Anotação automática de vídeo baseada na detecção de gestos em um ambiente de videoconferência

Tiago S. Nazaré, Rodrigo V. C. Beber, Moacir P. Ponti Jr.
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
13560-970 São Carlos, SP, Brasil
Project web page: http://www.icmc.usp.br/~moacir/project/VideoProcessing

***Resumo*—As filmagens de videoconferências e de salas de aula em ambiente de educação a distância muitas vezes precisam ser anotadas após serem gravadas. Uma anotação importante refere-se ao momento em que algum participante fez um gesto de levantar a mão. Nesse trabalho propomos um sistema para detectar esse gesto e anotar o vídeo. Ele usa como base um algoritmo conhecido de detecção de objetos, analisa cada frame e gera uma lista de classificação que permite identificar a existência de intervalos nos quais houve gestos de levantar a mão. Esse sistema foi aplicado em vídeos com situações diversas. Os resultados mostraram que o método proposto é eficaz em trabalhar sob variações de luminosidade e de plano de fundo, múltiplos gestos, câmeras em movimento e com a existência de gestos falsos.**

***Keywords*-Processamento de vídeo, detecção de gestos, anotação de vídeo.**

***Abstract*—Videoconference and distance learning videos are often annotated after being recorded. One of the important occurrences are the moment that someone makes a hand rising gesture. In this paper we propose a system to detect the hand rinsing gesture and annotate the video. A known object detection method is used, in which each frame is analysed and a classification list is generated in order to identify the existance of frames where a hand rising was detected. This system was tested in videos with several issues. The results show that the proposed method is capable of deal with illumination and background variations, detection of multiple true gestures, moving cameras and fake gestures.**

***Keywords*-Video processing, gesture detection, video annotation**

I. Introdução

O avanço da indústria de hardware possibilitou redução nos preços e melhorias nos sistemas de captura, transmissão e armazenamento de imagens e vídeos, como máquinas fotográficas, câmeras de vídeo, fibras ópticas e discos de alta capacidade. A facilidade de aquisição desses itens fez com que os mesmos passassem a ser cada vez mais utilizados para uso doméstico, em sistemas de vigilância, videoconferência etc.

A viabilidade de se trabalhar com dados visuais em várias áreas levou à busca por métodos para detecção e classificação automáticas de cenas e objetos, já que realizar essa tarefa de forma manual pode ser difícil e ineficiente.

Uma das situações na qual nos deparamos com problemas como os citados acima é a análise de filmagens de videoconferência e salas de aula em ambiente de educação a distância. Em vídeos desse tipo, muitas vezes é necessário detectar quando alguém pede a palavra para fazer um comentário ou pergunta, o que geralmente é feito através de um gesto de levantar a mão. Nesse caso, um sistema de anotação automática de vídeo que consiga detectar, com boa precisão, os intervalos em que tais gestos ocorrem, pode auxiliar nessa difícil tarefa de anotação.

Com objetivo de facilitar a anotação de vídeos com relação ao gesto de levantar a mão, propomos um sistema no qual detectamos faces e mãos abertas e verificamos se a posição relativa dos mesmos é compatível com o gesto de levantar a mão, ou seja, a mão se encontra em uma posição próxima ao rosto e acima da linha dos ombros. Para isso utilizamos uma combinação de métodos já existentes na literatura, em especial o método Viola-Jones melhorado [7].

Kölsch e Turk [5], [6] demonstraram que o detector visual de objetos proposto por Viola e Jones [11], usado inicialmente para detecção de faces, também podia ser utilizado de maneira muito eficiente na detecção de várias posturas de mão. Lienhart e Maydt [7] desenvolveram uma modificação mais eficiente que o Viola-Jones original e, embora Kölsch e Turk [5], [6] não tivessem abordado a modificação, utilizamos o algoritmo de Lienhart e Maydt na construção do nosso método.

Utilizando esse detector como base, construímos um sistema no qual cada frame é analisado individualmente. Por isso, o nosso sistema deve ser capaz de lidar bem com problemas referentes ao movimento de câmera, alterações na luminosidade e do plano de fundo, o que pode permitir, entre outras coisas, fazer a alternância entre câmeras (corte) e filmar com a câmera em movimento, variações essas que são comuns em uma situação real. Entretanto, a plateia deve ser enquadrada de frente, já que os detectores de objetos utilizados não possuem um grande suporte à variação de ângulo.

Entre as principais contribuições desse trabalho estão: o **método de detecção de gestos de mão aberta** e o **estudo das limitações do algoritmo Viola-Jones para essa aplicação**, além de uma **nova base de imagens de exemplos e contra-exemplos para detecção de mão aberta**, e um conjunto de vídeos que busca testar a robustez do método proposto às diversas situações citadas.

O texto está organizado da seguinte maneira: discutimos alguns trabalhos relacionados a técnicas de detecção de gestos de levantar a mão na seção II. Uma revisão do método de

detecção de objetos Viola-Jones e da modificação desenvolvida por Lienhart e Maydt são feitas na seção III. Demonstramos o funcionamento e a construção de nosso método na seção IV. Apresentamos a metodologia de testes na seção V. Os resultados obtidos e uma discussão sobre os mesmos, na seção VI. Na seção VII colocamos nossas conclusões.

II. Trabalhos relacionados

O interesse por sistemas que permitem detectar seres humanos e seus gestos ganhou, nos últimos anos, a atenção do público e dos pesquisadores ligados à área. Um dos segmentos que tem recebido bastante atenção é o referente a técnicas que fazem detecção do gesto de levantar a mão, bastante úteis em ambientes de classe de aula e videoconferência.

Entre os trabalhos anteriores que tentam resolver esse problema, Jie Yao e Jeremy Cooperstock [12] propõem um método no qual supõe-se que as cabeças das pessoas da plateia são enquadradas pela câmera em uma mesma linha horizontal, que a câmera está estática e a cor do fundo não varia. Então, procura-se por regiões, acima da linha das cabeças, nas quais há movimento e pele humana. Se uma região com essas características é encontrada uma reta é aproximada usando-se os pontos médios das bordas da região e se a inclinação da reta estiver entre 45 e 135 graus a região é considerada como uma mão levantada.

Xiaodong Duan e Hong Liu [2] propõem um método de detecção do gesto de levantar a mão baseado na análise da silhueta humana, em ambiente fechado. Esse método consegue trabalhar com pessoas se movendo em um grupo. Entretanto, a câmera não pode se mover, pois isso tornaria, segundo os próprios autores, difícil separar as pessoas do cenário. A arquitetura dessa abordagem consiste em: *foreground detection*: para detectar as partes da cena onde há movimento; *blob segmentation*: tenta isolar silhuetas humanas presentes nas partes da imagem onde há movimento; busca de regiões candidatas (*CR: candidate regions*): busca por componentes conexos situados no quarto superior da silhueta; extração de características: é realizada através do algoritmo *R-transform*; e classificação dos gestos de levantar a mão: verifica se o contorno das *CR* são de um braço ou uma mão levantada.

Nosso trabalho buscou melhorias para problemas dos métodos acima, especialmente a dependência de uma única câmera, o posicionamento estático da mesma, o enquadramento específico da plateia, a manutenção de um mesmo plano de fundo e de um mesmo padrão de luminosidade.

III. O detector de Viola-Jones e a modificação feita por Rainer Lienhart e Jochen Maydt

Nesse trabalho utilizamos a modificação feita por Rainer Lienhart e Jochen Maydt [7] do detector de objetos proposto por Paul Viola e Michael Jones [11] para detectar faces e mãos nos frames, para que então possamos verificar se a posição relativa das mesmas é compatível com um gesto de levantar a mão, como será descrito em detalhes na próxima seção.

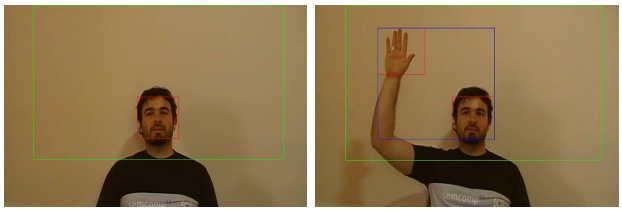
Viola e Jones [11] descreveram uma abordagem de aprendizado de máquina para detecção visual de objetos capaz de processar imagens com extrema rapidez e altos índices de detecção. Segundo os autores, seu trabalho apresenta três contribuições-chave: "*Integral Image*", uma forma de representação que permite um processamento rápido das características (*features*); algoritmo de aprendizagem baseado em AdaBoost [3] que, dentro de um conjunto extenso de elementos, seleciona um pequeno número de descritores visuais críticos; e a combinação de classificadores com complexidade crescente em cascata, permitindo descartar rapidamente regiões de fundo da imagem e gastar a capacidade de processamento com regiões que se assemelham a um objeto.

Lienhart e Maydt [7] introduziram duas modificações no algoritmo Viola-Jones. A primeira consiste em acrescentar um novo conjunto de características rotativas (*rotated haar-like features*) ao conjunto de características simples semelhantes à haar de Papageorgiou et al. [8], [9] (*set of simple haar-like features*) usadas por Paul Viola e Michael Jones. O segundo melhoramento refere-se ao aprimoramento do uso dos classificadores em cascata. Segundo os autores, o uso conjunto dessas duas inovações conseguiu apresentar resultados que indicam um aumento na performance geral de 23,8%, sendo 10% atribuído ao conjunto de características rotativas (*rotated features*) e 12,5% ao estágio de otimização dos classificadores (*stage post-optimization scheme*).

IV. O sistema de anotação automática

A. Método

O nosso sistema baseia-se em analisar separadamente cada frame do vídeo. Em cada um deles procuramos por faces e mãos através do método de detecção de objetos desenvolvido por Lienhart e Maydt. Após fazer a detecção das faces e mãos verificamos se a posição relativa entre elas é compatível com a de uma mão levantada. Consideramos que **em um gesto de mão levantada há uma mão próxima a uma face e acima da linha dos ombros**. Portanto, tomamos cada par possível, formado por uma mão e uma face, e verificamos se a mão se encontra em uma região próxima à face. Essa região foi considerada, experimentalmente, como o espaço ao redor da cabeça que corresponde a uma medida que é duas vezes e meia aquela da região da face horizontalmente, tanto para a esquerda como para a direita, cinco vezes verticalmente acima e metade verticalmente abaixo, como exemplifica a Figura 1.



(a) (b)

Figura 1. Imagem que exemplifica a região de busca. Em (a) temos a face destacada por um retângulo vermelho e a região de busca destacada por um retângulo verde. Em (b) além do destaque da face e da região de busca temos a mão, também destacada por um retângulo vermelho, e um retângulo azul que mostra a relação entre mão e face.

Os frames, nos quais são encontrados pelo menos um par face-mão que satisfaz as condições citadas acima, são considerados positivos.

Depois que todos os frames foram analisados temos uma lista com a classificação dos frames - positivos e negativos. Com base nessa lista tentamos encontrar intervalos nos quais há uma mão levantada. Os intervalos, nos quais há uma mão levantada, têm as seguintes características:

- O intervalo é maior ou igual a 1 segundo de vídeo;
- O primeiro e o último frames do intervalo são considerados positivos;
- Entre o primeiro e o último frame há no mínimo 80% de frames positivos;
- Não há no intervalo nenhuma sequência de frames negativos maior que 1 segundo de vídeo.

O Algoritmo 1 contém os passos para a detecção do gesto de levantar a mão.

Algorithm 1 Detecção de mão levantada

```
for cada frame do vídeo do
    encontre as faces na imagem.
    encontre as mãos esquerdas na imagem.
    espelhe verticalmente a imagem.
    encontre as mãos esquerdas na imagem.
    for cada face encontrada do
        for para cada mão encontrada do
            if a mão está próxima da face then
                marque o frame como positivo.
            end if
        end for
    end for
end for
```

B. Implementação e base de treinamento

A implementação desse sistema foi feita em linguagem C e fez uso da biblioteca de visão computacional OpenCV [1] versão 2.1. Essa biblioteca conta com cascatas já treinadas e armazenadas em arquivos XML, algumas delas para detecção de faces. Usamos uma delas (`haarcascade_frontalface_default.xml`) para fazer a detecção de faces. Além disso, também é possível criar novas cascatas fornecendo exemplos e contraexemplos, as quais também serão armazenadas em um arquivo XML.

Para fazer a detecção das mãos, como o OpenCV não conta com uma cascata já treinada, **coletamos 905 imagens de mãos esquerdas abertas** — com a ajuda de diversos voluntários fotografados, usando diversos planos de fundo, variações de iluminação e mudança na abertura dos dedos — que tiveram a região da mão marcada manualmente e utilizamos também **1000 imagens de contraexemplos**, das quais a maioria foi extraída de [10]. Na Figura 2 são mostradas imagens de exemplos e na Figura 3 de contraexemplos. A base de imagens está disponível no site do projeto [1].

[1]http://www.icmc.usp.br/~moacir/project/VideoProcessing/

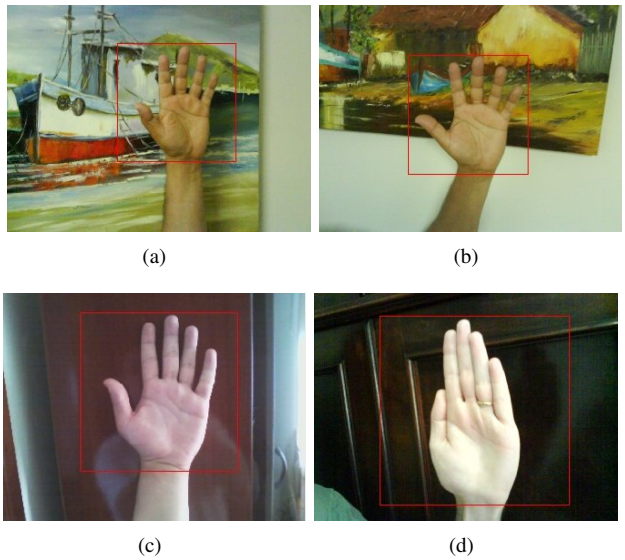


Figura 2. Imagens com exemplos de mãos levantadas. A região da mão, usada no treinamento, está destacada em vermelho.

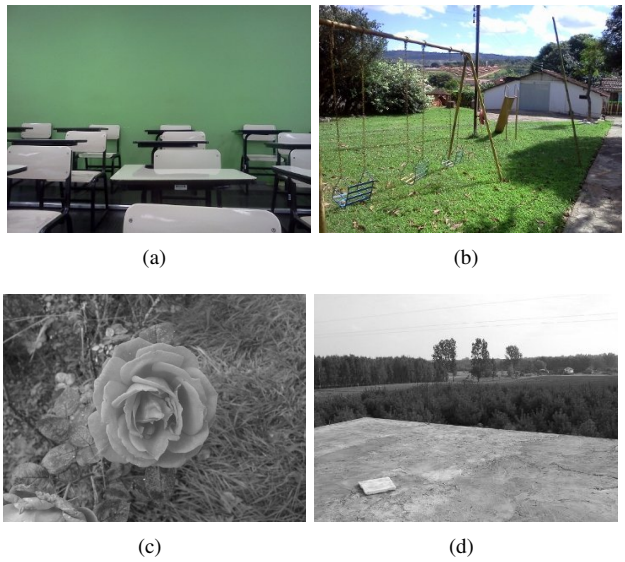


Figura 3. Imagens utilizadas como contraexemplos no treinamento. As imagens (a) e (b) foram obtidas pelos autores, enquanto (c) e (d) foram obtidas de [10].

Devido ao tempo de marcação dos exemplos e de treinamento das cascatas serem muito longos, treinamos uma cascata apenas para detecção de mãos esquerdas e para detectar mãos direitas processamos as imagens invertidas horizontalmente.

V. Experimentos

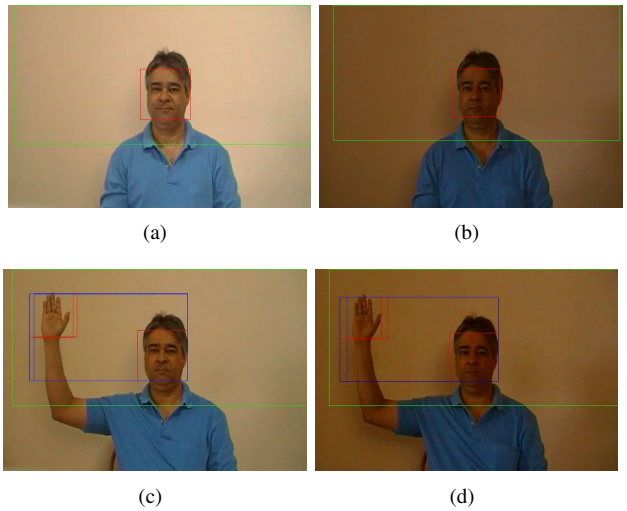
Foram produzidos diversos vídeos para testar o nosso método. Buscamos analisar a eficiência de nosso algoritmo em situações encontradas em videoconferências. As descrições dos testes e alguns frames que esclarecem as propriedades dos vídeos são apresentados na Tabela I.

Nos vídeos 1 e 3 buscamos testar a detecção de mãos levantadas quando há variação de iluminação (Figura 4). A variação

Tabela I
CARACTERÍSTICAS DOS VÍDEOS TESTADOS.

nº	Número de pessoas	Teste de robustez	Resolução	FPS
1	1	variação de iluminação	720x480	29
2	1	gestos como coçar a cabeça, espreguiçar	720x480	29
3	2	variação de iluminação e mútliplos gestos simultâneos	720x480	29
4	1	iluminação externa e movimento da câmera	720x480	29
5	2	múltiplos gestos simultâneos e pessoas em planos diferentes	720x480	29
6	2	múltiplos gestos simultâneos	720x480	29
7	1	oclusão parcial da mão e gestos diferentes do gesto de levantar a mão	640x480	24
8	1	oclusão parcial da mão e gestos diferentes do gesto de levantar a mão	640x480	25
9	1	oclusão parcial da mão	640x480	25
10	1	oclusão parcial da mão e gestos diferentes do gesto de levantar a mão	640x480	24
11	1	oclusão parcial da mão e gestos diferentes do gesto de levantar a mão	640x480	24
12	1	oclusão parcial da mão	640x480	24
13	1	oclusão parcial da mão e gestos diferentes do gesto de levantar a mão	640x480	24
14	1	variação de proximidade da mão	640x480	15
15	1	gestos diferentes do gesto de levantar a mão	640x480	30
16	13	múltiplos gestos simultâneos e pessoas em planos diferentes	1280x720	29

da iluminação gera uma mudança bastante drástica nas cores de frames próximos e com isso há também uma variação na cor de fundo, o que causa problemas em sistemas que se baseiam em buscar regiões onde há movimento para detectar gestos. Embora o nosso sistema não detecte movimento, os vídeos permitem verificar se há boa detecção nessas condições.



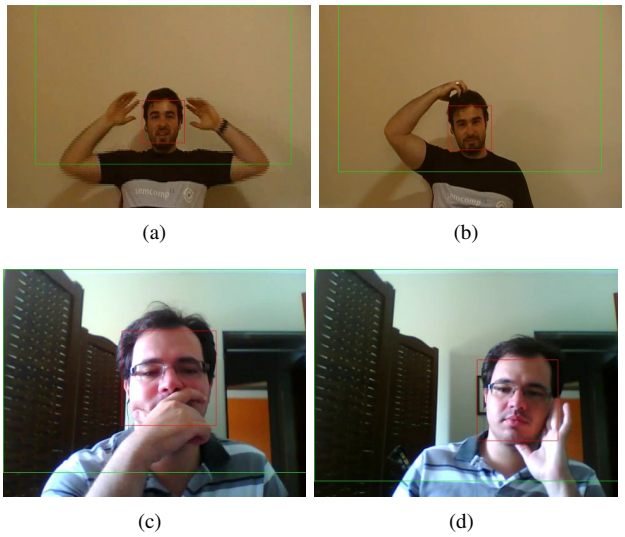
(a) (b) (c) (d)

Figura 4. Frames resultantes da detecção quando há variação de iluminação. As imagens (a) e (c) ilustram a detecção com maior luminosidade; (b) e (d) mostram a detecção no mesmo vídeo com menor luminosidade.

Outro problema é conseguir distinguir entre os gestos nos quais realmente há uma mão levantada e outros gestos. Por exemplo, em uma videoconferência ou uma aula, gestos como espreguiçar e coçar a cabeça são comuns e representam um desafio para um detector automático. Com os vídeos 2, 7, 8, 10, 11, 13 e 15 buscamos verificar como o nosso detector se comporta para esses casos (Figura 5).

Um dos maiores desafios de um sistema de detecção automática é o de fazer detecções quando há oclusão parcial. Para testar nosso sistema nessas situações utilizamos os vídeos de 7 a 13 (Figura 6).

Além dos problemas apresentados acima, em um ambiente



(a) (b) (c) (d)

Figura 5. Frames resultantes da detecção aplicada a imagens em que há outros gestos comuns em videoconferências e aulas.

de videoconferência e em salas de aula haverá naturalmente a ocorrência de múltiplos gestos simultâneos e as pessoas estarão posicionadas em planos diferentes. Os vídeos 3, 5, 6 e 16 trabalham pessoas posicionadas em planos diferentes e executando múltiplos gestos simultâneos (Figura 7).

Embora o foco do sistema seja em ambientes internos, o vídeo 4 foi gravado em ambiente externo para verificar se essas condições de filmagem têm um impacto significativo no desempenho. Um dos principais pontos observados foi a iluminação natural (luz do dia) que é bem diferente da iluminação artificial encontrada nos ambientes fechados. Esse vídeo também inclui movimentação da câmera (Figura 8).

VI. RESULTADOS E DISCUSSÃO

Nosso algoritmo foi aplicado em dezesseis vídeos, realizados sob uma ampla gama de variações, combinadas de diversas maneiras. Essas variações incluíram: iluminação, gestos diferentes daquele de levantar a mão, múltiplos gestos

Figura 6. Frames resultantes da detecção aplicada a frames em que há oclusão parcial.

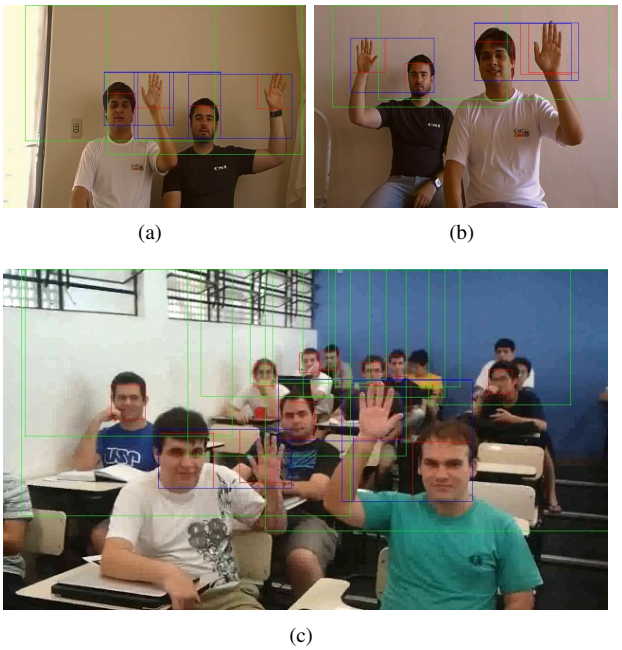


Figura 7. Frames resultantes da detecção quando há múltiplos gestos simultâneos.

simultâneos, pessoas em planos diferentes e oclusão parcial da mão, conforme se pode ver na Tabela I. Os resultados obtidos com a aplicação de nosso algoritmo a esses vídeos estão expostos na Tabela II, na qual apresentamos o número de verdadeiros positivos (*TP: true positive*), falsos positivos (*FP: false positive*), falsos negativos (*FN: false negative*), verdadeiros negativos (*TN: true negative*), frames por segundo (*FPS: frames per second*), precisão, acurácia e tempo de processamento.

O método proposto de detecção de gestos para anotação

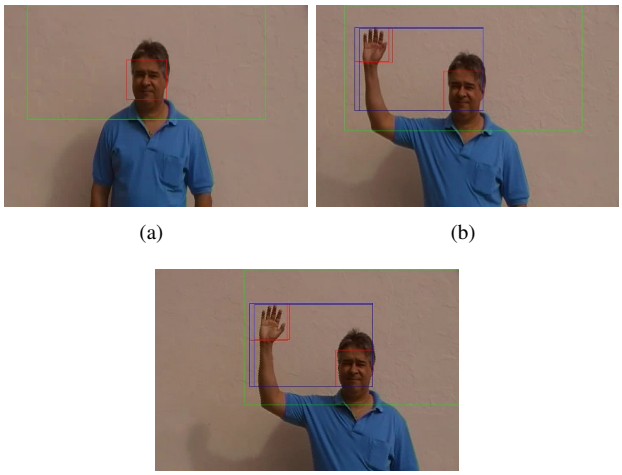


Figura 8. Frames resultantes da detecção em ambiente externo. (b) e (c) ilustram a movimentação da câmera.

automática de vídeo, baseado no algoritmo Viola-Jones melhorado, apresentou os seguintes resultados com relação à robustez:

- **Variação da iluminação**: nosso sistema não mostrou alterações de desempenho, com poucos falsos positivos e poucos falsos negativos nos vídeos 1 e 3, o que levou a uma boa acurácia e precisão. O sistema também não demorou para se adaptar ao novo padrão de luminosidade, como acontece com alguns dos sistemas baseados na detecção do plano de fundo.
- **Outros gestos comuns em um ambiente de videoconferência**: na presença de gestos como coçar a cabeça, espreguiçar e colocar a mão no rosto, o sistema também não teve redução de desempenho. A Figura 5 mostra exemplos de gestos não detectados.
- **Oclusão parcial**: o método não conseguiu realizar a detecção sob oclusão parcial. Isso se dá porque o detector de objetos utilizado não tem um bom suporte a oclusões e, nos casos em que a oclusão ocorre no canto da imagem, o detector não consegue centrar um quadrado na região onde ela se encontra e, por isso, não consegue detectá-la. Podemos ver alguns exemplos na Figura 6, em que os gestos não foram detectados, e podemos observar que os vídeos nos quais há uma quantidade considerável de oclusões (7 a 13) foram os vídeos de pior desempenho. Também notamos no vídeo 15, no qual não há oclusão da mão mas existem frames nos quais há oclusão do rosto pela mão, que houve um aumento nos falsos negativos.
- **Detecção de múltiplos gestos e variação na escala**: a detecção de múltiplos gestos também não afeta o desempenho, desde que não haja oclusões. Porém, houve falha de detecção em gestos muito distantes da câmera, pois nesses casos a mão se torna pequena na imagem e a detecção é dificultada, ver Figura 7.
- **Ambientes externos**: a iluminação natural não diminuiu

Tabela II
RESULTADOS DOS TESTES.

Vídeo #	TP	FP	FN	TN	FPS	Precisão	Acurácia	Tempo (s.)
1	838	4	10	811	29	0.9952	0.9915	545
2	401	7	6	887	29	0.9828	0.9900	291
3	1103	28	1	815	29	0.9752	0.9851	739
4	699	11	7	403	29	0.9845	0.9839	707
5	973	54	84	418	29	0.9474	0.9097	831
6	429	29	18	363	29	0.9366	0.9439	348
7	2	5	91	454	24	0.2857	0.8260	299
8	68	8	355	712	25	0.8947	0.6824	627
9	42	2	180	339	25	0.9545	0.6767	268
10	112	0	182	254	24	1.0000	0.6678	151
11	167	25	162	369	24	0.8697	0.7413	226
12	35	1	223	167	24	0.9722	0.4741	179
13	159	115	76	460	24	0.5802	0.7641	243
14	234	0	55	69	15	1.0000	0.8463	128
15	645	52	286	618	30	0.9253	0.7888	892
16	168	27	94	114	29	0.8615	0.6997	805

o nível de detecção, como podemos ver no resultado obtido no vídeo 4, Figura 8.

- **Movimentação da câmera**: a movimentação da câmera foi testada e não afetou o desempenho da detecção.

VII. CONCLUSÕES

Este estudo investigou e desenvolveu um sistema capaz de detectar gestos de levantar a mão e permitir a análise de filmagens de videoconferências. O método melhorado de Viola-Jones proposto Lienhart e Maydt [7] foi usado como base deste sistema. A análise frame a frame permitiu lidar bem com movimento de câmera, alterações de luminosidade e de plano de fundo, alternância entre câmeras e filmagens com a câmera em movimento.

A partir da análise frame a frame é gerada uma lista com a classificação dos frames — positivos e negativos — e assim verifica-se a existência de intervalos com gestos de mão levantada, realizando a anotação dos frames em que há esse gesto no vídeo.

Estes resultados mostraram que o sistema não foi afetado pelas **variações de iluminação, movimentação da câmera, gestos parecidos com o de levantar a mão e a existência de múltiplos gestos**. As variações de distância, quando acentuadas, podem afetar a detecção. **O sistema falha principalmente na ocorrência de oclusões**.

Por utilizar uma abordagem de análise individual dos frames e não informações temporais, o **método proposto foi menos afetado pelas variações de iluminação**, o que acontece com sistemas baseados na detecção do plano de fundo. A utilização da relação entre dois parâmetros — a existência de uma mão próxima a um rosto como forma de procurar obter uma análise mais precisa na busca de informações — é algo que poderá ser útil em outras aplicações do gênero. Porém, é importante perceber que os trabalhos aos quais nos comparamos fazem análise em tempo real, enquanto nosso projeto realiza o processamento *offline* por rodar três vezes, para cada frame, o detector de Viola-Jones.

Trabalhos futuros poderão tratar melhor a ocorrência de oclusões parciais e também otimizações para que se consiga trabalhar em tempo real. Nesse caso poderia ser estudado o uso de unidades gráficas de processamento (*GPU*), de grande potencial para essa aplicação [4], e a análise de uma quantidade limitada de frames por segundo.

AGRADECIMENTOS

Esse trabalho foi parcialmente financiado pela FAPESP (proc. n. 2011/16411-4). Agradecemos ao CNPq e a USP pelas bolsas de iniciação científica, e aos voluntários que participaram da criação dos vídeos.

REFERÊNCIAS

[1] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[2] Xiaodong Duan and Hong Liu. Detection of hand-raising gestures based on body silhouette analysis. In *Proceedings of the 2008 IEEE International Conference on Robotics and Biomimetics*, pages 1756–1761, Washington, DC, USA, 2009. IEEE Computer Society.

[3] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proc. 13th International Conference on Machine Learning (ICML-96)*, pages 148–156, 1996.

[4] J. Fung. Computer vision on the GPU with OpenCV. GPU Technology Conference, 2011.

[5] Mathias Kölsch and Matthew Turk. Analysis of rotational robustness of hand detection with a viola-jones detector. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004)*, volume 3, pages 107–110, 2004.

[6] Mathias Kölsch and Matthew Turk. Robust hand detection. In *6th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 614–619, Seoul, Korea, 2004.

[7] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *IEEE Proc. Int. Conf. Image Processing (ICIP 2002)*, pages 900–903, 2002.

[8] Anuj Mohan, Constantine Papageorgiou, and Tomaso Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:349–361, 2001.

[9] Constantine P. Papageorgiou, Michael Oren, and Tomaso Poggio. A general framework for object detection. In *Proceedings of the Sixth International Conference on Computer Vision (ICCV 1998)*, ICCV '98, pages 555–, Washington, DC, USA, 1998. IEEE Computer Society.

[10] Naotoshi Seo. Tutorial: Opencv haartraining (rapid object detection with a cascade of boosted classifiers based on haar-like features), 2011.

[11] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. 2001 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, pages 511–518, 2001.

[12] Jie Yao and Jeremy R. Cooperstock. Arm gesture detection in a classroom environment. In *Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision*, WACV '02, pages 153–, Washington, DC, USA, 2002. IEEE Computer Society.