Dario Scott

Identificação de Atividade de Voz Baseada em Vídeo

Fevereiro de 2010

Dario Scott

Identificação de Atividade de Voz Baseada em Vídeo

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre, pelo Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos.

Orientador:
Profa. Dra. Marta Becker Villamil

Universidade do Vale do Rio dos Sinos

Fevereiro de 2010

Ficha catalográfica

S425i Scott, Dario
Identificação de atividade de voz baseada em vídeo / por Dario Scott. – 2010.
63 f. : il. , 30cm.

Dissertação (mestrado) — Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, 2010.
"Orientação: Profª. Drª. Marta Becker Villamil".

1. Algoritmo – Computação. 2. Processamento de imagem. 3. Detecção de bordas. 4. Detecção de fala. 5. Modelo de cor e pele. I. Título.

CDU 004.421

Catalogação na Fonte:
Bibliotecária Vanessa Borges Nunes - CRB 10/1556

Dedico este trabalho
à minha esposa Ana, sempre companheira,
à minha filha Thaís, sempre viva no meu coração e
aos novos herdeiros que estão por vir.

Agradecimentos

Neste espaço eu gostaria de deixar registrado o meu agradecimento a todos aqueles que de alguma forma me apoiaram e contribuíram para que este trabalho chegasse a bom termo. Em primeiro lugar ao Professor Dr. Cláudio Rosito Jung que me mostrou ainda na conclusão da graduação as possibilidades da utilização da matemática no processamento de imagens digitais, seu profundo conhecimento desta área, agregado ao ambiente sempre agradável e estimulante para se trabalhar no laboratório do projeto da HP, somado às importantes observações e discussões ao longo dos 21 meses em que me orientou, acredito que tenham enriquecido este trabalho. Também agradeço a Professora Dra. Marta Becker Villamil que assumiu a orientação deste trabalho na sua fase final. Não posso deixar de mencionar aqui meus colegas de laboratório, Leandro Dhil, José Carlos Bins, Lucas Seewald, Rodrigo Schramm, Alessandro Parolin, Dante Augusto Blauth, Márcio Cappellari e Vicente Minotto que ao longo destes dois anos contribuíram para que fizéssemos os vídeos para testes, além da troca de dicas em C++ que sempre foram bem vindas. A todos meus colegas da turma de mestrado, em especial ao Marcos Kich o qual foi parceiro de muitos trabalhos. Agradeço a todos os professores do Programa de Pós-Graduação em Computação Aplicada da Unisinos, em especial aos Professores Drs. Patricia Jaques Maillard e Leonardo Dagnino Chiwiacowsky que também me auxiliaram com suas criticas pertinentes ao meu relatório de andamento do mestrado. Agradeço a Hewlett Packard, pelo seu programa de pesquisa, que financiou meu mestrado e ainda me deu a oportunidade de participar de várias videoconferências com pesquisadores de várias partes do mundo e reuniões de projetos de pesquisas para desenvolvimento de novos produtos.

Resumo

Atualmente, existem diversos trabalhos com as mais variadas abordagens relativas ao processamento de imagens digitais para detecção de atividade de voz (VAD). As suas aplicações perpassam diferentes áreas, como por exemplo, comandos de voz em veículos e videoconferência. A motivação deste trabalho constitui-se na construção de um algoritmo que contribua para o aperfeiçoamento das técnicas de processamento de imagens aplicadas para a detecção de atividade de voz em vídeos. A problemática envolvida já apresenta uma grande diversidade de abordagens. No entanto, o foco deste trabalho situa-se na busca de alternativas para a melhoria na extração de um modelo de cor de pele e não-pele e, a partir daí, extrair um classificador para identificar a atividade de fala com mais precisão. Algoritmos já existentes de identificação de face e de classificação dos lábios foram utilizados e aprimorados. Através da criação de *patches* abaixo dos olhos, foi criado um modelo para determinar as características individuais de cor de pele por meio de média e desvio padrão dos pixeis dos *patches* e da região da boca. Os resultados encontrados são apresentados baseados em duas abordagens. A primeira, quando se realiza o treinamento somente com imagens sem fala e uma segunda treinando fala e não fala para determinar o classificador de VAD. Este novo modelo de identificação de atividade de voz apresentou um grau de acerto em torno de 80% para a primeira abordagem e 90% para a segunda além de um baixo custo computacional em tempo real.

Processamento de Imagem, Detecção de Fala, Modelo de cor de pele

Abstract

Currently, there are several works with many different approaches to image processing for detection of voice activity (VAD). Its applications cross over different areas, such as voice commands in vehicles and videoconferencing. The motivation of this work consists in building an algorithm that contributes to the improvement of techniques image processing applied to detect voice activity on video. The issue already presents a great diversity of approaches. However, the focus of this work lies in finding alternatives to improve the extraction of a skin and non-skin color model and, from there, extract a classifier to identify the activity of speech more accurately. Existing algorithms of face detection and classification of the lips were used and improved. Through the creation of patches under the eyes, a model was created to determine the individual characteristics of skin color using the mean and standard deviation of the pixels of the patches and the mouth area. The results are presented based on two approaches. The first, when the training is conducted only with images without a speak activity and the second one, with speak and non-speak activities to determine the classifier VAD. This new model for voice activity identification presented a degree of accuracy around 80% for the first approach and 90% for the second, besides a low computational cost in real time.

Voice Activity Detection, Image processing, Skin color model

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AAM: *Active Appearance Models*
CIELAB: *Commission Internationale d'Eclairage L = Luminescence (density) a = red/green b = blue/yellow*
CIELUV: *Commission Internationale d'Eclairage L = Luminescence u = saturation v = hue angle*
CPD: *Coherent Patch Displacement*
EBGM: *Elastic Bunch Graph Matching*
FCMS: *Fuzzy C-Means Shape*
FLDA: *Fisher's Linear Discriminant Analysis*
GMM: *Gaussian Mixture Model*
HMM: *Hidden Markov Model*
HSV: *Hue Saturation Value*
KLT: *Kanade-Lucas-Tomasi*
LRT: *Likelihood Ratio Test*
MAD: *Median Absolute Deviation*
MFCC: *Mel-Frequency Cepstral Coefficients*
PCA: *Principal Component Analysis*
RGB: *Red Green Blue*
ROI: *Region Of Interest*
VAD: *Voice Activity Detection*

Sumário

1 Introdução

Devido ao avanço tecnológico, cada vez mais pessoas têm acesso a imagens digitais, seja por meio de câmeras, celulares, scanner entre outros aparelhos, que posteriormente podem ser modificadas através de programas de tratamento de imagens. Isto permitiu que profissionais das mais variadas áreas se beneficiassem desses avanços, que passaram a ser integrados no seu dia-a-dia. Mais comumente, esses recursos são utilizados nos campos de conhecimento onde a tecnologia está mais presente, como nas ciências da saúde e nas ciências exatas. Por outro lado, a disseminação das câmeras digitais através da redução de seu custo possibilitou uma maior abrangência dos usuários dessa tecnologia, e isso têm propiciado um aumento significativo da comunicação por meio de videoconferência.

De 1964 até hoje, a área de processamento de imagens vem crescendo vigorosamente. Além de aplicações no programa espacial, técnicas de processamento de imagens digitais são atualmente utilizadas para resolver uma variedade de problemas. Embora freqüentemente não relacionados, esses problemas comumente requerem métodos capazes de melhorar a informação visual para a análise e interpretação humanas (GONZALEZ, 2000).

Com o avanço tecnológico, aliado ao aumento da banda de dados e redução dos custos para acesso à internet, cada vez mais pessoas têm acesso a utilização de videoconferência. Por outro lado, a disseminação das câmeras digitais integradas nos microcomputadores através da redução de seu custo possibilitou uma maior abrangência dos usuários dessa tecnologia. Nesse sentido, diversos pesquisadores vêm dedicando seus esforços para conseguir melhorar a performance das videoconferências, cada vez mais utilizadas.

O interesse em métodos de processamento de imagens digitais está vinculado a duas aplicações principais:

- Melhoria de informação visual para interpretação humana;
- Técnicas de processamento para percepção através de máquinas.

Independentemente do tipo de aplicação, a matemática tem um papel fundamental nas técnicas de tratamento de imagens digitais, uma vez que toda imagem digital pode ser trabalhada como uma matriz.

Sabemos que as imagens digitais são definidas em pixel[1]. Nesse sentido, podemos trabalhar com essas imagens por meio de operações com matrizes. Realçar uma imagem está diretamente relacionado às modificações que realizamos sobre cada pixel da mesma.

A utilização de recursos computacionais para melhorar a qualidade de videoconferência é uma área em evidência e existem diversos estudos que abordam o assunto. Um dos grandes problemas em aplicações de videoconferência consiste na detecção de atividade de voz (VAD), ou seja, determinar a presença ou ausência do sinal de voz. O resultado do algoritmo de detecção de atividade de voz (VAD) pode ser utilizado para compartilhar o canal de voz com outras informações multimídia, que quando combinada com a localização original do som pode ser utilizada para redução de ruídos, possibilitando assim um incremento do sinal de voz.

A utilização de um algoritmo eficiente de VAD não se limita a aplicação de videoconferência, pode ser utilizado para reconhecimento da fala, comando de voz em veículos, etc.

Existem várias abordagens diferentes para o VAD. Algumas delas exploram a discriminação entre fala e ruído utilizando as informações de áudio (SOHN et al., 1999; NEMER; GOUBRAN; MAHMOUD, 2001, 2005). Outras abordagens podem utilizar a análise do movimento labial baseado somente em vídeo (SODOYER et al., 2006a; ONG; BOWDEN, 2008; SODOYER et al., 2009b), e uma terceira classe de técnica que explora a análise multimodal (LO; GOUBRAN; DANSEREAU, 2004; PITSIKALIS et al., 2006; PAPANDREOU et al., 2009; MARCEL JOHNNY MARIÉTHOZ; CARDINAUX, 2006; PAPANDREOU et al., 2007; CHIBELUSHI; DERAVI; MASON, 2002; MARAGOS; POTAMIANOS; GROS, 2008), combinando as informações de áudio e vídeo.

Muito embora o problema da localização do áudio possa ser contornado utilizando um *array* de microfones (BRANDSTEIN; SILVERMAN, 1997; DO; SILVERMAN, 2007)[2], o uso

[1] Pixel (abreviatura de *"picture element"*, elemento de figura) é a menor unidade de uma imagem, e quanto maior for o número de pixeles, melhor a resolução que a imagem terá; em imagens em tons de cinza, a quantidade desses tons de cinza é determinada em BITS, sempre demonstrados em potência de 2, ou seja, uma imagem com oito bits, por exemplo, refere-se a 2^8, que representa 256 tons de cinza variando entre o branco e o preto.

[2] *array* de microfones é uma seqüência de microfones dispostos linearmente eqüidistantes entre si, possibilitando através da diferença de tempo para captação do sinal de áudio triangular a distância do interlocutor.

da informação de vídeo para a melhor exploração de fonte de áudio é vantajoso quando não se dispõe desse recurso.

Neste trabalho vamos apresentar alguns resultados que foram obtidos através da demarcação geométrica da região da boca e identificação de atividade de voz, utilizando apenas informação de vídeo.

1.1 Problema

A identificação da fala é de extrema importância para aplicações como a videoconferência, pois ela permite melhorar a qualidade da mesma, não sendo necessária a transmissão de dados referente ao som quando a pessoa está em silêncio.

Apesar de ser um tema bastante estudado, existem alguns desafios que devem ser superados:

- mudanças de iluminação;
- sombras;
- diferenças nos tons de pele entre pessoas distintas;
- oclusões parciais da face;
- necessidade do processamento em tempo real.

Dos fatores apresentados, a questão da iluminação tem uma influência considerável para a classificação da cor de pele e não pele, portanto este é mais um trabalho que busca apresentar resultados significativos nessa área contribuindo com o avanço que temos acompanhado nos últimos anos.

1.2 Contribuição

Como dissemos anteriormente, a matemática é a base para o tratamento de imagens digitais, e sua utilização tem como objetivo fundamental extrair a parte fundamental do problema e formalizá-la em um contexto abstrato. Pensando dessa forma foi que procuramos encontrar uma solução que apresentasse uma resposta satisfatória ao problema sem a necessidade de cálculos complexos e custosos computacionalmente.

Portanto, o objetivo deste trabalho foi procurar definir a região da boca dentro de uma face identificada e desenvolver e implementar um algoritmo em C++ que possibilite sua execução em tempo real sem comprometer o processamento da máquina, procurando identificar se existe fala ou não do indivíduo analisado a cada quadro de uma seqüência de vídeo.

É importante que esse algoritmo não tenha um grande custo computacional para que ele não interfira no processamento de uma videoconferência, pois o mesmo deve ser executado em tempo real.

Como em uma seqüência de vídeo pode haver variação de iluminação e/ou posicionamento do objeto que está sendo analisado, no nosso caso a face humana, é importante que o modelo absorva essa variação e se ajuste a mesma, sem que para isso seja necessário um grande tempo de processamento.

Justifica-se nossa preocupação quanto ao custo computacional para a extração do parâmetro para análise da existência de atividade de voz, pelo fato de que em uma videoconferência, para que o modelo se adapte a uma variação de luminosidade ou distância da câmera por exemplo, ele deve processar muitos quadros de imagem.

Portanto, nossas contribuições neste trabalho são:

- desenvolvimento de um modelo de cor individual da pessoa que está sendo analisada, proporcionando assim, uma melhor discriminação entre pixels de pele e não pele;
- utilização de *patches* na região da face abaixo dos olhos, criando um parâmetro para análise do modelo de cor individual;
- utilização da média e desvio padrão dentro de uma janela temporal para extrair um identificador de fala ou silêncio dentro de um período;

A partir destas três premissas foi desenvolvido um novo modelo de identificação de atividade de voz com um grau de acerto em torno de 90%. É importante ressaltar ainda que a utilização de *patches* na região da face abaixo dos olhos, além de criar um identificador de cor de pele personalizado nos permite compensar uma variação de iluminação entre os dois lados da face, pois a média e o desvio padrão são calculados por *patches* permitindo assim que, mesmo que a região da boca se encontre numa situação em que exista uma variação de luminosidade entre o lado direito e esquerdo da face que está sendo analisada, esta se ajuste para verificar a variação na região da boca, como veremos no decorrer deste trabalho.

Como já foi dito anteriormente, a iluminação tem uma grande influência na classificação de cor de pele e não pele e essa foi uma das razões que nos levaram a utilizar *patches* na região da face abaixo dos olhos.

1.3 Estrutura do trabalho

Este trabalho está estruturado da seguinte maneira. O próximo capítulo apresenta uma revisão bibliográfica procurando abordar a identificação da face, identificação da boca/lábios e mais especificamente sobre algoritmo de VAD.

O método que se pretende adotar na resolução do problema é apresentado no capítulo 3, e os resultados encontrados nas abordagens feitas neste trabalho são mostrados no capítulo 4. Finalmente, as considerações finais e trabalhos futuros são apresentadas no capítulo 5.

2 Revisão Bibliográfica

Este capítulo apresenta os principais conceitos inerentes ao desenvolvimento deste trabalho, através de uma revisão bibliográfica dos temas: identificação de face e VAD. Uma vez que a grande maioria dos artigos especializados da área tem apontado para uma detecção da face, através da combinação de diversas técnicas, podemos subdividir o trabalho desses pesquisadores em duas grandes abordagens.

Um primeiro grupo que se utiliza de padrões geométricos para identificação da face humana e um segundo grupo que se utiliza dos padrões de cores para a mesma identificação. Apesar dessa divisão, um grande número desses trabalhos se utilizam de ambas para melhorar o grau de acerto dos modelos.

Especificamente no nosso trabalho, a identificação da face é necessária para se obter uma referência para determinação da localização da boca, que será o objeto de análise para extração de um classificador para VAD.

Existem muitos trabalhos na área de processamento de imagens que se utilizam da informação visual para analisar atividade de voz. Nesse sentido, esta revisão estará subdividida em três grupos, apesar deles se completarem.

- Identificação da Face;
- Identificação de Lábio e Boca;
- Detecção de Atividade de Voz;

No que se refere à detecção de atividade de voz, existem trabalhos que se utilizam somente da informação visual e outros que trabalham de uma forma combinada entre imagem e som.

Os algoritmos que processam a informação visual para extrair uma característica da imagem que possa quantificar o grau de confiança da ocorrência ou não de fala em um

determinado quadro têm em comum a necessidade de se identificar a região de interesse (ROI), que no nosso caso se trata de identificar a boca e/ou lábios.

A seguir serão apresentadas algumas técnicas de identificação de face que não é o foco deste trabalho, porém é utilizado como uma ferramenta facilitadora para identificar a região da boca.

2.1 Identificação da Face

Antes de identificar a boca e/ou lábio é necessário localizar a face. Existem diversos algoritmos que executam essa tarefa como por exemplo os algoritmos que utilizam características faciais individuais, tais como os olhos, o nariz e a boca. Por exemplo, Hsu e colaboradores (HSU; ABDEL-MOTTALEB; JAIN, 2002) propõem uma detecção à base de componentes da face que construa um "mapa da boca" baseado no espaço de cor de YCbCr conforme podemos verificar nas Figuras 1 e 2.

No trabalho de Hsu et al. (HSU; ABDEL-MOTTALEB; JAIN, 2002), os autores apresentaram bem a utilização de uma composição entre uma filtragem por cor de pele e identificação através de pontos reconhecidos da geometria da face humana.

Os modelos Markovianos ocultos, *("Hidden Markov Model"* HMM) foram e continuam sendo utilizados no reconhecimento da ação onde os dados são essencialmente excedentes de uma dimensão tempo. Nefian e Hayes III (NEFIAN; III, 1998) investigaram o desempenho do reconhecimento e da detecção de um HMM de uma dimensão para imagens da face em escala de cinza. Para imagens frontais da face, as regiões faciais significativas (YANG; KRIEGMAN; AHUJA, 2002) (cabelo, testa, olhos, nariz, a boca) vêm em uma ordem natural de cima para baixo, mesmo se as imagens se submetem a rotações pequenas no plano de imagem e/ou a rotações na perpendicular do plano ao plano de imagem. A cada uma das regiões faciais é atribuído um estado por esse modelo conforme apresentado na Figura 3.

Para utilização desse método de identificação da face é necessário realizar um treinamento com imagens frontais de diferentes povos tomados em diferentes circunstâncias de iluminação. Esse procedimento é necessário para se treinar a HMM.

Pode-se verificar na Figura 4 as divisões por regiões de diversas faces que foram detectadas por esse modelo. No artigo de Liang e colaboradores (LIANG X. LIU; NEFIAN,

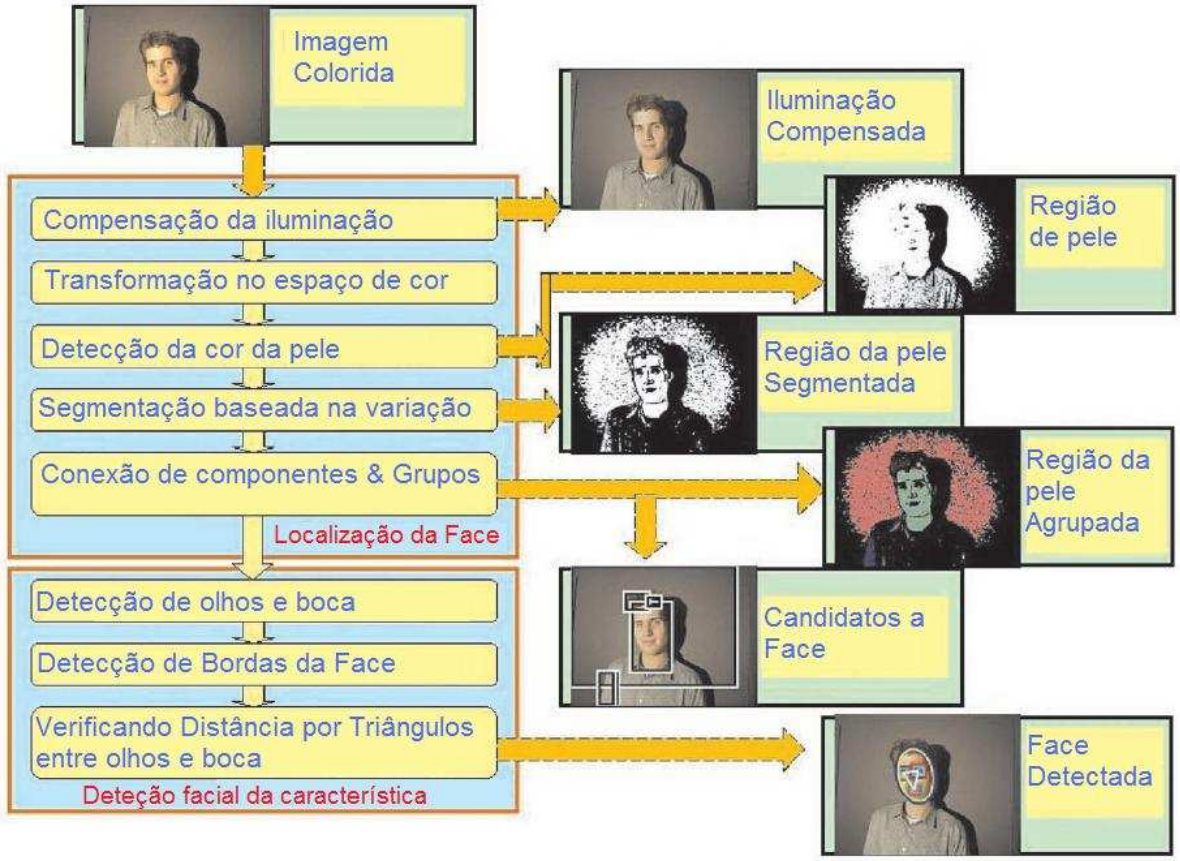


Figura 1: Processo para detecção de face (HSU; ABDEL-MOTTALEB; JAIN, 2002).

2002) também observa-se a utilização do HMM.

Um outro trabalho que também utiliza um filtro para cor de pele e reconhecimento da geometria da face humana é o de Heisele e Poggio (HEISELE; POGGIO, 2006), onde os autores apresentaram uma estrutura à base de componentes para a detecção e a identificação da face. A detecção da face e os módulos da identificação compartilham da mesma arquitetura hierárquica. Ambos consistem em duas camadas de classificadores, em uma camada com um conjunto de classificadores de componentes e uma outra camada com um único classificador da combinação. Os classificadores de componentes detectam independentemente cada região para identificar partes faciais na imagem. Suas saídas são passadas ao classificador da combinação que executa a detecção/identificação finais da face. Os autores descreveram um algoritmo que aprende automaticamente os dois conjuntos separados de componentes faciais para as tarefas da detecção e da identificação. Portanto, foi abordada uma detecção de face por aproximação dos componentes para a detecção da face, treinando um classificador para identificar características faciais (em particular, há um classificador para o canto direito da boca e outro para canto esquerdo). Apesar das vantagens destas aproximações para detectar a face, a identificação individual

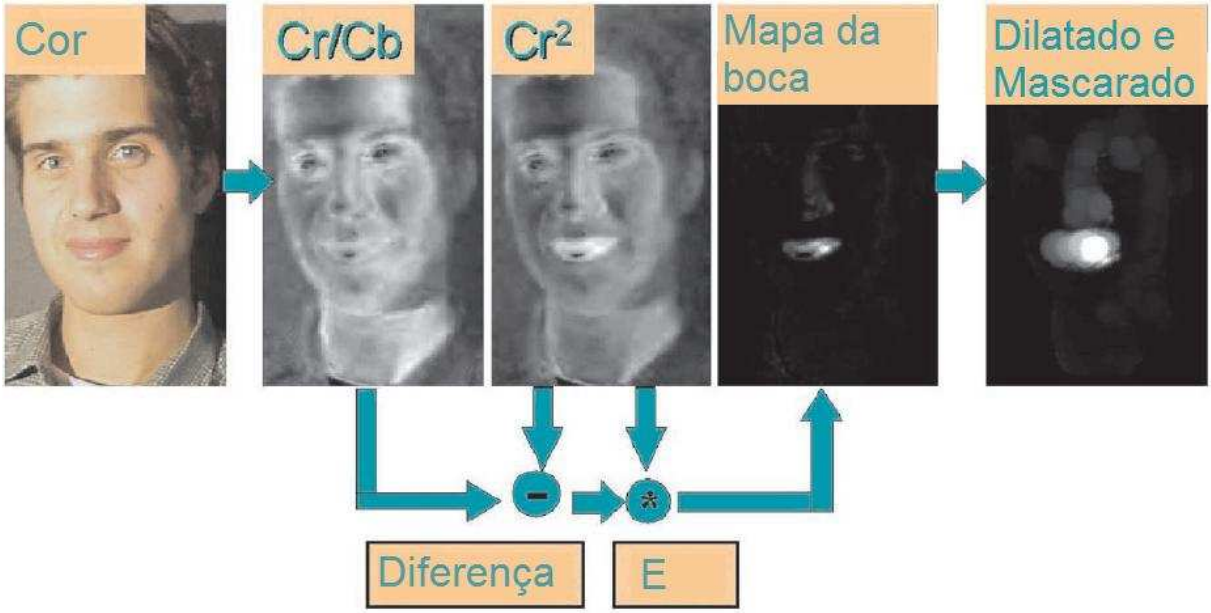


Figura 2: Processo de detecção da boca conforme (HSU; ABDEL-MOTTALEB; JAIN, 2002).

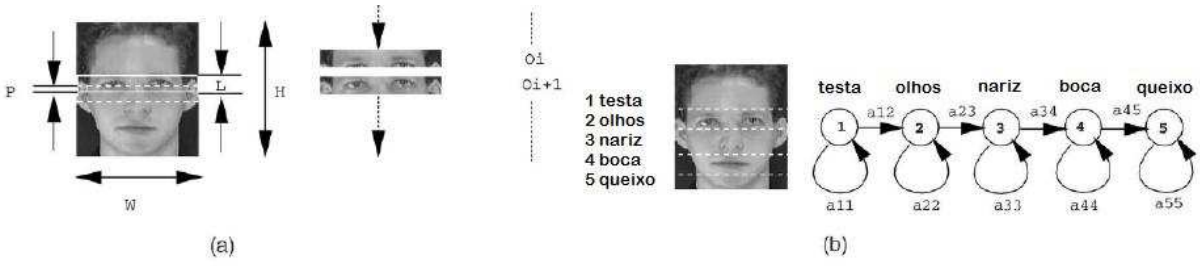


Figura 3: HMM para localização da face. (a) Vetor de observação para treinamento do HMM, onde cada exemplo de face é convertido numa seqüência de vetores de observação. Os vetores de observação são construídos a partir da janela WxL. A seqüencia de observação é construída fazendo uma varredura vertical com um passo de P pixels. (b) Apresenta a demarcação da área, quando são treinadas os 5 estados numa seqüência de vetores observados (YANG; KRIEGMAN; AHUJA, 2002).

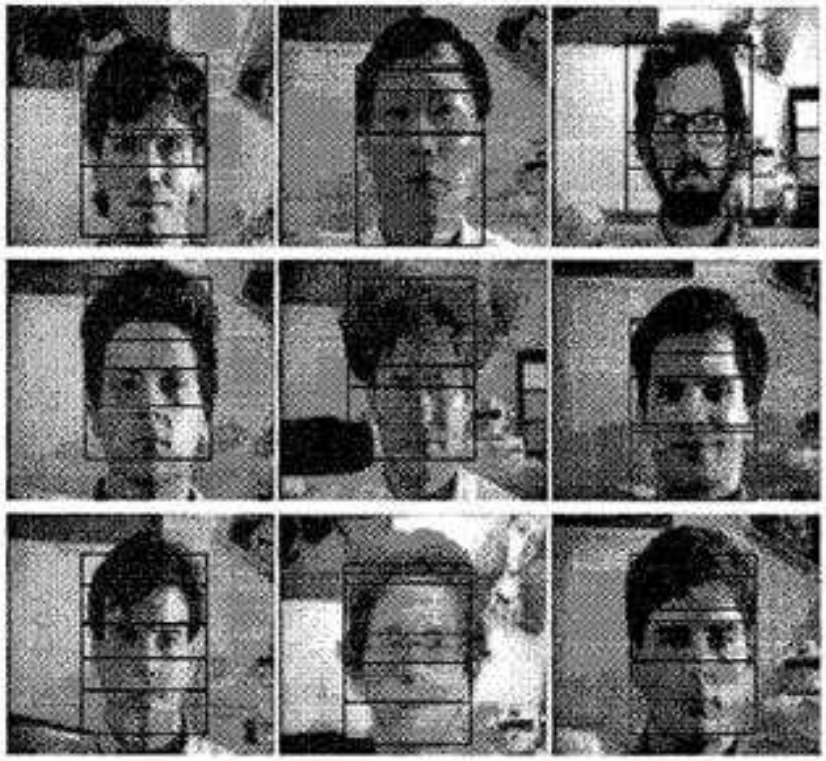


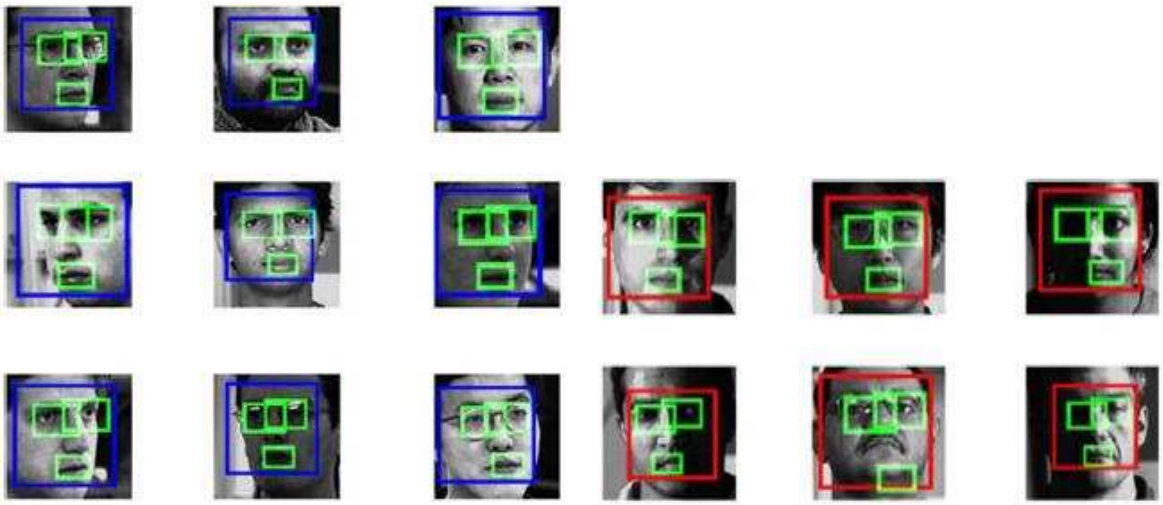
Figura 4: Resultado de Faces detectadas com (HMM) (NEFIAN; III, 1998).

dos componentes faciais não é geralmente muito exata.

Apesar de não se considerar muito exata a identificação de componentes faciais individualmente, pode-se observar através da Figura 5(a) e 5(b), que o resultado obtido na detecção da face por partes apresentou um bom resultado, mesmo quando a face não era frontal e também existindo uma grande variação da luminosidade entre as imagens.

A detecção de face proposta por Viola e Jones (VIOLA; JONES, 2004) utiliza o que os autores chamam de *"integral image"*, o que agiliza a varredura de um quadro da imagem em busca da região candidata à face. Este modelo também identifica a face por regiões. A idéia básica desse algoritmo é deslizar uma janela (W) na imagem para procurar os candidatos à face. Para isso os autores utilizam um conjunto de atributos denominado *"haar-like features"*, conforme podemos observar na Figura 6. Esse conjunto de atributos codificam as diferenças médias de intensidade nas regiões retangulares vizinhas da imagem.

Para que o cálculo seja rápido, é utilizada a *imagem integral* , contendo o somatório de qualquer região retangular da imagem original, conforme podemos verificar na Figura 7. A vantagem na utilização desses atributos é que eles conseguem representar pequenas estruturas sem se ater ao detalhe. Segundo os autores, esse tipo de atributo é menos



(a) (b)

Figura 5: (a)Faces identificadas, (b)Faces não identificadas (HEISELE; POGGIO, 2006).

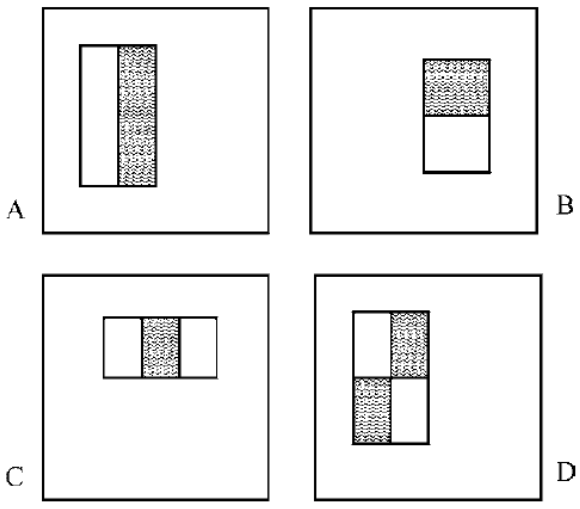


Figura 6: Tipos de *"haar-like features"* utilizadas por Viola e Jones (VIOLA; JONES, 2004) (A) e (B)Estrutura com dois retângulos, (C)Estrutura com três retângulos e (D)Estrutura com quatro retângulos.

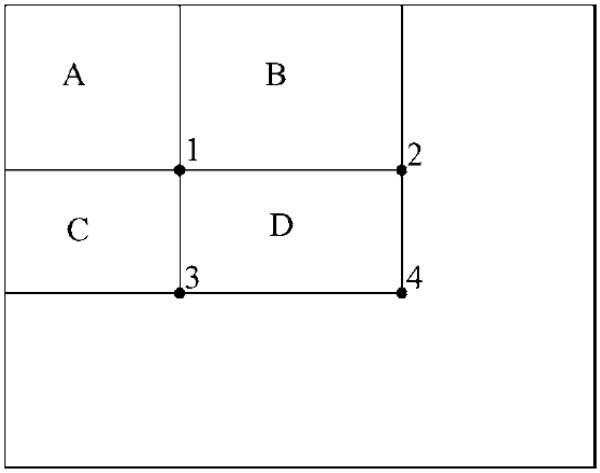


Figura 7: A soma dos pixels do retângulo D pode ser computado pelos quatro vetores de referência. O valor da imagem integral *"integral image"* em 1 é a soma dos pixels do retângulo A. O valor da localização 2 é dado por $A + B$, a localização 3 por $A + C$, e a localização 4 é dada por $A + B + C + D$. A somatória dos pixels de D podem ser computados por $4 + 1 - (2 + 3)$ (VIOLA; JONES, 2004).

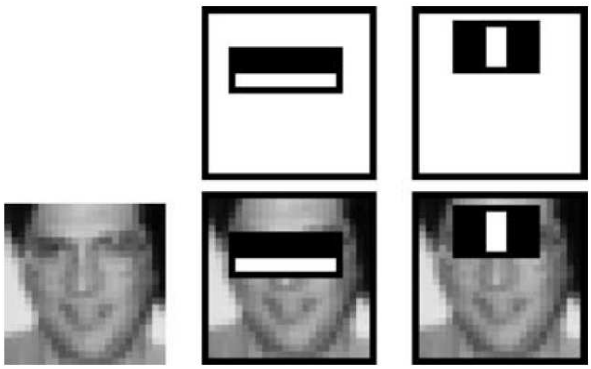


Figura 8: Exemplos de retângulos utilizados na detecção de face pelo AdaBoost reconhecendo o padrão da face humana (VIOLA; JONES, 2004).

sensível ao ruído e tolera pequenas variações na luminosidade.

A técnica de detecção através de padrões geométricos é denominada de *Haar-like features*. Uma forma retangular simples *Haar-like feature* pode ser definida da diferença entre a diferença da soma dos valores dos pixels das áreas dentro do retângulo como visto na Figura 6.

Para a detecção da face, a região dos olhos pode ser considerada uma área a ser pesquisada, a soma dos valores dos pixels em escala de cinza nessa região será maior se levarmos em consideração as áreas próximas como a área inferior dos olhos e o nariz. As *Haar-like features* utilizadas nessa busca seriam as mostradas na Figura 8. Todas as regiões da imagem que passem por esse teste seriam candidatas a serem classificadas como

face humana.

Uma maior combinação de retângulos poderia ser utilizada para classificar a face humana, porém uma maior combinação resultaria num processamento maior para se atingir o resultado. Viola e Jones utlizaram uma técnica denominada *AdaBoost (Adaptative Boosting)* que é um algoritmo de aprendizagem. É um algoritmo de reforço por filtragem e tem acesso a um modelo de aprendizagem fraca. O *AdaBoost* se ajusta de forma adaptativa em relação aos erros da hipótese fraca retornada pelo modelo de aprendizagem fraca. Como resultado, é combinada uma coleção de classificadores fracos para formar um classificador mais forte. O modelo apresentado por Viola e Jones para reconhecimento de face foi o adotado neste trabalho por apresentar um resultado satisfatório para o processamento em tempo real.

2.2 Identificação de Lábio e Boca

A identificação da região da boca e/ou lábio pode ser encontrada de diversas formas, inclusive eles são muito utilizados para a identificação da face como pudemos ver anteriormente. Podemos subdividir em três grupos a identificação de regiões da face de acordo com os padrões utilizados:

- Padrões geométricos;
- Padrões de cor;
- Combinação dos padrões geométricos e de cor;

2.2.1 Padrões geométricos

A Figura 9 mostra um fluxo de uma harmonização elástica da ligação do grafo (EBGM). Dada uma imagem, o grafo é colado à mesma, e então começa a busca local para extração dos lábios. Finalmente o grafo é extraído depois que todas as posições dos pontos da característica facial são combinadas.

Neste modelo, os autores (AOKI K. MASUDA; ARIKI, 2007) determinam a região provável da boca e extraem o contorno dos lábios, depois determinam se a pessoa está falando ou não através da relação entre altura e comprimento da boca conforme pode-se observar na Figura 10.

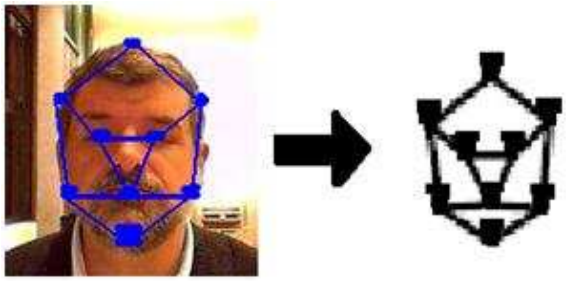


Figura 9: *Elastic Bunch Graph Matching* (EBGM) (AOKI K. MASUDA; ARIKI, 2007).

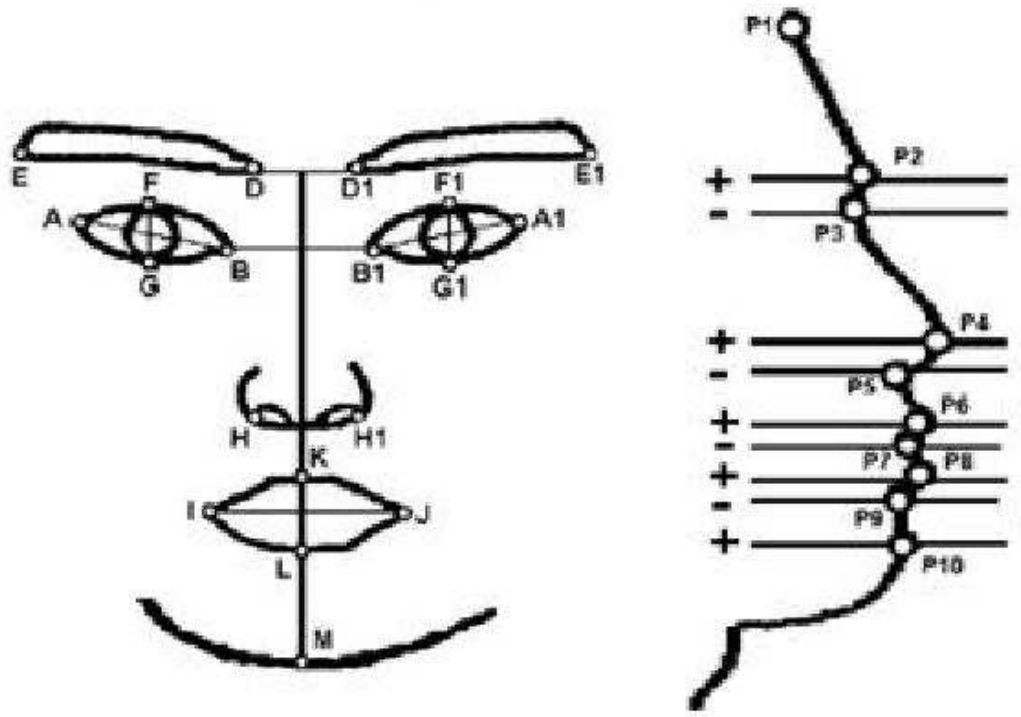
Figura 10: Razão labial (EBGM) (AOKI K. MASUDA; ARIKI, 2007).

Uma outra abordagem é dada no artigo de Pantic e Rothkrantz (PANTIC; ROTHKRANTZ, 2004) que nos mostra quais são os pontos referenciais da face humana que podem ser extraídos, conforme mostrado na Figura 11.

Eveno et al. (EVENO; CAPLIER; COULON, 2004) propuseram um algoritmo de segmentação labial quase-automático. Seu método utiliza uma *"jumping snake"* para obter o contorno inicial dos lábios, e um modelo paramétrico baseado em um conjunto de pontos do contorno é usado para representar a forma. Estes pontos de controle são seguidos através do tempo usando uma variação do rastreador KLT, e o contorno correspondente é obtido através do modelo paramétrico. Um inconveniente desta aproximação é a intervenção manual exigida para inicializar a curva.

2.2.2 Padrões de cor

Wang e seus colaboradores (WANG; LAU; LEUNG, 2004) propuseram um algoritmo para a extração automática do contorno do lábio utilizando a informação da cor. Inicialmente, segmentaram os lábios utilizando uma função *"Fuzzy c-means with shape"* (FCMS) nos espaços de cor CIELAB e da CIELUV, depois aplicaram um procedimento de equalização da luminância para corrigir mudanças da iluminação. O contorno do lábio foi extraído



E, E1: canto exterior da sobrancelha
D, D1: canto interior da sobrancelha
A, A1: canto exterior do olho
B, B1: canto interior do olho
F, F1: parte superior do olho
G, G1: parte inferior do olho
H, H1: canto exterior da narina
K: parte superior do lábio
L: parte inferior do lábio
I, J: canto da boca
M: ponta do queixo

P1 : parte superior da testa
P2 : arcada da sobrancelha
P3 : raiz do nariz
P4 : ponta do nariz
P5 : maxila superior
P6 : lábio superior
P7 : borda do lábio
(se a lingüeta é visível:
P7 : lingüeta da borda do lábio
P7′ : a ponta da lingüeta
P7′′: lingüeta da borda do lábio)
P8 : lábio inferior
P9 : maxila inferior
P10: ponta do queixo

Figura 11: Pontos de característica dos contornos dos componentes da face frontal e do perfil (PANTIC; ROTHKRANTZ, 2004).

usando 16 pontos de controle utilizando um procedimento de otimização. O procedimento pode ser executado em tempo real e apresenta uma exatidão elevada, mas as imagens da entrada são regiões escolhidas em torno dos lábios, que simplifica claramente o problema.

Pantic et al. (PANTIC; ROTHKRANTZ, 2004) também se utilizaram da cor para extração da boca. No artigo (RURAINSKY; EISERT, 2003) observa-se a identificação da boca e olhos para animação 3D para videoconferência. Aqui os autores também trabalham com uma segmentação de cor para distinguir a cor da boca da cor da pele. Conforme pode-se observar na Figura 13 o espaço de cor da boca no padrão RGB é diferente do espaço de cor da pele. Portanto, ao trabalhar com as médias dos espaços de cor, aumenta-se a fronteira entre elas.

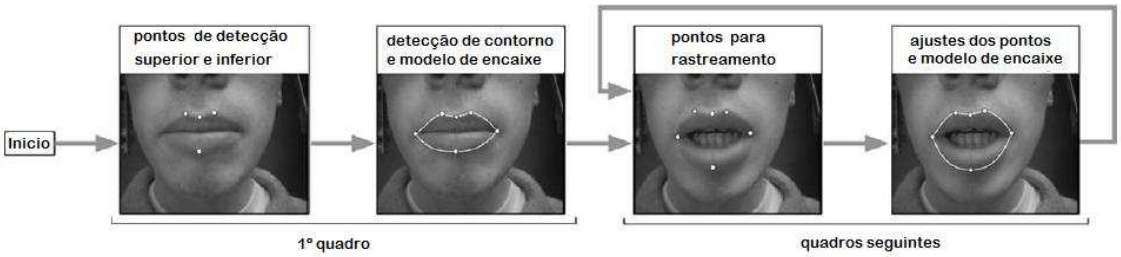


Figura 12: Proposta de (EVENO; CAPLIER; COULON, 2004).

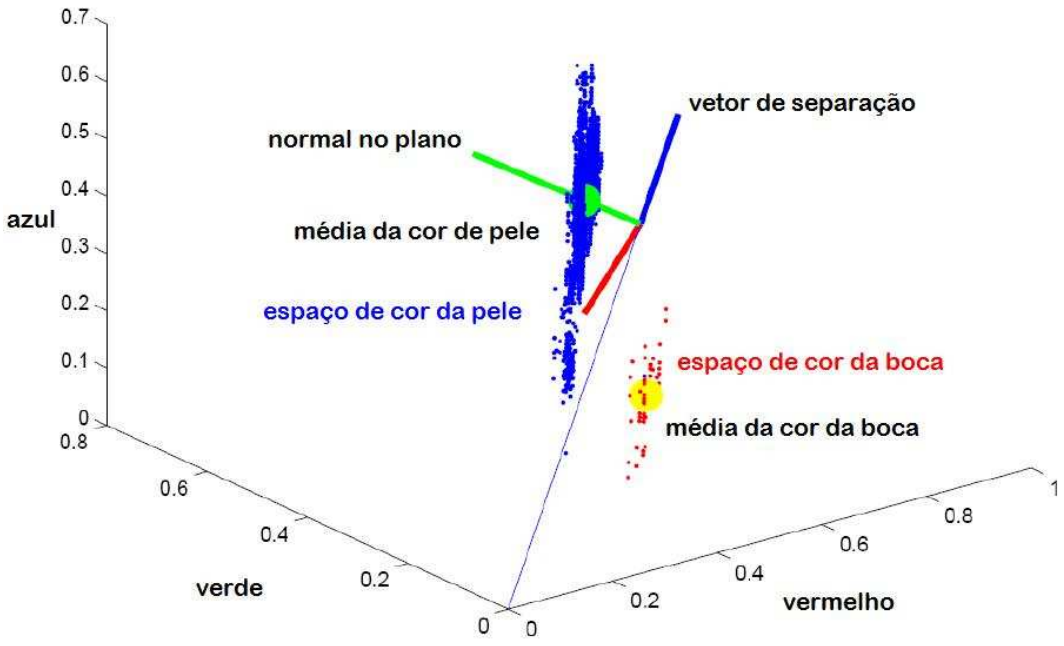


Figura 13: Plano de cor do RGB do frame um da seqüência de Akiyo (RURAINSKY; EISERT, 2003).

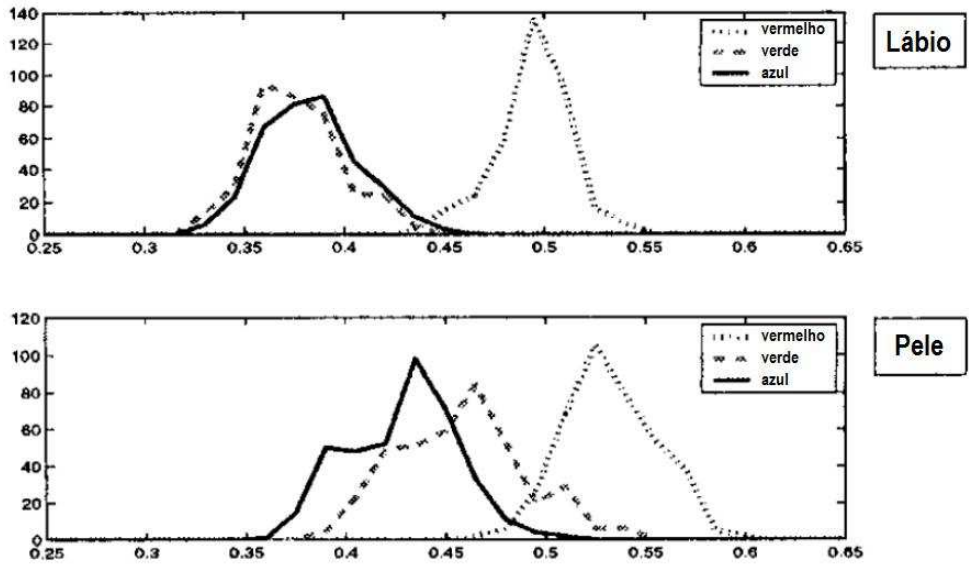


Figura 14: Histograma do canal de cor vermelho R (pontos), verde G (tracejado) e azul B (linha) (EVENO; CAPLIER; COULON, 2001).

A técnica do realce do lábio proposto por Eveno et al. (EVENO; CAPLIER; COULON, 2001) se baseia na diferença dos canais RGB conforme pode-se verificar na Figura 14. Os autores (EVENO; CAPLIER; COULON, 2004) também apresentam um modelo de rastreamento por pontos dos lábios conforme podemos verificar na Figura 12. Outro autor que também apresenta um rastreamento por pontos dos lábios é (ONG; BOWDEN, 2008) conforme podemos verificar na Figura 16.

Os autores reduzem o ruído da imagem através de um filtro e depois trabalham na redução da dependência da luminância. Segundo os autores, se $I_c(x, y)$ denota a intensidade do canal de cor $c \in \{R, G, B\}$ no pixel (x, y), e $I_l(x, y)$ denota o componente luminância (a componente Y no espaço de cor YCbCr). Os autores apresentam uma equação para a redução da luminância onde apresentam os parâmetros a e b que controlam o peso da luminância na lei da correção, onde $c \in \{R, G, B\}$. Segundo os autores, a precisão dos parâmetros a e b não é crítica. Os autores não detalham como chegaram aos valores assumidos de $a = 0.4$ e $b = 0.8$. Depois, calcula-se a componente de correção da luminância I'_c, como a pseudo matiz $h(x, y)$, e finalmente traça-se uma parábola passando por três pontos de controle.

Este modelo de segmentação de cor de pele foi o utilizado neste trabalho pelo bom resultado apresentado nos testes realizados e será mais detalhado no próximo capítulo.

2.2.3 Padrões geométricos e de cor

No trabalho de Pantic et al. (PANTIC; TOMC; ROTHKRANTZ, 2001) por exemplo, os lábios são extraídos aplicando-se um filtro para o domínio do vermelho, os autores aplicam uma linha de corte na distribuição dos pixeis de 60%, conforme pode-se verificar na Figura 15.

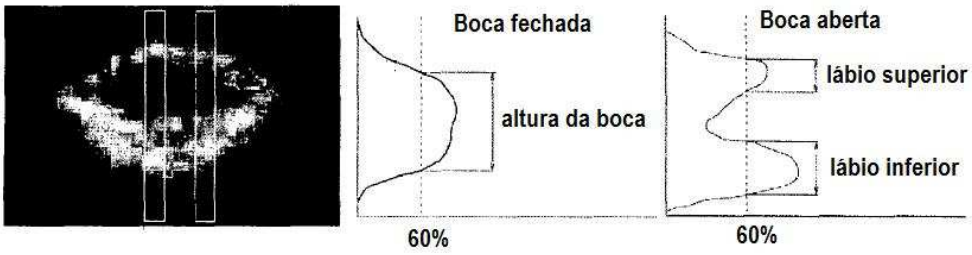


Figura 15: Processo para detecção de atividade de voz (PANTIC; TOMC; ROTHKRANTZ, 2001).

2.3 Detecção de Atividade de Voz

O primeiro problema em algoritmos de VAD baseados em vídeo é a extração da boca/lábios, e como descrito anteriormente, existem diversas abordagens que tratam dessa questão. Também existe um grande número de estudos para detecção de voz e direcionamento de microfone utilizando sinais acústicos e imagens faciais simultaneamente. Nos trabalhos relacionados com a extração dos lábios são utilizadas várias características como RGB, HSV, iluminação e vetor de movimento do contorno para identificar os mesmos.

Podemos verificar isso no trabalho de Aoki et al. (AOKI K. MASUDA; ARIKI, 2007), onde os autores utilizam o *"Elastic Bunch Graph Matching"* (EBGM) para uma extração mais refinada do contorno labial. Neste caso, o processo acústico, voz e ruído são discriminados por *"likelihood ratio test"* (LRT), e o processamento visual permite separar outras vozes ou ruído pela detecção do movimento labial.

Misturas de Gaussianas (GMM) também são extensamente utilizadas para detecção de atividade de voz através de sinal do microfone porque é um modelo relativamente fácil de ser treinado e geralmente poderoso. GMMs são expressas pelo sinal acústico x_t (MFCC: *"Mel-Frequency Cepstral Coefficients"*) no instante t. Utilizando-se uma função de distribuição normal $N(x_t; \mu_m; C_m)$ para modelar o $m-$ésimo componente da mistura,

misturando o vetor de média μ_m e a matriz de covariância C_m, temos:

$$P(x_t) = \sum_m P(m) N\left(x_t; \mu_m, C_m\right). \tag{2.1}$$

No caso apresentado pelos autores, o modelo foi treinado com dados de voz e dados de não voz. E a razão de probabilidade foi calculada por:

$$L(x_t) = log\left[\frac{P(x_t|voicemodel)}{P(x_t|non-voicemodel)}\right], \tag{2.2}$$

onde $Pr(x_t|voicemodel)$ e $Pr(x_t|non-voicemodel)$ são a probabilidade de voz e não voz respectivamente. Para evitar a interrupção da voz pelas pequenas pausas na fala, a razão de probabilidade de voz é suavizada pela expressão:

$$L'(x_t) = \frac{1}{n}\sum_{j=t-\frac{n}{2}}^{t+\frac{n}{2}} L(x_j), \tag{2.3}$$

que é então explorada para VAD.

Conforme descrito anteriormente, existe um grande número de trabalhos relacionados com VAD por vídeo, sozinha ou combinada com outras informações no processamento multimodal (geralmente, áudio). Todos estes algoritmos têm em comum o fato de que identificam a boca/lábios, extraindo algum tipo de característica que é utilizada para detectar a atividade da voz. Entretanto, podem diferir significativamente em como a boca e/ou os lábios são identificados, e em como a informação temporal é explorada.

Ong e Bowden (ONG; BOWDEN, 2008) propuseram um algoritmo de segmentação labial usando um conjunto de atributos lineares, que poderiam ser usados para VAD. Na perspectiva dos autores, um conjunto de pontos é colocado no contorno dos lábios (em uma imagem em tons de cinza), e cada ponto é acompanhado independentemente usando um atributo linear. De fato, os atributos múltiplos são agrupados em um único conjunto para melhorar a robustez do modelo. Apesar dos bons resultados apresentados por este algoritmo, ele exige uma inicialização manual dos pontos de característica.

Liu e Wang (LIU; WANG, 2004) utilizaram um algoritmo de segmentação do lábio para obter uma região do interesse (ROI) em torno da boca, e usaram a análise por componentes principais (PCA) para extrair um vetor da característica 6D. A distribuição de vetores da característica relacionada ao silêncio é modelada como uma única distribuição Gaussiana, e a característica relacionada à fala como uma mistura de distribuições Gaussianas. Na região de fronteira o VAD é executado como um problema de classificação. Os autores

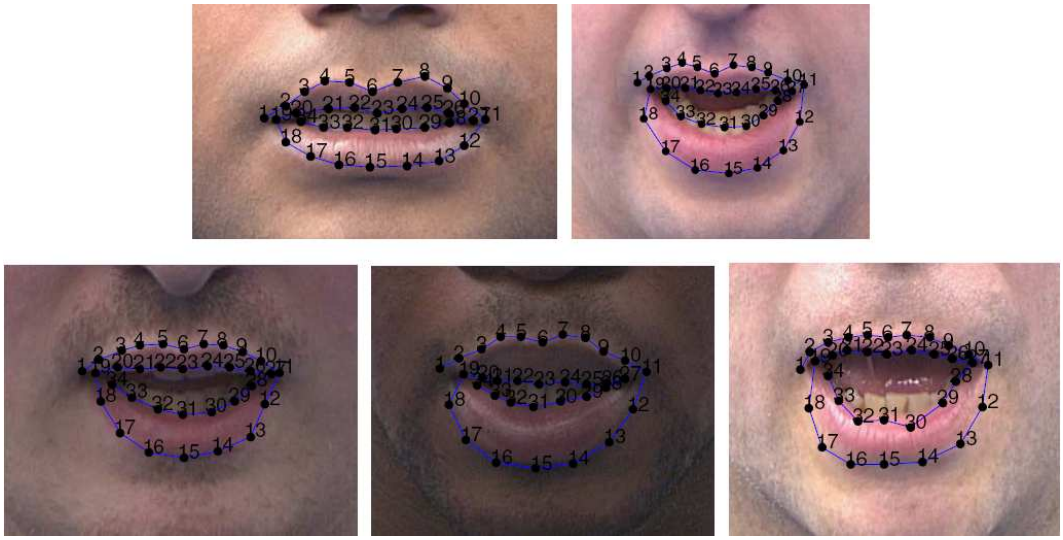


Figura 16: Imagem com 34 pontos de rastreamento em 5 sujeitos (ONG; BOWDEN, 2008).

reivindicam uma melhoria significativa sobre um algoritmo competidor (TANYER; OZER, 2000) (redução relativa de 98.4% no erro da quebra de sentença), mas sofre da sensibilidade potencial às mudanças da iluminação.

Sodoyer e seus colaboradores (SODOYER et al., 2006b) propuseram uma abordagem para detectar a atividade ou a não atividade de voz explorando a coerência entre o sinal acústico da fala e a movimentação labial. A informação visual empregada em seu trabalho é baseada na diferença temporal da largura e da altura interlabial, alisada com um filtro passa baixa *(low-pass)* que considera as amostras nos quadros precedentes. Apesar dos bons resultados conseguidos por esta aproximação, é difícil avaliar a influência dos erros quando da detecção da largura e da altura do lábio. De fato, o estágio da extração do lábio foi feito sob situações controladas (SODOYER et al., 2006b), que não se encontra em aplicações práticas de videoconferência.

Ainda, Sodoyer e seus colaboradores (SODOYER et al., 2009a) realizaram um estudo do movimento labial e a sua relação com a atividade de voz (VAD). Os autores utilizam as medidas de altura e largura do contorno interno dos lábios para análise de atividade de voz. Neste trabalho os autores utilizam uma situação controlada para extração do lábio conforme podemos verificar na Figura 17.

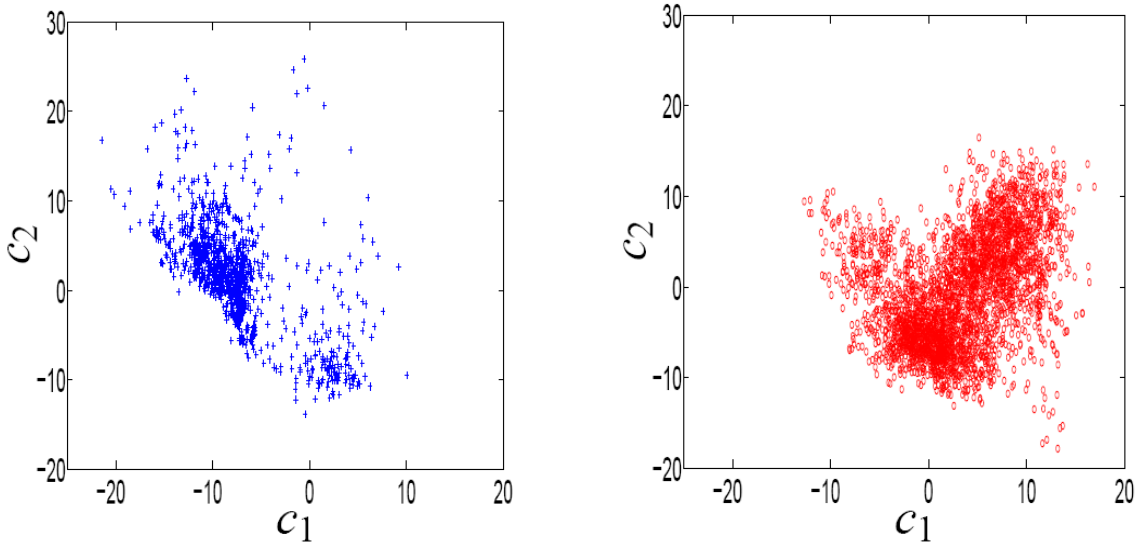
Aubrey e seus colegas (AUBREY et al., 2007) propõem dois detectores de atividade de voz visuais, baseados em modelos de aparência *(appearance models)* e filtro de retina, respectivamente. O primeiro dos métodos utiliza um modelo de aparência ativo (AAM) para extrair os lábios, e uma cadeia oculta de Markov (HMM) para VAD. O segundo detector emprega um filtro retinal em uma região em torno dos lábios, e a diferença

Figura 17: Ilustração da captura de áudio e vídeo em salas separadas (SODOYER et al., 2009a).

temporal é explorada para construir um valor para VAD. Baseado em suas experiências, os autores concluíram que o uso da informação do AAM *a-priori* era mais consistente com a detecção das seções do não-discurso que contêm movimentos complexos do lábio, e o filtro retinal era mais consistente na detecção do não-discurso onde os lábios mostram menos movimento. Deve-se observar que um AAM é geralmente demorado, enquanto que a filtragem retinal é muito mais rápida. Entretanto, o último exige um algoritmo auxiliar para obter a região labial. Pode-se verificar na Figura 18 que existe uma distribuição bem diferenciada para a condição de fala e de silêncio.

Borgström e Alwan (BORGSTROM; ALWAN, 2008) propuseram um modelo de contorno labial para reconhecimento de fala. Na perspectiva deles, tanto o lábio superior como o inferior são modelados por curvas parabólicas, obtidas por mínimos quadrados dos pontos de contorno do lábio. Contudo não fica claro no artigo como são extraídos os pontos de contorno dos lábios.

Apesar de existirem diversos algoritmos baseados em vídeo para VAD, vários desafios ainda estão abertos. O primeiro é conseguir extrair de uma forma robusta a boca/lábio sob uma condição de iluminação variada. O outro é a obtenção de uma medida da boca/lábio



(a) (b)

Figura 18: Distribuição em duas dimensões do atributo de (AUBREY et al., 2007). (a) Silêncio. (b) Fala.

que possa ser utilizada para determinar se existe ou não atividade de voz.

A contribuição deste trabalho é exatamente extrair a região da boca em imagens coloridas utilizando um modelo de pele adaptativo em tempo real (SCOTT et al., 2009). Para que isto ocorra, foi necessário identificar uma região da face onde pudéssemos extrair as características da cor de pele do sujeito em estudo e a partir desses dados, determinar se existe uma variação na região da boca. A evolução temporal da estimativa do padrão de pele na boca é então explorada para VAD. O nosso método será detalhado no próximo capítulo.

3 Modelo Proposto

O primeiro passo da proposta é identificar uma face[1] em um quadro de uma seqüência de vídeo. Existem diversas formas de identificar uma face em uma seqüência de vídeo (YANG; KRIEGMAN; AHUJA, 2002), neste caso utilizamos o detector proposto por Viola e Jones (VIOLA; JONES, 2001) por apresentar um bom resultado.

Uma vez que a face esteja identificada, conseguimos determinar o centro e a escala dela. Essas informações são essenciais para se determinar uma localização específica na face dentro da imagem obtida num quadro de uma seqüência de vídeo. De fato, temos o ponto $c(x, y)$ que determina o centro da face e o raio r que determina o tamanho da mesma. Se o indivíduo se movimenta diante da câmera o ponto c se altera, ao passo que se ele se distanciar da câmera o valor de r diminuirá, ocorrendo o inverso quando ele se aproxima da câmera. Essas são apenas algumas das variáveis que temos que acompanhar para conseguir rastrear as informações do indivíduo.

Para que não seja preciso processar toda a imagem, vamos determinar áreas de interesse e concentrar o processamento nelas. Ainda para incrementar a velocidade de execução podemos adotar uma região de interesse (ROI) para acompanhar a face ao invés de pesquisar em todo a região da imagem a cada quadro. Isto quer dizer que partindo do centro da face identificada determinamos uma área ao redor da mesma para pesquisar no próximo quadro.

Apresentamos um diagrama na Figura 19 de como se desenvolve o processo para a obtenção do classificador de cor de pele e VAD propostos. O treinamento para extração da média e desvio padrão se dá durante n quadros do vídeo. Muito embora esse parâmetro possa ser alterado para mais ou para menos, 75 quadros é o suficiente para determinar os valores da média e desvio padrão para inicialização do processo do classificador de VAD.

Baseado na posição e tamanho da face existem diversos dados antropométricos que

[1]Para simplificar o modelo vamos utilizar somente a primeira face identificada para analisar a atividade ou não de voz independente de outras faces que possam ser identificadas posteriormente.

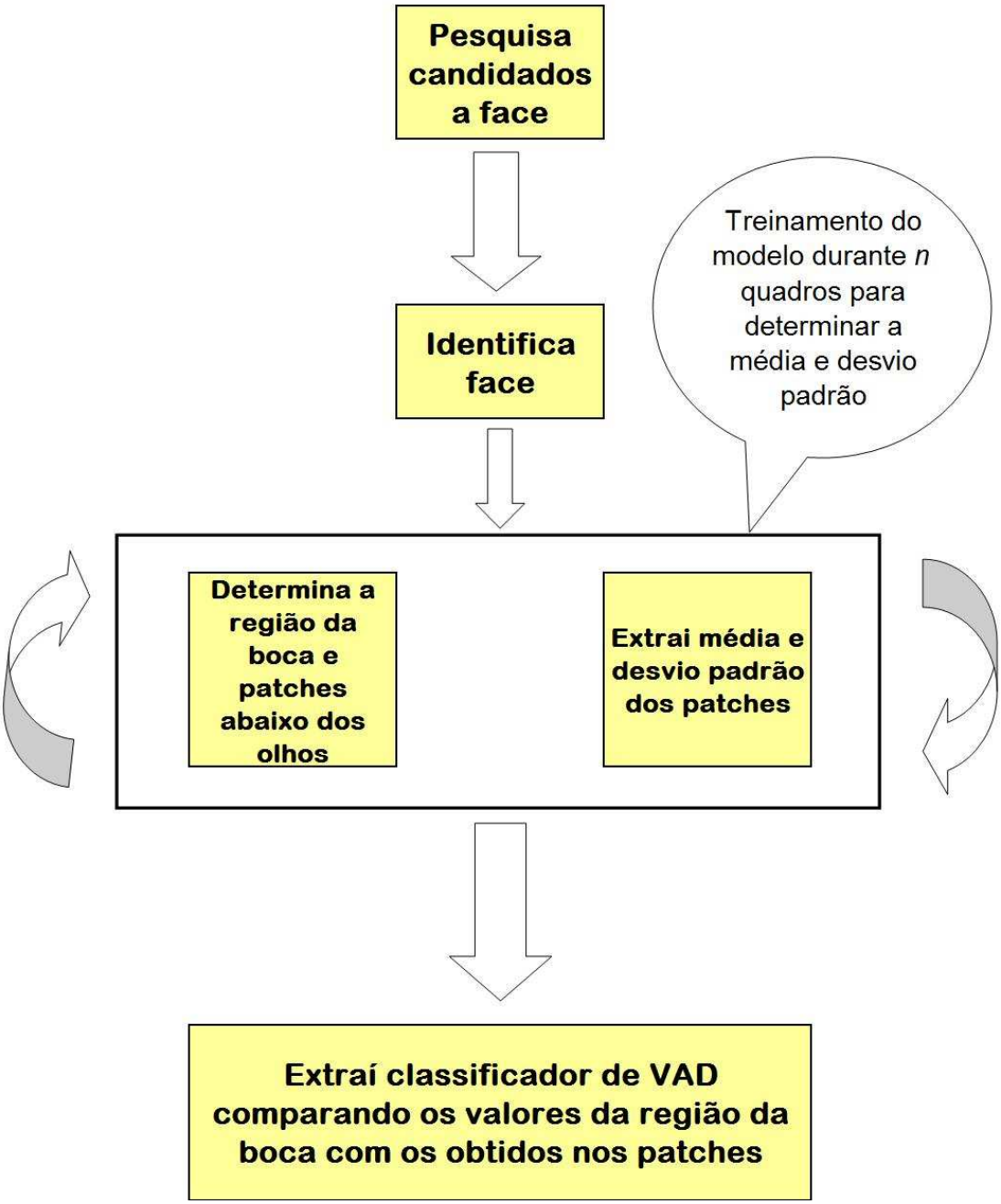


Figura 19: Processo para detecção de VAD (SCOTT et al., 2009).

Figura 20: Posição dos *patches* e região da boca na face (SCOTT et al., 2009).

nos permitem localizar aproximadamente os componentes da face como olhos, nariz e boca (FARKAS, 1995) *apud* (RURAINSKY; EISERT, 2003). No nosso caso, determinamos duas regiões abaixo dos olhos para extrair o modelo de cor de pele e depois determinamos aproximadamente a região da boca para realizar comparações que serão detalhadas posteriormente. A Figura 20 nos mostra as regiões de onde serão extraídos os dados de cor de pele abaixo dos olhos e a região de interesse, a boca, que se encontra preenchida de azul para os pixeis classificados como cor de pele e verde para não pele.

3.1 Discriminação em tempo real da pele/boca

Apesar da existência de diversos modelos para detectar pixeis da pele, usando uma variedade de espaços de cor, seu desempenho é afetado significativamente pela influência das condições de iluminação (SCHMUGGE et al., 2007). Em particular, um problema adicional levanta-se quando duas fontes diferentes iluminam ambos os lados da face (por exemplo luz natural em um lado e fluorescente no outro, quando uma pessoa senta ao lado de uma janela).

O objetivo neste trabalho é diferenciar pixeis da pele dos pixeis da boca como mostrado na Figura 20. Para essa finalidade, nós apresentamos uma modificação da técnica do realce do lábio proposto por Eveno et al. (EVENO; CAPLIER; COULON, 2001), utilizando pixeis da pele para um treinamento em tempo real.

Nesse sentido, foi necessário definir um modelo de identificação de pele para servir como referência. A melhor solução nos pareceu ser a região inferior aos olhos, por se tratar de uma região frontal, simétrica e com ausência de pelos. A definição dessa região

será detalhada mais adiante neste trabalho.

A primeira etapa é reduzir a dependência da luminância. Se $I_c(x, y)$ denota a intensidade do canal de cor $c \in \{R, G, B\}$ no pixel (x, y), e $I_l(x, y)$ denota o componente luminância (a componente Y no espaço de cor YCbCr), a redução da luminância é dada por:

$$I'_c(x, y) = \frac{I_c(x, y)}{I_c(x, y) + (b - a)I_l(x, y) + a}, \tag{3.1}$$

onde a e b são parâmetros que controlam o peso da luminância na lei da correção, e $c \in \{R, G, B\}$. Lembremos que os autores não detalham como chegaram aos valores de $a = 0.4$ e $b = 0.8$, também assumidos neste trabalho por apresentarem um bom resultado. Depois, calcula-se a componente de correção da luminância I'_c, como a pseudo matiz $h(x, y)$, que é computada através da seguinte equação.

$$h(x, y) = \frac{I'_R(x, y)}{I'_G(x, y) + I'_R(x, y)}. \tag{3.2}$$

Ainda seguindo a mesma abordagem, calcula-se a pseudo matiz normalizada

$$k(x, y) = \frac{h(x, y) - \min\{h(x, y)\}}{\max\{h(x, y)\} - \min\{h(x, y)\}}, \tag{3.3}$$

e traça-se uma parábola para cada ponto (x, y) que passam por três pontos de controle:

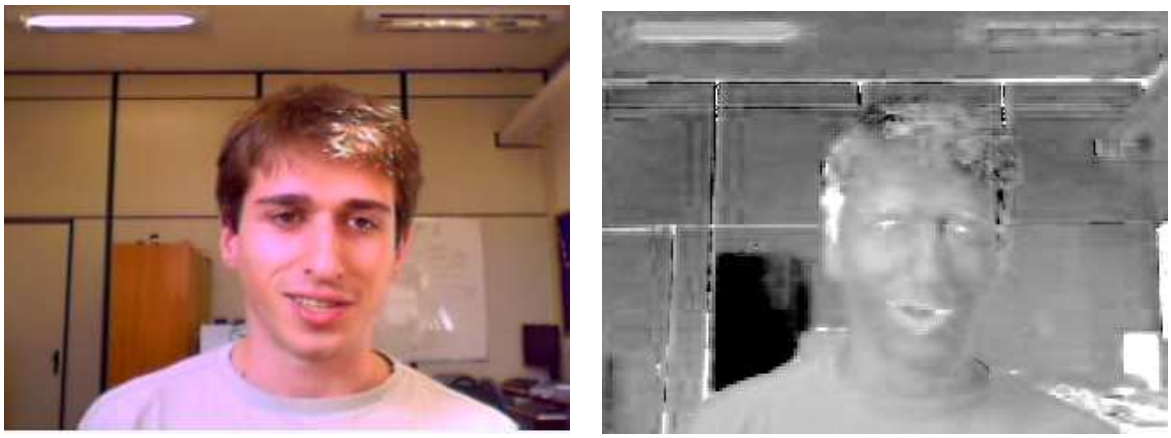
$$\begin{aligned} \boldsymbol{P}_1 &= (-\alpha k(x, y), I'_B(x, y) + \beta k(x, y)), \\ \boldsymbol{P}_2 &= (0, I'_G(x, y)), \\ \boldsymbol{P}_3 &= (1, \gamma k(x, y)), \end{aligned} \tag{3.4}$$

onde $\alpha = 0.4$, $\beta = 0.4$ e $\gamma = 2$ são parâmetros definidos pelos autores. A curvatura $c(x, y)$ dessa parábola é dada por

$$c(x, y) = \frac{I'_B(x, y) + \beta k(x, y)}{\alpha k(x, y)(\alpha k(x, y) + 1)} - \frac{I'_G(x, y)}{\alpha k(x, y)} + \frac{\gamma k(x, y)}{\alpha k(x, y) + 1}, \tag{3.5}$$

e apresenta valores mais baixos para pixeis da pele, e valores mais elevados para os pixeis relacionados com os lábios, conforme Figuras 21(b) e 22(b).

Entretanto, não há nenhuma menção dos autores em como ajustar um limiar nos valores de $c(x, y)$ para discriminar a pele e os lábios. De fato, tal limiar pode variar de imagem para imagem (e mesmo em pontos diferentes da mesma imagem), dependendo da condição de iluminação. Além disso, em situações com a boca aberta, os dentes são geralmente visíveis, e devem ser igualmente discriminados dos pixeis relacionados com a pele.



(a) (b)

Figura 21: Imagem sem barba: (a) original, (b) realçada

.

(a) (b)

Figura 22: Imagem com barba: (a) original, (b)realçada

.

Pode-se observar nas Figuras 21(b) e 22(b) como essa técnica realça a região da boca e olhos, também pode-se observar a variação que existe no realce da região da boca e olhos nessas duas imagens em função da iluminação e pessoa detectada.

Neste trabalho, duas regiões retangulares (fragmentos) são colocadas (uma abaixo de cada olho) na imagem, baseadas na face detectada e nas medidas antropométricas fornecidas por (FARKAS, 1995) *apud* (RURAINSKY; EISERT, 2003), representando as regiões que contêm pixeis de pele (as regiões logo abaixo dos olhos são menos prováveis conter qualquer tipo do pêlo facial). Mais especificamente, se (x_c, y_c)[2] é o centro da face e R é o

[2]O ponto (x, y) é dado pelas coordenadas da imagem, isto é, a origem está no ponto superior esquerdo, x que cresce de cima para baixo, e y que cresce da esquerda para a direita.

raio detectado, os pontos que definem os dois fragmentos abaixo dos olhos são dados por:

$$\begin{aligned} \boldsymbol{P}_{tl}^{l} &= (x_c, y_c - 0.3R), \quad \boldsymbol{P}_{br}^{l} = (x_c + 0.25R, y_c - 0.15R), \\ \boldsymbol{P}_{tl}^{r} &= (x_c, y_c + 0.15R), \quad \boldsymbol{P}_{br}^{r} = (x_c + 0.25R, y_c + 0.3R), \end{aligned} \tag{3.6}$$

Onde $\boldsymbol{P}_{tl}^{l}$ é o ponto superior esquerdo do fragmento esquerdo, e $\boldsymbol{P}_{br}^{l}$ é o ponto inferior direito do mesmo retângulo. Os pontos $\boldsymbol{P}_{tl}^{r}$ e $\boldsymbol{P}_{br}^{r}$ representam os pontos correspondentes do fragmento direito. Para os dois fragmentos, os valores da média μ_l (esquerda), μ_r (direita) e o desvio padrão σ_l (esquerdo), σ_r (direito) da curvatura $c(x, y)$, são calculados, e assume-se que a distribuição de valores da curvatura para pixeis relativos a pele é unimodal. De acordo com a desigualdade de Vysochanskij-Petunin (VYSOCHANSKIJ; PETUNIN, 1980), se a variável v apresenta uma distribuição unimodal com média μ e desvio padrão σ, então

$$\mathrm{P}\left(\frac{|v - \mu|}{\sigma} \geq k\right) \leq \frac{4}{9k^2}. \tag{3.7}$$

Baseado nesta desigualdade, calcula-se

$$\begin{aligned} d_l(x, y) &= \tfrac{|c(x,y) - \mu_l|}{\sigma_l}, \\ d_r(x, y) &= \tfrac{|c(x,y) - \mu_r|}{\sigma_r}, \end{aligned} \tag{3.8}$$

para testar o valor da curvatura $c(x, y)$ em qualquer pixel. Para os pixeis relativos à pele, tanto $d_l(x, y)$ como $d_r(x, y)$ devem ser baixos, dependendo da incidência da iluminação nesse pixel específico. Na nossa proposta, o pixel (x, y) é relacionado com a pele se:

$$\min\{d_l(x, y), d_r(x, y)\} \leq k, \tag{3.9}$$

e não é considerado pele caso contrário. Certamente, se k aumenta, o número de pixeis da não-pele classificados como pele aumentará. Por outro lado, os valores menores para k tendem a classificar os pixeis da pele como não-pele. Foi definido experimentalmente $k = 6.5$, que deve acomodar mais de 98% da distribuição de acordo com a inequação (3.7).

3.1.1 Atualização do modelo em tempo real

O modelo da pele é baseado nos parâmetros estatísticos (média e desvio padrão) que são computados dentro dos fragmentos da região abaixo dos olhos. Entretanto, estes parâmetros podem variar no tempo, devido aos movimentos da cabeça e/ou à mudança nas condições da iluminação. Denota-se a média μ_1 e o desvio padrão σ_1 para um fragmento dado (direito ou esquerdo) computados no quadro precedente, μ_2 e σ_2 denotam os mesmos parâmetros computados usando o quadro atual. A nova média μ e o desvio padrão σ que

considera a informação do quadro passado e do presente são dados por

$$\begin{aligned} \mu &= (1-w)\mu_1 + w\mu_2, \\ \sigma &= \sqrt{(1-w)\left(\sigma_1^2 + \mu_1^2\right) + w\left(\sigma_2^2 + \mu_2^2\right) - \mu^2}, \end{aligned} \tag{3.10}$$

onde $0 \leq w \leq 1$ é a taxa da atualização. Pode-se mostrar após alguma manipulação algébrica que se μ_1, σ_1 são computados usando um conjunto de N amostras, e μ_2, σ_2 são computados usando um conjunto de M amostras, então a média e o desvio padrão da união de ambos os conjuntos é obtido usando $w = M/(M+N)$ e $1 - w = N/(M+N)$.

Neste trabalho, μ e σ são inicializados com a média e o desvio padrão dos fragmentos no primeiro quadro, e atualizados em cada quadro ajustando $\mu_1 = \mu$ e $\sigma_1 = \sigma$ na equação (3.10), e usando μ_2, σ_2 como os parâmetros estatísticos computados no quadro atual. A taxa de atualização w foi ajustada experimental para $w = 0.25$. Deve-se igualmente observar que a atualização é feita para os fragmentos direito e esquerdo independentemente.

3.2 Identificação dos pixeis da boca

Na identificação da região dos lábios será utilizada a morfologia matemática, ramo do processamento e análise não linear de imagens que nos permite processar imagens com o objetivo de realçar, segmentar e detectar bordas, entre várias outras aplicações. Segundo Efford (EFFORD, 2000), na morfologia matemática, a erosão e a dilatação são a base para a maioria das operações que podemos utilizar no realce de imagem.

Matematicamente, erosão é definida por:

$$(A \ominus B) = \left[x : B_x \subset A\right], \tag{3.11}$$

onde A é a imagem original, B é o elemento estruturante e $A \ominus B$ consiste em todos os pontos x para os quais a translação de B por x encaixa no interior de A. Já a dilatação de uma imagem A e elemento estruturante B, será $A \oplus B$:

$$(A \oplus B) = \left[x \in A : B_x \cap x \neq \oslash\right], \tag{3.12}$$

A abertura consiste em aplicar uma erosão seguida de dilatação:

$$(A \circ B) = \left[(A \ominus B) \oplus B\right], \tag{3.13}$$

Enquanto que o fechamento consiste em aplicar uma dilatação seguida de uma erosão:

$$(A \bullet B) = [(A \oplus B) \ominus B], \tag{3.14}$$

O teste dado pela condição (3.9) é usado para identificar os pixeis relacionados à boca. Para essa finalidade, uma região retangular de interesse (ROI) em torno da boca, igualmente obtida através de uma proporcionalidade dessa região com relação à face (FARKAS, 1995) *apud* (RURAINSKY; EISERT, 2003), é computada baseando-se na posição e na escala da face detectada. Tal região é definida por

$$\boldsymbol{P}_{tl}^{m} = (x_c + 0.17R, y_c - 0.2R), \quad \boldsymbol{P}_{br}^{m} = (x_c + 0.45R, y_c + 0.2R), \tag{3.15}$$

onde $\boldsymbol{P}_{tl}^{m}$ e $\boldsymbol{P}_{br}^{m}$ denotam, respectivamente, o ponto superior esquerdo e o inferior direito da região de interesse (boca) retangular. Dentro dessa região, constrói-se uma imagem binária $B(x, y)$, atribuindo o valor 0 para os pixeis que satisfazem a condição (3.9), e 1 para os pixeis remanescentes.

Muito embora este procedimento forneça a discriminação da boca/pele, o ruído da imagem pode produzir pixeis isolados da não-pele e/ou furos na região da boca (desde que o parâmetro $k = 6$ foi escolhido para classificar uma quantidade menor de pixeis como pele). Para remover estes problemas, uma seqüência de operadores morfológicos é aplicada na região da boca. De fato, um operador de abertura é inicialmente aplicado para expandir os pixeis de pele e depois um de fechamento para remover pequenos buracos dentro da boca, visando eliminar o ruído. Desde que a seleção de $k = 6$ tende a produzir alguns furos na região da boca, o elemento estruturante para o fechamento é maior do que o utilizado para a abertura. Mais precisamente, o operador de fechamento é aplicado utilizando o elemento estruturante s $(1 + 2[W/10]) \times (1 + 2[H/25])$ (largura $\times$ altura), e de abertura com $(1 + 2[W/20]) \times (1 + 2[H/50])$.

A região da boca é demarcada por $W = 0.4R$ na largura e $H = 0.28R$ na altura, sendo essa a ROI da boca, e $[\cdot]$ denota o arredondamento para o inteiro mais próximo. Assim $B_m(x, y)$ denota a imagem binária após o processamento morfológico.

A Figura 23 ilustra uma imagem típica em uma aplicação da videoconferência. A face detectada e os fragmentos (retângulos) abaixo dos olhos usados para obter a distribuição de pixeis de pele na condição de silêncio são mostrados na Figura 23(a), junto com a região de interesse da boca (ROI). A imagem binária que contém os pixeis da não-pele é ilustrada na Figura 23(b) antes de aplicar os operadores morfológicos, e o resultado final é mostrado na Figura 23(c). Já a distribuição de pixeis de pele na condição de fala

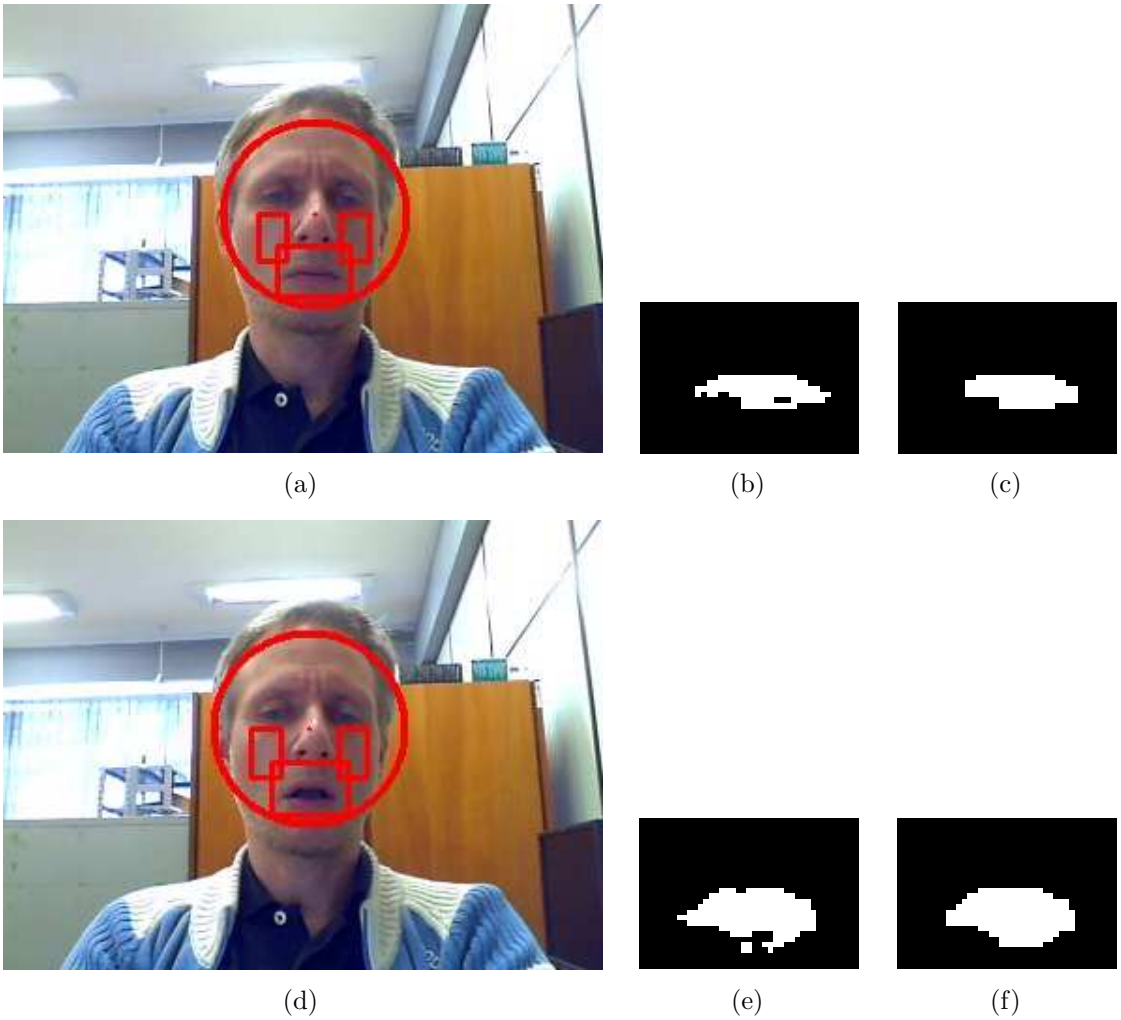


Figura 23: Posição dos fragmentos em uma face detectada (a), boca fechada binarizada antes (b) e depois (c) do processamento morfológico. Posição dos fragmentos em uma face detectada (d), boca aberta binarizada antes (e) e depois (f) do processamento morfológico.

é mostrada na Figura 23(d), junto com a região de interesse da boca (ROI). A imagem binária que contém os pixeis da não-pele é ilustrada na Figura 23(e) antes de aplicar os operadores morfológicos, e o resultado final é mostrado na Figura 23(f).

Claramente, a imagem binária $B_m(x, y)$ fornece uma estimativa da abertura da boca, desde que uma quantidade maior de pixeis da não-pele é esperada quando a boca for aberta. De fato, a medida da abertura proposta neste trabalho é o número médio de pixeis diferentes de pele na região da boca (ROI):

$$o = \frac{1}{HW} \sum_{(x,y)\in \mathrm{ROI}} B_m(x, y). \tag{3.16}$$

3.3 Detecção de Atividade de Voz

Como na maioria dos algoritmos baseados em vídeo para detecção de atividade de voz (VAD), a idéia principal deste trabalho é a determinação se há uma quantidade razoável de movimento da boca/lábio dentro de uma janela temporal. Considerando $o(t)$ a medida de abertura da boca no quadro t, a proposta deste trabalho é calcular a média $\mu(t)$ e o desvio padrão $\sigma(t)$ para uma janela temporal de T_w quadros conforme:

$$\begin{aligned} \mu(t) &= \frac{1}{T_w} \sum_{k=t-T_w+1}^{t} o(k), \\ \sigma(t) &= \sqrt{\frac{1}{T_w} \sum_{k=t-T_w+1}^{t} (o(k) - \mu(t))^2}, \end{aligned} \tag{3.17}$$

Espera-se que $\mu(t)$ e o $\sigma(t)$ apresentem valores maiores em janelas de tempo relacionadas à fala devido à variação existente na região da boca. Um teste inicial para VAD é:

$$\mu_t > k_1 \times \mu_B, \sigma_t > k_2 \times \sigma_B, \tag{3.18}$$

onde μ_B e σ_B são a média e o desvio padrão de $o(t)$ na região da boca em um período de treinamento sem fala. Os valores atribuídos para $k_1 \geq 1$ e $k_2 \geq 1$ para determinar se existe ou não atividade de voz foram obtidos empiricamente através de diversos testes realizados. Os valores $k_1 = 1.2$ e $k_2 = 1.5$ apresentaram um bom resultado, parecendo ser adequados.

Durante toda a pesquisa realizamos uma grande série de testes. No início do desenvolvimento deste trabalho, obtivemos bons resultados atribuindo os valores $k_1 = 1.2$ e $k_2 = 1.5$ na Equação 3.18. Porém, apesar de apresentarem um bom resultado, este identificador não era robusto, pois estes parâmetros poderiam ser mais adequados para um indivíduo que para outro.

Com o objetivo de deixar o modelo mais robusto, analisamos diversos quadros de uma mesma pessoa na condição de silêncio e fala e obtivemos o resultado apresentado na Figura 24 que nos mostra a distribuição da média (μ) e desvio padrão (σ) usando uma janela de tempo de 10 quadros.

Através de um longo período de fala e silêncio, temos os círculos vermelhos que representam a distribuição do par (μ, σ) representando a fala, e as marcas azuis representando o período de silêncio.

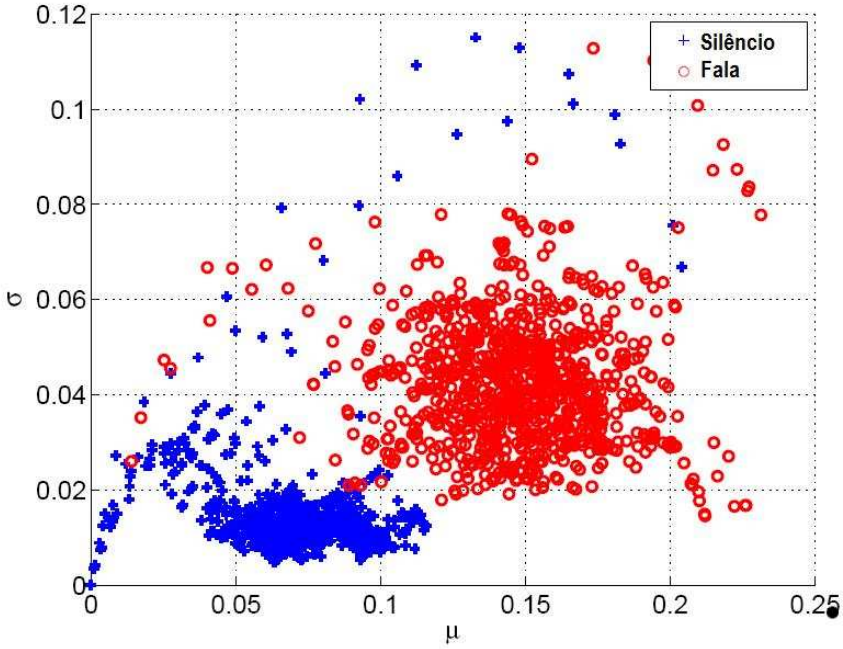


Figura 24: Diagrama de dispersão (μ, σ) para fala (círculos vermelhos) e silêncio (+ azul) em uma seqüencia de vídeo (SCOTT et al., 2009).

Pensando na detecção de atividade de voz (VAD) como um problema de classificação (fala e silêncio), a Figura 24 indica claramente que podemos utilizar o par (μ, σ) para discriminar as duas classes. Sendo assim, nos parece adequado a utilização de uma fronteira de decisão linear.

Para obter uma fronteira de decisão, podemos aplicar uma análise discriminante linear de Fisher (FLDA) para obter o melhor eixo de projeção. O eixo de projeção caracterizado pelo vetor $\boldsymbol{w} = (w_1, w_2)^T$, e a projeção do vetor é dada por

$$y = w_1\mu + w_2\sigma, \tag{3.19}$$

resultando uma variável escalar y. O histograma de y para silêncio e fala é mostrado na Figura 25, onde podemos observar que as duas classes apresentam uma boa separação.

Para determinar um atributo para identificar a atividade de voz partindo da distribuição do histograma mostrado na Figura 25, foram avaliados modelos que possam representar a função densidade de probabilidade de cada classe (embora a Figura 25 indique que uma Gaussiana pode ser adequada).

$$\begin{aligned} p_{\text{si}}(y) &= \frac{1}{\sigma_{\text{si}}\sqrt{2\pi}} \exp\left(\frac{(y-\mu_{\text{si}})^2}{2\sigma_{\text{si}}^2}\right), \\ p_{\text{sp}}(y) &= \frac{1}{\sigma_{\text{sp}}\sqrt{2\pi}} \exp\left(\frac{(y-\mu_{\text{sp}})^2}{2\sigma_{\text{sp}}^2}\right), \end{aligned} \tag{3.20}$$

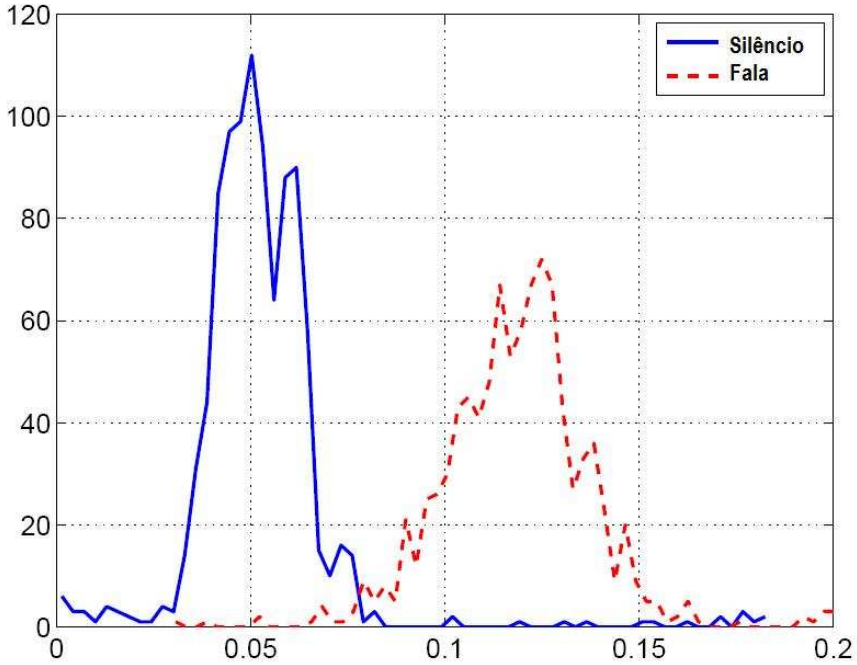


Figura 25: Histograma da transformada FLDA para silêncio e fala (SCOTT et al., 2009).

Onde $p_{\text{si}}(y)$ e $p_{\text{sp}}(y)$ são as funções de densidade de probabilidade de silêncio e fala respectivamente. Os parâmetros μ_{si}, σ_{si}, μ_{sp}, σ_{sp} correspondem à média e desvio padrão dessas distribuições.

Sendo assim, o problema de VAD pode ser formulado com a probabilidade posterior de uma amostra y sendo relacionada com a distribuição de silêncio e fala, que pode ser computada utilizando a regra de Bayes:

$$P(\text{sp}|y) = \frac{P(\text{sp})p_{\text{sp}}(y)}{P(\text{sp})p_{\text{sp}}(y) + P(\text{si})p_{\text{si}}(y)}, \tag{3.21}$$

onde $P(\text{sp})$ e $P(\text{si})$ são a priori a probabilidade da distribuição de fala e silêncio, sendo que as duas classes são consideradas com a mesma probabilidade, então $P(\text{sp}) = P(\text{si}) = 0.5$. Finalmente detectamos a atividade de voz se

$$P(\text{sp}|y) \geq 0.5. \tag{3.22}$$

Para a obtenção dos parâmetros μ_{si}, σ_{si}, μ_{sp}, σ_{sp} requeridos na Equação 3.20, deve-se realizar um treinamento nas duas condições, fala e silêncio. Para treinar a classe de silêncio durante um período em uma aplicação é relativamente fácil, basta solicitar ao indivíduo que permaneça em silêncio durante um certo número de quadros no início da aplicação. Por outro lado, forçar a fala somente para realizar o treinamento pode não ser

simples em algumas situações.

Uma alternativa para esses casos pode ser conseguida pela Equação 3.23 utilizando-se somente a distribuição relacionada ao silêncio $p_{\mathrm{si}}(y)$, e detectar a fala se a amostra cair fora da distribuição. De fato, um teste alternativo para relacionar a amostra y à fala é

$$\mathrm{y} - \mu_{\mathrm{si}} > k_s \sigma_{\mathrm{si}}, \tag{3.23}$$

por exemplo, a amostra da classe de silêncio é rejeitada quando ela difere mais que k_s desvio padrão da média. Este teste é simples de implementar mas não têm uma precisão tão boa comparado ao treinamento de amostra das duas classes, silêncio e fala (a comparação entre as duas abordagens será detalhada no próximo capítulo). Valores menores para k_s tendem a aumentar o número de classificação da amostra para a classe de fala, o oposto ocorre se aumentarmos o valor de k_s. Experimentalmente definimos $k_s = 5$ e utilizamos esse valor para os testes que serão apresentados.

Também, quando utilizamos somente amostras de silêncio para treinar o classificador, não é possível obter um ótimo eixo de projeção $(w_1, w_2)^T$ utilizando FLDA, desde que não haja nenhuma amostra para a classe de fala. Os testes executados utilizando diferentes indivíduos indicaram que a distribuição de (μ, σ) para as classes de fala e de silêncio são similares ao apresentado no diagrama de dispersão da Figura 24. Portanto, quando a fala não é utilizada para o período de treinamento, o mesmo eixo de projeção $w_1 = 0.5657$, $w_2 = 0.8246$ foi utilizado em todos os experimentos, obtendo os dados apresentados na Figura 24.

Nos testes das duas abordagens de aplicação de VAD, foi necessário estimar a média e desvio padrão em um (ou dois) parâmetros de treinamento. Em particular, a versão que se utiliza apenas do treinamento da classe de silêncio foi afetada significamente com a presença de *outliers* no ajuste do treinamento, uma vez que o desvio padrão pode ser sobrestimado (e diversas amostras de fala podem ser classificadas erroneamente como silêncio durante a fase de teste). Para atenuar o efeito dos possíveis *outliers* na versão que emprega somente o silêncio [3], o valor da média é estimado através da mediana e o desvio padrão utilizando o desvio absoluto mediano (MAD) (HUBER, 1981):

$$\begin{aligned} \mu &\approx \lambda = \mathrm{median}\,\{y_i\} \\ \sigma &\approx 1.4826 \mathrm{median}\,\{|y_i - \lambda|\} \end{aligned} \tag{3.24}$$

[3]Na versão que treina o silêncio e a fala, nossos testes indicaram que as estimativas tradicionais para o desvio padrão e média apresentaram uma performance melhor que a mediana e MAD.

Para realizar a avaliação quantitativa do modelo proposto, utilizamos uma série de vídeos que foram configurados manualmente nas duas classes (fala e não fala) para posterior comparação com o modelo de VAD proposto. Alguns resultados são apresentados no próximo capítulo.

4 Resultados

O modelo proposto necessita de treinamento para poder gerar os parâmetros que irão basear a identificação de atividade da região da boca ou não. Foram realizados dois testes distintos, o primeiro realizando o treinamento somente da condição de silêncio e o segundo efetuando o treinamento nas duas condições, fala e silêncio.

Os resultados foram avaliados através de inspeção visual verificando os quadros que foram corretamente identificados com atividade de voz ou silêncio. Podemos verificar nas Figuras 26 e 27 as regiões abaixo dos olhos de onde se extraem os parâmetros para análise comparativa com a região da boca. Neste caso, efetuamos o preenchimento na cor azul dessa região com os pixeis mais próximos à cor de pele e verde para os pixeis que mais se distanciaram. Sendo assim, podemos verificar que já existe um pequeno realce da região labial.

Como podemos observar na Figura 28 a aplicação dos operadores morfológicos auxiliam na definição da região da boca. Por outro lado, notamos na Figura 29 que quando a região da boca é pouco demarcada, ela desaparece após o tratamento morfológico. Daí a importância na definição do elemento estruturante.

4.1 Treinamento manual silêncio

O treinamento utilizando somente o vídeo na condição de silêncio já apresentou um bom resultado na avaliação de atividade de voz, conforme podemos constatar no quadro comparativo entre a proposta 1 (treinamento somente da condição de silêncio) e proposta 2 (treinamento na condição de silêncio e fala).

É interessante notar a importância de se utilizar uma estimativa robusta para a média e desvio padrão quando aplicamos a proposta 1. Por exemplo, o período de treinamento da classe de silêncio para a seqüência 1 conteve alguns *outliers*, que causaram uma sobrestimação do desvio padrão, utilizando-se uma estimativa tradicional (conduziu o PH

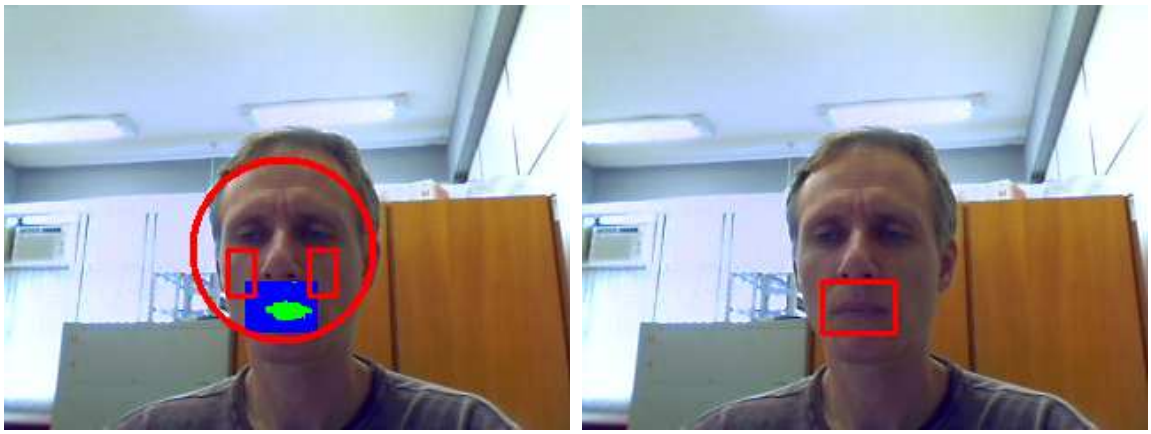


Figura 26: Fragmentos da região abaixo dos olhos e região da boca em silêncio (sem barba).

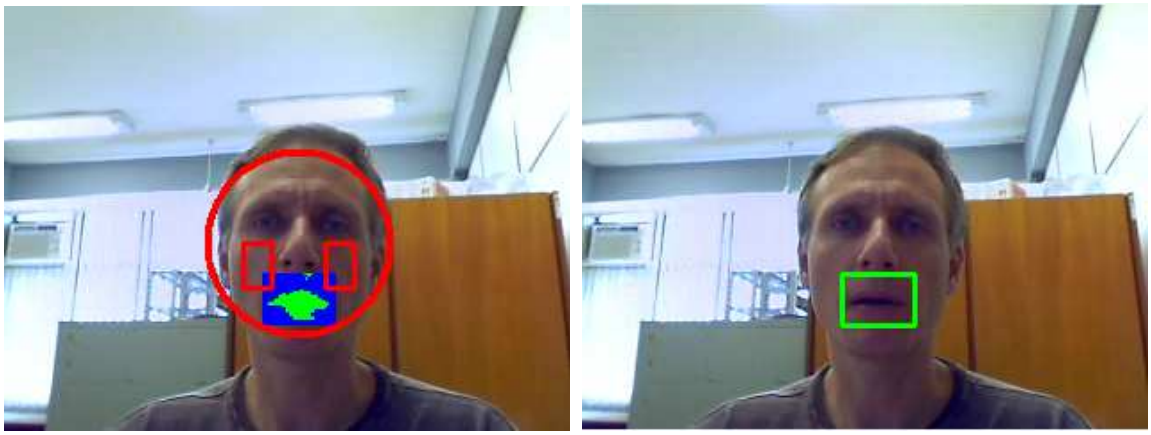


Figura 27: Fragmentos da região abaixo dos olhos e região da boca com fala (sem barba).

= 57.65%, que é muito próximo à classificar todas as amostras como silêncio). Usando uma estimativa robusta, esta taxa aumentou em torno de 80%, segundo as indicações da Tabela 3.

4.2 Treinamento manual silêncio e fala

O treinamento nas duas condições (fala e silêncio), proposta 2, requer do usuário duas etapas distintas de treinamento na seqüência de vídeo. Obviamente, neste modelo temos uma estimativa do classificador mais precisa da fronteira entre fala e silêncio. De fato, já era esperado um resultado melhor na proposta 2 que utiliza um treinamento para o modelo de silêncio e fala para treinar o classificador, enquanto que na proposta 1 só é realizado o treinamento na condição de silêncio. Contudo, a diferença de acerto entre as duas abordagens sempre foi abaixo de 10% em todas as sequências de vídeo, indicando que

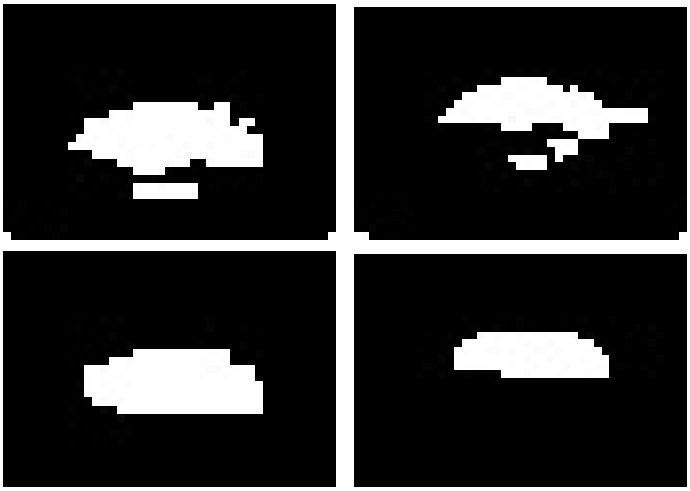


Figura 28: Seqüência do tratamento morfológico na região da boca bem demarcada.

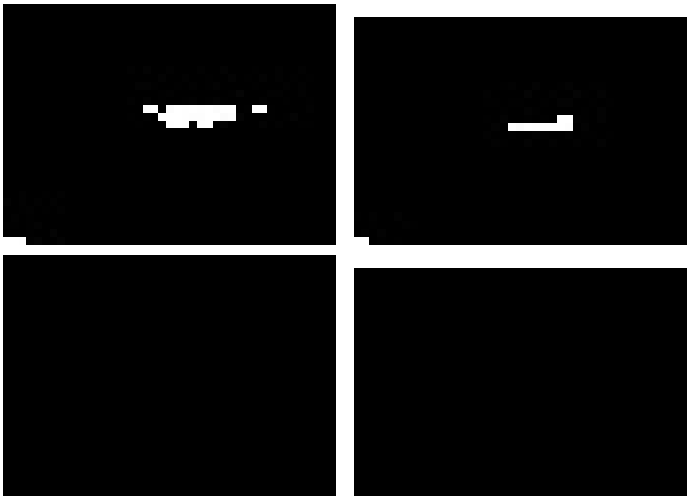


Figura 29: Seqüência do tratamento morfológico na região da boca pouco demarcada.

pode-se utilizar a abordagem 1 quando a solução não possibilitar ou não desejar realizar o treinamento de fala.

Para que possamos validar visualmente o grau de acerto do algoritmo, definimos que o retângulo que demarca a região da boca fosse mostrado em verde, Figura 27 e 31 quando a fala for detectada e em vermelho quando não houvesse uma grande variação conforme verificamos na Figura 26 e 30.

Figura 30: Fragmentos da região abaixo dos olhos e região da boca em silêncio (com barba).

Figura 31: Fragmentos da região abaixo dos olhos e região da boca com fala (com barba).

Como mostrado nas Figuras 30 e 31, podemos verificar que o algoritmo conseguiu compensar a questão da iluminação diferente para cada lado da face identificada.

Apesar de conseguir um bom resultado, existem algumas situações em que o modelo proposto tende a falhar. Isso ocorre quando não é possível identificar a face devido a uma inclinação ou rotação da mesma conforme podemos verificar na Figura 32(a). Uma falha também pode ocorrer quando a cor da pele se aproxima muito da cor do lábio conforme

(a) (b)

Figura 32: Situações em que o modelo proposto tende a falhar. (a) Inclinação da face. (b) Cor dos lábios muito próxima a cor da pele.

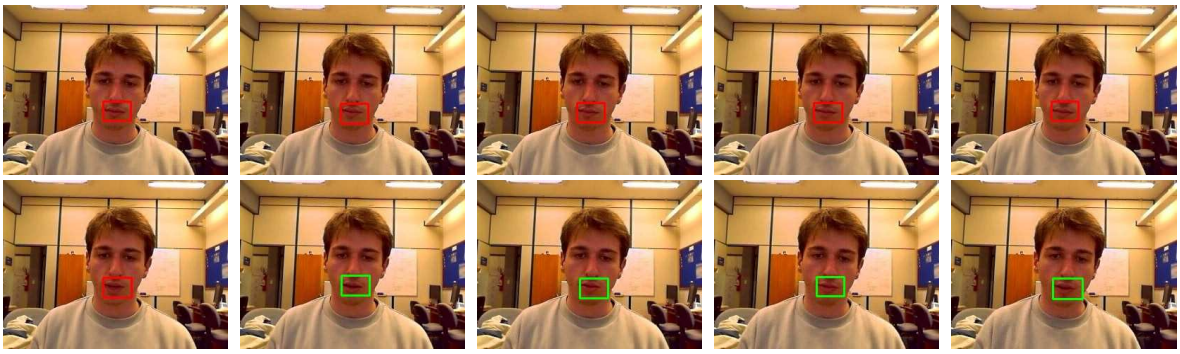


Figura 33: Seqüência 1 de vídeo.

exemplo na Figura 32(b).

Para ilustrar melhor o modelo proposto, apresentamos três seqüências de vídeo com a transição entre silêncio e fala nas Figuras 33, 34 e 35. Podemos verificar que entre a Figura 33 e a Figura 34 existe uma grande variação na iluminação, e mesmo assim o sistema identificou movimento labial. Já na Figura 35 verificamos que a utilização dos fragmentos abaixo dos olhos ajudaram a compensar a iluminação diferenciada entre os dois lados da face processada. Ainda na Figura 35, podemos verificar que esse modelo também apresenta um bom resultado para as pessoas com barba.

As Figuras 33, 34 e 35 ilustram alguns quadros das três seqüências vídeo analisadas neste trabalho. O retângulo em torno da boca é vermelho quando o silêncio é detectado, e altera-se para verde quando a fala é detectada. Em particular, uma pessoa com barba foi utilizada na seqüência 3, e os resultados de VAD giraram ao redor 89% para a abordagem 1 e a abordagem 2, conforme apresentado na Tabela 3.

Figura 34: Seqüência 2 de vídeo.

Figura 35: Seqüência 3 de vídeo.

Para avaliar os resultados desta proposta foram utilizadas três seqüências vídeo que foram classificadas manualmente para obter períodos de fala e de silêncio. As duas aproximações descritas neste trabalho foram apresentadas em dezembro de 2009 no 11th IEEE International Symposium on Multimedia em San Diego (SCOTT et al., 2009). As duas propostas, a primeira realiza apenas o treinamento do modelo de silêncio, nomeada proposta 1, e a segunda realiza o treinamento para um período de silêncio e outro para fala, nomeada proposta 2 foram analisadas, e comparadas quantitativamente nos termos da exatidão da detecção.

A Tabela 1 mostra algumas especificações dos vídeos, tais como o tamanho, o número de quadros por segundo *"frame rate"* e o número de amostras identificadas manualmente como fala e silêncio.

Tabela 1: Especificação dos vídeos

Vídeo	Tempo	*Frame Rate*	Exemplos		
			Fala	Silêncio	Total
1	$\approx$ 1 min	15 FPS	519	357	876
2	$\approx$ 1.5 min	14 FPS	500	604	1104
3	$\approx$ 40 sec	14 FPS	290	268	558

Tabela 2: Matriz de confusão das seqüências de vídeo analisadas

Vídeo	Classe	Exemplos	Abordagem 1		Abordagem 2	
			Detecção de		Detecção de	
			Fala	Silêncio	Fala	Silêncio
1	Fala	519	348	171	486	33
	Silêncio	357	7	350	25	332
2	Fala	500	475	25	474	26
	Silêncio	604	165	439	81	523
3	Fala	290	289	1	289	1
	Silêncio	268	57	211	53	215

Tabela 3: Percentual de acerto

Vídeo	Abordagem 1		Abordagem 2	
	PH	NPH	PH	NPH
1	79.79	82.64	93.38	93.32
2	82.79	83.84	90.31	90.69
3	89.61	89.19	90.32	89.94

A Tabela 2 apresenta uma matriz de confusão das três seqüências de vídeo, onde pode-se verificar o número de falsos positivos (quadro detectado como fala quando na realidade era silêncio) e falso negativos (quadros detectados como silêncio quando na realidade existia atividade labial, fala).

Na tabela 3 pode-se verificar o percentual de acerto (PH) e também o percentual de acerto normalizado (NPH). PH apresenta o percentual de acerto não levando em conta o tamanho da amostra em cada classe, enquanto que NPH considera na contagem o número de exemplos de cada classe. Mais especificamente, PH e NPH são dados por

$$\mathrm{PH} = 100\frac{T_p + T_n}{N_p + N_n}, \tag{4.1}$$

$$\mathrm{NPH} = \frac{100}{2}\left(\frac{T_p}{N_p} + \frac{T_n}{N_m}\right), \tag{4.2}$$

onde T_p e T_n são o número de resultados positivos e negativos, respectivamente, N_p e N_n são o número de exemplos positivos e negativos em cada seqüência de vídeo, respectivamente.

Conforme podemos observar nas Tabelas 2 e 3, as duas abordagens apresentam um grau de acerto acima de 80%, e a proposta 2 apresenta resultados melhores que a proposta 1 em todas as sequências de vídeo testadas.

Os resultados apresentados neste trabalho foram obtidos processando o algoritmo implementado em C++ e executado em um computador PC quad core 2.8GHz rodando

sob Windows XP. A média do tempo de execução da análise da seqüência de vídeo, desconsiderando a operação de I/O, foi de 33.8 ms por quadro, além dos 24.2 ms (72%) utilizados pelo detector de face. A resolução da seqüência de vídeo foi de 320 x 240 pixeis.

4.3 Treinamento automático dos modos de silêncio e fala

Com o objetivo de facilitar a utilização do modelo proposto, realizamos uma implementação de utilização do sinal de áudio (SOHN et al., 1999) para realizar o treinamento de forma automática para a condição de fala e silêncio, conforme apresentado em nossa proposta. Porém, dado que o objetivo deste trabalho é a identificação de atividade de voz somente por vídeo e houve uma baixa considerável no rendimento na execução em tempo real devido ao tratamento do sinal de áudio para treinamento do classificador, optamos por deixar esta implementação para um trabalho futuro.

5 *Considerações Finais e Trabalhos Futuros*

5.1 Considerações Finais

Considerando que o nosso objetivo inicial era apresentar um algoritmo de identificação de atividade de voz utilizando somente sinal de vídeo, a solução apresentada se mostrou bastante satisfatória nas duas abordagens apresentadas.

De uma forma geral, o modelo proposto de VAD apresentou bons resultados. Os resultados não foram satisfatórios quando a variação do valor do pixel entre o lábio e a pele é pequena, mas acredita-se que estes problemas possam ser superados através de um refinamento do filtro de cor de pele e os resultados de uma forma geral possam ser melhorados. A utilização dos fragmentos criados abaixo dos olhos se mostraram úteis para o ajuste da questão de variação de luminosidade variada na face.

O modelo foi implementado em C++ partindo-se da identificação da face feita pelo OpenCV (INTEL...,) (Open Computer Vision Library) da Intel®, depois determinou-se a possível localização da boca através da antropometria e os fragmentos abaixo dos olhos para detectar os pixeis de pele.

O treinamento deste modelo é realizado de forma manual para se obter os atributos da boca na condição de silêncio em uma seqüência de vídeo. Este treinamento captura uma seqüência de quadros em silêncio e uma seqüência com movimentação labial (fala). Em nossa proposta, foi pensado utilizar o áudio para realizar o treinamento de forma automática, porém, quando fizemos a integração o sistema se tornou muito mais lento devido ao tratamento de mais essa informação. Como o objetivo desde o início era o de identificar atividade de voz somente tratando vídeo, optamos por não incorporar o tratamento de áudio para realização do treinamento.

Sabe-se que existem diversas variáveis que podem influenciar no resultado final. O grande desafio é definir um atributo que seja ao mesmo tempo confiável e adaptável às

diversas condições de iluminação e também às diversas variações de faces.

5.2 Trabalhos Futuros

Para que não seja necessário efetuar o treinamento de uma forma manual, vamos procurar resolver a integração do sinal de áudio no algoritmo e possibilitar a realização do treinamento de uma forma automática para situação de fala e silêncio. Com a integração da informação de áudio (SOHN et al., 1999) no modelo de VAD proposto, acredita-se que seja possível realizar o treinamento de uma forma automática.

Uma outra questão que também poderá ser abordada no futuro é tentar refinar os parâmetros utilizados na classificação de pele e não pele para diminuir o problema encontrado nos indivíduos que tenham a cor da pele muito próxima a cor dos lábios.

Referências

AOKI K. MASUDA, H. M. T. T. M.; ARIKI, Y. Voice activity detection by lip shape tracking using ebgm. In: ACM. *Proceedings of the 15th international conference on Multimedia*. Augsburg, Germany, 2007. p. 561–564.

AUBREY, A. et al. Two novel visual voice activity detectors based on appearance models and retinal filtering. In: *Proceedings of European Signal Processing Confenrence*. [S.l.: s.n.], 2007. v. 1, p. 2409–2413.

BORGSTROM, B. J.; ALWAN, A. A low-complexity parabolic lip contour model with speaker normalization for high-level feature extraction in noise-robust audiovisual speech recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, v. 38, n. 6, p. 1273–1280, November 2008.

BRANDSTEIN, M. S.; SILVERMAN, H. F. A practical methodology for speech source localization with microphone arrays. *Computer Speech & Language*, v. 11, n. 2, p. 91–126, 1997.

CHIBELUSHI, C.; DERAVI, F.; MASON, J. A review of speech-based bimodal recognition. *Multimedia, IEEE Transactions on*, v. 4, n. 1, p. 23–37, Mar 2002. ISSN 1520-9210.

DO, H.; SILVERMAN, H. F. A fast microphone array srp-phat source location implementation using coarse-to-fine region contraction (cfrc). In: *Proceedings of Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics 2007 (WASPAA 2007)*. New Paltz, USA: [s.n.], 2007. p. 295–298.

EFFORD, N. *Digital Image Processing: A Practical Introduction Using Java (with CD-ROM)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2000. ISBN 0201596237.

EVENO, N.; CAPLIER, A.; COULON, P. Accurate and quasi-automatic lip tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 14, n. 5, p. 706–715, May 2004.

EVENO, N.; CAPLIER, A.; COULON, P.-Y. New color transformation for lips segmentation. In: *Multimedia Signal Processing, 2001 IEEE Fourth Workshop on*. [S.l.: s.n.], 2001. p. 3–8.

FARKAS, L. *Anthropometry of the Head and Face*. 2nd. ed. [S.l.]: Raven Press, 1995.

GONZALEZ, R. C. e. R. E. W. *Processamento de imagens digitais*. São Paulo (BR): Edgard Blücher, 2000. 509 p.

HEISELE, T. S. B.; POGGIO, T. A component-based framework for face detection and identification. *International Journal of Computer Vision*, Netherlands, v. 74, p. 103–215, Springer 2006.

HSU, R.-L.; ABDEL-MOTTALEB, M.; JAIN, A. Face detection in color images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 24, n. 5, p. 696–706, May 2002. ISSN 0162-8828.

HUBER, P. *Robust Statistics*. New York: John Wiley and Sons, 1981.

INTEL OpenCV, 15/02/2008. Disponível em: <http://www.intel.com/technology/computing/opencv/index.htm>.

LIANG X. LIU, Y. Z. X. P. L.; NEFIAN, A. V. Speaker independent audio-visual continuous speech recognition. In: *IEEE International Conference on Multimedia and Expo*. [S.l.: s.n.], 2002.

LIU, P.; WANG, Z. Voice activity detection using visual information. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. [S.l.: s.n.], 2004. v. 1, p. I–609–12 vol.1.

LO, D.; GOUBRAN, R.; DANSEREAU, R. Multimodal talker localization in video conferencing environments. In: *Haptic, Audio and Visual Environments and Their Applications, 2004. HAVE 2004. Proceedings. The 3rd IEEE International Workshop on*. [S.l.: s.n.], 2004. p. 195–200.

MARAGOS, P.; POTAMIANOS, A.; GROS, P. *Multimodal Processing and Interaction: Audio, Video, Text*. [S.l.]: Springer-Verlag, 2008.

MARCEL JOHNNY MARIÉTHOZ, Y. R. S.; CARDINAUX, F. Bi-modal face and speech authentication: a biologin demonstration system. In: *Workshop on Multimodal User Authentication (MMUA)*. [S.l.: s.n.], 2006.

NEFIAN, A. V.; III, M. H. H. Face detection and recognition using hidden markov models. In: *IEEE International Conference Image Processing*. [S.l.: s.n.], 1998.

NEMER, E.; GOUBRAN, R.; MAHMOUD, S. Robust voice activity detection using higher-order statistics in the lpc residual domain. *Speech and Audio Processing, IEEE Transactions on*, v. 9, n. 3, p. 217–231, Mar 2001. ISSN 1063-6676.

NEMER, E.; GOUBRAN, R.; MAHMOUD, S. Robust voice activity detection using higher-order statistics in the lpc residual domain. *IEEE Transactions on Speech and Audio Processing*, v. 9, n. 3, p. 217–231, March 2005.

ONG, E.; BOWDEN, R. Robust lip-tracking using rigid flocks of selected linear predictors. In: *Proc. 8th IEEE International Cong on Automatic Face and Gesture Recognition*. [S.l.]: IEEE, 2008.

PANTIC, M.; ROTHKRANTZ, L. Facial action recognition for facial expression analysis from static face images. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, v. 34, n. 3, p. 1449–1461, June 2004. ISSN 1083-4419.

PANTIC, M.; TOMC, M.; ROTHKRANTZ, L. J. M. A hybrid approach to mouth features detection. In: IEEE. *IEEE International Conference on Systems, Man, and Cybernetics*. [S.l.], 2001.

PAPANDREOU, G. et al. Multimodal fusion and learning with uncertain features applied to audiovisual speech recognition. In: *Proc. IEEE Workshop on Multimedia Signal Processing (MMSP-2007), Chania, Greece, Oct. 2007*. [S.l.: s.n.], 2007. p. 264–267.

PAPANDREOU, G. et al. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, v. 17, n. 3, p. 423–435, March 2009. ISSN 1558-7916.

PITSIKALIS, V. et al. Adaptive multimodal fusion by uncertainty compensation. In: *Proc. Int'l Conference on Spoken Language Processing (ICSLP-06), Pittsburgh, PA, Sep. 2006*. [S.l.: s.n.], 2006. p. 2458–2461.

RURAINSKY, J.; EISERT, P. Template-based eye and mouth detection for 3d video conferencing. *Lecture Notes in Computer Science*, Berlin: Heidelberg, p. 23–31, 2003.

SCHMUGGE, S. J. et al. Objective evaluation of approaches of skin detection using roc analysis. *Computer Vision and Image Understanding*, Elsevier Science Inc., New York, NY, USA, v. 108, n. 1-2, p. 41–51, 2007. ISSN 1077-3142.

SCOTT, D. et al. Video based vad using adaptive color information. In: *ISM '09: Proceedings of the 2009 11th IEEE International Symposium on Multimedia*. San Diego, California, USA: IEEE Computer Society, 2009. p. 80–87. ISBN 978-0-7695-3890-7.

SODOYER, D. et al. An analysis of visual speech information applied to voice activity detection. In: *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. [S.l.: s.n.], 2006. v. 1, p. I–I. ISSN 1520-6149.

SODOYER, D. et al. An analysis of visual speech information applied to voice activity detection. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2006. v. 1, p. I–I.

SODOYER, D. et al. A study of lip movements during spontaneous dialog and its application to voice activity detection. *Journal of the Acoustical Society of America*, v. 125, n. 2, p. 1184–1196, February 2009.

SODOYER, D. et al. A study of lip movements during spontaneous dialog and its application to voice activity detection. *The Journal of the Acoustical Society of America*, ASA, v. 125, n. 2, p. 1184–1196, 2009. Disponível em: <http://link.aip.org/link/?JAS/125/1184/1>.

SOHN, J. et al. A statistical model-based voice activity detection. *IEEE Signal Process. Lett*, v. 6, p. 1–3, 1999.

TANYER, S.; OZER, H. Voice activity detection in nonstationary noise. *IEEE Transactions on Speech and Audio Processing*, v. 8, n. 4, p. 478–482, Jul 2000.

VIOLA, P. A.; JONES, M. J. Rapid object detection using a boosted cascade of simple features. In: . Kauai, HI: IEEE Computer Society, 2001. p. 511–518. ISBN 0-7695-1272-0.

VIOLA, P. A.; JONES, M. J. Robust real-time face detection. *International Journal of Computer Vision*, Kluwer Academic Publishers, Hingham, MA, USA, v. 57, n. 2, p. 137–154, 2004.

VYSOCHANSKIJ, D. F.; PETUNIN, Y. I. Justification of the 3σ rule for unimodal distributions. *Theory of Probability and Mathematical Statistics*, v. 21, p. 25–36, 1980.

WANG, S.; LAU, W.; LEUNG, S. Automatic lip contour extraction from color images. *Pattern Recognition*, v. 37, n. 12, p. 2375 – 2387, 2004. ISSN 0031-3203.

YANG, M.-H.; KRIEGMAN, D. J.; AHUJA, N. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, p. 34–58, 2002.