Avaliação de técnicas de seleção de quadros-chave na recuperação de informação por conteúdo visual

Shênia Salvador de Pinho, Kleber J. F. Souza
Instituto de Ciências Exatas e Informática
PUC Minas
Guanhães, MG, Brasil
shenia.salvador@gmail.com, kleberjfsouza@gmail.com

***Resumo*—A recuperação de informação de conteúdo visual de vídeos digitais é uma forma de gerenciamento e busca de dados multimídias. Todavia, sem utilizar técnicas de seleção de quadros-chave no processo, todos os quadros do vídeo são utilizados, independente de sua relevância, gerando excesso de informações para armazenar. Neste trabalho é avaliado o comportamento de três métodos de seleção de quadro-chave: seleção de 1 em 1 segundo, 3 em 3 segundos e por pixel-a-pixel aplicado na recuperação de informação por conteúdo visual. Os resultados obtidos indicam que a relevância dos quadros-chaves selecionados pode ser determinante na obtenção de melhores resultados.**

***Palavras-chave*-Quadros-chave, Vídeo, Recuperação de Informação por Conteúdo.**

***Abstract*—The information retrieval of visual content of digital video is a form of management and search of multimedia data. However, selection techniques without using keyframes in the process, all the video frames are used, regardless of its relevance, generating too much information to store. In this paper we evaluate the performance of three selection methods keyframe: 1 selection in 1 second, 3 in 3 seconds and pixel-by-pixel applied in information retrieval by visual content. The results indicate that the relevance of selected key frames can be crucial in obtaining better results.**

***Keywords*-Keywords-Key frames, Video, Information Retrieval by Content.**

I. Introdução

Hoje em dia com as inovações tecnológicas que baratearam os preços dos dispositivos e da comunicação de dados, houve a popularização do compartilhamento de conteúdo visual e sua disponibilização em meio digitais, como exemplo o site *Youtube* que possibilita que seus usuários compartilhem dados multimídia.

Os usuários, ao publicarem seus vídeos na Internet, geralmente descrevem o assunto do vídeo através de uma descrição textual, ou seja, o próprio usuário retrata, ao publicar seu vídeo, o contexto ao qual ele se refere. De acordo com o trabalho de [1], a descrição do conteúdo do vídeo feita pelo usuário não é suficiente para realizar um controle eficaz de dados multimídias publicados. O motivo é que nem sempre o conteúdo original do vídeo condiz com o que o usuário descreveu, por isso surge a necessidade de se ter um controle destes dados publicados [2].

Outro cenário que deve ser levado em consideração está relacionado às diretrizes de sites referentes a publicação de vídeos que contém cenas de violência, nudez entre outras [3].

Para que estes conteúdos impróprios sejam publicados apenas com o consentimento do administrador do sistema [2], é muito importante ter um gerenciamento do conteúdo visual publicado [4] e esta automação pode ser realizada através de técnicas de recuperação de informação por conteúdo visual.

Sabendo disso, o presente trabalho visa avaliar o comportamento de diferentes técnicas de seleção de quadros-chave aplicada na recuperação de informação por conteúdo visual.

Segundo [5], no processo inicial da recuperação de informação pelo conteúdo visual é realizada a coleta de todos os quadros do vídeo. Entretanto os vídeos são compostos por inúmeros quadros, sendo necessário selecionar como base para a recuperação da informação, os quadros que melhor representem o vídeo [6].

A não utilização de técnicas de seleção de quadros-chave tem como consequência o excesso de informações armazenadas, pois todos os quadros são utilizados no processo de recuperação, independentemente de serem relevantes ou não, sendo elas: quadros negros, quadros repetidos, quadros sem relevância, entre outros.

Trabalhos que apresentaram ganhos com a utilização de técnicas de seleção de quadros-chave foi realizado [7] e [8] onde foi feito comparações entre técnicas de seleção de quadros-chave e obteve diminuição do custo computacional e armazenamento.

Outro trabalho que apresentou benefícios com a utilização de técnicas de seleção de quadros-chave foi o [9] que comparou seis técnicas com diferentes intervalos de tempo de seleção de quadros-chave sendo 0.5, 1.0, 1.5, 2.0, 2.5 e 3.0 segundos, pode-se observar que a técnica três segundos proporcionou melhora no desempenho e sabendo disso, faz-se necessário investigar os ganhos que se pode obter com a utilização de seleção de quadros-chave no processo de recuperação de informação pelo conteúdo visual, utilizando não só a técnica utilizada por [9], mas por outras técnicas de seleção de quadros-chave.

Este trabalho está organizado da seguinte forma: Primeiramente são apresentadas as técnicas de seleção de quadro-chave que foram utilizadas no trabalho, logo após, discute o teste dos algoritmos, o experimento que foi realizado e análise dos resultados. Para finalizar a conclusão da pesquisa e as

referências utilizadas.

II. TÉCNICAS DE SELEÇÃO DE QUADRO-CHAVE

A partir dos estudos realizados, foi possível definir as técnicas de seleção de quadros-chave que foram implementadas e analisadas no desenvolvimento deste trabalho. De acordo com os artigos estudados, foram selecionadas três técnicas são elas: comparação de quadro pixel-a-pixel, a técnica foi selecionada devido ser uma abordagem simples para comparar quadros do vídeo [7], seleção de quadros-chave a cada três segundos do vídeo, técnica escolhida, pois reduziu o custo computacional no trabalho [9]. Por fim, o método de seleção de quadros-chave a cada um segundo do vídeo [10], devido já ter sido utilizada no trabalho de Silva [11], o trabalho pelo qual foi utilizado seu processo de recuperação da informação por conteúdo visual.

A primeira técnica implementada foi à comparação de quadros pixel–a–pixel [7]. O algoritmo consiste em selecionar o primeiro quadro de um vídeo como quadro-chave e, a partir deste, calcular a distância Euclidiana entre todos os quadros seguintes. A distância entre o quadro-chave selecionado e o quadro que está sendo analisado é comparada com o limiar 80. O valor do limiar foi definido experimentalmente observando as diferenças visuais dos quadros-chave selecionados, ou seja, o que obteve entre os experimentos a menor taxa de quadros repetidos e a menor taxa de perda de quadros relevantes do vídeo.

Quando o valor do limiar na comparação na técnica de pixel–a–pixel é alcançado, o quadro que está sendo comparado é selecionado como o novo quadro-chave do vídeo e a partir deste momento, ele passa a ser a referência para as próximas comparações até a seleção do próximo quadro-chave.

O trabalho [9] utiliza o método de seleção de quadro-chave a cada três segundos do vídeo, ou seja, no intervalo de tempo a cada três em três segundos do vídeo é selecionado o quadro-chave. Por fim, a técnica de seleção de quadro-chave do trabalho de [10] foi implementada, onde é selecionado um quadro-chave a cada um segundo do vídeo. Podemos observar que as técnicas de seleção de quadro-chave a cada três segundos e um segundo do vídeo é feito a seleção por intervalos de tempo, todavia o método de seleção pixel–a–pixel foi realizado etapa de processamento de imagem nos quadros do vídeo.

Os quadros-chave selecionados em cada técnica de seleção foram utilizados no processo de recuperação de informação por conteúdo visual no trabalho [11].

O trabalho de [11] consiste em realizar comparações entre as métricas de indexação e recuperação de vídeos. Além disso, mostrou a quantidade de informações para se descrever uma imagem e como estas características influenciam diretamente a recuperação de informação por conteúdo.

III. TESTE DOS ALGORITMOS

Primeiramente foi realizado o teste dos algoritmos, com a finalidade de avaliar o funcionamento das técnicas de seleção de quadros-chave. Para o teste foram utilizados dois vídeos do projeto Open Video[1]. O Open Video é um projeto que visa à disponibilização gratuita de vídeos digitais para pesquisa de recuperação de multimídia.

Os vídeos aplicados no teste dos algoritmos são apresentados na Tabela I. São documentários da *University of Maryland* do ano de 1998, e foram obtidos no projeto Open Video, possuem som e cor.

Tabela I
VIDEOS DE TESTES

Vídeo	Nome do Vídeo	Duração	Total de Quadros
V1	*New Indians, Segment 08*	00:23	707
V2	*New Indians, Segment 09*	00:16	503

As três técnicas implementadas foram executadas em ambos os vídeos, com o objetivo de verificar a quantidade de quadros-chave que os algoritmos selecionavam. A Tabela II apresenta a quantidade de quadros-chave selecionados por cada técnica em cada vídeo.

Para verificar se tais resultados estavam corretos, na técnica de seleção 3 em 3 segundos e 1 em 1 segundo, foi realizado uma contagem manual da taxa de quadros por segundo do vídeo pela quantidade total dos quadros, sendo que, em ambos os vídeos esta taxa é de 29 quadros por segundo, ou seja, a cada 29 quadros na técnica de 1 em 1 segundo o quadro é selecionado como quadro-chave e cada 87 quadros para a técnica de 3 em 3 segundos. Na técnica pixel-a-pixel o resultado foi verificado pelas diferenças visuais entre os quadros selecionados. Interessante ressaltar que os quadros-chave selecionados pela técnica tiveram como resultado, quadros de tomadas distintas. Esta observação também foi verificada manualmente, apenas com a observação das características visuais dos quadros e tomadas.

Tabela II
RESULTADO DO TESTE

Vídeo	Pixel-a-Pixel	Seleção 3 seg.	Seleção 1 seg.
V1	5	8	24
V2	6	5	17

IV. EXPERIMENTOS

Depois da etapa de teste dos algoritmos, foi realizada uma análise das técnicas implementadas, utilizando a mesma abordagem de avaliação do trabalho de [11], a partir da base de quadros-chave selecionados foram analisados os acertos da busca de um objeto no vídeo. Vale ressaltar que o trabalho de [11] teve como objetivo principal a análise comparativa do uso de estruturas métricas para a indexação e a recuperação de vídeos utilizando vocabulário visual. O autor comparou três estruturas: *M-Tree*, *Slim-Tree* e *D-Index*.

Como o objetivo deste artigo não é avaliar tais estruturas métricas, uma destas estruturas foi escolhida para realização do experimento deste trabalho: a estrutura *D-Index*, é uma estrutura de indexação baseada em *hashing*. Seu objetivo é de

[1] http://www.open-video.org

minimizar o número de acessos a disco e número de cálculos de distâncias [11].

As técnicas de seleção de quadros-chave foram implementadas em C# no ambiente de programação Visual Studio 2010. Na realização dos experimentos foi utilizado o mesmo vídeo utilizado por [11]: *The Big Bang Theory* da CBS (*Columbia Broadcasting System*). Este vídeo tem duração de 19 minutos de 56 segundos, e possui 28.682 quadros.

Para o processo de recuperação de informação por conteúdo visual em vídeo, foram selecionados aleatoriamente 12 objetos do vídeo para fazer a busca pelo seu conteúdo visual, através de sua imagem e não de sua descrição textual, como se pode observar na Figura 1.

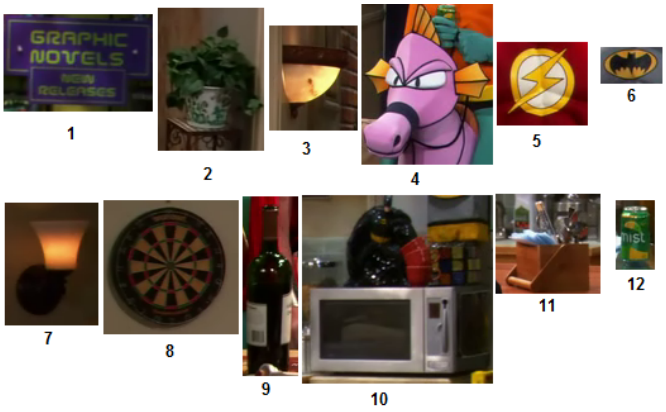

Figura 1. Objetos para consulta

Inicialmente foram executadas as 3 técnicas de seleção de quadros-chave no vídeo para selecionar as imagens que mais representavam o conteúdo. Sendo no processo de seleção de 1 em 1 segundo e 3 em 3 segundos foi feito com base na taxa de 23 quadros por segundo. Na Tabela III é apresentada a quantidade de quadros-chave selecionados por técnica:

Tabela III
RESULTADO DO TESTE

Pixel-a-Pixel	Seleção 3 seg.	Seleção 1 seg.
440	415	1247

Os quadros-chave selecionados pelas técnicas foram utilizados no processo de recuperação da informação por conteúdo visual no trabalho de [11]. Ou seja, foram aplicadas as 3 técnicas implementadas e o restante do processo de recuperação de informação por conteúdo visual é do trabalho de [11] como se pode observar na Figura 2.

Os objetos apresentados na Figura 1 foram utilizados no processo de busca. O objetivo do processo de recuperação é retornar as imagens que possuem os objetos pesquisados independente de suas posições.

A Figura 3 apresenta alguns exemplos de quadros que foram utilizados na base consulta dos objetos.

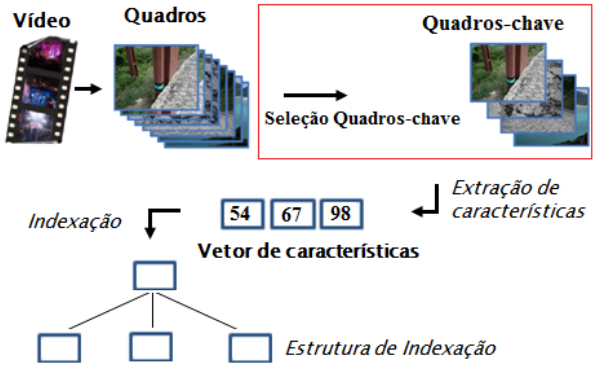

Figura 2. Indexação das características do vídeo

Figura 3. Quadros para consulta

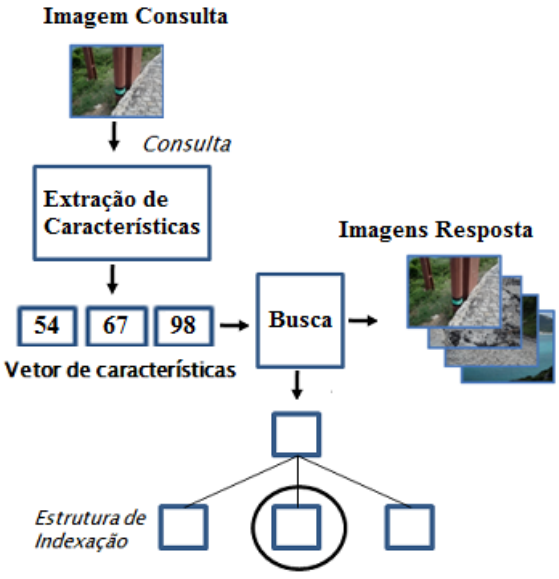

Figura 4. Recuperação de Informação por Conteúdo Visual [11]

A Figura 4 apresenta o processo de recuperação das imagens que possuem os objetos procurados no vídeo. Primeiramente, é feito um processo de extração das características da imagem do objeto que se deseja buscar no conteúdo do vídeo. Logo após, as características extraídas são armazenadas em um vetor de características que possui as informações sobre cor, textura e entre outras do conteúdo do vídeo [11], logo após é feita a busca na estrutura de indexação (gerada no processo de indexação das características do vídeo). Por fim, são retornadas todas as imagens que possuem características em comum com o objeto procurado.

V. ANÁLISE E RESULTADO

Na análise dos resultados das consultas, foi feito a "rotulação dos objetos verdadeiros" para cada técnica de seleção de quadros-chave, ou seja, manualmente foi contabilizada a frequência com que os objetos selecionados para recuperação da imagem aparecem nos quadros-chave selecionados do vídeo, de acordo com o mesmo foi realizada a comparação dos resultados.

A Tabela IV apresenta a análise quantitativa dos resultados da recuperação de informação visual da técnica de seleção de quadros-chave pixel-a-pixel.

Tabela IV
SELEÇÃO DE QUADRO-CHAVE TÉCNICA PIXEL-A-PIXEL

Consulta	rotulação dos objetos verdadeiros	Acertos	Revocação (%)
Objeto 01	11	9	81,82
Objeto 02	1	1	100
Objeto 03	13	3	23,08
Objeto 04	33	25	75,76
Objeto 05	61	29	47,54
Objeto 06	28	1	3,57
Objeto 07	20	13	65,00
Objeto 08	4	3	75,00
Objeto 09	4	3	75,00
Objeto 10	5	2	40,00
Objeto 11	5	0	0
Objeto 12	13	8	61,54
Média Revocação			54,02

A Tabela V apresenta a análise dos resultados da recuperação de informação visual da técnica de seleção de quadros-chave 3 em 3 segundos.

Por fim, a Tabela VI apresenta os resultados da recuperação de conteúdo visual técnica seleção 1 em 1 segundo.

Foram utilizadas métricas de revocação na comparação das três técnicas de seleção de quadros-chave, que avaliou a capacidade de se recuperar as imagens mais relevantes para o usuário. Para o resultado se calculou a quantidade de quadros que possui o objeto procurado, ou seja, os acertos, dividido pela quantidade de quadros da "rotulação dos objetos verdadeiros".

De acordo com as técnicas implementadas pode-se observar que o método pixel-a-pixel é o método que apresentou o melhor resultado de forma geral, obteve 54,02% da média de resultados relevantes retornados no processo de recuperação de

Tabela V
SELEÇÃO DE QUADRO-CHAVE TÉCNICA 3 EM 3 SEGUNDOS

Consulta	rotulação dos objetos verdadeiros	Acertos	Revocação (%)
Objeto 01	9	0	0
Objeto 02	2	0	0
Objeto 03	63	7	11,11
Objeto 04	41	9	21,95
Objeto 05	59	27	45,76
Objeto 06	55	0	0
Objeto 07	74	28	37,84
Objeto 08	8	1	12,50
Objeto 09	8	1	12,50
Objeto 10	8	7	87,50
Objeto 11	7	0	0
Objeto 12	26	18	69,23
Média Revocação			24,86

Tabela VI
SELEÇÃO DE QUADRO-CHAVE TÉCNICA 3 EM 3 SEGUNDOS

Consulta	*Groundtruth*	Acertos	Revocação (%)
Objeto 01	35	0	0
Objeto 02	07	0	0
Objeto 03	184	4	2,17
Objeto 04	131	2	1,53
Objeto 05	171	5	2,92
Objeto 06	157	23	14,60
Objeto 07	219	29	13,20
Objeto 08	21	2	9,52
Objeto 09	27	0	0
Objeto 10	19	0	0
Objeto 11	18	3	16,70
Objeto 12	80	12	15
Média Revocação			6,30

conteúdo visual. Obteve também 100% de acerto na consulta do objeto 02.

O método de seleção de 3 em 3 segundos obteve resultados de algumas consultas de forma satisfatória. Um exemplo é a consulta 12 onde se teve 18 acertos, tendo o melhor resultado nesta consulta entre as três técnicas. Entretanto, de forma geral alcançou apenas a média de 24,86% de resultados relevantes.

Por fim, o método de 1 em 1 segundo, que teve uma revocação média de apenas 6,30%, também apresentou vantagens, sendo uma delas a consulta do Objeto 06, onde a mesma acertou 23 quadros relevantes. Todavia, deve ser levado em consideração que a mesma técnica possuía uma base 157 quadros para a consulta, ou seja, em comparação com as outras técnicas, é a que possui uma maior quantidade de quadros na "rotulação dos objetos verdadeiros".

Pode-se observar que apesar da base de quadros selecionados pelas técnicas 1 em 1 segundo ser maior que a da técnica pixel-a-pixel, esta última obteve os resultados mais relevantes da consulta na maioria dos casos. Pode-se dizer que o resultado final da recuperação de informação por conteúdo visual possui uma dependência das entradas e das bases. Mesmo quando a base de quadros é maior, como o na técnica de seleção de 1 em 1 segundo, não houve influência, de uma maneira geral, no melhor resultado. Além disso, devido ao

fato dos dois primeiros métodos (i.e., 1 em 1 segundo e 3 em 3 segundos) não realizarem nenhum processamento sobre as imagens (apenas selecionou os quadros em relação ao tempo do vídeo) os resultados obtidos não foram tão relevante quanto à técnica pixel-a-pixel, que realiza a seleção dos quadros-chaves de maior relevância de acordo com um critério de pré-processamento.

VI. CONCLUSÕES

Com o compartilhamento de conteúdo visual na Internet, um grande volume de dados multimídias é publicado. Devido ao volume destes dados, faz-se necessário ter formas automáticas tanto para realizar o gerenciamento dos mesmos quanto para a realização de buscas, e esta automação pode ser realizada através de técnicas de recuperação por conteúdo visual.

De acordo com [10] este processo já é utilizado para o gerenciamento de dados multimídia. Entretanto, com a não aplicação técnicas de seleção de quadro-chave no processo inicial da recuperação por conteúdo de um vídeo, tem como consequência o excesso de informações armazenadas.

Portanto, o presente trabalho teve como objetivo avaliar o comportamento de diferentes técnicas de seleção de quadro-chave aplicado na recuperação de informação por conteúdo visual. Para isso, três técnicas de seleção de quadros-chave foram implementadas e avaliadas, são elas: pixel-a-pixel, 3 em 3 segundos e 1 em 1 segundo. Estas técnicas foram avaliadas utilizando o trabalho de [11], conforme apresentado no experimento realizado. De acordo com os resultados dos experimentos, verifica-se que a utilização de técnicas de seleção de quadros-chave na recuperação de informação por conteúdo visual de fato influenciam nos resultados obtidos. A técnica de seleção de quadros-chave pixel-a-pixel foi a técnica que obteve os melhores resultados, obtendo a media de 54,02% de resultados relevantes.

Este trabalho contribui para a área de processamento digital de imagens apresentando uma avaliação de técnicas de seleção de quadros-chave no processo recuperação de informação por conteúdo visual. Os resultados obtidos indicam que a quantidade de quadros selecionados não garante os melhores resultados, e, além disso, a relevância dos quadros selecionados pode ser determinante na obtenção de melhores resultados na recuperação de informação por conteúdo visual. Ou seja, algoritmos que realizam uma análise para selecionar os quadros relevantes, como a técnica pixel-a-pixel, tende a gerar melhores resultados.

REFERÊNCIAS

[1] A. P. B. Lopes, S. E. F. de Avila, A. N. A. Peixoto, R. S. Oliveira, M. de M. Coelho, and A. de Albuquerque Araújo, "Nude detection in video using bag-of-visual-features." in *SIBGRAPI*. IEEE Computer Society, 2009, pp. 224–231. [Online]. Available: http://dblp.uni-trier.de/db/conf/sibgrapi/sibgrapi2009.html#LopesAPOCA09

[2] C. C. Paiva and P. H. S. M. Serrano, "Critérios de categorização para os vídeos do youtube," in *X Congresso de Ciências da Comunicação na Região Nordeste*, São Luis, 2008.

[3] T. Deselaers, L. Pimenidis, and H. Ney, "Bag-of-visual-words models for adult image classification and filtering," in *19th International Conference on Pattern Recognition (ICPR 2008)*. Tampa, Florida, USA: IEEE, 2008, pp. 1–4.

[4] S. E. F. d. Avila, "Uma abordagem baseada em características de cor para a elaboração automática e avaliação subjetiva de resumos estáticos de vídeos," Master's thesis, Universidade Federal de Minas Gerais, 2008.

[5] M. L. R. Gomes, "Recuperação de vídeos por conteúdo com base em informação estáticas e dinâmicas," Master's thesis, Pontifícia Universidade Católica do Paraná, 2006.

[6] C. H. Morimoto and T. T. Santos, "Estruturação e indexação de vídeo digital," in *XVI Brazilian Symposium on Computer Graphics and Image Processing*, Salvador, BA, Brasil, 2003.

[7] C. A. F. Pimentel Filho, S. A. Celso, and T. A. Buck, "Integração de métodos baseados em diferença de quadros para sumarização do conteúdo de vídeos," in *WebMedia - XIV Simpósio Brasileiro de Sistemas Multimídia e Web*, vol. 2, Vila Velha, 2008, pp. 1–4.

[8] A. Santos, M. Bueno, A. Ferreira, J. Kelner, and I. SIQUEIRA, "Refinamento de reconstrução 3d através de seleção apropriada de keyframes," in *VIII Workshop de Realidade Virtual e Aumentada*, Uberaba-MG, 2011.

[9] G. A. Nascimento, M. G. Manzato, and R. Goularte, "Extraçãde quadros-chave como subsídio para personalização em vídeos digitais," in *Worksop de Iniciação Científica - XVI Simpósio Brasileiro de Sistemas Multimídia e Web, 2010*, vol. 2, Porto Alegre, 2010, pp. 105–107.

[10] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 591–606, 2009.

[11] H. B. Silva, "Análise de estruturas métricas para recuperação de vídeo utilizando vocabulário visual," Master's thesis, Pontifícia Universidade Católica de Minas Gerais, 2011.