Exploração de Métodos para Detecção Automática de Claquetes em *Rushes Videos*

Flávio Gonçalves Henriques de Souza
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
sflavio@dcc.ufmg.br

Tiago Oliveira Cunha
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
tocunha@dcc.ufmg.br

Arnaldo de Albuquerque Araújo
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
arnaldo@dcc.ufmg.br

***Resumo*—A presença de claquetes é frequente numa categoria de vídeos conhecida como *rushes*, que são vídeos gravados por produtoras de cinema e TV e ainda não editados. *Rushes* são potencialmente valiosos, mas largamente inexplorados. Uma tarefa que desempenha um importante papel para o gerenciamento eficiente desse tipo de vídeo é a segmentação automática. Espera-se que a capacidade de segmentar *rushes* automaticamente em cenas possa contribuir significativamente para o processo de edição das produções, uma vez que tal processo é realizado manualmente pela equipe de edição. Em *rushes*, um forte indicativo de inicio e fim de gravações de cenas é a presença de claquetes. Porém, o processo de reconhecimento de segmentos de vídeo contendo claquetes é complexo, devido à grande variação de cores e formatos que elas podem assumir. Para solucionar esse problema, o presente trabalho propõe um método para a detecção automática de claquetes baseado no uso de características espaciais e de cor (SIFT e HueSIFT), na representação dessas características por histogramas de palavras visuais (*Bags of Visual Features* - BoVF) e em aprendizado supervisionado (SVM). Os resultados dos experimentos realizados mostram que o método é competitivo quando comparado com os melhores resultados alcançados pelos trabalhos submetidos às tarefas de sumarização do TRECVID de 2007 e 2008.**

***Abstract*—The presence of clapboards is frequent in a category of videos known as rushes, which are unedited videos generated during video film recording. Rushes are potentially valuable, but largely unexplored. A task that plays an important role for the efficient management of this type of video is the automatic segmentation. It is expected that the ability to automatically segment rushes into scenes can significantly contribute to the editing process of this kind of material, since this process is performed manually by the editing team. In rushes, a good indicative of beginning and ending of the shooting of a scene is the presence of clapboards. However, the recognition of video segments containing clapboards is complex because of the wide range of colors and shapes that the clapboards can take. To solve this problem, this work proposes an automatic clapboard detection method based on spatial and color features (SIFT and HueSIFT) represented by a Bags of Visual Features - BoVF approach. This representation is robust to a series of transformations in the image and to occlusion, what makes it suitable for the detection of objets that can appear in different ways in the image, like clapboards. A supervised learning technique (SVM) is also employed. The results of the experiments shows that the method is competitive when compared to the best results obtained by the works submitted to the TRECVID summarization tasks of 2007 and 2008.**

***Palavras-chave*-detecção de claquetes; *rushes videos*; detecção de objetos**

I. Introdução

Com os recentes avanços em tecnologia, a produção, armazenamento e distribuição de conteúdo de vídeo nunca foi tão rápida e fácil. Não somente a indústria de cinema e TV tem produzido uma quantidade crescente de material videográfico, como também observamos um aumento significativo de produção desse tipo de material por parte de pessoas comuns.

Os vídeos produzidos durante uma gravação cinematográfica ou televisiva e ainda não editados são chamados de *rushes videos*. Durante a gravação de uma produção videográfica são registradas cenas de acordo com o roteiro. Normalmente, uma cena é gravada diversas vezes, com variações indicadas pelo diretor da produção. Além disso, os *rushes* possuem uma série de dados que são utilizados apenas para fins de marcação e separação entre gravações, por exemplo, segmentos de padrões de teste para calibração das cores da câmera, trechos com barras de cor e sequências que possuem claquetes, as quais identificam a produção e a cena gravada, entre outras coisas.

Rushes são potencialmente valiosos, mas largamente inexplorados, pois somente a equipe de produção original sabe exatamente o que o material contém. Além disso, os metadados são geralmente muito limitados, incluindo apenas poucas informações de indexação, tais como nome, equipe responsável, diretor e data.

Uma tarefa que desempenha um importante papel para o gerenciamento eficiente desse tipo de vídeo é a segmentação automática. Espera-se que a capacidade de segmentar *rushes* automaticamente em cenas possa contribuir significativamente para o processo de edição das produções, uma vez que tal processo é realizado manualmente pela equipe de edição.

Ao analisar a estrutura dos *rushes*, percebe-se que as claquetes funcionam como indicadores de início e fim de gravações de cenas e, assim, aparecem frequentemente nesse tipo de vídeo. Entretanto, conforme apontado por Liu *et al.* [1] e Pan *et al.* [2], o reconhecimento de segmentos de vídeo contendo claquetes é complexo, por causa da grande variação de cores e formatos que as claquetes podem assumir, o que as torna, muitas vezes, visualmente similares às cenas normais, como pode ser visto na Fig. 1.

Na literatura são encontradas diversas abordagens para a

Fig. 1. Quadros com diferentes padrões de claquetes.

detecção de claquetes, como o reconhecimento de padrões, o reconhecimento de caracteres, vetores de movimento, histogramas de cor e análise de áudio, entre outras. Nenhuma dessas abordagens, porém, mostra-se como solução definitiva para o problema.

Este trabalho concentra-se na tarefa de detecção de objetos em vídeos, mais especificamente na detecção de claquetes presentes em *rushes videos*, como os utilizados na TRECVID BBC *Rushes Summarization Task* [3], [4]. O método proposto é baseado no uso de características espaciais e de cor, representadas por histogramas de palavras visuais (*Bags of Visual Features* - BoVF) [5]. BoVF é um tipo de representação inspirada na técnica de *Bags of Words* (BoW) comumente utilizada em recuperação de informação para a representação de coleções de textos. Porém, ao invés de utilizar palavras, BoVF utiliza características visuais, em que cada característica visual representa uma região de interesse na imagem, gerada por descritores (no presente trabalho foram utilizados os descritores SIFT [6] e HueSIFT [7]). BoVF é uma representação de médio nível que tenta reduzir a diferença semântica entre as características de baixo nível e o conteúdo visual da imagem. Esse tipo de representação tem sido utilizado na literatura em vários cenários de detecção e classificação de padrões, alcançando bons resultados devido à sua robustez a oclusão e a uma série de transformações na imagem [8].

Além do uso de características de baixo nível e de representação de médio nível, o método proposto também emprega uma estratégia de aprendizado supervisionado. Uma base de dados de quadros de vídeo foi criada a partir dos vídeos da TRECVID BBC *Rushes Summarization Task*, separada em duas classes: "claquete" e "não claquete". Após a descrição dos quadros com SIFT e HueSIFT e sua representação por meio de BoVF, classificadores SVM [9] são treinados para classificar quadros como "claquete" ou "não claquete".

Os resultados dos experimentos realizados foram comparados com os trabalhos submetidos para a tarefa de sumarização do TRECVID de 2007 e 2008, mostrando que o método é competitivo em relação aos melhores resultados do TRECVID, considerando-se as métricas de precisão, revocação e acurácia.

O restante deste artigo é organizado da seguinte forma: a seção 2 descreve abordagens de detecção de claquetes em *rushes* encontradas em outros trabalhos, a seção 3 introduz o método proposto, a seção 4 indica a configuração dos experimentos realizados, a seção 5 apresenta os resultados dos experimentos e, finalmente, a seção 6 traz as conclusões do trabalho.

TABELA I
CATEGORIAS DE MÉTODOS DE DETECÇÃO AUTOMÁTICA DE CLAQUETES ENCONTRADOS NOS TRABALHOS DE SUMARIZAÇÃO DO TRECVID.

Método de Detecção	**Ocorrências em 2007**	**Ocorrências em 2008**	**Total**
Aprendizado supervisionado	0	3	3
Detecção não realizada	2	3	5
Análise de áudio	2	3	5
Comparação com uma base montada manualmente	1	5	6
Combinação de características	2	4	6
Não especificado	5	3	8
Outros (pouco explicados)	5	5	10
Total	17	26	43

II. REFERENCIAL TEÓRICO

Embora possa ocorrer em qualquer vídeo, a presença de claquetes é muito mais frequente em *rushes videos* – vídeos gravados por produtoras de cinema e TV e ainda não editados. A TRECVID BBC *Rushes Summarization Task* (2007 e 2008) disponibilizou para seus participantes uma base de dados contendo vídeos desse tipo. Devido à estrutura específica dos *rushes*, as abordagens de sumarização tiveram que realizar um pré-processamento nos vídeos, o que incluiu a remoção de segmentos de vídeo que não adicionavam informação útil ao sumário, entre os quais estavam os segmentos contendo claquetes. Desse modo, é possível encontrar nos trabalhos do TRECVID uma variedade de métodos e estratégias para a detecção de claquetes.

Em 2007, 22 grupos participaram da tarefa de sumarização do TRECVID, sendo que 17 destes publicaram artigos sobre seus trabalhos [4]. Em 2008 foram 31 participantes da tarefa de sumarização, com 26 trabalhos publicados [3]. Analisando-se cada um desses trabalhos publicados, é possível dividir as abordagens para detecção de claquetes em algumas categorias, de acordo com o ponto forte de cada abordagem. A Tabela I apresenta essa divisão.

Pelos dados da Tabela I, observa-se que os métodos de detecção mais usados foram os de "combinação de características" e de "comparação com uma base montada manualmente".

"Combinação de características" significa o uso de mais de uma técnica em conjunto, explorando propriedades identificadas nos quadros que contêm claquetes. Por exemplo, a partir da constatação de que muitas claquetes são constituídas por alguns caracteres ou números sobre um fundo preto/branco/cinza, Wang *et al.* [10] usam reconhecimento de caracteres e análise de cor da região em torno desses caracteres para encontrar possíveis quadros contendo claquetes. Já Laganière *et al.* [11] primeiro encontram picos no gráfico de densidade de características espaço-temporais, os quais podem representar os momentos em que a claquete é batida. Em seguida, analisam o histograma de cor da região central da imagem para verificar a saturação, pois as claquetes tendem a produzir histogramas com uma maioria de *pixels* pretos e brancos.

A "comparação com uma base montada manualmente" trata

de medir a distância entre um quadro desconhecido e outros quadros previamente escolhidos como exemplos de claquetes. Seguindo essa linha, Beran *et al.* [12] fazem uma anotação manual dos trechos de vídeo contendo lixo e aplicam sobre eles o algoritmo *K-means*, para agrupamento dos diversos tipos de lixo a serem removidos, incluindo claquetes. Um quadro desconhecido é analisado a partir da taxa de erro obtida na comparação deste com os *clusters*, via distância Euclidiana. Chasanis, Likas e Galatsanos [13] usam uma abordagem ainda mais simples, baseada em SIFT. Descritores SIFT são calculados para uma base de 150 quadros contendo claquetes. Posteriormente, os descritores SIFT também são calculados para quadros desconhecidos e comparados com aqueles da base. Se o número de resultados positivos das comparações for superior a um certo limiar pré-estabelecido, considera-se que o quadro contém claquete.

Métodos de detecção de claquetes por "análise de áudio" são também muito usados. O trabalho de Liu, Liu e Zhang [1] dá grande enfoque à detecção de claquetes, propondo um novo método de detecção a partir do cálculo de energia acústica. Os autores observaram que mais de 95% das sequências de vídeo do TRECVID de 2007 contêm apenas sons estacionários. Os 5% restantes podem incluir sons similares ao de uma claquete batendo, como sons de colisões e batidas em móveis, entre outros. A similaridade reside em que esses tipos de sons aparecem em ambientes calmos. Felizmente, a energia de alguns sons não aumenta tão rápido quanto o som de uma claquete, o qual pode ser, assim, discriminado.

Em 2008 percebe-se também o emprego de métodos de detecção baseados em "aprendizado supervisionado". Ren, Punitha e Jose [14] e Dumont e Mérialdo [15] trabalham com dois conjuntos de quadros de treinamento manualmente construídos ("claquete" e "não claquete") e classificadores SVM. Depois de treinados, os classificadores podem determinar se um certo quadro desconhecido contém ou não uma claquete.

Finalmente, nos dois anos da tarefa de sumarização do TRECVID observa-se o uso de métodos específicos, os quais foram agrupados na Tabela I na categoria "outros". Esses métodos, em sua maior parte, não foram bem detalhados nos trabalhos publicados, o que, infelizmente, dificulta sua compreensão e análise.

III. Um Método para Detecção de Claquetes em *Rushes Videos*

Neste trabalho, optou-se pelo foco em características visuais e, por isso, a análise dos quarenta e três trabalhos publicados para a tarefa de sumarização do TRECVID de 2007 e 2008 concentrou-se naqueles que utilizaram esse tipo de característica.

A análise desses trabalhos permite perceber que os métodos mais simples, em geral, são os que obtêm melhores resultados. Em resumo, as seguintes etapas podem ser delineadas para o processo de detecção de claquetes, tendo por base os pontos comuns dos trabalhos do TRECVID: *i)* montagem de uma base de dados de quadros contendo claquetes (e também não contendo, dependendo do caso); *ii)* aplicação de um ou mais descritores a cada quadro dessa base; *iii)* extração de quadros do vídeo desconhecido que se quer analisar; *iv)* aplicação dos mesmos descritores da etapa *ii* em cada quadro extraído do vídeo desconhecido; *v)* comparação dos quadros desconhecidos com os da base, usando as informações dos descritores; *vi)* avaliação do resultado da comparação, determinando se os quadros desconhecidos contêm claquete ou não.

A escolha dos descritores a serem utilizados e do método de comparação entre quadros varia nos trabalhos do TRECVID. Com relação aos descritores, chama a atenção o uso do descritor SIFT, que aparece em seis trabalhos. SIFT é um descritor que calcula pontos de interesse na imagem. Um ponto de interesse é um ponto (ou região) específico da imagem que apresenta variação significativa de intensidade em mais de uma direção. O descritor SIFT possui a propriedade de ser robusto com relação a certas transformações dos objetos na imagem, como rotação, oclusão e mudança de escala, o que é muito pertinente no caso de claquetes, tendo em vista as diferentes formas em que elas podem aparecer nos vídeos.

Para cada imagem, o descritor SIFT pode extrair uma quantidade diferente de pontos de interesse, sendo cada um deles representado por um vetor de tamanho fixo. Uma imagem descrita com SIFT é, portanto, um conjunto de vetores, conjunto este que pode apresentar qualquer tamanho, tornando, assim, impraticável a comparação direta entre imagens descritas com SIFT. A fim de resolver esse problema, foi adotada neste trabalho a estratégia de *Bag of Visual Features* - BoVF para quantificar a distribuição de pontos de cada quadro analisado, de maneira similar ao realizado por Christel *et al.* [16] – que denominam a estratégia de *Bag-of-Word* – e por Noguchi e Yanai [17] – que denominam a estratégia de *Bag-of-Keypoints* – em seus trabalhos para o TRECVID. A ideia básica por trás da estratégia de BoVF é representar cada quadro por um único vetor. Seu funcionamento se dá da seguinte maneira: primeiramente, uma determinada quantidade de pontos de interesse é escolhida aleatoriamente entre os quadros da base de dados montada, formando um vocabulário de BoVF; depois, para cada ponto de interesse de cada quadro é feito o cálculo da distância Euclidiana entre esse ponto e todos os pontos do vocabulário, atribuindo-se uma ocorrência àquele ponto do vocabulário que apresentar a menor distância; ao final, cada quadro pode ser então caracterizado por um histograma de características visuais (o BoVF) com tamanho igual ao do número de pontos do vocabulário. Dessa forma, a comparação entre quadros torna-se possível, pois trata-se apenas de comparar os histogramas que representam cada quadro.

A comparação entre os quadros e a análise do resultado dessa comparação podem ser feitas separadamente, como mostrado nas etapas descritas anteriormente, ou de maneira conjunta. No primeiro caso, a distância entre os quadros é calculada (etapa *v*) e um limiar é usado para determinar se o quadro comparado é claquete ou não (etapa *vi*). No segundo caso, é possível usar um classificador, que retorna o resultado da comparação em forma de classe, indicando se o quadro comparado contém claquete ou não. Por condensar duas etapas

em uma só, o uso de um classificador tende a facilitar o processo de detecção. Entre os classificadores empregados nos trabalhos do TRECVID, destaca-se o SVM, utilizado em cinco trabalhos.

Essas observações levaram à construção do método de detecção de claquetes proposto neste trabalho, que pode ser resumido nas seguintes etapas: *i)* montagem de uma base de dados de quadros com duas classes: "claquete" e "não claquete"; *ii)* aplicação do descritor SIFT a cada quadro da base; *iii)* criação de BoVF a partir dos vetores SIFT extraídos; *iv)* treinamento de um classificador SVM com os BoVF; *v)* aplicação do classificador a quadros desconhecidos para determinar a presença de claquetes nesses quadros.

Além do uso do descritor SIFT, também chama a atenção, nos trabalhos do TRECVID, a exploração de características de cor para a detecção de claquetes. Observa-se que treze trabalhos, entre os quarenta e três, empregam essa característisca em sua etapa de detecção, com certo destaque para o uso do espaço de cor HSV (*Hue, Saturation, Value*), presente em cinco trabalhos. Assim, considerou-se também a realização de experimentos com um descritor relacionado a cores.

As características de cor proporcionam uma representação global da imagem. Entretanto, conforme apontado por Noguchi e Yanai [17], é difícil detectar claquetes apenas com um método baseado em histogramas de cor, devido à grande variação de aparência das claquetes. Dessa forma, neste trabalho optou-se pelo uso de características de cor agregadas a pontos de interesse (que proporcionam uma representação local da imagem). O descritor escolhido para isso foi o HueSIFT, uma variante do descritor SIFT que leva em consideração a informação de cor do componente H do espaço HSV. A criação de BoVF também foi realizada para o HueSIFT.

IV. Configuração dos Experimentos

A base de dados para os experimentos foi montada a partir de 44 vídeos usados na tarefa de sumarização do TRECVID de 2007. 1054 quadros foram manualmente extraídos dos vídeos e divididos igualmente em duas classes: "claquete" e "não claquete". A escolha dos quadros para a composição da classe "claquete" foi realizada de modo a contemplar a diversidade de formatos, cores e posições das claquetes encontradas nos vídeos do TRECVID. Para isso, procurou-se selecionar quadros a partir de todos os 44 vídeos, de modo a coletar padrões de claquetes diferentes e situações de ocorrência diferentes (luminosidade maior/menor; claquete aberta/fechada; claquete próxima/distante da câmera; etc.). A classe "não claquete" foi construída de maneira a evitar a presença de qualquer quadro de claquete ou de lixo e também de forma a garantir a diversidade entre as imagens, assegurando que quadros de todos os 44 vídeos estivessem presentes.

Vetores de pontos de interesse foram calculados para cada quadro nas duas classes usando as implementações SIFT e HueSIFT fornecidas no *software* "ColorDescriptor" [7]. Em seguida, foram calculados os BoVF de cada quadro, tanto para os vetores extraídos com SIFT quanto para os vetores extraídos com HueSIFT. Tamanhos diferentes de vocabulário para os BoVF foram testados, usando 100, 500, 1.000, 2.000 e 4.000 pontos de interesse.

TABELA II
Resultados obtidos para os testes com o SIFT.

	Tamanho do vocabulário				
	100	500	1.000	2.000	4.000
Acurácia	83,93%	72,59%	76,69%	83,64%	83,26%
Precisão	81,97%	65,46%	66,97%	82,64%	84,43%
Revocação	88,76%	60,95%	68,95%	88,38%	84,19%

TABELA III
Resultados obtidos para os testes com o HueSIFT.

	Tamanho do vocabulário				
	100	500	1.000	2.000	4.000
Acurácia	78,88%	80,31%	79,17%	74,98%	81,93%
Precisão	79,69%	80,70%	82,97%	70,07%	86,59%
Revocação	79,62%	82,29%	76,00%	58,67%	75,62%

Para a etapa de classificação, realizada separadamente para o SIFT e para o HueSIFT, foi utilizada validação cruzada dos dados. Os BoVF da base de dados foram divididos em cinco *folds*. Quatro *folds* foram usados para treinamento de um classificador SVM e um *fold* foi usado para teste posterior com o modelo gerado pelo classificador. Neste trabalho optou-se pelo uso da biblioteca LIBSVM [18]. Os dados de treinamento e teste foram primeiramente colocados na mesma escala. Um subconjunto dos dados de treinamento foi então gerado e passado como parâmetro ao *script* grid.py, que acompanha o LIBSVM. Esse *script* determina os melhores parâmetros, "c" e "g", para o kernel RBF (*Radial Basis Function*) do classificador. Em seguida, os parâmetros encontrados pelo grid.py foram usados para a criação do modelo SVM a partir dos dados de treinamento. Finalmente, o modelo foi aplicado na predição das classes dos dados de teste.

V. Análise de Resultados dos Experimentos

Para a análise de resultados dos experimentos realizados com o método proposto, foram calculadas as médias da acurácia (*accuracy*), da precisão (*precision*) e da revocação (*recall*) [19] obtidas com a validação cruzada dos dados. A Tabela II resume os resultados conseguidos com o SIFT, enquanto a Tabela III resume os resultados conseguidos com o HueSIFT.

Observando os dados das tabelas, é possível perceber que o aumento do tamanho do vocabulário não necessariamente implica num aumento das taxas de acurácia, precisão e revocação. É relevante notar que, mesmo com um número pequeno de pontos, como 100 ou 500, é possível obter taxas em torno de 80%. A vantagem em usar um número menor de pontos é que o tempo necessário para a criação dos BoVF e o tempo necessário para a etapa de classificação são menores.

Outra informação importante a ser inferida a partir dos resultados apresentados é a de que não houve ganho considerável no uso do HueSIFT em relação ao SIFT. Para 100 pontos e 2.000 pontos, inclusive, o uso do SIFT mostrou-se melhor do que o uso do HueSIFT em todas as três métricas. Especula-se

que isso tenha acontecido devido à grande variação de cores entre uma claquete e outra. Com essa alta variação, mesmo uma avaliação local do componente de cor, como feito pelo HueSIFT, pode não ajudar a distinguir um ponto ou região da imagem pertencente a uma claquete de um outro ponto ou região que não pertença.

Algumas colunas das duas tabelas apresentam números bem próximos para as três métricas adotadas, como o SIFT com vocabulário de tamanho 4.000 (todas as métricas próximas a 84%) e o HueSIFT com vocabulário de tamanho 100 (todas as métricas próximas a 79%). Por outro lado, algumas colunas exibem taxas razoavelmente discrepantes entre as métricas, como o SIFT com vocabulário de tamanho 500 (revocação de 60,95% e acurácia de 72,59%) e o HueSIFT com vocabulário de tamanho 2.000 (revocação de 58,67% e acurácia de 74,98%). Considerando-se que a base de dados utilizada é balanceada, especula-se que esse fenômeno esteja ligado a duas das etapas do método: a criação dos BoVF e a criação do subconjunto de dados para a estimativa dos parâmetros do classificador.

Na etapa de criação dos BoVF, os pontos de interesse do vocabulário são escolhidos aleatoriamente entre todos os pontos dos quadros que compõem a base de dados (tanto da classe "claquete" quanto da classe "não claquete"). Dessa forma, é coerente pensar que essa escolha aleatória tenha levado à criação de vocabulários mais significativos para certos testes do que para outros, isto é, em alguns casos os pontos de interesse escolhidos para a formação do vocabulário foram mais representativos dos dados da base do que em outros casos.

A criação do subconjunto de dados para a estimativa dos parâmetros do classificador também pode ter influenciado na discrepância entre as taxas. Esse subconjunto é criado com o auxílio do *script* subset.py, que acompanha o LIBSVM. Apesar do subconjunto ser construído de maneira estratificada, os pontos de interesse que formam o subconjunto são escolhidos aleatoriamente. Assim, da mesma maneira que no caso da criação dos vocabulários dos BoVF, essa escolha aleatória pode levar à construção de subconjuntos mais significativos em certos momentos do que em outros, afetando a estimativa dos parâmetros do classificador e, consequentemente, o resultado final da classificação.

A execução de mais experimentos pode testar essas hipóteses, especialmente a execução de experimentos exaustivos com o mesmo tamanho de vocabulário.

Avançando para além da análise particular das duas etapas tratadas nos parágrafos anteriores, cabe destacar que todas as etapas do método proposto possuem parâmetros que podem ser variados, tais como: tamanho da base de dados construída, tamanho do subconjunto para estimativa dos parâmetros do classificador, número de *folds* usados na validação cruzada dos dados, entre outros. A análise do impacto da variação de cada um desses parâmetros no resultado final do método requer novos experimentos, com foco específico.

A. Comparação com os trabalhos do TRECVID

Para dar uma visão mais abrangente da relevância dos resultados obtidos com o método proposto, é importante compará-los com os resultados obtidos pelos participantes do TRECVID.

Dos quarenta e três trabalhos publicados para a tarefa de sumarização do TRECVID de 2007 e 2008, somente sete apresentaram medidas específicas para seus métodos de detecção de claquetes.

Liu, Liu e Zhang [1] publicaram o trabalho que traz o maior foco no problema de detecção, entre todos os quarenta e três. Os resultados reportados demonstram que o método proposto pelos autores atingiu um ótimo desempenho: precisão de 93,97% e revocação de 87,64%.

Christel *et al.* [16] fizeram uso de três características distintas: SIFT, histograma de cor no espaço HSV e reconhecimento de voz. Para o reconhecimento de voz, os autores reportaram precisão de apenas 17%, com revocação de 20%. Para o histograma de cor, a precisão encontrada foi de 65% e a revocação de 22%. Os melhores resultados foram obtidos com o uso de SIFT: 70% de precisão e 37% de revocação. Os autores testaram, ainda, a combinação das três características, com resultados de 31% de precisão e 58% de revocação.

Pan, Chuang e Hsu [2] também realizaram a detecção de claquetes através da combinação de características, obtendo resultados superiores: 68% de precisão e 61% de revocação. Contudo, o trabalho que se saiu melhor, na linha de combinação de características, foi o de Wang *et al.*: 70% de precisão e 71,8% de revocação.

No trabalho de Ren, Punitha e Jose [14], as claquetes foram divididas em dois grupos: claquetes grandes e claquetes pequenas. A precisão média obtida foi de 77,2% para a detecção de claquetes pequenas e 69,5% para claquetes grandes. Os autores comentaram ainda que a adição de uma etapa de pós-processamento, aplicando programação dinâmica, chegou a aumentar a precisão em 7% para claquetes pequenas e em 12% para claquetes grandes.

Valdés e Martínez [20] propuseram o treinamento de um detector de objetos da biblioteca OpenCV[1], o qual conseguiu taxas de acerto acima de 95% para a detecção de claquetes do conjunto de treinamento. Os autores reportaram, no entanto, que a alta variabilidade na posição, tamanho, rotação e iluminação das claquetes do conjunto de teste produziram uma redução não quantificada da precisão do detector.

Seguindo outro caminho, Kleban *et al.* [21] usaram uma árvore de vocabulário SIFT para a detecção de claquetes, obtendo uma taxa de verdadeiro positivo em torno de 90%, com uma taxa de falso positivo de menos de 2%. Além disso, comentaram também que o uso de uma técnica de "*smoothing*", para lidar com quadros borrados (*blurred*), chegou a aumentar a taxa de verdadeiro positivo, em certos casos, em até 10%.

A Fig. 2 apresenta um gráfico comparativo entre dois dos melhores resultados dos experimentos feitos com o método

[1] http://opencv.willowgarage.com/wiki/

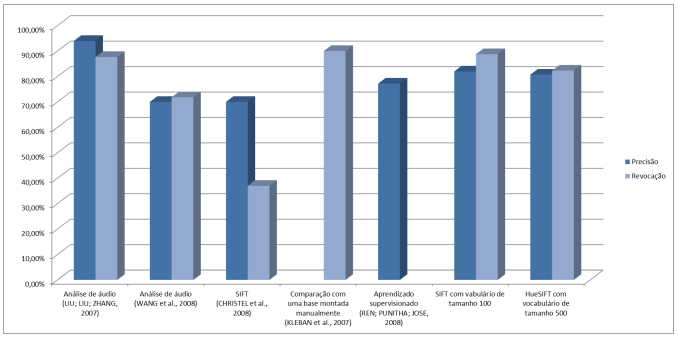

Fig. 2. Comparação entre dois dos melhores resultados obtidos nos experimentos e alguns dos melhores resultados encontrados nos trabalhos do TRECVID.

proposto neste trabalho e alguns dos melhores resultados encontrados nos trabalhos do TRECVID, comentados anteriormente. [2]

Observando o gráfico, nota-se que o método proposto é capaz de atingir taxas de precisão e revocação próximas àquelas apresentadas pelo trabalho do TRECVID que possui os melhores resultados para o processo de detecção de claquetes.

VI. Conclusões

Este trabalho apresentou um método para detecção de claquetes baseado em características espaciais e de cor (extraídas com os descritores SIFT e HueSIFT), representadas por uma estratégia de *Bags of Visual Features* - BoVF, a qual tenta reduzir a diferença semântica entre as características de baixo nível e o conteúdo visual das imagens.

O método empregou também uma estratégia de aprendizado supervisionado. Uma base de dados foi criada contendo uma classe positiva ("claquete") e outra negativa ("não claquete"). Após a criação dos BoVFs, classificadores SVM foram treinados para distinguir quadros que contêm claquetes de quadros que não contêm.

Os resultados experimentais mostraram que o método é competitivo quando comparado com os melhores resultados alcançados pelos trabalhos submetidos ao TRECVID.

De maneira geral, o descritor SIFT foi o que produziu os melhores resultados nos experimentos. Observou-se, com isso, que as características de cor combinadas com o descritor local (HueSIFT) não agregaram informação relevante para a detecção de claquetes. Trabalhos da literatura mostram, contudo, que o uso das características de cor separadamente pode ser promissor.

É válido ressaltar que o método proposto pode ser facilmente adaptado para lidar com a detecção de outros objetos além de claquetes, ampliando seu alcance para a solução do problema mais abrangente da detecção de objetos em vídeos. Essa constatação representa um forte incentivo para a realização de novas investigações e experimentos relacionados ao método apresentado, dando continuidade ao que foi feito neste trabalho.

[2]Infelizmente, nem todos os trabalhos do TRECVID reportaram seus resultados para a detecção de claquetes em termos da precisão e da revocação.

Referências

[1] Y. Liu, Y. Liu, and Y. Zhang, "The hong kong polytechnic university at trecvid 2007 bbc rushes summarization," in *Proceedings of the international workshop on TRECVID video summarization*. Germany: ACM, 2007.

[2] C.-M. Pan, Y.-Y. Chuang, and W. H. Hsu, "NTU TRECVID-2007 fast rushes summarization system," in *Proceedings of the international workshop on TRECVID video summarization*. Germany: ACM, 2007.

[3] P. Over, A. F. Smeaton, and G. Awad, "The trecvid 2008 BBC rushes summarization evaluation," in *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*. USA: ACM, 2008.

[4] P. Over, A. F. Smeaton, and P. Kelly, "The trecvid 2007 bbc rushes summarization evaluation pilot," in *Proceedings of the international workshop on TRECVID video summarization*. Germany: ACM, 2007.

[5] T. de Campos, M. Barnard, K. Mikolajczyk, J. Kittler, F. Yan, W. Christmas, and D. Windridge, "An evaluation of bags-of-words and spatio-temporal shapes for action recognition," in *IEEE WACV*, 2011.

[6] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE ICCV*. IEEE, 1999.

[7] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9.

[8] T. O. Cunha, F. G. H. Souza, A. A. Araujo, and G. L. Pappa, "Rushes Video Summarization Based on Spatio-temporal Features," in *Proceedings of the ACM SAC 27th Symposium on Applied Computing*. Italy: ACM, 2012.

[9] T. Joachims, "Aktuelles schlagwort: Support vector machines," *Künstliche Intelligenz*, vol. 4, 1999.

[10] T. Wang, Y. Gao, J. Li, P. P. Wang, X. Tong, W. Hu, Y. Zhang, and J. Li, "THU-ICRC at rush summarization of TRECVID 2007," in *Proceedings of the international workshop on TRECVID video summarization*. Germany: ACM, 2007.

[11] R. Laganière, R. Bacco, A. Hocevar, P. Lambert, G. Païs, and B. E. Ionescu, "Video summarization from spatio-temporal features," in *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*. USA: ACM, 2008.

[12] V. Beran, M. Hradiš, P. Zemčík, A. Herout, and I. Řezníček, "Video summarization at Brno university of technology," in *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*. USA: ACM, 2008.

[13] V. Chasanis, A. Likas, and N. Galatsanos, "Video rushes summarization using spectral clustering and sequence alignment," in *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*. USA: ACM, 2008.

[14] R. Ren, P. Punitha, and J. Jose, "Video redundancy detection in rushes collection," in *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*. USA: ACM, 2008.

[15] E. Dumont and B. Mérialdo, "Sequence alignment for redundancy removal in video rushes summarization," in *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*. USA: ACM, 2008.

[16] M. G. Christel, A. G. Hauptmann, L. Wei-Hao, M.-Y. Chen, J. Yang, B. Maher, and R. V. Baron, "Exploring the utility of fast-forward surrogates for bbc rushes," in *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*. USA: ACM, 2008.

[17] A. Noguchi and K. Yanai, "Rushes summarization based on color, motion and face," in *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*. USA: ACM, 2008.

[18] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[19] R. Kohavi and F. Provost, "Glossary of terms," *Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*, vol. 30, no. 2–3, 1998.

[20] V. Valdés and J. M. Martínez, "Binary tree based on-line video summarization," in *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*. USA: ACM, 2008.

[21] J. Kleban, A. Sarkar, E. Moxley, S. Mangiat, S. Joshi, T. Kuo, and B. Manjunath, "Feature fusion and redundancy pruning for rush video summarization," in *Proceedings of the international workshop on TRECVID video summarization*. Germany: ACM, 2007.