4
Método Proposto

4.1.
Descrição geral do método

O objetivo do método proposto consiste em avaliar o potencial dos FERNS para detecção das características faciais, em tempo real, em uma seqüência de imagens adquiridas a partir de uma webcam comum.

A primeira parte do algoritmo consta de uma fase de treinamento offline, na qual, primeiramente, são detectados pontos característicos em uma imagem frontal da face e geradas vistas das possíveis aparências de cada um dos keypoints extraídos. Os keypoints são detectados dentro das áreas de interesse (olho direito, olho esquerdo e narinas) delimitadas, manualmente, pelo usuário na imagem de treinamento. Após isso, é realizado o treinamento dos FERNS de forma a possibilitar o reconhecimento dos keypoints adquiridos, em tempo real, de diferentes pontos de vista.

Na fase em tempo real, para cada quadro do vídeo, o reconhecimento dos keypoints é realizado utilizando-se o classificador previamente treinado. Uma etapa de segmentação da face é realizada a fim de melhorar a fase de reconhecimento desses pontos característicos, reduzindo o número de outliers provenientes desse processo. Os keypoints reconhecidos são agrupados, de acordo com a sua proximidade, de forma a possibilitar a detecção das regiões das características faciais. As regiões estimadas são utilizadas como espaço de busca para as características da face.

Detalhes do método proposto e da sua implementação são descritos nas seções abaixo.

4.2.
Fase de Treinamento

Como descrito em [17], se o objeto possui uma pequena curvatura uniforme, os patches ao redor de um keypoint podem ser tratados como localmente planares

e suas distorções em relação às mudanças de perspectivas podem ser vistas como homografias. A fim de testar os efeitos da não-planaridade, Özuysal et al. [18] aplicaram os FERNS em imagens de face assumindo sua planaridade para geração das vistas sintéticas e, de acordo com os autores, bons resultados foram alcançados, como pode ser visto na figura 5. Dessa forma, uma única imagem frontal do rosto é suficiente para gerar um conjunto de amostras de possíveis aparências da face tomadas de diversos pontos de vista.

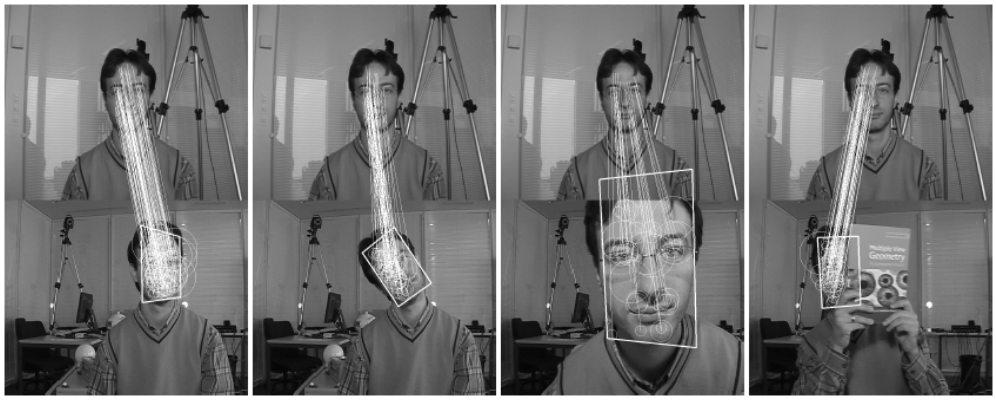

Figura 1: Correspondência de imagens de faces usando FERNS. Apesar da não planaridade das faces, o FERNS continua apresentando bons resultados. Imagem retirada de Özuysal et al. [18].

Como já mencionado no capítulo 3, a fase de treinamento começa pela detecção de pontos característicos (keypoints) na imagem de treinamento. A imagem de treinamento consta de uma foto frontal da face. O usuário delimita, manualmente, as áreas de interesse para o treinamento. As áreas de interesse utilizadas neste trabalho foram as regiões dos olhos e das narinas. Um exemplo de imagem de treinamento e algumas regiões delimitadas podem ser vistos na figura 6 (a) e 6 (b), respectivamente. A figura 6 (c) mostra os keypoints detectados nessa imagem. A detecção dos pontos característicos na imagem de treinamento é feita calculando-se o laplaciano da gaussiana da mesma maneira como foi explicada na seção 3.5. O número de keypoints extraídos pode ser escolhido pelo usuário. Nesta dissertação foram usados 400 keypoints.

Cada ponto característico extraído nessa imagem corresponde a uma classe diferente. As amostras de cada classe são dadas por um conjunto de pequenos retalhos da imagem (image patches) que representam as possíveis aparências do

keypoint. Esse conjunto de treinamento é construído como descrito na seção 3.4 do capítulo 3. Essas amostras são utilizadas no treinamento dos FERNS que retornam a probabilidade a posteriori de um patch pertencer a cada uma das classes aprendidas durante a fase de treinamento. Diferentemente do que acontece com um objeto planar como o da figura 1, a face possui movimentos de rotação restritos. Dessa forma, os ângulos utilizados para o cálculo das vistas sintéticas foram restritos ao intervalo de e não ao intervalo como proposto na seção 3.4. Dessa maneira, o conjunto de treinamento criado automaticamente através de homografias representa melhor as características que devem ser reconhecidas ao longo do vídeo adquirido em tempo real.

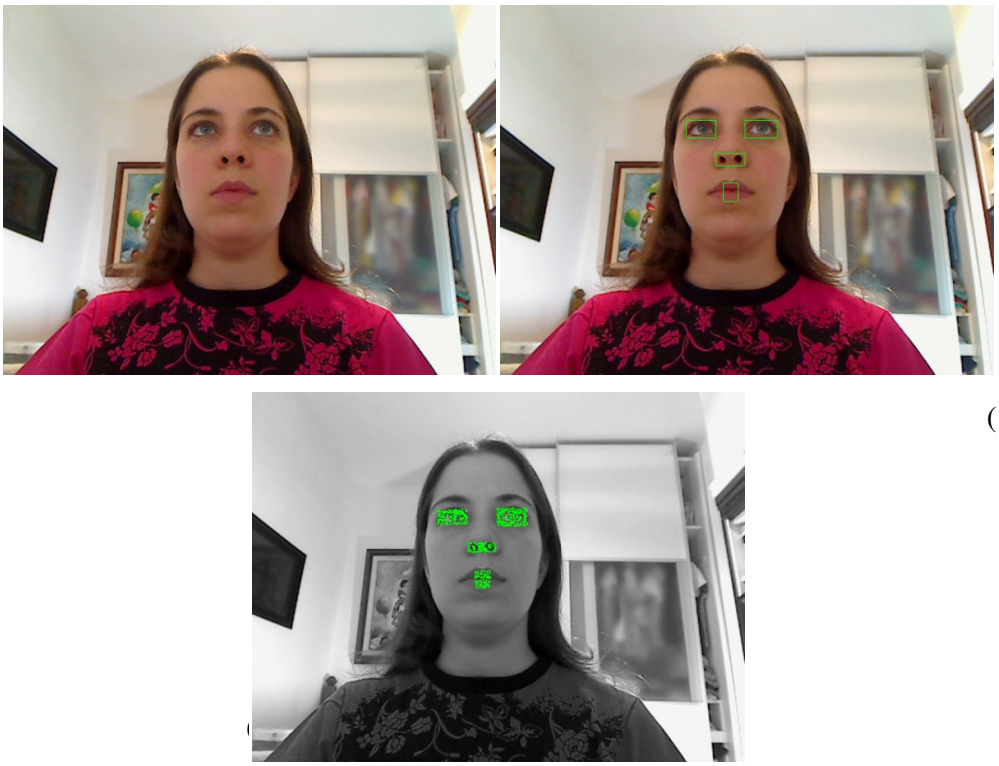

Figura 2: (a) Imagem de treinamento. (b) Seleção das áreas de interesse. (c) Pontos característicos extraídos da imagem de treinamento.

Como mencionado anteriormente, na seção 3.1, um único FERN não é suficiente para classificar todos os keypoints aprendidos. Dessa forma, diversos FERNS são utilizados com esse propósito. Nesta dissertação foram utilizados 30 FERNS com 12 testes binários por FERN. Com isso, as respostas de cada FERNS são combinadas de forma a possibilitar a classificação. Além disso, são utilizadas

quatro octaves, a fim de aumentar a robustez do classificador. Quanto maior o número de FERNS e de testes binários associados a cada um deles, maior é a confiabilidade do classificador. No entanto, a fase de treinamento acaba consumindo muito tempo, prejudicando a aplicabilidade do sistema.

4.3. Segmentação da face

A segmentação de face é uma tarefa importante e essencial em diversas aplicações, tais como reconhecimento de faces, rastreamento de pessoas e segurança. Nesse trabalho a segmentação da face é um pré-processamento necessário para eliminação dos outliers provenientes da fase de reconhecimento dos keypoints obtidos durante a fase de treinamento das características faciais.

Diversos métodos de detecção de face já foram propostos na literatura e muitos dos algoritmos desenvolvidos são baseados na cor da pele e exploram essa informação de forma a localizar e extrair a região da face.

Nesse trabalho a localização e segmentação da face ao longo do vídeo é feita com base nas amostras de cor da pele extraídas da imagem de treinamento pelo usuário. Após a delimitação de algumas regiões da pele, calcula-se, para cada um dos canais R, G e B, a média e o desvio padrão do total de pixels obtidos pelo usuário. Considere mR, mG, mB as médias e σR, σG, σB os desvios padrões da amostra de pixels de cada componente R, G e B, respectivamente. Um pixel i da imagem com componentes RGB iguais a Ri, Gi e Bi é classificado como pertencente à face se as condições (1), (2) e (3) forem satisfeitas. Caso contrário, o pixel é classificado como não pertencente à face.

$$Ri \in [mR - \sigma R, mR + \sigma R] \quad (1)$$

$$Gi \in [mG - \sigma G, mG + \sigma G] \quad (2)$$

$$Bi \in [mB - \sigma B, mB + \sigma B] \quad (3)$$

Essa abordagem fornece uma primeira segmentação da região da face. No entanto, como qualquer algoritmo que leva em consideração a cor da pele, outras regiões da imagem que possuem a mesma cor também serão classificadas como pele, gerando detecções erradas. Considerando que a face a ser segmentada

representa a parte mais significativa da imagem, encontra-se os contornos de cada região detectada como face e atribui a face correta à maior área interna dentre os contornos obtidos. Um exemplo dessa abordagem pode ser vista na figura 7.

Figura 3: (a) Imagem de treinamento. (b) imagem segmentada pela cor da pele. (c) Contornos detectados. (d) Extração do maior contorno e detecção da face.

Um dos problemas da abordagem por cor pode ser vista na figura 8. A sobrancelha se funde com a região do olho e ambos se unem ao fundo. Dessa forma, quando o interior do maior contorno é preenchido, a região do olho é cortada e com isso a detecção do olho não será possível ou será erroneamente detectada.

Figura 4: Região do olho se une ao fundo e não pode ser detectada corretamente.

A solução utilizada nesta dissertação para este problema foi encontrar o fecho convexo da região da face (em branco). No entanto, mais outliers são provenientes da fase de reconhecimento, uma vez que parte do fundo é levada em consideração. Um exemplo da segmentação da face e do fecho convexo correspondente pode ser visto na figura 9.

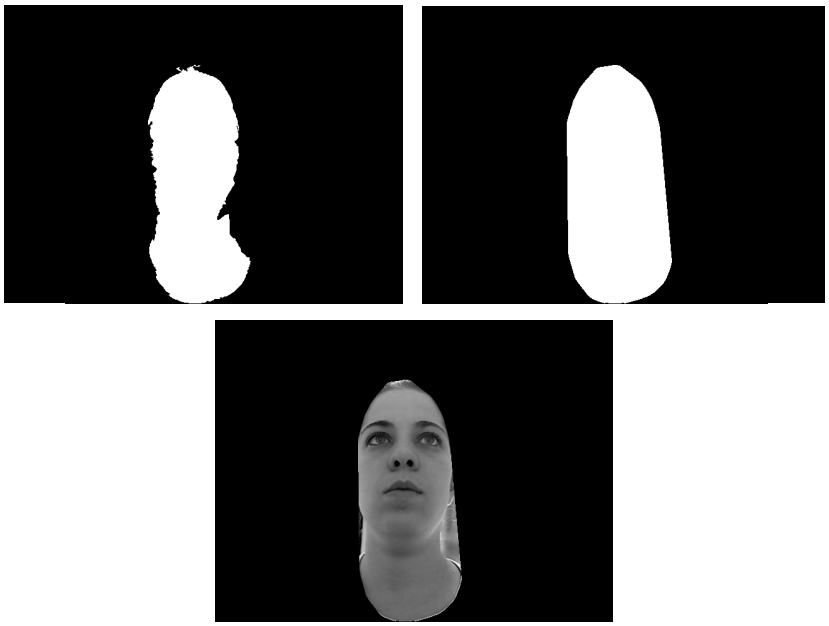

Figura 5: (a) Face segmentada pela cor de pele. (b) Fecho convexo da face segmentada. (c) Face segmentada usando o fecho convexo.

4.4. Reconhecimento de keypoints

Uma vez que o classificador tenha sido previamente treinado, é possível utilizá-lo para o reconhecimento de patches em imagens obtidas em tempo real por uma webcam. Para cada quadro do vídeo é feita a segmentação da face como descrito na seção 4.2. Após isso, pontos característicos são extraídos da imagem da face segmentada (em tons de cinza). Para cada keypoint extraído é feito o seguinte procedimento: pega-se o patch associado ao keypoint e passa por cada um dos FERNS do classificador. Cada FERN retorna a probabilidade desse patch pertencer a cada uma das classes aprendidas durante a fase de treinamento, como explicado nas seções 3.2 e 3.3. As respostas de cada FERN são combinadas e retornam o keypoint correspondente na imagem de treinamento. Esse processamento pode ser visto na figura 10.

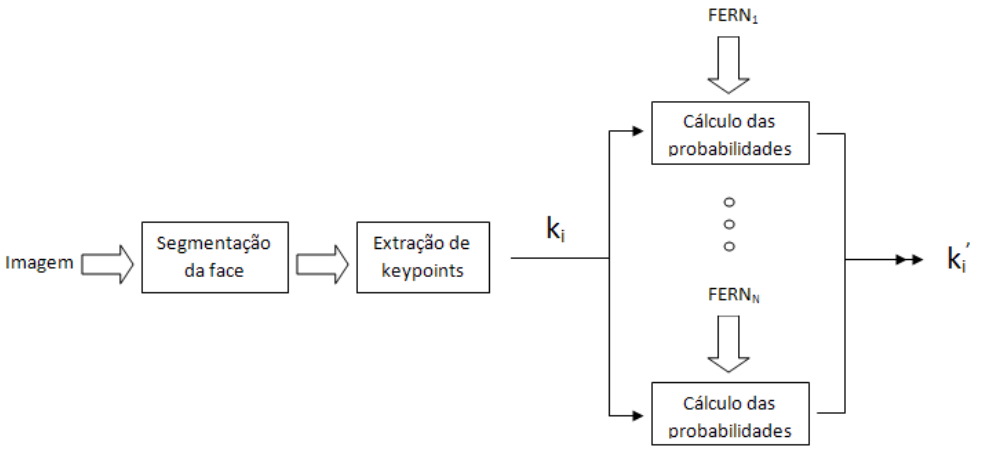

Figura 6: Processo de classificação de um keypoint.

4.5. Clusterização e detecção das regiões das características faciais

A simetria da face e a homogeneidade que apresenta em determinadas regiões pode acabar gerando outliers no processo de reconhecimento dos keypoints. Um keypoint pertencente à íris, por exemplo, pode ser reconhecido erradamente como sendo outro ponto na mesma íris. Além disso, um outro erro bastante comum é um keypoint do olho esquerdo ser reconhecido como estando

no olho direito. Se considerarmos que a maior parte dos keypoints reconhecidos se encontra mais concentrada nas áreas de interesse em questão (olho direito, olho esquerdo ou narinas), esses problemas podem ser contornados e não afetam significativamente o sistema. Dessa maneira, basta agruparmos keypoints próximos e decidirmos qual dos grupos melhor representa a característica em questão.

Os keypoints de cada característica facial são agrupados de forma independente. Isso é possível, pois durante a fase de treinamento atribui-se a cada keypoint detectado um marcador dizendo se aquela característica pertence ao olho esquerdo, ao direito ou às narinas. Dessa forma, quando é feito o reconhecimento, o keypoint carrega esse marcador e é possível saber à qual característica esse keypoint pertence. Duas abordagens diferentes foram utilizadas nesta dissertação para detectar as regiões onde estão localizadas as características faciais.

O primeiro método utilizado na detecção das regiões das características da face baseia-se no fato de que a maior parte dos keypoints reconhecidos pelo classificador deve estar localizada na região da característica treinada. Dessa forma, basta agrupar os keypoints próximos e pegar o maior cluster como sendo a região desejada, ou seja, a região que corresponde à característica em questão. Isso é feito, de maneira independente, para cada uma das características tratadas, isto é, olho esquerdo, olho direito e narinas.

A segunda abordagem, além de considerar a organização no espaço, leva em conta a probabilidade que os FERNS retornam quando um keypoint é reconhecido. Se a probabilidade retornada pelos FERNS for muito pequena, provavelmente aquele keypoint foi reconhecido de forma errada. Após realizar o agrupamento dos keypoints pela sua proximidade, soma-se as probabilidades de todos os keypoints pertencentes à um cluster (agrupamento). Isso é feito para cada um dos clusters obtidos e aquele que obtiver a maior soma é considerado o candidato mais forte, pois possui a maior probabilidade. Isso implica que existe uma probabilidade maior dos keypoints daquele cluster terem sido reconhecidos de forma correta. Esse procedimento também é realizado de forma independente para cada uma das características treinadas.

4.6. Rastreamento baseado nos clusters e reinicialização do sistema

Os dez primeiros quadros de um vídeo são utilizados para inicializar o processo de detecção das características faciais. Os FERNS são bastante confiáveis na detecção das características em imagens frontais, pois essas são muito semelhantes à imagem utilizada para o treinamento do classificador. Por isso, considerando que nos primeiros quadros do vídeo a face se encontra na posição frontal, o sistema pode ser inicializado automaticamente. Após o processo de reconhecimento de keypoints em um quadro, para cada uma das características faciais em questão faz-se: aplica a clusterização e utiliza um dos métodos descritos na seção anterior para determinar a região onde se encontra uma determinada característica. Após isso, calcula-se o centro da região detectada e atribui-se essa posição à característica em questão. Esse procedimento é realizado nos dez primeiros quadros. Com isso é possível calcular a posição média onde a característica foi detectada. Esse procedimento é realizado para cada uma das características e os valores obtidos são utilizados para inicialização do sistema.

Para melhorar a eficiência do sistema, ao se processar um quadro pode-se considerar as detecções das regiões das características faciais encontradas no quadro anterior para determinação se a detecção foi ou não correta. Assume-se que não existem mudanças muito grandes entre dois quadros consecutivos. Com isso, se a distância entre o ponto encontrado em dois quadros consecutivos for muito grande, então um erro aconteceu e a posição encontrada no quadro anterior é mantida. Caso contrário, considera-se que a detecção foi correta. No entanto, se a detecção se mantiver errada por muitos quadros consecutivos o sistema se perde e precisa ser reinicializado. Isso é possível garantindo-se que a inicialização do sistema foi correta, caso contrário poderia gerar problemas na detecção das regiões das características faciais.

A reinicialização do sistema é feita da mesma maneira que a inicialização, isto é, consideram-se os dez quadros seguintes e calcula a posição média para cada uma das características. Para garantir a corretude na reinicialização, o ideal é que o usuário retorne à posição de frente para a câmera, onde a detecção utilizando os FERNS é mais confiável. Caso contrário, a reinicialização pode apresentar resultados errados.

4.7.
Detecção das características faciais

4.7.1.
Detecção das narinas

A região de busca pelas narinas é realizada no interior da região detectada anteriormente, como explicado nas seções 4.5 e 4.6. Uma maneira simples de encontrar as narinas é buscar pelas duas regiões mais escuras dentro da região onde a busca esta sendo realizada. Para isso é feito um threshold que separa as partes mais escuras do histograma. Como as narinas se encontram mais ou menos no centro do rosto, a região ao redor dessa característica é bem homogênea. Apesar disso, não pode-se considerar que apenas as duas regiões da narina serão encontradas, pois dependendo do valor do threshold utilizado essas regiões podem ficar divididas. Além disso, a região de busca pode ter outros resíduos, como por exemplo, um pedaço do olho, o que pode acontecer dependendo da movimentação feita pela pessoa. Para solucionar o problema da região da narina estar dividida, é utilizado um operador de dilatação com o objetivo de unir regiões disjuntas que se encontram próximas. No entanto, resíduos que estão longe dessa região e realmente não fazem parte das narinas não desaparecem com esse processo, muito pelo contrário, eles ficam mais destacados ainda. Se o número de regiões detectadas após a dilatação for igual a dois, então encontra-se os contornos dessas regiões e utiliza a posição central de cada um para representar o centro das narinas. No entanto, se o número de regiões detectadas após a operação de dilatação for maior do que dois, significa que ruídos não pertencentes às narinas foram detectados. Nesse caso, para localizar as narinas considera-se que essas representam as duas maiores regiões detectadas. Logo, basta aplicar um operador de erosão até que restem apenas duas regiões escuras na imagem. No entanto, o elemento estuturante utilizado para a operação morfológica de erosão não pode ser o mesmo utilizado para a dilatação, caso contrário as regiões poderiam acabar desaparecendo e a detecção não seria possível. Após o processo de erosão, quando existirem apenas duas regiões, o mesmo procedimento explicado anteriormente é utilizado para detectar o centro das narinas.

4.7.2.
Detecção das pupilas

A detecção das pupilas é um pouco mais complicada do que a detecção das narinas, pois quando é realizado o threshold grande quantidade de ruído ainda é encontrada. Além disso, a região dos cílios e sobrancelhas atrapalham a detecção exata das pupilas. Duas abordagens foram utilizadas com o propósito de detectar as pupilas da maneira mais correta possível.

A primeira, e mais simples das abordagens, consiste em detectar a região mais escura dentro da região de busca. Mais uma vez, o procedimento utilizado é a aplicação de um threshold. Após isso calcula-se a elipse que se ajusta melhor à região mais escura detectada. Considera-se então, o centro da elipse como sendo o centro da pupila. Isso é feito para os dois olhos independentemente.

A segunda abordagem, visa melhorar ainda mais a detecção das pupilas. A primeira parte desse método consiste em aplicar o threshold e encontrar a região mais escura. Após isso, da mesma maneira que na primeira abordagem, é realizado o ajuste da elipse à região mais escura detectada. No entanto, a diferença se deve ao processamento realizado no interior da elipse encontrada. Como pode ser visto na figura 11, a região da pupila tem uma concentração maior de pixels pretos na vertical depois de realizado o threshold.

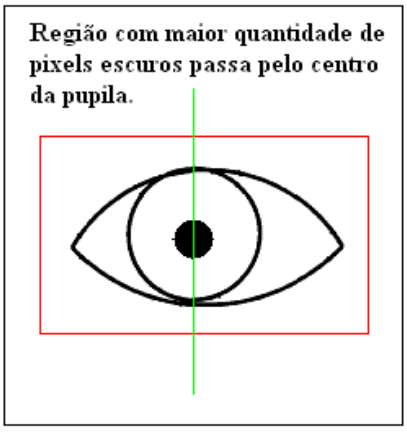

Figura 7: Representação do olho mostrando que a maior quantidade de pixels por coluna passa pelo centro da pupila.

Dessa maneira, percorre-se a região de busca contando quantos pixels pretos existem em cada uma das colunas. Dessa forma, a posição x é dada pela coluna que contém a maior concentração de pixels pretos, enquanto a posição y é dada pela posição y do centro da elipse.