornecem os histogramas que compõem o descritor. Em (b) a grade é orientada de acordo com o vetor gradiente do vértice em análise, referente a região da testa da estátua (Fig. 3).

Com a grade polar centralizada e orientada de acordo com o vetor $\vec{\mathbf{m}}_i$, o descritor de cada vértice $i$ consiste em um vetor, composto pelos histogramas de orientações referentes às 8 regiões da grade polar. Cada amostra $c_j$, referente ao vértice adjacente $j$, é adicionada ao histograma ponderada pela magnitude do gradiente e por uma Gaussiana circular simétrica em torno do centro geométrico $\vec{\mathbf{r}}_i$ do superpixel $i$:

$$c_j = ||\vec{\mathbf{m}}_j|| e^{-\left(\frac{||\vec{\mathbf{r}}_j - \vec{\mathbf{r}}_i||}{2R}\right)^2}. \qquad (5)$$

Utilizando-se quatro orientações para o histograma de cada setor, o descritor se configura como um vetor de $8 \times 4 = 32$ componentes, que é normalizado em valores entre $[0, 1]$.

*1) Confronto:* o conceito de grafos ponderados pode ser utilizado para definir como os pares vencedores são selecionados. Os pesos $W$ quantizam as conexões entre todos os vértices de duas imagens representadas como grafos de região, $V_1$ e $V_2$.



(a) p = 4; k = 2 (b) p = 3; k = 1 (c) p = 3; k = 4

Fig. 5. Confronto direto entre descritores: (a) objetos de escalas distintas, o par estéreo constituído pela vista esquerda original (superior) e a vista direita reduzida por um fator 2 (inferior exibida no aspecto original para melhor visualização); (b) par estéreo com a vista direita rotacionada em $60^o$ no sentido anti-horário; (c) duas imagens de RM (ressonância magnética) oriundas de cortes próximos com características de iluminação ligeiramente distintas, imagem inferior em menor intensidade. Superpixels fora da escala de cinza representam as correspondências.

O peso $w_{i,j}$ da aresta $e_{i,j}$ ($i \in V_1$ e $j \in V_2$) tem como valor a combinação entre três pesos:

$$w_{i,j} = (1-(1-\alpha_C w_{i,j}^{\{C\}})(1-\alpha_S w_{i,j}^{\{S\}})(1-\alpha_I w_{i,j}^{\{I\}}))^2, \quad (6)$$

onde o peso $w_{i,j}^{\{C\}}$ é o relativo ao produto interno entre os descritores de $i$ e $j$, $w_{i,j}^{\{S\}}$, e $w_{i,j}^{\{I\}}$, as tradicionais Gaussianas orientadas pelas distâncias entre as posições e intensidades/cores dos superpixels. $\alpha_C$, $\alpha_S$ e $\alpha_I$ ponderam seus respectivos pesos. Para ser considerado um par, $i$ e $j$ devem ter $w_{i,j}$ como peso de valor máximo para ambos, evitando o casamento de um vértice de uma imagem com mais de um vértice em outra imagem.

Em seus limites, a equação (6) trabalha como uma função lógica. Com $\alpha$'s iguais a 1, quando um dos pesos tende a 1, toda a equação tende a 1, e quando dois pesos tendem a 0 toda equação tende ao valor do peso restante ao quadrado. Os termos $\alpha_C$, $\alpha_S$ e $\alpha_I$, ostentando valores menores do que 1, evitam o casamento entre vértice guiado apenas por correspondências de cor ou posição.

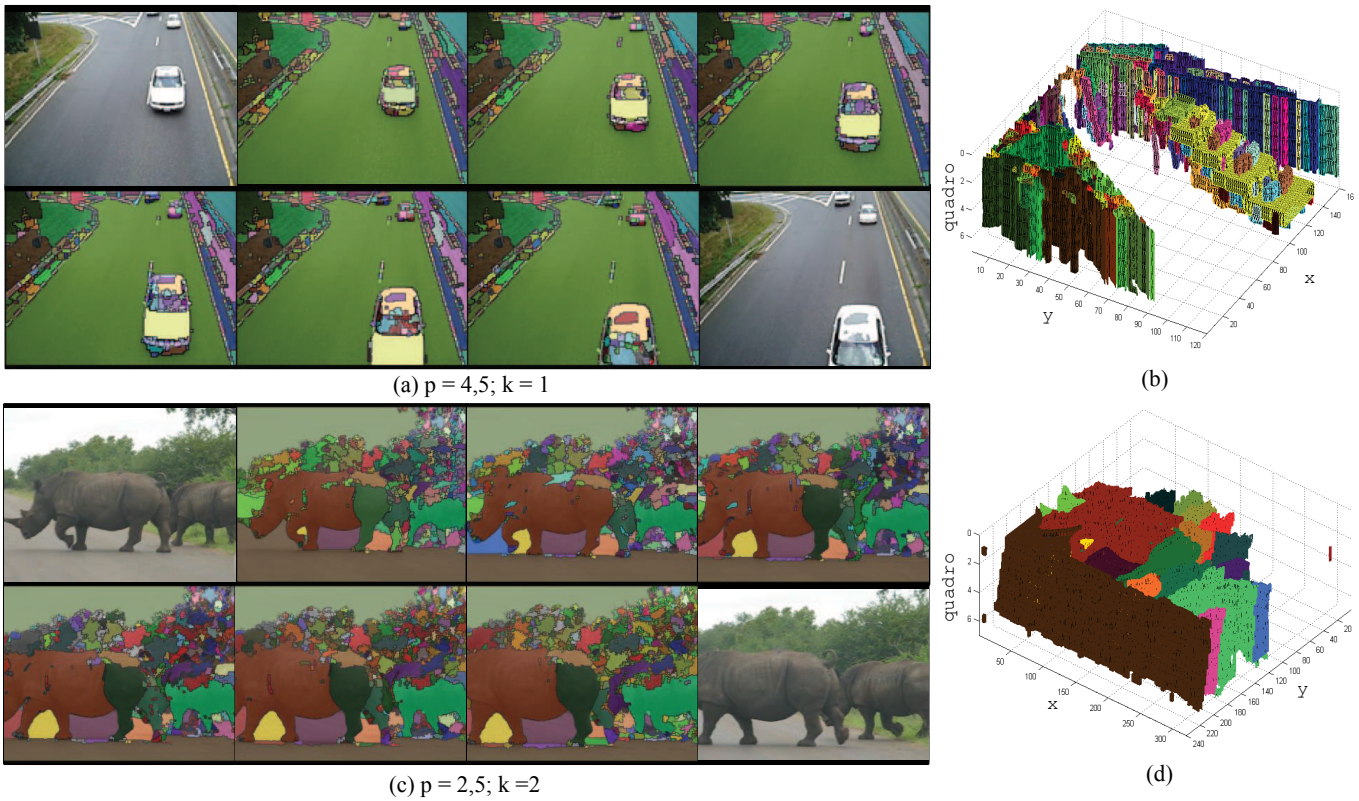
## III. Resultados e Discussão

Os resultados encontram-se em uma avaliação qualitativa. A fim de demonstrar a capacidade discriminatória oferecida pelo algoritmo proposto, o confronto entre imagens e seus pares sujeitos a transformações em escala, rotação e contraste, foi realizado (Fig. 5 (a), (b) e (c), respectivamente). Essas são transformações básicas às quais um objeto pode estar submetido entre dois quadros consecutivos de um vídeo, que em comparação às exibidas aqui (Fig. 5), são esperadas em maior grau de sutileza. O confronto emprega apenas o peso referente ao produto interno entre o descritores das imagens confrontadas, ou seja, $\alpha_S$ e $\alpha_I$ assuem valores nulos.

As transformações de rotação e de iluminação tiveram resultados mais satisfatórios, pois uma vez mantidas as proporções de área e diferença de intensidade entre os superpixels, o descritor proposto é um bom representante para os vértice. Os erros apresentados estão relacionados às regiões próximas. Já a variação em escala teve, visualmente, o pior desempenho. A detecção de bordas por meio de filtros de máscaras de tamanho $3 \times 3$ torna as regiões formadas sensíveis às mudanças de escala e a grandes variações no contraste, pois bordas não detectadas podem significar a união de regiões em uma escala que não se apresentam unidas em outra escala, podendo alterar significantemente a posição de um vértice e, por consequência, seu descritor.

Aplicando-se o descritor proposto no confronto de regiões quadro a quadro de vídeos (Fig. 6 (a) e (b)), em conjunto com as características de cor e posição dos superpixels (sendo $\alpha_C = 1$, $\alpha_S = 0,5$ e $\alpha_I = 0,5$, heuristicamente definidos a princípio), observa-se uma boa manutenção da coerência temporal entre superpixels fixos ou em movimento, mesmo diante de um confronto que parte de uma análise pontual, vértice a vértice, sem considerar o movimento homogêneo de um grupo de vértices pertencentes a um objeto, por exemplo. A contribuição do peso gerado pelo produto interno dos descritores vértice a vértice não se limita ao casamento de correspondências entre quadros, podendo ser empregado como relação de semelhança entre regiões para uma segmentação em um volume sobre-segmentado por meio da técnica discutida (Fig. 6(b) e (d)). Retorna-se então a um problema de supervoxels [2, 3], porém, com um decréscimo no esforço computacional e informações adicionais quanto ao padrão de movimento e vizinhança de superpixel

As queixas quanto à instabilidade de uma segmentação quadro a quadro são relevantes [4], principalmente as promovidas pela *watershed* [18]. Entretanto, o foco deste trabalho é a adaptação da SIFT para a criação de descritores locais diretamente no domínio dos grafos de regiões. Trabalhos futuros terão como objetivo a detecção desses pontos-chaves como vértices no domínio dos grafos de região, que podem ser distribuídos hierarquicamente, por superpixels que poderão ser obtidos por técnicas mais robustas e eficientes [18]. O algoritmo de agrupamento proposto é o ponto de partida para a adaptação dos espaços de escala de [10], para a detecção de pontos especiais (pontos-chave) que se preservem durante mudança na escala de objetos e, para o caso estudado, transformações em sua geometria e intensidade de cor.




(a) p = 4,5; k = 1 (b)

(c) p = 2,5; k =2 (d)

Fig. 6. (a) e (c) exibem sequências a segmentação de 6 quadros de dois vídeos, nas quais foram utilizadas o descritor proposto ($\alpha_C = 1$ na equação (6)) e características de posição e cor ($\alpha_S = 0,5$ e $\alpha_I = 0,5$) para relacionar os superpixels entre quadros em sequência. Um limiar mínimo de 0.5 para o peso de um par vencedor também é utilizado. Regiões sem correspondência recebem um novo rótulo, representado por uma nova cor. As imagens (b) e (d) são as representações em 3D (espaço-tempo) de alguns supervoxels criados a partir de (a) e (b), respectivamente, apresentando os mesmos padrões de cores.

## IV. Conclusões

Foi apresentada uma proposta de algoritmo que traz conceitos da transformação SIFT para o domínio dos grafos de região, criando uma nova propriedade para o rastreamento de superpixels e de relações quadro a quadro entre esses. Os resultados inciais se mostram promissores, uma vez que o descritor proposto casa com boa coerência correspondências entre regiões de imagens de quadros de vídeo, aproveitando a simplificação oriunda da transformação de uma imagem de pixels para um grafo de regiões (superpixels).

## Referências

[1] A. Vazquez-Reina, S. Avidan, H. Pfister, and E. Miller, "Multiple hypothesis video segmentation from superpixel flows," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 268–281.

[2] C. Xu and J. J. Corso, "Evaluation of super-voxel methods for early video processing," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1202–1209.

[3] T. Nagahashi, H. Fujiyoshi, and T. Kanade, "Video segmentation using iterated graph cuts based on spatio-temporal volumes," in *Computer Vision–ACCV 2009*. Springer, 2010, pp. 655–666.

[4] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2141–2148.

[5] S. Hickson, S. Birchfield, I. Essa, and H. Christensen, "Efficient hierarchical graph-based segmentation of rgbd videos," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 344–351.

[6] J. Chang, D. Wei, and J. Fisher, "A video representation using temporal superpixels," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2051–2058.

[7] W. Wang and R. Nevatia, "Robust object tracking using constellation model with superpixel," in *Computer Vision–ACCV 2012*. Springer, 2013, pp. 191–204.

[8] F. Yang, H. Lu, and M.-H. Yang, "Robust superpixel tracking," *Image Processing, IEEE Transactions on*, vol. 23, no. 4, pp. 1639–1651, 2014.

[9] W. Brendel and S. Todorovic, "Video object segmentation by tracking regions," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 833–840.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[11] F. Tang and H. Tao, "Object tracking with dynamic feature graph," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005, pp. 25–32.

[12] S. Battiato, G. Gallo, G. Puglisi, and S. Scellato, "Sift features tracking for video stabilization," in *Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on*. IEEE, 2007, pp. 825–830.

[13] Y.-Z. Song, C. Li, L. Wang, P. Hall, and P. Shen, "Robust visual tracking using structural region hierarchy and graph matching," *Neurocomputing*, vol. 89, pp. 12–20, 2012.

[14] Y. Zhao, Y. Duan, X. Nie, Y. Huang, and S. Luo, "Experts-shift: Learning active spatial classification experts for keyframe-based video segmentation," in *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*. IEEE, 2011, pp. 622–627.

[15] J. B. Roerdink and A. Meijster, "The watershed transform: Definitions, algorithms and parallelization strategies," *Fundamenta informaticae*, vol. 41, no. 1, pp. 187–228, 2000.

[16] J. Stawiaski, "Mathematical morphology and graphs: Application to interactive medical image segmentation," Ph.D. dissertation, Ph. D. dissertation, Paris School Mines, Paris, France, 2008.

[17] R. C. Gonzalez and R. E. Woods, "Digital image processing," 2002.

[18] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274–2282, 2012.