Método Supervisionado de Classificação e Identificação de Padrões de Pele em Imagens e Vídeos

Thiago Rateke[1,2], Antonio Carlos Sobieranski[1,2], Eros Comunello[1,2]

[1]LAPIX – Laboratório de Processamento de Imagens e Computação Gráfica
UFSC – Universidade Federal de Santa Catarina

[2]UNIVALI – Universidade do Vale do Itajaí

Abstract. *In this paper we present a supervised method for skin classification and identification in video sequences. The user trains the system by selecting in a frame some typical positive skin pixels that will be used as a reference for the construction of a nonlinear distance metric. In this learning process the global optimum is obtained by induction employing higher polynomial terms of the Mahalanobis distance, extracting nonlinear features of the skin pattern distributions. These nonlinear features are then used to classify the frames captured from the camera, identifying all skin and non-skin regions on the scene. We adopt an strategy which enables this method to run in real-time after some iterations. We also compare our classification method against vector norm (L^2) and Mahalanobis distance, showing a better classification.*

Resumo. *Neste artigo é apresentado um método supervisionado de classificação e identificação de padrões de pele em seqüências de vídeo. O usuário treina o sistema através da seleção de padrões típicos de pele em um determinado frame, que será utilizado como referência na construção de uma métrica de distância não-linear. Neste processo de aprendizado o ótimo global é indutivamente obtido através da distância polinomial de Mahalanobis, extraindo as características não-lineares dos padrões de pele. Estas características não-lineares são então utilizadas para classificar os frames capturados de uma câmera, identificando todos os padrões de pele e não-pele na cena. Foi adotada uma estratégia que possibilita ao método ser executado em tempo real logo após algumas iterações. O método de classificação foi comparado em relação à norma vetorial (L^2) e distância de Mahalanobis, demonstrando uma melhor classificação para os padrões de pele.*

1. Introdução

Nos últimos anos a área de Processamento Digital de Imagens (PDI) vem ganhando importância no cenário global com muitas aplicações de interatividade e entretenimento no contexto de interação humano-computador. Esta área tem agregado muitos dispositivos tecnológicos tais como câmeras de infra-vermelho e ambientes imersivos, entretanto ainda apresenta um custo elevado para aplicações do mundo real. Soluções viáveis podem ser encontradas pelo uso de câmeras aliadas com a Visão Computacional (VC) e técnicas de processamento de imagem. Câmeras de computador ou popularmente *webcams* são acessíveis em qualquer computador e o estágio atual em termos de hardware (CPU, resolução e frames por segundo – *fps*) possibilita o

desenvolvimento de aplicações efetivas com tempo de resposta muito próximo ao tempo real.

Embora a VC e as técnicas de PDI tenham alcançado um nível elevado desde então, uma aplicação muito esperada desde então é a possibilidade de interpretação gestual pelo uso de uma simples câmera. A fim de identificar algumas dessas ações, uma tarefa primária é a identificação e a segmentação dos padrões correspondentes à pele e suas diferentes variações ao longo do tempo. Uma das principais dificuldades envolvendo essa tarefa é a variação de luz em diferentes ambientes (luzes, sombras, interior e exterior de ambientes) ou mesmo com a própria movimentação da pessoa **[4]**. Outros problemas são relacionados com as características das câmeras, como os diferentes sensores e lentes utilizados. Um exemplo que pode ser apresentado é a cor reproduzida por uma câmera CCD, que possui dependência com a reflectância espectral e varia de acordo com a sensibilidade do sensor da câmera. Problemas de etnia também dificultam a elaboração de um modelo genérico, uma vez que a cor de pele de varia de pessoa para pessoa, e em diferentes grupos étnicos e regiões **[4]**.

Outro problema relacionado, que por sua vez corresponde a efetividade de um modelo matemático, diz respeito às diferentes funções de discriminação usadas para medir a real diferença entre os vetores de cores e os padrões de pele. Essa discriminação é uma função que visa quantificar cada entrada do vetor cores, dando um valor escalar correspondente à similaridade ou qual próximo de um padrão de cor de pele este vetor se encontra. Há muitas maneiras diferentes de calcular uma função de similaridade para um algoritmo de reconhecimento de tonalidades de pele, mais especificamente cores **[8]**. As funções de similaridade mais comuns são as métricas de distância baseadas na norma L^2, calculada a partir de dois vetores ($\| u - g \|$) **[7]**. Outras funções de similaridade tem o foco na geometria do espaço de cores, que desempenham um papel importante na medição de similaridade **[2]**. Para cada modelo de espaço de cores, tem-se uma métrica de distância apropriada para ser usada para aperfeiçoar a relação de percepção entre geometria e similaridade. Para muitos casos, a grande quantidade de combinações possíveis entre espaço de cores e métricas de distância pode tornar difícil a escolha de uma boa função de discriminação. Como conseqüência, alguns métodos de segmentação de imagem precisam de muitos parâmetros de entrada para compensar a limitação da função de similaridade utilizada **[5]**.

A expressividade de uma função de similaridade pode ser melhorada pelo uso de algum conhecimento prévio sobre algum objeto na imagem. Esse objeto ou "conhecimento" pode ser considerado um tipo de informação extra conforme demonstrado por **[1]**, que é denominado de restrições *pairwise*. Estas restrições são usadas como informação de entrada nos classificadores *Knn*, possibilitando criar uma superfície de discriminação nos espaços de cores (RGB, HSV, etc). Uma vez obtidas as funções de discriminação estas são usadas para classificar os vetores de cores em valores bimodais nas seguintes faixas: 1 para vetores de cores similares e 0 para não similares. Uma abordagem interessante para uso desses classificadores é apresentada em **[7].** Essa abordagem é comparada com outros métodos estado-da-arte, demonstrando uma melhor classificação para o contexto de agrupamento de dados e classificação de padrões.

No entanto, os padrões de pele podem apresentar uma variabilidade considerável no espaço de cores, tornando o processo de classificação dependente de um classificador rigoroso. Trabalhos relacionados onde a identificação ou segmentação de tonalidades de

pele podem ser encontrados na literatura: Em **[6]** um método para a segmentação de pele em tempo real é apresentado, onde o método proposto é capaz de adaptar-se à mudanças na cor de pele e até mesmo a diferentes condições de luz. Como primeiro passo o método usa um modelo de cor de pele genérico, e um detector de bordas é utilizado para localizar as fronteiras ou picos de gradiente na cena. Essas variações de gradiente são usadas como evidência qualitativa das regiões de pele na cena, no entanto necessitando de operações de morfologia matemática para correção de pequenos pontos de falha na imagem. Em **[9]** uma abordagem para a detecção de pele em imagens é apresentada. Esse método também utiliza um treinamento do modelo de cor de pele para classificar uma imagem em regiões de pele ou não. Os pixels selecionados nessa primeira classificação são então usados para criar um novo modelo intermediário, que é usado para identificar os próximos frames. Uma metodologia baseada em redes neurais (mais especificamente RCE) é apresentada em **[10]**. Esse último destina-se a segmentos de região da mão para um sistema baseado em gestos de interação homem-máquina. Nesse trabalho é apresentado também um estudo onde diferentes espaços de cores foram usados (RGB, HSI, CIE-lab espaço de cores), demonstrando que as cores da pele humana apresentam maior variação de intensidade do que os componentes de cor.

Nesse trabalho é proposto um novo método para a segmentação de cor da pele em imagens ou mesmo em seqüências de vídeo. Essa abordagem é composta por dois procedimentos: primeiro, uma etapa de treinamento é realizada onde o padrão de pele é capturado "in-situ" a partir de um determinado frame/imagem e um mapa não-linear topológico é definido. Na segunda etapa, cada vetor de cor e em cada frame é classificado nesse mapa topológico, gerando uma imagem de intensidades, correspondendo às probabilidades de tons de pele ou *background*. Os resultados obtidos demonstram que o método proposto é capaz de identificar padrões de pele de uma ou várias pessoas na cena (modelo genérico étnico), com desempenho muito próximo ao tempo real após algumas iterações do processo de classificação e geração do modelo topológico.

2. Metodologia

A abordagem proposta é composta por duas etapas, sendo: treinamento e classificação, conforme demonstrado no diagrama da Figura 1. A etapa de treinamento (figura 1 – parte superior) é um processo de aprendizado supervisionado, onde os padrões de pele correspondentes são definidos. Esses padrões de pele são então usados para o aprendizado de uma métrica de distância, necessária para criar os mapas topológicos que serão utilizados na etapa de classificação (figura 1 – parte inferior). Na subseção **2.1**, a métrica de distância usada para classificar os padrões de pele é apresentada. A etapa de treinamento para a criação dos mapas topológicos é detalhada em **2.2**. O processo de classificação e a estratégia usada para alcançar o desempenho próximo a tempo-real é discutida na subseção **2.3**.

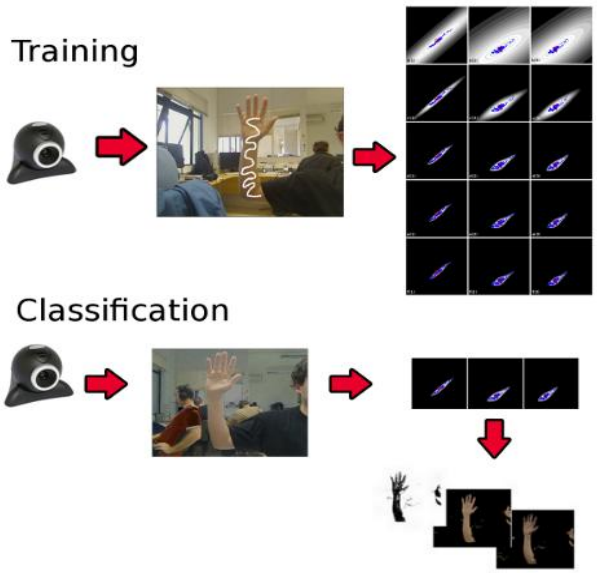

Figura 1. Diagrama geral da metodologia proposta.

2.1. Polynomial Mahalanobis

Neste trabalho algumas limitações apresentadas pelas metodologias *pairwise* podem ser corrigidas através do uso da distância polinomial de Mahalanobis **[3]**. Esta métrica de distância possui por característica a representação de padrões não-lineares utilizando para isto somente um conjunto de entrada específico na etapa de aprendizado. A discriminação não-linear é guiada por um grau polinomial (*q*-order) que determina o quão rigorosa a classificação será em relação ao padrão de entrada. Esta ordem polinomial também possibilita a eliminação de classificações falso positivos, quando altas ordens são utilizadas, aumentando a rigorosidade da classificação.

A distância polinomial pode ser calculada tomando por base a distância convencional de Mahalanobis, também conhecida por distância estatística. Esta distância estatística pondera os valores de similaridade de acordo com a variação estatística em cada componente vetorial, dada pela equação:

$$d_M(x,y) = \|x-y\|_A = \sqrt{(x-y)^T A^{-1}(x-y)},$$

onde *dm(x,y)* é a distância entre dados dois vetores de cores (***x*** e *y*) e $\mathbf{A}^{-1}$ é a matriz de covariância inversa computada a partir de uma distribuição multivariada de entrada (neste caso, os padrões de pele). Como pode ser observado, a equação acima também se reduz à norma vetorial, caso **A** seja uma matriz identidade.

O polinômio de Mahalanobis pode ser obtido diretamente pela distância de Mahalanobis, que é a primeira de ***q*** ordens polinomiais. O principal conceito do polinômio de Mahalanobis é mapear o conjunto de treinamento em termos polinomiais de ordem maior, gerando mapas polinomiais de alta ordem. Por exemplo, considere um conjunto de treinamento de padrões de pele **S: {$\mathbf{p_1}$, $\mathbf{p_2}$, $\mathbf{p_3}$, …$\mathbf{p_N}$}** sendo vetores de cores multidimensionais, onde **N** é a cardinalidade de **S**. O primeiro passo é calcular a distância de Mahalanobis entre dois vetores de cores quaisquer (*x e y*), mapeando todas

as bases ***m***-dimensionais $\mathbf{p_k} = \{\mathbf{p_{k1}}, \mathbf{p_{k2}}, \ldots, \mathbf{p_{km}}\}$ do conjunto de treinamento **S**, para cada vetor de cor **(k=1,…, N)** em todos os termos polinomiais de ordem menores ou iguais a **q**. Por exemplo, considerando um vetor 2-dimensional $\mathbf{p_k} = \{\mathbf{p_{k1}}, \mathbf{p_{k2}}\}$) ocorreria o seguinte mapeamento:

$$(p_{k_1}, p_{k_2}, p_{k_1}^2, p_{k_2}^2, p_{k_1} p_{k_2})$$

De um ponto de vista computacional, as ordens polinomiais ***q*** podem ser obtidas diretamente a partir do mapeamento do conjunto de entrada em seus respectivos termos polinomiais, e usá-los na equação convencional de Mahalanobis. No entanto, para grandes valores de dimensão ***m*** e ordem ***q*** a distância polinomial de Mahalanobis torna-se computacionalmente inviável. Uma forma para computar termos polinomiais de grande ordem é pela utilização do seguinte framework projetivo, proposto por **[3]**:

$$d_{PM}(x, y) = d_{M_{\sigma^2}}(x, y) + \sum_{l=1}^{L} d_{M_{\sigma^2}}\left(g_l^i, g_l^j\right)$$

onde o primeiro termo é a distância convencional de Mahalanobis adicionada a um pequeno valor $\boldsymbol{\sigma^2}$ para anular casos onde a matriz inversa não existir, $\mathbf{L > 0}$ é o número máximo de ordens polinomiais que podem ser geradas (onde $\mathbf{q} = \mathbf{2^L}$), e os argumentos $\mathbf{g_l}$ são as próximas projeções em termos polinomiais de **x** e **y** (sendo **i** e **j**, respectivamente) em seus termos polinomiais. A distância polinomial de Mahalanobis é detalhada no artigo original, em **[3]**.

2.2. Etapa de Treinamento e Calibração do Modelo

A etapa de treinamento é uma fase composta por dois sub-procedimentos: definição dos dados de entrada e calibração do modelo. A definição dos dados de entrada é a seleção qualitativa dos padrões de pele na cena. Esses padrões de pele – designado de conjunto ***S*** (como na seção 2.1) – serão utilizados para o aprendizado da métrica de distância. Como apresentado previamente, o padrão *L* da equação 2 precisa ser definido. Esse parâmetro controla o número de projeções no mapa polinomial que deve ser gerado para a próxima etapa (procedimento de classificação).

A calibração do modelo é uma fase onde a métrica de distância é efetivamente estabelecida, um para cada ordem ***q*** até a ordem *L*. Como a métrica PMD é um método projetivo, todas as ordens inferiores construídas *($q \leq L$)* estarão disponíveis para serem usadas sem uma nova etapa de treinamento do modelo. O usuário apenas define a ordem *q* do mapa que deve ser utilizada, desde que essa ordem ***q*** seja menor ou igual a *L*. Os *frames* podem ser classificados através de um cálculo de similaridade em cada vetor de cores e em um mapa de ordem *q* específico. Quanto maior a ordem do mapa polinomial selecionado pelo usuário, mais restritivas serão as respostas fornecidas pela classificação dos vetores de cores. Essa propriedade natural foi explorada nesta abordagem com o objetivo de reduzir a dependência de um valor de *thresholding* normalmente exigido em outras abordagens relacionadas (*pairwise*). O processo seguinte é uma etapa de normalização utilizada para melhorar a função de discriminação obtida pelo método, dada por:

$$I_{(x,y)} = e^{(-\lambda d_{PM}(x,y))}$$

onde d_{PM} é a distância polinomial de Mahalanobis previamente descrita na equação 2, e $\lambda > 0$ é um valor de parâmetro de contraste, usado para controlar o grau de restrição do mapa topológico. Como consequência, a equação acima resulta em valores de intensidade $\boldsymbol{I}$ para um plano cartesiano não-binários variando em um intervalo [0-1], onde vetores de cores não correspondentes às tonalidades de pele serão atribuídos a valores próximos de 0, e vetores similares próximos a 1.

Na figura 2 são apresentados os mapas topológicos obtidos pelo processo de normalização e uma breve comparação em relação à norma vetorial (L^2 – distância Euclideana), distância de Mahalanobis indicada na Equação 1, e a métrica PMD descrita na Equação 2. Também é apresentado o conjunto de treinamento S definido por um padrão de pele, representado pelos pontos em azuis. O resultado de cada coluna representa também o espaço de cores RGB, decomposto em nas facetas RG, RB, GB.

Essa decomposição é usada aqui apenas para fins de visualização, e todas as informações de dimensionalidade são usadas em $\boldsymbol{S}$ na fase de calibração do modelo. Na linha (a) é apresentada a métrica da distância Euclideana e ao mesmo tempo dois problemas comuns: a necessidade de um valor de *threshold* para delinear o espaço em regiões de pele e não-pele (bimodal), resultando em muitos pontos de falso positivo. Regiões brancas indicam alta similaridade e em contrapartida demostrado por regiões em escuro (zonas de baixa similaridade). É demostrado que o aumento ou relaxamento do parâmetro de *thresholding* não deve fornecer uma classificação robusta do padrão de pele, e alguns pontos de falso positivo possivelmente serão incluídos. Na linha (b) é apresentado a métrica MD (primeira ordem q) utilizando o processo de normalização descrito pela equação 1. É importante observar que o mapa topológico ainda necessita de um limiar (*threshold*) para realizar a classificação entre as regiões de pele e não-pele.

Na linha (c) da Figura 2 é mostrado o mapa polinomial de ordem 2 e suas respectivas decomposições. Nessa ordem q é possível obter uma boa discriminação entre as áreas positivas (branco) e as negativas (preto). O mapa polinomial reduz a necessidade de um valor limiar (*thresholding*), aumentando a qualidade da classificação. Na sequência de linhas (d), (e) e (f), são apresentadas a ordem 4, ordem 8 e ordem 16, respectivamente (2^2, 2^3 e 2^4). Como pode-se perceber quanto maior a ordem polinomial utilizada no mapa, maior a rigorosidade da classificação em relação ao conjunto de treinamento **S**. Para as projeções polinomiais foram utilizados $L = 5$ e parâmetro de contraste $\lambda = 1$.

2.3. Processo de Classificação

O processo de classificação é uma etapa onde cada frame é classificado nos mapas topológicos previamente definidos. Esse processo é apresentado pelo diagrama da Figura 2 – inferior, e nessa etapa o usuário seleciona a ordem polinomial $\boldsymbol{q}$ do mapa topológico que será usado na classificação. Cada pixel em um determinado frame é verificado neste mapa, e um valor escalar correspondente a sua similaridade é obtido.

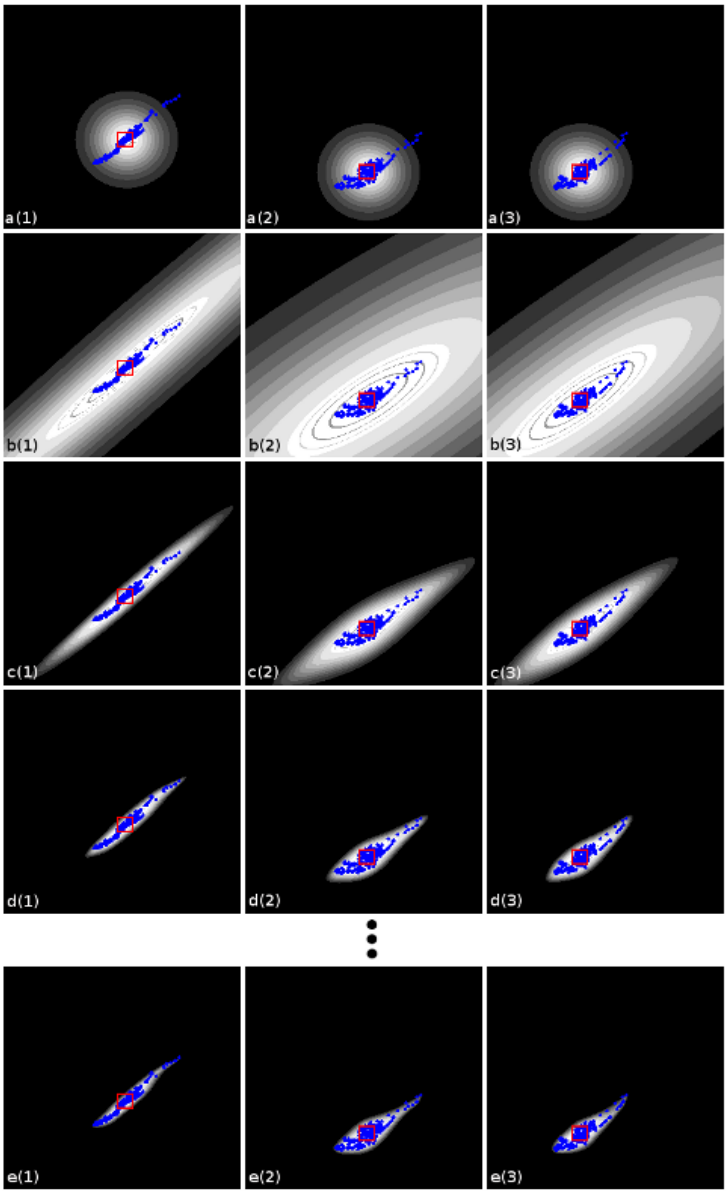

Figura 2. Mapas topológicos obtidos a partir de um determinado conjunto de treinamento de tonalidades de pele.

Esse processo, no entanto, pode-se apresentar custoso sendo que cada pixel em cada frame necessitará ser verificado, e muitas vezes repetidamente dependendo da variabilidade da imagem ou frame de entrada. A fim de tornar esse processo passível se ser executado próximo de tempo real, a seguinte estratégia foi adotada:

- Uma estrutura que possibilite representar todo o espaço RGB foi construída. Para esta estrutura 3D, outra dimensão foi adicionada com o objetivo de comportar também ***N*** ordens dimensionais do polinômio. Em outras palavras, uma estrutura 4D foi definida onde o espaço RGB existe em diferentes ordens polinomiais de tempo, onde cada coordenada *rgb* em uma dada ordem dimensional *q* possui um valor correspondente de similaridade (*tipo ponto flutuante*). Esta estrutura é denominada de **M**;

- Com a execução do programa, a estrutura é alocada em memória, ocupando aproximadamente 128 mb para cada ordem dimensional. Nesta etapa a estrutura é inicializada com valores -1, sinalizando que o índice *rgbq* ainda não foi computado no polinômio;
- Quando o processo de classificação inicia, um histograma é calculado para o primeiro *frame*, onde todos os vetores de cores não repetidos são computados no polinômio. Para cada índice *rgbq*, um escalar é retornado pelo mapa polinomial, e este valor é armazenado na estrutura **M**;
- A partir do segundo *frame*, o procedimento é calcular o histograma da imagem de entrada e cada vetor de cor é primeiramente verificado em **M**. Somente serão recalculados no mapa polinomial as coordenadas de índice *rgbq* ausentes em **M**.

Após algumas iterações e *frames* analisados, a estrutura **M** contém informação suficiente para possibilitar a execução da metodologia acima de 25 frames por segundo (dependendo da resolução da imagem e características do dispositivo de captura), ou um vídeo em execução onde somente as regiões correspondentes à pele treinada anterioremente aparecem. Com o aparecimento de novos objetos na cena ou com a movimentação da câmera da sua posição inicial somente as coordenadas *rgbq* não existentes em **M** serão computadas.

2.4. Resultados Preliminares

Experimentos preliminares demonstraram a efetividade da metodologia proposta para o contexto de identificação e segmentação de padrões de pele em imagens e cenas de vídeo. Diferentemente de outras abordagens, o ambiente de testes utilizado é um ambiente heterogêneo, composto de vários objetos sob diferentes condições de luminosidade.

Figura 3. Ambiente de testes utilizado no experimento. Em (a) imagem luz visível, em (b) a identificação dos padrões de pele pela sobreposição da imagem de intensidade em (c).

Na figura 4 são apresentadas algumas sequências de *frames* coletadas ao longo da execução da captura através de uma *webcam* convencional no mesmo ambiente apresentado na figura 3 (no entanto, em outro ângulo de visão). Neste exemplo, o processo de treinamento foi realizado sobre uma determinada cena onde algumas pessoas estavam presente. As tonalidades de pele foram coletadas a partir estas pessoas de forma aleatória, e um modelo genérico para identificação de pessoas da etnia caucasianos foi gerado.

Figura 4. Ambiente de testes utilizado no experimento. Diferentes *frames* capturados sobre o mesmo conjunto de treinamento.

A ordem polinomial utilizada neste experimento foi q=3 e λ = 1 (Figura 4), apresentando um tempo médio de execução na ordem dos 27 *fps,* para um *quad-core Q9550 2.83 GHz* com sistema operacional *Linux OpenSuse 11.3 x86_64.* A linguagem de programação utilizada foi C++ utilizando como ambiente gráfico wxWidgets e a biblioteca *openCV* para captura do vídeo. Nestes experimentos a implementação não se utilizou de nenhuma estratégia ou abordagem de paralelismo.

3. Conclusão e Discussões

Nesse artigo foi apresentado um método supervisionado de identificação segmentação de padrões de pele para processamento de vídeo em tempo real. Os resultados obtidos demonstraram uma melhor discriminação quando comparados com a norma vetorial (L^2) e também em relação a distância de Mahalanobis, que pondera o cálculo da distância de acordo com as covariância. A estratégia utilizada na nesta abordagem utiliza uma estrutura de armazenando 4D onde somente os posições *rgbq* presentes em um frame de entrada são computados, e logo após somente utilizados em caso de nova requisição. Esta estrutura embora necessite maior espaço de armazenamento em disco possibilitou a execução em 27 fps em um dos experimentos realizados em um ambiente não-controlado. Os resultados preliminares mostraram que o método pode ser usado para interação humano-computador, utilizando para isso câmeras convencionais (*webcams*) e computadores pessoais. Como trabalho futuro, propõem-se a extensão desta metodologia para a identificação e reconhecimento de padrões gestuais, possibilitando

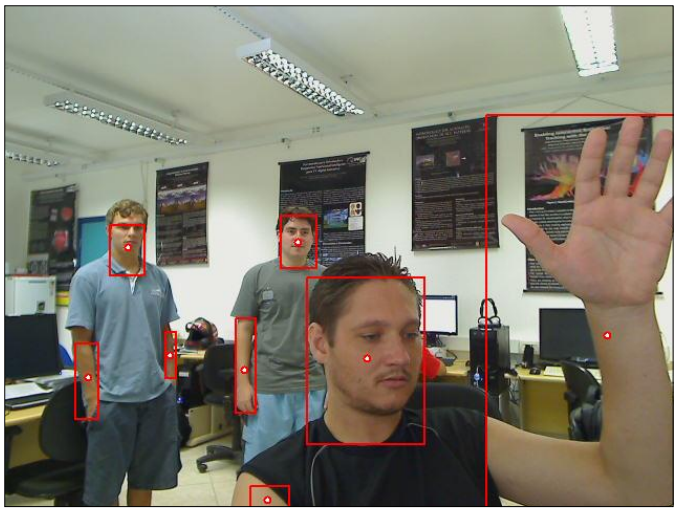

Figura 5. Demonstração da etapa seguinte à identificação dos padrões de pele: *tracking* dos objetos de interesse na cena para interpretação gestual.

assim a elaboração de um ambiente onde o humano possa interagir com o computador. Com isso, objetiva-se um meio de interação onde se utilize uma linguagem mais natural do que somente *mouse* e *teclado*, conforme ilustrado na Figura 5.

Referências

[1] E.P. Xing, A.Y. Ng, M.I. Jordan, S. Russell. Distance metric learning with application to clustering with side-information. Advances in NIPS, MIT Press, Cambridge, Mam USA (2003), 505–512.

[2] Foley, J., van Dam, A., Feiner, S., and Hughes, J. Computer Graphics: Principles and Practice in C, 2/E. Addison-Wesley, 1996.

[3] G. Grudic. J. Mulligan. Outdoor path labeling using polynomial Mahalanobis distance. Robotics: Science and Systems II Conference (2006).

[4] Kakumanu, P., Makrogiannis, S., and Bourbakis, N. A survey of skin-color modeling and detection methods. Pattern Recogn. 40, 3 (2007), 1106–1122.

[5] Koepfler, G., Lopez, C., and Morel, J. M. A multiscale algorithm for image segmentation by variational method. SIAM J. Numer. Anal. 31, 1 (1994), 282–299.

[6] Li, B., Xue, X., and Fan, J. A robust incremental learning framework for accurate skin region segmentation in color images. Pattern Recogn. 40, 12 (2007), 3621–3632.

[7] S .Xiang, F. Nie, C. Zhang. Learning a Mahalanobis distance metric for data clustering and classification. Pattern Recognition (2008).

[8] Stauffer, C. Learning a probabilistic similarity function for segmentation. In IEEE Workshop on Perceptual Organization in Computer Vision (POCV2004)

[9] Sun, H.-M. Skin detection for single images using dynamic skin color modeling. Pattern Recogn. 43, 4 (2010), 1413–1420.

[10] Yin, X., Guo, D., and Xie, M. Hand image segmentation using color and rceneural network. Robotics and Autonomous Systems 34, 4 (2001), 235–250.