UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

SYLVIO ALEXANDRE BIASUZ BLOCK

ESTUDO DE ASSINATURAS DIGITAIS PARA IDENTIFICAÇÃO DE VÍDEOS

TRABALHO DE CONCLUSÃO DE CURSO

CURITIBA
2015

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

SYLVIO ALEXANDRE BIASUZ BLOCK

ESTUDO DE ASSINATURAS DIGITAIS PARA IDENTIFICAÇÃO DE VÍDEOS

TRABALHO DE CONCLUSÃO DE CURSO

CURITIBA
2015

SYLVIO ALEXANDRE BIASUZ BLOCK

ESTUDO DE ASSINATURAS DIGITAIS PARA IDENTIFICAÇÃO DE VÍDEOS

Trabalho de conclusão de curso apresentado como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação do Departamento Acadêmico de Informática da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Rodrigo Minetto

Coorientador: Prof. Dr. Ricardo Dutra da Silva

CURITIBA
2015

AGRADECIMENTOS

Gostaria de agradecer a Deus e aos meus pais pelo apoio e amor incondicional. E também aos meus orientadores, Prof. Dr. Rodrigo Minetto e Prof. Dr. Ricardo Dutra da Silva pela orientação e confiança, sempre mostrando-se dispostos e colaborativos. O autor também gostaria de agradecer a equipe do projeto QoSTREAM, de número 295220, FP7-MC-IRSES e também Universal-CNPq-Brazil, de número 444789/2014-6.

RESUMO

Block, Sylvio Alexandre Biasuz. Estudo de Assinaturas Digitais para Identificação de Vídeos. 34f. Trabalho de Conclusão de Curso – Departamento Acadêmico de Informática, Universidade Tecnológica Federal do Paraná. Curitiba, 2015.

Uma assinatura de vídeo é um descritor único extraído a partir do conteúdo do vídeo, para identificação e recuperação do mesmo. Neste trabalho é apresentado um estudo comparativo de três métodos, Hua *et. al.* [20], Lee and Yoo [18] e Cook [3] para geração de assinaturas de vídeos. Esses métodos utilizam características espaciais, como luminância e gradiente da imagem, e características temporais, como movimentação da cena e de objetos. Neste trabalho foi utilizada a base de vídeos LIVE Video Quality, juntamente com vídeos alterados pelo autor. Essas alterações foram produzidas por compressão, transmissão, transformações geométricas e outras distorções, intencionais ou não intencionais, para avaliar a capacidade de identificação de vídeos dos métodos estudados.

Palavras Chaves: Assinatura digital de vídeos, identificação de vídeos, recuperação de vídeos, descritores de vídeos.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

A disseminação de dispositivos móveis de gravação de vídeo e também a adição desta funcionalidade aos celulares contribuiu para um rápido crescimento da produção de vídeos [12]. O crescimento pode ser verificado em redes sociais e sites especializados, como é o caso do YouTube, que conta com mais de 100 milhões de vídeos visualizados por dia [1]. Estes vídeos podem ser acessados cada vez mais e por mais pessoas pois, segundo a International Telecomunication Union [16], em 2014 estimavam-se aproximadamente três bilhões de usuários de internet no mundo, ou seja, cerca de 40% da população mundial.

Com tamanho volume de vídeos gerados e acessados, percebe-se que o seu correto tratamento do armazenamento, bem como descrição e sua identificação, são tópicos de interesse para computação. As áreas da computação que estudam a recuperação e detecção de cópias são: A área de, *Content-Based Video Retrieval* (CBIR) que em tradução livre significa recuperação de vídeos baseada em conteúdo, trata do processo de geração de assinatura de vídeo e de identificação (recuperação) de vídeos [8]. Esse ramo da computação abrange desde o desenvolvimento de algoritmos para geração de assinaturas, o estudo de equivalência (*matching*) entre as assinaturas até a busca e recuperação de vídeos. A área de *Content-Based Copy Detection* (CBCD) que em tradução livre significa detecção de cópias baseada em conteúdo, trata da identificação de cópias de vídeos usando assinaturas digitais. O interesse por identificar automaticamente um vídeo também provém da possibilidade de localizá-lo em diferentes bases de dados [19]. Esse ponto é importante, pois a facilidade de cópia e disseminação na internet facilita burlar questões de direitos autorais.

Deste modo pode-se utilizar algumas técnicas para recuperação de vídeos, dentre elas a utilização de rótulos (*tags*). Nesta técnica, a descrição e identificação dos vídeos são feitas manualmente, utilizando-se palavras-chave (rótulos). Por exemplo, enquanto uma pessoa pode rotular um vídeo com as seguintes palavras: “praia”, “sol”, “litoral”, outra pessoa pode utilizar: “mar”, “oceano”, “férias”. Dessa forma, o processo de identificar um vídeo torna-se lento, visto que é preciso comparar diversas palavras e sinônimos.

[1] Disponível em https://www.youtube.com/yt/press/pt-BR/statistics.html. Visualizado em 9 de Julho de 2015

Além disso o processo de descrever um vídeo pode tornar-se redundante e impreciso [1], Sem mencionar a quantidade de informações contidas em um único exemplar [22] e a facilidade de edição de seu conteúdo. Em vídeos de alta resolução e duração, a análise pixel a pixel torna-se demasiadamente custosa.

Dificuldades tais como as discutidas geraram a demanda por métodos automáticos, compactos e precisos de identificação única de vídeos. Dentre estes pode-se citar a assinatura digital, que é uma técnica que extrai características específicas de um vídeo, tais como luminância, gradiente da imagem e movimentação de câmera e objeto, entre outras. A partir dessa informação, é possível comparar as assinaturas para verificar se existe cópia total ou parcial de seu conteúdo, evitando redundância de informações e métodos de força bruta.

A dificuldade encontrada ao propor uma assinatura digital é a possibilidade de um vídeo manter seu conteúdo principal mas ter alterações em características não essenciais ao conteúdo tais como: mudança na compressão do arquivo digital, filmagens em telas de cinema ou de televisão, alteração de cores, adição de legendas, subtração do fundo do vídeo, inserção de tarjas (nas partes superior, inferior e laterais do vídeo), alteração do tamanho (altura e largura), entre outras.

Na Figura 1(a) pode-se observar um quadro extraído de um vídeo, publicado no site Youtube, corresponde a filmagem de um vídeo mostrado em uma televisão. Quando comparado com o original, Figura 1(b), percebe-se diferenças de enquadramento dos personagens, de resolução, no tamanho e também nas cores. No entanto, o conteúdo principal do vídeo é mantido.

(a) Frame do vídeo copiado. (b) Frame do vídeo original.

Figura 1: Exemplo de cópia de vídeo. O vídeo copiado (a) possui alteração de características quando comparado ao vídeo original (b). Fonte: Youtube (copyright Fox Broadcast Company).

Este trabalho apresenta um estudo a respeito da identificação única para um vídeo a partir da criação de uma assinatura digital, técnica que identifica um vídeo através de características como cor, forma, duração, textura e movimentação de objetos ou câmera. Diversos métodos abordando as mais variadas características foram propostos para executar tal tarefa [19], buscando sempre uma assinatura compacta, robusta e que minimize o custo computacional da identificação. No decorrer do trabalho serão apresentados os métodos de Hua *et. al.* [20], Lee e Yoo [18] e Cook [3] utilizados para gerar as assinaturas e a comparação entre eles na identificação de vídeos alterados.

1.1 Objetivo Geral

Este trabalho tem como principal objetivo realizar um estudo comparativo de métodos para assinatura digital de vídeos.

1.2 Objetivos Específicos

- Estudar e comparar as técnicas de assinatura de vídeo por distribuição de intensidade [20], por distribuição de gradientes [18] e por diferença entre quadros [3] ;
- Verificar a possibilidade de utilização das assinaturas digitais estudadas, para identificação e recuperação de cópias, parciais e completas de vídeos;
- Analisar a eficiência das assinaturas digitais na identificação de vídeos com efeitos de edição e com adição de ruídos.

1.3 Estrutura do Documento

Este trabalho está estruturado como segue. No Capítulo 2 são descritos os fundamentos básicos e a revisão de trabalhos correlatos. No Capítulo 3 são apresentados os algoritmos utilizados no projeto. No Capítulo 4 são apresentadas a metodologia, a base de dados utilizada e são mostrados os resultados obtidos. No Capítulo 5 são feitas as considerações finais.

2 Estado da Arte

Neste capítulo serão abordados alguns dos principais conceitos e definições necessárias para entendimento de assinaturas digitais para vídeos e suas implicações.

2.1 Definição de Quadro

Um quadro (*frame*) é uma imagem em uma unidade de tempo dentro do domínio de tempo de um vídeo. Deste modo um quadro é uma função $\mathbb{I}_t(x, y)$ que representa a intensidade de um píxel, em um tempo t, com coordenadas espaciais (x, y) [15].

2.2 Definição de Vídeo

Um vídeo de tamanho n é uma sequência de quadros $V = (\mathbb{I}_1, \mathbb{I}_2, \ldots, \mathbb{I}_n)$, relacionados temporalmente, em que $\mathbb{I}_t$ representa o t-ésimo quadro. Em um vídeo é possível observar dois tipos principais de propriedades. A amostra espacial, referente às dimensões como altura e largura, e a amostra temporal, referente à relação de tempo entre os quadros de um vídeo, como a Figura 2 ilustra tais propriedade.

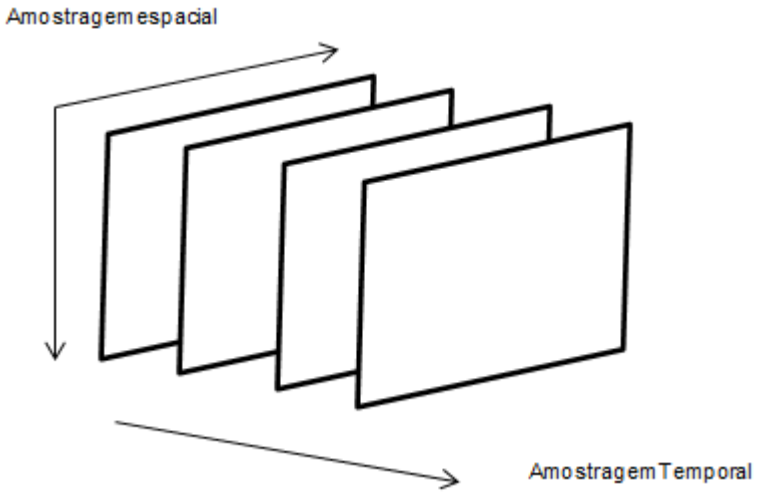


Figura 2: Propriedades de vídeo. Fonte: Autoria própria

Tendo em vista que um vídeo conecta temporalmente os quadros que o constituem, pode-se dividir um vídeo em intervalos intermediários. A Figura 3 ilustra a divisão do vídeo em

cena, tomada e quadros, sendo que uma cena é composta por tomadas e cada tomada por quadros [14].

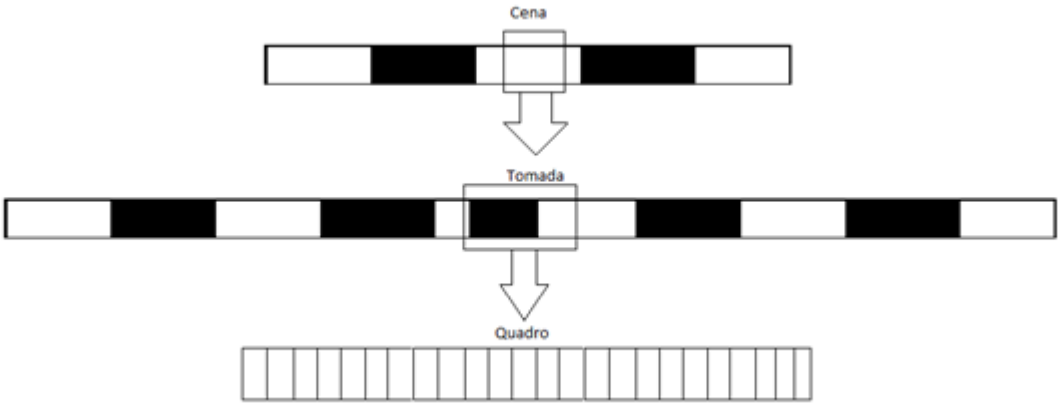


Figura 3: Divisão de vídeo. Fonte: Autoria própria

2.3 Definição de Assinatura de Vídeo

Uma assinatura (ou descritor) de vídeo pode ser definida como um vetor de características que visa identificar unicamente um vídeo juntamente com uma medida de similaridade [12]. Como escopo deste projeto, distintas assinaturas de vídeos serão avaliadas. Para tanto, é necessário caracterizar o que se espera delas.

2.3.1 Características da Assinaturas de Vídeo

A busca de uma assinatura digital capaz de identificar um vídeo com alta precisão e baixo custo computacional conduziu a uma grande variedade de técnicas [8, 9, 19]. A maioria utiliza propriedades temporais e espaciais para a extração de um descritor confiável. A assinatura de vídeo deve possuir ao menos três características, segundo [6, 18, 21]:

- robustez: a assinatura de um determinado vídeo deve ser altamente similar à assinatura do mesmo vídeo sujeito a alterações (distorções) de conteúdo;
- unicidade: as assinaturas de dois vídeos diferentes devem ser proporcionalmente diferentes;

- eficiência de busca: a assinatura digital deve ser compacta e de baixo custo computacional para que seja eficiente na busca de vídeos em banco de dados.

Se a assinatura digital gerada por determinada técnica atender esses critérios, ela é considerada viável para ser utilizada na identificação de cópia completa ou parcial de vídeos.

Como descrito no Capítulo 1, existem várias características de um vídeo que podem ser alteradas, dificultando a construção de soluções computacionais adequadas [5]. As diferentes técnicas utilizadas para gerar assinaturas em vídeos divergem tanto nas características usadas para extrair seus parâmetros quanto nas técnicas empregadas para avaliar o grau de similaridade entre as diferentes assinaturas.

2.4 Descritores

Do ponto de vista das características que os descritores exploram, é possível dividir as técnicas em dois grandes grupos. O primeiro refere-se às técnicas que utilizam descritores globais para extrair as propriedades dos vídeos e o segundo é o grupo de técnicas que utilizam descritores locais para o mesmo propósito.

2.4.1 Descritores Globais

Estes descritores buscam sintetizar informações contidas nos vídeos a partir de características únicas e tratam cada quadro do vídeo como um objeto único, ou seja, as informações usadas representam o quadro como um todo. Geralmente, possuem uma codificação e implementação mais simples, bem como possuem um bom desempenho computacional.

Os descritores globais utilizam-se de características como distribuição de cores [3] [20], bordas [18] e também histogramas de cor [11] que representam um quadro. Essas técnicas mostram-se robustas sobre distorções geométricas como rotação, translação ou até mesmo recortes. Contudo, são mais suscetíveis a distorções que atingem o espaço de cores, tais como inversão de cores e incremento ou decremento na luminância. Os algoritmos testados neste projeto (Capítulo 3) são todos globais e foram escolhidos por possuir resultados amplamente

divulgados, além de soluções simples e viáveis para o objetivo de identificação e recuperação de vídeos.

2.4.2 Descritores Locais

As técnicas que utilizam descritores locais exploram elementos da imagem, tais como pontos de interesse, bordas ou objetos específicos. Nesse contexto busca-se caracterizar o comportamentos das regiões ao longo do tempo tornando descritores mais robustos a distorções de geometria, de cores e de texturas. Porém, o custo computacional é mais alto em relação aos descritores globais e também é maior a complexidade de implementação.

Essas técnicas procuram encontrar pontos de interesse que possam ser rastreados ao longo do vídeo produzindo descritores relevantes e imunes a distorções. O descritor local deve possuir a capacidade de repetir-se ao longo do tempo. Deste modo ao sofrer transformações espaciais o descritor mantêm-se robusto. Pode-se citar a detecção de bordas em regiões específicas e também a técnica de *Space-Time Interest Points* [1] [7] que procura regiões no quadro onde ocorrem movimentações mais abruptas.

2.5 Trabalhos Prévios

Além da divisão entre descritores globais e locais, também é possível subdividir os descritores em relação as características extraídas dos vídeos. Alguns métodos utilizam abordagens espaciais para produzir seus descritores, calculando medidas baseadas em níveis de cinza e informações sobre bordas. Outros métodos utilizam medidas temporais, explorando informações entre quadros separados temporalmente. Na sequência será apresentada uma revisão dos métodos presentes na literatura.

Dos métodos que utilizam descritores baseados em características espaciais, pode-se citar os trabalhos de Hua *et. al.* [20] que calcularam uma medida ordinal que reflete a distribuição de níveis de cinza para cada quadro. A assinatura foi desenvolvida para ser robusta a diferentes formatos de compressão, taxa de quadros, assim como diferenças no tamanho

e composição espacial dos quadros. Lee e Yoo [17, 18] propuseram um método baseado na orientação do centroide do gradiente para ser utilizada sobre transformações geométricas, tais como rotação e translação, e ruído Gaussiano. Uma abordagem similar é apresentada por Massoudi *et. al.* [10], que propõem uma assinatura baseada na orientação do gradiente e foi testada sobre transformações como compressão, recortes e borramento. Su *et. al.* [21] propuseram um método baseado em regiões de atenção, o qual foi avaliado em vídeos contendo distorções de resolução, de quantidade de quadros por segundo e de adição de logotipos. Radhakrishnan and Bauer [13] apresentam um método que utiliza a Decomposição de Valores Singulares (SVD) que busca robustez sobre transformações geométricas, compressões e alterações nas taxas de quadros.

Além disso tem-se os métodos que utilizam descritores baseados em características temporais. Cook [3] propôs uma assinatura temporal, que utiliza a diferença de luminância entre quadros para medir como a informação contida no vídeo é alterada através do tempo. A simplicidade deste método é explorada para atingir eficiência computacional. O autor afirma que o método é robusto tanto sobre transformações geométricas tanto quanto não geométricas. No trabalho de Kim e Vasudev [2], a intensidade média dos quadros é computada e comparada com a mesma medida em um quadro subsequente, implicando em uma assinatura espaço-temporal. Chen e Stentiford [4] comparam a média do nível de cinza para dois quadros consecutivos. Hampur *et. al.* [5] propõem um método baseado na captura de movimento entre dois quadros subsequentes, utilizando como métrica o mínimo da diferença da soma absoluta dos pixeis entre os quadros.

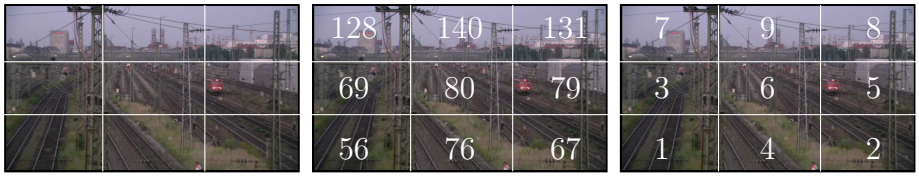
Neste trabalho foram estudados os métodos de Hua *et. al.* [20], Lee e Yoo [18] com o objetivo de avaliar a robustez e unicidade de seus descritores. Os métodos abrangem as técnicas mais comumente utilizadas para a geração de assinaturas digitais através de descritores globais.

3 Técnicas de Assinatura Digital

Nesta seção serão apresentados e discutidos como os métodos testados neste trabalho geram as assinaturas digitais.

3.1 Assinatura de Vídeo por Distribuição de Intensidade

Em 2004, Hua *et. al.* [20] propuseram um método baseado na intensidade de níveis de cinza de um quadro para gerar uma assinatura, com este propósito, os autores dividem os quadros de um vídeo V em $M \times N$ blocos de tamanho igual. Constituindo o vetor de blocos $\mathbf{b} = (b_1, b_2, ...b_{M \times N})$, como mostrado na Figura 4(a). Fonte: Autoria própria.



(a) (b) (c)

Figura 4: Cálculo do descritor de Hua *et. al.* [20]: (a) divisão do quadro em blocos; (b) nível de cinza médio em cada bloco; (c) o descritor $\mathbf{d} = (7, 9, 8, 3, 6, 5, 1, 4, 2)$ é formado pela permutação que ordena os níveis de cinza dos blocos de (b).

A média da intensidade do nível de cinza em cada bloco b_i, de um quadro $\mathbb{I}$, para $i = \{1, \ldots, M \times N\}$, é calculada como:

$$g_i = \frac{\sum\limits_{x,y \in b_i} \mathbb{I}(x, y)}{|b_i|}, \tag{1}$$

onde $|b_i|$ é o número de pixeis que compõem o bloco b_i, ver Figura 4(b).

Obtém-se o descritor $\mathbf{d}$ através da ordenação, em ordem crescente, do vetor da média da intensidade do nível de cinza $\mathbf{g} = (g_1, g_2, ...g_{M \times N})$ e nomeando-os com inteiros consecutivos $1, 2, \ldots, M \times N$. Esse vetor ordenado também é conhecido como *medida ordinal*. Em outras palavras o valor do descritor d_i é o valor do inteiro que atribui-se ao bloco b_i durante a ordenação, como visto na Figura 4(c).

A assinatura de um vídeo é composta pela sequência de descritores de cada quadro. Os autores afirmam que a assinatura de um vídeo $\mathbf{k}$ onde $\mathbf{f}^k = \langle \mathbf{d}_1^k, \mathbf{d}_2^k, \ldots, \mathbf{d}_n^k \rangle$ é robusta sobre distorções no espectro de cores e propriedades como altura e largura.

3.2 Assinatura de Vídeo por Distribuição de Gradientes

Lee e Yoo [18] propuseram uma assinatura baseada na distribuição dos gradientes dos quadros. Inicialmente calcula-se o gradiente, de um quadro $\mathbb{I}$ para cada ponto (x, y) da seguinte forma:

$$\nabla\mathbb{I}(x,y) = \begin{bmatrix} \mathbb{G}_x \\ \mathbb{G}_y \end{bmatrix} = \begin{bmatrix} \partial\mathbb{I}/\partial x \\ \partial\mathbb{I}/\partial y \end{bmatrix} = \begin{bmatrix} \mathbb{I}(x+1,y) - \mathbb{I}(x-1,y) \\ \mathbb{I}(x,y+1) - \mathbb{I}(x,y-1) \end{bmatrix}. \tag{2}$$

Dado um gradiente, sua magnitude e sua orientação são obtidas, respectivamente, por:

$$\omega(x,y) = \sqrt{\mathbb{G}_x^2 + \mathbb{G}_y^2} \qquad \theta(x,y) = \tan^{-1}\left(\frac{\mathbb{G}_y}{\mathbb{G}_x}\right). \tag{3}$$

Os quadros do vídeo V são divididos em $M \times N$ blocos de tamanhos iguais e calcula-se a orientação do centroide dos gradientes para cada bloco. Especificamente, o descritor do centroide d_i relativo ao bloco b_i, $i = \{1, \ldots, M \times N\}$, é dado por:

$$d_i = \frac{\sum\limits_{x,y \in b_i} \omega(x,y)\theta(x,y)}{\sum\limits_{x,y \in b_i} \omega(x,y)}. \tag{4}$$

Na Figura 5, observa-se o esquema de normalização do vídeo utilizado por Lee e Yoo [18], onde antes de se iniciar a extração do descritor, o vídeo tem sua cadência de quadros padronizada, assim como suas dimensões, e por fim é convertido para escala de cinza. Posteriormente é iniciado o processo de geração da assinatura de gradiente.

Os autores afirmam que a assinatura está intimamente relacionada com a distribuição de bordas nos quadros, fornecendo informações visuais relevantes sobre o conteúdo dos quadros, como os limites de objetos. A assinatura é robusta sobre modificações globais na intensidade dos *pixels*, tais como alteração no brilho, cor e contraste.

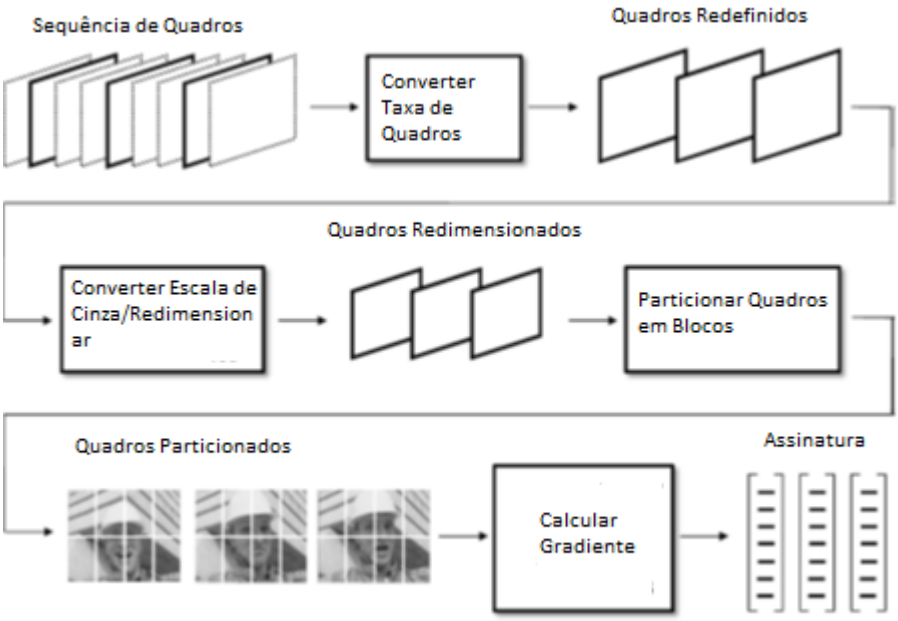


Figura 5: Assinatura por distribuição de gradiente. Fonte: Lee, Sunil e Yoo, Chang D. (2008. P 984). Traduzido pelo Autor.

3.3 Assinatura de Vídeo por Diferença entre Quadros

Cook [3] propôs, em 2011, uma assinatura baseada na informação temporal contida entre dois quadros deslocados no tempo. Sejam $\mathbb{J}$ e $\mathbb{I}$ dois quadros de um vídeo, não necessariamente subsequentes. A diferença entre eles é dada por

$$dY = \sum_{x,y \in \mathbb{I},\mathbb{J}} |\mathbb{I}(x,y) - \mathbb{J}(x,y)|. \tag{5}$$

Os autores também calculam a soma dos níveis de cinza do quadro $\mathbb{I}$

$$Y = \sum_{x,y \in \mathbb{I}} \mathbb{I}(x,y). \tag{6}$$

O descritor para cada quadro é então composto pelas duas medidas, $\mathbf{d} = \langle dY, Y \rangle$.

Os autores afirmam que o valor dY reflete a mudança global dos quadros. Alterações que persistem entre quadros, tais como translação, rotação, espelhamento, legendas, sobreposições estáticas, brilho, contraste e alterações no espectro de cores, são canceladas quadro a quadro, deixando um registro relativamente consistente das mudanças no vídeo.

4 Metodologia

Nesta seção serão apresentados a base de dados e as métricas de avaliação, assim como os resultados obtidos.

4.1 Base de Dados

Neste trabalho foi utilizada a base de vídeo *LIVE Video Quality* (LIVE-VQD), que contém dez vídeos de referência com alta qualidade e sem compressão, estes vídeos contém movimentação de câmera e de objetos, sendo que a quantidade de quadros varia entre 250 e 500. Além dos 10 vídeos de referência, a base conta ainda com 150 vídeos distorcidos, sendo 15 vídeos distorcidos a partir de cada vídeo de referência. As distorções contidas nas bases são: Compressão MPEG-2, compressão H.264, simulação de transmissão de dados comprimidos com H.264 através de uma rede IP e perda de dados na transmissão *wireless*. Dessa forma cada vídeo de referência possue 15 vídeos distorcidos, sendo que entre esses vídeos encontram-se diferentes níveis das quatro distorções citadas. Na Figura 6 estão os quadros iniciais dos dez vídeos de referência.

Neste trabalho ainda foram incluídas 14 distorções adicionais,ver Figura 7, por vídeo de referência : rotação, recorte, espelhamento, legendas, marcas d'água, alteração de cores, aumento e diminuição do quadro, adição de bordas e aumento e diminuição na taxa de quadros, compressão dos quadro por JPEG. Totalizando uma base de 300 vídeos. Geralmente tais alterações são deliberadamente realizadas sobre vídeos proprietários com a intenção de despistar métodos de detecção de cópias.

4.2 Experimentos

Os experimentos foram realizados através da consulta da assinatura de um vídeo V^q contra a assinatura de cada vídeo V^t na base de dados. O vídeo de consulta é uma sequência de quadros $V^q = \langle \mathbb{I}_1, \mathbb{I}_2, \ldots, \mathbb{I}_n \rangle$ e o vídeo alvo é uma sequência de quadros $V^t = \langle \mathbb{J}_1, \mathbb{J}_2, \ldots, \mathbb{J}_m \rangle$, $n \ll m$. As assinaturas para esses vídeos são as sequências $\mathbf{f}^q = \langle \mathbf{d}_1^q, \mathbf{d}_2^q, \ldots, \mathbf{d}_n^q \rangle$ e $\mathbf{f}^t = \langle \mathbf{d}_1^t, \mathbf{d}_2^t, \ldots, \mathbf{d}_m^t \rangle$, respectivamente.

Figura 6: Quadros iniciais dos dez vídeos de referência. Fonte: LIVE Video Quality (LIVE-VQD).

Seja $\mathbf{d}^q \in \langle \mathbf{d}_1^q, \mathbf{d}_2^q, \ldots, \mathbf{d}_n^q \rangle$ e $\mathbf{d}^t \in \langle \mathbf{d}_1^t, \mathbf{d}_2^t, \ldots, \mathbf{d}_n^t \rangle$. A distância L_1 entre os descritores $\mathbf{d}^q = \langle d_1^q, d_2^q, \ldots, d_k^q \rangle$ e $\mathbf{d}^t = \langle d_1^t, d_2^t, \ldots, d_k^t \rangle$ é dada por:

$$L_1(\mathbf{d}^q, \mathbf{d}^t) = ||\mathbf{d}^q - \mathbf{d}^t|| = \sum_{i=1}^{k} |d_i^q - d_i^t|. \tag{7}$$

A distância entre as assinaturas do vídeo de consulta e de uma subsequência de quadros do vídeo alvo são calculadas por:

$$D(\mathbf{f}^q, \mathbf{f}_j^t) = \frac{1}{nk} \sum_{i=1}^{n} L_1(\mathbf{d}_i^q, \mathbf{d}_{j+i}^t) \tag{8}$$

Figura 7: Quadros com 9 das 14 distorções, da esquerda para a direita e de cima para baixo: blur (ofuscamento), adição de borda vermelha, inversão de cores, recorte central do quadro, espelhamento, compressão JPEG do quadro, rotação para a direita, adição de legenda e, por fim, adição de marca d'água. Fonte: adaptado de LIVE Video Quality (LIVE-VQD).

onde o fator nk é utilizado para normalização e $j = \{0, \ldots, m - n\}$ são as possíveis posições onde a consulta pode ser comparada com o vídeo alvo.

Os valores obtidos para os descritores de cada método foram normalizados para o intervalo $[0, 1]$. Os ajustes dos parâmetros de cada método como descritos originalmente pelos autores podem ser vistos na Tabela 1.

Tabela 1: Parâmetros dos algoritmos e informações das assinaturas

Algoritmo	Intervalo do Descritor	Número de blocos	Número de propriedades por bloco	Tamanho do descritor
Hua *et. al.* [20] (Ordinal)	$[1:9]$	9	1	9
Lee and Yoo [18] (Gradient)	$[-\pi/2 : +\pi/2]$	8	1	8
Cook [3] (Temporal)	$[0 : \lvert\mathbb{I}\rvert \times 255]$	1	2	2

Com o objetivo de avaliar as assinaturas testadas, foram computadas as curvas de Característica de Operação do Receptor (no inglês Receiver Operator Characteristic) e pre-

cisão-revocação. Dado um limiar (*threshold*) τ, no intervalo $[0, 1]$, se $D(\mathbf{f}^q, \mathbf{f}_j^t) \leq \tau$ então assume-se que o vídeo V^q corresponde à subsequência do vídeo V^t que se inicia no quadro j. Ao comparar-se a consulta com todos os vídeos da base de dados foram computadas as medidas de precisão (P), revocação (R) (também denominada taxa de verdadeiros positivos (TPR)), e taxa de falsos positivos (FPR)

$$P = \frac{TP}{TP + FP} \qquad R(TPR) = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN} \tag{9}$$

onde TP é o número de verdadeiros positivos (acertos), FP é o número de falsos positivos, FN é o número de falsos negativos e TN é o número de verdadeiros negativos. Para um limiar específico, um verdadeiro positivo ocorre quando a consulta buscada de um vídeo corresponde a outra assinatura de um vídeo distorcido sendo ambos do mesmo vídeo de referência. Se eles não forem correspondentes, tem-se um falso negativo. Um falso positivo ocorre quando há a correspondência de assinaturas oriundas de vídeos de diferentes vídeos de referência, de outro modo ocorre um verdadeiro negativo.

Foram considerados três cenários para se avaliar a robustez e a unicidade das assinaturas:

1. Vídeos com compressão e erros de transmissão como aqueles contidos na base de dados LIVE-VQD;

2. Vídeos com compressão e erros de transmissão comuns, juntamente com vídeos distorcidos intensionalmente pelo autor;

3. Variação do tamanho da consulta sobre os vídeos do cenário (2).

As curvas de precisão-revocação e ROC para 50 consultas aleatoriamente selecionadas são apresentadas nas Figuras 8-10. A Figura 8 apresenta os resultados no cenário (1) considerando consultas que têm tamanho igual a 25 quadros. Todos os métodos saíram-se muito bem para estes tipos de distorções. O método de Hua *et. al.* [20] alcançou 100% de revocação e de precisão.

O comportamento não é o mesmo para o cenário (2), onde incluem-se os vídeos intencionalmente distorcidos, como mostra a Figura 9. Percebe-se que a performance cai rapidamente

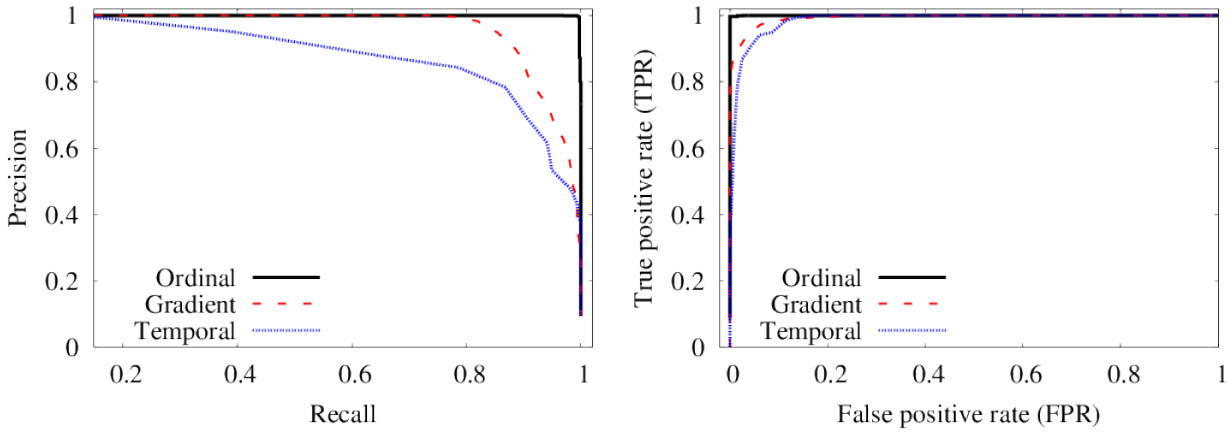


Figura 8: Performance dos métodos: curvas de precisão-revocação (esquerda) e ROC (direita) para consultas com **25** quadros considerando o cenário (1).

para estes tipos de distorções. Os métodos de Hua *et. al.* [20] e Lee e Yoo [18] alcançaram resultados semelhantes, sendo ambos mais robustos que o método de Cook [3].

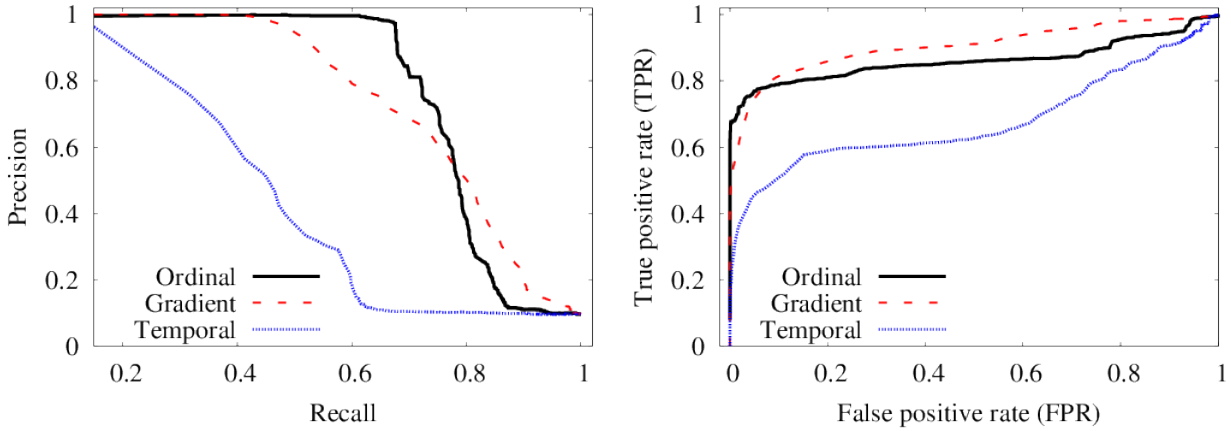


Figura 9: Performance dos métodos: curvas de precisão-revocação (esquerda) e ROC (direita) para consultas com **25** quadros considerando o cenário (2).

Com o objetivo de verificar a influência do tamanho da consulta na recuperação de vídeos, foram utilizados os vídeos do cenário (2) e consultas com 100 e 200 quadros. Os resultados dos testes podem ser vistos nas Figuras 10 e 11. Percebe-se que a performance de todos os algoritmos melhoraram nitidamente com consultas de 100 e 200 quadros. No entanto, o tamanho da consulta impacta diretamente no custo de computação dos métodos, visto que é preciso maior quantidade de cálculos para comparar as assinaturas. Esse ponto de troca

mostra-se muito importante na eficiência da busca em bases de dados.

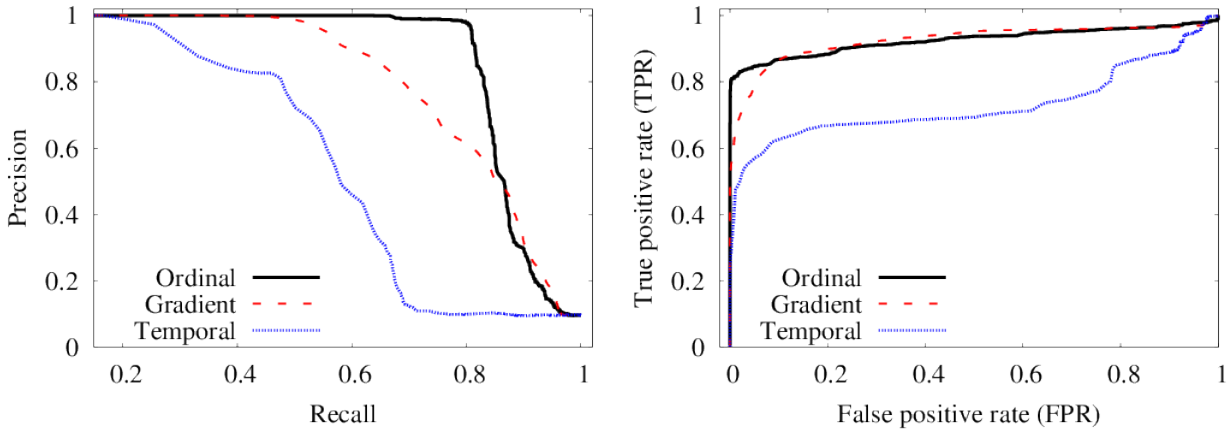


Figura 10: Performance dos métodos: curvas de precisão-revocação (esquerda) e ROC (direita) para consultas com **100** quadros considerando o cenário (3)

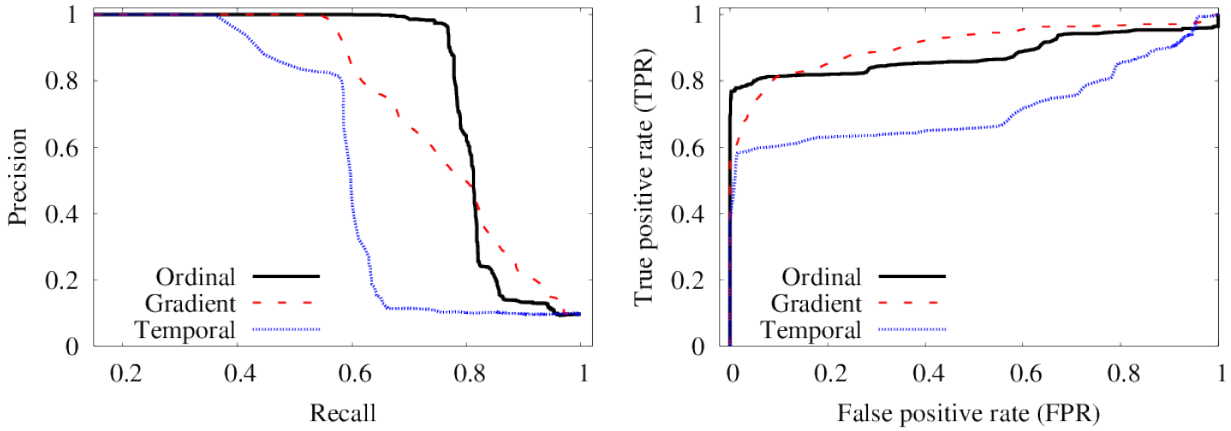


Figura 11: Performance dos métodos: curvas de precisão-revocação (esquerda) e ROC (direita) para consultas com **200** quadros considerando o cenário (3)

Durante os testes também foram comparadas as curvas de distâncias L_1 entre duas assinaturas. A análise destas curvas permite observar o comportamento das assinaturas em diferentes cenários. O comportamento esperado é que a distância L_1 de dois descritores oriundos de um mesmo vídeo de referência, possuam um mínimo evidente, abaixo do limiar estabelecido, configurando assim uma correspondência. Por outro lado, para descritores provenientes de vídeos de referência diferentes espera-se que a distância L_1 não apresente mínimo abaixo do limiar.Para elaborar esse teste foram realizadas três comparações.

Primeiramente foram comparadas as assinaturas geradas entre o vídeo bs1 e o vídeo bs1_blur que são o mesmo vídeo porém o segundo conta com a adição de ruído do tipo blur. Pode-se observar a diferença entre os três métodos na Figura 12, onde verifica-se que os três métodos conseguiram identificar precisamente o ponto onde ocorre a correspondência entre a consulta e o vídeo alvo, caracterizando um verdadeiro positivo. Este ponto é representado pela linha vertical "Consulta".

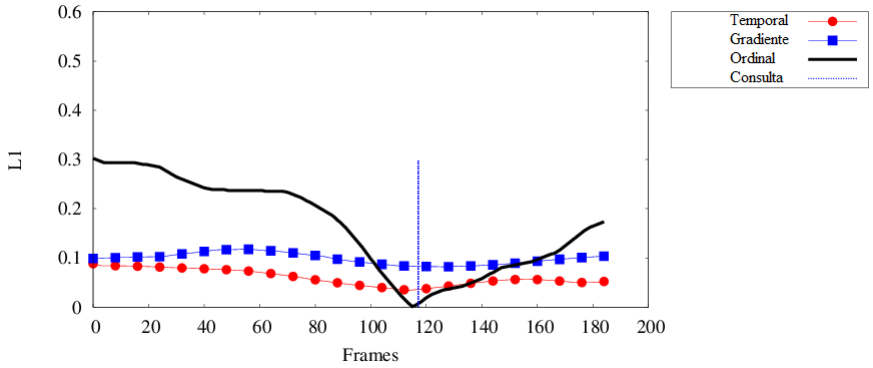


Figura 12: Distância L_1 entre vídeos de mesma referência com mínimo evidente. As linhas horizontais representam as técnicas e a linha vertical o ponto de início da consulta.

O segundo teste comparou a distância L_1 entre os vídeos bs1 e tr1, ou seja, dois vídeos completamente diferentes e sem a adição de distorções. Na Figura 13 observa-se que nas três curvas não se percebe um mínimo aparente, evidenciando que tratam-se de vídeos distintos, caracterizando um verdadeiro negativo.

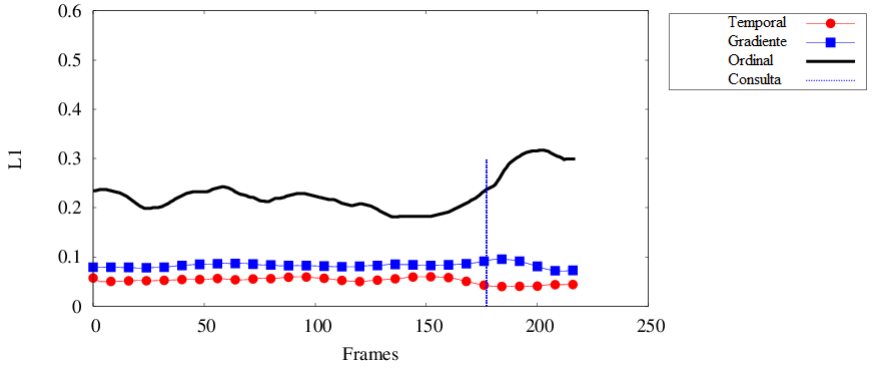


Figura 13: Distância L_1 entre vídeos de diferentes referências sem mínimo evidente. As linhas horizontais representam as técnicas e a linha vertical o ponto de início da consulta.

Por último foram comparados os vídeos bs1 e bs1_Crop, ou seja, o vídeo de referência e sua cópia com recorte central. O comportamento esperado era que fosse encontrado um mínimo aparente, mas ele não ocorreu, evidenciando um falso negativo. Ver Figura 14.

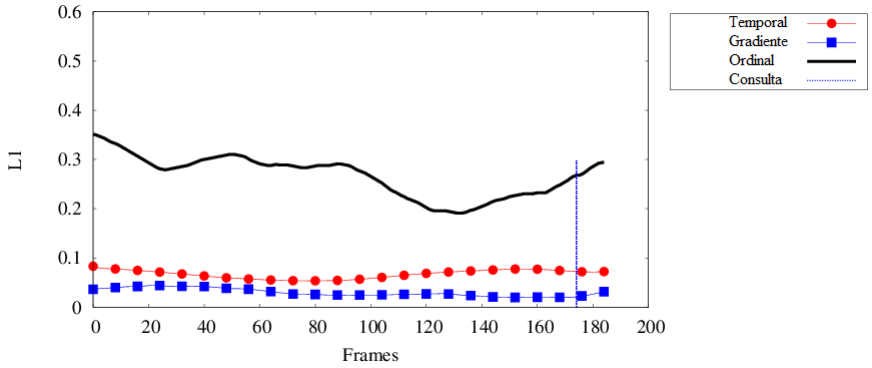


Figura 14: Distância L_1 entre vídeos de mesma referência sem mínimo evidente. As linhas horizontais representam as técnicas e a linha vertical o ponto de início da consulta.

Ao observar os gráficos de distância L_1 percebe-se que os valores de L_1 para a técnica de intensidade estão sempre maiores, mas quando encontra uma correspondência este valor cai drasticamente. Dessa forma, são corroborados os gráficos de precisão/revocação onde esta técnica obteve o melhor desempenho. Isto ocorre pois, com um mínimo acentuado, torna-se fácil para a assinatura identificar uma correspondência. Nas outras técnicas, os valores de L_1 não obtêm grande variação, dificultando a identificação de correspondência. Percebe-se que nas técnicas de gradiente e temporal os valores de L_1 estão sempre muito próximos de zero, mesmo em situações onde não existe correspondência. Este fato pode indicar um dos fatores que levaram a uma precisão inferior nas curvas de precisão/revocação e ROC.

5 Conclusão

Neste trabalho relatou-se uma avaliação experimental dos métodos de Hua *et. al.* [20], Lee e Yoo [18], baseados em diferentes atributos, para a computação de assinaturas de vídeos digitais. Os métodos foram testados na base de vídeo *LIVE Video Quality* (LIVE-VQD) que possui qualidade reconhecida, a fim de investigar o efeito de distorções de vídeo e erros de transmissão, com o objetivo de recuperação de vídeo. Deste modo foram inclusas 14 distorções de preservação de conteúdo, a fim de simular cópias de vídeo modificados. Os experimentos mostraram que os métodos são robustos à tipos comuns de compressão e distorção, tais como os vídeos da base de dados LIVE Video Quality, mas são sensíveis a cópias de vídeo modificados propositalmente. Nos experimentos desenvolvidos neste trabalho, os melhores resultados foram obtidos utilizando-se o método da medida ordinal. Trabalhos futuros podem abordar a utilização conjunta de diferentes descritores para avaliar se existe melhora de desempenho, sendo possível também avaliar o desempenho de recuperação de vídeos em grande bases de dados.

A contribuição principal do projeto ocorreu durante o processo de desenvolvimento do artefato textual, pois foi necessária extensa pesquisa na área, assim como no processo de testes.

Referências

[1] Felipe dos Santos Pinto de Andrade. Combinação de descritores locais e globais para recuperação de imagens e vídeos por conteúdo. *Universidade Estadual de Campinas (UNICAMP). Instituto de Computação*, 2012.

[2] B. Changick Kim e Vasudev. Spatiotemporal sequence matching for efficient video copy detection. *IEEE Trans. on Circ. and Systems for Video Tech.*, 15(1):127–132, 2005.

[3] R. Cook. An efficient, robust video fingerprinting system. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, July 2011.

[4] Li Chen e F.W.M. Stentiford. Video sequence matching based on temporal ordinal measurement. *Pattern Recognition Letters*, 19:1824–1831, 2008.

[5] Hyun Kiho e Bolle Ruud M. Hampapur, Arun. Comparison of sequence matching techniques for video copy detection. *Storage and Retrieval for Media Databases*, 4676:194–201, 2001.

[6] Jean-Michel Jolion Isabelle Simand e Denis Pellerin, Stéphane Bres. Spatio-Temporal Signatures for Video Copy Detection. In *Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS)*, pages 421–427, September 2004.

[7] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[8] Chen Li Joly Alexis Laptev-Ivan Buisson Olivier Gouet-Brunet Valerie Boujemaa Nozha e Stentiford Fred Law-To, Julien. Video copy detection: A comparative study. In *ACM International Conference on Image and Video Retrieval*, CIVR, pages 371–378, 2007.

[9] Jian Lu. Video fingerprinting for copy identification: from research to industry applications. *Media Forensics and Security*, 7254:725402–725402–15, 2009.

[10] Lefebvre F. Demarty C. Oisel L. e Chupeau B. Massoudi, A. A video fingerprint based

on visual digest and local fingerprints. In *IEEE Int. Conf. on Image Processing*, pages 2297–2300, 2006.

[11] Kankanhalli Mohan S e Narasimhalu A Desai Mehtre, Babu M and Guo Chang Man. Color matching for image retrieval. *Pattern Recognition Letters*, 16(3):325–331, 1995.

[12] Otavio Augusto Bizetto Penatti. Estudo comparativo de descritores para recuperação de imagens por conteúdo na web. *Universidade Estadual de Campinas (UNICAMP). Instituto de Computação*, 2009.

[13] C. Radhakrishnan, R. e Bauer. Robust video fingerprints based on subspace embedding. In *IEEE ICASSP*, pages 2245–2248, 2008.

[14] Thiago Teixeira Santos. Segmentação automática de tomadas em vídeo. *Universidade Estadual de São Paulo (USP). Instituto de Computação*, 2004.

[15] Nielsen Cassiano Simoes. Detecção de algumas transições abruptas em sequencias de imagens. *Mestrado, Instituto de Computação, UNICAMP, Campinas*, 5, 2004.

[16] ITU Statistics. Key ict indicators for developed and developing countries and the world (totals and penetration rates). *URL: http://www. itu. int/ITUD/ict/statistics/at_glance/KeyTelecom. html*, 29:2012, 2014.

[17] C.D. Sunil Lee e Yoo. Robust video fingerprinting based on affine covariant regions. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1237–1240, 2008.

[18] C.D. Sunil Lee e Yoo. Robust video fingerprinting for content-based video identification. *IEEE Trans. on Circuits and Systems for Video Technology*, 18(7):983–988, 2008.

[19] Li Li Xianglin Zeng e Maybank S. Weiming Hu, Nianhua Xie. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 41(6):797–819, Nov 2011.

[20] Xian Chen e Hong-Jiang Zhang Xian-Sheng Hua. Robust video signature based on

ordinal measure. In *International Conference on Image Processing*, volume 1, pages 685–688, Oct 2004.

[21] Tiejun Huang e Wen Gao Xing Su. Robust video fingerprinting based on visual attention regions. In *IEEE ICASSP*, pages 1525–1528, 2009.

[22] Rui Yong Huang Thomas S e Mehrotra Sharad Zhuang, Yueting. Adaptive key frame extraction using unsupervised clustering. In *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, volume 1, pages 866–870. IEEE, 1998.

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS LONDRINA