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

JORDY JACKSON ANTUNES DA ROCHA,
JULIA ULSON TRETEL,
RODRIGO MACHADO

ANÁLISE COMPARATIVA DE ALGORITMOS PARA ASSINATURA DIGITAL DE VÍDEOS

TRABALHO DE CONCLUSÃO DE CURSO

CURITIBA
2018

JORDY JACKSON ANTUNES DA ROCHA,
JULIA ULSON TRETEL,
RODRIGO MACHADO

ANÁLISE COMPARATIVA DE ALGORITMOS PARA ASSINATURA DIGITAL DE VÍDEOS

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Sistemas de Informação da Universidade Tecnológica Federal do Paraná, como requisito parcial para a obtenção do título de Bacharel.

Orientador: Prof. Dr. Rodrigo Minetto
Universidade Tecnológica Federal do Paraná

Coorientador: Prof. Dr. Ricardo Dutra da Silva
Universidade Tecnológica Federal do Paraná

CURITIBA
2018

Agradecimentos

TODO AGRADECIMENTOS

RESUMO

Rocha, J. J. A, Tretel, J. U, Machado, R.. Análise Comparativa de Algoritmos para Assinatura Digital de Vídeos. 2018. 23 f. Trabalho de Conclusão de Curso – Bacharelado em Sistemas de Informação, Universidade Tecnológica Federal do Paraná. Curitiba, 2018.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur luctus ante nec sem pretium, vel tincidunt arcu imperdiet. Interdum et malesuada fames ac ante ipsum primis in faucibus. Donec auctor, nunc sed elementum mattis, urna ex commodo metus, nec mattis metus felis at turpis. Pellentesque tincidunt metus eros, in dapibus libero imperdiet in. Sed sit amet ipsum venenatis leo bibendum mollis eget non erat. Ut ut mauris a nisl euismod semper. Sed pharetra, dui eu tempus vulputate, neque nulla varius quam, eget consequat diam ante euismod dolor. Interdum et malesuada fames ac ante ipsum primis in faucibus. Quisque vitae tincidunt nisi, vel pulvinar nunc. Vestibulum quam neque, bibendum quis iaculis at, finibus ut neque. Maecenas tincidunt eget arcu vel aliquet. Proin non iaculis ante. Etiam blandit quam at arcu consectetur, vitae volutpat elit blandit. Nam in ornare nisi, quis egestas diam. In vestibulum mauris neque, vitae iaculis mauris tincidunt vitae.

Palavras-chave: Assinatura Digital de Vídeo, Descritor

LISTA DE FIGURAS

LISTA DE TABELAS

Sumário

1 Introdução

A Detecção de Cópias Baseada em Conteúdo é um método utilizado para a proteção de propriedade intelectual de mídias digitais. Ele consiste em extrair uma assinatura da mídia original e de uma mídia de teste, e então compará-las para identificar que se trata do mesmo conteúdo ou não. Esta abordagem assume que a mídia contenha alguma informação única que possa ser utilizada para identificar uma cópia Kim e Vasudev(4).

Suponha que uma plataforma online de publicação de vídeos (como YouTube ou Vimeo) receba vários pedidos vindos de criadores de conteúdo digital (como estúdios de cinema, cineastas independentes ou "vloggers") solicitando a retirada de vídeos da plataforma alegando o infringimento de direitos autorais. Segundo estes criadores, seus filmes estão sendo reproduzidos parcial ou integralmente na plataforma.

Enquanto o número de pedidos é baixo, não há grandes problemas em definir se realmente se trata de uma cópia: Basta assistir ao vídeo original e à suposta cópia, e então comparar os conteúdos. A medida que o número de pedidos cresce ou que os a extensão dos vídeos seja cada vez mais longa (filmes tendem a ter pelo menos duas horas de duração), torna-se claramente inviável a realização desse processo de forma manual.

Para realizar esta tarefa de forma automática, podem ser utilizados métodos de Detecção de Cópias Baseada em Conteúdo para a criação uma base de dados de assinaturas, para então ser realizada a busca e verificação de cópias.

Um dos serviços comerciais mais utilizados para este fim é o *Content ID* da Audible Magic. O método utilizado cria uma assinatura baseando-se no aúdio de uma mídia digital (podendo então ser utilizado tanto para vídeos, quanto para músicas), e disponibiliza um banco de dados global de assinaturas de conteúdos protegidos por propriedade intelectual. Este serviço é usado por grandes corporações como Facebook, SoundCloud e Vimeo Compliance. . . (5).

No entanto, o uso do áudio para a criação de assinaturas pode não ser o melhor atributo utilizado na criação de uma assinatura, pois vídeos (principalmente longas-metragens) tendem a ter versões dubladas na língua de cada país onde são distribuídos.

Para combater este problema, foram propostos inúmeros métodos que utilizam informações contidas nos quadros dos vídeos para a geração das assinaturas. Este trabalho propõe-se a fazer uma revisão da literatura de métodos para Detecção de

Cópias Baseada em Conteúdo, selecionar seis candidatos que possuam abordagens distintas, compará-los e avaliar quais atributos do conteúdo digital devem ser escolhidos a fim de otimizar a acurácia na detecção de cópias.

1.1 Objetivo Geral

Este trabalho tem como objetivo comparar descritores para assinatura digital de vídeos.

1.2 Objetivos Específicos

- Implementar 6 algoritmos, locais e globais, para assinatura digital de vídeos.
- Propor um novo descritor que combine características espaciais e temporais do vídeo.
- Compilar uma base de vídeos para a realização dos testes e avaliação dos algoritmos.
- Comparar os algoritmos com base na semelhança da assinatura e contrastar os métodos de comparação utilizados.

2 Definições

Nas próximas seções estão definidos conceitos utilizados neste trabalho, como quadro (*frame*), vídeo e tomada, além da definição de assinatura de vídeo e algumas características consideradas importantes na geração destas.

2.1 Definição de Vídeo

2.1.1 Quadro

Um quadro (ou *frame*) I é uma matriz de altura h e largura w. Cada ponto $I[x,y]$ representa uma intensidade de pixel [Simoes et al.(6)]. Além disso, um quadro possuí um determinado tempo, que representa o instante em que aparece em um vídeo. Portanto, seguindo a definição de Simoes et al.(6), um vídeo é representado por uma sequência de quadros, todos dispostos através da amostra temporal.

Um vídeo pode ser segmentado em cenas, que por sua vez possuem diferentes tomadas, compostas por diferentes quadros em sequência, conforme Figura 1. As tomadas de um vídeo, como pode ser observado na Figura 1, são definidas por um ou mais quadros capturados continuamente, representando uma ação ininterrupta no tempo e no espaço, de acordo com Davenport, Smith e Pincever(7). Um conjunto de tomadas semelhantes, por sua vez, é intitulado de *cena*.

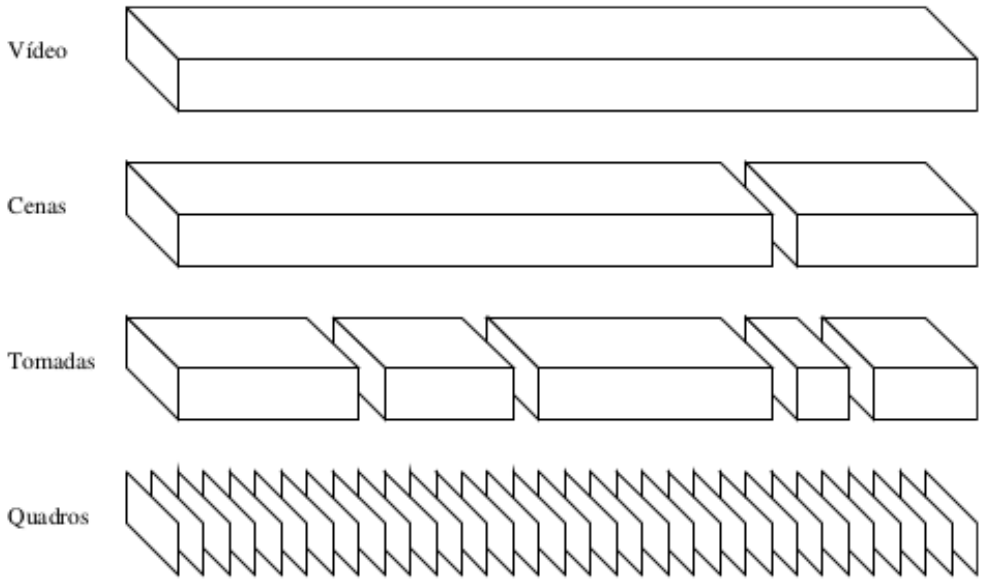


Figura 1 – Estrutura de um vídeo. Referência: Santos(1)

2.2 Definição de Assinatura

Uma assinatura de vídeo é definida como um vetor de características que representa um vídeo e o diferencia de outros Lee e Yoo(8). Outro termo utilizado para fazer essa referência é o de descritor, uma vez que a assinatura irá descrever um determinado vídeo. As características do vídeo podem ser obtidas de diferentes formas, as quais serão apresentadas na seção 5.5.

Para um algoritmo de geração de assinatura ser considerado eficiente, é importante que três características sejam consideradas: robustez, singularidade e eficiência de busca. De acordo com Lee e Yoo(8), uma assinatura é considerada robusta caso o descritor gerado para um vídeo modificado seja similar ao descritor do vídeo original. A singularidade é a capacidade do algoritmo de gerar assinaturas diferentes para vídeos perceptivelmente diferentes. Por fim, eficiência de busca é a capacidade da assinatura ser utilizada por uma aplicação para buscas em banco de dados de larga escala. Nesta monografia, são avaliados apenas a robustez e a singularidade das assinaturas.

2.3 Definição de Quadro de Cena

De acordo com Mao et al.(3), um quadro de cena pode ser representado por quadros que sigam duas características: deve ser uma imagem, como definido em

2.1.1; e os mesmos objetos, juntamente com um *background* altamente similar, devem pertencer a uma mesma cena.

[ARRUMAR: Procurar outras definições de quadros de cena]

2.4 Tipos de descritores de vídeo

2.5 Descritores de imagem

Em visão computacional pode-se descrever imagens utilizando algoritmos de identificação de pontos de interesse, sendo conhecidos como descritores locais. Há também descritores globais, que consideram características da imagem como um todo, e não apenas pontos de interesse. Cada abordagem tem vantagens e desvantagens distintas frente a cada tipo de variação ou ataque às imagens, porém, têm o mesmo objetivo, gerar uma descrição do imagem, normalmente tendo por saída um vetor, que pode ser considerado a assinatura da imagem. Idealmente, um descritor de imagens, seja global ou local, deve gerar assinaturas com robustez mediante os principais tipos de variação fotométricas, como borrados, iluminação, ruído e compressão JPEG e também em relação a variações geométricas, como rotação, escala (zoom), translação e ponto de referência (viewpoint).

2.5.1 Descritores Locais

Descritores locais utilizam características locais de cada imagem, comumente denominadas pontos e regiões de interesse para gerar sua saída. Uma imagem pode ter vários pontos ou regiões de interesse. Os pontos de interesse podem ser determinados a partir de regiões que possuem uma acentuada variação na orientação do gradiente dos objetos presentes em uma imagem, como por exemplo o enquadramento de uma porta ou o pico de uma montanha. As regiões de interesse são determinadas pelos pixels ao redor de um ponto de interesse e ajudam a determinar os limites de cada ponto de interesse. Aproveitando o exemplo anterior, uma região se interesse seria a região ao redor da intersecção do enquadramento de uma porta ou passagem.

Existem algoritmos de descrição local que são baseados em pontos de interesse, algoritmos que são baseados em regiões de interesse e algoritmos que são baseados tanto em pontos de interesse quanto em regiões de interesse. O que determina a escolha do algoritmo é o tipo de ataque ou distorção que se deseja combater, detectar ou evitar com o descritor (Krystian Mikolajczyk and Cordelia Schmid).

Descritores locais são robustos contra variações fotométricas (borrados, ilumina-

ção, cores, ruído e compressão JPEG) e normalmente mais custosos computacionalmente que os descritores globais (A VANISHING POINT-BASED GLOBAL DESCRIPTOR FOR MANHATTAN SCENES). Um descritor local consiste normalmente de três etapas: detecção das características, descrição das características e combinação das características, de acordo com (Zen Chen and Shu-Kuo Sun 2010).

Na primeira etapa do processo são determinados os pontos e regiões de interesse, como por exemplo o detector de pontos de interesse SIFT, que utiliza uma janela circular na imagem. Esta abordagem com o SIFT, entretanto, não é robusta contra variações geométricas. Na sequência, é definida uma janela de tamanho fixo e é utilizado um descritor das características de região para essa janela. Com todas as regiões de interesse descritas, a última etapa é definir uma função para combinar as diferentes regiões de interesse e assim gerar o vetor de descrição da imagem. (Zen Chen and Shu-Kuo Sun 2010)

2.5.2 Descritores Globais

Ao contrário dos descritores locais, descritores globais trabalham com a imagem como um todo, dessa forma, não são priorizados os pontos de interesse. Entretanto, existem trabalhos [citar quais] que utilizam regiões de interesse na implementação dos seus algoritmos, como é o caso do FrameDiff, que será apresentado na seção [indicar seção].

As regiões de interesse para descritores globais são determinadas de maneira diferente dos descritores locais. Para os globais, essas regiões normalmente são divisões da imagem, sem considerar de maneira discriminativa o conteúdo de cada região. Cada algoritmo de descrição global implementa a divisão da imagem de maneira diferente, como por exemplo o [algoritmo xxxxx] que divide a imagem em nove retângulos de igual tamanho, e também o [algoritmo yyyyy] que cria um círculo no centro da imagem e o divide em forma de fatias de pizza. Uma variação do [algoritmo yyyyy], por exemplo, determina uma elipse a partir do centro da imagem e também realiza a divisão em forma de pizza, resultando nas regiões de interesse. Todavia, existem algoritmos que não utilizam nenhum tipo de divisão ou regiões de interesse, trabalhando apenas com as características da imagem.

Uma das principais vantagens dos descritores globais é ter menor custo computacional do que a maioria dos descritores locais, segundo [fonte], devido à quantidade inferior de cálculos e passos necessários para a extração da assinatura. Além disso, há uma tendência dos descritores globais gerarem assinaturas mais robustas quando considerados os ataques geométricos às imagens, de acordo com [citar fonte], devido

à natureza dos algoritmos. Pode-se utilizar, por exemplo, o caso de ataques do tipo rotação e translação, conforme demonstra a imagem [imagem tal]. Descritores locais usualmente não conseguem gerar assinaturas robustas para esses casos, pois os pontos e regiões de interesse ocupariam uma posição diferente na imagem alterada em relação à imagem original, entretanto, um descritor global leva vantagem justamente por não considerar onde está a região de interesse, mas sim as suas características. O exemplo também vale para os algoritmos que não utilizam a divisão da imagem em regiões.

2.6 Descritores de vídeo

As duas classes de descritores apresentadas são descritores de imagens, e de acordo com a definição de vídeo, segundo [citar fonte], um vídeo é uma sequência de imagens. Os descritores locais e globais, entretanto, não podem garantir a singularidade, umas das principais características de um descritor, quando a entrada do algoritmo é um vetor de imagens ao invés de uma imagem isolada. É evidente que os algoritmos vão gerar as assinaturas para as suas entradas, baseadas em cada frame que compõe o vídeo, porém, não há garantias que o descritor consiga produzir uma assinatura que seja suficientemente semelhante para determinar se um vídeo que sofreu ataques geométricos e/ou fotométricos é a cópia de outro vídeo original.

O trabalho de [fonte] introduz duas abordagens para contornar o problema da singularidade da assinatura. De acordo com os autores, é possível verificar os elementos que estão presentes em uma ou mais cenas para auxiliar os algoritmos a produzir uma assinatura mais precisa dos vídeos.

2.6.1 Descritores espaciais

2.6.2 Descritores temporais

3 Estado da Arte

3.1 Detecção de Cópias Baseada em Conteúdo

Segundo Jiang et al.(9), o método de detecção de cópias baseada em conteúdo, do inglês *Content Based Copy Detection* (CBCD), tem se tornado uma alternativa à marca d'água, conforme ilustra a Figura 2, para identificar e proteger vídeos e seus direitos autorais através da criação de uma assinatura digital para o conteúdo. Apesar de o método CBCD vir sendo largamente utilizado em diversas aplicações, a detecção de cópias é um grande desafio e o trabalho de Jiang et al.(9) expõe determinadas limitações do método.

O conteúdo a ser procurado pode sofrer uma intensa redução em sua qualidade entre o vídeo original e a cópia, podendo até mesmo sofrer alterações, como por exemplo, distorções e transformações. É uma tarefa difícil buscar cópias através de mecanismos baseados em pesquisa por quadros sem uma ferramenta que consiga unir corretamente a duração de cada trecho a ser pesquisado com o vídeo original. Por último, é necessária uma representação compacta e eficiente das assinaturas para que se construa uma ferramenta de busca e detecção de cópias para grandes sistemas.

Figura 2 – Exemplo de marca d'água em uma imagem.

3.2 Recuperação de Vídeo Baseada em Conteúdo

De acordo com Law-To et al.(10), recuperação de vídeo baseado em conteúdo, do inglês *Content Based Video Retrieval* (CBVR) é o procedimento para gerar, pesquisar e analisar assinaturas digitais. Essa área de estudo produz os algoritmos que geram os identificadores digitais, estuda a similaridade entre assinaturas e também busca trechos em vídeos. Ainda segundo Law-To et al.(10), CBVR tem foco em procurar e encontrar vídeos em uma mesma categoria, como por exemplo, jogos de futebol ou vídeos sobre balões.

3.3 Recuperação de Imagens Baseada em Conteúdo

Conforme Gudivada e Raghavan(11), recuperação de imagens baseada em conteúdo, em inglês *Content Based Image Retrieval* (CBIR), é um sistema que auxilia na recuperação e extração de imagens de acordo com o conteúdo da imagem. Para representar uma imagem, o sistema CBIR se baseia em elementos visuais como cores, texturas e formas Vikhar e Karde(12). Há diversas aplicações que se beneficiam desta tecnologia, como: previsão do tempo, serviços de informações geográficas, design de interiores, galerias de arte, etc Gudivada e Raghavan(11).

4 Trabalhos relacionados

Existem diversos trabalhos correlatos que discutem assuntos como assinatura digital de vídeos, formas de obter as descrições, analisar e comparar assinaturas, métodos para realizar buscas em vídeos, descritores globais e locais, etc.

Em 1999, quando a visualização de vídeos na internet ainda estava em sua infância, Indyk, Iyengar e Shivakumar(13) já buscava métodos para encontrar vídeos pirateados na web. Este propôs um algoritmo para geração de assinatura temporal baseada nos limites das cenas de um vídeo. Embora seja boa para encontrar filmes inteiros, esta técnica não é apropriada para vídeos curtos, que dominam as redes sociais de vídeos (aproximadamente 4 minutos) comScore... (14).

Desde então, várias técnicas têm sido usadas para a geração de assinaturas. Coskun, Sankur e Memon(15) propôs o conceito de funções de *hash* como uma ferramenta para identificação de vídeo, criando uma algoritmo espaço-temporal baseado no diferencial da luminância entre regiões de quadros. Outras abordagens incluem o uso de descritores globais que utilizam a distribuição da intensidade de movimento e cor Hampapur, Hyun e Bolle(16), além de medidas ordinais Hua, Chen e Zhang(17), que se provaram robustas para variadas resoluções, mudanças de iluminação e formatos de vídeos.

Abordagens que utilizam características locais também foram extensamente pesquisadas, como em Joly, Buisson e Frelicot(18), cujo algoritmo apresentado busca ser eficiente para buscas em grandes bases de dados, tanto em velocidade quanto a qualidade. Há também a pesquisa de Law-To et al.(19), que usa o algoritmo de Harris para encontrar pontos de interesse no vídeo e criar uma assinatura compacta.

Andrade et al.(20) fez um estudo comparativo entre descritores globais e locais e mostrou como unir os dois tipos de descritores através de algoritmos genéticos. Andrade et al.(20) também fez vários experimentos mostrando que a combinação de descritores globais e locais são complementares e se usados em conjunto, produzem resultados superiores quando comparados com o uso individual de cada tipo de descritor.

Hu et al.(21) discorre sobre a indexação e recuperação de conteúdo em vídeos. O trabalho apresenta métodos para analisar a estrutura de vídeos, segmentação de cenas, extração de quadros-chave, características de movimento, mineração de informações em vídeos, mensuramento de similaridade e relevância entre assinaturas digitais, pesquisa de conteúdo em vídeos entre outros.

5 Metodologia

A metodologia do projeto consiste das seguintes etapas:

1. Revisão da literatura;
2. Definição e obtenção da base de vídeos
3. Definição dos algoritmos;
4. Implementação dos algoritmos;
5. Definição do método de comparação;
6. Implementação do método de comparação;
7. Validação das implementações;
8. Definição dos experimentos;
9. Experimentos;
    a) Compilação de uma nova base de vídeos;
    b) Geração dos vídeos com distorções;
    c) Geração da assinatura para todos os vídeos;
    d) Aplicação do método de comparação;
10. Análise dos resultados;

5.1 Revisão da Literatura

Foi realizada uma revisão da literatura pertinente à pesquisa. Este método de revisão da literatura foi escolhido pois define formas de identificar, interpretar e avaliar pesquisas disponíveis coerentes com o tema.

Além de artigos recentes relacionados à área, as referências de Block(2) também foram utilizadas na revisão e passaram pelo processo de filtragem, uma vez que este trabalho baseia-se no de Block(2).

5.1.1 Definições

O primeiro passo da revisão é a definição das perguntas que devem ser respondidas, das palavras-chave a serem utilizadas e dos critérios de exclusão de artigos. O intuito da revisão é definido a seguir:

- encontrar algoritmos para tratar diferentes problemas relacionados à descrição de vídeos e compreender seu funcionamento;
- encontrar métodos de *matching* de descritores;
- encontrar e compilar uma base de vídeos para realização dos testes;

Além das referências de Block(2), foram escolhidos repositórios em português e em inglês para a busca das palavras-chave. Os repositórios estão definidos na tabela 1 enquanto as palavras-chave na tabela 2.

Língua	Repositórios	Endereço
Inglês	IEEE	<http://ieeexplore.ieee.org/Xplore/home.jsp>
	ScienceDirect	<http://www.sciencedirect.com/>
	ACM	<http://dl.acm.org/>
	Google Scholar	<https://scholar.google.com.br/>
Português	Periódico CAPES	<http://www-periodicos-capes-gov-br.ez48.periodicos.capes.gov.br>
	SciELO	<http://www.scielo.org>

Tabela 1 – Lista de repositórios utilizados.

Português	Inglês
Detecção de cópia baseada em conteúdo	Content-Based Video Copy Detection
Descritor de vídeo	Video fingerprint
Detecção de cópia de video	video copy detection

Tabela 2 – Palavras-chave utilizadas na revisão sistemática.

Para auxiliar na indexação dos artigos achados e na sua utilização como base para este trabalho, os artigos foram divididos em categorias, como mostra a tabela 3.

	Categoria	Descrição
1	Descritores Globais	Algoritmos que produzem descritores baseados em características globais de um vídeo
2	Descritores Locais	Algoritmos que produzem descritores baseados em características locais de um vídeo
3	Comparação de Descritores	Métodos para comparação de descritores

Tabela 3 – Categorização dos artigos.

5.2 Definição e Obtenção da Base de Vídeos

Foi escolhida a base UCF50 - Action Recognition Data Set Reddy e Shah(22) e embora seja utilizada para reconhecimento de movimento humano baseado em ações, como por exemplo ciclismo, natação, caminhada com o cachorro, TaiChi, etc., os motivos para sua escolha foram a quantidade de vídeos e sua licença de livre utilização.

Do total de 6.681 itens, foram selecionados 1.264, pois os vídeos estavam fragmentados em diversas cenas, ou seja, apenas uma cena de cada vídeo foi utilizada. Cada cena tem uma duração média de 3 a 9 segundos.

Para cada vídeo selecionado foram aplicadas 14 distorções, totalizando 18.960 itens, sendo destes, 17.696 vídeos resultantes das distorções.

[escrever das 14 distorções]

5.3 Definição e implementação dos algoritmos

Como este trabalho visa continuar a pesquisa em descritores de vídeo baseando-se em Block(2), os primeiros algoritmos estudados e implementados são os três algoritmos já implementados por ele, descritos nas Seções 5.6, 5.7 e 5.8.

Embora Block(2) tenha optado por comparar apenas algoritmos de descrição global, devido a sua simplicidade de implementação, neste trabalho serão utilizados tanto descritores locais quanto globais.

POR QUE DOS OUTROS TRÊS ALGORITMOS? [sao recentes e locais]

Foi escolhida a linguagem Python (versão 2.7.12) para a implementação dos algoritmos por ser multiplataforma e por sua simplicidade na integração com bibliotecas de alta-performance implementadas em linguagens de baixo nível (C). Isto foi importante pois serão utilizadas as bibliotecas *OpenCV* (versão 3.0.0) e *NumPy* (versão 1.12.1) para manipulação e operação de imagens.

Para a compilação da base de vídeos foi utilizada a biblioteca MagicImage (versão 7.0.7-8-Q16-x64 - Windows), que incluí os programas *convert* e *ffmpeg*, utilizados para conversão de formatos, aplicação de distorções e transformações de vídeo e imagem respectivamente.

5.4 Definição e Implementação do método de comparação

pegar base de assinaturas distrocidas, e para cada distorçao, calcular a distancia para todas as assinaturas originais

pegar as duas menores distancias para cada distorçao de cada vídeo

se a dif do primeiro colocado (primera distancia) for diferente o suficiente para o segundo colocado , entao podemos assumir (podemos?) que encontramos a assinatura do vídeo original

EMBASAR ESSAS PARADAS LOUCAS

5.5 Validação das implementações

Antes de realizar-se qualquer experimento, é necessário que os algoritmos implementados passem por uma etapa de validação que consiste em comparar seus resultados com os dos artigos de base. Para isso serão revisados os artigos dos quais os algoritmos foram retirados, além de códigos disponibilizados através destes trabalhos.

A validação ocorrerá em duas etapas:

1. teste de software;

2. comparação das saídas dos algoritmos implementados para o trabalho e as implementações originais, dada a mesma entrada.

Estudaremos em detalhes os algoritmos das seguintes seções.

5.6 Assinatura de vídeo baseada na medida ordinal

O algoritmo proposto por (17) baseia-se na intensidade dos *pixels* de cada quadro para compor a assinatura. O que se propõe é que, primeiramente, a taxa de amostragem, ou seja, a taxa de quadros por segundo (FPS, do inglês *frames per second*) do vídeo de entrada seja padronizada, para que a assinatura gerada fique mais compacta e seja tolerante a diferentes formatos de compressão, por exemplo. Além disso, o vídeo é convertido para escala de cinza.

Após esse pré-processamento, para cada quadro é realizada a divisão em $M \times N$ blocos, como pode ser observado na Figura 3, bem como o cálculo da intensidade média para cada um dos blocos. Esses valores são então colocados em ordem crescente e representam cada elemento que compõe a assinatura.

Figura 3 – Exemplo com a divisão em blocos, cálculo das intensidades médias e ordem atribuída a cada valor. Referência: (2)

5.7 Assinatura de vídeo baseada em gradientes

Este algoritmo, proposto por (8), utiliza a distribuição dos gradientes para geração de assinaturas. O primeiro passo é definir uma taxa de quadros por segundo (fps) fixa, além da conversão para escala de cinza. Outro procedimento utilizado é o redimensionamento dos quadros, tornando o método robusto independente da mudança de resolução do vídeo. Em seguida, o gradiente dos *pixels* de cada quadro é calculado da seguinte forma. Encontram-se os gradientes $\mathbb{G}x$ e $\mathbb{G}y$ através da equação 1.

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

O quadro é então dividido em $M \times N$ blocos, para quais é determinado o valor do centroide dos gradientes, criando assim um vetor com $(M \times N)$ elementos.

Para isso, é necessário encontrar a magnitude $w(x,y)$ e a orientação $\Theta(x,y)$, conforme mostra a Equação 2.

$$w(x,y) = \sqrt{\mathbb{G}x^2 + \mathbb{G}y^2} \qquad \Theta(x,y) = tan^{-1}\left(\frac{\mathbb{G}y}{\mathbb{G}x}\right) \tag{2}$$

Em seguida o centroide para cada bloco é obtido a partir do somatório do produto da magnitude e orientação, dividido pela somatória de todas as magnitudes daquele bloco, como pode ser observado na Equação 3:

$$[i] = \frac{\sum_{x,y \in b[i]} w(x,y)\Theta(x,y)}{\sum_{x,y \in b[i]} w(x,y)} \tag{3}$$

5.8 Assinatura de vídeo baseada na diferença entre quadros

O algoritmo proposto por (23) utiliza características globais, como luminância e diferença de luminância intra-quadros, para uma assinatura robusta (ao utilizar medidas que refletem a estrutura temporal do conteúdo) e eficiente (ao utilizar medidas simples de serem calculadas e comparadas). Para cada quadro de um vídeo são coletadas as seguintes características:

- *timestamp*, o tempo do quadro relativo ao início do vídeo
- Luminância Total (Y), soma da luminância de todos os pixels de um frame
- Luminância Máxima (Y_{max}), o valor do pixel mais brilhante do quadro
- Área do Frame (A), largura $\times$ altura do vídeo, útil para normalização
- Luminância diferencial (dY), a diferença absoluta de luminância pixel a pixel do quadro atual como quadro que estava visível há 100 milisegundos, a diferença resultante é somada (como mostra a equação 4).

$$dY = \sum_{x,y \in \mathbb{I},\mathbb{J}} |\mathbb{I}(x,y) - \mathbb{J}(x,y)| \tag{4}$$

Após a obtenção das características primárias, estas passam por um processo de refinamento no qual os vetores Y e dY são passados por filtros passa-baixa (figura 4). Além disso, duas outras características são derivadas das características principais e visam medir o quão imóvel uma sequência de frames é. Para isso, são definidas as medidas "Quietude" (equação 6) e "Créditos" (equação 5).

$$Quietude = 100 \times \left(\sqrt{\frac{\ln \frac{dY}{A}}{\ln 256}}\right) \tag{5}$$

$$Creditos = 100 \times \frac{\frac{Y_{max}}{256} + \left(1 - \left(\frac{\ln \frac{Y}{A}}{\ln 256}\right)^2\right)}{2} \tag{6}$$

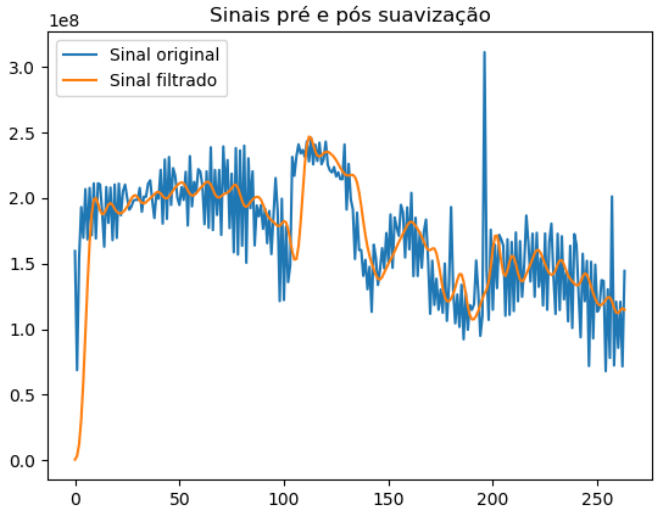


Figura 4 – Linha azul mostra os valores originais do vetor dY e a linha alaranjada mostra os valores pós filtro passa-baixa

A figura 5 mostra a característica dY plotada para um vídeo original e sua versão distorcida com efeitos de desfoque (*blur*) e adição de legenda, além de mostrar os valores de dY para outro vídeo não relacionado.

Para realizar a comparação entre duas assinaturas, Cook(23) propõe o uso da distância de Manhattan normalizada, como descrito na fórmula 7. Além disso, é proposto o uso de uma combinação das características Y e dY na comparação, pois separadas elas obteram 0.579% e 0.157% de falsos positivos, respectivamente, enquanto que a combinação das duas características obteve apenas 0.018%.

$$Distancia(a,b) = \frac{\sum_i |a_i - b_i|}{|a|} \tag{7}$$

5.9 Assinatura de vídeo baseada em wavelets

Esta abordagem foi escolhida por ter sido projetada especialmente para ser robusta a uma variedade de ataques fotométricos e de pós-produção, como modificações em contraste, brilho, contaminação por ruído e desfoque, inserção de logos, bordas e mudança de formato do quadro. Para se tornar ainda mais robustas a estes ataques, Dutta, Saha e Chanda(24) também descreve um fluxo de pré-processamento.

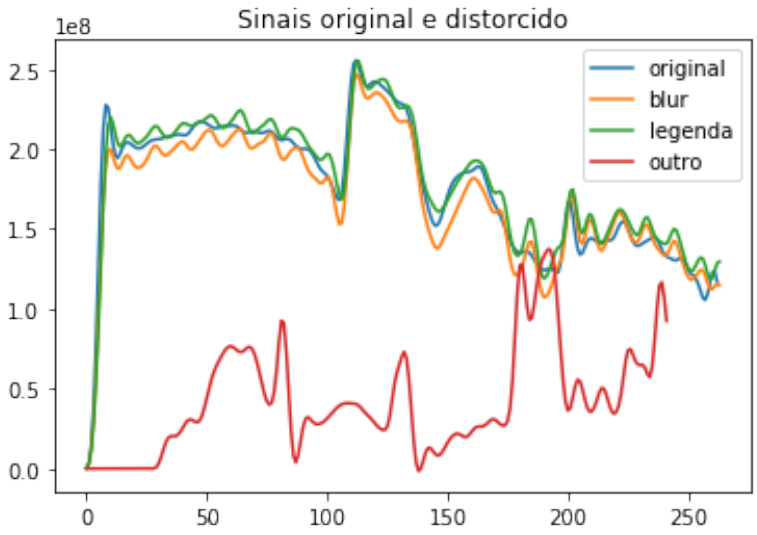


Figura 5 – Comparação entre os vetores de característica dY gerados para um vídeo, uma cópia com o efeito de desfoque, uma cópia com uma legenda inserida, e um outro vídeo qualquer

Após a etapa de pré-processamento e para ser usado como entrada para este algoritmo, os vídeos devem ser transformados para escala de cinza e ter suas intensidades normalizadas para o intervalo $[0,1]$. A assinatura proposta por Dutta, Saha e Chanda(24) é composta de uma parte baseada na transformada de Haar e em outra baseda na distribuição espacial de gradientes.

5.9.1 Transformada de Haar

A transformada de Haar consiste em separar um sinal (ou quadro, neste caso) em 4 partes:

- LL, que contém $1/4$ dos dados originais, removendo os detalhes.
- LH, que contém a derivada na horizontal do quadro
- HL, que contém a derivada na vertical do quadro
- HH, que contém a derivada na diagonal do quadro

Ela pode ser aplicada de forma recursiva n vezes, usando o quadro I como entrada inicial do algoritmo e o LL como entrada das chamadas subsequentes. Um exemplo do resultado da transformada de Haar pode ser visto na Figura 6.

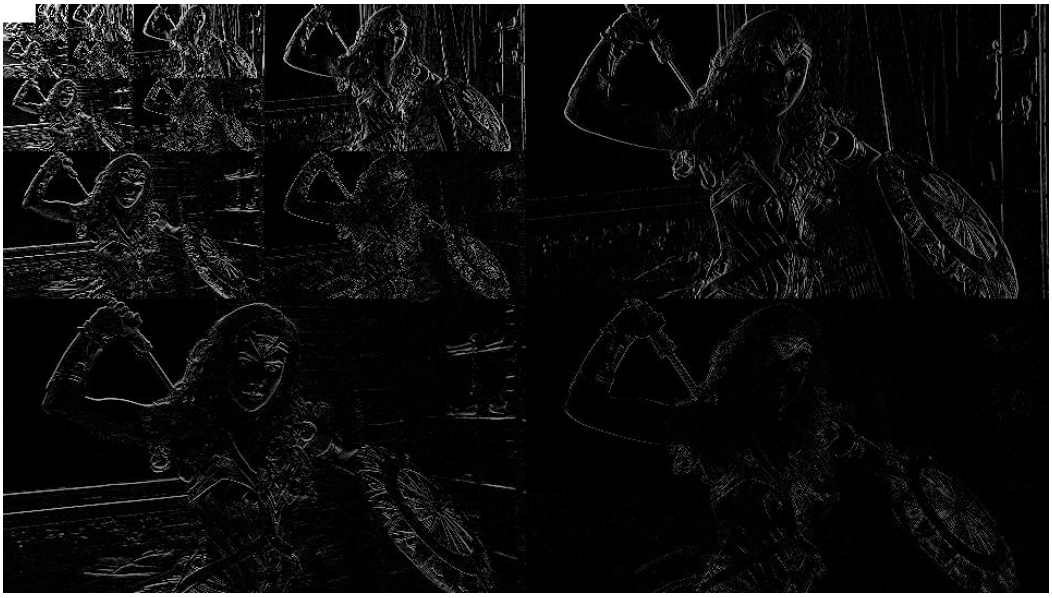
5.9.2 Assinatura baseada em wavelets

1. Seja I a imagem obtida após conversão para escala de cinza e normalização
2. Para i de 1 até n, onde n é o número de iterações da transformada de Haar

   a) Aplicar a transformada de Haar sobre I para obter um vetor com (LL, LH, HL, HH)
   b) Computar energia de LH, HL, HH [1]
3. Computar a energia da subimagem I, do último valor de LL e concatenar estes valores em um vetor,
4. Concatenar os valores de energia obtidos nos passos anteriores para obter o descritor

a. Quadro original

b. Quadro em escala de cinza



c. Após transformada de Haar

Figura 6 – Sequência de transformações feitas pelo algoritmo utilizando um quadro do filme Mulher Maravilha

5.9.3 Assinatura baseada em distribuição espacial de gradientes

1. Dado um quadro I já pré-processado, aplicar a transformada de Haar com um nível e obter um vetor com LL,LH,HL,HH

[1] $\frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}|\mathbb{I}(x,y)|$

2. Computar o gradiente de LL[2]
3. Dividir a imagem de gradiente em N_p partições com o mesmo tamanho
4. Para cada partição, computar um histograma de gradiente
5. Concatenar os histogramas para obter o descritor

5.10 Assinatura baseada em quadros de cena

Outra abordagem, proposta por Mao et al.(3), é baseada na assinatura de quadros de cena. De acordo com os autores, os quadros de cena podem ser *intraframes*, ou seja, quadros que iniciam tomadas, quanto *interframes*, contanto que sigam as características descritas em 2.3.

O algoritmo fundamenta-se na ideia de que as chances de existirem cinco quadros de cena seguidos é extremamente baixa, por isso são selecionados apenas os cinco primeiros quadros de cena de um vídeo. A forma como estes são determinados é descrita na Seção 5.10.1.

5.10.1 A extração de assinatura

A obtenção da assinatura, para Mao et al.(3), é realizada para todos os quadros, para então serem comparadas e selecionadas. Como pode ser observado no Diagrama **??**, os quadros passam por um pre-processamento, onde o componente de luminância é obtido. O quadro então é recortando, mantendo-se apenas sua região central e, por fim, redimensionado para o tamanho definido de $3/4$QCIF, ou seja, (108×132).

5.10.2 Diminuição do espaço de memória utilizado

O artigo também propõe uma alternativa para diminuir o espaço de memória utilizado para armazenar as assinaturas, visto que o banco de dados dos vídeos pode ser grande. Para isso, é proposta uma técnica chamada qualificação quaternária, na qual os valores são classificados de acordo com um *threshold*.

[2] A imagem LL é usada pois contém menos ruído que a imagem original graças à transformada de Haar

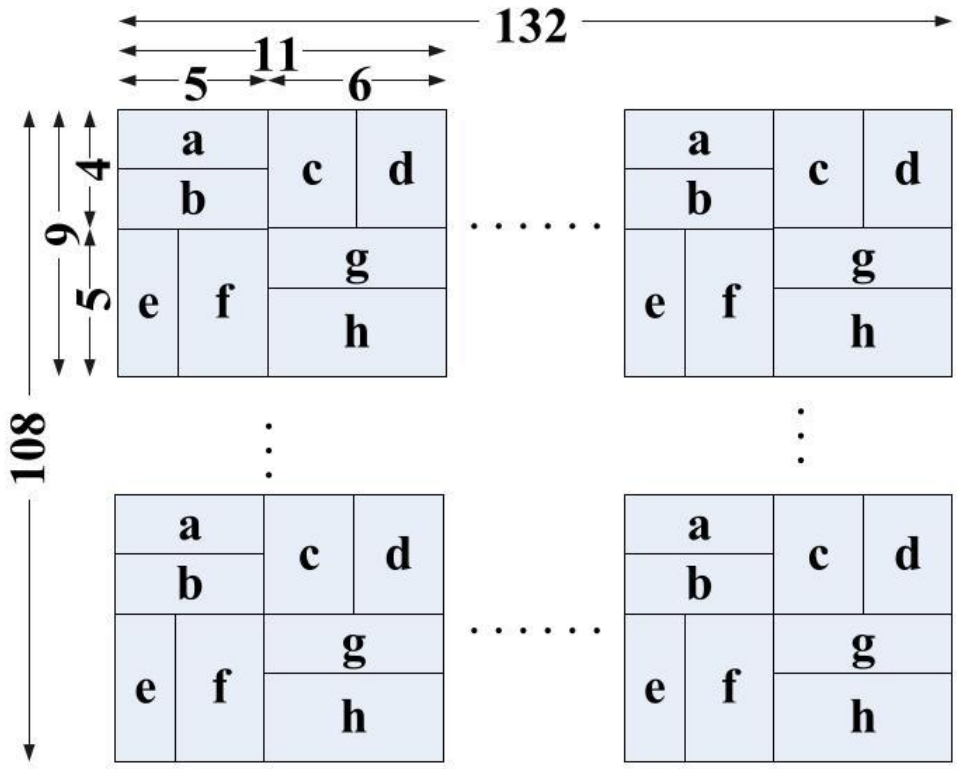


Figura 7 – Divisão da imagem para cálculo dos elementos diferenciais. Referência: Mao et al.(3)

Referências

1 SANTOS, T. T. **Segmentaçao automática de tomadas em vıdeo**. Tese (Doutorado) — Universidade de Sao Paulo, 2004.

2 BLOCK, S. A. B. Estudo das assinaturas digitais para a identificação de vídeos. . 2015. 28 f. trabalho de conclusão de curso (graduação). **Universidade Tecnológica Federal do Paraná**, Curitiba, 2015.

3 MAO, J. et al. A method for video authenticity based on the fingerprint of scene frame. 2015.

4 KIM, C.; VASUDEV, B. Spatiotemporal sequence matching for efficient video copy detection. **IEEE Transactions on Circuits and Systems for Video Technology**, IEEE, v. 15, n. 1, p. 127–132, 2005.

5 COMPLIANCE Automation for Media Sharing Platforms. <http://www.audiblemagic.com/compliance-service/#how-it-works>. Acessado em 14/03/2018.

6 SIMOES, N. C. et al. Detecçao de algumas transiçoes abruptas em sequencias de imagens. **Mestrado, Instituto de Computaçaoo, UNICAMP, Campinas**, v. 5, 2004.

7 DAVENPORT, G.; SMITH, T. A.; PINCEVER, N. Cinematic primitives for multimedia. **IEEE Computer graphics and Applications**, v. 11, n. 4, p. 67–74, 1991.

8 LEE, S.; YOO, C. D. Robust video fingerprinting based on affine covariant regions. In: IEEE. **Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on**. [S.l.], 2008. p. 1237–1240.

9 JIANG, M. et al. Pku-idm@ trecvid 2011 cbcd: content-based copy detection with cascade of multimodal features and temporal pyramid matching. In: **TRECVID Workshop: NIST**. [S.l.: s.n.], 2011.

10 LAW-TO, J. et al. Video copy detection: a comparative study. In: ACM. **Proceedings of the 6th ACM international conference on Image and video retrieval**. [S.l.], 2007. p. 371–378.

11 GUDIVADA, V. N.; RAGHAVAN, V. V. Content based image retrieval systems. **Computer**, IEEE, v. 28, n. 9, p. 18–22, 1995.

12 VIKHAR, P.; KARDE, P. Improved cbir system using edge histogram descriptor (ehd) and support vector machine (svm). In: IEEE. **ICT in Business Industry & Government (ICTBIG), International Conference on**. [S.l.], 2016. p. 1–5.

13 INDYK, P.; IYENGAR, G.; SHIVAKUMAR, N. **Finding pirated video sequences on the internet**. [S.l.], 1999.

14 COMSCORE Releases January 2014 U.S. Online Video Rankings. Disponível em: <http://www.comscore.com/Insights/Press-Releases/2014/2/comScore-Releases-January-2014-US-Online-Video-Rankings>.

15 COSKUN, B.; SANKUR, B.; MEMON, N. Spatio–temporal transform based video hashing. **ieee Transactions on Multimedia**, IEEE, v. 8, n. 6, p. 1190–1208, 2006.

16 HAMPAPUR, A.; HYUN, K.; BOLLE, R. M. Comparison of sequence matching techniques for video copy detection. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Electronic Imaging 2002**. [S.l.], 2001. p. 194–201.

17 HUA, X.-S.; CHEN, X.; ZHANG, H.-J. Robust video signature based on ordinal measure. In: IEEE. **Image Processing, 2004. ICIP'04. 2004 International Conference on**. [S.l.], 2004. v. 1, p. 685–688.

18 JOLY, A.; BUISSON, O.; FRELICOT, C. Content-based copy retrieval using distortion-based probabilistic similarity search. **ieee Transactions on Multimedia**, IEEE, v. 9, n. 2, p. 293–306, 2007.

19 LAW-TO, J. et al. Robust voting algorithm based on labels of behavior for video copy detection. In: ACM. **Proceedings of the 14th ACM international conference on Multimedia**. [S.l.], 2006. p. 835–844.

20 ANDRADE, F. d. S. P. de et al. Combinação de descritores locais e globais para recuperação de imagens e vídeos por conteúdo. Campinas, SP, 2012.

21 HU, W. et al. A survey on visual content-based video indexing and retrieval. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, IEEE, v. 41, n. 6, p. 797–819, 2011.

22 REDDY, K. K.; SHAH, M. Recognizing 50 human action categories of web videos. **Machine Vision and Applications**, Springer, v. 24, n. 5, p. 971–981, 2013.

23 COOK, R. An efficient, robust video fingerprinting system. In: IEEE. **Multimedia and Expo (ICME), 2011 IEEE International Conference on**. [S.l.], 2011. p. 1–6.

24 DUTTA, D.; SAHA, S. K.; CHANDA, B. An attack invariant scheme for content-based video copy detection. **Signal, Image and Video Processing**, v. 7, n. 4, p. 665–677, 2013. ISSN 1863-1711. Disponível em: <http://dx.doi.org/10.1007/s11760-013-0482-x>.