ISSN: 1984-3151

ANÁLISE COMPARATIVA DE DETECTORES DAS CARACTERÍSTICAS LOCAIS DE UMA IMAGEM NA RECUPERAÇÃO DE VÍDEOS

COMPARATIVE ANALYSIS OF LOCAL FEATURES DETECTORS OF IMAGES IN THE VIDEO RETRIEVAL

Juliana do Couto Matilde[1]; Márcio Sandro Silvestre[2]; Henrique Batista da Silva[3]

1 Bacharel em Ciência da Computação. Centro Universitário de Belo Horizonte - UniBH/MG, 2011. Companhia Energética de Minas Gerais. Belo Horizonte, MG. julianarcouto@gmail.com.

2 Bacharel em Ciência da Computação. Centro Universitário de Belo Horizonte - UniBH/MG, 2011. . Attps Informática. Belo Horizonte, MG. marciosilvestre@gmail.com.

3 Mestre em Ciência da Computação. PUC/MG, 2011. Doutorando na Universidade Federal de Minas Gerais. Belo Horizonte, MG. henrique.silva@dcc.ufmg.br.

Recebido em: 06/04/2012 - Aprovado em: 09/07/2012 - Disponibilizado em: 30/07/2012

RESUMO: A Internet tem se tornado um importante repositório de vídeos. Obter a informação desejada diante da enorme quantidade de dados é, muitas vezes, uma tarefa trabalhosa se não houver um sistema de busca eficiente. O sistema de recuperação baseado em evidências textuais é, atualmente, o mais utilizado para recuperação de vídeos. Porém, ele depende de fatores subjetivos impostos pela equipe que descreve o seu conteúdo e pelo usuário que busca pela informação, o que sugere o estudo de novas abordagens de recuperação da informação baseado em conteúdo, onde as características locais das imagens do vídeo são detectadas e descritas, produzindo um vocabulário visual. O objetivo deste trabalho é explorar essa abordagem, efetuando uma análise comparativa do desempenho de três desses detectores de características. Os resultados obtidos permitem afirmar que o detector Hessian-Affine obtém melhores resultados que os detectores Harris-Affine e MSER para o tipo de imagem pesquisada.
PALAVRAS-CHAVE: Recuperação da Informação. Recuperação de Vídeos baseado em conteúdo. Descrição de características de imagem. Detectores de características locais de imagem.

ABSTRACT: The Internet has become an important repository of videos. Get the desired information on the huge amount of data can become a chore unless we have an efficient search system. The retrieval system based on textual evidence is currently the most widely used for retrieval of videos, but it depends on subjective factors imposed by the team that describes the content and the user searching for information, which suggests the study of new approaches to information retrieval based on content, where the local characteristics of the video images are detected and described producing a visual vocabulary. The goal of this paper is to explore this approach by making a comparative analysis of the performance of three of these feature detectors. The results have revealed that the Hessian-Affine detector showed better results than Harris-Affine and MSER detectors for the type of image retrieved.
KEYWORDS: Information Retrieval. Video retrieval based on content. Description of imaging characteristics. Detectors local image features.

---

1 INTRODUÇÃO

Atualmente os vídeos têm se tornando cada vez mais populares, o advento da banda larga no Brasil e sua vasta aplicabilidade em sistemas de busca, bibliotecas digitais, programas de TV, sistemas de segurança, educação à distância, entretenimento, entre outros,

contribuem para a crescente disseminação desse tipo de mídia na Internet.

Com a elevada quantidade de informações disponibilizadas, torna-se cada vez mais trabalhoso encontrar a informação que se procura, sendo necessário o emprego de ferramentas que facilitem esse trabalho. Os Sistemas de Recuperação de Informação se tornam, então, aliados do usuário, contribuindo para uma busca mais rápida e eficiente.

A subárea da Recuperação da Informação aplicada neste trabalho é a Recuperação de Vídeos por meio de busca pelo conteúdo visual de seus quadros, uma vez que vídeos é o conjunto de vários quadros (imagens estáticas em sequência exibidas por um dado espaço de tempo).

Existem, na literatura, diversas abordagens para recuperação de imagens e vídeos, entre elas, destacam-se duas:

1. Recuperação de imagens baseada em evidências textuais, em que a informação é buscada no texto ao redor da imagem (SILVA; LOBATO, 2008). Esta é, atualmente, a abordagem mais utilizada nos sistemas de recuperação de imagens e vídeo. Existem vários trabalhos que exploram essa técnica e alguns dos sites mais populares se utilizam deste método, porém ele apresenta uma deficiência pelo fato da descrição do conteúdo visual variar de acordo com a percepção da pessoa que descreve o vídeo e do usuário que está efetuando a busca (TORRES *et al*, 2008).

O site Youtube, mundialmente popular, é um exemplo de um sistema que utiliza a recuperação de vídeos baseado em evidências textuais, uma vez que ele retorna os vídeos pela ocorrência das palavras digitadas pelo usuário no texto que descreve o vídeo, e não pelo seu conteúdo visual.

2. Recuperação de imagens baseada no conteúdo, onde são extraídas e processadas as características contidas nos quadros do vídeo, tais como as cores, textura, forma, entre outras.

A metodologia utilizada na literatura segmenta, primeiramente, o vídeo em quadros. Desses quadros (a quantidade dos quadros pode ser reduzida para diminuição do volume, processo conhecido como sumarização) são então extraídas características indexadas em uma estrutura para busca posterior.

No processo de recuperação, através de um vídeo ou imagem de consulta, é realizado o mesmo processo de detecção e extração de características e, posteriormente, efetua-se a busca na estrutura de indexação. Serão retornados um ou mais vídeos que são visualmente semelhantes ao vídeo de consulta, ou até mesmo, um segmento do vídeo no qual tal imagem de consulta esteja presente.

No entanto, para descrição do conteúdo de um vídeo existem diversas abordagens na literatura. O tipo mais simples de representar os quadros que compõe o vídeo utiliza descritores de características de imagem de baixo nível (propriedades dos pixels), tais como cor, textura, região, ou borda, segundo (BROWNE; SMEATON, 2005). Outra abordagem utiliza descritores de características locais de uma imagem, como descritores de canto e regiões conexas de uma imagem.

Este trabalho tem foco nesta última abordagem, pois, como foi dito, os sistemas disponíveis para recuperação de vídeos ainda possuem deficiências, devido à técnica de trabalhar apenas com informações textuais. Sabe-se que estas são passíveis de falha humana, uma vez que os dados digitados não seguem um padrão e dependem exclusivamente da interpretação de quem descreve e do usuário.

Portanto, este trabalho visa explorar a recuperação de vídeo, baseada em conteúdo. Além de efetuar uma análise da utilização de diferentes detectores de características para descrição do conteúdo do vídeo,

pois todo o seu vocabulário visual é produzido pelo próprio sistema de forma automatizada.

Assim, o objetivo geral é realizar uma análise comparativa entre detectores de características locais de uma imagem na recuperação de vídeos. E os objetivos específicos são:

- Desenvolver um programa computacional para executar a segmentação e sumarização de um vídeo;
- Realizar a detecção e descrição das características dos quadros do vídeo, utilizando os detectores Haris-Affine, Hessian-Affine e MSER;
- Fazer a indexação e busca das imagens contidas nos vídeos e análise comparativa dos resultados obtidos.

Na segunda seção do artigo são apresentados os conceitos e trabalhos relacionados que fundamentam esta pesquisa, incluindo métodos de segmentação e sumarização de vídeos, além de detecção e descrição de características visuais. A terceira seção apresenta a metodologia utilizada para recuperação de vídeos baseada no conteúdo. A quarta seção inclui detalhes da implementação, descrição dos experimentos e a análise dos resultados. Por fim, na quinta seção são apresentadas as conclusões do trabalho.

2 TRABALHOS RELACIONADOS

Os vídeos são formados pela sequência de quadros que ao serem reproduzidos em determinada faixa de tempo dão a ilusão de movimento. Normalmente os vídeos são compostos de 24 a 30 quadros por segundo para que essa movimentação seja perceptível ao olho humano (AVILA, 2008).

Conforme ilustrado na FIG. 1, os quadros são a menor fração de um vídeo e o conjunto deles forma uma tomada, que, em outras palavras é uma sequência de quadros registrados pela mesma câmera. A sequência de tomadas forma uma cena, que se assemelha pela proximidade temporal. Por fim, o vídeo é formado pela sequência de cenas. (AVILA, 2008).

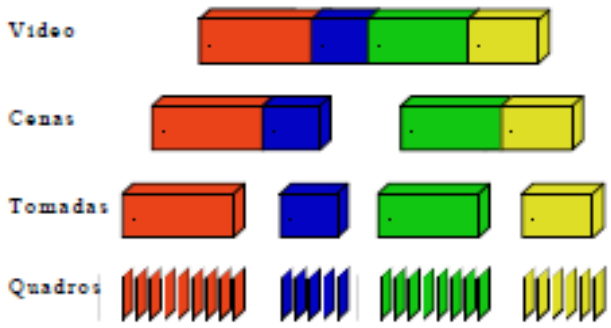

Figura 1 - Estruturação de vídeo
Fonte - AVILA, 2008, p. 9.

Os vídeos são compostos de muita informação e para que seja possível fazer a recuperação de seu conteúdo é necessário realizar um pré-processamento sobre as suas imagens: segmentação, sumarização, extração das características visuais e, na metodologia utilizada neste trabalho, a representação por um vocabulário visual. As etapas desse processo são descritas nas próximas seções.

2.1 SEGMENTAÇÃO

Segundo Ávila (2008) a segmentação de um vídeo é, geralmente, o primeiro passo para se trabalhar com suas imagens. Existem na literatura duas abordagens de segmentação de vídeo: A segmentação do vídeo em tomadas, que consiste em se identificar o limite entre tomadas consecutivas, que são uma sequência de quadros do vídeo gravados por uma mesma câmera e apresentam conteúdo visual muito similar; outra abordagem é a segmentação em quadros, onde há a separação do vídeo em sua menor parte, os quadros.

A segmentação é a decomposição do vídeo em um conjunto de imagens e uma importante preparação para a etapa posterior, onde é realizado um resumo do conteúdo original do vídeo.

2.2 SUMARIZAÇÃO

Sumarização é o processo de resumo do conteúdo original de um vídeo, tendo como objetivo a redução do volume de conteúdo sem perder informações relevantes, de forma a agilizar a busca. Tal processo pode ser dividido em 2 categorias: *keyframes* ou *vídeo skim* (AVILA, 2008). A primeira se caracteriza pela escolha de quadros-chave do vídeo original e resulta em um resumo estático. A segunda resulta em resumos dinâmicos, pois são selecionadas tomadas por meio de similaridade ou por uma relação temporal entre os quadros.

Após a segmentação e sumarização do vídeo é necessário utilizar técnicas para identificar e descrever as características dos quadros gerados por seus resumos.

2.3 DETECÇÃO DE CARACTERÍSTICAS

As características locais são padrões da imagem que se diferenciam da região imediatamente vizinha. Elas são associadas a mudanças de uma ou mais propriedades da imagem, como intensidade, cor e textura (TUYTELAARS; MIKOLAJCZYK, 2008).

As informações são retiradas de uma região em torno de uma característica local e convertidas em descritores. O conjunto de descritores forma uma representação da imagem que permite a identificação de objetos nela contidos.

Porém, na fase de interpretação das características pode ter variações, como por exemplo, no ponto de vista, escala ou rotação. Portanto, as regiões descritas pelos pontos devem ser covariantes, como apresentado no trabalho de Sivic e Zisserman (2003).

Regiões covariantes, segundo Sivic e Zisserman (2003), são regiões que devem corresponder às mesmas características da imagem original sob pontos de vista diferentes, ou seja, sua forma não é fixa, mas adapta-se automaticamente, de modo que as proporções aproximadas da imagem original sejam mantidas.

Vários detectores de regiões covariantes têm sido propostos na literatura. Dentre eles, são utilizados neste trabalho, os detectores Harris-Affine, Hessian-Affine e MSER.

Harris-Affine é utilizado para detectar cantos invariantes na imagem (que são regiões de alta curvatura). O resultado da aplicação do detector de Harris-Affine em duas imagens da mesma cena é apresentado na FIG. 2, na qual pode-se observar que apesar da deformação afim, as formas da região correspondem claramente.

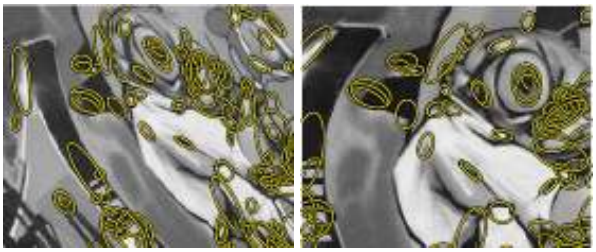

Figura 2 - Harris-Affine: regiões geradas por dois diferentes pontos de vista em uma cena
Fonte - TUYTELAARS; MIKOLAJCZYK, 2008, p. 226.

Hessian-Affine também pode ser considerado como um detector de cantos invariantes, porém ele detecta regiões parcialmente oclusas com mais eficiência que o detector Harris-Affine. O resultado do detector Hessian-Affine em duas imagens da mesma cena é apresentado na FIG 3, observe que seu grau de cobertura é maior que o anterior.

Figura 3 - Hessian-Affine: regiões geradas para duas visões de cena
Fonte - TUYTELAARS; MIKOLAJCZYK, 2008, p. 234.

E finalmente, o detector MSER (Maximally Stable Extremal Region) é utilizado para detectar regiões conexas da imagem por meio da intensidade do brilho de seus pixels. Segundo Tuytelaars e Mikolajczyk (2008) a palavra "Extremal" se refere à propriedade de que todos os pixels dentro de uma região possuem brilho mais intenso (regiões de brilho extremo) ou menos intenso (regiões extremamente escuras) que todos os pixels vizinhos. A expressão "Maximally Stable" descreve a propriedade otimizada no processo de seleção de limiar. Como visto em Mikolajczyk e Schmid (2005) a detecção realizada pelo MSER está relacionada ao limiar uma vez que cada região extrema é um componente conexo à imagem após a aplicação do limiar. As características detectadas pelo MSER se localizam principalmente nas bordas de regiões, resultando em detecções de localização mais precisa se comparadas às detecções de outros detectores. Seus resultados podem ser visualizados na FIG. 4.

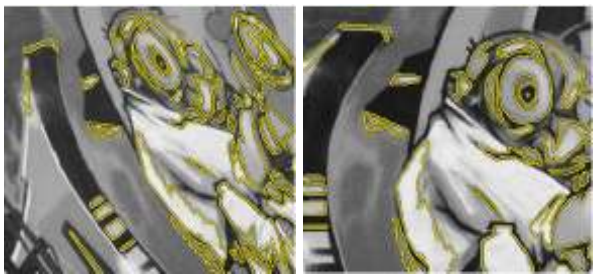

Figura 4 - Regiões detectadas pelo MSER
Fonte - TUYTELAARS; MIKOLAJCZYK, 2008, p. 240.

2.4 DESCRIÇÃO DE CARACTERÍSTICAS

Após a detecção das regiões locais em uma imagem o procedimento habitual é extrair um descritor de cada uma das regiões, ou seja, descrevê-las de forma padronizada para serem utilizadas no processo de busca. Um dos descritores mais utilizados na literatura é o SIFT (Scale Invariant Feature Transform), proposto por Lowe (2004).

O descritor SIFT transforma informações das imagens em coordenadas invariantes à escala relativas às regiões detectadas. Ele é formado por um vetor contendo os valores dos histogramas de orientação que representam gradientes de uma região da imagem. Esse vetor possui 128 dimensões e é criado para cada região de interesse na imagem (LOWE, 2004).

No entanto, a quantidade de informações contidas em cada quadro do vídeo ainda é muito grande para que sejam armazenadas, portanto é importante classificar os descritores em classe de palavras visuais. Criando assim um vocabulário visual, conforme apresentado em (SIVIC; ZISSERMAN, 2003).

Desta forma, com o objetivo de reduzir o volume de informação gerado no processo de extração de características dos quadros do vídeo, nem todos os descritores serão armazenados na estrutura de indexação, e sim apenas um conjunto destes descritores, que irão compor o vocabulário visual do vídeo, que será utilizado tanto na indexação, como na consulta.

2.5 INDEXAÇÃO DAS CARACTERÍSTICAS

Finalmente, as informações devem ser armazenadas e é utilizada uma estrutura de arquivo invertido para armazenar as palavras visuais do vocabulário. Essa estrutura é composta de um índice, onde são armazenadas as palavras do vocabulário visual, e cada palavra contém uma referência a uma lista de

quadros do vídeo que contém aquela palavra (COELHO; LAMARQUE; BERTHIER, 2001).

As informações indexadas serão utilizadas na fase final, onde é possível comparar se a imagem a ser buscada encontra-se na estrutura de indexação e exibir o resultado ao usuário (SIVIC; ZISSERMAN, 2003).

3 METODOLOGIA

Como apresentado anteriormente, os vídeos são compostos de muitas informações. Para que seja viável trabalhar com a grande quantidade de imagens é necessário dividir o processo nas etapas de segmentação e sumarização do vídeo, extração de características e finalmente, na etapa de indexação, em que uma estrutura é utilizada para armazenar o conteúdo do vídeo.

A etapa de segmentação e sumarização consiste na técnica de gerar um resumo estático do vídeo, pois ela resulta na redução da quantidade de informação que será indexada.

Primeiramente o vídeo é segmentado em quadros, ou seja, todo o seu conteúdo é decomposto em quadros.

Em seguida ocorre a sumarização, que consiste na extração de quadros-chave do vídeo. Neste trabalho foi selecionado um quadro por segundo do vídeo. O objetivo é que esse resumo represente todo o vídeo. Os quadros-chave selecionados são usados para criar o vocabulário visual e posteriormente para indexar o conteúdo do vídeo. No entanto, para criar o vocabulário visual, apenas uma amostra destes quadros-chave é obtida.

Segundo Tuytelaars e Mikolajczyk (2008) quando não se sabe o tipo de imagem a ser buscada é importante combinar diferentes tipos de detectores para se obter um melhor resultado. Pois eles possuem aplicabilidades diferentes. Portanto, neste trabalho, foram selecionados detectores que vão compor três diferentes técnicas de extração das características dessas imagens: Harris-Affine, Hessian-Affine e Maximaly Stable Extremal Regions (MSER) que foram abordados com mais detalhes em (TUYTELAARS, MIKOLAJCZYK, 2008).

Harris-Affine e Hessian-Affine são detectores de cantos (pontos na imagem de alta curvatura), baseado nos detectores Harris-Laplace e Hessian-Laplace, portanto invariantes à escala e também invariantes às transformações afins. Detectores afins covariantes conseguem detectar as regiões de interesse mesmo quando a imagem sofre uma transformação de escala com intensidades diferentes em cada direção (MIKOLAJCZYK; SCHMID, 2005).

MSER ou Maximally Stable Extremal Region é o terceiro detector a ser utilizado, ele detecta componentes conexos em uma imagem, na qual foi aplicada um limiar.

As características extraídas precisam ser descritas de forma padronizada para serem usadas nas etapas seguintes. Este processo é realizado pelo SIFT, que após ser configurado com os três detectores, Harris-Affine, Hessian-Affine e MSER, irá gerar, para cada quadro, um conjunto de vetores de 128 dimensões que descrevem aquele quadro de acordo com o seu detector, ou seja, são gerados três conjuntos de vetores por quadro.

Em função do alto custo computacional de processar e armazenar essa grande quantidade de informações se faz necessário extrair um novo resumo a partir dos resultados obtidos até a etapa atual. Neste trabalho é usada a mesma técnica abordada por SIVIC e ZISSERMAN, (2003), onde foi retirada uma amostra aleatória de aproximadamente 10% de cada conjunto de vetor obtido. Essa amostra é considerada como um resumo do vídeo baseado no detector.

Sendo assim, a amostra retirada dos vetores de características irá compor o vocabulário visual do

vídeo. As características extraídas serão transformadas em palavras visuais e armazenadas em um arquivo invertido.

Na estrutura de arquivo invertido, cada quadro do vídeo será considerado como um documento do arquivo invertido (análogo aos documentos dos sistemas de recuperação de informação textual), que será representado por um vetor, onde estão contidas as informações que descrevem seu conteúdo. O arquivo invertido representa o conjunto de todos os quadros e é estruturado como um índice, onde está listada uma entrada para cada palavra do vocabulário visual.

Para efetuar a busca por similaridade serão comparados os descritores do quadro de referência com todas as imagens do arquivo invertido, retornando os quadros que possuírem vetores mais próximos ao da referência. Esse processo se assemelha ao de busca por texto usado atualmente pelo Google (BRANSKY, 2004).

É importante salientar que para garantir a eficácia da comparação entre imagens é necessário não só que as características das imagens estejam bem descritas, mas também que a função de similaridade seja a mesma utilizada na etapa anterior.

4 EXPERIMENTOS

4.1 DETALHES DO DESENVOLVIMENTO DO PROGRAMA

Para o desenvolvimento do programa utilizado neste trabalho foi utilizada a linguagem C# na IDE do Visual Studio 2010, da Microsoft. Também foram utilizadas bibliotecas de segmentação de vídeo e de descrição de características de imagens desenvolvidas pelos autores Tuytelaars e Mikolajczyk (2008), e disponível para download na Internet.

A FIG. 5 mostra a Tela Inicial do programa exibindo as opções “Selecionar Vídeo”, que abre uma caixa de diálogo de seleção para o vídeo a ser processado, contendo: “Processar”, que inicia o processo de segmentação, sumarização, extração de características dos quadros, montagem do vocabulário visual e estrutura de indexação; e “Selecionar Imagem”, que abre uma caixa de diálogo de seleção para que seja informada a imagem que deve submeter à busca na estrutura de indexação obtida após o processamento do vídeo.

Figura 5 - Tela Inicial do Programa

A FIG. 6 mostra o programa reproduzindo o vídeo selecionado.

Figura 6 - Reprodução do Vídeo Selecionado

A Figura 7 mostra a caixa de diálogo de seleção da imagem a ser buscada no vídeo.

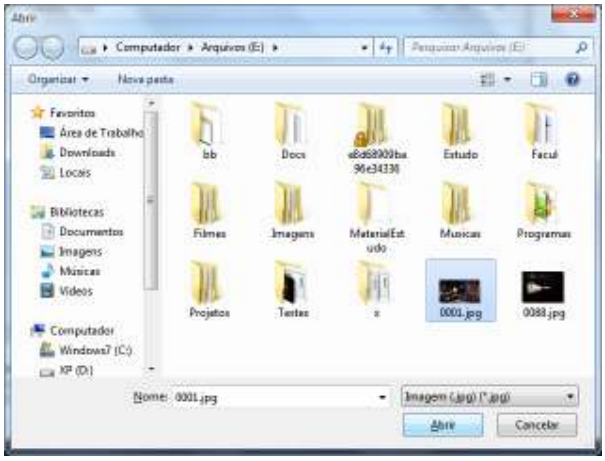

Figura 7 - Seleção da imagem de busca

Os resultados do processamento do programa são analisados através dos arquivos textos gerados em cada etapa do processo.

4.2 PARÂMETROS DE TESTES

Foi utilizado no teste um trecho de um capítulo da série The Big Bang Theory, da CBS (Columbia Broadcasting System). O vídeo possui 2min23seg de duração e taxa de 29 quadros/s. Realizada a segmentação foi obtido um total de 4293 quadros. Após a sumarização, com a retirada de um quadro por segundo, resultou-se o seguinte resumo: 4293 / 29 = 148 quadros (aproximadamente).

Os detectores de características Harris-Affine, Hessian-Affine e MSER foram aplicados separadamente em cada quadro-chave do vídeo, produzindo assim, um conjunto de regiões de interesses no quadro, baseado na característica do detector utilizado (canto ou região). Depois de geradas as regiões de interesse em cada quadro é necessário utilizar um descritor de regiões para gerar os vetores de características, que irão representar os quadros do vídeo. Para descrever as regiões identificadas pelos detectores foi utilizado o descritor SIFT (Scale Invariant Feature Transform), que produz um vetor de características de 128 dimensões. O SIFT é aplicado em cada região de interesse detectada nos quadros do vídeo, gerando um vetor de características para cada uma destas regiões.

Para compor as palavras visuais foi retirado, aleatoriamente, dez por cento dos vetores gerados por cada detector. Essa amostra representa todo o vídeo e foi armazenada em um arquivo invertido.

No processo de busca foram selecionadas para a consulta seis regiões retiradas de seis diferentes quadros do vídeo. A TAB. 1 quantifica a ocorrência de cada uma dessas regiões com base na análise visual de todos os quadros do vídeo, considerados para este trabalho. Portanto, as regiões selecionadas, somadas, aparecem 221 vezes nos quadros do vídeo. A diferença entre o somatório do número de vezes em que as regiões aparecem e o total de quadros selecionados se explica pela aparição, no mesmo quadro, de mais de uma região de consulta, fato que retornaria o mesmo quadro em busca de regiões diferentes.

Tabela 1

Quantidade de quadros por Região de busca

Consulta	Qtd
Região 1	4
Região 2	32
Região 3	32
Região 4	11
Região 5	67
Região 6	75
Total	221

4.3 MÉTRICAS

Para uma análise de desempenho dos resultados individuais, obtidos em cada detector de características, foram usados dois índices de avaliação: a precisão e a revocação, ambos abordados por Cardoso (2004).

A precisão é a fração dos documentos recuperados que são relevantes para a pesquisa. Neste experimento, 100% de precisão significa que todas as imagens recuperadas são relevantes, enquanto 0% revela que, de todas as imagens recuperadas, nenhuma é relevante. É importante observar que se nenhuma imagem for recuperada a precisão será de 100%, portanto é fundamental combinar os resultados de precisão e revocação para se obter melhor análise do desempenho de cada detector.

A revocação relaciona o número de resultados relevantes em relação ao total de itens relevantes, ou seja, uma alta taxa de revocação significa que a maior parte dos itens relevantes foi obtida. Neste experimento, 100% de revocação significam que todas as imagens relevantes foram encontradas corretamente, enquanto 0% revela que nenhuma imagem relevante foi encontrada.

4.4 Análises dos Resultados por Detector

Todos os quadros selecionados que compõem o vídeo analisado foram processados utilizando os três detectores. Os resultados estão apresentados na TAB. 2 e comentados nas sessões seguintes.

Tabela 2

Resultados obtidos por detector

Detector	Qtd	Acerto	FP	FN	R (%)	P (%)
Harris-Affine	221	48	0	173	22	100
Hessian-Affine	221	84	0	137	38	100
MSER	221	65	26	156	29	71

Os dados da TAB. 2 exibem as seguintes informações: Qtd (Quantidade de consultas realizadas); Acerto (Quantidade de imagens recuperadas corretamente); FP (Falso Positivo - quantidade de imagens recuperadas, mas que não correspondem ao objeto pesquisado); FN (Falso Negativo - quantidade de imagens que continham o objeto procurado, mas que não foram recuperadas); R(%) (taxa de Revocação); P(%) (taxa de Precisão).

A taxa de precisão é dada pela razão entre o número de acertos e a soma entre o próprio número de acertos com a quantidade de falsos positivos. A taxa de revocação é dada pela razão entre o número de acerto e a soma entre o próprio número de acertos com a quantidade de falsos negativos.

Os resultados apresentados na TAB. 2 mostram que o detector Harris-Affine demonstrou ser o menos eficiente na busca pelos objetos selecionados em relação a seu número de acertos. Ele obteve o pior desempenho entre os três analisados. O desempenho do detector Hessian-Affine foi o melhor dentre todos os detectores em relação ao número de acertos. Manteve o nível máximo de precisão, análogo ao resultado do Harris-Affine.

Tanto o Harris-Affine quanto o Hessina-Affine apresentaram taxas de precisão de 100%. Estes valores indicam que não houve erro em nenhuma das consultas retornadas, ou seja, todos os quadros do vídeo retornados continham o objeto da consulta. No entanto, o alto número de falsos negativos contribuiu para os valores baixos de revocação, pois diversos quadros em que o objeto de consulta estava presente não foram retornados como resultado da consulta.

Em relação ao detector MSER, os valores obtidos nos experimentos foram melhores daqueles valores obtidos pelo detector Harris-Affine, em relação ao número de acertos. No entanto, foi o único experimento que retornou quadros do vídeo que não continham o objeto da consulta realizada (imagens classificadas como falsos positivos), o que impactou diretamente em sua taxa de precisão. Sendo esta, abaixo dos experimentos realizados com os demais detectores.

Além dos resultados falsos positivos, os experimentos com o MSER também apresentaram uma alta taxa de

falsos negativos. Fator que contribuiu para sua baixa taxa de revocação.

Para melhor compreensão dos dados, o gráfico da FIG. 8 apresenta os resultados de Precisão e Revocação obtidos nos experimentos realizados para cada detector. No gráfico fica claro notar que todos os três detectores analisados apresentaram altas taxas de precisão, porém com baixos valores de revocação. Tal comportamento se deve justamente pela grande quantidade de quadros do vídeo que não foram retornados nas consultas realizadas (falsos negativos).

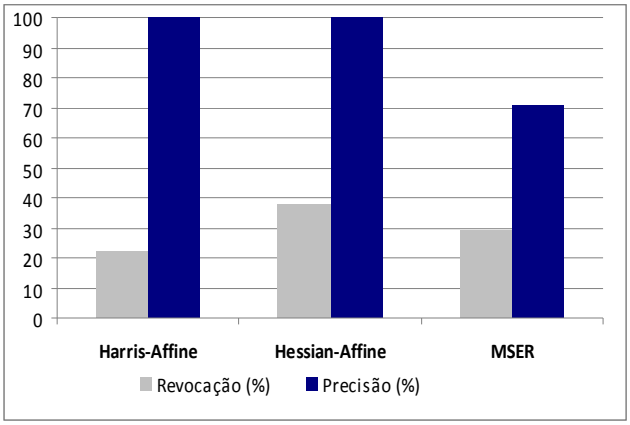

Figura 8 - Desempenho geral de cada detector

5 CONCLUSÕES

Uma grande quantidade de vídeos está disponível na Internet. Técnicas de indexação e recuperação eficazes deste tipo de conteúdo são essenciais para os usuários.

Neste trabalho foi efetuada uma análise comparativa entre os detectores de características locais de uma imagem: Harris-Affine, Hessian-Affine e MSER. Com base nos resultados obtidos é possível afirmar que, de maneira geral, o detector Hessian-Affine obtém melhores resultados que os detectores Harris-Affine e MSER. O ganho de desempenho, entretanto, está associado a um maior custo computacional para detecção das regiões de interesse por meio do detector Hessian-Affine.

Os resultados dos experimentos realizados mostram que todos os detectores avaliados apresentaram altas taxas de precisão, principalmente os detectores Hessian-Affine e Harris-Affine, que obtiveram 100% neste critério. Os experimentos realizados com o detector MSER com uma taxa de precisão um pouco menor, de 71%. No entanto, todos os experimentos realizados apresentaram resultados baixos para a taxa de revocação, ou seja, muitos quadros não foram retornados nas consultas realizadas (falsos negativos). Tal comportamento poderia ser justificado pela qualidade do vocabulário visual criado (índice do arquivo invertido), que pode não ter sido representativo o suficiente para o vídeo indexado.

Diante dos dados expostos e experimentos realizados, verificou-se um grande impeditivo para expansão das técnicas analisadas: o custo computacional elevado para a descrição de um vídeo e ainda maior para a indexação desses dados em arquivo invertido.

Testes realizados levaram até 18 horas de processamento para se obter os descritores das imagens, o vocabulário visual e a indexação por meio do arquivo invertido, considerando que o vídeo escolhido para este trabalho foi de pouco mais de 2 minutos de duração.

A recuperação de imagens e vídeos baseados em conteúdo é uma evolução natural do processo de Recuperação da Informação nesses formatos de mídias. Com os avanços cada vez mais rápidos das tecnologias e o aumento da necessidade desse tipo de recuperação de vídeos – que como foi relatado ao longo do trabalho produz resultados mais eficazes, eliminando a subjetividade da pesquisa com base em evidências textuais – acredita-se que em breve este será o mecanismo de recuperação de vídeos e imagens mais utilizado.

Como sugestões de trabalhos futuros, é importante ressaltar que experimentos com vocabulários visuais de diferentes tamanho, ou até mesmo outra abordagem para criação do vocabulário visual (como por meio de clusterização) poderiam melhorar a sua relevância e possivelmente impactando na melhoria dos resultados.

Experimentos com outros tipos de detectores presentes na literatura e com uma base de dados maior também poderiam ser realizados. Visto que o tempo de processamento torna-se um fator que dificulta testes com grandes bases de dados. No entanto, tais experimentos são importantes, considerando que, com a expansão da Internet, já é possível ter vídeos de horas de duração em sites como o Youtube, por exemplo.

AGRADECIMENTOS

Primeiramente a Deus, por trilhar nossos rumos e permitir este momento. À Prof.ª Magali Maria de Araújo Barroso, pelas minuciosas revisões e críticas construtivas. Aos demais professores pelo apoio, doação constante e pelos ensinamentos profundos. Por fim, a todos que colaboraram direta ou indiretamente na execução deste trabalho, nossos sinceros agradecimentos.

REFERÊNCIAS

AVILA, S. E. F. *Uma abordagem baseada em características de cor para a elaboração automática e avaliação subjetiva de resumos estáticos de vídeos*. Belo Horizonte: UFMG, 2008. 152p.

BRANSKY, R. M. *Recuperação de informações na web*. Perspect. Ciênc. Inf., Belo Horizonte, v.9, n.1, 2004. p . 70 – 87.

BROWNE, P.; SMEATON A. *Video retrieval using dialogue, keyframe similarity and video Objects*. Dublin City University, Irlanda, 2005. 4p.

CARDOSO, O. N. P. *Recuperação da Informação*. Lavras: UFLA, 2004., 6p.

COELHO, T. A. S.; LAMARQUE V.; BERTHIER R. N. *Recuperação de imagens na Web baseada em múltiplas evidências textuais*. Belo Horizonte: UFMG, 2001. 15p.

LOWE, D. G. *Distinctive Image Features from Scale-Invariant Keypoints*. University of British Columbia, Vancouver, Canadá, 2004. 28 p.

LOWE, D. G. *Object Recognition from Local Scale-Invariant Features*. University of British Columbia, Vancouver, Canadá, 1999. 8p.

MIKOLAJCZYK, K.; SCHMID, C. *A performance evaluation of local descriptors*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, v. 27, n. 10, 2005. p. 1615-1630.

SILVA, A.B.; LOBATO, M. C. C. *Recuperação de imagens na Web baseada em informações textuais*. Belo Horizonte: UFMG, 2008. 7p.

SIVIC J.; ZISSERMAN A. *Video Google: A Text Retrieval Approach to Object Matching in Videos*. United Kingdom: University of Oxford, 2003. 8 p.

TORRES R. S. et al. *Recuperação de imagens: Desafios e novos rumos*. Campinas: UNICAMP, 2008. 237p.

TUYTELAARS, T.; MIKOLAJCZYK, K. *Local Invariant Feature Detectors: A Survey*. Foundations and Trends. 2008. 104p.