5
Metodologia

O objetivo principal da dissertação é criar um método capaz de detectar a emoção de um usuário através de imagens frontais de sua face, usando as técnicas de ASM e SVM, ambos treinados com algoritmo supervisionado e não incremental.

Mais detalhadamente, essa detecção é composta de várias sub-tarefas, sendo elas:

1. O treinamento de um modelo ASM;
2. O treinamento de um modelo SVM;
3. Análise de uma imagem frontal da face de um indivíduo:
   (a) Encontrar a forma, ou seja, os pontos de referência que definem o rosto na imagem com o ASM;
   (b) Extrair dessa forma os atributos de entrada para o SVM;
   (c) Predizer a emoção com SVM treinado.

O diagrama na Figura 5.1 ilustra o terceiro item descrito acima de maneira gráfica.

Nos capítulos anteriores, uma descrição é feita para os algoritmos tradicionais de ASM e SVM, que são usados no trabalho. Portanto, esse capítulo se limita a detalhar como é feito o treinamento dos modelos e os caminhos tomados para tal e também apresentar o programa criado que utiliza o método proposto para reconhecimento de emoções. Além disto, esse capítulo apresenta vários ajustes práticos feitos no sistema para melhorar o reconhecimento.

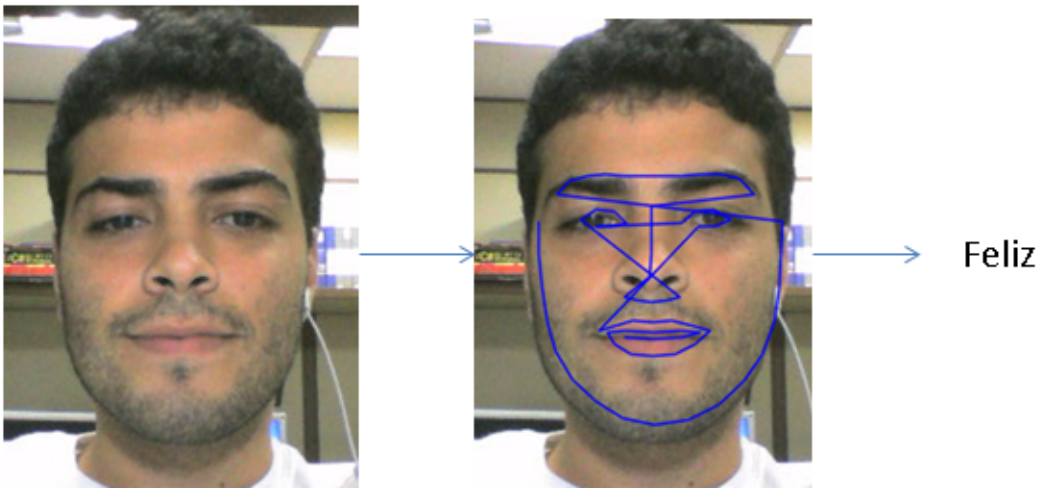


Figura 5.1: À esquerda a foto de um rosto, no centro, os pontos de referência encontrados para a foto e, à direita a classificação de emoção dessa imagem

5.1 O Treinamento do Modelo ASM

A criação do modelo ASM é feita utilizando a implementação disponibilizada por Milborrow(2008), que desenvolveu um *framework* para a criação e uso desse tipo de modelo. Seu *framework* dispõe dos seguintes programas principais:

- STASM é o localizador de formas em imagem, isto é, dos pontos de referência que representam um objeto. Ele necessita de arquivos de configuração previamente criados, como o arquivo do modelo ASM;
- TASM é o criador de ASM. Com ele é possível criar ASM a partir de qualquer banco de imagens, desde que exista a anotação dos pontos de referência para cada imagem do banco.
- ASM padrão disponibilizado por Milborrow treinado com o banco de imagens MUCT (29);

O ASM padrão disponibilizado por Milborrow (*op. cit.*) havia sido treinado com imagens inadequadas para o escopo da dissertação, pois o banco de imagens MUCT consiste em imagens de pessoas posando para uma foto no estilo 3x4, algumas felizes, outras neutras, mas com pouca variação. Entretanto, como descrito na seção de Modelos Flexíveis, o SSM consegue se deformar dentro das variações existentes no banco de dados analisado. Para a tarefa proposta é necessário que o ASM tenha a capacidade de identificar uma maior variedade de expressões faciais para que também seja capaz de capturar formas do rosto de pessoas tristes, nervosas, surpresas, entre outras.

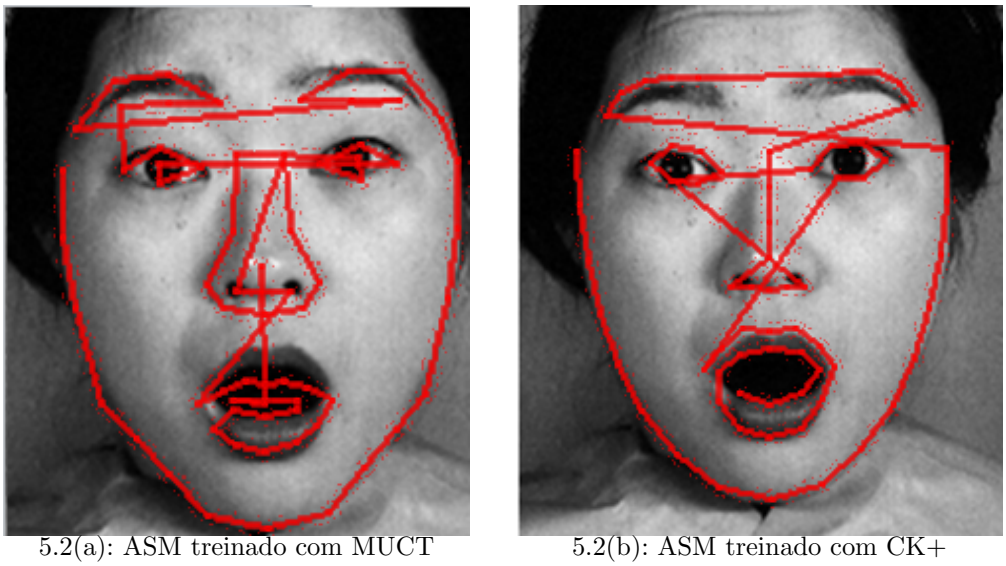
Em decorrência disso, o ASM foi treinado com o banco de imagens Cohn-Kahnade+ (CK+)(22) (27), um banco próprio para o estudo de emoções, que fornece mais de 10 mil imagens frontais de rostos de 123 indivíduos distintos expressando alguma emoção (neutro, felicidade, tristeza, raiva, surpresa, medo e desgosto). Tais fotos possuem 68 pontos de referência anotados automaticamente com um *Active Appearance Model* (AAM).

Para aumentar o banco de imagens, foi utilizada a técnica de espelhamento de imagens. Apesar de parecer algo redundante, o espelhamento de imagens adiciona novas imagens que não estavam presentes anteriormente, pois a assimetria dos rostos e diferenças de iluminação e de pose fazem com que a imagem espelhada carregue informações diferentes da original. Além disso, sabemos que uma face qualquer, quando invertida (espelhada com eixo vertical), continuará sendo uma forma plausível de uma face, devido ao eixo de simetria da face. Com o uso dessa técnica, dobra-se o número de imagens do banco, fazendo com que o ASM seja treinado com mais informações e tenha uma maior capacidade de generalização e adaptação (30).

A partir do banco de imagens CK+, é criado, então, um ASM compatível com o STASM e o resultado é um modelo que consegue se deformar para encontrar pontos de referência em fotos de pessoas expressando várias emoções fortes. O resultado, em comparação com o ASM provido por Milborrow (*op. cit.*), é superior para fotos onde a pessoa possui fortes expressões no rosto. Como pode ser visto na Figura 5.2, da mesma imagem foram extraídos pontos faciais com dois modelos: o padrão do STASM e o treinado com o CK+. Podemos ver que os contornos se tornaram bem mais justos em regiões importantes, com enfoque na boca e nariz. É importante ressaltar que o padrão de marcação de pontos faciais é diferente para os dois modelos. Então, o ASM utilizado daqui em diante é o que foi treinado com o CK+.

5.2 O Treinamento do Modelo SVM

Uma vez que o modelo ASM foi criado, tornou-se possível montar um banco de dados para treinamento do modelo responsável pela classificação. Apesar de possuir o banco de dados CK+ anotado, optamos por não utilizá-lo no momento da criação do modelo SVM. Essa opção foi feita pois os pontos de referência presentes no banco de imagens CK+ foram obtidos de maneira diferente do modo como seriam obtidos os novos pontos na hora da execução. As anotações do CK+ foram feitas utilizando um AAM; e, se o ASM criado pelo TASM fosse aplicado no CK+, as formas encontradas seriam diferentes



5.2(a): ASM treinado com MUCT 5.2(b): ASM treinado com CK+

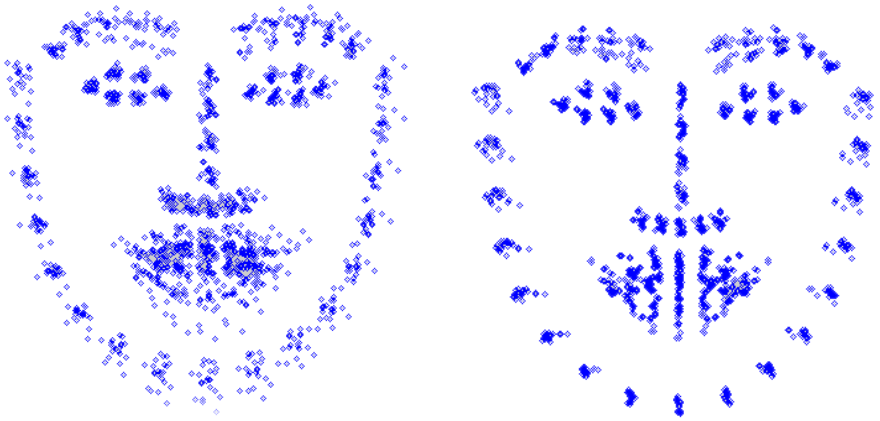
Figura 5.2: Resultado da execução de dois ASM diferentes, o padrão do STASM treinado com MUCT 5.2(a) e o treinado com CK+ 5.2(b)

das oficiais do banco de dados. Desta maneira, o resultado seria um modelo em que os dados de treinamento e os dados reais são extraídos de maneira diferente, o que adicionaria ruído ao processo, diminuindo sua capacidade de predição.

Por esse fato, abandonamos a anotação do CK+ para questões futuras e aplicamos o ASM em outro banco de imagens faciais, o RaFD. Esse banco de imagens possui aproximadamente 1500 imagens frontais de 70 indivíduos diferentes com oito emoções diferentes. Os resultados de todas as formas encontradas para cada uma das imagens foram guardados em um banco de dados de onde atributos poderiam ser modelados. O mesmo foi feito em um subconjunto de imagens do CK+.

Para reduzir ruídos, as formas do banco de dados foram alinhadas com o uso do GPA, pois pequenas variações independentes do estado emocional de um indivíduo poderiam atrapalhar a classificação. Um exemplo da melhoria encontrada após o alinhamento das formas pode ser visto na Figura 5.3. Essa etapa trouxe mais uniformidade às formas exaltando o que realmente era importante: o deslocamento de locais expressivos, como a boca, sobrancelhas, etc. Isso ocorreu porque, com o alinhamento, certos aspectos ruidosos, como a distância do indivíduo até a câmera e possíveis translações e rotações da cabeça, foram ignorados.

Uma vez que o banco de dados havia sido criado e alinhado, tornou-se necessário definir mais precisamente conceitos importantes para a tarefa de



5.3(a): Formas não alinhadas

5.3(b): Formas alinhadas

Figura 5.3: Diferença entre não alinhar as formas 5.3(a) e com o alinhamento feito 5.3(b)

aprendizado de máquina.

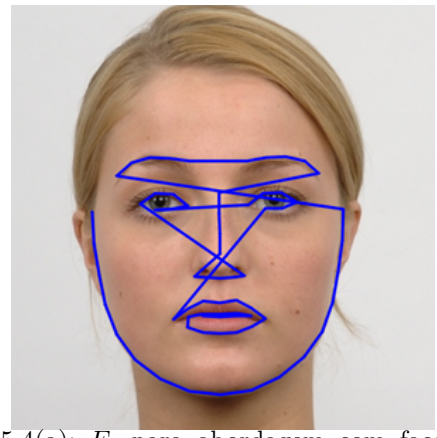
Para o treinamento do SVM, optamos por utilizar atributos de deslocamento. Esses atributos estão relacionados à alteração e por isso necessitam de dois momentos, um estado inicial E_i e um final E_f. Esse tipo de atributo, ao invés dos instantâneos, está relacionado ao quanto um atributou mudou, como, por exemplo, o quanto o canto da boca se deslocou do estado inicial até o final ou o quanto a sobrancelha se deslocou entre os dois momentos. Os atributos de alteração foram utilizados pelo fato de que apresentaram melhores resultados desde os protótipos mais primitivos e também porque são comumente usados para a tarefa de reconhecimento de emoções.

Duas abordagens foram tomadas para E_i:

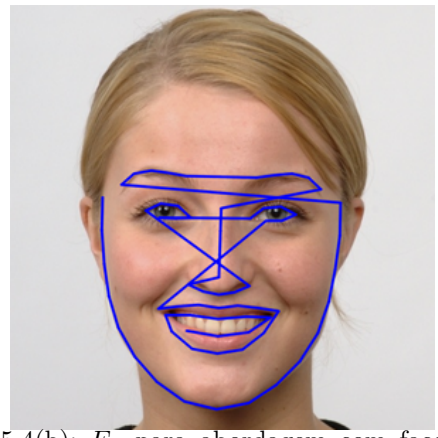
- A face neutra do mesmo indivíduo;
- A forma média do ASM.

As Figuras 5.4 e 5.5 ilustram as duas formas de comparação para um mesmo indivíduo. Enquanto na abordagem de face média, Figura 5.5, utiliza-se uma face média comum para todas as comparações de E_f, independente do indivíduo, na abordagem de face neutra, Figura 5.4, o E_i sempre conterá o mesmo indivíduo do E_f com expressão neutra. As respectivas Figuras também exibem as formas extraídas dos estados, delineadas com cor azul.

É interessante ressaltar que tais abordagens, da face neutra e forma média, em linhas gerais, já foram pesquisadas em outros estudos, *e.g.*(15) e (41). Mas até o momento, estas duas faces nunca foram testadas utilizando um mesmo banco de dados para treinamento, algoritmo e base de testes.



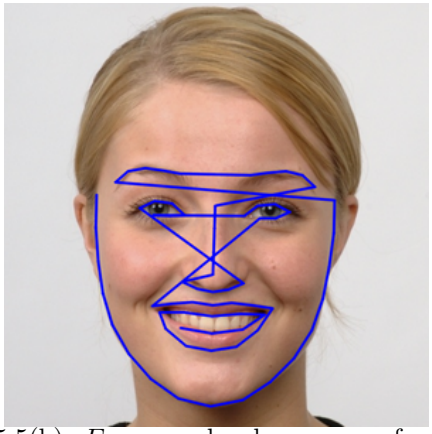
5.4(a): E_i para abordagem com face neutra



5.4(b): E_f para abordagem com face neutra

Figura 5.4: E_i e E_f para a abordagem de face neutra como base de comparação. Imagens do banco de imagens RaFD (24).

5.5(a): E_i para abordagem com face média



5.5(b): E_f para abordagem com face média

Figura 5.5: E_i e E_f para a abordagem de face média como base de comparação. Imagem à direita vinda do banco de imagens RaFD (24).

Ignoramos, nesse momento, que quando a forma média é usada como estado inicial, o modelo adquire a capacidade de predizer a expressão neutra, pois uma das principais premissas da dissertação é fazer uma comparação entre as duas opções para estado inicial. Então, a tarefa é formalizada da seguinte forma:

Sendo f_i a forma referente ao E_i e f_f a referente ao E_f superimpostas, temos a definição do vetor de atributos, V, como:

$$V = f_f - f_i, \text{ ou seja,}$$
$$V = (x_{1_{f_f}} - x_{1_{f_i}}, y_{1_{f_f}} - y_{1_{f_i}}, x_{2_{f_f}} - x_{2_{f_i}}, y_{2_{f_f}} - y_{2_{f_i}}, ..., x_{n_{f_f}} - x_{n_{f_i}}, y_{n_{f_f}} - y_{n_{f_i}})$$

- Atributos de entrada: Vetor V, composto de valores numéricos reais iguais ao deslocamento de todos os pontos faciais entre o f_f e f_i. No caso, como são 68 pontos de referência, o vetor de atributos contém 136 posições (x e y para cada ponto),
- Saída: O rótulo de uma das seguintes emoções, felicidade, tristeza, raiva, desgosto, surpresa e medo.

A Figura 5.6 ilustra o deslocamento dos pontos de referência entre duas formas, onde os círculos vermelhos representam f_f, os losangos azuis f_i e o segmento negro entre eles, o deslocamento entre cada ponto correspondente das duas formas, ou seja, $V = f_f - f_i$.

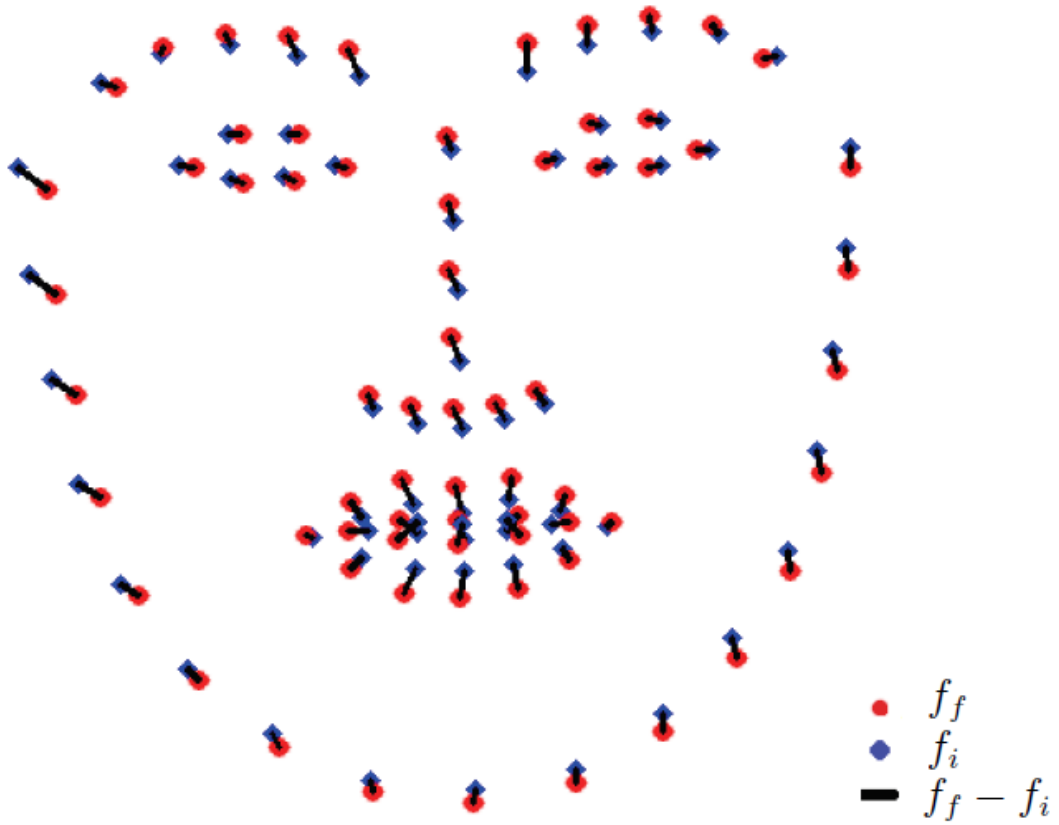


Figura 5.6: Um exemplo de deslocamento dos pontos faciais entre dois estados

Dois modelos diferentes foram treinados, um para cada abordagem de estado inicial. A implementação utilizada foi a LIBSVM (5) que suporta SVM multiclasse (*one-versus-one*), essencial para a predição de várias emoções. Os atributos foram normalizados para a escala entre -1 e 1 e os parâmetros de treinamento foram custo = 32 e *gamma* = 0.5. Os resultados específicos dos SVMs treinados podem ser vistos nas seções 6.3.1 e 6.3.2.

5.3 Aplicação de Reconhecimento de Emoções

O objetivo secundário do presente trabalho é desenvolver uma aplicação utilizando o método de reconhecimento de emoções desenvolvido para validar o seu uso em tempo real.

Tal aplicação foi desenvolvida em C++ utilizando como suporte a OpenCV (4), uma biblioteca com funções próprias para visão computacional que também fornece boa interface para uso de *webcam*. Como descrito na seção anterior, existem duas abordagens possíveis para o estado inicial, a face neutra do indivíduo e a forma média do ASM.

Como a face neutra do indivíduo alcançou resultados mais promissores desde o início do estudo, optamos por desenvolver a aplicação com a face neutra como o estado inicial.

Existe uma significativa diferença entre o modelo capaz de reconhecer a emoção e uma aplicação que o utiliza. No primeiro, estamos trabalhando com um modelo que foi treinado com dados vindos de um ambiente controlado. As fotos dos bancos de imagens foram cuidadosamente preparadas e avaliadas para que se encontrassem em um nível de qualidade específico. Já no segundo caso, não existe tal controle de qualidade da imagem, pois o usuário está simplesmente sendo capturado pela *webcam*, o que possibilita vários cenários desfavoráveis à predição, tais como o mal posicionamento da câmera e a fraca luminosidade no ambiente.

Por isso, durante o desenvolvimento dos protótipos da aplicação, percebemos que certos ajustes precisariam ser feitos para alcançar um bom resultado para a predição de emoção. O algoritmo ASM se mostrou pouco estável para a aquisição de pontos faciais adquiridos de um *frame* para outro e isso causava muitos erros de classificação; pois, devido a pequenas alterações de luminosidade no ambiente ou movimentos do rosto do usuário, as formas encontradas pelo ASM variavam muito. Para ter mudanças mais suaves nas formas encontradas pelo ASM e obter predições mais corretas, as seguintes técnicas foram implementadas:

- Uma janela deslizante das últimas formas encontradas pelo algoritmo ASM;
- Uma janela deslizante das últimas predições realizadas pelo algoritmo SVM.

Com o primeiro ajuste, a média das últimas n formas encontradas é usada para a predição. Com isso, os erros discrepantes devido às interferências citadas

anteriormente são suavizados. Já com o segundo, a contagem das últimas predições suaviza erros pontuais de classificação, fazendo com que ela possua mais segurança. A desvantagem de tais adaptações é que as respostas demoram mais para se concretizarem, pois, em trechos de transição, informações de momentos anteriores estão presentes nas janelas deslizantes. No entanto, ao utilizar janelas de tamanho pequeno, $n = 5$, por exemplo, encontramos um bom *tradeoff* entre a latência para a predição de emoção e a redução de erros.

Dessa forma, o *loop* principal da aplicação tem o seguinte funcionamento:

1. Inicialização dos recursos necessários (*Webcam*, arquivos de configuração do ASM, SVM);
2. Aquisição da imagem da *webcam*;
3. Execução do ASM sobre imagem capturada;
4. Com o resultado dos pontos faciais adquiridos em 3, calculamos a forma média das últimas cinco formas encontradas;
5. Extraímos os atributos de deslocamento entre as formas e os classificamos com o SVM;
6. Incluindo a classe encontrada em 5, calculamos qual emoção aconteceu mais vezes nas últimas cinco predições;
7. Exibimos na tela a classe que mais ocorreu ultimamente, resultante do passo 6;
8. Repetimos a partir de 2.

Para a aplicação pretendida nesta dissertação tal abordagem é satisfatória. No entanto, para aplicações que têm o reconhecimento de emoções apenas como parcela do funcionamento geral, é importante que um processo mais elaborado seja criado. Por esse motivo, recomendamos o uso de *threads* e paralelismo. Isso é importante porque a tarefa envolvida é computacionalmente cara e pode se tornar um gargalo.

Além das adaptações descritas acima, adotamos a amplicação de emoções para que a aplicação tivesse capacidade de reconhecer emoções mais sutis do que as vistas nos bancos de imagens com os quais os modelos foram treinados. A amplicação de emoções (34) funciona através da alteração dos vetores de deslocamento de pontos faciais por um determinado fator. Essa alteração pode ser feita em regiões específicas, como olhos e bocas, através de um vetor de

pesos. Para a aplicação, fizemos um vetor de pesos iguais, que aumenta em 20% todos os deslocamentos encontrados. Tal valor foi definido empiricamente através de tentativas e erros e após perceber que um fator muito alto de amplificação acarretava em distorções nas formas, o que diminuía a acurácia de predição do SVM. A Figura 5.7 mostra a diferença entre uma forma normal, com cor verde e a ampliada, vermelha.

Figura 5.7: Um exemplo de amplicação de emoção, na qual a forma verde representa o normal e a vermelha, a ampliada

Como a classificação é feita em seis emoções básicas que não incluem a expressão neutra, tornou-se necessário uma maneira de identificá-la, que é detalhada a seguir. Uma vez que o usuário registra como estado inicial sua face neutra, ela é armazenada e, antes das predições ocorrerem, é feito um teste que verifica se o erro quadrático médio entre a face neutra e a atual ultrapassa um limite pré-definido. Se não ultrapassar, as formas se modificaram pouco e, portanto, o usuário está neutro. Caso contrário, o modelo SVM é utilizado para a predição.