Rastreamento facial em vídeo com aplicação em monitoramento e Segurança

Alessandro G. F. Fior, Maurício P. Segundo, Luciano Silva, Olga R. P. Bellon*
Grupo IMAGO de Pesquisa – Universidade Federal do Paraná
Caixa Postal 19092 – CEP 81531-980 – Curitiba/PR – Brasil
{agff05,mauricio,luciano,olga}@inf.ufpr.br

Resumo

Neste trabalho, apresentamos uma ferramenta para rastreamento de faces em seqüências de vídeo que é componente de um sistema de monitoramento e Segurança baseado em informações multibiométricas. A ferramenta tem como diferencial uma forte integração com a etapa de detecção de faces, anterior ao rastreamento. Inicialmente, a aplicação busca por faces no vídeo combinando detecção de movimento, filtragem de pele, e um classificador baseado em características de Haar para diferenciar imagens faciais e não-faciais. Este mesmo classificador é aplicado para o rastreamento, uma vez que as características de Haar mantêm informações do padrão das faces. Nenhuma informação adicional é necessária para o rastreamento, apenas as características já calculadas durante a etapa de detecção. Os experimentos foram realizados em uma base própria de seqüências de vídeo adquiridas em um ambiente interno. Com a abordagem proposta, a porcentagem de faces com a posição corretamente estimada foi de 94,18%.

1. Introdução

Os sistemas de reconhecimento biométrico através de faces em vídeo [4, 5, 9] compreendem, de modo geral, três fases: (1) detecção de faces [10, 11], (2) rastreamento das faces ao longo do vídeo [12] e (3) reconhecimento da identidade das faces [2]. O rastreamento tem como função garantir a continuidade das faces em uma seqüência de vídeo, assegurando que uma mesma face encontrada em vários quadros do vídeo pertence à mesma pessoa. Para isso, o rastreador estima a localização de uma face mesmo que ela não seja encontrada pelo detector em um ou mais quadros.

Nesse trabalho é proposto um método de rastreamento facial cujo diferencial é a integração com a etapa de detecção. O método é baseado em características de Haar [10], amplamente utilizadas para a detecção facial. O mesmo conjunto de características aplicado na detecção é utilizado no rastreamento, que é mais robusto a variações ao longo das seqüências de vídeo (*e.g.* variações de iluminação e expressão) do que a detecção.

Os métodos de rastreamento de alvos disponíveis na literatura utilizam diferentes tipos de informação, como por exemplo, histograma da intensidade dos pixels do alvo [3] ou a covariância da intensidade e/ou gradiente dos pixels do alvo [8]. Ao utilizar algumas destas abordagens, é necessário calcular características específicas para o rastreamento, enquanto a abordagem proposta utiliza as mesmas características já calculadas para a detecção de faces.

O objetivo desde trabalho é monitorar pessoas que circulam em um determinado ambiente. Neste contexto, para a realização dos experimentos, foi criada uma base própria contendo seqüências de vídeo. A criação dessa base foi necessária pois não havia nenhuma base pública disponível com as características desejadas: pessoas andando em direção à câmera, com variações de iluminação, expressão e dimensão das faces.

O método de rastreamento proposto é apresentado na Seção 2 deste artigo. Os resultados experimentais são mostrados na Seção 3, seguidos de algumas considerações finais e referências utilizadas.

2. Rastreamento facial em vídeo

O método de rastreamento proposto neste artigo é baseado em uma cascata de classificadores formados por características de Haar. Estes classificadores são gerados inicialmente para realizar a detecção de faces [10], mas fornecem um vetor característico que permite diferenciar a região detectada das demais regiões da imagem.

2.1. Características de Haar para detecção facial

Neste trabalho, 7 tipos de características de Haar foram utilizadas [6]. Seus valores representam a diferença do somatório de pixels em diferentes áreas de uma máscara, que varia conforme o seu tipo. Estas características são aplicadas com diferentes dimensões e posições na imagem.

* Os autores gostariam de agradecer ao CNPq, CAPES e FINEP pelo suporte financeiro.

Quando estas características são aplicadas em padrões similares (*e.g.* faces), estes valores tendem a ser próximos. Logo, eles podem ser utilizados para identificar se uma imagem pode ou não ser de um determinado padrão. Embora uma única característica de Haar não seja suficiente para diferenciar grandes quantidades de imagens de faces e não-faces, classificadores mais eficientes podem ser obtidos combinando várias características.

A partir de uma base contendo imagens de faces e não-faces, o algoritmo *AdaBoost* [10] é aplicado para encontrar as características que melhor diferenciam imagens de faces e não-faces e distribuir estas características em classificadores. Os primeiros classificadores gerados possuem menos características que os seguintes, e assim consecutivamente. A Fig. 1 mostra dois exemplos de classificadores com diferentes quantidades de características de Haar.

(a) (b)

Figura 1. Exemplo de classificadores contendo (a) 2 e (b) 3 características de Haar.

Para a detecção, uma subjanela percorre a imagem selecionando como faces apenas as regiões aceitas por todos os classificadores gerados, que são aplicados em ordem crescente de complexidade para melhorar o desempenho do método. Outra forma de otimizar a detecção consiste em realizar a busca apenas em regiões da imagem que apresentem movimento [1] e cor de pele [7].

2.2. Rastreamento facial

Quando uma face f é detectada, esta pode ser representada como um vetor de L classificadores, e cada classificador como um vetor de valores das características de Haar $C_i^f = [h_f(i,1), h_f(i,2), ..., h_f(i,M_i)]$, onde C_i^f é o i-ésimo classificador da face f, contendo M_i características de Haar, e $h_f(i,j)$ é o valor da j-ésima característica deste classificador.

Para melhorar o desempenho do rastreador, apenas um subconjunto dos classificadores é utilizado. Segundo Yao e Li [12], quando um detector é construído em cascata, os primeiros classificadores utilizam características grosseiras, enquanto os últimos classificadores representam melhor a face. Por este motivo, neste trabalho apenas os L' últimos classificadores são usados para o rastreamento.

Esta representação é gerada para cada subjanela em uma vizinhança em torno da face detectada. A distância E entre uma face f e uma subjanela g é dada pela soma das distâncias Euclidianas entre os classificadores em f e seus respectivos classificadores em g, como mostrado na Eq. 1:

$$E(f,g) = \sum_{i=L-L'+1}^{L} \sqrt{\sum_{j=1}^{M_i} (h_f(i,j) - h_g(i,j))^2} \quad (1)$$

A posição e o tamanho da face rastreada correspondem aos valores da subjanela g do quadro atual com a menor distância da face f. A representação da face é atualizada para o próximo quadro segundo a Eq. 2:

$$f_{t+1} = \alpha g + (1-\alpha) f_t \quad (2)$$

onde α é a taxa de aprendizado do rastreador, f_t é a representação atual da face, e f_{t+1} é a nova representação.

Entretanto, estas atualizações serão necessárias somente se a face rastreada não for detectada no quadro atual. Se a face for detectada, o rastreamento é utilizado apenas para garantir que esta pertence à seqüência do mesmo indivíduo. Neste caso, a posição e o tamanho da face a ser rastreada são atualizados com as informações da face detectada h, e a representação é atualizada segundo a Eq. 3:

$$f_{t+1} = \beta h + (1-\beta) f_t \quad (3)$$

onde β é a taxa de aprendizado usada para faces detectadas.

3. Resultados experimentais

Para a realização dos experimentos, foi criada uma base de validação contendo 20 seqüências de vídeo de 5 indivíduos. Cada seqüência possui em média 127 quadros, e foi adquirida por uma câmera de segurança AXIS 207MW a 18Hz com resolução de 1280×720 pixels. A Fig. 2 mostra um dos quadros de uma seqüência da base.

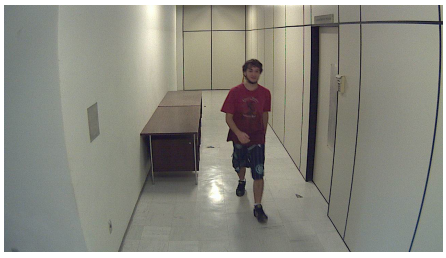

Figura 2. Exemplo de um quadro de uma seqüência de vídeo da base.

Estas seqüências contêm um indivíduo andando em um corredor, e as faces apresentam variações como oclusões parciais e diferenças de iluminação, escala, pose e expressões. Algumas destas variações são mostradas na Fig. 3.

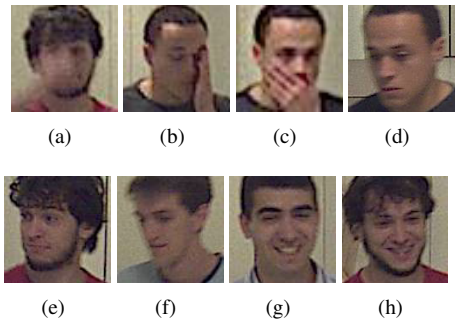

Figura 3. Imagens apresentando (a)-(c) oclusões parciais, (d)-(f) variações de pose, e (g)-(h) expressões faciais.

3.1. Detecção facial

Para treinar o detector, foi utilizada uma base de treino com 4483 imagens de faces e 74065 imagens de não faces. As imagens de treino são diferentes das contidas na base de validação. A cascata resultante contém 264 características de Haar distribuídas em 28 classificadores.

Em todas as seqüências da base de validação as faces são detectadas. Os dois primeiros classificadores da cascata rejeitam em torno de 82% das imagens de não-faces. Em média, apenas 4 características foram aplicadas por subjanela devido à organização em cascata dos classificadores.

3.2. Rastreamento facial

Para esta etapa dos experimentos, cada face foi marcada manualmente nas seqüências para serem utilizadas como referência para o rastreamento. A partir destas referências, foi possível medir o erro em pixels do deslocamento nos eixos x e y, e da escala para cada seqüência. Aplicando um limiar de tolerância neste erro, é possível determinar se a face foi rastreada corretamente. Esse limiar é necessário porque podem haver pequenas diferenças entre a marcação manual e a posição determinada pelo rastreador. Nos resultados apresentados nessa Seção, utilizamos um limiar de 15% do tamanho da face. Ou seja, se a distância entre os centros das faces rastreada e de referência ou a diferença entre seus tamanhos forem maiores do que 15% do tamanho da face de referência, então o rastreamento é considerado incorreto.

As seqüências de vídeo foram utilizadas para determinar os parâmetros α e β. Para estimar a taxa de aprendizado α do rastreador, o valor de β foi definido como 0 e diferentes valores de α foram testados. A Fig. 4 mostra a porcentagem média de acerto do rastreador ao longo do tempo para alguns destes valores testados. O tempo 0 representa a primeira face rastreada, e o tempo 1 a última face da seqüência. As outras faces foram distribuídas ao longo deste intervalo. Como pode ser observado, valores mais altos para α apresentam um melhor comportamento ao longo do tempo.

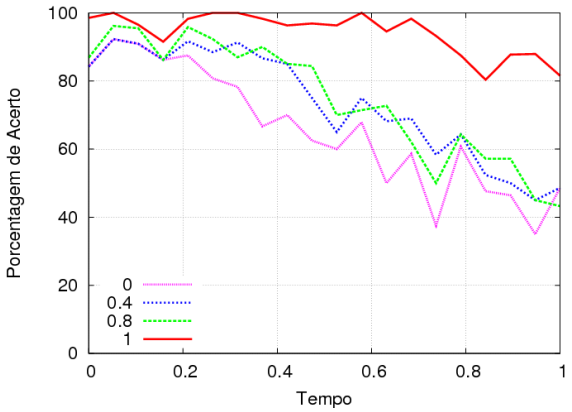

Figura 4. Erro médio do rastreador para diferentes valores de α.

O desempenho do rastreamento diminui no final das seqüências devido à baixa qualidade das imagens. A qualidade diminui por dois motivos: (1) borrões causados por movimentação acentuados, e (2) câmera com foco fixo, que deixa os objetos próximos a ela desfocados. A Fig. 5 mostra alguns exemplos destas imagens.

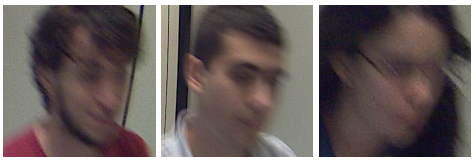

Figura 5. Exemplos de quadros do final das seqüências com baixa qualidade.

Para estimar a taxa de aprendizado β para faces detectadas, o valor de α foi definido como 0. Foram testados para β os mesmos valores testados para α. A Fig. 6 mostra a porcentagem média de acerto ao longo do tempo para alguns destes valores, e, assim como para o parâmetro α, valores mais altos para β apresentam um desempenho melhor.

Em nossos experimentos, observamos que os valores das taxas de aprendizado estão relacionados à velocidade de captura da câmera. Quanto maior a quantidade de quadros por segundo, menor é a variação entre quadros consecutivos, diminuindo a necessidade de atualizar o modelo de rastreamento. Conseqüentemente, as taxas de aprendizado podem ser menores, e os modelos tornam-se mais robustos a variações no vídeo.

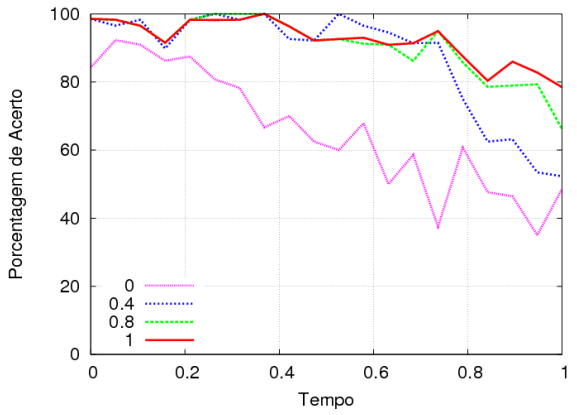

Figura 6. Erro médio do rastreador para diferentes valores de β.

A Fig. 7 mostra as faces detectadas e rastreadas de uma seqüência de vídeo da base. Como pode ser observado, muitos quadros não tiveram suas faces detectadas devido a variações de pose e oclusões, mas o rastreador foi capaz de estimar corretamente suas posições.

Figura 7. Faces detectadas (com borda preta) e rastreadas em uma das seqüências.

Computamos as porcentagens médias de acerto do rastreador para diferentes limiares de erro considerando a melhor configuração obtida a partir dos experimentos anteriores ($\alpha = 1$, $\beta = 1$). Para o limiar de tolerância de 15% utilizado nos experimentos anteriores, localizamos corretamente a face em 94,18% dos quadros, para um limiar de 10%, esta porcentagem cai para 83,41%, e considerando um limiar de 40%, a porcentagem de acerto é de 98,82%.

4. Considerações finais

Neste trabalho apresentamos um método de rastreamento de faces que utiliza as mesmas características aplicadas para a detecção facial. O método foi capaz de relacionar as faces detectadas em todas as seqüências de vídeo, e estimar corretamente a posição da face em aproximadamente 94% dos quadros, considerando um limiar de erro de 15%. Os experimentos mostram que o rastreamento é mais robusto que a detecção quando as faces apresentam variações como oclusões parciais, ou diferenças de iluminação, pose e expressões. Esta abordagem está inserida em um sistema de monitoramento e Segurança multibiométrico, responsável pela localização da informação a ser utilizada pelo reconhecimento facial.

Referências

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *CVIU*, 73(3):428–440, 1999.

[2] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and machine recognition of faces: a survey. In *Proc. IEEE*, volume 83, pages 705–741, 1995.

[3] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE PAMI*, 24(5):603–619, 2002.

[4] A. Hadid and M. Pietikäinen. An experimental investigation about the integration of facial dynamics in video-based face recognition. *Electronic Letters on Computer Vision and Image Analysis*, 5(1):1–13, 2005.

[5] V. Krueger and S. Zhou. Exemplar-based face recognition from video. In *Proc. FG*, pages 182–187, 2002.

[6] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Proc. ICIP*, pages 900–903, 2002.

[7] P. Peer, J. Kovac, and F. Solina. Human skin colour clustering for face detection. In *Proc. EUROCON*, 2003.

[8] F. Porikli, O. Tuzel, and P. Meer. Covariance tracking using model update based on lie algebra. In *Proc. CVPR*, pages 728–735, 2006.

[9] J. Steffens, E. Elagin, and H. Neven. PersonSpotter - fast and robust system for human detection, tracking and recognition. In *Proc. FG*, pages 516–521, 1998.

[10] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Computer Vision*, 57(2):137–154, 2004.

[11] M. H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE PAMI*, 24(1):34–58, 2002.

[12] Z. Yao and H. Li. Tracking a detected face with dynamic programming. In *Proc. CVPRW*, volume 5, pages 63–70, 2004.