Rastreamento eficiente de faces em um subespaço wavelet

Rogério Schmidt Feris

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA OBTENÇÃO DO GRAU DE MESTRE
EM
CIÊNCIA DA COMPUTAÇÃO

Área de Concentração : Ciência da Computação

Orientador : Prof. Dr. Roberto Marcondes Cesar Junior

- São Paulo, Maio de 2001 -

Rastreamento eficiente de faces em um subespaço wavelet

Este exemplar corresponde à redação final
da dissertação devidamente corrigida e
apresentada por Rogério Schmidt Feris e
aprovada pela Comissão Julgadora.

São Paulo, 11 de maio de 2001

Banca Examinadora :

- Prof. Dr. Roberto Marcondes Cesar Junior (orientador) - IME-USP
- Prof. Dr. Carlos Hitoshi Morimoto - IME-USP
- Prof. Dr. Luciano da Fontoura Costa - IFSC-USP

Ao meu amigo de sempre, tio Antônio

Agradecimentos

Ao meu amigo e orientador Prof. Dr. Roberto Marcondes Cesar Junior, por tudo o que com ele aprendi, em ambiente propício à pesquisa de excelência, aliado ao permanente apoio afetivo/emocional.

Ao pesquisador Dr. Volker Krüger (*University of Maryland*), idealizador da técnica utilizada neste trabalho, pela sua amizade e por ter acompanhado com muito interesse minha pesquisa, através de contatos freqüentes via e-mail.

Ao Prof. Dr. Carlos Hitoshi Morimoto, em especial, por ter acompanhado meu trabalho, e pelas frutíferas conversações técnicas (e não técnicas) que tivemos.

Ao Teo, companheiro de trabalho, e amigo em todas as horas.

A todo pessoal do laboratório de processamento de imagens do IME-USP, pela acolhida, abertura, e disponibilidade em me ajudar.

Ao Sérgio, que colaborou com a formação do banco de seqüências de imagens utilizado neste trabalho, e com a realização de testes pertinentes ao módulo de rastreamento de faces.

À FAPESP (99/01487-1), pelo fundamental suporte financeiro dispensado à execução desta pesquisa.

À minha família, especialmente a meus pais e a minha irmã, pelo apoio constante, e por vibrarem sempre com as minhas conquistas pessoais e profissionais.

À minha namorada Evelyn, com especial carinho, pelo apoio, paciência e compreensão, neste período de árduo trabalho.

A Deus, sentido da minha vida, em quem encontro força e motivação para alcançar meus objetivos.

Resumo

O reconhecimento computacional de faces humanas a partir de seqüências de vídeo emerge nos dias atuais como um instigante e desafiador campo de pesquisa. Os resultados dos estudos realizados neste tema encontram diversas aplicações em sistemas de segurança, autenticação de usuários, interação homem-máquina e outras.

Em geral, três etapas são discriminadas em um sistema de reconhecimento de faces a partir de vídeo: (1) detecção e rastreamento da face do usuário; (2) segmentação e normalização da face em cada quadro da seqüência; e (3) extração de características e classificação. O objetivo do presente trabalho consiste no estudo e desenvolvimento de algoritmos de visão computacional relativos à primeira etapa, com ênfase no problema de rastreamento de faces.

A detecção da presença da face do usuário, em uma seqüência de imagens, foi realizada utilizando-se um procedimento baseado em um modelo estatístico de cor da pele. Em contrapartida, para efetuar o rastreamento da face, realizamos experimentos com uma técnica mais sofisticada, recentemente proposta na literatura, denominada *Gabor wavelet networks* (GWNs). O problema de localização e rastreamento de características faciais também foi tratado, visto que o módulo de reconhecimento, atualmente sendo implementado por outro aluno de mestrado [Campos, 2001], utiliza essas características para efetuar a classificação da face.

Como principal contribuição deste estudo, salientamos a proposição de um método eficiente de rastreamento de faces, realizado em um subespaço wavelet de baixa dimensionalidade. Mais especificamente, esse método constitui um aperfeiçoamento, em termos de eficiência, da técnica de rastreamento baseada em GWNs. A comparação de ambas abordagens, bem como a avaliação de performance da técnica introduzida, considerando aspectos como precisão, robustez e eficiência, serão discutidas ao longo deste trabalho.

Abstract

Computational human face recognition from video sequences emerges nowadays as an instigating and challenging research field. The results obtained from the study conducted over this topic encompass several applications in security systems, user authentication, human-machine interaction and so forth.

In general, three steps are distinguished in a face recognition system from video sequences: (1) user's face detection and tracking; (2) face segmentation and normalization in each frame; and (3) feature extraction and classification. The objective of the present work consists in studying and developing computer vision algorithms relative to the first step, with emphasis on the face tracking problem.

The user's face detection in an image sequence was performed by using a method based on a statistical skin-color model. On the other hand, to accomplish face tracking, we have carried out experiments with a more sophisticated technique, proposed recently in the literature, which is called Gabor wavelet networks (GWNs). The problem of locating and tracking facial features was also considered, since the recognition module, which is being developed by another master's student [Campos, 2001], uses these features to perform face classification.

As the main contribution of this thesis, we highlight the proposal of an efficient method for face tracking, carried out in a low-dimensional wavelet subspace. More specifically, this method poses a considerable enhancement over the tracking based on GWNs. The comparison between both approaches, as well as the performance assessment of the introduced technique, considering aspects such as precision, robustness and efficiency, are discussed along this work.

Sumário

1	Introdução	1
1.1	Objetivos	2
1.2	Contribuições	4
1.3	Organização da Dissertação	5
2	Revisão Bibliográfica	6
2.1	Detecção de Faces	6
2.1.1	Imagens Estáticas	7
2.1.2	Imagens Dinâmicas	9
2.2	Rastreamento de Faces	10
2.2.1	Rastreamento 2D	11
2.2.2	Rastreamento 3D	15
3	Introdução à Teoria de Wavelets	16
3.1	Transformada Contínua em Wavelets	18
3.2	Transformada Discreta em Wavelets	20
3.3	Wavelet Networks	21

4	Representação de Faces Utilizando GWNs	23
4.1	Representação Baseada em GWNs	24
4.2	Cálculo Direto dos Pesos	27
4.3	Reparametrização de uma GWN	31
4.4	Características da Representação	33
5	Deteção e Rastreamento de Faces	35
5.1	Deteção de Faces	35
5.1.1	Modelo Estatístico de Cor da Pele	35
5.1.2	Segmentação da Face	37
5.2	Rastreamento de Faces Utilizando GWNs	39
5.3	Localização e Rastreamento de Características Faciais	40
6	Rastreamento de Faces em um Subespaço Wavelet	44
6.1	Metodologia	46
6.2	Distância entre Pesos	47
6.3	Algoritmo	48
7	Resultados Experimentais	50
7.1	Avaliação da Técnica de Rastreamento de Faces	50
7.1.1	Análise de Precisão	50
7.1.2	Análise de Robustez	52
7.1.3	Análise de Eficiência	56
7.2	Aplicação: Reconhecimento de Faces a partir de Vídeo	58
7.2.1	Formação do Banco de Dados	58
7.2.2	Deteção da Face	59

<i>SUMÁRIO</i>	iii
7.2.3 Posicionamento das Características Faciais	59
7.2.4 Rastreamento da Face e das Características Faciais	60
7.2.5 Seleção e Normalização dos Melhores Quadros	61
8 Conclusões	62
A Lista de Símbolos	64
B Lista de Abreviaturas	67

Lista de Figuras

1.1	Extração e normalização das características faciais utilizadas para o reconhecimento do usuário.	4
2.1	Abordagens utilizadas no problema de detecção de faces.	7
2.2	Abordagens utilizadas no problema de rastreamento de faces.	11
3.1	(a) Exemplos de funções $g_{u,b}(t)$, com diferentes deslocamentos e frequências. (b) Exemplos de funções $\psi_{a,b}(t)$, com diferentes deslocamentos e escalas. . .	19
4.1	Função de Gabor 2D considerada (parte imaginária).	25
4.2	(a) Imagem de uma face. (b) Representação obtida pela GWN com 52 wavelets. (c) Posição das 16 wavelets de maior peso.	27
4.3	(a) Imagem de uma face. (b) Representações obtidas utilizando GWNs com 32, 52, 100 e 320 wavelets.	28
4.4	(a) Imagem de uma face. (b) Reconstrução com base na equação (4.6). (c) Reconstrução com pesos ótimos.	31
4.5	(a) Imagem da face em uma pose diferente. (b) Representação wavelet reparametrizada. (c) Posição das 16 wavelets de maior peso.	32
5.1	Amostra de cor da pele.	36
5.2	(a) Agrupamento de cores no espaço RGB. (b) Agrupamento de cores no espaço cromático.	36

5.3	(a) Imagem de entrada. (b) Imagem obtida a partir do modelo de cor da pele.	37
5.4	Exemplo de modelos de cor de pele e não-pele no caso unidimensional. Note que a distribuição de probabilidade de cor não-pele é uma constante, dada por um valor de limiar.	38
5.5	(a) Imagem resultante da limiarização. (b) Aplicação do operador de fechamento e filtro da mediana.	38
5.6	Face média e sua representação wavelet.	42
5.7	Localização de características faciais em diferentes indivíduos.	42
5.8	Localização de características faciais sob diferentes expressões faciais e mudanças de iluminação.	43
5.9	Rastreamento de características faciais utilizando GWNs.	43
6.1	Mapeamentos entre o espaço de imagens $\mathbb{L}^2(\mathbb{R}^2)$ e o subespaço wavelet $\mathbb{R}^N$	45
7.1	Rastreamento de faces no subespaço wavelet, utilizando-se uma GWN com 116 wavelets.	51
7.2	Representações obtidas com 116, 51, 22 e 9 wavelets.	51
7.3	Estimação do parâmetro de posição c_x em cada quadro da seqüência.	53
7.4	Estimação do parâmetro de posição c_y em cada quadro da seqüência.	54
7.5	Estimação do parâmetro referente ao ângulo da face em cada quadro da seqüência.	55
7.6	Deslocamento da face em cada quadro da seqüência.	56
7.7	Comparação dos tempos exigidos por ciclo entre o rastreamento no espaço de imagens e o método proposto.	57
7.8	Detecção da face e localização das características faciais.	60
7.9	Rastreamento de olhos, nariz e boca. Note que o método é robusto a piscar de olhos, mudança homogênea de iluminação e diferentes expressões faciais.	61

Lista de Tabelas

2.1	Comparação de diferentes técnicas de rastreamento de faces.	12
2.2	Notação utilizada.	12

Capítulo 1

Introdução

Recentemente, com os avanços das técnicas de visão computacional, diversas pesquisas têm sido realizadas para implementar sistemas computacionais capazes de detectar a presença de pessoas em imagens, acompanhá-las, reconhecê-las e interpretar seus gestos. Exemplos concretos desse fato são evidenciados através da promoção de diversos congressos internacionais como o *IEEE Face and Gesture Recognition* e o *Perceptual User Interfaces*, além do considerável aumento do número de artigos publicados sobre esse assunto em anais de congressos e nas revistas especializadas em visão computacional.

Os resultados obtidos a partir dessas pesquisas são passíveis de aplicabilidade em diferentes áreas. Sistemas de reconhecimento de faces, por exemplo, podem ser utilizados na identificação de criminosos, busca de crianças desaparecidas, ou mesmo para permitir acesso a pessoas autorizadas a indústrias ou a bancos. Em geral, tais sistemas exigem, como pré-requisito, um módulo de detecção e rastreamento de faces, quando seqüências de vídeo são utilizadas. Mesmo quando somente imagens estáticas são tratadas, faz-se necessário um procedimento de localização da face para que o reconhecimento seja efetuado.

A detecção e o rastreamento de outras partes do corpo humano, além da face, é comumente requerida em sistemas de reconhecimento de gestos, os quais podem ser aplicados no ensino de movimentos, como dança ou atividades esportivas, e até mesmo para auxiliar pessoas portadoras de dificuldades especiais, traduzindo sinais para a linguagem falada. Nesse contexto, vale salientar, também, os sistemas de reconhecimento de expressões faci-

ais, que possibilitam às máquinas reagir apropriadamente ao estado emocional do usuário.

A utilização desses sistemas, de forma integrada, certamente proporcionará novas formas de interação homem-máquina, mais naturais e eficazes. Atualmente, a comunicação homem-computador é caracterizada pela utilização de dispositivos como teclado, mouse e por interfaces gráficas (*Graphical User Interfaces* - GUIs), baseadas em janelas, ícones e menus. Tendo em vista a atual possibilidade de aquisição de computadores pessoais eficientes, com entradas de áudio e vídeo a um custo relativamente baixo, surge o desafio de desenvolver novas interfaces mais naturais, em que as máquinas sejam capazes de “escutar” e “enxergar” o usuário, com o auxílio de técnicas de processamento da fala, visão computacional, aprendizado de máquina e outras.

Esse novo paradigma de interação homem-máquina é comumente referido como “interfaces de percepção do usuário” (*Perceptual User Interfaces* - PUIs)[Turk, 1998]. Nesse sentido, as interfaces baseadas em visão computacional cumprem um papel fundamental, concedendo às máquinas a capacidade de perceber visualmente o usuário, permitindo sua identificação e o reconhecimento de suas ações. Em geral, tais interfaces concentram-se em prover respostas para as seguintes questões[Cipolla and Pentland, 1998]: *Onde* está o usuário? *Quem* é o usuário? *O que* ele deseja, ou está fazendo, e *por quê*?

Deve-se mencionar que esse tema, além de constituir tópico atual de pesquisa em diferentes centros acadêmicos, tem atraído a atenção de vários líderes empresariais. Como exemplo, temos a criação do grupo de “Interfaces Baseadas em Visão” (<http://www.research.microsoft.com/vision/VisionBasedInterfaces>) no centro de pesquisa da Microsoft em Redmond, o projeto “BlueEyes” (<http://www.almaden.ibm.com/cs/blueeyes>) da IBM em Almaden e muitos outros.

1.1 Objetivos

O principal objetivo do presente trabalho consiste no estudo e desenvolvimento de algoritmos de visão computacional para a detecção e rastreamento de faces em seqüências de vídeo. Além disso, este trabalho está inserido em um projeto mais amplo, em desenvolvimento no departamento de Ciência da Computação do IME-USP, o qual visa a reconhecer pessoas a partir de seqüências de imagens.

Assim, assumimos que o usuário esteja localizado na frente de um computador, sendo filmado por uma câmera, em princípio estática, que adquire imagens coloridas. Supomos, também, um ambiente fechado, razoavelmente controlado, em que o usuário movimenta-se naturalmente diante da câmera.

Para efetuar-se o reconhecimento a partir da seqüência de imagens, três etapas são discriminadas: (1) detecção e rastreamento da face do usuário; (2) segmentação e normalização da face em cada quadro da seqüência; e (3) extração de características¹ e classificação. A primeira etapa consiste em, primeiramente, verificar se o usuário está presente diante da câmera, mediante a detecção da sua face. Uma vez detectada, a face é precisamente localizada e rastreada ao longo da seqüência de imagens. Logo após, em cada quadro da seqüência, a imagem da face é extraída e normalizada, caracterizando a segunda etapa. A normalização deve considerar aspectos geométricos (e.g. deformações rígidas da face) e radiométricos (e.g. variação nas condições de iluminação). Finalmente, a terceira etapa consiste no reconhecimento do usuário, através de um processo de extração de características e classificação da face.

Este trabalho concentra-se na primeira etapa, com maior ênfase no módulo de rastreamento de faces. O problema de localização e rastreamento de características faciais também será tratado, visto que módulo de reconhecimento, atualmente sendo implementado por outro aluno de mestrado do nosso grupo[Campos, 2001], utiliza essas características para efetuar a classificação da face.

Uma descrição detalhada do nosso projeto de reconhecimento de faces a partir de vídeo pode ser consultada no artigo[Feris et al., 2000b], incluído em anexo. Neste momento, é importante apenas salientar que o reconhecimento é realizado com base nas regiões dos olhos, nariz e boca do usuário. A figura 1.1 ilustra a extração e a normalização dessas características em um quadro de uma seqüência de imagens.

O tema focalizado neste trabalho, isto é, a tarefa de detectar a presença de uma face (e de suas características faciais) e acompanhá-la ao longo de uma seqüência de vídeo, constitui um tópico de pesquisa complexo e desafiador. Dificuldades aparecem com a

¹O termo “características” está sendo utilizado no sentido de “*features*” de reconhecimento estatístico de padrões, sendo mais amplo que o termo “características faciais”, o qual denota regiões como os olhos, nariz e boca do usuário.

Figura 1.1: Extração e normalização das características faciais utilizadas para o reconhecimento do usuário.

presença de ruído nas imagens, mudanças nas condições de iluminação, e deformações rígidas e não rígidas da face, como variação de pose, escala e expressão facial.

A abordagem que escolhemos para tratar esses problemas baseia-se em uma técnica recentemente proposta por [Kruger and Sommer, 2000], denominada *Gabor wavelet networks* (GWNs). Tal técnica possibilita decompor a imagem de uma face em uma combinação linear de Gabor wavelets, em que os pesos e parâmetros das wavelets são determinados mediante um processo de otimização, visando preservar o máximo de informação da imagem. Esta representação apresenta propriedades de invariância com relação a distorções afins da face e variação homogênea das condições de iluminação.

1.2 Contribuições

As principais contribuições deste trabalho estão discriminadas abaixo:

- Amplo levantamento bibliográfico, com o estudo dos principais métodos que compõem o estado da arte no que se refere à detecção e rastreamento de faces em seqüências de imagens.
- Experimentos com detecção de faces, utilizando-se um modelo estatístico de cor de pele para localizar regiões de interesse e um procedimento de correlação para verificar a presença de uma face nessas regiões[Feris et al., 2000a].

- Aplicação da técnica baseada em *Gabor wavelet networks* para a localização e rastreamento de características faciais[Feris and Cesar, 2001].
- Proposição de um esquema para reconhecimento de faces a partir de seqüências de vídeo[Feris et al., 2000b], juntamente com o aluno de mestrado responsável pela implementação do módulo de reconhecimento.
- Proposição e avaliação de um método eficiente de rastreamento de faces, realizado em um subespaço wavelet de baixa dimensionalidade[Feris et al., 2001]. Essa é a principal contribuição original deste trabalho, e constitui um aperfeiçoamento da técnica de rastreamento de faces baseada em GWNs.

1.3 Organização da Dissertação

Esta dissertação está organizada da seguinte forma: inicialmente, no capítulo 2, será apresentada uma revisão bibliográfica sobre detecção e rastreamento de faces em seqüências de vídeo. O capítulo 3 apresenta uma breve introdução à teoria de wavelets, importante para o entendimento da representação de faces baseada em GWNs, a qual é detalhada no capítulo 4. O capítulo 5 trata dos problemas de detecção e rastreamento de faces, descrevendo, também, um procedimento para localização e rastreamento de características faciais. No capítulo 6, o tópico de rastreamento de faces é aprofundado, mediante a proposição de uma técnica eficiente baseada em um subespaço wavelet. A avaliação de performance da técnica introduzida, bem como os resultados obtidos no contexto da aplicação considerada, estão descritos no capítulo 7. Por fim, o capítulo 8 traça algumas conclusões sobre o tema, mencionando o desenvolvimento de trabalhos futuros. Em anexo, constam os principais artigos publicados neste período.

Capítulo 2

Revisão Bibliográfica

Diversos trabalhos já foram propostos para realizar a detecção e o rastreamento de faces em seqüências de imagens. Neste capítulo, iremos apresentar um panorama geral das principais técnicas utilizadas neste tópico de pesquisa, comentando suas vantagens e desvantagens.

2.1 Detecção de Faces

A detecção automática de faces humanas em imagens conduz a questões de extrema complexidade na área de visão computacional. De fato, dificuldades advêm de fatores extrínsecos, como ruídos na imagem, variação das condições de iluminação, oclusão, bem como de fatores intrínsecos, como deformações e mudanças na aparência da face, devido, por exemplo, a diferentes expressões faciais. Assim, a maioria dos sistemas propostos na literatura impõem algumas restrições na aplicação considerada (e.g. processando imagens bem comportadas ou assumindo a face frontal na imagem).

A seguir, apresentaremos, primeiramente, os principais métodos de detecção de faces utilizados em imagens estáticas, para somente após discutir técnicas relacionadas com seqüências de vídeo. A figura 2.1 ilustra as diferentes abordagens que discutiremos ao longo desta seção.

Figura 2.1: Abordagens utilizadas no problema de detecção de faces.

2.1.1 Imagens Estáticas

Basicamente, as técnicas propostas para a detecção de faces em imagens estáticas podem ser classificadas em abordagens baseadas em características e abordagens holísticas. As abordagens baseadas em características realizam o processo de detecção da face através da busca por características faciais (e.g. olhos, nariz e boca), levando em consideração, por exemplo, a relação geométrica entre essas características. Em contrapartida, as abordagens holísticas utilizam a informação da face como um todo para efetuar a detecção.

Abordagens Baseadas em Características

As primeiras técnicas propostas para a detecção de faces em imagens estáticas basearam-se na procura por características faciais mediante análise de formas[Sakai et al., 1969, Govindaraju et al., 1989]. Basicamente, essas técnicas processam o mapa de bordas da imagem considerada, sendo sensíveis a ruído.

A utilização de técnicas baseadas em casamento de padrões (*template matching*), em geral, não é adequada para detectar características faciais, visto que tais características não são rígidas. Com base nesse aspecto, o trabalho de [Yuille et al., 1989] propõe a utilização de padrões deformáveis para detectar elementos faciais. Essa abordagem, no

entanto, depende de uma correta inicialização de parâmetros e demanda elevado custo computacional.

Métodos similares utilizam modelos de contornos ativos [Blake and Isard, 1998], os quais constituem curvas que podem ser deformadas ou atraídas a um contorno da imagem, em resposta a um sistema de forças. Entretanto, esses métodos também dependem de uma boa inicialização para detectar corretamente uma característica facial.

Outras técnicas, como a utilização de modelos estatísticos, que exploram a relação geométrica das características faciais [Coelho et al., 1999], ou mesmo técnicas baseadas em correlação [Zelinsky and Heinzmann, 1996], já foram propostas na literatura. No entanto, no que se refere à detecção de faces em imagens estáticas, as abordagens holísticas são utilizadas, em geral, com maior eficácia.

Abordagens Holísticas

No processo de detecção de múltiplas faces em imagens estáticas, a maioria das abordagens holísticas utilizam-se de uma janela, a qual é deslocada na imagem, em várias escalas, extraindo regiões que são processadas por um método específico.

O trabalho de [Moghaddam and Pentland, 1997] utiliza PCA (*Principal Component Analysis*) e modelos probabilísticos para efetuar a detecção de faces nessas regiões. A referida técnica baseia-se na decomposição de uma imagem em uma combinação linear de *eigenfaces* [Turk and Pentland, 1991], que definem um espaço vetorial de dimensionalidade reduzida, apropriado para representação de faces. Os testes foram conduzidos apenas com imagens bem comportadas. Os principais problemas inerentes à representação baseada em PCA incluem sua sensibilidade em relação à variação de iluminação e pose da face, bem como a necessidade de alinhamento das imagens em um sistema de coordenadas comum.

Abordagens baseadas em redes neurais foram utilizadas com sucesso para detectar faces em imagens complexas. O trabalho de [Sung and Poggio, 1998] utiliza uma técnica de agrupamentos (*clustering*) para modelar a distribuição de faces, a qual é limitada por padrões não-face que costumam resultar em falsos alarmes. Com base em distâncias definidas em relação a este modelo, um perceptron multicamada é treinado para determinar a presença ou ausência de faces em regiões da imagem. Dentro desta abordagem, também

destaca-se o trabalho de [Rowley et al., 1998a], que utiliza um conjunto de redes neurais, juntamente com algumas heurísticas, para efetuar a detecção frontal de faces. Mediante o re-treinamento de padrões incorretamente classificados como faces, o sistema alcançou bons resultados em imagens complexas. Recentemente, uma extensão deste trabalho foi proposta para tratar a detecção de faces rotacionadas [Rowley et al., 1998b], utilizando-se uma rede neural específica para se determinar a orientação da face.

Outras abordagens, as quais consideram a detecção de faces em imagens estáticas coloridas, utilizam informação de cor [Wu et al., 1999] para selecionar regiões de interesse, ao invés de deslocar uma janela na imagem, como nas técnicas supracitadas. Tais abordagens são eficientes e apropriadas para o processamento de seqüências de imagens, sendo comentadas a seguir.

2.1.2 Imagens Dinâmicas

A despeito das técnicas supracitadas considerarem apenas imagens estáticas, requerendo elevado custo computacional, nos últimos anos, vários trabalhos concernentes à análise de seqüências de vídeo foram desenvolvidos. Neste caso, em geral, as imagens são de menor qualidade, mas os métodos podem dispor de mais informação, como movimento, por exemplo.

Basicamente, duas etapas podem ser discriminadas nos métodos de detecção de faces em seqüências de vídeo: redução do espaço de busca e verificação da face. A primeira etapa consiste na seleção eficiente de regiões candidatas na imagem. Na segunda etapa, a presença ou ausência de uma face é verificada nas regiões selecionadas, configurando o processo de detecção.

Embora abordagens diferentes já tenham sido propostas, as técnicas que compõem o estado da arte na literatura utilizam essas duas etapas para processar seqüências de vídeo com maior eficiência.

Redução do Espaço de Busca

As principais técnicas utilizadas para reduzir o espaço de busca por faces são baseadas em modelos de cor da pele humana [Yang et al., 1997, Kjeldsen and Kender, 1996]. Essas

abordagens são simples e demandam um esforço computacional mínimo, sendo adequadas para o processamento em tempo-real. Além disso, a utilização de modelos de cor possibilita a localização de faces de diferentes raças[Yang et al., 1997], sob diferentes orientações. A principal limitação da referida abordagem é evidenciada na presença de forte variação das condições de iluminação.

Deve-se ressaltar que, além de faces, outros objetos com cor da pele são selecionados através da utilização dessas técnicas. De fato, a segunda etapa faz-se necessária para verificar a presença de uma face nas regiões (*blobs*) de cor da pele selecionadas.

Outras abordagens para redução do espaço de busca por faces em imagens incluem subtração de *background*[Turk, 1996] e informação de movimento[McKenna et al., 1996]. Tais técnicas, no entanto, somente são adequadas para câmeras estáticas e cenas comportadas, com poucos objetos em movimento.

Vale salientar, ainda, a utilização da informação de profundidade[Feyrer and Zell, 1999] e de câmeras infra-vermelho (para localização das pupilas[Morimoto et al., 1998]), as quais também possibilitam a seleção de regiões candidatas.

Verificação da Face

Após a redução do espaço de busca, a segunda etapa é realizada com a verificação da presença ou ausência da face em cada região de interesse selecionada.

As principais técnicas utilizadas nesta etapa são baseadas em informações como textura [Dai and Nakano, 1996], piscar de olhos[Bala et al., 1997], simetria[Sun et al., 1998], forma [Feyrer and Zell, 1999], correlação com padrões[Darrel et al., 1998] e outras.

Em geral, quanto maior a integração de informações, maior é a robustez da detecção, porém, também maior é o custo computacional exigido.

2.2 Rastreamento de Faces

O rastreamento de faces consiste no acompanhamento da face detectada ao longo da seqüência de vídeo. Em outras palavras, a face deve ser constantemente localizada em cada

Figura 2.2: Abordagens utilizadas no problema de rastreamento de faces.

quadro da sequência. Essa tarefa pode ser realizada no plano, configurando o rastreamento 2D, ou no espaço tridimensional, em que a posição e a orientação 3D da face são estimadas.

A figura 2.2 ilustra as abordagens que discutiremos ao longo dessa seção. Uma comparação dos trabalhos ora comentados é mostrada na tabela 2.1, a qual foi adaptada de [Toyama, 1998b]. As colunas dessa tabela denotam, da esquerda para direita, a referência ao trabalho, o objeto sendo rastreado (qualquer face, face particular, etc.), o método utilizado, o tempo de processamento e o *hardware* considerado. A notação utilizada está descrita na tabela 2.2.

2.2.1 Rastreamento 2D

A maior dificuldade inerente à pesquisa em rastreamento de faces consiste em efetuar esse processo com robustez e eficiência, permitindo o processamento em tempo-real. A seguir, apresentaremos as principais técnicas propostas para o rastreamento de faces 2D, discutindo aspectos como robustez e eficiência, com base na integração de informações.

Tabela 2.1: Comparação de diferentes técnicas de rastreamento de faces.

Rastreamento 2D de faces				
Trabalho	Objeto	Método	Tempo	Hardware
[Birchfield, 1998]	BCPE	C B	30Hz	Pm 200
[Crowley and Schwerdt, 1999]	QF	C F	15Hz	SGI
[McKenna et al., 1996]	QF	M T	5Hz	DC
[Darrel et al., 1998]	QF	C P T	12Hz	2 SGIs
[Turk, 1996]	FO	B	33Hz	Pm 166
[Kruger and Sommer, 2000]	FP	T	25Hz	Pm450
[Hager and Belhumeur, 1998]	FP	T	30Hz	Pm 120
[Morimoto et al., 1998]	QF	F	30Hz	Pm II 333
[Maurer and Malsburg, 1996]	FP	F	NTR	Sun 20
Rastreamento 3D de faces				
[Cascia et al., 2000]	FP	T	5Hz	SGI O2
[Stiefelhagen and J.Yang, 1997]	QF	F	30Hz	HP-9000

Tabela 2.2: Notação utilizada.

Métodos de rastreamento de Faces 2D			
C	Cor	B	Borda (contorno)
M	Movimento	F	Característica (<i>Feature</i>)
P	Profundidade	T	Modelo Holístico (<i>Template</i>)
Objeto		Tempo/Hardware	
BCPE	<i>Blob</i> de cor da pele elíptico	NTR	Acima de 1Hz
QF	Qualquer face	Pm X	Pentium XMhz
FO	Contorno da face e dos ombros	SGI	Silicon Graphics
FP	Face particular	DC	Datacube MaxVideo 250
		Sun X	Sun Sparc X
		HP	Hewlett-Packard

Integração de Informações

Basicamente, os métodos propostos para o rastreamento de faces 2D utilizam uma ou mais das seguintes informações [Toyama, 1998b] (listadas, aproximadamente, em ordem crescente de tempo de processamento): *cor*, *movimento*, *profundidade*, *bordas*, *características da face* e *modelo holístico da face*.

As técnicas de rastreamento de faces baseadas em cor consistem, essencialmente, em acompanhar um *blob* de cor da pele ao longo da seqüência de imagens, com a possível utilização de filtros de Kalman para estimar a posição da face em cada quadro da seqüência [Olivier et al., 1997]. Entre outros, citamos os trabalhos de [Yang and Waibel, 1996] e [Raja et al., 1998] que apresentam um modelo estatístico de cor robusto a pessoas de diferentes raças e adaptável a diferentes condições de iluminação. Bons resultados também são apresentados no sistema proposto por [Birchfield, 1998], que utiliza restrições geométricas em adição à informação de cor para rastrear faces em tempo-real. O trabalho de [Crowley and Schwerdt, 1999] apresenta um sistema robusto de rastreamento de faces para compressão de vídeo, baseado em cor e correlação. Ainda neste contexto, é válido mencionar o trabalho de [Zarit et al., 1999] que mostra uma comparação entre dois métodos de segmentação de cor da pele em cinco diferentes espaços de cor.

Além de simplicidade e eficiência, as técnicas baseadas em cor permitem uma maior robustez com relação a deformações rígidas e não rígidas da face. No entanto, essas técnicas falham na presença de forte variação das condições de iluminação e podem ser afetadas por outros objetos com cor da pele na imagem.

O rastreamento de faces baseado na informação de movimento [McKenna et al., 1996] é comumente empregado na presença de ruídos na imagem e fundos (*background*) complexos. Porém, como já mencionado, essas técnicas, em geral, assumem a utilização de câmeras estáticas. Outra informação importante na segmentação de faces em imagens com fundos complexos é a profundidade, utilizada no trabalho de [Darrel et al., 1998].

O trabalho de [Turk, 1996] utiliza informação de borda para localizar e rastrear a face e os ombros do usuário em tempo-real. O mapa de bordas é computado a partir da imagem resultante da subtração de *background*, sendo que apenas um usuário deve estar presente diante da câmera. Outro trabalho baseado em bordas foi proposto por

[Silva et al., 1995] para a localização e o rastreamento de características faciais. Essa técnica falha em diversas situações, como presença de óculos e cabelos cobrindo a testa.

As abordagens baseadas em modelos holísticos da face utilizam um conjunto desses modelos, correspondendo às diferentes poses da face, ou então um único modelo deformável, que se ajusta à face em cada quadro da sequência de imagens. Por exemplo, o recente trabalho de [Kruger and Sommer, 2000], o qual detalharemos ao longo desta dissertação, utiliza um modelo wavelet da face que pode ser deformado eficientemente em uma imagem.

A técnica proposta por [Hager and Belhumeur, 1998] considera modelos paramétricos de geometria e iluminação para efetuar o rastreamento de faces em tempo-real. Embora a face tenha sido modelada como um objeto rígido sob movimentos restritos, o método mostrou-se robusto a mudanças de expressão facial e pequenas rotações fora do plano. Neste contexto, um *framework* genérico é proposto para realizar o rastreamento eficiente de objetos que trata problemas como variação de pose, mudanças na iluminação e oclusão. Como limitação dessa técnica, salientamos os modelos de iluminação que são dependentes do usuário.

As abordagens baseadas em características efetuam o rastreamento de características (e.g. olhos, nariz, boca) ou pontos faciais específicos. Por exemplo, o sistema proposto por [Morimoto et al., 1998] realiza o rastreamento da face mediante a localização de pupilas em cada quadro da sequência. Esse sistema, o qual utiliza fontes de luz infra-vermelho, é altamente confiável e adequado para aplicações em interfaces homem-máquina. Como limitação, citamos sua sensibilidade a movimentos bruscos da face do usuário.

O trabalho de [Maurer and Malsburg, 1996] descreve um sistema para rastrear pontos faciais utilizando a técnica *elastic graph matching*. Neste método, filtros de Gabor, com diferentes frequências e orientações, são aplicados em algumas posições do interior da face, formando vetores de características, denominados *jets*. A face é então representada como um grafo, em que os nós correspondem aos *jets* e as arestas codificam informação geométrica. O rastreamento da face é realizado mediante um procedimento de casamento de grafos (*graph matching*) em cada quadro. A principal desvantagem dessa abordagem é o alto custo computacional requerido, o que conduz a uma taxa de processamento inadequada para aplicações em tempo-real.

2.2.2 Rastreamento 3D

Embora o rastreamento 3D de faces não pertença ao escopo deste trabalho, é válido mencionar, brevemente, algumas técnicas relativas a esse assunto. Basicamente, essas técnicas podem ser divididas em duas abordagens: baseadas em modelo e baseadas em características.

Abordagens Baseadas em Modelo

As abordagens baseadas em modelo [Cascia et al., 2000, Edwards et al., 1998] utilizam um modelo tridimensional da face para determinar a posição e a orientação 3D da face em cada quadro da seqüência. Por exemplo, o trabalho de [Cascia et al., 2000] utiliza um modelo cilíndrico no qual é projetada a textura da face. A determinação da posição e orientação do cilindro em cada quadro configura o processo de rastreamento. Em geral, as abordagens baseadas em modelo são robustas a diferentes deformações da face, mas demandam elevado custo computacional, não sendo, na maioria dos casos, passíveis de execução em tempo-real.

Abordagens Baseadas em Características

As abordagens baseadas em características determinam a orientação 3D da face a partir da posição de determinadas características faciais na imagem [Jebara and Pentland, 1997, Toyama, 1998a]. Por exemplo, o trabalho de [Stiefelhagen and J. Yang, 1997] determina a posição das pupilas, narinas e cantos da boca em cada quadro da seqüência, obtendo a orientação 3D da face a partir dessa informação. Geralmente, as abordagens baseadas em características são eficientes e permitem o processamento em tempo-real, mas são sensíveis a falhas quando os pontos característicos não são localizados (e.g. em virtude de oclusão ou variação de iluminação).

Capítulo 3

Introdução à Teoria de Wavelets

Apresentaremos, neste capítulo, uma breve introdução à teoria de wavelets, sob a ótica da análise de sinais. A formalização dessa teoria foi concretizada na década de 80 [Grossmann and Morlet, 1984, Mallat, 1989] e no início da década de 90 [Daubechies, 1990], com base na generalização de conceitos já conhecidos, oriundos de diversos campos de pesquisa, como geofísica, análise e compressão de sinais, física e matemática. Desde então, a teoria de wavelets tem atraído a atenção de diversos pesquisadores, encontrando aplicações em diferentes áreas, particularmente em processamento de imagens e visão computacional [Costa and Cesar, 2001, Cesar and Costa, 1998]. Para uma revisão aprofundada sobre o tema, recomenda-se ao leitor o livro de Ingrid Daubechies [Daubechies, 1992], o qual apresenta uma análise matemática detalhada desta teoria.

O termo “wavelets” foi originariamente introduzido por J. Morlet, sendo a base matemática de suas idéias formalizada pelo físico teórico A. Grossmann [Grossmann and Morlet, 1984]. Os dados sísmicos estudados por Morlet exibiam conteúdos de frequência que mudavam rapidamente ao longo do tempo, para os quais a transformada de Fourier não era adequada como ferramenta de análise.

De fato, a transformada de Fourier não permite uma análise local do conteúdo de frequência do sinal. Eventos que venham a ocorrer em intervalos de tempos distintos, e mesmo bastante remotos, contribuem de maneira global para a transformada, afetando a representação como um todo. A equação (3.1) ilustra a transformada de Fourier de um sinal contínuo $f(t) : \mathbb{R} \rightarrow \mathbb{R}$. É importante notar que a transformada está baseada na

integração de todo o sinal para o cálculo de cada frequência.

$$F(u) = \int f(t)e^{-i2\pi ut} dt \quad (3.1)$$

A transformada janelada de Fourier, por outro lado, permite uma análise da frequência do sinal localmente no tempo. Neste caso, uma janela de observação é deslocada no domínio do tempo, e a transformada de Fourier da porção “visível” do sinal é calculada para cada posição da janela. Considerando $j(t)$ a função que cumpre o papel de janela, podemos definir a transformada janelada de Fourier de um sinal contínuo $f(t)$:

$$F(u, b) = \int j(t - b)f(t)e^{-i2\pi ut} dt \quad (3.2)$$

Quando a função janela $j(t)$ é definida como sendo uma função Gaussiana, a transformada janelada de Fourier é denominada transformada de Gabor. Sugere-se ao leitor a consulta do artigo [Gabor, 1946] para uma visão mais aprofundada dessa transformada.

Considerando-se $g_{u,b}(t) = j(t - b)e^{-i2\pi ut}$, a equação (3.2) pode ser escrita como:

$$F(u, b) = \int f(t)g_{u,b}(t)dt = \langle f, g_{u,b} \rangle \quad (3.3)$$

Assim, a transformada consiste no cálculo de produtos internos entre o sinal $f(t)$ e a função $g_{u,b}(t)$, a qual determina o deslocamento da janela de observação (parâmetro b) e a frequência da exponencial complexa (parâmetro u).

A transformada janelada de Fourier, porém, pode não ser adequada para a análise de determinados sinais, tendo em vista que o tamanho da janela de observação permanece constante para todas as frequências. Morlet introduziu uma nova transformada, denominada transformada em wavelets, na qual o tamanho da janela varia com a frequência, possibilitando que eventos de alta frequência possam ser localizados com uma maior resolução temporal, ao passo que componentes de baixa frequência possam ser analisados com maior resolução no domínio de Fourier. Intuitivamente, um evento de baixa frequência precisa de uma janela maior no domínio do tempo para ser observado.

3.1 Transformada Contínua em Wavelets

Definiremos, inicialmente, a transformada em wavelets considerando um sinal contínuo $f(t) : \mathbb{R} \rightarrow \mathbb{R}$. A generalização para sinais 2D $f(x, y)$ ou imagens, objeto de nosso estudo, pode ser obtida, tendo sido definida no trabalho de [Antoine et al., 1993]. A transformada contínua em wavelets de $f(t)$ é dada pela equação:

$$F(a, b) = \int f(t) \psi_{a,b}(t) dt \quad (3.4)$$

Nesta equação, os parâmetros $a > 0$ e b variam continuamente, sendo que as funções $\psi_{a,b}(t)$ são denominadas wavelets e definidas da seguinte forma:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi \left(\frac{t-b}{a} \right) \quad (3.5)$$

Claramente, tais funções são derivadas a partir de dilatações e translações de uma wavelet mãe ψ , a qual deve satisfazer o critério de admissibilidade:

$$C_\psi = 2\pi \int \frac{|\bar{\psi}(u)|^2}{|u|} du < \infty \quad (3.6)$$

onde $\bar{\psi}(u)$ é a transformada de Fourier de $\psi(t)$. Se $\bar{\psi}$ é uma função contínua, então C_ψ pode ser finito somente se $\bar{\psi}(0) = 0$, isto é, $\int \psi(x) dx = 0$.

Observando a equação (3.4), percebemos que a transformada em wavelets depende de dois parâmetros a e b , que correspondem, respectivamente, às informações de escala e tempo. É comum associar-se o inverso da variável de escala com a frequência. Assim, o sinal 1D $f(t)$ é desdobrado em uma transformada 2D $F(a, b)$, permitindo uma análise tempo-frequência do sinal, de forma semelhante à transformada janelada de Fourier. O semi-plano bidimensional (a, b) é comumente denominado *espaço de configurações* ou *plano tempo-escala* na literatura.

A equação (3.4) pode ser reescrita como o produto interno de $f(t)$ com uma função wavelet $\psi_{a,b}(t)$, isto é, $F(a, b) = \langle f, \psi_{a,b} \rangle$. Basicamente, este produto interno irá “medir” o conteúdo do espaço de configurações do sinal $f(t)$ em torno do ponto (a, b) nesse espaço.

É importante salientarmos a principal diferença entre a transformada em wavelets e a transformada janelada de Fourier, a qual reside na forma das funções $\psi_{a,b}(t)$ e $g_{u,b}(t)$, como ilustra a figura 3.1. Todas as funções $g_{u,b}(t)$ apresentam a mesma largura, a despeito

Figura 3.1: (a) Exemplos de funções $g_{u,b}(t)$, com diferentes deslocamentos e frequências. (b) Exemplos de funções $\psi_{a,b}(t)$, com diferentes deslocamentos e escalas.

do parâmetro de frequência u . Em contraste, a largura (no domínio do tempo) das funções $\psi_{a,b}(t)$ varia com a frequência. Mais especificamente, eventos de altas frequências conduzem a funções $\psi_{a,b}(t)$ mais finas (maior resolução temporal), enquanto que eventos de baixas frequências conduzem a funções $\psi_{a,b}(t)$ mais largas no domínio do tempo (maior resolução frequencial). Essa propriedade possibilita, por exemplo, que singularidades (eventos curtos de alta frequência) sejam analisadas com uma maior resolução temporal que outros eventos de mais longa duração.

A transformada em wavelets pode ser vista como a representação de um objeto, o qual pode ser uma função, imagem, sinal de som, etc. A equação (3.4), de fato, está decompondo $f(t)$ em uma superposição de wavelets $\psi_{a,b}(t)$, sendo que os coeficientes nessa superposição são exatamente obtidos pela transformada em wavelets de $f(t)$. Neste caso, é possível reconstruir $f(t)$ perfeitamente a partir de sua representação, como mostra a equação abaixo:

$$f(t) = \frac{1}{C_\psi} \int \int \langle f, \psi_{a,b} \rangle \psi_{a,b}(t) \frac{da db}{a^2} \quad (3.7)$$

A reconstrução, como pode ser notado, é obtida mediante uma integração sobre todo o espaço de configurações.

3.2 Transformada Discreta em Wavelets

Claramente, a representação baseada na transformada em wavelets contínua é altamente redundante. Entretanto, essa redundância pode ser reduzida ou eliminada mediante a restrição da escolha da wavelet ψ e pela discretização do espaço de configurações. Este processo origina a transformada discreta em wavelets, em que o sinal $f(t)$ é expressado como uma combinação linear de wavelets $\psi_{a,b}(t)$. Neste caso, os parâmetros a e b assumem valores discretos.

Vamos considerar, deste modo, uma discretização do espaço de configurações através do conjunto $S \subset \mathbb{R}^+ \times \mathbb{R}$, assumindo que $a \in \mathbb{R}^+$ e $b \in \mathbb{R}$ assumem valores discretos. Logo, podemos definir $B_\psi = \{\psi_{a,b} | (a, b) \in S\}$ como uma família discreta de wavelets.

É importante notar que o espaço de configurações não deve ser discretizado uniformemente, como na transformada de Fourier janelada discreta [Daubechies, 1990]. A discretização do parâmetro de translação depende da escala da wavelet. Assim, para altas frequências, as wavelets são transladadas por passos pequenos, enquanto que, para frequências baixas, a translação é feita com passos maiores.

Devido à discretização do espaço de configurações, a transformada em wavelets discreta inversa não está diretamente disponível [Daubechies, 1990]. Assim, surgem questões de como o sinal $f(t)$ pode ser reconstruído e se existem algoritmos numericamente estáveis para computar a transformada.

As respostas a essas questões dependem da escolha de ψ e da maneira como o espaço de configurações é discretizado. Basicamente, uma escolha adequada (que possibilita a reconstrução de $f(t)$) conduz a duas configurações distintas: sistemas discretos redundantes (frames) [Daubechies, 1990] e bases de wavelets, as quais podem ser ortogonais ou não ortogonais [Vetterli and Kovacevic, 1995].

No caso em que a família de wavelets B_ψ constitui uma base ortogonal, o sinal $f(t)$ pode ser expresso como:

$$f(t) = \sum_{(a,b) \in S} \langle \psi_{a,b}, f \rangle \psi_{a,b}(t) \quad (3.8)$$

Por exemplo, escolhendo-se como wavelet mãe a conhecida função de Haar, e mediante uma discretização apropriada do espaço de configurações, é possível gerar uma base

ortornormal de wavelets para o espaço $\mathbb{L}^2(\mathbb{R})^1$ [Daubechies, 1992].

No caso de bases não ortogonais ou sistemas discretos redundantes, o sinal $f(t)$ é reconstruído utilizando-se uma família de wavelets duais $\tilde{B}_\psi = \{\tilde{\psi}_{a,b} | (a,b) \in S\}$:

$$\begin{aligned} f(t) &= \sum_{(a,b) \in S} \langle \tilde{\psi}_{a,b}, f \rangle \psi_{a,b}(t) \\ &= \sum_{(a,b) \in S} \langle \psi_{a,b}, f \rangle \tilde{\psi}_{a,b}(t) \end{aligned} \quad (3.9)$$

Duas famílias de funções B_ψ e $\tilde{B}_\psi$ são duais se para cada $\psi_i \in B_\psi$ e $\tilde{\psi}_j \in \tilde{B}_\psi$ é verificada a condição:

$$\langle \psi_i, \tilde{\psi}_j \rangle = \delta_{i,j} = \begin{cases} 1, & \text{se } i = j \\ 0, & \text{se } i \neq j \end{cases} \quad (3.10)$$

A seção 4.2 apresentará esses conceitos com mais detalhes.

Finalmente, é válido mencionar o clássico artigo de [Mallat, 1989], que apresenta a transformada discreta em wavelets em um framework denominado “análise multiresolução”, propondo, concomitantemente, esquemas eficientes para o cálculo dessa transformada com base na teoria de filtragem.

3.3 Wavelet Networks

Em nosso trabalho, iremos utilizar a decomposição em wavelets de uma forma diferente das abordagens supracitadas. Nossa representação será baseada em uma técnica proposta por [Zhang and Benviste, 1992] denominada *wavelet network*. Nesta técnica, um sinal $f(t)$ é representado por um número fixo de wavelets, que correspondem aos nós de uma rede neural. Os coeficientes das wavelets são representados pelos pesos sinápticos desta rede. Mediante um procedimento de otimização, os pesos e parâmetros das wavelets são determinados de forma a melhor aproximar o sinal considerado. Sob a ótica da análise de sinais, $f(t)$ é aproximado por uma combinação linear de wavelets²:

$$\hat{f}(t) = \sum_{i=1}^M w_i \psi_{a_i, b_i}(t) \quad (3.11)$$

¹Uma função $f(t) : \mathbb{R} \rightarrow \mathbb{R} \in \mathbb{L}^2(\mathbb{R})$ se $\int |f(t)|^2 dt < \infty$

²Neste trabalho, não estamos interessados na reconstrução perfeita do sinal.

É importante salientar que os parâmetros a e b das wavelets são obtidos a partir do espaço contínuo de configurações, através do processo de otimização. A precisão da aproximação é dada pela escolha do número M de wavelets.

A transformada em wavelets pode ser facilmente estendida para o caso bidimensional [Antoine et al., 1993], com vistas à análise de imagens, por exemplo. Neste caso, porém, além dos parâmetros de dilatação e translação, um parâmetro adicional de rotação é utilizado na transformada. No próximo capítulo, iremos detalhar uma técnica baseada em *Gabor Wavelet Networks* para representação da imagem de uma face.

Capítulo 4

Representação de Faces Utilizando GWNs

A escolha da representação computacional de objetos em imagens cumpre um papel fundamental no desempenho da aplicação considerada, determinando aspectos como eficiência e robustez.

No que se refere à representação 2D de faces, duas abordagens são particularmente discriminadas[Brunelli and Poggio, 1993]: representações baseadas em características (*feature-based*) e representações baseadas em padrões (*template-based*). As representações baseadas em características[Cox et al., 1996, Yow and Cipolla, 1997] descrevem o objeto como uma coleção de suas características, as quais podem constituir medidas geométricas[Cox et al., 1996], ou até mesmo um conjunto de filtragens aplicadas localmente na imagem [Wiskott et al., 1997]. Em contraste, as abordagens baseadas em padrões [Rowley et al., 1998a, Turk and Pentland, 1991], caracterizam-se por representações holísticas, utilizando, por exemplo, todo o conjunto de *pixels* do objeto.

Neste capítulo, apresentaremos, com base no trabalho de [Kruger and Sommer, 2000], uma representação 2D de objetos baseada em *Gabor wavelet networks* (GWNs). Tal representação apresenta propriedades de ambas abordagens, baseada em características e baseada em padrões.

Basicamente, as GWNs possibilitam que a imagem discreta de uma face seja aproxi-

mada por uma combinação linear de Gabor wavelets, em que os pesos e parâmetros das wavelets são determinados mediante um processo de otimização, visando a preservar o máximo de informação da imagem. A representação obtida é caracterizada pela abordagem baseada em características, tendo em vista que a face é representada por uma coleção de Gabor wavelets específicas. De fato, somente os pesos e parâmetros otimizados de tais funções são armazenados na representação. Por outro lado, as GWNs permitem a reconstrução da face a partir do conjunto de Gabor wavelets, de modo que a representação possa ser tratada de forma holística, caracterizando a abordagem baseada em padrões. É interessante notar que o número de características utilizadas na GWN permite controlar a precisão da reconstrução da face.

Entre outras vantagens, a representação baseada em GWNs apresenta propriedades de invariância com relação a distorções afins da face e variação homogênea das condições de iluminação. Essas propriedades são evidenciadas no módulo de rastreamento de faces, o qual detalharemos adiante. A seguir, apresentaremos a técnica baseada em GWNs para representação de faces, elucidando suas principais características.

4.1 Representação Baseada em GWNs

Para definir uma *Gabor wavelet network*, consideramos primeiramente uma função wavelet mãe $\psi : \mathbb{R}^2 \rightarrow \mathbb{R}$, definida pela parte imaginária de uma função de Gabor 2D, caracterizada pela equação abaixo. A figura 4.1 ilustra graficamente essa função.

$$\psi(x, y) = \exp\left(-\frac{1}{2}(x^2 + y^2)\right) \sin(x) \quad (4.1)$$

Seja $\Psi = \{\psi_{\mathbf{n}_1}, \dots, \psi_{\mathbf{n}_M}\}$ uma família de M Gabor wavelets, derivadas a partir de dilatações, translações e rotações de ψ , com a forma:

$$\begin{aligned} \psi_{\mathbf{n}_i}(x, y) &= \exp\left(-\frac{1}{2}[s_{x_i}((x - c_{x_i})\cos\theta_i - (y - c_{y_i})\sin\theta_i)]^2 \right. \\ &\quad \left. + [s_{y_i}((x - c_{x_i})\sin\theta_i + (y - c_{y_i})\cos\theta_i)]^2\right) \\ &\quad \times \sin(s_{x_i}((x - c_{x_i})\cos\theta_i - (y - c_{y_i})\sin\theta_i)) \end{aligned} \quad (4.2)$$

com $x, y \in \mathbb{R}$ e o vetor de parâmetros $\mathbf{n}_i = (c_{x_i}, c_{y_i}, \theta_i, s_{x_i}, s_{y_i}) \in \mathbb{R}^5$, onde c_{x_i}, c_{y_i} denotam a translação da Gabor wavelet, s_{x_i}, s_{y_i} denotam a dilatação e θ_i denota a orientação.

Figura 4.1: Função de Gabor 2D considerada (parte imaginária).

De acordo com [Zhang and Benviste, 1992], qualquer função $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, pertencente ao espaço das funções quadraticamente integráveis, isto é, $f \in \mathbb{L}^2(\mathbb{R}^2)$, pode ser representada por uma *wavelet network*.

Seja f a imagem contínua de uma face, a qual pode ser tomada como sendo uma função de $\mathbb{L}^2(\mathbb{R}^2)$. Vamos supor, sem perda de generalidade, que f possui média zero. A representação wavelet de f é obtida mediante a minimização da função energia abaixo:

$$E = \min_{\mathbf{n}_i, w_i \forall i} \|f - (\sum_{i=1}^M w_i \psi_{\mathbf{n}_i})\|^2 \quad (4.3)$$

com relação aos pesos $w_i \in \mathbb{R}$ e aos parâmetros $\mathbf{n}_i \in \mathbb{R}^5$ das Gabor wavelets.

Assim, o conjunto de wavelets:

$$\begin{aligned} \Psi &= \{\psi_{\mathbf{n}_1}, \dots, \psi_{\mathbf{n}_M}\} \text{ e o vetor} \\ \mathbf{w} &= (w_1, \dots, w_M) \end{aligned}$$

definem uma *Gabor wavelet network* $(\Psi, \mathbf{w})$ para uma função específica f .

Obtemos, então, a representação wavelet de f , a qual corresponde à aproximação $\hat{f}$ da imagem original:

$$\hat{f} = \sum_{i=1}^M w_i \psi_{\mathbf{n}_i} \quad (4.4)$$

A equação (4.3) também pode ser aplicada considerando-se uma imagem discreta I^1

¹Nós utilizamos a notação $f, g, \dots$ para referenciar funções contínuas, ao passo que a notação $I, J, \dots$ refere-se explicitamente a imagens discretas em níveis de cinza.

(e.g. f amostrada). Assim, uma *Gabor wavelet network* é definida através de um vetor M -dimensional de pesos w_i e um vetor M -dimensional de Gabor wavelets ψ_{n_i} , em que os pesos w_i e os vetores de parâmetros n_i são escolhidos tal que a soma ponderada das Gabor wavelets ψ_{n_i} aproxima de forma ótima a imagem discreta I .

Deve-se mencionar que o trabalho de [Daugman, 1988] utiliza a equação (4.3) para determinar os pesos ótimos de um conjunto fixo de wavelets não ortogonais. Com a técnica baseada em GWNs, não somente os pesos, mas também os parâmetros ótimos de cada wavelet são determinados a partir do espaço contínuo de configurações.

Assim, uma única wavelet, com pesos e parâmetros otimizados, aproxima precisamente a estrutura local da imagem. Essa é a principal vantagem sobre o espaço discreto de configurações, em que uma combinação de wavelets deve ser utilizada para efetuar tal aproximação [Wiskott et al., 1997].

Dada uma imagem discreta I de uma face, a otimização da GWN, isto é, a minimização da função energia na equação (4.3), é realizada pelo método de otimização não linear Levenberg-Marquardt [Press et al., 1986]. Esse método possibilita determinar os parâmetros de um modelo contínuo² de modo a interpolar um conjunto de dados discretos. Em nosso caso, o modelo contínuo refere-se à combinação linear de Gabor wavelets. Os dados discretos correspondem ao conjunto de *pixels* da imagem I . Com o algoritmo de Levenberg-Marquardt, é possível determinar os pesos e parâmetros das wavelets (isto é, os parâmetros do modelo) de forma a interpolar os dados discretos da imagem da face. Mais especificamente, dado um conjunto de parâmetros iniciais do modelo, o algoritmo gradualmente (em vários ciclos) ajusta tais parâmetros até que o mínimo da função energia na equação (4.3) seja encontrado. É importante mencionar que esse método pode resultar na determinação de mínimos locais.

Como exemplo, consideremos a imagem da face ilustrada na figura 4.2(a). Utilizando uma GWN com somente 52 Gabor wavelets, inicializadas de forma homogênea na região interna da face, é possível gerar uma representação wavelet da face, a qual está ilustrada na figura 4.2(b). É importante mencionar que esta imagem reconstruída, também denominada *Gabor wavelet template* (GWT), é uma versão discreta do modelo wavelet contínuo da face. A figura 4.2(c) ilustra a posição das 16 wavelets de maior peso, após o

²O modelo depende dos parâmetros de forma não linear

Figura 4.2: (a) Imagem de uma face. (b) Representação obtida pela GWN com 52 wavelets. (c) Posição das 16 wavelets de maior peso.

procedimento de otimização.

A precisão da representação da face é determinada pelo número M de wavelets da GWN. A escolha de M deve ser feita com base na aplicação desejada. Por exemplo, para problemas que requerem processamento em tempo real, um número pequeno de wavelets pode ser escolhido, gerando-se uma representação compacta e eficiente. No próximo capítulo, iremos mostrar que tal representação é adequada para o problema de rastreamento de faces, pois permite eficiência e adaptação a diferentes indivíduos. Por outro lado, se a aplicação exigir o processamento de informação mais precisa da face, então um valor mais alto para M deve ser escolhido.

Esta variabilidade da GWN na representação da informação da face é denominada atenção progressiva[Tsotsos, 1990]. A figura 4.3 ilustra este conceito, mostrando representações de uma face com diferentes níveis de precisão. É interessante notar que a GWN constitui uma *RBF network* (*Radial Basis Functions*)[Kosko, 1992], permitindo generalização dos dados de treinamento quando um pequeno número de funções base são utilizadas.

4.2 Cálculo Direto dos Pesos

Como já mencionado, no trabalho de [Daugman, 1988], a equação (4.3) é utilizada para o cálculo dos pesos de um conjunto de Gabor wavelets fixas. Visto que a otimização, em geral, caracteriza-se por um processo lento, iremos apresentar uma forma eficiente para o cálculo direto dos pesos de uma GWN. Esta seção servirá de base para a proposição de

Figura 4.3: (a) Imagem de uma face. (b) Representações obtidas utilizando GWNs com 32, 52, 100 e 320 wavelets.

um novo método de rastreamento de faces, o qual será apresentado no capítulo 6, como principal contribuição do presente trabalho.

Inicialmente, é importante detalhar alguns conceitos apresentados na seção 3.2. Nesse sentido, vamos considerar uma família de wavelets $\Phi = \{\phi_1, \phi_2, \dots, \phi_N\}$ ³ que constituem uma base para $\mathbb{L}^2(\mathbb{R}^2)$, e uma função $f \in \mathbb{L}^2(\mathbb{R}^2)$, isto é, $f \in \langle \Phi \rangle$. Então existe um conjunto de pesos $\{w_i\}$ tal que:

$$f = \sum_{i=1}^N w_i \phi_i \quad (4.5)$$

No caso das wavelets $\{\phi_i\}$ formarem uma base ortogonal, os pesos $\{w_i\}$ são facilmente calculados pelo produto interno de f com cada wavelet ϕ_i em $\mathbb{L}^2(\mathbb{R}^2)$:

$$w_i = \langle f, \phi_i \rangle \quad (4.6)$$

Portanto, neste caso, a reconstrução de f é dada por:

$$f = \sum_{i=1}^N \langle f, \phi_i \rangle \phi_i \quad (4.7)$$

No entanto, quando a família de wavelets Φ não constitui uma base ortogonal, o cálculo dos pesos $\{w_i\}$ deve ser realizado de forma diferente. Uma maneira eficiente consiste em utilizar uma base dual de wavelets $\tilde{\Phi} = \{\tilde{\phi}_1, \tilde{\phi}_2, \dots, \tilde{\phi}_N\}$ [Vetterli and Kovacevic, 1995,

³A notação Φ será utilizada para designar wavelets genéricas, ao passo que a notação Ψ denotará a utilização de Gabor wavelets.

Daubechies, 1990], escolhida de forma a satisfazer a condição de biortogonalidade:

$$\langle \phi_i, \tilde{\phi}_k \rangle = \delta_{i,k} = \begin{cases} 1, & \text{se } i = k \\ 0, & \text{se } i \neq k \end{cases} \quad (4.8)$$

Claramente, quando a família de wavelets Φ constitui uma base ortogonal, então $\phi_i = \tilde{\phi}_i, \forall i$. No caso de uma base não ortogonal, os pesos $\{w_i\}$ são calculados eficientemente pelo produto interno de f com cada wavelet dual $\tilde{\phi}_i$ em $\mathbb{L}^2(\mathbb{R}^2)$:

$$w_i = \langle f, \tilde{\phi}_i \rangle \quad (4.9)$$

Assim, neste caso, a reconstrução de f é dada por:

$$f = \sum_{i=1}^N \langle f, \tilde{\phi}_i \rangle \phi_i \quad (4.10)$$

É interessante notar que a utilização de uma base dual também permite reconstruir f através da equação:

$$f = \sum_{i=1}^N \langle f, \phi_i \rangle \tilde{\phi}_i \quad (4.11)$$

Quando $f \notin \langle \Phi \rangle$, as equações acima podem ser utilizadas da mesma forma, mas a reconstrução de f não é perfeita. No entanto, o cálculo de pesos $\{w_i\}$, acima descrito, configura uma projeção ortogonal de f em $\langle \Phi \rangle$, caracterizando uma aproximação ótima da função no sentido do erro quadrado médio.

Utilizando os conceitos discutidos até o momento, podemos, agora, concentrar em nosso objetivo: determinar eficientemente os pesos $\{w_i\}$ para uma família de Gabor wavelets $\Psi = \{\psi_{n_1}, \dots, \psi_{n_M}\}$ de uma GWN.

Tendo em vista que as Gabor wavelets $\{\psi_{n_i}\}$ não são ortogonais, os pesos $\{w_i\}$ serão calculados com a utilização de uma base dual $\tilde{\Psi} = \{\tilde{\psi}_{n_1}, \dots, \tilde{\psi}_{n_M}\}$, em que $\langle \psi_{n_i}, \tilde{\psi}_{n_k} \rangle = \delta_{i,k}$.

Considerando M o número de wavelets de uma GWN, vamos definir a matriz $\mathbf{A}_{M \times M}$, cujo termo geral é dado por $A_{ij} = \langle \psi_{n_i}, \psi_{n_j} \rangle$. Essa matriz é denominada matriz de

interferência das wavelets. As wavelets duais $\tilde{\psi}_{\mathbf{n}_i}$ podem ser então definidas, conforme a equação:

$$\tilde{\psi}_{\mathbf{n}_i} = \sum_{j=1}^M (A^{-1})_{ij} \psi_{\mathbf{n}_j} \quad (4.12)$$

De fato, tal definição satisfaz a condição de biortogonalidade $\langle \psi_{\mathbf{n}_i}, \tilde{\psi}_{\mathbf{n}_k} \rangle = \delta_{i,k}$, conforme mostrado a seguir:

$$\begin{aligned} \langle \psi_{\mathbf{n}_i}, \sum_{j=1}^M (A^{-1})_{kj} \psi_{\mathbf{n}_j} \rangle &= \int_{\mathbb{R}^2} \psi_{\mathbf{n}_i}(\mathbf{x}) \left[\sum_{j=1}^M (A^{-1})_{kj} \psi_{\mathbf{n}_j}(\mathbf{x}) \right] d\mathbf{x} \\ &= \sum_{j=1}^M (A^{-1})_{kj} \left[\int_{\mathbb{R}^2} \psi_{\mathbf{n}_i}(\mathbf{x}) \psi_{\mathbf{n}_j}(\mathbf{x}) d\mathbf{x} \right] \\ &= \sum_{j=1}^M (A^{-1})_{kj} \langle \psi_{\mathbf{n}_i}, \psi_{\mathbf{n}_j} \rangle \\ &= \sum_{j=1}^M (A^{-1})_{kj} A_{ji} \\ &= \delta_{i,k} \end{aligned} \quad (4.13)$$

Finalmente, podemos computar os pesos ótimos da GWN para aproximar uma função $f \in \mathbb{L}^2(\mathbb{R}^2)$. O cálculo dos pesos $\{w_i\}$ e a aproximação $\hat{f}$ são dados pelas equações:

$$w_i = \langle f, \tilde{\psi}_{\mathbf{n}_i} \rangle \quad (4.14)$$

$$\hat{f} = \sum_{i=1}^M w_i \psi_{\mathbf{n}_i} \quad (4.15)$$

A figura 4.4(a) ilustra a imagem de uma face e a figura 4.4(c) mostra sua reconstrução com base nos pesos ótimos, considerando uma GWN com 320 wavelets. Se os pesos fossem calculados com a equação (4.6) teríamos o resultado da reconstrução ilustrado na figura 4.4(b).

Figura 4.4: (a) Imagem de uma face. (b) Reconstrução com base na equação (4.6). (c) Reconstrução com pesos ótimos.

4.3 Reparametrização de uma GWN

Mostramos, anteriormente, como uma imagem discreta de uma face pode ser representada por uma combinação linear de Gabor wavelets, com parâmetros otimizados. Agora, iremos mostrar que essa representação pode ser distorcida de forma a *casar* com uma nova imagem da face, de modo que as wavelets estejam posicionadas nas mesmas características faciais da imagem original. Esse procedimento, denominado *reparametrização de uma GWN*, é realizado através da aplicação de uma transformação afim adequada na representação wavelet da face. O rastreamento de faces, a ser descrito no próximo capítulo, está baseado nesse conceito.

Como exemplo, considere a imagem da face ilustrada na figura 4.2(a). Seja G a GWN otimizada no interior da imagem dessa face. Considere, agora, a imagem da mesma face em uma pose diferente, como ilustra a figura 4.5(a). No processo de reparametrização, o conjunto de wavelets de G é registrado corretamente nas mesmas características faciais na imagem distorcida. É importante salientar que a reparametrização de uma GWN consiste na determinação dos parâmetros (translação, rotação, escala e cisalhamento) de qualquer deformação afim aplicada à imagem original. A figura 4.5(b) ilustra a representação discreta (GWT) reparametrizada, a qual foi obtida com 52 Gabor wavelets. A figura 4.5(c) mostra a posição das 16 wavelets de G com maior peso na imagem.

A reparametrização de uma GWN em uma nova imagem, isto é, a determinação dos parâmetros afins corretos, é feita com a utilização de uma superwavelet[Szu et al., 1992]. Seja $\Psi = (\psi_{n_1}, \dots, \psi_{n_M})$, $\mathbf{w} = (w_1, \dots, w_M)$ uma GWN otimizada em uma face. Uma

Figura 4.5: (a) Imagem da face em uma pose diferente. (b) Representação wavelet reparametrizada. (c) Posição das 16 wavelets de maior peso.

Gabor superwavelet (GSW) $\Psi_{\mathbf{n}} : \mathbb{R}^2 \rightarrow \mathbb{R}$ é definida como uma combinação linear de wavelets $\psi_{\mathbf{n}_i}$ tal que:

$$\Psi_{\mathbf{n}}(\mathbf{x}) = \sum_{i=1}^M w_i \psi_{\mathbf{n}_i}(\mathbf{S}\mathbf{R}(\mathbf{x} - \mathbf{c})) \quad (4.16)$$

em que os parâmetros do vetor $\mathbf{n}$ da GSW Ψ definem a matriz de dilatação $\mathbf{S}$, a matriz de rotação $\mathbf{R}$ e o vetor de translação $\mathbf{c}$ sendo que:

$$\mathbf{S} = \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix}, \mathbf{R} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \mathbf{c} = (c_x, c_y)^T$$

Assim, a *Gabor superwavelet* $\Psi_{\mathbf{n}}$ é também uma wavelet que possui os parâmetros típicos de dilatação s_x, s_y , translação c_x, c_y e rotação θ . Dada uma nova imagem J , é possível deformar arbitrariamente a superwavelet através da otimização de seu vetor de parâmetros, de forma que a representação wavelet *case* com a face na imagem J . Este procedimento é realizado pela minimização da função energia abaixo, utilizando-se o algoritmo de Levenberg-Marquardt:

$$E = \min_{\mathbf{n}} ||J - \Psi_{\mathbf{n}}||^2 \quad (4.17)$$

É importante notar que os parâmetros de uma wavelet incluem somente translação, dilatação e rotação. No entanto, o parâmetro de cisalhamento pode também ser incluído, permitindo qualquer deformação afim da GSW $\Psi_{\mathbf{n}}$. Assim, o parâmetro s_{xy} é adicionado ao vetor $\mathbf{n}$ e a matriz de escala é modificada para:

$$\mathbf{S} = \begin{pmatrix} s_x & s_{xy} \\ 0 & s_y \end{pmatrix}$$

Considerando-se uma face com ≈ 60 pixels de largura, a inicialização dos parâmetros da superwavelet pode diferenciar dos parâmetros corretos por aproximadamente ± 10 px. nas direções x e y , 20% em escala e $\pm 10^\circ$ em rotação [Kruger and Sommer, 2000].

A reparametrização de uma GWN pode ser realizada em diferentes indivíduos, tendo em vista sua propriedade de generalização. De acordo com nossos experimentos, a reparametrização de uma GWN em uma imagem do mesmo indivíduo em que a mesma foi otimizada é sempre possível, mesmo considerando diferentes expressões faciais. No entanto, não podemos garantir que uma GWN otimizada em uma face específica possa ser reparametrizada em qualquer outra pessoa. Este problema será tratado na próxima seção, no item referente à localização de características faciais.

4.4 Características da Representação

A representação de faces baseada em GWNs apresenta diversas peculiaridades. A escolha da função de Gabor como wavelet mãe justifica-se pelo fato dos filtros de Gabor constituírem bons detectores de características em imagens [Manjunath and Chellapa, 1993], além de possibilitarem o melhor compromisso entre resolução de frequência e espacial (princípio de Heisenberg). Outro aspecto importante a ser ressaltado é que a utilização de filtros de Gabor em análise de imagens é biologicamente motivada, pois experimentos mostram que esses filtros modelam campos receptivos de células presentes no córtex visual humano [Daugman, 1985].

As principais características da representação de faces baseada em GWNs estão discriminadas abaixo:

- A representação apresenta propriedades de invariância com relação a deformações afins da face e variação homogênea das condições de iluminação. Essa segunda propriedade decorre do fato das Gabor wavelets possuírem média zero.
- O grau de precisão da representação depende do número de filtros da GWN, permitindo generalização dos dados de treinamento ou codificação de informação mais precisa da face.

- Os parâmetros das wavelets são determinados mediante um processo de otimização, de modo a representar o máximo de informação da imagem, considerando um determinado número de filtros.
- Os pesos ótimos de uma GWN estão diretamente relacionados com as respostas dos filtros de Gabor, refletindo a estrutura local da imagem.
- Os filtros de Gabor são bons detectores de características em imagens. Durante o processo de otimização, alguns filtros são atraídos por características locais da imagem, codificando informação geométrica. Os parâmetros de tais filtros tendem a assumir a mesma posição, orientação e escala das bordas da imagem.
- A representação baseada em GWNs é extremamente compacta. Por exemplo, o espaço ocupado pela representação ilustrada na figura 4.2 consiste em apenas 1040 bytes.

Outras vantagens e limitações desta representação serão apresentadas ao longo deste trabalho, particularmente nos experimentos realizados com detecção e rastreamento de faces e de características faciais.

As *Gabor wavelet networks* permitem representar qualquer objeto em uma imagem, embora o enfoque deste estudo seja restringido à representação de faces. No trabalho de [Kruger and Sommer, 2000], esta técnica foi eficazmente utilizada para rastreamento de faces, reconhecimento independente de expressão facial e estimação de pose. No próximo capítulo, iremos tratar o problema de detecção e rastreamento de faces e de características faciais em seqüências de imagens.

Capítulo 5

Detecção e Rastreamento de Faces

Conforme já mencionado, nosso objetivo consiste em, primeiramente, detectar a presença de uma face em uma sequência de vídeo específica. Mostraremos, neste capítulo, que um modelo estatístico de cor de pele pode ser utilizado com eficácia para realizar esta tarefa. Após, iremos descrever o módulo de rastreamento de faces, bem como um procedimento para localizar e rastrear características faciais do usuário, utilizando GWNs.

5.1 Detecção de Faces

Tendo em vista os bons resultados obtidos pelos sistemas de detecção de faces baseados em cor, optamos pela escolha de tal abordagem. Basicamente, nosso método pode ser dividido em duas etapas: geração de um modelo estatístico de cor da pele e segmentação da face com base no modelo gerado.

5.1.1 Modelo Estatístico de Cor da Pele

A primeira etapa do método consiste em um treinamento supervisionado a partir de amostras de cor da pele humana. Um banco de 40 imagens de faces coloridas foi utilizado para realizar o treinamento. Tais imagens foram obtidas de pessoas de diferentes raças, idades e sexo, com variação das condições de iluminação. A figura 5.1 ilustra o processo de treinamento, em que uma região de cor da pele é selecionada, e sua representação

Figura 5.1: Amostra de cor da pele.

Figura 5.2: (a) Agrupamento de cores no espaço RGB. (b) Agrupamento de cores no espaço cromático.

RGB armazenada. A partir dos dados de treinamento, verificamos que as cores da pele humana formam um agrupamento (*cluster*) no espaço de cores, conforme ilustrado na figura 5.2(a).

É comum pensarmos que estas cores diferem significativamente de pessoa para pessoa, tendo em vista a diversidade de raças existentes. No entanto, o que realmente ocorre é uma maior diferença de brilho/intensidade e não de cor[Yang et al., 1997]. Assim, é possível reduzir a variância do agrupamento de cores da pele através da normalização de intensidade: $a = R/(R + G + B)$, $b = G/(R + G + B)$. As cores (a, b) são conhecidas como cores cromáticas, ou cores “puras”. A figura 5.2(b) ilustra o agrupamento de cores no espaço cromático.

De acordo com [Yang et al., 1997], a distribuição de cores da pele no espaço cromático pode ser eficazmente aproximada por um modelo Gaussiano. Assim, a partir dos dados

Figura 5.3: (a) Imagem de entrada. (b) Imagem obtida a partir do modelo de cor da pele.

de treinamento, podemos determinar os parâmetros -média e matriz de covariância - de uma Gaussiana bivariada, configurando o modelo de cor da pele.

5.1.2 Segmentação da Face

O procedimento de segmentação consiste em identificar as regiões (*blobs*) de cor da pele em uma imagem de entrada, com base no modelo gerado. Inicialmente, todos os pixels da imagem são convertidos para o espaço cromático. Mediante comparação com o modelo Gaussiano, uma nova imagem, em níveis de cinza, é gerada, sendo que a intensidade de cada pixel nesta nova imagem representa sua probabilidade de possuir cor da pele. A figura 5.3 ilustra este processo.

Para identificar os *blobs* de cor da pele, a imagem em níveis de cinza é convertida em uma imagem binária, com o ajuste de um valor de limiar (*threshold*) determinado experimentalmente.

Deve-se mencionar que essa técnica constitui um caso particular da teoria de decisão Bayesiana. De acordo com a regra de Bayes, um *pixel* é classificado com cor de pele se:

$$\begin{aligned}
 P(C_1)f(h|C_1) &\geq P(C_2)f(h|C_2) \\
 \text{ou} \\
 f(h|C_1) &\geq \frac{P(C_2)f(h|C_2)}{P(C_1)}
 \end{aligned} \tag{5.1}$$

onde C_1 é a classe “pele”, C_2 é a classe “não pele”, $P(C_1)$ e $P(C_2)$ são as probabilidades *a priori* dessas classes e $f(h|C_1)$ e $f(h|C_2)$ são as funções de densidade de probabilidade

Figura 5.4: Exemplo de modelos de cor de pele e não-pele no caso unidimensional. Note que a distribuição de probabilidade de cor não-pele é uma constante, dada por um valor de limiar.

Figura 5.5: (a) Imagem resultante da limiarização. (b) Aplicação do operador de fechamento e filtro da mediana.

condicional da variável aleatória h , dado que o *pixel* pertence à classe C_1 e C_2 , respectivamente.

Em nosso caso, na equação (5.1), $f(h|C_1)$ corresponde ao modelo Gaussiano de cor da pele, enquanto que $P(C_2)f(h|C_2)/P(C_1)$ corresponde ao valor de limiar determinado experimentalmente. A figura 5.4 ilustra esse conceito.

Em geral, devido a ruídos e distorções na imagem de entrada, o resultado da limiarização pode gerar regiões de cor da pele particionadas (componentes não-conexos), além de permitir a existência de *pixels* de cor da pele isolados na imagem binária. Para solucionar esses problemas, utilizamos uma operação morfológica de fechamento seguida da aplicação do filtro da mediana. A figura 5.5 ilustra este processo.

A identificação dos limites dos *blobs* de cor da pele é então realizada através do algoritmo de crescimento de regiões. Regiões menores que 1% do tamanho da imagem são

descartadas. Além disso, são levados em consideração aspectos estruturais da face, eliminando regiões que não apresentam a estrutura de uma face. Isto é feito de uma forma simples, considerando a relação entre a altura e a largura do retângulo que define a região de cor da pele.

É importante salientar que as regiões determinadas podem corresponder a objetos com cor da pele diferentes de uma face. Logo, a verificação da presença ou ausência de uma face em tais regiões faz-se necessária. Isto é feito através de um procedimento simples de correlação que visa a localizar olhos nas regiões de cor da pele[Feris et al., 2000a]. Caso existam várias faces na imagem, aquela que estiver mais próxima da câmera será segmentada.

Outros experimentos com relação à verificação da presença ou ausência de uma face em regiões de cor da pele foram realizados utilizando descritores de Fourier (vide artigo em anexo) [Campos et al., 2000b]. No entanto, para obtermos um sistema realmente robusto, a integração de várias outras informações (movimento, simetria, etc.) deve ser realizada.

5.2 Rastreamento de Faces Utilizando GWNs

Após o procedimento de detecção, a face segmentada é acompanhada nas imagens subsequentes, ao longo da seqüência de vídeo. Este processo de rastreamento é realizado com a técnica baseada em GWNs, sendo robusto a transformações afins da face e mudanças homogêneas da iluminação.

Inicialmente, a representação wavelet é obtida para o *blob* de cor da pele determinado. A reparametrização da GWN pode ser então aplicada à seqüência de imagens, caracterizando o rastreamento da face. Assim, para cada quadro J_t referente ao instante t , os parâmetros $\mathbf{n}$ da superwavelet $\Psi_{\mathbf{n}_t}$ são otimizados de acordo com a função energia abaixo:

$$E = \min_{\mathbf{n}_t} ||J_t - \Psi_{\mathbf{n}_t}||^2 \quad (5.2)$$

O vetor de parâmetros $\mathbf{n}_{t-1}$ é utilizado como valor inicial para a otimização no quadro J_t . Como as mudanças na imagem são pequenas de quadro para quadro, o processo de otimização converge rapidamente. Os valores iniciais para $\mathbf{n}_0$ no primeiro quadro são derivados a partir do *blob* de cor da pele.

O modelo adotado considera a face como um objeto planar, observado sob projeção ortográfica. Em consonância com o trabalho de [Hager and Belhumeur, 1998], verificamos que embora a face tenha sido modelada como um objeto rígido, com movimentos limitados, o algoritmo é capaz de rastrear o usuário ainda que este mude sua expressão ou realize rotações leves fora do plano.

5.3 Localização e Rastreamento de Características Faciais

A localização e rastreamento de características faciais constitui um tópico de pesquisa clássico em visão computacional. Nosso interesse neste problema reside no fato de que o reconhecimento de uma face pode ser realizado eficientemente utilizando apenas suas características faciais, como por exemplo, olhos, nariz e boca. Existem evidências biológicas[Moghaddam and Pentland, 1994] de que os olhos cumprem o papel mais importante no reconhecimento de uma face. O trabalho de [Campos et al., 2000a] demonstra a boa performance de um sistema de reconhecimento baseado apenas na região dos olhos, ao ser comparado com outro sistema que considera a face inteira. Além disso, trabalhos como [Moghaddam and Pentland, 1994, Brunelli and Poggio, 1993] mostram que o reconhecimento baseado em características faciais pode alcançar melhores resultados, tendo em vista que esta abordagem torna-se invariante a cortes de cabelo, penteados, maquiagem, etc.

Neste contexto, iremos apresentar um método baseado na reparametrização de uma GWN para localização e rastreamento de características faciais. Nossos experimentos referentes ao procedimento de localização foram conduzidos no banco de faces Yale, o qual consiste de 15 diferentes indivíduos, com 11 imagens por pessoa, apresentando diferentes expressões faciais (<http://cvc.yale.edu/projects/yalefaces/yalefaces.html>). Como este banco de dados contém apenas imagens em níveis de cinza, a região da face foi segmentada manualmente. Experimentos com imagens coloridas e detecção automática da face serão apresentados no próximo capítulo.

No capítulo anterior, foi mencionado que uma GWN de pouca precisão, isto é, com

um pequeno número de funções base, pode ser reparametrizada em diferentes indivíduos. No entanto, esta reparametrização não pode ser garantida em qualquer pessoa, tendo em vista que a mesma depende da similaridade entre a face de teste e a face em que a GWN foi otimizada, bem como do número de wavelets consideradas.

A solução que propomos para este problema consiste em otimizar a GWN em uma face média. Com este objetivo, foi selecionado um conjunto de 15 faces do banco de faces Yale, cada qual correspondendo a um indivíduo específico. As faces deste conjunto foram então registradas manualmente e a média calculada. A figura 5.6 ilustra a face média obtida, juntamente com sua representação wavelet, considerando apenas 52 funções base.

Utilizando a GWN otimizada nessa face média, a reparametrização funcionou perfeitamente em todas imagens dos indivíduos presentes no banco de faces Yale. No entanto, é provável que esta técnica falhe ao considerarmos bancos de faces com um número bem maior de indivíduos.

Com vistas a realizar a localização de características faciais, pontos característicos são inicialmente localizados na face média. Os pontos que consideramos correspondem às pupilas, centro do nariz e centro da boca. Após, a GWN otimizada na face média é reparametrizada na imagem da face de teste. Nesta imagem, os pontos característicos são determinados mediante a aplicação de uma transformação afim adequada aos pontos originalmente marcados na face média. Os parâmetros corretos desta transformação são obtidos a partir do vetor de parâmetros da superwavelet $(s_x, s_y, s_{xy}, c_x, c_y, \theta)$, o qual é determinado pela equação (5.2) no processo de reparametrização.

A figura 5.7 ilustra a localização de características faciais em três indivíduos do banco de faces Yale. É interessante notar que o método é robusto à presença de óculos e barba na imagem da face. A figura 5.8 ilustra os resultados considerando imagens do mesmo indivíduo sob diferentes expressões faciais e mudanças de iluminação.

O rastreamento das características faciais é realizado mediante a aplicação de uma transformação afim aos pontos localizados, em cada quadro da sequência¹. Claramente,

¹É importante salientar que não processamos diretamente os olhos, nariz e a boca do usuário; apenas obtemos uma estimação da posição dessas características. Ainda assim, optamos pelo uso do termo “rastreamento de características faciais” para denotar o acompanhamento dos quatro pontos faciais de interesse no modelo wavelet.

Figura 5.6: Face média e sua representação wavelet.

Figura 5.7: Localização de características faciais em diferentes indivíduos.

os parâmetros corretos de cada transformação são derivados do vetor de parâmetros da superwavelet, em cada quadro. A figura 5.9 ilustra o rastreamento de características faciais em alguns quadros de uma seqüência de imagens.

Esse procedimento considera a geometria global da face, sendo robusto a deformações das características faciais, tais como piscar de olhos e sorriso, as quais constituem, em geral, uma situação crítica para a maioria das abordagens locais. Neste sentido, o método não exige a condição de alta correlação entre quadros adjacentes, como é exigido em técnicas baseadas em casamento de padrões (*template matching*), por exemplo.

É importante notar que o método pode ser facilmente generalizado para rastrear pontos característicos adicionais ou até mesmo regiões, como polígonos arbitrários ao redor dos olhos, nariz e boca. No capítulo 7, iremos apresentar alguns experimentos relacionados com esta abordagem.

Figura 5.8: Localização de características faciais sob diferentes expressões faciais e mudanças de iluminação.

Figura 5.9: Rastreamento de características faciais utilizando GWNs.

Capítulo 6

Rastreamento de Faces em um Subespaço Wavelet

Neste capítulo, iremos apresentar, como principal contribuição científica deste trabalho, um método eficiente para rastreamento de faces baseado em GWNs. Basicamente, nosso objetivo consiste em realizar o rastreamento da face em um subespaço wavelet de dimensionalidade reduzida, visando a uma maior eficiência do processo.

Inicialmente, vamos considerar uma GWN $(\Psi, \mathbf{w})$ com $\Psi = \{\psi_{\mathbf{n}_1}, \dots, \psi_{\mathbf{n}_N}\}$ e $\mathbf{w} = (w_1, \dots, w_N)$. Conforme mencionado anteriormente, essa representação pode ser interpretada como uma superwavelet $\Psi_{\mathbf{n}}(\mathbf{x}) = \sum_{i=1}^N w_i \psi_{\mathbf{n}_i}(\mathbf{S}\mathbf{R}(\mathbf{x} - \mathbf{c}))$, onde N é o número de wavelets utilizadas na representação e $\mathbf{n} = (c_x, c_y, \theta, s_x, s_y, s_{xy})$ é o vetor de parâmetros da superwavelet, o qual define a matriz de dilatação $\mathbf{S}$, a matriz de rotação $\mathbf{R}$ e o vetor de translação $\mathbf{c}$. Como visto na seção 5.2, o rastreamento da face pode ser obtido mediante a deformação afim da superwavelet $\Psi_{\mathbf{n}}$ em cada quadro da seqüência, conforme mostra a equação (5.2).

Claramente, esse método realiza um típico casamento de padrões (*template matching*) no espaço de imagens, em que o padrão, dado pela imagem obtida a partir da superwavelet, é distorcido gradualmente até *casar* com a face no quadro atual da seqüência. É interessante notar que os pesos w_i das wavelets em $\Psi_{\mathbf{n}}$ são constantes nesse processo. Assim, a deformação afim é capturada apenas pela deformação das wavelets, ao passo que o vetor de pesos permanece invariante.

Figura 6.1: Mapeamentos entre o espaço de imagens $\mathbb{L}^2(\mathbb{R}^2)$ e o subespaço wavelet $\mathbb{R}^N$.

Com essa propriedade, concluímos que o rastreamento no espaço de imagens, acima descrito, pode também ser realizado com base apenas nos pesos das wavelets, isto é, em um espaço de dimensionalidade reduzida $\mathbb{R}^N$, o qual denominaremos *subespaço wavelet*.

De acordo com a seção 4.2, sabemos que uma imagem $g \in \mathbb{L}^2(\mathbb{R}^2)$ pode ser mapeada em um vetor $\mathbf{w} \in \mathbb{R}^N$ através do conjunto de wavelets duais $\tilde{\Psi} = \{\tilde{\psi}_{n_1}, \dots, \tilde{\psi}_{n_N}\}$. Tal mapeamento corresponde à projeção ortogonal de $g \in \mathbb{L}^2(\mathbb{R}^2)$ no subespaço $\langle \Psi \rangle \subset \mathbb{L}^2(\mathbb{R}^2)$. Analogamente, a reconstrução $\hat{g} \in \mathbb{L}^2(\mathbb{R}^2)$, isto é, o mapeamento de $\mathbf{w}$ em $\mathbb{L}^2(\mathbb{R}^2)$, é obtido através do conjunto de wavelets Ψ . A figura 6.1 ilustra com clareza esses mapeamentos. Os espaços vetoriais envolvidos estão discriminados abaixo:

- Espaço de imagens $\mathbb{L}^2(\mathbb{R}^2)$, no caso contínuo. Uma imagem $g \in \mathbb{L}^2(\mathbb{R}^2)$ define um ponto nesse espaço.
- Espaço $\langle \Psi \rangle$, cuja base é definida pelas wavelets da representação. A reconstrução $\hat{g} = \sum_{i=1}^N w_i \psi_{n_i}$ define um ponto nesse espaço. Claramente, $\langle \Psi \rangle$ constitui um subespaço do espaço de imagens.
- Subespaço wavelet $\mathbb{R}^N$. O vetor de pesos $\mathbf{w}$, obtido pela projeção ortogonal de g em $\langle \Psi \rangle$, define um ponto nesse espaço.

Esta discussão é perfeitamente válida para o caso de imagens discretas, as quais serão tratadas a seguir.

6.1 Metodologia

Para descrever o rastreamento no espaço $\mathbb{R}^N$, vamos considerar, inicialmente, uma GWN $(\Psi, \mathbf{v})$ otimizada na imagem de uma face, a qual desejamos rastrear em uma seqüência de vídeo. Como já mencionado, o vetor de pesos $\mathbf{v}$ é obtido através da projeção ortogonal da imagem da face no subespaço $\langle \Psi \rangle$. Deste modo, temos o padrão da face mapeado no conjunto de pesos $\mathbf{v} \in \mathbb{R}^N$, os quais denominaremos *pesos de referência*.

Basicamente, o rastreamento no subespaço wavelet é realizado com a deformação afim do subespaço $\langle \Psi \rangle$, sendo aplicada até que o vetor de pesos $\mathbf{w} \in \mathbb{R}^N$, obtido pelo mapeamento ortogonal do quadro atual nesse subespaço, seja o mais próximo do vetor de pesos de referência $\mathbf{v}$. Esse procedimento assemelha-se com o método de rastreamento no espaço de imagens, com a diferença de que o processamento é realizado eficientemente no subespaço $\mathbb{R}^N$.

O mapeamento de imagens no $\mathbb{R}^N$ é feito com um baixo custo computacional, através de um pequeno número de filtragens locais com as wavelets. De acordo com a seção 4.2, considerando uma imagem I , temos que $w_i = \langle I, \tilde{\psi}_{\mathbf{n}_i} \rangle$, em que $\tilde{\psi}_{\mathbf{n}_i} = \sum_j (A^{-1})_{ij} \psi_{\mathbf{n}_j}$. Simples transformações conduzem à equação abaixo:

$$w_i = \sum_j (A^{-1})_{ij} \langle I, \psi_{\mathbf{n}_j} \rangle \quad (6.1)$$

Assim, os pesos ótimos w_i são obtidos a partir de uma combinação linear de filtragens com as wavelets, onde os coeficientes são dados pela inversa da matriz $A_{ij} = \langle \psi_{\mathbf{n}_i}, \psi_{\mathbf{n}_j} \rangle$. É interessante notar que a matriz A_{ij} é, exceto por um fator de escala, invariante com relação a transformações afins da GWN¹. Isto permite que tal matriz seja calculada previamente, otimizando o cálculo dos pesos w_i , os quais são calculados eficientemente com a equação (6.1), através de uma aplicação local $\langle I, \psi_{\mathbf{n}_i} \rangle$ de cada uma das N Gabor wavelets $\psi_{\mathbf{n}_i}$, seguida por uma multiplicação com a inversa da matrix A_{ij} .

Seja $\mathbf{n} = (c_x, c_y, \theta, s_x, s_y, s_{xy})$ um vetor de parâmetros afim que configura uma parametrização para o subespaço $\langle \Psi \rangle$. Conforme descrito acima, o rastreamento no subespaço wavelet é obtido pela mudança gradual desses parâmetros até que o mapeamento do qua-

¹Os produtos internos $\langle \psi_{\mathbf{n}_i}, \psi_{\mathbf{n}_j} \rangle$ são alterados apenas por um fator de escala quando $\mathbf{n}_i$ e $\mathbf{n}_j$ são modificados pela mesma transformação afim.

dro atual J em $\mathbf{w} \in \mathbb{R}^N$ seja o mais próximo dos pesos de referência $\mathbf{v}$. Em outras palavras, devemos otimizar os parâmetros $\mathbf{n}$ com relação à função energia:

$$E = \min_{\mathbf{n}} \|\mathbf{v} - \mathbf{w}\|_{\Psi} \quad (6.2)$$

$$\text{com } w_i = \sum_j \frac{1}{s_x \cdot s_y} (A^{-1})_{ij} < J, \psi_{\mathbf{n}_j}(\mathbf{SR}(\mathbf{x} - \mathbf{c})) > \quad (6.3)$$

onde $\mathbf{S}$ é a matriz de dilatação, $\mathbf{R}$ é a matriz de rotação e $\mathbf{c}$ é o vetor de translação, todos definidos através do vetor $(c_x, c_y, \theta, s_x, s_y, s_{xy})$. Durante o rastreamento, essa otimização é feita para cada quadro da seqüência. Como existe pouca diferença entre dois quadros adjacentes, a otimização é rápida. Novamente, para minimizar a função energia (6.2), o algoritmo Levenberg-Marquardt foi utilizado.

6.2 Distância entre Pesos

Na equação (6.2), utilizamos a notação $\|\mathbf{v} - \mathbf{w}\|_{\Psi}$ para denotar a distância entre os vetores $\mathbf{v}$ e $\mathbf{w}$ no subespaço $\mathbb{R}^N$. No entanto, ainda não é claro como essa medida de distância deve ser calculada. De acordo com o trabalho de [Wiskott et al., 1997], a distância Euclidiana poderia ser utilizada entre esses dois vetores, mas tal medida trata igualmente as wavelets da representação, ainda que essas funções apresentem diferentes escalas.

Nós então propomos uma nova medida de distância, a qual é baseada na distância Euclidiana entre vetores do espaço de imagens. Assim, definimos a diferença $\|\mathbf{v} - \mathbf{w}\|_{\Psi}$ como a distância Euclidiana entre as duas respectivas imagens reconstruídas:

$$\|\mathbf{v} - \mathbf{w}\|_{\Psi} = \left\| \sum_{i=1}^N v_i \psi_{\mathbf{n}_i} - \sum_{j=1}^N w_j \psi_{\mathbf{n}_j} \right\|_2 \quad (6.4)$$

Várias transformações são então realizadas:

$$\begin{aligned} \|\mathbf{v} - \mathbf{w}\|_{\Psi} &= \left\| \sum_{i=1}^N v_i \psi_{\mathbf{n}_i} - \sum_{j=1}^N w_j \psi_{\mathbf{n}_j} \right\|_2 \\ &= \left[\int \left(\sum_{i=1}^N v_i \psi_{\mathbf{n}_i}(\mathbf{x}) - \sum_{j=1}^N w_j \psi_{\mathbf{n}_j}(\mathbf{x}) \right)^2 d\mathbf{x} \right]^{\frac{1}{2}} \end{aligned} \quad (6.5)$$

$$\begin{aligned}
&= \left[\int \left(\sum_{i=1}^N \delta_i \psi_{\mathbf{n}_i}(\mathbf{x}) \right)^2 d\mathbf{x} \right]^{\frac{1}{2}} \quad \text{com } \delta_i = (v_i - w_i) \\
&= \left[\int \left(\sum_{i=1}^N \sum_{j=1}^N \delta_i \delta_j \psi_{\mathbf{n}_i}(\mathbf{x}) \psi_{\mathbf{n}_j}(\mathbf{x}) \right) d\mathbf{x} \right]^{\frac{1}{2}} \\
&= \left[\sum_{i=1}^N \sum_{j=1}^N \delta_i \delta_j \int \psi_{\mathbf{n}_i}(\mathbf{x}) \psi_{\mathbf{n}_j}(\mathbf{x}) d\mathbf{x} \right]^{\frac{1}{2}} \\
&= \left[\sum_{i,j} \delta_i \delta_j \langle \psi_{\mathbf{n}_i}, \psi_{\mathbf{n}_j} \rangle \right]^{\frac{1}{2}} \\
&= [\delta^T A_{ij} \delta]^{\frac{1}{2}} \quad \text{com } \delta = (\delta_1 \cdots \delta_N)^T
\end{aligned} \tag{6.6}$$

Claramente, a medida de distância é caracterizada no espaço $\mathbb{R}^N$, sendo calculada eficientemente. A matriz de produtos escalares $A_{ij} = \langle \psi_{\mathbf{n}_i}, \psi_{\mathbf{n}_j} \rangle$ é a mesma matriz presente na eq. (6.1) e na seção 4.2. É interessante notar que, no caso das wavelets $\{\psi_{\mathbf{n}_i}\}$ serem ortonormais, A_{ij} é a matriz unitária e a equação (6.5) descreve a mesma medida de distância proposta em [Wiskott et al., 1997].

6.3 Algoritmo

Com o intuito de tornar mais clara a seqüência de etapas envolvidas no método de rastreamento de faces em um subespaço wavelet, apresentamos, a seguir, um algoritmo em alto nível do referido método. Por questões de simplicidade, o algoritmo considera a presença de uma face no primeiro quadro de uma seqüência de imagens, supondo-se conhecida a sua posição, orientação e escala.

Dada uma seqüência de imagens $\{J_1, J_2, \dots, J_N\}$:

→ Otimize a GWN $(\Psi, \mathbf{v})$ na face da imagem J_1 , obtendo, desta forma, os pesos de referência $\mathbf{v}$.

→ Inicialize o vetor de parâmetros $\mathbf{n}_1 = (c_x, c_y, \theta, s_x, s_y, s_{xy})$ com a posição, escala e orientação da face na imagem J_1 . Isto configura uma parametrização inicial do subespaço $\langle \Psi \rangle$.

Para i de 2 até N faça

→ Calcule $\mathbf{w}$, projetando J_i em $\langle \Psi \rangle$ com a equação (6.3).

Enquanto $\|\mathbf{v} - \mathbf{w}\|_\Psi$ não é mínimo faça

→ Ajuste o vetor de parâmetros $\mathbf{n}_i$ (deformação do subespaço $\langle \Psi \rangle$) com o método Levenberg-Marquardt.

→ Calcule $\mathbf{w}$, projetando J_i em $\langle \Psi \rangle$ com a equação (6.3).

Fim Enquanto

→ Imprima o vetor $\mathbf{n}_i$: posição, escala e orientação da face no quadro J_i .

Fim Para

A técnica proposta apresenta um considerável aumento de performance quando comparada ao rastreamento no espaço de imagens. Essa vantagem, em termos de eficiência, deve-se ao fato de que o rastreamento é realizado em um subespaço wavelet de dimensionalidade reduzida, promovendo uma alta redução de dados. Além disso, deve-se notar que a reconstrução do padrão da face, bem como o cálculo da diferença *pixel a pixel* entre imagens, são fatores críticos em termos de tempo de processamento, dispensados no rastreamento baseado nos pesos das wavelets. O próximo capítulo apresentará uma avaliação quantitativa dessa técnica.

Capítulo 7

Resultados Experimentais

Neste capítulo, iremos descrever os experimentos realizados com as técnicas introduzidas neste trabalho, bem como discutir os resultados obtidos. Inicialmente, apresentaremos uma avaliação quantitativa do método de rastreamento de faces proposto no capítulo 6, considerando aspectos como atenção progressiva, robustez e eficiência. Logo após, os conceitos apresentados nessa dissertação serão reunidos e aplicados no problema de reconhecimento de faces a partir de seqüências de vídeo.

7.1 Avaliação da Técnica de Rastreamento de Faces

O método proposto para rastreamento de faces em um subespaço wavelet foi testado com sucesso em várias seqüências de imagens. Todas as seqüências de teste apresentavam um indivíduo em movimento, mais ou menos frontal à câmera, de forma que suas características faciais estivessem sempre visíveis. Os experimentos foram conduzidos visando a analisar a performance do rastreamento com relação a fatores como precisão, robustez e eficiência.

7.1.1 Análise de Precisão

A precisão do método de rastreamento foi avaliada considerando o conceito de atenção progressiva, isto é, variando o número de wavelets na representação. Inicialmente, grava-

Figura 7.1: Rastreamento de faces no subespaço wavelet, utilizando-se uma GWN com 116 wavelets.

Figura 7.2: Representações obtidas com 116, 51, 22 e 9 wavelets.

mos uma seqüência de imagens de uma pessoa se movendo diante da câmera com diferentes expressões faciais. No interior da face desse indivíduo, uma GWN com 116 Gabor wavelets foi otimizada e utilizada para estimar os parâmetros afins de referência, em cada quadro da seqüência. A figura 7.1 ilustra alguns quadros do resultado da aplicação do novo método de rastreamento de faces nessa seqüência. É interessante notar a robustez do método com relação a deformações afins da face e a variações de expressão facial.

Com o objetivo de investigar o princípio da atenção progressiva no rastreamento da face, foram extraídas GWNs de menor precisão, a partir da representação de 116 wavelets, contendo somente as 51, 22 e 9 wavelets de maior peso, normalizadas com relação à escala. A figura 7.2 ilustra tais representações. Basicamente, queremos utilizar essas GWNs para analisar o decréscimo de precisão ao diminuir o número de wavelets na representação.

Assim, o rastreamento da face no subespaço wavelet foi realizado com cada uma das

GWNs citadas acima. As figuras 7.3, 7.4 e 7.5 ilustram a estimação dos parâmetros de posição c_x , c_y e o ângulo θ da face, respectivamente. Em cada gráfico, os parâmetros de referência são denotados pela linha pontilhada, ao passo que a linha contínua refere-se aos parâmetros estimados com 51, 22 e 9 wavelets. Os resultados obtidos com o rastreamento baseado no espaço de imagens são similares.

Mediante a análise visual desses gráficos, concluímos que a precisão do rastreamento depende diretamente do número de Gabor wavelets utilizadas na representação. De fato, a medida que diminuimos o número de wavelets, temos um decréscimo na precisão da estimação dos parâmetros em cada quadro da sequência. Por outro lado, temos um correspondente aumento de eficiência no rastreamento, como veremos adiante. Logo, a escolha do número de filtros depende da aplicação considerada, podendo ser ajustada de acordo com os recursos computacionais disponíveis.

7.1.2 Análise de Robustez

De acordo com nossos experimentos, o rastreamento no subespaço wavelet apresenta a mesma robustez do rastreamento baseado no espaço de imagens. Ambas abordagens são robustas à variação homogênea das condições de iluminação, bem como a deformações afins da imagem da face. É interessante notar que variações de expressão facial e rotações leves da face fora do plano são, em geral, corretamente aproximadas pelo modelo afim baseado em wavelets.

Os testes realizados mostraram, também, a robustez do método com relação à velocidade da face. Em outras palavras, o rastreamento é realizado mesmo na presença de considerável diferença entre dois quadros adjacentes.

O gráfico ilustrado na figura 7.6 mostra os deslocamentos (em *pixels*) da posição da face na sequência considerada. Tais deslocamentos foram calculados, em cada quadro, através dos parâmetros de referência c_x e c_y . Basicamente, considerando (c_x, c_y) a posição da face no quadro atual e (c_{xold}, c_{yold}) a posição da face no quadro anterior, o deslocamento d da face para o quadro atual foi calculado conforme a equação:

$$d = \sqrt{(c_x - c_{xold})^2 + (c_y - c_{yold})^2} \quad (7.1)$$

Observando o gráfico, percebemos que o deslocamento da face é relativamente alto em

Figura 7.3: Estimação do parâmetro de posição c_x em cada quadro da sequência.

Figura 7.4: Estimação do parâmetro de posição c_y em cada quadro da sequência.

Figura 7.5: Estimação do parâmetro referente ao ângulo da face em cada quadro da sequência.

Figura 7.6: Deslocamento da face em cada quadro da sequência.

determinados quadros. Quando isso ocorre, o rastreamento é realizado com robustez, mas o tempo requerido pelo procedimento de reparametrização é aumentado, devido ao maior número de ciclos no processo de otimização.

7.1.3 Análise de Eficiência

A resolução das imagens mostradas na figura 7.1 é 160×120 *pixels* e o tamanho da face na qual a GWN foi otimizada é 50×65 *pixels*. Utilizando-se somente 9 wavelets, o tempo de computação para cada ciclo do algoritmo Levenberg-Marquardt é de 15ms, em um Athlon 1GHz rodando Linux. Uma performance ainda mais alta pode ser alcançada se considerarmos imagens da face de menor resolução, ou mesmo, considerando menos parâmetros (e.g. somente translação e rotação).

É importante salientar que o número de ciclos para cada quadro depende da distância entre os parâmetros iniciais e o mínimo local, sendo diretamente relacionado com a velocidade da face. O aumento do número de wavelets na representação promove um rastreamento mais preciso, mas também mais lento. Por exemplo, utilizando-se 51 wavelets, foi exigido um tempo de computação de 85ms por ciclo, em cada quadro da sequência considerada.

Em comparação com o rastreamento baseado no espaço de imagens, observamos, na

Figura 7.7: Comparação dos tempos exigidos por ciclo entre o rastreamento no espaço de imagens e o método proposto.

técnica proposta, uma economia de aproximadamente 50% do custo computacional exigido em cada ciclo do método Levenberg-Marquardt, com um pequeno aumento no número de ciclos requeridos. Acreditamos que a utilização de outros algoritmos, como os métodos Condensation[Isard and Blake, 1998] ou o *Sequential Importance Sampling* (SIS)[Li, 2000] possivelmente guiarão para um aumento ainda maior de eficiência. O gráfico ilustrado na figura 7.7 mostra a comparação dos tempos exigidos por ciclo entre o rastreamento no espaço de imagens e o método proposto.

Deve-se mencionar que o cálculo dos tempos de computação dependem sobremaneira do *hardware* e do sistema operacional utilizado. Além disso, os resultados foram apresentados considerando uma sequência de vídeo específica. A despeito da obtenção de resultados similares em outras sequências de teste, torna-se difícil traçar uma conclusão genérica.

7.2 Aplicação: Reconhecimento de Faces a partir de Vídeo

Os conceitos apresentados neste trabalho foram reunidos visando uma aplicação específica: reconhecimento de faces a partir de seqüências de vídeo. Basicamente, um sistema foi configurado com as seguintes etapas:

- Detecção da face;
- Posicionamento das características faciais;
- Rastreamento da face e das características faciais;
- Seleção e normalização dos melhores quadros.

As características faciais normalizadas são então enviadas ao módulo de reconhecimento, incumbido da identificação do usuário. O esquema desse método proposto para reconhecer pessoas a partir de vídeo está descrito no artigo [Feris et al., 2000b] em anexo.

7.2.1 Formação do Banco de Dados

Para realizar os testes do nosso sistema, integrado com o módulo de reconhecimento, um banco de seqüências de vídeo foi gerado, solicitando-se a colaboração de 10 pessoas do departamento. As imagens, bem como os experimentos realizados, serão armazenados em um CD, com a devida autorização das pessoas envolvidas.

Para cada pessoa, duas seqüências, de treinamento e de teste, foram gravadas em tempos distintos, com um intervalo de uma semana. É interessante notar as mudanças na aparência do indivíduo, após esse intervalo, como diferentes penteados, presença ou ausência de barba, óculos, etc.

Inicialmente, solicitamos que cada pessoa se posicionasse na frente de um computador, com o intuito de ser filmada por uma câmera colorida. Assumimos que a imagem da face do usuário seja frontal, para facilitar o processo de detecção. Em seguida, pedimos a cada pessoa para mover-se naturalmente diante da câmera, configurando um processo

de interação com a máquina. Nesse processo, as características faciais - olhos, nariz e boca - devem estar presentes em todos os quadros da sequência.

Ao final, um total de 20 sequências de imagens foram gravadas. Deve-se mencionar que este banco de dados está sendo expandido, de forma que uma avaliação mais completa e precisa será apresentada em trabalho futuro.

A seguir, discutiremos os resultados obtidos com cada módulo do sistema, no contexto da aplicação considerada.

7.2.2 Detecção da Face

Com relação à detecção de faces, o método mostrou-se robusto ao identificar regiões de cor da pele de pessoas de diferentes raças, idades e sexo. Entretanto, cores de pele escura, quando comparadas com cores de pele clara, corresponderam a pixels de menor intensidade na imagem em níveis de cinza. Isto se deve ao fato de que, no treinamento supervisionado, utilizou-se um maior número de cores de pele clara. Ainda assim, as cores de pele escura são classificadas corretamente quando da identificação de regiões de cor da pele. Em geral, o método falha sob alta variação das condições de iluminação ou quando a face está muito próxima de outros objetos com cor da pele.

A verificação da presença ou ausência da face na região de cor de pele foi realizada, como já comentado, utilizando-se uma técnica simples de correlação. Visando a aumentar a robustez desse módulo, pretendemos ainda utilizar uma combinação de outras informações, como movimento, simetria, profundidade, forma, e outras.

7.2.3 Posicionamento das Características Faciais

Uma vez que a face tenha sido detectada, sua informação de escala é obtida e a região de cor da pele é convertida para uma imagem em níveis de cinza. As características faciais são então localizadas através do processo de reparametrização da GWN, otimizada em uma face média. A posição e a escala do *blob* da face são utilizadas como parâmetros iniciais na reparametrização. A figura 7.8 ilustra a detecção de uma face no primeiro quadro de uma sequência de imagens, bem como a localização de suas características faciais.

Figura 7.8: Detecção da face e localização das características faciais.

Consideramos que o método utilizado para o posicionamento das características faciais apresenta duas limitações. A primeira está relacionada com a reparametrização da GWN, visto que esse procedimento é adequado apenas quando bancos de dados com poucos indivíduos são tratados. A segunda limitação é derivada do fato de que a distância entre características faciais varia de pessoa para pessoa, ocasionando, em alguns casos, resultados imprecisos. Assim, seria interessante dispormos de um procedimento incumbido de ajustar a posição dos pontos característicos após a reparametrização da GWN. Como exemplo, nosso método poderia ser utilizado para a inicialização de contornos ativos [Blake and Isard, 1998], objetivando uma segmentação precisa das características faciais. Atualmente, uma aluna de doutorado do nosso grupo de pesquisa está investigando novos métodos para solucionar esse problema.

7.2.4 Rastreamento da Face e das Características Faciais

O rastreamento da face e das características faciais foi realizado corretamente nas seqüências de imagens consideradas, mostrando-se robusto a deformações afins da face, diferentes expressões faciais e mudanças homogêneas das condições de iluminação. A figura 7.9 ilustra o rastreamento de olhos, nariz e boca de uma seqüência específica, apresentando os quadros 60, 98 e 221.

De acordo com nossos experimentos, o rastreamento é realizado com maior precisão quando a GWN é otimizada na face do indivíduo considerado. A GWN otimizada na face média também possibilita o rastreamento, mas com menor precisão. Assim, após o posicionamento das características faciais, a representação wavelet da face é obtida, e o módulo de rastreamento acionado.

Figura 7.9: Rastreamento de olhos, nariz e boca. Note que o método é robusto a piscar de olhos, mudança homogênea de iluminação e diferentes expressões faciais.

7.2.5 Seleção e Normalização dos Melhores Quadros

Durante o processo de rastreamento, é possível que os pontos característicos não sejam precisamente localizados em determinados quadros. Isso ocorre quando o usuário realiza rotações leves fora do plano, modifica sua expressão facial, ou mesmo quando ocorrem mudanças nas condições de iluminação.

Assumindo que as características faciais são corretamente posicionadas inicialmente, é possível selecionar os melhores quadros e eliminar aqueles em que a localização dos pontos característicos não foi efetuada precisamente. Basicamente, se a reparametrização da GWN não foi realizada com precisão, isto é, se a função energia na equação (6.2) assumir um valor maior que um limiar, então o quadro pode ser eliminado.

Ao final, as características faciais dos quadros selecionados são extraídas, normalizadas, e enviadas ao módulo de reconhecimento. A normalização foi realizada aplicando-se uma transformação afim adequada nas regiões selecionadas, de acordo com os correspondentes parâmetros da superwavelet. Outros aspectos como normalização de iluminação precisam ser tratados posteriormente.

Capítulo 8

Conclusões

Neste trabalho, foram apresentados algoritmos de visão computacional para a detecção e rastreamento de faces humanas em seqüências de imagens. Mais especificamente, esse estudo concentrou-se na etapa de rastreamento de faces, sendo proposta, como principal contribuição, uma técnica eficiente para esse problema, baseada em um subespaço wavelet. Tal técnica foi avaliada considerando aspectos como precisão, robustez e eficiência. O problema de localização e rastreamento de características faciais, bem como alguns aspectos relativos à normalização de imagens, também foram comentados, visando a uma aplicação específica: reconhecimento de faces a partir de seqüências de vídeo.

Diversos experimentos foram realizados com a técnica baseada em *Gabor Wavelet Networks*, recentemente proposta na literatura. Entre outras vantagens, verificamos a robustez dessa técnica com relação a deformações rígidas e não rígidas da face no plano, bem como à variação homogênea das condições de iluminação. O compromisso entre precisão e eficiência é determinado pela escolha do número de wavelets na representação. Logo, essa escolha deve ser feita com base nas condições do problema e nos recursos computacionais disponíveis.

Os conceitos apresentados no presente estudo foram reunidos, ao final, objetivando a aplicação considerada, isto é, o reconhecimento de faces a partir de seqüências de vídeo. A pesquisa que subsidiou este trabalho possibilitou resposta a algumas questões, porém suscitou inúmeras outras, inerentes a toda investigação científica. A seguir, apresentamos tópicos a serem tratados em trabalho futuro:

- A verificação da presença ou ausência de uma face em um *blob* de cor de pele deve ser aperfeiçoada com a integração de novas informações como movimento, simetria, forma e outras.
- Pesquisas devem ser realizadas para se obter uma representação genérica de uma face baseada em GWNs. Atualmente, a representação utilizando-se uma face média é adequada apenas para banco de dados com poucos elementos.
- Durante o rastreamento, fatores como rotações leves fora do plano, variação nas condições de iluminação e mudanças de expressão facial são, em geral, bem aproximados pelo modelo wavelet afim. No entanto, é possível que, nessas condições, exista imprecisão no procedimento de posicionamento das características faciais. Esse problema também ocorre na localização dessas características utilizando-se uma GWN otimizada na face média, tendo em vista que a distância entre características faciais varia de pessoa para pessoa. Assim, seria interessante dispormos de um procedimento incumbido de ajustar a posição dos pontos característicos após a reparametrização da GWN, possibilitando uma segmentação precisa das características faciais. Esse tópico está sendo tratado atualmente em nosso grupo de pesquisa.
- Quando a face do usuário realiza rotações consideráveis fora do plano, o método de rastreamento falha. A solução para esse problema seria estender o atual modelo afim da face para uma representação wavelet 3D.
- O rastreamento de faces ainda pode ser aperfeiçoado em termos de eficiência, com a utilização de métodos tais como o Condensation[Isard and Blake, 1998] ou o *Sequential Importance Sampling* (SIS)[Li, 2000], ao invés do algoritmo de Levenberg-Marquardt.
- É interessante comparar os resultados obtidos neste trabalho com outras técnicas que compõem o estado da arte na literatura.

Apêndice A

Lista de Símbolos

$\langle f, g \rangle$	Produto interno em $\mathbb{L}^2(\mathbb{R}^n)$: $\langle f, g \rangle = \int f(t)g(t)dt$
$\mathbb{L}^2(\mathbb{R})$	Espaço das funções quadraticamente integráveis $f : \mathbb{R} \mapsto \mathbb{R}$
$\mathbb{L}^2(\mathbb{R}^2)$	Espaço das funções quadraticamente integráveis $f : \mathbb{R}^2 \mapsto \mathbb{R}$
$g_{u,b}$	Função janela deslocada no tempo (parâmetro b) e modulada por uma exponencial complexa (parâmetro u)
ψ	Wavelet mãe
$\psi_{a,b}$	Wavelet ψ deslocada no tempo (parâmetro b), com uma escala específica (parâmetro a)
C_ψ	Condição de admissibilidade das wavelets
$\mathbf{c}$	Vetor de translação
$\mathbf{R}$	Matriz de rotação

S	Matriz de escala
$\mathbf{n}$	Vetor de parâmetros da wavelet; $\mathbf{n} = (c_x, c_y, \theta, s_x, s_y)$ Vetor de parâmetros da superwavelet $\mathbf{n} = (c_x, c_y, \theta, s_x, s_y, s_{xy})$
$\psi_{\mathbf{n}_i}$	Gabor wavelet com vetor de parâmetros $\mathbf{n}_i$
$\mathbf{w}, \mathbf{v}$	Vetores de pesos $\mathbf{w} = (w_1, \dots, w_N)$, $\mathbf{v} = (v_1, \dots, v_N)$
f, g	Funções contínuas
$\hat{f}$	Aproximação de f ; Representação wavelet
I, J	Imagens discretas em níveis de cinza
$\Psi, \tilde{\Psi}$	Base biortogonal de Gabor wavelets $\Psi = \{\psi_{\mathbf{n}_1}, \dots, \psi_{\mathbf{n}_N}\}$ e $\tilde{\Psi} = \{\tilde{\psi}_{\mathbf{n}_1}, \dots, \tilde{\psi}_{\mathbf{n}_N}\}$
$\Phi, \tilde{\Phi}$	Base biortogonal de wavelets $\Phi = \{\phi_1, \dots, \phi_N\}$ e $\tilde{\Phi} = \{\tilde{\phi}_{\mathbf{n}_1}, \dots, \tilde{\phi}_{\mathbf{n}_N}\}$
$\langle \Psi \rangle$	Espaço vetorial definido pelas Gabor wavelets
E	Função energia
$(\Psi, \mathbf{w})$	<i>Gabor wavelet Network</i>
M, N	Número de wavelets
$\delta_{i,k}$	Função de Dirac $\delta_{i,k} = \begin{cases} 1, & \text{se } i = k \\ 0, & \text{se } i \neq k \end{cases}$

A_{ij}	Matriz de interferência de wavelets
Ψ_n	Superwavelet
δ	Diferença entre dois vetores de pesos $\delta = (\mathbf{v} - \mathbf{w})$
$\mathbb{R}^N$	Subespaço wavelet

Apêndice B

Lista de Abreviaturas

GWN	<i>Gabor wavelet network</i>
GWT	<i>Gabor wavelet template</i>
GSW	<i>Gabor superwavelet</i>
PCA	<i>Principal component analysis</i>
RBF	<i>Radial basis functions</i>

Referências Bibliográficas

- [Antoine et al., 1993] Antoine, J., Carrete, P., Murenzi, R., and Piette, B. (1993). Image analysis with two-dimensional continuous wavelet transform. *Signal Processing*, 31:241–272.
- [Bala et al., 1997] Bala, L., Talmi, K., and Liu, J. (1997). Automatic detection and tracking of faces and facial features in video sequences. In *Picture Coding Symposium*, Berlin, Germany.
- [Birchfield, 1998] Birchfield, S. (1998). Elliptical head tracking using intensity gradients and color histograms. In *Proc. of IEEE CVPR*, pages 232–237, Santa Barbara, California.
- [Blake and Isard, 1998] Blake, A. and Isard, M. (1998). *Active Contours*. Springer.
- [Brunelli and Poggio, 1993] Brunelli, R. and Poggio, T. (1993). Face recognition: Features versus templates. *IEEE Transactions on PAMI*, 15(10):1042–1052.
- [Campos, 2001] Campos, T. (2001). Técnicas de seleção de características com aplicações em reconhecimento de faces. Master’s thesis, Universidade de São Paulo, São Paulo, Brasil.
- [Campos et al., 2000a] Campos, T., Feris, R., and Cesar, R. (2000a). Eigenfaces versus eigeneyes: First steps towards performance assessment of representations for face recognition. In *Lecture Notes in Artificial Intelligence*, volume 1793, pages 193–201, Acapulco, Mexico.
- [Campos et al., 2000b] Campos, T., Feris, R., and Cesar, R. (2000b). Improved face x non-face discrimination using fourier descriptors through feature selection. In *Brazilian*

- Symposium on Computer Graphics and Image Processing*, pages 28–35, Gramado-RS, Brazil.
- [Cascia et al., 2000] Cascia, M., Sclaroff, S., and Athitsos, V. (2000). Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *IEEE Transactions on PAMI*, 22(4):322–336.
- [Cesar and Costa, 1998] Cesar, R. and Costa, L. (1998). Neural cell classification by wavelets and multiscale curvature. *Biological Cybernetics*, 79(4):347–360.
- [Cipolla and Pentland, 1998] Cipolla, R. and Pentland, A. (1998). *Computer Vision for Human-Machine Interaction*. Cambridge University Press.
- [Coelho et al., 1999] Coelho, P., Esperança, C., and Oliveira, A. (1999). An image processing and believe network approach to face detection. In *Brazilian Symposium on Computer Graphics and Image Processing*, pages 177–186, Campinas-SP, Brazil.
- [Costa and Cesar, 2001] Costa, L. and Cesar, R. (2001). *Shape Analysis and Classification: Theory and Practice*. CRC Press Book Series on Image Processing.
- [Cox et al., 1996] Cox, I., Ghosn, J., and Yianilos, P. (1996). Feature-based face recognition using mixture-distances. In *Proc. of IEEE CVPR*, pages 209–216, Seattle, WA.
- [Crowley and Schwerdt, 1999] Crowley, J. and Schwerdt, K. (1999). Robust tracking and compression for video communication. In *ICCV'99 Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, Corfu, Greece.
- [Dai and Nakano, 1996] Dai, Y. and Nakano, Y. (1996). Face-texture model-based on sgld and its application in face detection in a color scene. *Pattern Recognition*, 29(6):1007–1017.
- [Darrel et al., 1998] Darrel, T., Gordon, G., Woodfill, J., and Harville, M. (1998). A virtual mirror interface using real-time robust face tracking. In *Proc. of IEEE Face and Gesture Recognition*, Nara, Japan.
- [Daubechies, 1990] Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005.

- [Daubechies, 1992] Daubechies, I. (1992). *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia, Pennsylvania.
- [Daugman, 1985] Daugman, J. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized two-dimensional visual cortical filters. *Optical Society of America*, 2(7):1160–1168.
- [Daugman, 1988] Daugman, J. (1988). Complete discrete 2d gabor transform by neural networks for image analysis and compression. *IEEE Transactions on ASSP*, 36(7):1169–1179.
- [Edwards et al., 1998] Edwards, G., Taylor, C., and Cootes, T. (1998). Learning to identify and track faces in image sequences. In *Proc. of IEEE Face and Gesture Recognition*, pages 260–265, Nara, Japan.
- [Feris et al., 2000a] Feris, R., Campos, T., and Cesar, R. (2000a). Detection and tracking of facial features in video sequences. In *Lecture Notes in Artificial Intelligence*, volume 1793, pages 127–135, Acapulco, Mexico.
- [Feris et al., 2000b] Feris, R., Campos, T., and Cesar, R. (2000b). A project for face recognition from video sequences using gwn and eigenfeature selection. In *Workshop on Artificial Intelligence and Computer Vision*, pages 141–145, Atibaia - SP, Brazil.
- [Feris and Cesar, 2001] Feris, R. and Cesar, R. (2001). Detection and tracking of facial landmarks using gabor wavelet networks. In *Lecture Notes in Computer Science*, Rio de Janeiro, Brasil. Springer-Verlag Press.
- [Feris et al., 2001] Feris, R., Krüger, V., and Cesar, R. (2001). Efficient real-time face tracking in wavelet subspace. In *ICCV'2001 Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, Vancouver, Canada.
- [Feyrer and Zell, 1999] Feyrer, S. and Zell, A. (1999). Tracking and pursuing persons with a mobile robot. In *ICCV'99 Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, Corfu, Greece.
- [Gabor, 1946] Gabor, D. (1946). Theory of communication. *Institute of Electrical Engineers*, 93(26):429–457.

- [Govindaraju et al., 1989] Govindaraju, V., Sher, D., Srihari, R., and Srhari, S. (1989). Locating human faces in newspaper photographs. In *Proc. of IEEE CVPR*, pages 549–554.
- [Grossmann and Morlet, 1984] Grossmann, A. and Morlet, J. (1984). Decomposition of hardy functions into squared integrable wavelets of constant shape. *SIAM J. Math. Analysis*, 15:723–736.
- [Hager and Belhumeur, 1998] Hager, G. and Belhumeur, P. (1998). Efficient region tracking with parametric models of geometry and illumination. *IEEE Transactions on PAMI*, 20(10):1025–1039.
- [Isard and Blake, 1998] Isard, M. and Blake, A. (1998). Condensation - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28.
- [Jebara and Pentland, 1997] Jebara, T. and Pentland, A. (1997). Parametrized structure from motion for 3d adaptative feedback tracking of faces. In *Proc. of IEEE CVPR*.
- [Kjeldsen and Kender, 1996] Kjeldsen, R. and Kender, J. (1996). Finding skin in color images. In *Proc. of IEEE Face and Gesture Recognition*, pages 312–318, Killington, VT.
- [Kosko, 1992] Kosko, B. (1992). *Neural Networks and Fuzzy Systems*. Prentice Hall.
- [Kruger and Sommer, 2000] Kruger, V. and Sommer, G. (2000). Gabor wavelet networks for object representation. Technical Report 2002, University Of Kiel, Germany.
- [Li, 2000] Li, B. (2000). Human and object tracking and verification in video. Technical Report CS-TR-4140, Center for Automation Research, University of Maryland.
- [Mallat, 1989] Mallat, S. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on PAMI*, 11(7):674–693.
- [Manjunath and Chellapa, 1993] Manjunath, B. and Chellapa, R. (1993). A unified approach to boundary perception: edges, textures, and illusory contours. *IEEE Transactions on Neural Networks*, 4(1):96–107.

- [Maurer and Malsburg, 1996] Maurer, T. and Malsburg, C. (1996). Tracking and learning graphs and pose on image sequence of faces. In *Proc. International Conference on Artificial Neural Networks*, Bochum.
- [McKenna et al., 1996] McKenna, S., Gong, S., and Collins, J. (1996). Face tracking and pose representation. In *British Machine Vision Conference*, Edinburgh.
- [Moghaddam and Pentland, 1994] Moghaddam, B. and Pentland, A. (1994). Face recognition using view-based and modular eigenspaces. In *Automatic Systems for the Identification and Inspection of Humans*, volume 2277. SPIE.
- [Moghaddam and Pentland, 1997] Moghaddam, B. and Pentland, A. (1997). Probabilistic visual learning for object representation. *IEEE Transactions on PAMI*, 19(7):696–710.
- [Morimoto et al., 1998] Morimoto, C., Koons, D., Amir, A., and Flickner, M. (1998). Real-time detection of eyes and faces. In *Workshop on Perceptual User Interfaces*, pages 117–120, San Francisco, California.
- [Olivier et al., 1997] Olivier, N., Pentland, A., and Berard, F. (1997). Lafter: Lips and face real-time tracker. In *Proc. of IEEE CVPR*.
- [Press et al., 1986] Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (1986). *Numerical Recipes, The Art of Scientific Computing*. Cambridge University Press.
- [Raja et al., 1998] Raja, Y., McKenna, S., and S.Gong (1998). Tracking and segmenting people in varying lighting conditions using colour. In *Proc. of IEEE Face and Gesture Recognition*, Nara, Japan.
- [Rowley et al., 1998a] Rowley, H., Baluja, S., and Kanade, T. (1998a). Neural network-based face detection. *IEEE Transactions on PAMI*, 20(1):23–38.
- [Rowley et al., 1998b] Rowley, H., Baluja, S., and Kanade, T. (1998b). Rotation invariant neural network-based face detection. In *Proc. of IEEE CVPR*, pages 38–44, Santa Barbara, California.
- [Sakai et al., 1969] Sakai, T., Nagao, M., and Fujibayashi, S. (1969). Line extraction and pattern recognition in a photograph. *Pattern Recognition*, 1:233–248.

- [Silva et al., 1995] Silva, L., Aizawa, K., and Hatori, M. (1995). Detection and tracking of facial features. In *Proc. of SPIE Visual Communications and Image Processing*, Taiwan.
- [Stiefelhagen and J. Yang, 1997] Stiefelhagen, R. and J. Yang (1997). Gaze tracking for multimodal human computer interaction. In *Proc. of ICASSP*, Munich, Germany.
- [Sun et al., 1998] Sun, Q., Huang, W., and Wu, J. (1998). Face detection based on color and local symmetry information. In *Proc. of IEEE Face and Gesture Recognition*, Nara, Japan.
- [Sung and Poggio, 1998] Sung, K. and Poggio, T. (1998). Example-based learning for view-based human face detection. *IEEE Transactions on PAMI*, 20(1):39–50.
- [Szu et al., 1992] Szu, H., Telfer, B., and Kadambe, S. (1992). Neural network adaptive wavelets for signal representation and adaptation. *Optical Engineering*, 31(9):1907–1961.
- [Toyama, 1998a] Toyama, K. (1998a). "look, ma - no hands!" hand-free control with real-time 3d face tracking. In *Workshop on Perceptual User Interfaces*, pages 117–120, San Francisco, California.
- [Toyama, 1998b] Toyama, K. (1998b). Prolegomena for robust face tracking. Technical Report MSR-TR-98-65, Microsoft Corporation, Redmond, WA.
- [Tsotsos, 1990] Tsotsos, J. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sci.*, 13:423–469.
- [Turk, 1996] Turk, M. (1996). Visual interaction with lifelike characters. In *Proc. of IEEE Face and Gesture Recognition*, Killington, VT.
- [Turk, 1998] Turk, M. (1998). Moving from guis to puis. Technical Report MSR-TR-98-69, Microsoft Corporation, Redmond, WA.
- [Turk and Pentland, 1991] Turk, M. and Pentland, A. (1991). Face recognition using eigenfaces. In *Proc. of IEEE ICPR*, pages 586–591.

- [Vetterli and Kovacevic, 1995] Vetterli, M. and Kovacevic, J. (1995). *Wavelets and Sub-band Coding*. Prentice Hall.
- [Wiskott et al., 1997] Wiskott, L., Fellous, J., Krüger, N., and Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Transactions on PAMI*, 19(7):775–779.
- [Wu et al., 1999] Wu, H., Chen, Q., and Yachida, M. (1999). Face detection from color images using a fuzzy pattern matching method. *IEEE Transactions on PAMI*, 21(6):557–563.
- [Yang et al., 1997] Yang, J., Lu, W., and Waibel, A. (1997). Skin-color modeling and adaptation. Technical report, CMU-CS-97-146.
- [Yang and Waibel, 1996] Yang, J. and Waibel, A. (1996). A real-time face tracker. In *Proc. of IEEE Workshop on Applications of Computer Vision*, pages 142–147, Sarasota, Florida.
- [Yow and Cipolla, 1997] Yow, K. and Cipolla, R. (1997). Feature-based human face detection. *Image and Vision Computing*, 15(9):713–735.
- [Yuille et al., 1989] Yuille, A., Cohen, D., and Hallinan, P. (1989). Feature extraction from faces using deformable templates. In *Proc. of IEEE CVPR*, pages 104–109.
- [Zarit et al., 1999] Zarit, B., Super, B., and Quek, F. (1999). Comparison of five color models in skin pixel classification. In *ICCV'99 Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, Corfu, Greece.
- [Zelinsky and Heinzmann, 1996] Zelinsky, A. and Heinzmann, J. (1996). Real-time visual recognition of facial gestures for human-computer interaction. In *Proc. of IEEE Face and Gesture Recognition*, pages 351–356, Killington, VT.
- [Zhang and Benviste, 1992] Zhang, Q. and Benviste, A. (1992). Wavelet networks. *IEEE Transactions on Neural Networks*, 3(6):889–898.

Anexo

Principais artigos publicados

Efficient Real-Time Face Tracking in Wavelet Subspace

Rogério Schmidt Feris

Volker Krüger

Roberto Marcondes Cesar Jr.

Second International Workshop on Recognition, Analysis and
Tracking of Faces and Gestures in Real-time Systems
(RATFG-RTS 2001)

ICCV'2001 International Conference on Computer Vision

13-14 de Julho, 2001

Vancouver, Canada

Efficient Real-Time Face Tracking in Wavelet Subspace

Rogério S. Feris, Roberto M. Cesar Jr
Department of Computer Science
University of São Paulo
Rua do Matão, 1010, São Paulo-SP, Brazil
{rferis | cesar}@ime.usp.br

Volker Krüger
Center for Automation Research
University of Maryland
College Park, MD 20740
vok@cfar.umd.edu

Abstract

In this article we present a new method for visual face tracking that is carried out in wavelet subspace. Firstly, a wavelet representation for the face template is created, which spans a low-dimensional subspace of the image space. The video sequence frames where the face is tracked are then orthogonally projected into this low-dimensional subspace. This can be done efficiently through a small number of applications of the wavelet filters. All further computations are performed in wavelet subspace, which is isomorphic to the image subspace spanned by the sets of wavelets in the representation. Robustness w.r.t. facial expression and affine deformations, as well as the efficiency of our method, are demonstrated in various experiments.

1. Introduction

This paper addresses the issue of affine real-time face tracking. Real-time (RT, 25/30 Hz) is a major requisite for many human-computer-interface (HCI) and surveillance applications, as well as for tele-conferencing and tele-teaching tasks. For gesture, gaze and pose estimation applications, tracking has to be not only fast, but also precise. A variety of tracking approaches already exists [5, 2, 1]. In this work, we will use Gabor Wavelet Networks (GWN) [6, 4] in order to represent the face to be tracked. It is worth saying that we have already discussed a GWN-based real-time tracking method in [5]. The method discussed in the present paper differs from that approach in the sense that it is performed in a low-dimensional wavelet subspace, posing a considerable enhancement in terms of efficiency.

Gabor Wavelet Networks (GWN) combine the advantages of RBF networks with the advantages of Gabor wavelets. Objects are represented through a linear combination of Gabor wavelets where the parameters of each of the Gabor functions (such as orientation, position and scale) are optimized to reflect the particular local image structure.

The use of Gabor Wavelet Networks has several advantages, namely:

1. By their very nature, Gabor wavelet networks are invariant to some degree to affine deformations and homogeneous illumination changes,
2. Gabor filters are good feature detectors [8, 9] and the optimized parameters of each Gabor wavelet reflect the underlying image structure,
3. The Gabor wavelet weights are directly related to the Gabor filter responses and thus also reflect the underlying local image structure,
4. The precision of the representation can be varied to any desired degree ranging from a coarse to an almost photo-realistic representation by simply varying the number of used wavelets. Depending on the available computer power and the necessary tracking precision, the number of wavelets can be dynamically varied.

Properties 2-4 allow us to define a subspace of the image space. Its basis is given by the selectively chosen Gabor wavelets. Property 4 allows an easy projection of any image into the subspace, while property 3 assures the great sparseness of the wavelet representation. We will discuss each single point in section 2. In section 3, we will introduce our subspace tracking approach, discuss the details and conclude the paper with the experiments in section 4 and concluding remarks.

2. Introduction to Gabor Wavelet Networks

The basic idea of the wavelet networks was first stated in [11], and the use of Gabor functions is inspired by the fact that they are recognized to be good feature detectors [8, 9].

To define a GWN, we start out, generally speaking, by taking a family of N odd Gabor wavelet functions $\Psi =$

Figure 1. Face images reconstructed with different number of wavelets

$\{\psi_{\mathbf{n}_1}, \dots, \psi_{\mathbf{n}_N}\}$ of the form

$$\begin{aligned} \psi_{\mathbf{n}}(x, y) = & \exp \left(-\frac{1}{2} \left[s_x ((x - c_x) \cos \theta - (y - c_y) \sin \theta) \right]^2 \right. \\ & \left. + \left[s_y ((x - c_x) \sin \theta + (y - c_y) \cos \theta) \right]^2 \right) \\ & \cdot \sin \left(s_x ((x - c_x) \cos \theta - (y - c_y) \sin \theta) \right), \quad (1) \end{aligned}$$

with $\mathbf{n} = (c_x, c_y, \theta, s_x, s_y)^T$. Here, c_x, c_y denote the translation of the Gabor wavelet, s_x, s_y denote the dilation and θ denotes the orientation. The choice of N is arbitrary and related to the degree of desired representation precision of the network. In order to find the GWN for a function $f \in \mathbb{L}^2(\mathbb{R}^2)$ (f dc-free, w.l.o.g.) the energy functional

$$E = \min_{\mathbf{n}_i, w_i \text{ for all } i} \|f - \sum_i w_i \psi_{\mathbf{n}_i}\|_2^2 \quad (2)$$

is minimized with respect to the weights w_i and the wavelet parameter vector $\mathbf{n}_i$. The two vectors

$$\Psi = (\psi_{\mathbf{n}_1}, \dots, \psi_{\mathbf{n}_N})^T \quad \text{and} \quad \mathbf{w} = (w_1, \dots, w_N)^T$$

define then the *Gabor wavelet network* $(\Psi, \mathbf{w})$ for function f .

In other words, considering a discrete image I^1 , a Gabor wavelet network is defined through a N -dimensional vector of weights w_i and a N -dimensional vector of Gabor wavelets $\psi_{\mathbf{n}_i}$, where the weights w_i and the parameter vectors $\mathbf{n}_i$ are chosen such that the weighted sum of Gabor wavelets $\psi_{\mathbf{n}_i}$ approximates the discrete gray value image I optimally.

From the optimal wavelets Ψ and weights $\mathbf{w}$ of the Gabor wavelet network, the function f can be (closely) reconstructed by a linear combination of the weighted wavelets:

$$\hat{f} = \sum_{i=1}^N w_i \psi_{\mathbf{n}_i} = \Psi^T \mathbf{w}.$$

¹We use the notation $f, g, \dots$ to generally refer to continuous functions, while we use $I, J, \dots$ when we explicitly refer to discrete gray value images.

Clearly, the quality of the image representation and reconstruction depends on the number N of wavelets and can be varied to reach almost any desired precision. An example of reconstruction can be seen in fig. 1. In this example, a family of 216 wavelets has been distributed over the inner face region of the very right image I by minimizing equation (2). Different reconstructions $\hat{I}$ obtained by the application of equation (3) for different values of N are shown.

2.1. Direct Calculation of Weights

The weights w_i of a GWN are directly related to the local filter responses of the Gabor filters $\psi_{\mathbf{n}_i}$. Gabor wavelet functions are not orthogonal, thus implying that, for a given family Ψ of Gabor wavelets, it is not possible to calculate a weight w_i by a simple projection of the Gabor wavelet $\psi_{\mathbf{n}_i}$ onto the image. In fact, a family of dual wavelets $\tilde{\Psi} = \{\tilde{\psi}_{\mathbf{n}_1}, \dots, \tilde{\psi}_{\mathbf{n}_N}\}$ has to be considered. The wavelet $\tilde{\psi}_{\mathbf{n}_j}$ is the dual wavelet of the wavelet $\psi_{\mathbf{n}_i}$ iff $\langle \psi_{\mathbf{n}_i}, \tilde{\psi}_{\mathbf{n}_j} \rangle = \delta_{i,j}$. With $\tilde{\Psi} = (\tilde{\psi}_{\mathbf{n}_1}, \dots, \tilde{\psi}_{\mathbf{n}_N})^T$, we can write $[\langle \Psi, \tilde{\Psi} \rangle] = \mathbf{I}$. In other words, given $g \in \mathbb{L}^2(\mathbb{R}^2)$ and a GWN $\Psi = \{\psi_{\mathbf{n}_1}, \dots, \psi_{\mathbf{n}_N}\}$, the optimal weights for g that minimize the energy in eq. (2) are given by $w_i = \langle g, \tilde{\psi}_{\mathbf{n}_i} \rangle$. It can be shown that $\psi_{\mathbf{n}_i} = \sum_j (\Psi^{-1})_{i,j} \psi_{\mathbf{n}_j}$, where $\Psi_{i,j} = \langle \psi_{\mathbf{n}_i}, \psi_{\mathbf{n}_j} \rangle$.

The above equations allow us to define the operator

$$\mathcal{T}_{\Psi} : \mathbb{L}^2(\mathbb{R}^2) \mapsto \langle (\psi_{\mathbf{n}_1}, \dots, \psi_{\mathbf{n}_N}) \rangle \quad (3)$$

as follows: given a set Ψ of optimal wavelets of a GWN, the operator $\mathcal{T}_{\Psi}$ represents an orthogonal projection of a function g onto the closed linear span of Ψ (see eq. (3) and fig. 2), i.e.

$$\hat{g} = \mathcal{T}_{\Psi}(g) = g \tilde{\Psi} \Psi = \sum_{i=1}^N w_i \psi_{\mathbf{n}_i}, \quad \text{with } \mathbf{w} = g \tilde{\Psi}. \quad (4)$$

3. Face Tracking in Wavelet Subspace

The wavelet representation described in the previous section may be effectively used for affine face tracking. Basi-

Figure 2. A function $g \in \mathbb{L}^2(\mathbb{R}^2)$ is mapped by the linear mapping $\tilde{\Psi}$ into the vector $\mathbf{w} \in \mathbb{R}^N$. The mapping of $\mathbf{w}$ into $\mathbb{L}^2(\mathbb{R}^2)$ is achieved with the linear mapping Ψ . Both mappings constitute an orthogonal projection of a function $g \in \mathbb{L}^2(\mathbb{R}^2)$ into the subspace $\langle \Psi \rangle \subset \mathbb{L}^2(\mathbb{R}^2)$.

cally, this task is achieved by affinely deforming a GWN so that it matches the face image in each frame of a video sequence. The affine deformation of a GWN is carried out by considering the entire wavelet network as a single wavelet, which is also called *superwavelet* [5]. Let $(\Psi, \mathbf{w})$ be a Gabor wavelet network with $\Psi = (\psi_{\mathbf{n}_1}, \dots, \psi_{\mathbf{n}_N})^T$ and $\mathbf{w} = (w_1, \dots, w_N)^T$. A superwavelet $\Psi_{\mathbf{n}}$ is defined as a linear combination of the wavelets $\psi_{\mathbf{n}_i}$ such that

$$\Psi_{\mathbf{n}}(\mathbf{x}) = \sum_i w_i \psi_{\mathbf{n}_i}(\mathbf{S}\mathbf{R}(\mathbf{x} - \mathbf{c})), \quad (5)$$

where the vector $\mathbf{n} = (c_x, c_y, \theta, s_x, s_y, s_{xy})^2$ defines the dilation matrix $\mathbf{S}$, the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{c}$, respectively. The affine face tracking is then achieved by deforming the superwavelet $\Psi_{\mathbf{n}}$ in each frame J , so that its parameters $\mathbf{n}$ are optimized with respect to the energy functional E (see eq. (2)):

$$E = \min_{\mathbf{n}} \|J - \Psi_{\mathbf{n}}\|_2^2 \quad (6)$$

Clearly, this method performs a typical pixel-wise pattern matching in image space, where the template corresponds to the wavelet representation, which is affinely distorted to match the face in each frame. It is interesting to note that the wavelet weights w_i (see eq. (5)) are constant under the deformation of the template. Therefore, the affine deformation is captured only by the deformation of the wavelets, while the weight vector remains invariant.

We thus claim that the tracking in image space described above may also be achieved in the low-dimensional space $\mathbb{R}^N$, which is isomorphic to the image subspace $\langle \Psi \rangle$, as illustrated by fig. 2. As it can be seen there, both spaces are related through the matrices Ψ and $\tilde{\Psi}$, respectively.

As the first step, consider a GWN $(\Psi, \mathbf{v})$ that is optimized for a certain face image. As previously mentioned,

²To include the shear in the parameter set of the wavelets, see [5].

the optimal weight vector $\mathbf{v}$ is obtained by an orthogonal projection of the facial image into the closed linear span of Ψ . Hence, we say that the face template was mapped into the weights $\mathbf{v} \in \mathbb{R}^N$, which we will call *reference weights*.

We mentioned before that the wavelet template gets affinely deformed for tracking in image space. Analogously, the tracking in wavelet subspace is performed by affinely deforming the subspace $\langle \Psi \rangle$, until the weight vector $\mathbf{w} \in \mathbb{R}^N$, obtained by the orthogonal mapping of the current frame into this subspace, is closest to the reference weight vector $\mathbf{v}$. In fact, this procedure performs roughly the same pattern matching as before, but this is done efficiently in the low-dimensional space $\mathbb{R}^N$.

The mapping of images into $\mathbb{R}^N$ is carried out with low computational cost through a small number of local filtrations with the wavelets. Recall from section 2.1 that $w_i = \langle I, \tilde{\psi}_{\mathbf{n}_i} \rangle$, where $\tilde{\psi}_{\mathbf{n}_i} = \sum_j (\Psi^{-1})_{i,j} \psi_{\mathbf{n}_j}$. This is equal to the following equation:

$$w_i = \sum_j (\Psi^{-1})_{i,j} \langle I, \psi_{\mathbf{n}_j} \rangle. \quad (7)$$

Thus, the optimal weights w_i are derived from a linear combination of wavelet filtrations, where the coefficients are given by the inverse of matrix $\Psi_{i,j} = \langle \psi_{\mathbf{n}_i}, \psi_{\mathbf{n}_j} \rangle$. It can be shown that the matrix $\Psi_{i,j}$ is, except for a scalar factor, invariant with respect to affine transformations of the wavelet network. It can therefore be computed off-line and beforehand. Hence, the weights w_i are computed efficiently with eq. (7) through a local application $\langle I, \psi_{\mathbf{n}_i} \rangle$ of each of the N Gabor wavelets $\psi_{\mathbf{n}_i}$, followed by a $N \times N$ matrix multiplication.

Let $\mathbf{n} = (c_x, c_y, \theta, s_x, s_y, s_{xy})$ be an affine parameter vector which configures a parameterization for the subspace $\langle \Psi \rangle$. As we described before, tracking in wavelet subspace is achieved by gradually changing these parameters until the projection of the current frame J onto $\mathbf{w} \in \mathbb{R}^N$ is closest to the reference weights $\mathbf{v}$. In other words, we must optimize the parameters $\mathbf{n}$ with respect to the energy functional:

$$E = \min_{\mathbf{n}} \|\mathbf{v} - \mathbf{w}\|_{\Psi} \quad \text{with} \quad (8)$$

$$w_i = \sum_j \frac{1}{s_x \cdot s_y} (\Psi^{-1})_{i,j} \langle J, \psi_{\mathbf{n}_j}(\mathbf{S}\mathbf{R}(\mathbf{x} - \mathbf{c})) \rangle \quad (9)$$

where $\mathbf{S}$ is the dilation matrix, $\mathbf{R}$ is the rotation matrix and $\mathbf{c}$ is the translation vector, all defined through the vector $(c_x, c_y, \theta, s_x, s_y, s_{xy})$. During tracking, this optimization is done for each successive frame. As there is not much difference between adjacent frames, the optimization is fast. To minimize the energy functional (8), the Levenberg-Marquardt algorithm was used.

In equation (8) we used the notation $\|\mathbf{v} - \mathbf{w}\|_{\Psi}$ to refer to a distance between vectors $\mathbf{v}$ and $\mathbf{w}$ of the subspace

$\langle \Psi \rangle$. However, it is not yet clear how this distance measure should be calculated. One could use the Euclidean distance between two vectors, as done in [10]. However, that distance measure misses any interpretation in this context and all wavelets are treated as equal, even though they might be of different scales.

Therefore we propose a different distance measure, which is based on the Euclidean distance between the two corresponding points in the wavelet subspace $\langle \Psi \rangle$. We thus define the difference $\|\mathbf{v} - \mathbf{w}\|_\Psi$ as the Euclidean distance between the two respective images $\hat{f}$ and $\hat{g}$:

$$\begin{aligned}\|\mathbf{v} - \mathbf{w}\|_\Psi &= \|\mathcal{T}_\Psi(f) - \mathcal{T}_\Psi(g)\|_2 \\ &= \left\| \sum_{i=1}^N v_i \psi_{\mathbf{n}_i} - \sum_{j=1}^N w_j \psi_{\mathbf{n}_j} \right\|_2\end{aligned}\quad (10)$$

Various transformations lead to

$$\begin{aligned}\|\mathbf{v} - \mathbf{w}\|_\Psi &= \left[\sum_{i,j} (v_i - w_i)(v_j - w_j) \langle \psi_{\mathbf{n}_i}, \psi_{\mathbf{n}_j} \rangle \right]^{\frac{1}{2}} \\ &= (\mathbf{v} - \mathbf{w})^t \Psi_{i,j} (\mathbf{v} - \mathbf{w}).\end{aligned}\quad (11)$$

The matrix of pairwise scalar products $\Psi_{i,j} = \langle \psi_{\mathbf{n}_i}, \psi_{\mathbf{n}_j} \rangle$ is the same matrix as the one in section 2.1 and eq. (7). Note that if the wavelets $\{\psi_{\mathbf{n}_i}\}$ were orthogonal, then Ψ would be the unity matrix and eq. (11) would describe the same distance measure as proposed in [10].

Compared to tracking in image space, the method presented here poses a considerable enhancement in terms of efficiency, as it provides a great data reduction, considering that tracking is performed in a low dimensional wavelet subspace. Moreover, it spares out the computationally demanding template reconstruction and pixel-wise sum-of-squared-difference computation required in the image-based tracking.

4. Experiments

The proposed approach for affine face tracking was successfully tested on various image sequences. All test sequences showed a person in motion, more or less frontal to the camera, so that the facial features are always visible.

Our experiments demonstrate the ability of our method to track a face as it undergoes changes in pose and expression. Although we have represented the face as a rigid object undergoing limited motion, we realized that different face expressions and small depth variations exhibited by facial features are enough to be well-approximated by the affine wavelet model. Furthermore, since Gabor wavelets are DC free, our approach also showed robustness with respect to homogeneous illumination variations.

Concerning tracking precision, we have evaluated the proposed method with respect to the number of wavelets used in the representation. As the first step, we recorded a video sequence that shows a person under different poses and facial expressions³. On the face of that person, a GWN with 116 Gabor wavelets was optimized and used to estimate the “ground truth” affine parameters in each frame. Figure 3 illustrates some frames of our subspace tracking results, using this wavelet representation. Note that the tracking method is robust to facial expressions variations as well as affine deformations of the face image.

Now, we want to analyze the decrease of precision when using a smaller number of wavelets in the representation. For that, from the large GWN above, GWNs that contained only the largest 51, 22 and 9 wavelets, sorted according to decreasing normalized weights, were used for tracking. The graphs in figure 4 depict the estimation of the face parameters x-position, y-position and angle θ as well as the ground-truth in each frame.

In the graphs we can see that the tracking precision increases with an increasing number of applied wavelets. On the other hand, the tracking speed is improved when the number of wavelets is small. Clearly, the number of applied wavelets is task dependent and can be dynamically changed.

The resolution of the image frames showed in figure 3 is 160x120 pixels and the size of the inner face region in which the GWN was optimized is 50x65 pixels. Using only 9 wavelets, the computing time for each Levenberg-Marquardt cycle was 15ms on a 1GHz Linux-Athlon. Higher performance is achieved for smaller face regions, or fewer parameters (e.g. just translation and rotation). It is worth saying that the number of cycles for each frame depends on the distance of the initial parameters from the local minimum, which is directly related to object speed. Increasing the number of wavelets in the representation leads to a more precise but slower tracking. For instance, using 51 wavelets, a computing time of 85ms per cycle was required in each frame.

Our method is closely related to the one presented in [5]. In comparison to that method we observed in our experiments a speedup of a factor two for each single Levenberg-Marquardt cycle, but with a slight increase in the number of needed cycles. However, we do believe that the use of other algorithms, such as the Condensation method [3] or the Sequential Importance Sampling (SIS) method [7], will even lead to better results in terms of efficiency.

5. Conclusions

In this paper we have presented a tracking method, that is based on Gabor Wavelet Networks. The GWNs are in-

³see <http://www.ime.usp.br/~rferis>

Figure 3. Sample frames of our wavelet subspace tracking. Note that the tracking method is robust to facial expressions variations as well as affine deformations of the face image.

variant to affine deformations so that they leave the reference vector of a template face constant. Furthermore, the direct relationship between the wavelet coefficients and the wavelet filter responses, which is widely used for multi-resolution analysis and for motion estimation, allows to map an image into the low-dimensional space $\mathbb{R}^N$, where all subsequent computations can be carried out. Furthermore, the careful selection of the wavelets for defining a GWN ensures that almost all visual information is preserved (see fig. 1 and [4]).

The method has the further advantage, that its precision and computation time can be adapted to the needs of a given task. When fast and only approximate tracking is needed, a small number of filtrations is usually sufficient to realize tracking. When high precision tracking is needed, the number of wavelets can be gradually increased. This implies on the one hand more filtrations, but ensures on the other hand a higher precision, as we have shown in the experimental section.

So far, the optimization of the wavelet networks has been done with a Levenberg-Marquardt algorithm. We think that the application of Condensation or Sequential Importance Sampling (SIS) instead would lead to a further speed-up. This will be evaluated in future work.

6. Acknowledgements

Rogério Feris is grateful to FAPESP (99/01487-1). Roberto M. Cesar Jr. is grateful to FAPESP (99/12765-2), CNPq (300722/98-2) and TSI-ENST Paris.

References

- [1] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Real-time tracking of non-rigid objects using mean shift. In *IEEE Conf. Computer Vision and Pattern Recognition, CVPR*, volume 2, pages 142–149, Hilton Head Island, SC, June 13-15, 2000.
- [2] G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(10):1025–1039, 1998.
- [3] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.
- [4] V. Krüger, Sven Bruns, and G. Sommer. Efficient head pose estimation with gabor wavelet networks. In *Proc. British Machine Vision Conference*, Bristol, UK, Sept. 12-14, 2000.
- [5] V. Krüger and G. Sommer. Affine real-time face tracking using gabor wavelet networks. In *Proc. Int. Conf. on Pattern Recognition*, Barcelona, Spain, Sept. 3-8, 2000.
- [6] V. Krüger and G. Sommer. Gabor wavelet networks for object representation. Technical Report 2002, Institute of Computer Science, University of Kiel, 2000.
- [7] Baoxin Li. Human and object tracking and verification in video. Technical Report CS-TR-4140, Center for Automation Research, University of Maryland, May 2000.
- [8] B.S. Manjunath and R. Chellappa. A unified approach to boundary perception: edges, textures, and illusory contours. *IEEE Trans. Neural Networks*, 4(1):96–107, 1993.
- [9] R. Mehrotra, K.R. Namuduri, and R. Ranganathan. Gabor filter-based edge detection. *Pattern Recognition*, 52(12):1479–1494, 1992.
- [10] L. Wiskott, J. M. Fellous, N. Krüger, and C. v. d. Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):775–779, July 1997.
- [11] Q. Zhang and A. Benveniste. Wavelet networks. *IEEE Trans. Neural Networks*, 3(6):889–898, Nov. 1992.

Figure 4. Estimation of the face parameters x-position, y-position and angle θ in each frame, using GWNS with 9, 22 and 51 wavelets. The ground-truth is depicted to illustrate the decrease of precision when considering few wavelets.

Locating and Tracking Facial Landmarks Using Gabor Wavelet Networks

Rogério Schmidt Feris

Roberto Marcondes Cesar Jr.

ICAPR'2001 International Conference on Advances in Pattern
Recognition

Lecture Notes in Computer Science (LNCS)
Springer-Verlag press

11-14 de Março, 2001

Rio de Janeiro, Brasil

Locating and Tracking Facial Landmarks Using Gabor Wavelet Networks

Rogério S. Feris and Roberto M. Cesar Junior

Department of Computer Science, University of São Paulo,
Rua do Matão, 1010, 05508-900 São Paulo-SP, Brazil
{rferis,cesar}@ime.usp.br

Abstract. A new approach for locating and tracking facial landmarks in video sequences is introduced in this paper. Our method is based on Gabor wavelet networks, an effective technique that represents a discrete face template as a linear combination of 2D Gabor wavelet functions. This wavelet representation allows positioning of facial landmarks (e.g. eyes, nose and mouth), even in the presence of glasses, beard and different facial expressions. The feature tracking is robust to homogeneous illumination changes and affine deformations of the face image. Moreover, the tracking approach considers the overall geometry of the face, thus being robust to deformations such as eye blinking and smile, which is usually a critical situation to most local-based traditional methods.

Keywords: Computer Vision, Facial Feature Tracking, Gabor Wavelets.

1 Introduction

Computational face recognition is an important research problem in computer vision, presenting many applications such as in human-computer interaction, security systems and surveillance. There are two main approaches to face recognition by computers, namely, static and dynamic. While the former is related to recognizing people in still images, the latter addresses the problem of detecting, tracking, extracting information and recognizing moving people in digital video sequences. As far as the problem of tracking a face in a video sequence is concerned, there are three distinct approaches: (1) tracking the whole face region; (2) tracking the head outline (e.g. using active contour models); (3) tracking a set of feature points or facial landmarks, generally defined by the eyes, the nose and the mouth. The last approach presents, among its attractives, the potentiality for allowing faster recognition in real-time applications and the fact of presenting a psychophysical inspiration because of the well-known importance that these landmarks present to human perception. This paper introduces an approach for locating and tracking facial feature points, which has proven to be robust and even suitable for real-time applications.

The technique is based on a recent approach for face representation called Gabor wavelet network (GWN) [1]. The GWN represents the face as a linear combination of 2D Gabor wavelet functions, whose parameters (position,

scale and orientation) and weights are determined optimally so that the maximum of image information is preserved for a given number of wavelets.

Once that the GWN has been optimized, it can be repositioned with its wavelet parameters undergoing an affine transformation in order to match a target face image. The repositioning procedure is a key concept of this paper, being applied with two different purposes, namely, (1) for locating the facial feature points of the detected face-like blob in an initial frame and (2) for performing facial feature tracking.

Basically, our approach can be divided in three subsequent steps:

- *Face detection*, which is performed automatically by skin-color blob detection [2]. Once that a face-like blob is located, a simple correlation procedure is used to decide whether the blob actually corresponds or not to a true face.
- *Facial feature points positioning*, which is done automatically by matching a GWN, optimized considering a mean face, to the initial face-like blob;
- *Tracking of face and facial feature points* by the GWN. The tracking algorithm may be even executed in real-time. In this case, a suitable number of wavelets in the representation must be chosen with respect to the available computational resources.

We show that our method is able to locate facial features points even in the presence of glasses, beard, and different facial expressions. The tracking approach is robust to homogeneous illumination changes and affine deformations of the face image. Moreover, since we consider the overall geometry of the face, it is robust to facial feature deformations such as eye blinking and smile.

The remainder of this paper is organized as follows. Section 2 reviews some techniques related to our work. Section 3 introduces the Gabor wavelet networks for face representation and advantages of this technique are discussed. Section 4 is concerned with the repositioning of a GWN, which allows face tracking. In section 5, the positioning of facial landmarks is presented. The facial feature tracking is described in section 6. Experimental results and some discussions are presented in section 7. Finally, section 8 concludes this paper with some remarks on further research directions.

2 Related Work

Many approaches have been proposed to locate and track faces and facial features in video sequences. Recently, color-based systems have been widely used to accomplish this task. For instance, we can cite the work of Jie Yang and Alex Waibel [3], which presents a real-time face tracker based on a statistical skin-color model [4]. Another example is the work of Stiefelhagen and Yang [5], that describes a color-based method for detection and tracking of specific facial features (pupils, nostrils and lipcorners). The use of color to

track faces and facial features has advantages such as face pose invariance and real-time processing. On the other hand, this approach is, in general, not robust to illumination changes.

Liyanage Silva et. al. [6] used an edge-based approach to locate and track facial features in image sequences. This method is based on the fact that a higher edge concentration is observed in the center of facial features (eyes, nose and mouth), while slightly outside of such features a less edge concentration is observed. The method is simple but it fails in several situations, such as in the presence of cluttered backgrounds, glasses and hair covering the forehead.

Another different approach is presented in the work of Thomas Maurer and Christoph Malsburg [7], which describes a system for tracking facial features with the elastic graph matching technique. In this method, Gabor filters are applied in some facial feature locations (selected by hand), forming a feature vector, or a jet, for each face position. The face is then modeled as a graph, in which the nodes correspond to the jets and the edges encode face geometrical information. Facial feature tracking is performed by a graph matching procedure in each frame. The main disadvantage of this approach is the high computational cost required.

Our approach uses a wavelet representation for the face image that is even sparser than the Gabor jet representation. Also it differs from the one introduced by Mallat or Daubechies [8,9]. In fact, it is based on a wavelet network concept, which will be explained in the next section.

3 Face Representation Using GWN

Wavelet networks [10], or wavenets, were proposed as an alternative to feed-forward neural networks for function approximation. In this section we will show, with basis on the recent work of Kruger and Sommer [1], that this mathematical tool may be used to approximate a discrete face template, providing an effective face representation.

To define a Gabor wavelet network, we start by taking a family of M 2D odd-Gabor wavelet functions $\Psi = \{\psi_{\mathbf{n}_1}, \dots, \psi_{\mathbf{n}_M}\}$ of the form

$$\begin{aligned} \psi_{\mathbf{n}}(x, y) = & \exp\left(-\frac{1}{2}[s_x((x - c_x)\cos\theta - (y - c_y)\sin\theta)]^2\right. \\ & \left.+ [s_y((x - c_x)\sin\theta + (y - c_y)\cos\theta)]^2\right) \\ & \times \sin(s_x((x - c_x)\cos\theta - (y - c_y)\sin\theta)) \end{aligned} \quad (1)$$

with the parameter vector $\mathbf{n} = (c_x, c_y, \theta, s_x, s_y)$, where c_x, c_y denote the translation (position) of the Gabor wavelet, s_x, s_y denote the dilation (scale) and θ denotes the orientation.

In order to obtain a wavelet representation for a face image f , the weights and parameters of each wavelet are determined optimally, by means of a fitting technique, which minimizes the energy function

$$E = \min_{\mathbf{n}_i, w_i \forall i} \|f - (\sum_i w_i \psi_{\mathbf{n}_i} + \bar{f})\|_2^2 \quad (2)$$

with respect to the weights $w_i \in R$ and wavelet parameters $\mathbf{n}_i \in R^5$. In the equation above, $\bar{f}$ is the DC-value of f . The Levenberg-Marquard gradient descent method [11] was employed to determine the optimal wavelet network for the face template. The method might get stuck in local minima and a careful selection of the initial parameters is important.

Then, we can say that the two optimized vectors $\Psi = (\psi_{\mathbf{n}_1}, \dots, \psi_{\mathbf{n}_M})^T$ and $\mathbf{w} = (w_1, \dots, w_M)^T$ define an optimized Gabor Wavelet Network ($\Psi, \mathbf{w}$) for a specific face image f . The wavelet representation for f may be considered as the reconstruction of the original image and it is given by:

$$\hat{f} = \sum_{i=1}^M w_i \psi_{\mathbf{n}_i} + \bar{f} \quad (3)$$

The precision of the face representation is determined by the number M of used wavelets. For instance, if we use a short number of wavelets, we obtain a coarse GWN, which may work well in different individuals and may be suitable for real-time applications. As we increase the number of wavelets, the representation becomes more specific, encoding more precise object information. Another aspect of the GWN representation is that it is invariant to some degree to affine deformations of the face image, as we will see in the next section. Furthermore, since the odd-Gabor Wavelets are DC-free, they are invariant to some degree to homogeneous illumination changes.

The Gabor functions are biologically motivated [12] and provide the best possible tradeoff between spatial and frequency resolution (Heisenberg principle). Besides, they act as good feature detectors, encoding texture and geometrical information in the representation.

Figure 1(a) shows a face template and its discretized representation is illustrated in figure 1(b), which we call the Gabor wavelet template (GWT). This representation was obtained by using a GWN of just $M = 52$ odd-Gabor wavelets, initialized in the inner face region. Figure 1(c) shows the position of the 16 largest wavelets, after optimization.

4 Face matching by the GWN

In the previous section, we have shown how a continuous wavelet representation for a face template is obtained based on a Gabor Wavelet Network. Now, we will see how this representation can match a new face image so that its wavelets are registered on the same facial features as in the original

Fig. 1. (a) The face template. (b) The wavelet representation obtained by the GWN. (c) Position of the 16 largest wavelets.

image. This process, which is called GWN repositioning, is done by applying a suitable affine deformation on the entire wavelet representation. It will be used both for positioning and for tracking the facial landmarks.

For instance, consider the face template shown in figure 1(a) and let G be its optimized GWN. Now, consider this face image in a different pose as shown in figure 2(a). In the repositioning process, the set of wavelets of G are positioned correctly on the same facial features in the distorted image. It is important to emphasize that the GWN repositioning may determine the parameters (translation, scale, rotation and shearing) of any affine deformation applied to the original image. Figure 2(b) shows the repositioned discrete face template representation (GWT), with 52 odd-Gabor wavelet functions. Figure 2(c) illustrates the position of the 16 largest wavelets of G in the image.

Fig. 2. (a) Face template in a different pose. (b) Repositioned wavelet representation. (c) Position of the 16 largest wavelets.

The repositioning of a GWN in a new image, i.e., the determination of the correct affine parameters, is established by using a superwavelet [13]. Let $\Psi = (\psi_{\mathbf{n}_1}, \dots, \psi_{\mathbf{n}_M})$, $\mathbf{w} = (w_1, \dots, w_M)$ be an optimized GWN. A Gabor superwavelet $\Psi_{\mathbf{n}}$ (GSW) may be defined as a linear combination of the wavelets $\psi_{\mathbf{n}_i}$ such that

$$\Psi_{\mathbf{n}}(\mathbf{x}) = \sum_i w_i \psi_{\mathbf{n}_i}(\mathbf{S}\mathbf{R}(\mathbf{x} - \mathbf{c})) \quad (4)$$

where the parameters of vector $\mathbf{n}$ of the GSW Ψ define the dilation matrix $\mathbf{S}$, the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{c}$ with:

$$\mathbf{S} = \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix}, \mathbf{R} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \mathbf{c} = (c_x, c_y)^T.$$

Thus, a Gabor superwavelet $\Psi_{\mathbf{n}}$ is again a wavelet that has the typical parameters, i.e. dilation s_x, s_y , translation c_x, c_y and rotation θ . So, for a given

new image g , we may arbitrarily deform the superwavelet by optimizing its parameter vector $\mathbf{n}$ so that the wavelet representation matches the face in image g . This is done by minimizing the energy function below, using the Levenberg-Marquard algorithm:

$$E = \min_{\mathbf{n}} \|g - \Psi_{\mathbf{n}}\|_2^2 \quad (5)$$

It is important to note that the parameters of a wavelet include only translation, dilation and rotation. Even so, we may include shearing and thus allow any affine deformation of GSW $\Psi_{\mathbf{n}}$. For this, we add the parameter s_{xy} to vector $\mathbf{n}$ and rewrite the scaling matrix:

$$\mathbf{S} = \begin{pmatrix} s_x & s_{xy} \\ 0 & s_y \end{pmatrix}.$$

5 Locating Facial Landmarks

In this section, we address the problem of locating facial landmarks by using the GWN repositioning procedure described in the previous section. Our experiments were carried out on the Yale face database (<http://cvc.yale.edu/projects/yalefaces/yalefaces.html>), which consists of 15 different subjects with 11 images per person, showing different gestures. Because the database contains only grey-level images, we segmented the face region by hand. Experiments with color images and automatic face detection will be presented in the next section.

When we optimize a coarse GWN (with a short number of wavelets) on a face template, we obtain a wavelet representation that may be repositioned in different individuals. According to our experiments, the repositioning procedure never fails when the target face image is obtained from the same person with different facial expressions in which the GWN was optimized. However, we can not guarantee that a coarse GWN optimized considering a specific individual can always be repositioned in any person. The repositioning process depends heavily on the similarity between the person in which the GWN was optimized and the test person as well as on the number of used wavelets.

The solution that we propose for this problem is to optimize the GWN considering a mean face. We have normalized and averaged a set of 15 faces of the Yale database, corresponding each one to a different individual.

Using the GWN optimized considering this mean face, the repositioning worked well on all individual images of the Yale database. We are now investigating how to address the repositioning problem in any person considering very large face databases.

The proposed method for locating facial landmarks is now described. Initially, facial feature points are located in the mean face. We have considered the pupils, center of nose and center of mouth. Then, the GWN related to the mean face is repositioned in the target face image. In order to determine

the facial landmarks in the test face image, we apply a suitable affine transformation to the initial facial feature points of the mean face. The correct parameters of this transformation are obtained from the superwavelet parameter vector $(s_x, s_y, s_{xy}, c_x, c_y, \theta)$, which is determined by equation (5) in the repositioning process.

The obtained results show the robustness of the method. Figure 3 illustrates the positioning of facial landmarks in three individuals of the Yale database. It is worth saying that the method may work even in the presence of glasses and beard. Furthermore, figure 4 illustrates the results considering face images of the same person under different expressions and illumination changes. Section 7 presents more discussions about this technique.

Fig. 3. Facial landmarks positioning in different individuals.

Fig. 4. Facial feature positioning under different expressions and illumination changes.

6 Tracking Facial Features

The GWN repositioning described in section 4 may be applied to an image sequence, allowing affine face tracking. Thus, for each frame J_t at time step t , the Gabor superwavelet $\Psi_{\mathbf{n}_t}$ is optimized according to the energy function:

$$E = \min_{\mathbf{n}_t} \|J_t - \Psi_{\mathbf{n}_t}\|_2^2 \quad (6)$$

The parameter vector $\mathbf{n}_{t-1}$ is used as initial value for optimization in the frame J_t . As image changes are small from frame to frame, the optimization process converges quickly. Initial values for $\mathbf{n}_0$ in the first frame are derived from a color blob information.

Our tracking method assumes that the facial landmarks have been correctly determined in the first frame of the sequence, as described in the previous section. We then apply, in each frame, the suitable affine transformation to the located feature points, performing facial landmark tracking. The parameters of the affine transformation are obtained by means of the superwavelet parameter vector in each frame of the sequence.

It is important to emphasize that the procedure to track facial landmarks considers the overall face geometry, thus being robust to deformations such as eye blinking and smile, which is usually a critical situation to most local-based traditional approaches. It is also important to note that the introduced approach can be straightforwardly generalized in order to track additional feature points or even regions, such as arbitrary polygons around the eyes, nose and mouth.

7 Experiments and Discussion

As discussed in section 1, our approach can be divided in three subsequent steps: face detection, facial landmarks positioning and tracking of face and facial landmarks. The first step is performed by a skin-color approach as well as by a simple correlation procedure to verify the presence of a face in the located skin-blob [2]. Once the face was detected, its scale information is obtained and the color face region is converted to a grey-level image. Facial landmarks are then located by repositioning a GWN into the face region. The position and scale of the face-like blob are used as initial parameters in the repositioning procedure. Finally, face and facial landmarks are tracked along the video sequence as described in the previous section.

We have tested our method in different color video sequences, obtaining good results (<http://www.ime.usp.br/~rferis>). Figure 5 shows, in the left illustration, the detection of a face in the initial frame. The right illustration shows the facial landmarks positioning, which was achieved by repositioning the GWN optimized considering the mean face of the Yale database. Tracking of regions around the facial landmarks (eyes, nose and mouth) is illustrated in figure 6, which presents the frames 60, 98 and 221 of a specific video sequence. Note that the method is robust to eye blinking, homogeneous illumination changes and different facial expressions. We are still verifying the performance of the system so that future work will cover more experimental results as well as comparison with other systems.

Concerning facial landmarks positioning, our method has basically two limitations. The first one is related to GWN repositioning, for this procedure may fail when large databases are involved. The second limitation is derived

from the fact that the distance among facial features varies from person to person. This may lead to imprecise facial landmarks location in some cases. So, it would be interesting to have another procedure that could adjust the position of the selected facial feature points after the GWN repositioning. These topics belong to our ongoing research work.

The tracking of facial landmarks may be imprecise under strong 3D face pose variation. On the other hand, it showed to be robust and even suitable for real-time applications, when a small number of wavelets is considered in the representation. We intend to use this approach for face recognition from video sequences.

Fig. 5. Locating a face and its facial landmarks.

Fig. 6. Tracking eyes, nose and mouth.

8 Conclusions

This paper described a method for locating and tracking facial landmarks using Gabor Wavelet Networks. The method is based on a continuous wavelet representation of a discrete face template, which is invariant to some degree to illumination changes and affine deformations of the face image.

The obtained results confirmed the robustness of the method. Positioning of facial landmarks in the initial frame may be accomplished even in the presence of glasses, beard and different facial expressions. The tracking approach considers the overall geometry of the face so that it is robust to facial feature deformations such as eye blinking and smile. As future work, we intend to use the GWN to perform face detection.

Acknowledgements

Roberto M. Cesar Junior is grateful to FAPESP for the financial support (98/07722-0 and 99/12765-2), to “pro-reitoria de pesquisa” and to “pro-reitoria de pós-graduação” - USP, as well as to CNPq (300722/ 98-2). Rogerio Feris is grateful to FAPESP (99/01487-1).

We are grateful to Volker Kruger for providing the source code of GWN technique and for discussions. Some images in the paper were derived from the Yale face database.

References

1. Kruger V. and Sommer G. (1999) Affine real-time face tracking using a wavelet network. Proceedings of ICCV'99 Workshop Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, Corfu, Greece.
2. Feris R., Campos T. and Cesar R. (2000) Detection and tracking of facial features in video sequences. Lecture Notes in Artificial Intelligence, vol. 1793, pp. 127-135, Springer-Verlag.
3. Yang J. and Waibel A. (1996) A real-time face tracker. Proceedings of the Third IEEE Workshop on Applications of Computer Vision, pp. 142-147, Sarasota, Florida.
4. Yang J., Lu W. and Waibel A. (1997) Skin-color modeling and adaptation. CMU CS Technical Report, CMU-CS-97-146.
5. Stiefelhagen R. and Yang J. (1996) Gaze tracking for multimodal human computer interaction. University of Karlsruhe. Available at <http://werner.ira.uka.de/ISL.multimodal.publications.html>
6. Silva L., Aizawa K. and Hatori M. (1995) Detection and tracking of facial features. Proceedings of SPIE Visual Communications and Image Processing, Taiwan.
7. Maurer T. and Malsburg C. (1996) Tracking and learning graphs and pose on image sequences of faces. Proceedings of Int. Conf. on Artificial Neural Networks, Bochum.
8. Mallat S. (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Analysis and Machine Intelligence, 11(7):674-693.
9. Daubechies I. (1990) The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. Informat. Theory, 36(5):961-1004.
10. Zhang Q. and Benviste A. (1992) Wavelet networks. IEEE Trans. on Neural Networks, 3(6):889-898.
11. Press W., Flannery B., Teukolsky S. and Vetterling W. (1986) Numerical Recipes, The Art of Scientific Computing, Cambridge University Press, UK.
12. Daugman J. (1985) Uncertainty relation for resolution in space, spatial frequency, and orientation optimized two-dimensional visual cortical filters. Journal Opt. Soc. Am., 2(7):1160-1168.
13. Szu H., Telfer B. and Kadambe S. (1992) Neural network adaptive wavelets for signal representation and adaptation. Optical Engineering, 31(9):1907-1961.

A Framework for Face Recognition from Video Sequences Using GWN and Eigenfeature Selection

Rogério Schmidt Feris

Teófilo Emídio de Campos

Roberto Marcondes Cesar Jr.

I Workshop on Artificial Intelligence and Computer Vision

International Joint Conference IBERAMIA'2000 and SBIA'2000

19-22 de Novembro, 2000

Atibaia, Brasil

A Framework for Face Recognition from Video Sequences Using GWN and Eigenfeature Selection

Teófilo Emídio de Campos
Rogério Schmidt Feris
Roberto Marcondes Cesar Junior

{teo, rferis, cesar}@ime.usp.br
<http://www.vision.ime.usp.br/~creativision>
Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo (USP),
Rua do Matão, 1010, PO-Box 66.281,
05315-970, São Paulo, SP, Brasil

Abstract. This work describes an ongoing project related to a new method to perform face recognition from video sequences. Faces are detected and tracked in a video sequence using Gabor wavelet networks. This process also allows locating and extracting facial feature regions around the eyes, nose and mouth. A modified Karhunen-Loève transform, defined with the aid of an automatic feature selection procedure, is used for feature extraction and face recognition. Several preliminary results are discussed.

Keywords: face detection, tracking of facial features, face recognition, Gabor Wavelet Networks, Principal Components Analysis and Feature Selection.

1 Introduction

Face recognition has emerged as an instigating research field both under the machine and the biological vision systems point of views. In fact, this research field has attracted intense and growing attention by the vision research community, finding many important practical applications, such as in security and human-machine interaction. This paper describes an ongoing project, including some preliminary results, about a face recognition system based on video sequences.

The methods for face recognition can be broadly grouped in two classes, i.e. static and dynamic approaches. The former concerns face recognition in single (static) images, while the latter concentrates on people possibly moving in video sequences. The static

approach has developed many interesting and powerful techniques, succeeding in different applications. Nevertheless, as far as recognition in video sequences is concerned, much work still remains to be done [13].

Face recognition in video sequences often involves four important steps: (1) face detection; (2) face tracking along the video sequence; (3) feature extraction; and (4) recognition. It should be clear that the large amount of data involved in video sequences represent a challenge for real-time implementation of these four steps. In order to circumvent this problem, not all face regions will be taken into account by the foreseen system. Instead, the system is based on extracting information from important facial feature regions (FFRs) defined around specific landmarks, i.e. the eyes, the nose and the mouth.

Figure 1 Overview of the project.

Figure 1 illustrates an overview of the system project. Using Gabor wavelet networks (GWN), the important facial landmarks (FLs) are detected and

tracked, allowing the FFRs to be segmented and extracted. The features used for classification are extracted from the FFRs using the Karhunen-Loève transform. Nevertheless, instead of taking the eigenvectors (or eigenfeatures) associated to larger eigenvalues, an automatic feature selection algorithm will be applied in order to define a better transform from the training set, by choosing the most discriminative eigenfeatures. Once that the feature extraction transform has been defined from this procedure, it can be applied to the FFRs in order to generate the feature vectors that allow classifying the faces in each frame. Finally, a super classifier is supposed to collect the output classification results from each frame in order to recognize the face based on several classification results (one from each frame, in a voting scheme). All these steps are represented in Figure 1, which also includes a systematic performance assessment module that will be carried out in order to validate the system.

This work is organized as follows. Section 2 describes the GWN approach for face detection, tracking and location of the facial landmarks. Section 3 concentrates on feature extraction and selection. Finally, the face classification in each frame, the super classifier and performance assessment are topics presented in Section 4.

2 Detection and Tracking of Facial Features

The detection and tracking of facial features in video sequences is an important step of our approach, since it provides the normalized facial feature images, which are required by both the training and classification procedures. The topics concerned to this section can be divided in three parts: detection of faces, facial features and normalization of facial features. It is important to say that the user is supposed to move his head naturally while the camera acquires the image sequence, with non-controlled illumination conditions.

Face detection is performed by using a statistical skin-color model to segment face-candidates as well as a simple correlation procedure to verify the presence or absence of a face in each detected skin-color blob. Once a face is detected, its approximated scale and position are computed, and the skin-color region is converted into a gray-level image. These procedures are described in [1].

Facial features, i.e. eyes, nose and mouth, are then located and tracked along the video sequence. The method that we have used to accomplish this task is based

on Gabor wavelet networks (GWNs) [2], an effective technique for object representation. Basically, the idea of GWN is to represent a face image as a linear combination of 2D Gabor wavelets, whose parameters (position, scale and orientation) are stored in the network nodes, while the linear coefficients are represented as the synaptical weights. The weights and wavelet parameters are determined optimally so that the maximum of image information is preserved for a given number of wavelets. This wavelet representation allows detection of facial features even in the presence of glasses, beard and different facial expressions. Feature tracking is robust to homogeneous illumination changes and affine deformations of the face image. Moreover, the approach considers the overall geometry of the face, thus being robust to deformations such as eye blinking and smile, which is usually a critical situation to most local-based traditional methods.

Figure 2 illustrates the face and facial feature detection, whereas figure 3 shows the tracking process in a specific video sequence¹. The technique based on Gabor wavelet networks provides, in each frame, the orientation, scale, translation and shearing of facial features. Thus, using these parameters, we can segment and normalize each facial feature by applying a correct affine transformation in each frame. The resulting data are used by both the training procedure and the classification, as we will describe in the next sections.

Figure 2 Detection of face and facial features.

Figure 3 Tracking of facial features.

¹ <http://www.ime.usp.br/~cesar/creativision/demo/>

3 Feature Space Generation

In order to perform recognition, our system generates a feature space according to the method schematized in figure 4. Next two subsections contain further details about our methodology.

3.1 Eigenfeatures

After estimating the facial features locations, followed by segmentation and normalization of facial features regions from all the frames of the training set, the resultant images are used to train four principal components analysis (PCA) feature extractors (also called Karhunen-Loève extension), one for each facial feature, because PCA generates a representation better suited to classification procedures.

Figure 4 Feature space generation.

These PCA transforms are obtained through the method described by Turk and Pentland [3], but using only images of facial features instead of whole face images thus obtaining eigenfeatures. The terms eigenfeatures, eigeneyes, eigennoses and eigenmouths were created by Moghaddam and Pentland in [11]. The

advantages of using only facial features includes computational cost, because smaller images require less processing to train the PCA; and recognition accuracy, due to the curse of dimensionality [9]. These advantages were reported in [4], and Brunelli and Poggio [7] also confirms these results for the template matching approach.

In order to perform dimensionality reduction, Turk and Pentland [3], simply select the first d eigenvectors of the covariance matrix of the training faces set ($d < m$, where m is the total number of eigenvectors). In this work, we apply a feature selection algorithm over a combination of all the eigenfeatures (eigenrighteyes, eigenlefteyes, eigennoses and eigenmouths).

Besides the well-know advantages of feature selection, like those discussed in [9], one of the motivations for performing automatic eigenfeature selection are the results obtained in [8], in which the performance of a PCA for face recognition system was improved when the first three eigenfaces were not used. The authors in [8] claim that there are some evidences that these eigenfaces are influenced by illumination changes, and not to inter-class variations. This fact provides evidence that it is possible to obtain better results by applying an automatic feature selection method to the eigenfeatures, instead of using the first d eigenvectors of each facial feature.

3.2 Feature Selection

Briefly, automatic feature selection is an optimization technique that, given a set of m features, attempts to select a subset of size d ($d < m$) that lads to the maximization of some criterion function, e. g. the distance among the training classes.

In [12], we described a feature selection strategy that produces good classification results without dependence of a specific classifier. These results are possible because that algorithm selects a feature set that maximizes distances among elements belonging to different classes and minimize distances among elements that belong to the same class. An advantage of that technique is that it is possible to obtain these results independently of the shape of distribution of the patterns in the feature space.

Briefly, the technique proposed in [12] is composed by the sequential floating search methods [5], using, as criterion function, the tolerance-based fuzzy distance [6]. The membership value of each pattern was defined as being inversely proportional to the distance between it

and the prototype of its class. We defined the prototype of a class as the mean vector of the training samples of that class.

4 Classification

In order to perform classification of video sequences, the feature space is obtained by transforming the input video sequence using the PCA transform of section 3. Each individual frame is classified and a super-classifier is used to give the final classification result of the video sequence.

An evaluation procedure will be used to assess and improve the performance by determining which classifier is the best between the K -nearest-neighbor (Knn) and the minimum distance to the prototype. The evaluation system will also be used to determine the value of the parameter K for the Knn classifier. These classifiers are described in [10].

We use the voting scheme [9] as a super-classifier (combination method), but we plan to assess the performance of other super-classifiers, like those described by Jain et al. in [9].

Our ongoing work also includes performance assessment using the "leaving-one-out" strategy [10] and the mean of the correct classification rates.

We have obtained promising results applying the feature selection technique described in previous section over eye images and using a Knn classifier. Our preliminary tests were carried out using 174 images of eyes from 29 people, 6 images per person. The resolution of the images is 64 x 64 pixels. All images were used to train a principal components analysis system (PCA), obtaining 468 eigenvectors.

The ASF feature selection was done to select the best (according to the tolerance-based distance [6]) $d=30$ eigenfeatures. The obtained feature set was evaluated using the Knn classifier, for $K=3$, trained using all the images and testing using one image per person. The recognition accuracy was 97.7%, while, by using the first 30 eigenvectors (the traditional PCA classification method), the obtained recognition accuracy was 94.8%. This first result shows that our feature selection strategy is promising.

5 Concluding Remarks

We presented a new method to perform face recognition from video sequences. Our approach is based on: (1) detection using a statistical skin-color model to segment faces and GWN to detect facial landmarks; (2) tracking of facial landmarks using GWN and (3) recognition using eigenfeatures selection.

The main contributions of this work are the eigenfeatures selection for people recognition, and the association among GWN, PCA and feature selection using fuzzy distance.

As described before, the preliminary results of the detection and tracking systems as well as the results of the eigenfeature selection system are very promising. As future work, we plan to perform tests comparing the performance of eigenfeatures selection versus eigenfeatures without automatic feature selection. We also plan to compare it with other systems concerning face recognition from video sequences.

Acknowledgements

T. E. Campos is grateful to FAPESP for the financial support (99/01488-8), R. S. Feris is grateful to FAPESP (99/01487-1) and R. M. Cesar J. is grateful to FAPESP (98/07722-0) as well as to CNPq (300722/98-2).

References

- [1] R. S. Feris, T. E. Campos and R. M. Cesar Jr.; "Detection and tracking of facial features in video sequences"; *Lecture Notes in Artificial Intelligence*, vol. 1793, pp. 127-135, Springer-Verlag, April 2000.
- [2] V. Kruger and G. Sommer; "Affine real-time face tracking using a wavelet network"; Presented at the *ICCV'99 Workshop Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, Corfu, Greece, September 1999.
- [3] M. A. Turk and A. P. Pentland; "Face recognition using eigenfaces"; *Proc. of the IEEE on Computer Soc. Conf.*, 1991.
- [4] T. E. Campos, R. S. Feris, R. M. Cesar Jr.; "Eigenfaces versus Eigeneyes: First Steps Toward Performance Assessment of Representations for Face Recognition"; *Lecture Notes in Artificial Intelligence*, vol. 1793, pp 197-206, Springer-Verlag press, April 2000.

- [5] P. Pudil, J. Novovicová and J. Kittler, "Floating search methods in feature selection", *Pattern Recognition Letters* 15 (1994), pp. 1119-1125.
- [6] R. Lowen and W. Peeters; "Distances between fuzzy sets representing gray level images"; *Fuzzy sets and Systems*, 99(2), pp. 135-150, 1998.
- [7] R. Brunelli, T. Poggio; "Face recognition: features versus templates"; *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10), pp. 1042-1052, October 1993.
- [8] P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman; "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection"; *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), pp. 711-720, July 1997.
- [9] A. Jain, R. P. W. Duin, J. Mao; "Statistical pattern recognition: a review"; *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), pp. 4-37, January, 2000.
- [10] R. Duda, P. Hart, "Pattern Classification and Scene Analysis", Wiley, New York, 1973.
- [11] Baback Moghaddam, Alex Pentland; "Face Recognition using View-Based and Modular Eigenspaces"; Automatic Systems for the Identification and Inspection of Humans, *SPIE* Vol. 2277, July 1994.
- [12] T. E. Campos, Isabelle Bloch, R. M. Cesar Jr.; "Feature Selection Based on Fuzzy Distances Between Clusters: First Results on Simulated Data", *submitted to Second International Conference on Advances in Pattern Recognition*, Rio de Janeiro, Brazil, March, 2001. Preprint available at: <http://www.ime.usp.br/~cesar/creativision/>.
- [13] R. Chellappa, C. L. Wilson and S. Sirohey; "Human and Machine Recognition of Faces: A Survey"; *Proceedings of the IEEE*, 83(5), pp. 703-740, May 1995.

Improved Face x Non-Face Discrimination using Fourier Descriptors through Feature Selection

Teófilo Emídio de Campos

Rogério Schmidt Feris

Roberto Marcondes Cesar Jr.

SIBGRAPI'2001 International Brazilian Symposium on Computer
Graphics, Image Processing and Computer Vision

IEEE Computer Society Press

16-20 de Outubro, 2000

Gramado, Brasil

Improved Face $\times$ Non-Face Discrimination Using Fourier Descriptors Through Feature Selection

TEÓFILO EMÍDIO DE CAMPOS
ROGÉRIO SCHIMIDT FERIS
ROBERTO MARCONDES CESAR JUNIOR

Instituto de Matemática e Estatística
Universidade de São Paulo (IME/USP),
Rua do Matão, 1010, PO-Box 66.281,
05315-970 São Paulo, SP, Brasil
{teo, rferis, cesar}@ime.usp.br

Abstract. This work presents a new method to discriminate face from non-face images using Fourier descriptors. The first step of our approach consists in applying a horizontal edge detection filter in the input image, followed by the extraction of a 1D signal from the computed edge map. Then we calculate the Fourier descriptors from this signal and classify the image using statistical classifiers. In order to improve our results, we applied a feature selection algorithm. Preliminary performance assessment results have shown that this approach is superior than traditional transform based methods. Besides, these results showed that our method might be used to develop a fast face detection system.

1 Introduction

Human face detection is an active research area in the computer vision community, with important applications in automatic face recognition, visual surveillance and man-machine interfaces. Furthermore, a very rich literature on biological face recognition can be found due to the importance of this task for humans and other animals.

Among a large number of proposed techniques, we can cite the work of Rowley, Baluja and Kanade [1], which uses a neural network to perform frontal face detection. Sung and Poggio [2] also presented a technique for locating vertical frontal views of human faces, with an example-based learning approach. These approaches provide good results, but, like the majority of the methods of the literature, their computational cost is very large, being not suitable for real-time face detection.

Recently, the work of Wu, Chen and Yachida [3] described an effective method to detect faces in color images based on the fuzzy theory. Another color-based approach to detect and to track faces is described in the

work of Feris, Campos and Cesar [11]. These color-based approaches are efficient with respect to real-time processing. But they are not robust to presence of lots of skin color objects near the face in the image. Further problems of this approach include the fact that color images involves larger data sets than gray scale images and that the most of security systems use gray scale cameras instead of color cameras.

In this paper, we propose a new method to discriminate face from non-face images using Fourier descriptors. The preliminary obtained results are very promising, showing that we can use our approach to perform a task such as face detection, by sliding a window at different image locations and scales, or by detecting skin-color blobs, i.e., face candidates.

In order to discriminate face from non-face images, we have used Fourier descriptors, an effective technique to perform shape analysis, which has been successfully applied in a large number of different problems (see the paper of Zahn and Roskies [4] for further information). Basically, given an image to be classified as face or non-face, we first apply a horizontal edge detector, followed

by the extraction of a 1D signal from the computed edge map. We then calculate Fourier descriptors from this signal and classify the image using the Fourier descriptors to form a feature vector. Preliminary results showed that our approach might be very useful to develop a fast and robust face detection system, mainly because only a few coefficients have been used in the classification. To improve our results, we applied a feature selection algorithm based on a search method proposed recently, and using, as the criterion function, the performance of the minimum distance to the prototype classifier.

The remainder of this paper is organized as follows. Section 2 describes the classification method and the assessing experiments, while Section 3 presents the experimental results. Finally, in Section 4 follows a conclusion, where we address further research directions.

2 The face/non-face discrimination method

The experiments carried out in order to investigate the proposed technique are summarized schematically in figure 1. The basic idea of the proposed approach is to extract a 1D signal from the horizontal edge map and to characterize the face patterns by a set of Fourier descriptors extracted from this signal. The obtained features have been evaluated by training a statistical classifier and classifying a test set in order to obtain the classification accuracy. In order to improve the system's performance, it was applied a feature selection algorithm over the Fourier coefficients. Each step is described below.

2.1 Face database formation

A set of 219 almost upright face images and of 219 non-face images has been used. The images do not present necessarily the same illumination conditions because different face databases that can be found in the Internet have been merged.

Some images created for our database have also been included. The latter images have been acquired using a FujifilmDX-7 digital camera, with resolution of 640 X 480 pixels. The images have been hand-cropped so that only eyes nose and mouth are taken into account. The face images database and the pictures considered have the property that all the obtained face crops are larger than 128 x 128 pixels. An example is shown in figure 2.

2.2 Non-face database formation

The non-face image database has been created by randomly searching for images in the WWW. A vision research software environment, called Synergos, has been

implemented (see [5]), which includes computational tools for searching and storing WWW images.

Synergos includes a web-worm-like robot that makes the access to HTML documents, i.e. that implements the http protocol in order to exchange data with WWW servers. The resources that Synergos may extract by parsing a HTML document include other HTML documents, downloadable files, images, ftp links, emails, newsgroups and so on. An important feature of this approach is that the nature of the obtained images vary a lot, for the search engine navigates through many different web pages depending only on the hypertext structure (i.e. the URL's of the visited page).

Figure 1 General overview of the experiments architecture.

Once that a WWW page is visited, the Synergos gets the images therein. In order to normalize the database, the following procedure is adopted: Synergos selects only images larger than 256 X 256. Then, it chooses randomly a block of size 256 X 256 from the selected image and this block is stored in the database. In order to verify that no face image has been selected, a human operator examines the database (though images that present faces, among other features, are allowed).

Figure 2 Input image obtained with a digital camera (a) and its respective cropped face image containing only eyes, nose and mouth (b).

2.3 Size normalization

All images from both databases are normalized with respect to size, so that all images are 128 X 128 before the subsequent steps. There are many reasons for performing size normalization. First, for classification purposes, between faces and non-faces, it is important to normalize size because comparisons between features of the feature space are actually made.

Furthermore, the search for a minimum size for feature extraction is also very important in order to save computational demand. In fact, because the proposed approach rely on the discrete Fourier transform, it is also

important to have signals of suitable size, so that Fast Fourier Transform (FFT) algorithms can be adopted.

Finally, normalizing the image size provides a homogenization for using the same parameter values among the different imaging algorithms for all input images. Size normalization is a step commonly used in many different face detection procedures, such as those described by Sung and Poggio [2].

2.4 Detecting of horizontal edges

Although some previous works on face detection and recognition have already suggested the use of image edges (like Moghaddam and Pentland [6], and Kondo and Yan [7]), the experiment reported in this work shows that very good correct discrimination rates between faces and non-faces can be achieved using only the horizontal edges of face images.

The horizontal edges preserve mainly the information of eyes, eyebrows, mouth and the region around the nostrils. There are two basic approaches for exploring the information associated to horizontal edges: using an edge emphasized image (e.g. differentiated); and using an edge map (hence, a binary image), which can be obtained, for instance, by thresholding or by searching for local maxima of the edge emphasized image.

Figure 3 Horizontal edge map.

The traditional methods for edge enhancement and detection include the gradient based and Laplacian based techniques, such as Roberts, Sobel and the Laplacian-of-Gaussian. In the experiment reported in this paper, a simple Sobel mask for detecting horizontal edges has been used, providing very good results.

Figure 3 presents the edge map of figure 2 (b) using Sobel. It is important to note that the visual information regarding the face structure is preserved.

2.5 1D signal formation

The main idea behind the method proposed here is to characterize the edge patterns formed from the edge detection step. An approach borrowed from 2D shape analysis techniques is based on extracting some 1D signature from the pattern followed by a linear transform of this signal, say the Fourier or the wavelet transform (see Antoine et. al. for further details [8]).

There are many different ways of defining the signal, such as Scholl diagrams, shape matrices, the pattern image itself and so on. Because the experiment discussed in this work assumes vertical and frontal faces, an invertible representation of the edge map can be obtained taking the first column of the image, followed by the second, the third, and so on.

Let $h(p, q)$ be the $m \times n$ edge map, with $h(p, q) = 1$ if the pixel (p, q) is an edge point and 0 otherwise, where $p = 0, 1, \dots, (m-1)$ and $q = 0, 1, \dots, (n-1)$. Therefore, the characterizing signal is defined as:

$$g(p + mq) = h(p, q)$$

Clearly, $g(i)$, $i = 0, 1, \dots, (mn)$ is a discrete time binary signal, or an impulse train. Figure 4 shows the signal $g(i)$ obtained from the edge map of Figure 3.

Figure 4 1D signal (impulse train) representing the edge map of figure 3.

2.6 Fourier descriptors

The problem that must be tackled now is how to characterize signals $g(i)$ obtained from the edge maps in order to discriminate faces from non-faces. Clearly, a simple point-to-point comparison between the signals would not lead to good results for two very similar edge maps could have very few impulses coinciding exactly at the same time locations.

A preferable approach is to define and obtain the Fourier descriptors of $g(i)$ [4] and to use them as a feature vector together with a statistical pattern classifier. There are many different Fourier descriptors (FD's) in the literature, and they may differ a lot depending on the

problem in hand. In this work, we define the FD's of $g(i)$ as:

$$G(s) = \log \left(\left| \sum_{i=1}^{mn} g(i) e^{-j2\pi \frac{si}{mn}} \right| \right), s \in S, S \subset F$$

where F is the set of all the FD's, S is a subset of F . The FD's are based on the modulus of the complex coefficients of the Fourier transform of $g(i)$. Adopting the modulus equips the FD's with the nice property of invariance with respect to time shifting of $g(i)$ (which only affects the phase of the Fourier transform). The $\log$ function is important because the coefficient amplitudes may vary a lot, which could be a problem when using them together in feature vector comparisons by the statistical classifier. The $\log$ function attenuates such large coefficient variations. Finally, it is important to note that the FD's defined above do not include the 0-th DC component, which only depends on the number of edge points in the image, despite their spatial organization.

2.7 Feature selection

The above mentioned feature extraction process produces a large number of features (FD's) that can be used for classification. Nevertheless if we use a large number of FD's, the performance of the statistical classifier will decrease with respect to execution time and to recognition rate. The execution time increases with the number of features because of the measurement cost. The recognition rate can decrease because of redundant features and of the fact that small number of features can alleviate the course of dimensionality when the number of training samples is limited. On the other hand, a reduction in the number of features may lead to a loss in the discrimination power and thereby lower the accuracy of the recognition system. Jain, Duin and Mao's paper [14] made a good review about these facts.

Once that real-time face detection is aimed, it is desirable to define a feature space that can lead to good classification rate with a simple and fast classifier. Therefore, it is necessary to keep our pattern representation (Fourier space) as small as possible. In this context, the problem is how to select the best FD's set.

In order to investigate this question, we made some tests with the first 30 FD's obtained through our technique. Basically, given the desirable feature set size d , these tests are divided into 3 categories, that will be described in the next subsections.

2.7.1 First d FD's

This is the most straightforward ways to select features in a Fourier space for image analysis. It is based on the fact that the most of the signals' energy from real images are concentrated on the low frequencies, so we can discard high frequencies, which are redundant for doing classification.

2.7.2 Largest d FD's

This approach is often adopted in the context of recognition using wavelets. It is based on the idea that the most important coefficients of a linear transform are the largest coefficients. Therefore, the FD's are chosen by doing a threshold. In case of wavelet research, an analogous technique is known as wavelet shrinkage.

In our case, we selected these features by first determining the mean vector of all the samples of the face class and taking the index of the d largest elements of this vector. Alternative voting schemes can be adopted if the features are not so homogeneous for specific classes.

2.7.3 Feature selection using search methods

Automatic feature selection is an optimization technique that consists on, given a set of m features, select a subset of size d that leads to the maximization of some criterion function. In the case of this work, we used the performance of a classifier as criterion function, as explained in section 3.

There are many different feature selection methods. A well know algorithm is the Branch and Bound (BB), that is an optimal method for monotonic feature sets. Nevertheless its computational cost is prohibitive for large feature sets: in the worst case, its time complexity is exponential on the dimension of the feature space.

In 1997, Jain and Zongker published a review paper with a taxonomy and a comparative study about the feature selection methods [10]. According to them, probably the most effective feature selection techniques was the sequential floating search methods (SF), proposed by Pudil et al. in 1994 [12].

Basically, in the case of forward search (SFFS), the algorithm starts with a null feature set and, for each step, the best feature that satisfies some criterion function is included with the current feature set. The algorithm also verifies if the criterion can be improved if some feature is excluded. Therefore, the SFFS proceeds dynamically increasing and decreasing the number of features until the desired d is reached. The backward search works analogously, but it starts with the full feature set. The time complexity of these methods is linear.

In 1999, Somol et al. proposed some improvements on these algorithms [13] and created the adaptive floating search methods (ASF). The difference between this new approach and the previous is on the number of features that can be inserted or excluded from the feature set. The new algorithms can determine dynamically this number, while the earlier tests one feature per step. The advantage of ASF is that its results are closer to the optimum solution than the results of SF methods. However, for large values of d , ASF is very slow.

Both methods do not have the restriction that the feature space has to be monotonic, like the BB method.

For these reasons, we choose to do tests with automatic feature selection algorithms based on SF and ASF. Our tests comparing the performance of feature spaces generated by these four aforementioned methods will be explained in section 3. We omit further detail about the SF and ASF algorithms because of paper length restrictions. The reader is referred to [12 and 13].

2.8 Statistical classifier

A statistical classifier together with the feature vector formed by the FD's can be used in order to discriminate between face and non-face images. We implemented the "minimum distance to the prototype" statistical classifier in which one prototype per class is defined, where each prototype is the average of all training samples. The advantage of this classifier is its computationally efficiency (its time complexity is $O(n)$, for n being the number of features). This classifier is further described in the Duda and Hart book [9].

Besides the simplicity of this classifier, it is possible to reach a good correct classification performance. This fact is illustrated in the next section.

3 Experimental results

We have used the "minimum distance to the prototype" statistical classifier to obtain our results. The database contains 219 face images and 219 non-face images, all of them have resolution normalized to 128 x 128 pixels.

Our tests are summarized in the next tables. The results considered the recognition rate on classifying faces and non-faces in percentages. We performed tests varying the size of the desired feature set d , and using the four methods to determine the feature set.

Let Λ be the learning set of our classifier, T the test set and U be the full set of the database. We did tests with $\Lambda \cap T = U$ and others tests with $|\Lambda| = 66,67\%$ of $|U|$ and $|T|$

= 33,33% of $|U|$ and $\Lambda \cap T = \{\}$. The elements of the sets Λ and T was chosen randomly.

In order to apply the feature selection (SF and ASF methods), we used as criterion function the classification rate of the statistical classifier based on minimum distance to the prototype. This criterion function perform the classification on sets that $\Lambda \cap T = \{\}$. Therefore, the SF and ASF methods work to maximize the classification rate using learning and testing sets with no intersection.

Table 1 shows the some of the obtained results for a face database $|U| = 219$ and non-face database $|U| = 219$. Figure 5 resumes of the results of the minimum distance to the prototype classifier when tested in images that was not in the learning set ($\Lambda \cap T = \{\}$) for all the values of d tested.

4 Concluding remarks

We have proposed a new method to discriminate face from non-face images using Fourier descriptors that show very promising results.

The best results were obtained through ASF methods, being superior to the traditional methods to perform classification using Fourier coefficients. The most traditional approach for selecting the FD's coefficients is by using the first d 's and it presented the worst results (except for $d \geq 15$, in this case, the "largest d " is the worst).

Another important fact is that, for all tested methods, the best results was obtained for $d < 20$. This result confirms that increasing the number of features does not guarantee that the performance of the classifier also increases.

The main difference between the performance of the SF and ASF was 4,3478%, for $d=13$ and $d=15$. But in the worst case (concerning execution time), SF spent 2 seconds do reach the best subset (for $d=25$), while, in the same computer, ASF spent 4 hours, 22 minutes and 10 seconds (for $d=15$). These facts show that if the priority is to obtain fast results on the feature selection, it is better to choose the SF algorithms. It is important to see that even if SF is not the best feature selection algorithm, its results are better than the results of "First d " and "Largest d ", for all cases.

An important fact is that 30 is the total number of features available and, for this reason, all the methods leads to the same result when $d=30$.

It is important to emphasize three facts about the face database:

- these tests were done using hand cropped face images, without precise registration;

- the faces were not exactly upright;
- the considered database is composed by several face databases obtained with different equipment and under different illumination conditions.

Table 1 Classification rate tables (%).

$d=3$	$\Lambda \cap T=U$	$\Lambda \cap T=\{\}$
First d	66,1836	71,0145
Largest d	66,1836	71,0145
SF	69,5652	75,3623
ASF	69,5652	75,3623

$d=9$	$\Lambda \cap T=U$	$\Lambda \cap T=\{\}$
First d	74,8792	73,9130
Largest d	74,8792	73,9130
SF	78,7440	76,8116
ASF	81,1594	78,2609

$d=15$	$\Lambda \cap T=U$	$\Lambda \cap T=\{\}$
First d	79,2271	59,4203
Largest d	79,7101	57,9710
SF	85,5072	75,3623
ASF	86,4734	79,7101

$D=21$	$\Lambda \cap T=U$	$\Lambda \cap T=\{\}$
First d	82,6087	50,7246
Largest d	81,6425	49,2754
SF	85,5072	75,3623
ASF	85,5072	75,3623

$d=27$	$\Lambda \cap T=U$	$\Lambda \cap T=\{\}$
First d	85,5072	56,5217
Largest d	84,5411	55,0725
SF	84,5411	65,2174
ASF	84,0580	66,6667

As expected, these database characteristics are very prejudicial to the classifier's performance. Nevertheless, they show the robustness of our system with natural conditions for face detection.

Recognition rate with with no intersection between learning and testing sets

Figure 5 Results of the classifier for $\Lambda \cap T = \{\}$.

The best result obtained was the recognition rate of 86,4734% ($d=15$ and $\Lambda \cap T=U$ with ASF), which is a very promising result. Nevertheless, the main contribution of this work is the speed of the method. For vectors with only 3 dimensions, it is possible to detect faces with an accuracy of 74,3243% minimum distance to the prototype classifier.

As a future work, we plan to apply other feature extraction methods, other criterion function for feature selection and to perform these tests using other classifiers.

5 Acknowledgements

Teófilo E. Campos is grateful to FAPESP for the financial support (99/01488-8), Rogério S. Feris is grateful to FAPESP (99/01487-1) and Roberto M. Cesar J. is grateful to FAPESP (98/07722-0) as well as to CNPq (300722/98-2).

We are grateful to P. Somol, P. Pudil, J Novovicová and P. Paclík for providing some source codes.

6 References

- [1] H.A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(1) (1998), 23--38.
- [2] K.K. Sung and T. Poggio, "Example-based learning for view-based human face detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(1) (1998) 39--55.
- [3] H. Wu, Q. Chen, and M. Yachida, "Face Detection From Color Images Using a Fuzzy Pattern Matching Method", *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21(6) (1999) 557--563.
- [4] C.T. Zahn & R.Z. Roskies, "Fourier descriptors for plane closed curves", *IEEE Trans. on Computers* C21 (1972), 269--281.
- [5] O. M. Bruno, R.M. Cesar Jr., L.A. Consularo, and L. da F. Costa, "Automatic Feature Selection for Biological Shape Classification in SYNERGOS", *Proc. Brazilian Conference on Computer Graphics, Image Processing*

and Vision, (SIBGRAPI-98, Rio de Janeiro - RJ, Out 1998), IEEE Computer Society Press (1998), 363--370.

[6] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7) (1997), 696--710.

[7] T. Kondo, H. Yan, "Automatic human face detection and recognition under non-uniform illumination", *Pattern Recognition* 32(10) (1999), 1707--1718.

[8] J.-P. Antoine, D. Barache, R.M. Cesar Jr., L. da Fontoura Costa, "Shape characterization with the wavelet transform", *Signal Processing*, 62(3) (1997), 265--290.

[9] R.O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, NY, 1973.

[10] A. Jain and D. Zongker, "Feature selection - evaluation, application, and small sample performance", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2) (1997), 153--158.

[11] R. S. Feris, T. E. Campos and R. M. Cesar J., "Detection and tracking of facial features in video sequences", *Lecture Notes in Artificial Intelligence*, vol. 1793 (2000) 127--135, Springer-Verlag.

[12] P. Pudil, J. Novovicová and J. Kittler, "Floating search methods in feature selection", *Pattern Recognition Letters* 15 (1994), 1119--1125.

[13] P. Somol, P. Pudil, J. Novovicová and P. Paclík, "Adaptive floating search methods in feature selection", *Pattern Recognition Letters* 20 (1999), 1157--1163.

[14] A. Jain, R. P. W. Duin and J. Mao, "Statistical pattern recognition: a review", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, no. 1 (2000), 4--37.

Detection and Tracking of Facial Features in Video Sequences

Rogério Schmidt Feris

Teófilo Emídio de Campos

Roberto Marcondes Cesar Jr.

Mexican International Conference on Artificial Intelligence

Lecture Notes in Artificial Intelligence (LNAI)

Springer-Verlag press

10-14 de Abril, 2000

Acapulco, Mexico

Detection and Tracking of Facial Features in Video Sequences

Rogério Schmidt Feris
Teófilo Emídio de Campos
Roberto Marcondes Cesar Junior

{rferis, teo, cesar} @ime.usp.br
CreatiVision - Creative Vision Research Group
Department of Computer Science
DCC-IME-USP, University of São Paulo
Rua do Matão, 1010, São Paulo, SP, 05508-900, Brazil
Phone: +55 (011) 818 6235, Fax: +55 (011) 818 6134
© Springer-Verlag

Abstract. This work presents a real time system for detection and tracking of facial features in video sequences. Such system may be used in visual communication applications, such as teleconferencing, virtual reality, intelligent interfaces, human-machine interaction, surveillance, etc. We have used a statistical skin-color model to segment face-candidate regions in the image. The presence or absence of a face in each region is verified by means of an eye detector, based on an efficient template matching scheme. Once a face is detected, the pupils, nostrils and lip corners are located and these facial features are tracked in the image sequence, performing real time processing.

1 Introduction

The detection and tracking of faces and facial features in video sequences is a fundamental and challenging problem in computer vision. This research area has many applications in face identification systems, model-based coding, gaze detection, human-computer interaction, teleconferencing, etc.

In this paper, we present a real time system that performs facial features detection and tracking in image sequences. We are interested in developing intelligent interfaces and this system is the first step to give computers the ability to look at people, which may be a good way to improve human-computer interaction.

We have used a color-based approach to detect the presence of a human face in an image. A Gaussian distribution model was generated by means of a supervised training, using a set of skin-color regions, obtained from a color face database.

Basically, the input image is compared with the generated model in order to identify face candidates, i.e., skin-color regions. The presence or absence of a face in each region is verified by using an eye detector based on an efficient template matching scheme. There is biological evidence that eyes play the most important role in human face detection [1].

Once a face is detected, the pupils, nostrils and lip corners are located and tracked in the video sequence. The system is very fast and robust to small face pose and rotation variation.

The remainder of this paper is organized as follows. Section two reviews some techniques related to our work. Section three describes the statistical skin-color model

and section four explains how it was used to segment face-candidate regions. Section five describes the eye detector used to perform the face verification process. In section six, the detection and tracking of facial features is described. The obtained results are showed in section seven. Finally, in section eight follows a conclusion and we address further research directions.

2 Related Work

Automatic detection of facial features in video sequences is performed, in general, after a face is detected in a frame. Many approaches have been proposed to detect faces in static scenes, such as Principal Component Analysis [1,2], clustering [3] and neural nets [4]. These techniques obtained good detection / false alarm rates, but they are computational expensive and hardly achieve real-time performance.

Liyanage Silva et. al. [5] proposed a method, which they called edge pixel counting, to detect and track facial features in video sequences. This method is based on the fact that a higher edge concentration is observed in the vicinity of facial features, while slightly outside of such features a less edge concentration is observed. In spite of its simplicity, the method fails in the presence of cluttered backgrounds, glasses, hat, etc.

Recently, several color-based systems have been proposed to perform face detection and tracking in image sequences. Processing color is much faster than processing other facial features and, furthermore, color is orientation invariant under certain lighting conditions. This property makes motion estimation much easier since only a translation model is needed.

The work of Jie Yang and Alex Waibel [6] presents a statistical skin-color model, adaptable to different people and different lighting conditions. This work was extended to detect and track faces in real time [7], achieving a rate of 30+frames/second using a HP-9000 workstation with a framegrabber and a Canon VC-C1 camera.

Nuria Oliver et. al. [8] and Trevor Darrell et. al. [9] also obtained good results using skin-color to detect and track people in real time. Stiefelhagen and Yang [10] presented a method similar to ours, which detects and tracks facial features in image sequences. The main differences lie in face verification process. They have used iterative thresholding to search the pupils whereas we have adopted an efficient template matching scheme to detect eyes.

3 Skin-color Model

The statistical skin-color model [6] is generated by means of a supervised training, using a set of skin-color regions, obtained from a color face database, with 40 face images. Such images were obtained from people of different races, ages and gender, with varying illumination conditions. Figure 1 illustrates the training process, in which a skin-color region is selected and its RGB representation is stored. It was verified, using training data, that skin-colors are clustered in color space, as illustrated in Figure 2.

It is common to think that different people have skin-colors which differ significantly from each other due to the existence of different races. However, what really occurs is a larger difference in brightness / intensity and not in color [6]. Thus,

it is possible to reduce the variance of the skin-color cluster through intensity normalization:

$$a = \frac{R}{R + G + B} \quad (1)$$

$$b = \frac{G}{R + G + B} \quad (2)$$

The colors (a,b) are known as chromatic or “pure” colors. Figure 3 illustrates the skin-color cluster in chromatic space.

Fig. 1. Selected skin-color region

Fig. 2. Cluster in color space

Fig. 3. Cluster in chromatic space

According to [6], the skin-color distribution in chromatic space can be approximated by a Gaussian model $N(m, \Sigma^2)$, where $m=(a_{avg}, b_{avg})$ and:

$$a_{avg} = \frac{1}{N} \sum_{i=1}^N a_i \quad (3)$$

$$b_{avg} = \frac{1}{N} \sum_{i=1}^N b_i \quad (4)$$

$$\Sigma = \begin{bmatrix} \sigma_{aa} & \sigma_{ab} \\ \sigma_{ba} & \sigma_{bb} \end{bmatrix} \quad (5)$$

The Gaussian model, which was generated from training data, is illustrated in Figure 4.

Fig. 4. Gaussian Model

4 Image Segmentation

The segmentation process consists in identifying skin-color regions in the input image (the first frame of the video sequence) based on the generated model.

Initially, all pixels in input image are converted to the chromatic space. Using the Gaussian model, a gray-level image is generated and the intensity of each pixel in this new image represents its probability to have skin-color. Figure 5 illustrates this process.

Since skin-color pixels present a stronger response in the gray-level image, we can identify such regions by a thresholding process. The threshold value was chosen empirically.

In general, due to noise and distortions in input image, the result of thresholding may generate non-contiguous skin-color regions. Furthermore, after this operation, spurious pixels may appear in the binary image. In order to solve these problems, first we have applied a morphological closing operator to obtain skin-color blobs. Subsequently, a median filter was used to eliminate spurious pixels. Figure 6 shows the obtained binary image after thresholding (left illustration) and the result of morphological closing operation followed by a median filtering (right illustration).

Fig. 5. Skin-color identification

Fig. 6. Thresholding and morphological closing followed by median filtering

Boundaries of skin-color regions are determined using a region growing algorithm in the binary image. Regions with size less than 1% of image size are eliminated. Furthermore, structural aspects are considered so that regions that do not present the face structure are removed.

In the next section, we will show a method to decide whether each detected skin-color region corresponds or not to a face.

5 Verification Process

The main goal in identifying skin-color regions in the input image is to reduce the search space for faces. However, it is important to note that not all detected regions contain faces. Some correspond to parts of human body, while other correspond to objects with colors similar to those of skin. Thus, it is necessary to verify the presence or absence of a face in each detected region.

To accomplish this task, we have used an eye detector based on an efficient template matching scheme. It is worth saying that there is biological evidence that eyes play the most important role in human face detection [1]. The eye template used is illustrated in Figure 7. Its resolution is 57x15 with 256 gray levels.

Fig. 7. Eye template

The eye detector procedure first converts the skin-color region to a gray-level image, in order to perform template matching. The most important property in our

technique is that the eye template is resized according to the skin region size. In general, existing methods use several templates with different scales to accomplish the detection. For instance, the work of Brunnelli and Poggio [11] uses five templates with scales 0.7, 0.85, 1, 1.5 and 1.35 to detect eyes in an image. Although these methods have obtained good results, they are computational expensive and not suitable to real time processing.

Our method estimates the person eye size, resizing the eye template. Thus, the template matching is performed only once, using the upper portion of the detected skin-color region. Basically, correlation coefficients ρ are computed in different positions within the skin region:

$$\rho = \frac{\text{cov}(x, y)}{\sigma(x)\sigma(y)} \quad (6)$$

where x is the eye template and y is the image patch in a specific position within the skin region. The function cov in the equation above returns the covariance between x and y . The correlation coefficient ρ assume values ranging from -1 to 1.

A threshold value, which was determined empirically, was used to decide eye detection. When eyes are found within the skin-color region, we report the presence of a face. The system then detects some facial features and proceeds in the tracking module, as described in the next section.

It is important to note that the above strategy detects the first frame in which the face appears in frontal view.

6 Detection and Tracking of Facial Features

Once a face is detected, the system proceeds with the search for pupils, nostrils and lip corners. After, these facial features are tracked in the video sequence. The approach used is similar to the work of Stiefelhagen and Yang [10].

6.1 Searching the Pupils

Assuming a frontal view of the face initially, we locate the pupils by looking for two dark pixels that satisfy certain geometric constraints and lie within the eye region detected by template matching.

6.2 Searching the Lip Corners

First, the approximate positions of lip corners are predicted, using the position of the face and the pupils. A small region around those points is then extracted and a Sobel horizontal edge detector is applied. The approximate horizontal boundaries of the lips are determined by first computing the vertical integral projection P_v of this horizontal edge image:

$$P_v(y) = \sum_{x=1}^H E_h(x, y), \quad 0 \leq y \leq W \quad (7)$$

where $E_h(x,y)$ is the intensity function of the horizontal edge image, and W and H are the width and height of the search region, respectively.

The approximate left and right boundaries of the lips are located where P_v exceeds a certain threshold t or respectively falls below that threshold. We choose t to be the average of the projection P_v . The vertical position of the left and right lip corners are simply located by searching for the darkest pixel along the columns at the left and right estimated boundaries of the lips in the search region.

6.3 Searching the Nostrils

Similar to searching the pupils, the nostrils are located by searching for two dark regions, which satisfy certain geometric constraints. The search region is restricted to an area below the pupils and above the lips.

Figure 7 illustrates the detection of facial features.

Fig. 7. Detection of facial features

6.4 Tracking

For tracking pupils and nostrils in the video sequence, simple darkest pixel finding in the search windows around the last positions is used. In order to track lip corners, the same detection process is performed in a small region around the last lip corners position.

7 Obtained Results

The proposed real time facial feature tracker obtained good results in different video sequences. Initially, a color-based frontal view face detection is performed. This procedure fails, in general, when the eye template scale is incorrectly estimated. This case occurs when the skin region contains, beyond a face, other human body parts, such as a hand occluding the face, or other objects. We have also noted that faces may be missed when they are too small or under strong varying illumination.

The face detection procedure has proven to be robust in identifying skin-colors of people with different races, ages and gender. It is worth saying that dark-skin regions had smaller skin-color probabilities, since most persons in the face database had light skin.

Once a face is detected, its facial features are located and tracked in the image sequence. If there is more than one face in the first frame, only the bigger face is

considered.

During the tracking process, we verified that the system fails when a large movement is performed by the person. Furthermore, due to varying illumination, some features may be missed in the sequence. On the other hand, the tracking module showed to be robust to small face pose and rotation variations.

Figure 8 shows the tracking process in some frames of a image sequence, which can be observed in <http://www.ime.usp.br/~cesar/creativision>.

Fig. 8. Tracking Process

8 Conclusions

This paper described a real time system for detection and tracking of facial features in video sequences. A statistical skin-color model was used to detect face-candidate regions and an efficient template matching scheme was applied to report the presence or absence of a face in these regions. Given a detected face, pupils, nostrils and lip corners are located and tracked in the image sequence. The system may be used in different applications, such as teleconferencing and human-computer interaction.

Future work includes pose estimation and 3D head tracking by using the positions of pupils, nostrils and lip corners. Furthermore, we intend to improve the system robustness developing a recovery module and considering motion to increase the accuracy of detection and tracking of facial features.

Acknowledgments

Roberto M. Cesar Jr. is grateful to FAPESP for the financial support (98/07722-0), as well as to CNPq (300722/98-2). Rogério Feris and Teófilo Campos are grateful to FAPESP (99/01487-1 and 99/01488-8).

We are grateful to Carlos Hitoshi Morimoto (IME-USP) for useful discussions.

References

- [1] B. Moghaddam and A. Pentland. "Face Recognition using View-Based and Modular Eigenspaces". Proc. of Automatic Systems for the Identification and Inspection of Humans, SPIE vol. 2277, July 1994.
- [2] B. Moghaddam and A. Pentland. "Probabilistic Visual Learning for Object Detection". In Fifth International Conference on Computer Vision, pp. 786-793, Cambridge, Massachusetts, June 1995
- [3] K. Sung and T. Poggio. "Example-base Learning for View-based Human Face Detection" C.B.C.L. Paper no. 112, MIT, 1994.
- [4] H. Rowley, S. Baluja and T. Kanade. "Neural Network-based Face Detection". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp. 23-38, January 1998.
- [5] L. Silva, K. Aizawa and M. Hatori. "Detection and Tracking of Facial Features". Proc. of SPIE Visual Communications and Image Processing, Taiwan. May, 1995.
- [6] J. Yang, W. Lu and A. Waibel. "Skin-Color Modeling and Adaptation". CMU CS Technical Report, CMU-CS-97-146. May, 1997.
- [7] J. Yang and A. Waibel. "A Real-time Face Tracker". Proc. of the Third IEEE Workshop on Applications of Computer Vision, pp. 142-147, Sarasota, Florida, 1996.
- [8] N. Oliver, A. Pentland and F. Berard. "LAFTER: Lips and Face Real-time Tracker with Facial Expression Recognition". Proc. of Computer Vision and Pattern Recognition Conference, 1997.
- [9] C. Wren, A. Azarbayejani, T. Darrell and A. Pentland. "Pfinder: Real-time Tracking of the Human Body". Proc. SPIE, vol. 2615, pp. 89-98, 1996.
- [10] R. Stiefelhagen and J. Yang. "Gaze Tracking for Multimodal Human Computer Interaction". University of Karlsruhe, 1996. Available at <http://werner.ira.uka.de/ISL.multimodal.publications.html>.
- [11] R. Brunelli and T. Poggio. "Face Recognition: Features versus Templates". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 10, pp. 1042-1052. October 1993.