DIEGO RUIVO CUEVA

FUSÃO COMPUTACIONAL DE OBSERVAÇÕES AFETIVAS

Dissertação apresentada à Escola
Politécnica da Universidade de São
Paulo para obtenção do título de
Mestre em Ciências.

São Paulo

2013

DIEGO RUIVO CUEVA

FUSÃO COMPUTACIONAL DE OBSERVAÇÕES AFETIVAS

Dissertação apresentada à Escola
Politécnica da Universidade de São
Paulo para obtenção do título de
Mestre em Ciências.

Área de Concentração:
Engenharia Mecânica

Orientador:
Prof. Dr. Fábio G. Cozman

São Paulo
2013

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, de julho de 2013.

Assinatura do autor _____

Assinatura do orientador _____

FICHA CATALOGRÁFICA

Cueva, Diego Ruivo

Fusão computacional de observações afetivas / D.R. Cueva.

-- versão corr. -- São Paulo, 2013.

93 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia Mecânica.

1. Interação homem-máquina (Aspectos cognitivos) I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia Mecânica II. t.

Aos meus pais, Heloise e Nelson.

AGRADECIMENTOS

Ao professor Dr. Fábio Cozman, que não apenas recebeu com entusiasmo e de braços abertos o projeto em seu laboratório, mas compartilhou a experiência necessária para transformar uma área de pesquisa tão aberta em uma contribuição acadêmica.

Ao professor Dr. Marcos Barretto que, desde a graduação, tem como cruzada me apresentar uma visão diferenciada da engenharia, criativa e sem restrições.

Ao grande amigo Rafael Gonçalves, parceiro constante nessa área de pesquisa tão jovem, pelas inúmeras e riquíssimas discussões colaborativas sobre referências, algoritmos, o universo e todas as coisas.

Aos meus pais e meu irmão Daniel, que desde infância mostraram quão divertida, desafiadora e ilimitada a engenharia poderia ser.

À Juliana e Arthur pela colaboração na revisão de trechos em inglês.

À Marília, pela paciência, ajuda com revisões, discussões e apoio incondicional nos finais de semana e feriados ocupados.

Aos amigos e parentes que de uma ou outra forma estiveram envolvidos na produção deste trabalho.

À FAPESP e CNPq, pelo auxílio em publicações e bolsa de estudos.

Resumo

Este trabalho tem como objetivo discutir o estado da arte e propor melhorias em uma área de grande potencial e de crescimento recente na computação: a análise do estado emocional do usuário de um sistema de informação e a aplicação dessa observação em cenários ricos de interação homem-máquina.

Para foco em um escopo tratável, escolheu-se o estudo das técnicas atuais de captação afetiva através de três sensores diferentes (expressões faciais, expressões vocais e contexto semântico) e de como sua combinação poderia resultar em dados mais robustos sobre o estado emocional do usuário, em uma aplicação conhecida como “Fusão Multimodal”.

Nesse sentido, o trabalho levanta não apenas a bibliografia recente sobre o estado da arte da aquisição de dados nas três áreas e em fusão, mas também aplica algumas das ferramentas existentes (tanto comerciais como abertas) para melhor compreensão do nível atual da tecnologia acessível ao público acadêmico.

Dada a inexistência de uma aplicação facilmente disponível para análise de emoções no contexto semântico, desenvolveu-se uma ferramenta dedicada à pesquisa, nomeada emoCrawler, a qual utiliza redes sociais para avaliação do caráter emocional que palavras indiretamente podem acarretar em uma discussão. O emoCrawler busca avaliar as emoções evocadas em usuários de Twitter e as compara com um dicionário emocional conhecido, baseado em contribuições acadêmicas diversas, gerando um resultado quantitativo útil ao escopo do trabalho.

Por fim, o trabalho lançou mão da ferramenta emoCrawler e das outras ferramentas selecionadas para análise de expressões faciais e vocais na criação de um sistema de fusão sensorial que aumente a confiabilidade global da emoção detectada. Para isso, inicialmente um corpus de áudio e vídeo disponível foi tratado e reduzido para o escopo da pesquisa. Em seguida, diferentes técnicas de fusão baseadas em algoritmos de classificação foram aplicadas (redes neurais, *Naive Bayes* e *Support Vector Machine*), com abordagens distintas de treinamento para busca de resultados melhores dos que os encontrados nas aplicações unimodais.

Os resultados verificados de fato indicam para uma melhor taxa de detecção quando é aplicado o algoritmo de fusão. Nesse processo, a presença de dados fornecidos pelo emoCrawler parece ter contribuído positivamente.

Abstract

This work aims at discussing the state of the art and proposing improvements in a field of great potential and recent growth in computing: the analysis of the emotional state of the user of an information system and the application of such observations in rich scenarios of human-machine interaction.

In a tractable scope of research, this work evaluates the current techniques of affective observation applied on three different inputs (facial expressions, speech features and semantic context) and comprehends how their combination can result in more robust affirmations on the actual emotional status of the user – an application called “Multimodal Fusion”.

In this sense, the work presents not only the recent literature on the state of the art in data acquisition in the three aforementioned areas and in fusion, but also applies some of the existing tools (both commercial and open) for a better understanding of the current level of technology available to the academic audience.

Given the lack of an application readily available for the analysis of semantic context in emotions, a dedicated piece of software to the research was developed, entitled emoCrawler, which uses social networks as a means to evaluate the emotional signature that words may leave in a discussion. emoCrawler aims at evaluating emotions evoked in Twitter users and comparing them with an emotion dictionary built around academic contributions, generating a quantitative result useful to this research.

Finally, this work made use of emoCrawler and the other selected tools for both facial and speech expressions in the creation of a sensor fusion system to enhance the global reliability of the affective detection. To achieve this, an audiovisual database was initially processed and condensed. Then, different fusion techniques based on classification algorithms with distinct learning approaches (neural networks, naïve Bayes and *Support Vector Machine*) were applied in the pursuit of better outcomes than the unimodal applications.

The final results indeed point to a better rate of detection when fusion algorithms are applied. In this process, the data provided by emoCrawler seems to have contributed positively.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

Lista de Símbolos

Introdução.....	1
1.1 Objetivos	4
1.2 Contribuição	5
1.3 Aplicações	5
1.4 Estrutura do Texto	6
2 Contextualização e Revisão Bibliográfica.....	8
2.1 Histórico do Estudo de Emoções	8
2.2 Conceitos em Computação Afetiva.....	10
2.2.1 O Processo Emocional	13
2.2.2 A Definição da Emoção	15
2.3 As Emoções na Face	17
Técnicas Computacionais	19
2.4 As Emoções na Voz	20
Técnicas Computacionais	20
2.5 As Emoções na Semântica	21
Técnicas Computacionais	22
2.6 Análise Multimodal de Emoções	24
Técnicas Computacionais	26
2.7 Classificadores	30
2.7.1 Redes Neurais	30
2.7.2 <i>Naive Bayes</i>	33
2.7.3 <i>Support Vector Machine</i>	34
3 Abordagem Proposta	38
3.1 Visão Geral.....	38
3.2 Fusão por Classificação.....	39
Técnicas de Fusão Aplicadas	41
3.3 emoCrawler: Detecção Semântica de Emoções	43
3.3.1 Motivação.....	44
3.3.2 A Problemática da Semântica e o Uso do Twitter	44
3.3.3 Arquitetura	45
3.3.4 O Dicionário Emocional	48
3.3.5 O Contexto Temporal.....	49
3.3.6 A Questão do Janelamento.....	51
4 Experimentos e Resultados	53
4.1 Ferramental	53

4.1.1 Corpus	53
4.1.2 Detecção de Emoções na Face: eMotion	54
4.1.3 Detecção de Emoções na Voz: EmoVoice.....	56
4.2 Experimentos.....	57
4.3 Resultados	59
4.3.1 Fusão com Redes Neurais	60
4.3.2 Fusão com <i>Naive Bayes</i>	61
4.3.3 Fusão com <i>Support Vector Machine</i>	62
4.3.4 Tempo de Execução em Classificação	63
4.3.5 Influência do emoCrawler.....	64
4.3.6 Tempo de Execução no emoCrawler	67
5 Comentários Finais	68
5.1 Publicações.....	69
5.2 Trabalhos Futuros.....	70
6 Referências Bibliográficas	72

LISTA DE FIGURAS

Figura 1.1. Interfaces gráficas pioneiras foram produzidas pela Xerox entre 1980-1985. Surgem conceitos de metáforas, ícones e janelas. Fonte: [THOCP, 2007].....	1
Figura 1.2. Alguns esforços buscam investigar o efeito psicológico de interagir com um robô com feições humanas e o ganho consequente em usabilidade. Ainda assim, pouco se desenvolve em relação à compreensão das emoções do usuário pelo sistema. Fonte: [Hellweg & Carroll, 2011]	2
Figura 2.1. Campo AV. Adaptado de [Laros, 2005].	12
Figura 2.2. Superfície Catastrófica [Sander et al, 2005].	13
Figura 2.3. Fluxo do processo emocional: causas, avaliações, mudanças fisiológicas, tendência/ação/expressão, regulação.	14
Figura 2.4. Regras do modelo Ortony-Clore-Collins (Ortony et al, 1988, apud [Höök, 2012])	15
Figura 2.5. Expressões emocionais básicas (no sentido horário: raiva, medo, desagrado, surpresa, felicidade, tristeza (Ekman, 1975, apud [Höök, 2012]))	18
Figura 2.6. Experimento para identificação da relação de modos emocionais. O gráfico ilustra a frequência relativa de respostas “triste” em função do contínuo de faces, quando a vocalização associada era “triste”, “neutra” ou “feliz”. Fonte: [Campanella e Belin, 2007].	25
Figura 2.7 Modelo Omnibus: um exemplo de arquitetura para fusão de sensores. Modificado de [Bedworth & O’Brien, 1999].	27
Figura 2.8 Fusão de sensores ao nível de <i>features</i> (a) e decisão (b). Inspirado em [Pantic & Rothkrantz, 2003].....	29
Figura 2.9. Modelo de Neurônio Artificial.....	31
Figura 2.10. Estrutura de Rede Neural Probabilística. Adaptado de [MathWorks01]...	32
Figura 2.11. Grupos separados por SVM. Os elementos sombreados são os vetores de suporte.	35

Figura 2.12. O fundamento por trás das função kernel. A transformação para um espaço de dimensão maior pode tornar o problema original (a) separável por um hiperplano (b) .	36
Figura 2.13. Processo de classificação SVM multiclasse tipo “um contra todos”.	37
Figura 3.1. Área de pesquisa. A etapa de pré-processamento e os itens hachurados foram desenvolvidos no escopo do trabalho.	39
Figura 3.2. FFBN utilizada para fusão de expressões emocionais (12, 7 e 4 nós respectivamente em cada camada).	42
Figura 3.3. Fluxo de Informações no emoCrawler. O processo de estimação corresponde à média dos valores gerados para diferentes expressões.	46
Figura 3.4. Estabilidade do emoCrawler em função do número de amostras (N) – resultados sem normalização dos pesos.	47
Figura 3.5. Avaliação emocional de diferentes palavras-chave em função do tempo decorrido (em minutos) – resultados sem normalização por pesos.	50
Figura 4.1. Exemplos de amostras do eNTERFACE.	53
Figura 4.2. Aplicativo eMotion em funcionamento.	55
Figura 4.3. Vídeo classificado por humano como "Felicidade", mas detectado incorretamente pelo <i>software</i> eMotion. A linha laranja representa o valor médio de “Felicidade” para todo o vídeo. As ordenadas representam o nível de detecção, enquanto as abscissas indicam a progressão de quadros de vídeo (N).	56
Figura 4.4. Vídeo classificado por humano como "Felicidade" e detectado corretamente pelo <i>software</i> eMotion. A linha laranja representa o valor médio de “Felicidade” para todo o vídeo. As ordenadas representam o nível de detecção, enquanto as abscissas indicam a progressão de quadros de vídeo (N).	56

LISTA DE TABELAS

Tabela 2.1. Relação entre sistemas do organismo e as funções e componentes da emoção. Adaptado de [Scherer, 2005].....	16
Tabela 3.1. Listagem de expressões do dicionário emocional. Diversas interjeições e <i>emoticons</i> estão presentes no dicionário, mas foram omitidos nesta listagem.	48
Tabela 4.1 Algumas expressões semânticas contidas no banco eNTERFACE 2005. Os itens em negrito nas frases representam palavras processadas pelo emoCrawler. .	54
Tabela 4.2. Exemplo de ficha de resultados.	59
Tabela 4.3. Taxa percentual de acertos em análise unimodal – conjuntos com número de amostras variável.	59
Tabela 4.4. Comparativo das medições individuais com a fusão multimodal em redes neurais: taxas percentuais de acerto para cada método.	61
Tabela 4.5. Comparativo das medições individuais com a fusão multimodal em <i>Naive Bayes</i> : taxas percentuais de acerto para cada método.	61
Tabela 4.6. Comparativo das medições individuais com a fusão multimodal em SVM multiclasse: taxas percentuais de acerto para cada método.	62
Tabela 4.7. Matriz de confusão para a técnica SVM multiclasse. Nas linhas encontram-se os valores previstos pelo classificador.	62
Tabela 4.8. Tempo médio para treinamento e posterior execução de cada classificador, em segundos.	63
Tabela 4.9 – Resultados da execução do emoCrawler para palavras diversas (não contidas no <i>corpus</i>), sem qualquer etapa de classificação. Aproximadamente 1500 buscas foram realizadas por palavra. Itens com (*) indicam expressões contidas no dicionário emocional e que, por consequência, são prontamente identificadas.	64
Tabela 4.10 – Avaliação da eficiência do emoCrawler sobre o grupo de teste para a FFBP: taxas percentuais de acerto em cada caso.	66

LISTA DE ABREVIATURAS

PNN	<i>Probabilistic Neural Network</i>
FFBP	<i>Feedforward Backpropagation</i>
SVM	<i>Support Vector Machine</i>
AV	<i>Arousal – Valence</i>
API	<i>Application Programming Interface</i>

LISTA DE SÍMBOLOS

Alfabeto Romano

$A, B, I, E1, \dots, En$	Eventos quaisquer
kb	Termo constante em <i>Naive Bayes</i>
W	Matriz de coeficientes para definição de um hiperplano
X	Vetor variável de um hiperplano
u	Vetor de termos independentes em um hiperplano
d	Distância de um hiperplano a seu vetor de suporte (SVM)
Sk	Sensores de aquisição de expressões emocionais ($k=1, \dots, n$)
Er	Expressões emocionais detectadas pelos Sk ($r=1, \dots, m$)
Pkr	Probabilidade de um sensor Sk determinar corretamente uma expressão emocional
Pr	Probabilidade de acerto de uma determinada expressão com uma abordagem de medição combinada
f	Função de transferência resultante de um classificador
$[v], [adj], [p], [s]$	Elementos sintáticos: verbos, adjetivos, preposições e substantivos.
$EstFrase$	Estimativa quantitativa de expressão emocional gerada pelo emoCrawler para uma frase.
$EstPalavraN$	Estimativa quantitativa de expressão emocional gerada pelo emoCrawler para uma palavra N em uma frase.

Alfabeto Grego

ρ	Termo margem em SVM
--------	---------------------

Introdução

Desde o início do desenvolvimento da computação, caminham em paralelo as questões de desenvolvimento tecnológico em *hardware*, *software* e interface com usuário. Entretanto, ainda que nas últimas duas décadas o avanço da eletrônica e dos aplicativos tenha sido intenso, o processo de interação entre homem e máquina tem mantido abordagens consideradas canônicas.

Se por um lado *mouses*, teclados e telas sensíveis ao toque têm mantido presença de diversas formas durante muitos anos, os paradigmas estruturados de *software* para limitação de casos de uso como rótulos, ícones e menus estão ainda mais enraizados na própria existência da computação (Figura 1.1). Há pouco espaço para interação flexível.

De fato, interagir com uma máquina exige do ser humano a definição de um processo de conversão de suas necessidades em uma sequência objetiva e contida de procedimentos unitários definidos pela interface disponível. Não há como processar comandos enviados, detectar frustrações ou satisfações do usuário e reagir a elas.

Figura 1.1. Interfaces gráficas pioneiras foram produzidas pela Xerox entre 1980-1985. Surgem conceitos de metáforas, ícones e janelas. Fonte: [THOCP, 2007]

Alguns desenvolvimentos da última década têm, contudo, sido promissores na busca de uma interação mais natural com o indivíduo. As chamadas NUI's (*Natural User Interfaces*), por exemplo, são um passo positivo e renovado para a melhora dessa interação. Interfaces multi-toque, com resposta tátil a movimentos manuais sobre a tela são um exemplo disso. A popularização recente de telefones celulares e *tablets* com tal tecnologia mostra seu apelo e a urgência do ser humano por experiências de computação que mimetizem a realidade, que se confundam com a coisa física.

Não é de se espantar que dispositivos ainda mais complexos de reconhecimento de movimentos corporais, voz e rosto, tenham sido incorporados no cotidiano em *videogames*, tornando-se imensamente populares. Tecnologias de mapeamento

infravermelho do ambiente e reconstrução tridimensional de objetos estão permitindo experiências antes reservadas à ficção científica. Contudo, transitando pela grande maioria dessas novas maneiras de acesso à máquina, é possível perceber que faltam subsídios para que a melhoria da interação passe do simples mecânico para o intelectual. De fato, faltam contribuições essenciais de inteligência artificial (IA) para que a interação seja mais rica e complexa.

A inteligência artificial foi a promessa de revolução tecnológica nas décadas de 60 e 70 [Russel & Norvig, 2003]. Acreditava-se que em algumas dezenas de anos conseguiríamos criar um ente artificial que não apenas passasse com louvor o teste de Turing [Turing, 1950], mas que de fato raciocinasse e compreendesse seu mestre no auxílio de tarefas cotidianas. Desenhos animados, filmes e literatura projetavam um futuro próximo onde raciocínio natural e artificial embaralhavam-se em um grande acúmulo de expectativas (Figura 1.2).

Contudo, enquanto as promessas de robôs conscientes permaneciam no folclore coletivo, na academia e na indústria percebia-se que as dificuldades acerca desse problema seriam fatalmente intransponíveis sem um salto qualitativo na abordagem da computação¹. As expectativas sobre o futuro da IA foram então reprimidas, aplicadas, sobretudo, em sistemas especialistas, nos quais algoritmos têm alguma autonomia para

Figura 1.2. Alguns esforços buscam investigar o efeito psicológico de interagir com um robô com feições humanas e o ganho consequente em usabilidade. Ainda assim, pouco se desenvolve em relação à compreensão das emoções do usuário pelo sistema. Fonte: [Hellweg & Carroll, 2011]

avaliar situações específicas e tomar decisões flexíveis. Entretanto, mesmo em aplicações tão restritas, a falta do caráter procedural na IA era frequentemente visto com maus olhos pela indústria, onde o controle total sobre os estados de um algoritmo é tratado como uma necessidade em aplicações de segurança (controle de veículos,

¹ Tal período da história é considerado o primeiro “Inverno da IA” [Russel & Norvig, 2003].

aeronaves, plantas industriais, etc.). Havia o receio em ceder processos decisórios à máquina.

De qualquer forma, caminhando para o desenvolvimento de sistemas especialistas, a inteligência artificial refreou sua ousada premissa original de compreender as minúcias da interação humana.

A interação entre dois ou mais indivíduos envolve a compreensão do contexto semântico do que é conversado, a observação de gestos e das expressões faciais, da intensidade, frequência e timbre da fala e o atrelamento desses todos à identificação de objetos físicos que são apresentados durante essa interação. São processos complexos, simultâneos, e que envolvem a fusão de dados provindos de diferentes aparatos biológicos para suporte de um único objetivo: o foco para manter a conversa, obtendo reações e informações desejadas.

Para a máquina, emular a capacidade humana de contextualização de uma conversa e de reação flexível ao significado semântico seria uma forma mais natural e produtiva de lidar com o problema de interação: se o computador fosse capaz de se utilizar dos recursos dos diversos sinais linguísticos e não linguísticos presentes no cotidiano humano, (tais como expressões faciais, o tom da voz e o contexto afetivo da interação), seria possível obter máquinas capazes de compreender de forma mais adequada as necessidades e dificuldades particulares do usuário, suas frustrações, motivações e individualidades de caráter.

Parte fundamental dessa compreensão envolve a modelagem das emoções do usuário, tema principal do presente trabalho. Estimativas conservadoras [Birdwhistell, 1972] afirmam que algo em torno de 65% das informações captadas em um diálogo não estão presentes no significado objetivo do que é dito, o que indica que grande parte das habilidades de comunicação do indivíduo é perdida quando o mesmo tenta extrair determinado resultado de uma máquina.

Motivados por essas questões, diversos trabalhos buscam elaborar algoritmos computacionais para detecção afetiva, seja ela extraída de elementos faciais, gestuais ou vocais, de forma a contribuir para um avanço incremental nos problemas da interação. Em geral, tais trabalhos suportam-se sobre a amplamente referenciada classificação facial elaborada por Paul Ekman [Ekman & Oster, 1979], a qual distingue seis blocos de

expressões emocionais básicos: felicidade, tristeza, desagrado, medo, raiva e surpresa. É uma abordagem simples, mas abrangente; um suporte sobre o qual mais complexas análises podem ser incorporadas.

Na última década, com o vasto crescimento da base de dados opinativos da Internet (*blogs*, fóruns, redes sociais) alguns trabalhos têm igualmente tentado observar emoções a partir de elementos semânticos básicos da conversa, comparando-os com a enorme base de textos disponível na rede. Utiliza-se neles o senso comum dos usuários como forma de obtenção de dados.

Posto isso, ainda que o campo da computação afetiva desenvolva-se rapidamente, as técnicas existentes para o tratamento das emoções apresentam-se atualmente limitadas. A dificuldade de adquirir robustez torna esses sistemas pouco confiáveis e frequentemente irrealizáveis em aplicações de alta complexidade. Além disso, a abordagem unimodal (isto é, observação de uma única fonte de emoção) ignora a relação intrínseca entre essas diversas entradas afetivas, cuja relevância é discutida e estudada em detalhes por Scherer e Ellgring [Scherer & Ellgring, 2007].

Tais desafios trazem consigo a herança da cultura do desenvolvimento de uma inteligência artificial pouco focada em aspectos cognitivos. Em última análise, a interação robusta e a inteligência artificial estão intimamente conectadas: a melhoria da interação homem-máquina depende do avanço na inteligência artificial ao mesmo tempo em que o avanço da inteligência artificial depende da melhoria na riqueza da interação homem-máquina.

1.1 Objetivos

Dado o contexto da interação inteligente homem-máquina, este trabalho propõe investigação e desenvolvimento na área de detecção de expressões emocionais por algoritmos computacionais.

Nesse âmbito, avalia-se a robustez de soluções individuais de detecção afetiva e propõe-se o uso combinado dessas ferramentas - neste caso atuando sobre face, voz e semântica do discurso - através de fusão de dados por meio de diferentes algoritmos classificadores. No caso das entradas de face e voz, as informações são extraídas de

trechos de vídeo processados a partir de ferramentas de detecção de emoções já existentes, consideradas referência na área de pesquisa.

No que se refere à entrada sensorial de semântica do discurso, propõe-se o desenvolvimento de uma ferramenta nova e específica deste trabalho, denominada “emoCrawler” para colaboração e aumento da diversidade no processo de fusão de emoções.

Os testes são realizados através de banco de vídeos que correspondem ao ambiente potencial em que o sistema global de detecção poderia atuar. Neste primeiro passo, contudo, não se busca o processamento dos dados em tempo real, ainda que seja um objetivo viável na continuação das pesquisas.

1.2 Contribuição

Como contribuição técnica, este trabalho busca avaliar objetivamente diferentes técnicas de fusão sensorial aplicadas à determinação da emoção expressa pelo usuário. Além disso, o trabalho desenvolve uma ferramenta dedicada para interpretação semântica de emoções, de forma que sua combinação com sensores faciais e vocais em fusão coloca nova luz sobre a possibilidade de detecção robusta de emoções presentes em discursos diretos.

1.3 Aplicações

O movimento da indústria e da academia em busca de experiências artificiais com as quais o usuário possa se identificar em um nível emocional tem aumentado significativamente na última década, graças ao crescimento do poder computacional e da sedimentação de uma base teórica sólida. Com a volta do enfoque da IA em processos focados no humano, diversos desafios técnicos inéditos têm sido colocados em foco. Nesse sentido, novas tecnologias e abordagens de programação (tal como [SOAR, 2006]) têm permitido resultados pioneiros e motivantes na área, os quais têm tido intensos ciclos de desenvolvimento para o oferecimento de novas soluções.

Algumas das áreas em que trabalhos relacionados a esta pesquisa são aplicados incluem:

- Segurança em operações industriais (prevenção de acidentes, análise da

condição psicomotora do operador de máquina);

- Entretenimento (análise de motivação/frustração, correção de mecanismos de recompensa psicológica em jogos eletrônicos);
- Produtividade em computação (controle semântico de computadores baseado em conceitos abstratos com compreensão e adaptação às necessidades do usuário);
- Problemas abertos que frequentemente requerem forte interação humana (“balcão” de informações, comando de veículos tripulados);
- Qualidade de serviço (detecção de estado emocional do atendente e cliente para rotas de conversa em suporte técnico e vendas por telefone);
- Auxílio na obtenção de diagnóstico psicológico (uso clínico, criminalístico, como sugerido por [Ekman, 2001]).

1.4 Estrutura do Texto

Este trabalho está organizado de forma a apresentar inicialmente o estado da tecnologia disponível para reconhecimento emocional e, em seguida, sugerir novas abordagens para melhoramento das técnicas existentes.

No Capítulo 2 é apresentado um breve histórico da relação da sociedade com as emoções e de como o ser humano vem tentando formalizar sua sistemática através de modelos. São apresentados os conceitos necessários ao desenvolvimento dos sensores e os trabalhos que trazem para o mundo computacional sistemas de detecção afetiva por meio de face, voz, semântica e através de algoritmos de fusão. Por fim, justifica-se e descreve-se conceitualmente o ferramental de modelagem matemática utilizado no trabalho.

O Capítulo 3 apresenta o modelo proposto para lidar com a problemática da detecção de emoções humanas. São apresentados os algoritmos aplicados, seu contexto no trabalho global e a justificativa das abordagens consideradas. O capítulo também apresenta uma introdução ao sistema emoCrawler, desenvolvido no contexto exclusivo deste trabalho para aumento da riqueza dos resultados em fusão multimodal. Nesse sentido, é apresentada sua motivação, as técnicas utilizadas e breves ensaios realizados durante sua construção.

O Capítulo 4 enfoca essencialmente os experimentos propostos e os resultados obtidos. No item 4.1, são apresentadas as ferramentas produzidas por terceiros utilizadas no trabalho, assim como o *corpus* necessário à validação dos testes. O item 4.2 do capítulo apresenta os resultados obtidos através da fusão multimodal com uso de redes neurais, *Naive Bayes* e *Support Vector Machine*, avaliando parte dos resultados obtidos em face a presença ou não do emoCrawler.

Por fim, o Capítulo 5 apresenta conclusões sobre os trabalhos realizados. O direcionamento de pesquisas futuras também é discutido.

2 Contextualização e Revisão Bibliográfica

Neste capítulo são apresentados conceitos essenciais a esta dissertação, tanto do ponto de vista humano e psicológico, quanto tecnológico e procedimental. Para isso, expõe-se inicialmente um breve panorama da história do estudo de emoções, além de seu papel na ciência moderna. Em seguida, detalhes sobre alguns dos elementos técnicos e ferramentas matemáticas utilizados durante a pesquisa são abordados.

2.1 Histórico do Estudo de Emoções

A tentativa de entendimento das emoções, suas funções e implicações no ser vivo tem paralelo com a história do autoconhecimento da humanidade. A emoção surge para o homem como um mecanismo de diálogo.

Um dos primeiros formalismos significativos sobre o uso e consequência da emoção de que se tem registro está presente na Retórica de Aristóteles, Livro II [Aristóteles, s.d]:

“Os fatos não se apresentam sob o mesmo prisma a quem ama e a quem odeia, nem são iguais para o homem que está indignado ou para o calmo, mas, ou são completamente diferentes ou diferem segundo critérios de grandeza.”

Ainda que através de uma abordagem aplicada e pouco científica, Aristóteles apresentava como o estado emocional do expositor e do ouvinte poderiam ser utilizados como ferramenta de manipulação argumentativa, realizando ali uma separação entre o estado emocional real e a expressão emocional controlada.

Mesmo que diversos povos tenham empreendido descrição de emoções através da Idade Média - árabes, por exemplo, frequentemente buscavam observações médicas - coube sobretudo à teologia europeia refletir filosoficamente sobre suas questões primárias, suportados por um arcabouço greco-romano de reflexões [King, 2010]. O contexto medieval altamente religioso influenciava tais retóricas, exaltando a volição como fonte das emoções. A visão puramente estoica de outrora era questionada. Santo Agostinho (354-430), por exemplo, clama em A Cidade de Deus [St. Agostinho, apud (King, 2010)] que ainda que diversas emoções sejam distúrbio, fruto da punição divina,

outras têm caráter genuíno e devem ser cultivadas: o reflexo de compaixão, o “temer a Deus”, etc..

Não houve grande ganho na área com o início do empirismo da Idade Moderna. Considerado como de baixa relevância e como uma expressão de incapacidade racional, o estudo das emoções foi frequentemente desconsiderado por pesquisadores. Seria apenas em 1872, com Charles Darwin, em seu trabalho “*The Expression of the Emotions in Man and Animals*” [Darwin, 1872] que maior atenção seria dada ao tema. Nele, Darwin tenta suportar sua então conturbada teoria da evolução com aspectos do surgimento das emoções nos animais. As emoções nos seres vivos, assim como tantas outras características, teriam surgido gradativamente através de longos ciclos de seleção natural, encerrando-se como vestigiais no ser humano.

Paralelamente, o surgimento da psicologia iniciava a abertura de novas oportunidades para o estudo afetivo. William James escreve, em 1884, o artigo fundamental no campo: “*What is an Emotion*” [James, 1884], o qual não fornecia respostas definitivas, mas trazia profundas reflexões filosóficas sobre a sequência de eventos que acarretam a geração da emoção. James valorizava a influência da emoção sobre a psique humana, escrevendo que eventos “poderiam ser tão excitantes emocionalmente que conseguiriam praticamente deixar uma cicatriz nos tecidos cerebrais” (apud [Muramatsu & Hanoch, 2005], tradução do autor).

Sigmund Freud (1856- 1939) também trataria emoções como itens importantes na psique, podendo, em grande intensidade ou repressão causar distúrbios diversos no pensamento cognitivo. Para ele, emoções estariam associadas a pensamentos e percepções inconscientes, surgindo antes do autoconhecimento de eventos [Freud, apud (Watson & Rayner, 1920)]. As emoções dão pistas sobre a opinião, assim como Aristóteles exploraria milhares de anos antes.

Durante o período de dominância do behaviorismo de Skinner (1950-1975), o aspecto mental das emoções foi novamente rejeitado. O behaviorismo valorizava o empirismo e a observação, de forma que a dificuldade de obter dados objetivos sobre qualquer comportamento emocional tornava difícil alguma contribuição.

A partir da década de 80, alguns trabalhos começam a refletir sinteticamente sobre o passado. Em 1980, notadamente, a publicação de “*Feeling and Thinking: Preferences*

Need No Inferences” por Robert Zajonc argumentava que sistemas cognitivo e afetivo teriam características independentes, com o segundo tendo maior força no indivíduo do que o primeiro. Resultados experimentais reforçariam a ideia de que o julgamento afetivo poderia preceder no tempo as operações cognitivas e de percepção associadas às emoções [Zajonc, 1980]. O artigo causou furor e contribuiu como diversos outros para que o papel do afeto e dos processos emocionais voltasse ao primeiro plano nas discussões em psicologia.

Atualmente, o preconceito sobre a falta de “pureza” do tema tem sido reduzido admiravelmente, sobretudo com a popularização da psicologia no cotidiano e com a introdução de produtos e soluções tecnológicas diversas suportados pelas várias teorias emocionais [Broek & Westerink, 2009]. Em suma, vê-se que o legado das teorias do passado se complementa com novos enfoques das abordagens em psicologia e neurociência cognitiva, tais como os vistos nas contribuições de Damasio [Bosse et al, 2008].

2.2 Conceitos em Computação Afetiva

O termo “Computação Afetiva” abrange as linhas de pesquisa que buscam tanto o desenvolvimento de sistemas que consigam compreender e lidar com emoções humanas (tópico deste trabalho), quanto sistemas que consigam exprimir emoções humanas de uma forma contextual e relevante. Termo cunhado por Rosalind Picard em um artigo homônimo de 1995 [Picard, 1995], a computação afetiva vem ganhando espaço em aplicações diversas a partir do ano 2000. Frequentemente associada à indústria do entretenimento, a pesquisa da compreensão emocional e comportamental tem na realidade aplicações igualmente amplas nas áreas de medicina (tratamento de síndromes, terapia), segurança de operações (avaliação contínua de operadores de máquinas, motoristas), qualidade de serviços, entre outras.

Dentre os diversos suportes teóricos para a computação afetiva, a linha de estudo das Teorias Cognitivas (*Appraisal Theories*), discutida por [Roseman, 2001] e [Schorr, 2001], fornece um paradigma adequado ao escopo de diversos trabalhos relacionados à cognição na área da computação [Scherer, 2010], ao passo que oferece um modelo que explica as diferenças comportamentais de cada indivíduo, ao mesmo tempo em que

determina aspectos comuns a todos. Para essa linha de pensamento, os processos de elicitación de emoções são comuns a todas as pessoas, mas o desenvolvimento desses processos varia individualmente, respeitando a experiência de vida de cada um. Além disso, prevê que a geração de cada emoção é centrada na avaliação do evento com o qual o indivíduo tem que lidar, em comparação com seus objetivos, expectativas e seu senso de responsabilidade sobre o acontecimento, itens particularmente enfatizados na abordagem de Roseman.

Contudo, para que tais processos e conceitos façam sentido em termos de modelagem lógica, surge a necessidade de se classificar e gerir os estados emocionais possíveis. Uma possível abordagem apresentada em diversos trabalhos é a realização da classificação de emoções de formas qualitativamente discretas, como descrito por Goleman (apud [Martinez-Miranda & Aldea, 2005]). Entretanto, além da evidente dificuldade de gerenciar um conjunto independente de termos frequentemente subjetivos, essa alternativa apresenta problemas na transição de estados – em como definir quais emoções originam outras e como descrever quantitativamente essas variações.

Para outros autores, como [Laros, 2005], as emoções devem ser modeladas e observadas em um espaço contínuo e com dinâmica definida, no qual parâmetros ditam a mudança contínua de humor de um estado para outro, tomando como base indicadores diversos.

Um dos modelos mais simples nesse sentido é conhecido como *Arousal-Valence* (AV). Nele, define-se inicialmente um plano em que as abscissas representam a valência de uma dada emoção (isto é, se é qualitativamente positiva ou negativa) e as ordenadas representam o grau de ativação (agitação) gerado pela emoção. Sobre esta base ortogonal observa-se que é possível mapear diferentes estados emocionais como regiões do espaço. Assim, ainda que, por exemplo, “raiva” e “tristeza” tenham qualidade de valência negativa, ambas são diferenciadas por sua intensidade, apresentando-se em quadrantes diferentes. A Figura 2.1 mostra um exemplo de plano AV:

Figura 2.1. Campo AV. Adaptado de [Laros, 2005].

Tomando como base a superfície A.V., alguns trabalhos implementam dinâmicas de transição contínuas. Em [Gonçalves et al 01, 2011], por exemplo, uma superfície com algumas características análogas ao plano A.V. é curvada de forma a gerar um parabolóide centrado em (0,0) e o estado emocional é modelado como uma partícula sem massa, deslizando sobre a superfície, atraída para o centro.

Há evidentemente limitações na abordagem bidimensional. Além de “*Arousal/Intensity*” e “*Valence*”, diversos outros parâmetros devem ser incorporados para um mapeamento adequado de uma quantidade abrangente de emoções, obtendo-se assim um espaço multidimensional com potencial de explosão de estados. Em [Sander et al, 2005], há a descrição de “*Superfícies Catastróficas*”, que representam a dinâmica de transição de emoções em situações específicas e mais complexas. Um exemplo bastante referenciado atribuído a [Zeeman, 1979] corresponde à modelagem do comportamento de um cão quando enfrentado por um adversário de força desconhecida. Nele, uma curva de histerese descreve transições abruptas de comportamento, como o incremento da raiva levar ao ataque ou o medo levar a fuga. A Figura 2.2, elaborada por Sander, ilustra um exemplo de modelagem complexa. É possível observar as transições abruptas e regiões inacessíveis descritas por Zeeman:

Figura 2.2. Superfície Catastrófica [Sander et al, 2005].

2.2.1 O Processo Emocional

Para que seja possível estimar o comportamento emocional como um sistema causal, é necessário compreender os estágios que desencadeiam a emoção e esforços necessários para remover o indivíduo de seu estado neutro. Nesse sentido, um conceito fundamental na modelagem emocional encontra-se no chamado “Processo Emocional”. Ainda que diversas teorias se confrontem sobre aspectos diversos dos processos que desencadeiam emoções, alguns componentes de fluxo operacional aparecem na maioria das abordagens, ainda que com diferentes nomes [Planalp, 1999]. A Figura 2.3 ilustra o fluxo mais comumente observado na literatura:

Figura 2.3. Fluxo do processo emocional: causas, avaliações, mudanças fisiológicas, tendência/ação/expressão, regulação.

Dessa forma, diz-se que todo processo emocional tem um objeto gerador interno ou externo (causa), o qual, ao ser exposto ao indivíduo, origina um procedimento de análise interior para atribuição do significado emocional (avaliação). A partir desse significado, o indivíduo experimenta mudanças corporais (tensão muscular, dilatação de pupila, variação de frequência cardíaca, etc.) e a intenção de realizar uma ação como resposta àquela emoção (tendência). Finalmente, o indivíduo realiza um processo reflexivo a fim de lidar com o evento emocional experimentado (regulação). Assim, por exemplo, uma criança que tem seu brinquedo preferido retirado de si (causa) realiza um processo interno de averiguação do ocorrido, ocasionando tristeza (avaliação) e exteriorizando choro (mudanças fisiológicas). Suas reações seguintes podem ser diversas, como tentar pegar o brinquedo de volta ou tentar atacar o indivíduo em posse do mesmo (ação). Após algum tempo, a criança reavaliará a importância da situação, tentando lidar com o ocorrido, investindo no confronto ou buscando uma nova distração (regulação).

Em Ortony et al, 1998 (apud [Höök, 2012]) a apresentação de uma abordagem mais completa para definição de classificação de emoções gera uma das mais conhecidas abordagens do processo emocional, discutindo particularmente o ambiente das causas. A Figura 2.4 ilustra um exemplo de fluxograma de regras para manipulação emocional elaborado por Ortony:

Se questionadas sobre o assunto, diferentes pessoas definirão o tópico de diferentes formas, baseadas em seu contexto social e aprendizado de vida. Particularmente, emoções são frequentemente confundidas com as definições de "sentimento", "humor" e "atitude", itens distintos e tão complexos quanto no desenvolvimento da personalidade humana. Percebe-se a dificuldade da definição quando até mesmo o primeiro texto relevante da área, "*What is an Emotion*", de Willian James (inicialmente mencionado na seção 2.1), traz incoerência de definições [Scherer, 2005].

A definição considerada neste trabalho corresponde a dos trabalhos de [Scherer, 2001], os quais afirmam que emoções são "um episódio de mudanças inter-relacionadas e sincronizadas nos estados de todos ou da maioria dos cinco subsistemas do organismo em resposta a avaliação de um evento de estímulo interno ou externo relevante ao organismo" (t.d.), onde tais "subsistemas" correspondem aos componentes dedicados do sistema nervoso (processamento de informações, suporte, execução, ação, monitoração), como ilustrado na Tabela 2.1:

Tabela 2.1. Relação entre sistemas do organismo e as funções e componentes da emoção. Adaptado de [Scherer, 2005].

Função da emoção	Subsistema do organismo e substrato principal	Componente emocional
Avaliação de objetos e eventos	Processamento de informações (SNC)	Componente cognitivo (avaliação)
Regulação do sistema	Suporte (SNC, SNE, SNA)	Componente neurofisiológico (sintomas corporais)
Preparação e direção da ação	Execução (SNC)	Componente motivacional (tendências de ação)
Comunicação de reação e intenção comportamental	Ação (SNS)	Componente de expressão motora (expressão vocal e facial)
Monitoração de estados internos e interação organismo-ambiente	Monitoração (SNC)	Componente de sensação subjetiva (experiência emocional)
Nota: SNC: sistema nervoso central; SNE: sistema neuroendócrino; SNA: sistema nervoso autônomo; SNS: sistema nervoso somático		

Nota-se dessa abordagem, por exemplo, que um "sentimento" atuaria como um componente da experiência emocional, com função prolongada de monitoramento e regulação do processo.

Além de definir o conceito, Scherer sugere alguns parâmetros relevantes que poderiam separar a emoção de outros processos afetivos, entre eles "velocidade de mudança" (emoções estariam passando por constantes reajustes através de avaliações), "intensidade" (emoções seriam intensas, tomando conta de aspectos do comportamento) e "duração" (emoções seriam curtas, de modo a não taxar demasiadamente o organismo e permitir flexibilidade no comportamento). O presente trabalho suporta-se nessas três características para a classificação automática de emoções.

Finalmente, é importante ressaltar que este trabalho busca a detecção da emoção do usuário com base na expressão emocional gerada pelos diversos mecanismos biológicos, a qual pode ou não corresponder de fato à emoção sentida. O termo "expressão" corresponde neste contexto, portanto, ao que é medido diretamente pelos diferentes subsistemas computacionais. O sucesso dessa medida desconsidera uma avaliação aprofundada sobre qualquer tentativa do usuário de esconder seu estado emocional.

2.3 As Emoções na Face

A expressão facial tem aspecto extremamente importante na sociedade humana. É através dela que não apenas emoções genuínas são apresentadas, mas diversas demonstrações manipulativas e de convenção social são geradas. Por meio do longo ciclo evolutivo da espécie, a face humana tornou-se também complexa em sua expressividade, apresentando um conjunto de 43 músculos que podem ser movimentados a qualquer momento para a construção de uma expressão qualquer [Foreman, 2003].

Um sorriso ou expressão de desdém podem ter enorme influência na interação entre indivíduos, de forma que se torna instintivo em qualquer conversa o foco visual nos olhos do interlocutor e na região periférica, em busca de sinais de aprovação ou rejeição de hipóteses.

Paul Ekman é considerado um dos pesquisadores com contribuições mais significativas no campo compreensão de emoções através da face [Foreman, 2003]. Em [Ekman & Friesen, 1978], propõe-se a divisão do rosto em *Action Units* (AUs), elementos básicos relacionados ao padrão de musculatura que podem possuir algum movimento durante a expressão de uma emoção. Dessa forma, a ativação de uma sequência específica de AUs poderia descrever uma determinada expressão ou uma combinação de expressões. Combinações comuns, como “Medo + Raiva” e “Tristeza + Medo” são contempladas pelo modelo.

Ekman também é responsável por definir um conjunto padrão de expressões emocionais simples - Felicidade, Medo, Tristeza, Raiva, Surpresa e Desagrado (Figura 2.5) – as quais são base de grande parte dos algoritmos computacionais de detecção facial [Bettadapura, 2012]. Um dos motivos para o largo uso de tal conjunto encontra-se em [Ekman, 1972], trabalho no qual o autor propõe, através de experimentos, que emoções básicas têm caráter universal e não são aprendidas por mimetismo. Assim, as soluções de *software* suportadas por essa metodologia poderiam ser mais simples, sem perda de generalidade.

A afirmação de universalidade de Ekman tem sido alvo de diversas contestações ao longo dos anos, tal como em [Naab & Russel, 2007].

Figura 2.5. Expressões emocionais básicas (no sentido horário: raiva, medo, desagrado, surpresa, felicidade, tristeza (Ekman, 1975, apud [Höök, 2012])

Técnicas Computacionais

No processo de identificação de emoções por reconhecimento facial, [Bartlett et al., 1999] implementaram alguns dos primeiros algoritmos bem sucedidos para automação computacional do processo, comparando diferentes técnicas de obtenção de dados. No trabalho em questão, um algoritmo possibilitou, após um período de aprendizagem, o reconhecimento de cerca de 90% das emoções apresentadas em fotografias fornecidas em condições controladas ao computador.

Entretanto, ainda que o reconhecimento de imagens estáticas tenha tido sucesso relativo já no início da década de 1990 [Bettadapura, 2012], a questão do reconhecimento em tempo real para vídeo provou-se problemática durante muitos anos, visto que os algoritmos de processamento aplicados a cada quadro são intensos e bastante dependentes da capacidade computacional disponível.

Nas últimas décadas, o aumento do poder computacional deu origem a diversas vertentes para detecção em vídeo. Entre elas, vê-se simples adaptações das soluções estáticas, além de abordagens que buscam incorporar o conhecimento sobre as mudanças temporais na estimação. Dentre essas soluções, pode-se mencionar o ajuste de modelos tridimensionais (tal como em [Azcarate et al, 2005]), as quais sobrepõem uma malha de vértices sobre elementos de fácil distinção na face monitorada, tentando realizar a movimentação dos pontos conforme a face real deforma-se. Outra abordagem encontra-se na avaliação pura do fluxo óptico da imagem (inicialmente investigado em [Essa & Pentland, 1997]), observando a velocidade, o deslocamento e as deformações dos diversos agrupamentos de pixels. Os padrões detectados são registrados e expressões são associadas a eles. Finalmente, com a popularização e o aumento da facilidade de aquisição de equipamento para registro da componente tridimensional do ambiente em um vídeo, alguns trabalhos têm buscado realizar toda a detecção das expressões diretamente de um modelo capturado através de configurações de câmeras estereoscópicas, luz estruturada, entre outros. Nesse caminho, [Sun & Yin, 2008] apresentam uma abordagem espaço-temporal sobre um Modelo de Markov Oculto, obtendo até aproximadamente 90% de acerto na detecção.

Contudo, ainda que as contribuições na área tenham sido abundantes no passado recente, grande parte dos trabalhos assumem condições excessivamente controladas para captura de vídeo. Não apenas restrições de luminosidade e de exposição da face são empecilhos, mas também a complexidade das expressões detectadas geralmente é baixa. Além disso, são poucos os trabalhos que tentam discutir a qualidade de autenticidade ou falsidade das expressões apresentadas pelo indivíduo (como [Valstar et al, 2007]).

2.4 As Emoções na Voz

Desprezando a semântica embutida, emoções transitam na voz através de timbre, energia e frequência base da onda sonora. É a manipulação consciente ou inconsciente desses aspectos sonoros que permite que uma mãe tenha uma conversa tranquilizadora com seu bebê e que um político possa transmitir segurança e certeza em seu discurso.

A neurociência mostra que a vocalização determina assinaturas emocionais processadas em tempo real por regiões dedicadas do córtex auditivo cerebral [Thomson, 2009]. Há, portanto, enorme peso da modulação vocal na compreensão do estado emocional de um indivíduo pelo próximo.

Pesquisas sobre a relevância e impacto da transmissão de expressões emocionais pela voz são em geral recentes, considerando-se a relevância do tema. Entre os trabalhos mais antigos, pode-se mencionar [Scherer, 1981], no qual o pesquisador realiza um extenso apanhado de diversos experimentos em que indivíduos são convidados a detectar a emoção presente em discursos predeterminados. Scherer notou que os grupos obtiveram uma taxa de acerto médio de 60%, considerada alta dada a inclusão de emoções complexas nos experimentos realizados (como amor, ciúme e orgulho). Em [Scherer, 1995], o autor mantém o interesse na área, fazendo uma discussão detalhada da expressão de emoções na voz e música e trazendo à luz as influências da acústica sobre a detecção emocional.

Técnicas Computacionais

Os algoritmos para detecção de emoções em traços vocais normalmente buscam padrões específicos nas variações do discurso, tentando ignorar aspectos absolutos tais como a intensidade da voz [Vogt & André, 2005]. Com isso, conseguem eliminar

conclusões enviesadas provocadas pelos diversos procedimentos de captura do áudio. Para tal intuito, frequentemente empregam o uso de algoritmos de classificação probabilística em busca de padrões sonoros, tais como *Naive Bayes* e *Support Vector Machine*.

Nos últimos 15 anos, diversos projetos têm buscado encontrar um *feature-set* relevante para a extração de expressões. [Eyben,2009] é um dos trabalhos frequentemente referenciados que discute o assunto. Um apanhado abrangente sobre o estado da arte e sobre os desenvolvimentos incrementais na área também é feito por [Schuller et al, 2011].

Frequentemente, trabalhos discutem a classificação emocional na voz através da valência do tom do discurso (positiva, neutra ou negativa), enquanto outros tentam buscar comportamentos mais bem definidos (tais como a identificação de emoções de Ekman); [Vogt e André, 2005] fazem um comparativo dessas alternativas sobre um mesmo classificador, discutindo sua eficiência.

Atualmente, alguns trabalhos tem tentado incorporar detecção automática em experimentos sociais. [Rachuri et al, 2010], por exemplo, apresenta uma plataforma para avaliação de emoções baseada em telefones celulares. Trabalhando no problema contrário, alguns pesquisadores tentam gerar emoções [Oudeyer, 2002] a partir da voz para aplicações em robótica. [Kim et al, 2007], por sua vez, discute a implementação de sistemas computacionais do ponto de vista de tratamento temporal das observações emocionais.

2.5 As Emoções na Semântica

A riqueza da transmissão de emoções através do contexto verbal pode ser sobrepujante. Contudo, para que seja de fato possível compreender a extensão da influência das palavras na compreensão afetiva, é necessário remover qualquer alcance do tom de voz, expressão facial ou contexto físico da conversa. Para se destituir desses elementos, uma boa forma de observar o fenômeno emocional semântico é analisar a literatura.

A relevância da literatura dramática é prova de que é possível instigar reações emocionais intensas no leitor através de nada além da junção de palavras em uma ordem

coerente. Quando, por exemplo, a icônica personagem Augusto Matraga, de Guimarães Rosa recorda de sua família [Rosa, 1946]:

“E, aí, Nhô Augusto se lembrou da mulher e da filha. Sem raiva, sem sofrimento, mesmo, só com uma falta de ar enorme, sufocando. Respirava aos arrancos, e teve até medo, porque não podia ter tento nessa desordem toda, e era como se o corpo não fosse mais seu. Até que pôde chorar, e chorou muito, um choro solto, sem vergonha nenhuma, de menino ao abandono. E, sem saber e sem poder, chamou alto, soluçando: “– Mãe... Mãe...”.”

Fica evidente o trabalho do autor em concatenar os objetos do discurso de forma a descrever o momento de Matraga com grande magnitude emocional. Em primeiro plano, nota-se a escolha de expressões explícitas para a caracterização da sensação da personagem (“teve medo”, “chorou muito”, “soluçando”). Além disso, observa-se a utilização de figuras de linguagem (“menino ao abandono”, “era como se o corpo não fosse mais seu”), as quais embutem ainda mais intensidade e cor no discurso, de forma a criar o resultado emocional desejado, conjunto e compacto.

O que se observa é que, diferentemente de outras formas de expressão emocional, a semântica da conversa possui caráter menos emergente e espontâneo do que a expressão facial ou vocal (ainda que possam ser controladas). Há, em graus diversos, a preconcepção do discurso emocional pelo indivíduo, de forma que frequentemente é ajustado durante a conversação com intuito de colaborar com o convencimento do ouvinte, como ferramenta retórica [Averill, 2001].

Recentemente, o estudo do comportamento emocional em textos tem recebido grande atenção, ao passo que ciências humanas podem usar os novos arcabouços de produção digital das redes sociais online para compreensão dos mecanismos de interação entre indivíduos ou grandes grupos. As redes sociais apresentam-se como enormes bancos de dados de comportamento, extremamente ricos em variabilidade, prontamente disponíveis para pesquisadores de todas as áreas.

Técnicas Computacionais

O entendimento semântico na interação inteligente (diga-se, compreensão da fala natural ou de textos pelo computador) ainda é um problema em aberto, com soluções

apenas parciais. O primeiro trabalho significativo nessa área é frequentemente associado a [Charniak, 1993], no qual o autor sugere o uso de análise sintática estatística (“*statistical parsing*”) para extração e classificação de significados. Entretanto, o desenvolvimento de técnicas robustas prosseguiu de forma lenta nos anos seguintes à publicação da obra, sobretudo devido a problemas na representação dos dados obtidos através da análise. De fato, a dificuldade de representação do conhecimento transmitido através da fala é ainda tarefa complexa igualmente em estudo. Particularmente, [Brachman, 2004] traz discussões e algumas aplicações voltadas para raciocínio em inteligência artificial (campo conhecido como “KR&R” - *Knowledge Representation and Reasoning*).

Tem se observado nos últimos anos a redução das expectativas com sistemas de detecção semântica, com a produção de estruturas dedicadas a identificar aspectos específicos de uma conversa ou texto. Nesse sentido, agências de propaganda buscam automaticamente a reação de clientes a um determinado produto em fóruns na Internet; centrais de *telemarketing* avaliam conversas em busca de sinais de *stress* do operador ou do cliente, dado o desenvolvimento da conversa. São aplicações que se aproveitam do conteúdo sem que se voltem ao problema abrangente da compreensão genérica.

Da mesma forma, algumas aplicações dedicadas práticas têm buscado detectar apenas emoções em discursos, para aplicações diversas. [Ptaszynski et al., 2009] e [Tokuhsa et al., 2008] são alguns exemplos de trabalhos que se utilizam da Internet como banco de dados para compreensão emocional de palavras, com resultados animadores. Para os trabalhos dessa área, os maiores desafios encontram-se na detecção dos chamados “termos afetivos indiretos” [Strapparava et al, 2006], os quais correspondem a expressões não explícitas de emoção (tópico explorado na seção 3.3.1 deste texto). O problema neste caso consiste em encontrar um mapeamento entre um elemento textual aparentemente neutro e uma expressão emocional definida.

Da mesma forma que as ciências humanas avaliam as redes sociais em busca de padrões de comportamento, as novas pesquisas em detecção automática de emoções tem nelas um arcabouço para não apenas validação de hipóteses e algoritmos, mas também para treinamento de sistemas com aprendizado supervisionado.

2.6 Análise Multimodal de Emoções

Muito das expressões emocionais naturais correntes no ser humano são consideradas multimodais [Feldman & Rimé, 1991], isto é, geradas e expressadas em conjunto por diversos mecanismos biológicos: ao errar um pênalti, um jogador de futebol frequentemente leva a mão à cabeça ao mesmo tempo em que exprime frustração na face; um cliente com raiva associa um tom de voz agressivo a uma expressão facial correspondente ao reclamar seus direitos.

Dado o aspecto significativo da expressão global afetiva, trabalhos recentes têm observado as relações entre as diversas componentes que alimentam o processo emocional. Nesse campo, [Scherer e Ellgring, 2007] realizam investigações profundas sobre o tema, aplicando metodologias para a busca de padrões de englobamento em características vocais, faciais e gestuais em uma dada expressão facial. Além de Scherer, [Campanella e Belin, 2007] realizam uma discussão dos estudos cognitivos, suportando a correlação entre voz e expressões faciais na demonstração de emoções, como ilustra um dos experimentos realizados pelos autores, apresentado na Figura 2.6.

Nesse experimento, indivíduos são convidados a apertar botões “triste”/“feliz” presentes em uma bancada ao verem um vídeo e escutarem o som associado. O gráfico resultante dos experimentos mostra a frequência relativa de vídeos classificados como “tristes” quando a vocalização associada era “triste”, “ausente” ou “feliz”. Verifica-se claramente a influência da alteração no modo vocal sobre a classificação.

Figura 2.6. Experimento para identificação da relação de modos emocionais. O gráfico ilustra a frequência relativa de respostas “triste” em função do contínuo de faces, quando a vocalização associada era “triste”, “neutra” ou “feliz”. Fonte: [Campanella e Belin, 2007].

Observa-se em trabalhos a constatação de que o aparato biológico atua globalmente tanto na geração quanto na detecção da emoção pelos diversos subsistemas. Assim, após o processo de avaliação emocional, os processos fisiológicos são desencadeados de forma paralela, ainda que através de mecanismos não completamente esclarecidos. Da mesma forma, o reconhecimento de uma emoção de um terceiro pelo ser humano deve ser realizado em conjunto pelos subsistemas audiovisual e tátil e não apenas a partir da conclusão dos dados extraídos de cada experiência sensorial. A complexidade aumenta com o aparecimento de indícios de que a importância de cada modo de detecção deve ser diferente para diferentes emoções - em [De Silva et al, 1997], um segundo experimento realizado com 18 indivíduos mostrou que quando submetidos separadamente às partes acústica e visual de uma emoção, os sujeitos identificavam mais facilmente algumas emoções a partir do áudio (tristeza, medo), enquanto outras a partir do vídeo (felicidade, raiva).

Uma particularidade relevante corresponde ao estudo dos intervalos e atrasos na expressão de uma emoção pelos diferentes mecanismos (face, vocalização, etc.). Enquanto expressões faciais podem perdurar apenas por instantes – as chamadas micro

expressões – assinaturas vocais podem durar por diversas palavras. Da mesma forma, a emoção presente no contexto do que é dito pode perdurar por diversos minutos ou pelo discurso inteiro. Tem-se assim, um sistema em que a abordagem de fusão deve lidar com escalar temporais distintas, aguardando a resposta de sensores em diferentes frequências de amostragem.

Técnicas Computacionais

O problema genérico de fusão de informações com busca de melhores resultados do que uma fonte única é abordado há séculos, com registros de estudos de modelos de política econômica datando de 1786 [Rao, 2004].

Hoje, a fusão de dados de sensores em um nível abrangente é frequentemente estudo da engenharia de sinais, com objetivo de fornecer a um sistema de controle qualquer informações mais confiáveis ou completas do que as captadas por um sensor operando individualmente [Hall, 1992]. Para que isso seja possível, técnicas de fusão de dados dependem não apenas de sensores que forneçam representações do mundo sob diferentes pontos de vista, mas também de um algoritmo robusto que selecione as informações mais relevantes de cada entrada de forma a melhorar essa representação.

Em alto nível, a literatura apresenta diversas arquiteturas para a implantação de sistemas de fusão sensorial², particularmente em tempo real, tais como o Waterfall Process Model [Markin et al, 1997 (apud Esteban et al, 2005)], a Arquitetura LAAS [Alami et al, 1998] e o modelo Omnibus [Bedworth & O'Brien, 1999] (Figura 2.7).

² A maior parte dessas arquiteturas, as quais se propõem a formalizar as etapas de captura e tratamento dos dados, surgem historicamente das demandas em robótica para controle de motores, planejamento de trajeto e visão de máquina.

Figura 2.7 Modelo Omnibus: um exemplo de arquitetura para fusão de sensores. Modificado de [Bedworth & O'Brien, 1999].

Contudo, o processo de análise dos dados dentro das diferentes arquiteturas é menos generalista, exigindo frequentemente abordagens particulares ao tipo de informação tratada. Para isso, sistemas de fusão de sinais lançam mão de uma larga gama de ferramentas matemáticas, tais como redes Bayesianas, Dempster–Shafer, filtros de Kalman e técnicas de classificação. Estas últimas, particularmente, apesar de frequentemente apresentarem dificuldades no uso em tempo real, permitem que sensores cujas saídas sejam enviesadas construam de forma supervisionada um sistema de fusão com base em nada além de dados de treinamento, algo bastante genérico e utilizado como base para este trabalho (Capítulo 3).

Quando se trata especificamente de detecção multimodal de emoções, verifica-se que os trabalhos relacionados a sistematizações computacionais são diretamente dependentes da evolução dos sistemas de detecção unimodais. As pesquisas encontram-se frequentemente em sua infância e, ainda que trabalhos documentados a mais de uma década possam ser localizados, os resultados apresentam-se geralmente poucos e esparsos.

[Pantic & Rothkrantz, 2003] oferece um dos mais referenciados trabalhos no tema, apresentando justificativas de sua importância e recomendações para as pesquisas na área. Em seu texto são levantados pontos conceituais sobre em que nível a fusão das entradas deve ocorrer, assunto retomado em [Chetty & Wagner, 2008]. Pantic argumenta que a fusão dos dados emocionais pode ocorrer apenas em um nível de características (*feature-level*) ou de tomada de decisão (*decision-level*).

Em um sistema orientado em nível de características, os diversos traços obtidos pelos sensores podem ser fundidos conjuntamente, em busca de padrões globais. Na abordagem de tomada de decisão, por sua vez, cada subsistema realiza a detecção individual, para que depois as conclusões parciais sejam fundidas (Figura 2.8). Pantic comenta que a abordagem em nível de decisão é normalmente mais utilizada em computação afetiva por motivos diversos, dentre eles a dificuldade de definição de métricas entre os diferentes sensores, as questões temporais de janelamento e os resultados de pesquisas aparentemente melhores. Contudo, salienta que ainda que a abordagem faça sentido computacionalmente, não representa a maneira de processar do cérebro biológico.

Figura 2.8 Fusão de sensores ao nível de *features* (a) e decisão (b). Inspirado em [Pantic & Rothkrantz, 2003]

Outro trabalho relevante foi proposto por [De Silva et al, 1997], quando averiguou as diferentes importâncias das modalidades sonoras e visuais na detecção de emoções. Sua proposta era utilizar os experimentos realizados com indivíduos na construção de uma matriz de pesos para realização de média ponderada das entradas de áudio e vídeo em cada uma das emoções. Alguns anos depois, trabalhos aplicados começam a surgir, tais como os apresentados em [Chen & Huang, 2000] e [De Silva & Ng, 2000] os quais realizam a fusão de dados da face e da voz no processamento, com o segundo obtendo bons resultados a partir do uso de redes Bayesianas na avaliação conjunta.

Na fusão de maior número de sensores, trabalhos apresentados em [Castellano et al., 2007] apresentam a fusão de dados de voz, face e expressões corporais, apresentando mais de dez por cento de aprimoramento em comparação com as abordagens unimodais.

2.7 Classificadores

Para que seja possível a análise conjunta das diversas entradas sensoriais tratadas neste trabalho, o processo de fusão de informações sobre emoções exige a utilização de ferramentas matemáticas que possibilitem a avaliação da confiabilidade de cada entrada e a decisão de qual expressão emocional o usuário apresenta em um dado momento. Dentre diversas abordagens conhecidas em fusão, mostra-se bastante popular um grupo de técnicas definido como “algoritmos de classificação”.

Tais algoritmos permitem abordar o problema genérico da separação de informações variadas em subconjuntos com propriedades semelhantes, a partir de regras definidas durante o processo de treinamento do classificador (aprendizagem supervisionada)³. No treinamento, um subconjunto com classificações conhecidas é alimentado ao algoritmo, de forma a servir como referência para a criação de regras para dados futuros. A grande qualidade dessa abordagem encontra-se na capacidade de realizar o processo de separação sem qualquer preconceção ou conhecimento em relação ao sistema modelado. Dessa forma, a modelagem do problema é realizada “*inline*”, durante o próprio processo de treinamento do classificador.

Dada a relevância de tais ferramentas no presente trabalho, nos itens 2.7.1 a 2.7.3 a seguir, os três tipos de classificadores utilizados nesta pesquisa são listados e detalhados.

2.7.1 Redes Neurais

Com sua teoria matemática fundamentada em 1943 através de trabalho por [McCulloch & Pitts, 1943], as redes neurais artificiais são inspiradas pelo comportamento de sua contraparte biológica presente nos sistemas nervosos animais, sobre a qual a ciência havia feito grandes contribuições nas décadas anteriores. A ferramenta surge a partir da hipótese de que uma simplificação matemática adequada dos elementos neurais pudesse manter a essência do potencial de processamento do cérebro humano, realizando tarefas genéricas.

³ Utiliza-se a nomenclatura frequentemente aplicada em aprendizado de máquina, na qual o problema de classificação é associado sempre a aprendizado supervisionado (aprendizado não supervisionado, o qual não faz parte do escopo desta pesquisa, é associado ao problema de “*clustering*”).

Nas redes neurais mais comumente utilizadas, cada nó (ou “neurônio”) é uma função de transferência parcial a qual recebe múltiplos sinais de entrada, realiza sua soma ponderada e aplica uma função de saturação à saída (para que os resultados das sucessivas camadas de neurônios não “explodam” em tamanho) [Rojas, 1996]. O processo de construção do neurônio consiste assim de aprendizagem da definição dos pesos W_{ki} que multiplicam as entradas. A Figura 2.9 ilustra o modelo de nó.

Figura 2.9. Modelo de Neurônio Artificial.

A partir desse modelo base, os neurônios organizam-se em camadas, de modo que cada saída em um elemento possa estar conectada a uma das entradas de outro, configurando assim na sua aplicação mais comum um grafo acíclico. É a partir dessas diversas camadas com número variável de neurônios que se formam estruturas que consideram não-linearidades do sistema modelado e permitem que entradas genéricas e ricas em informação sejam levadas a uma ou mais saídas desejadas.

Entretanto, ainda que fundamentadas em tais conceitos simples, as redes neurais foram pouco exploradas por várias décadas a partir de sua concepção, sobretudo devido ao problema intrínseco de obtenção dos pesos W_{ki} necessários para cada utilização do algoritmo. A retomada em sua aplicação ocorreu a partir da década de 1980, ressaltando-se seu uso em detecção de padrões para identificação de objetos e rostos em fotos e na fusão de sensores com qualidades distintas. Essa retomada em estudo foi resultado, sobretudo, da popularização de um método eficiente de treinamento de rede - conhecido como *backpropagation* – sugerido por Rumelhart [Rumelhart et al, 1986].

No *backpropagation*, utilizado especialmente em redes do tipo *feed-forward*, os pesos dos neurônios são iniciados de forma aleatória, com valores usualmente entre 0 e

1. Em seguida, cada elemento do conjunto de treinamento é alimentado à rede primeiramente a partir de sua entrada e, em seguida, retornando do alvo conhecido para a entrada. A diferença entre o obtido na passagem de ida e o esperado é utilizado como métrica de correção dos pesos para novos valores, em um processo iterativo, o qual é encerrado de acordo com algum critério de parada. O processo resultante é bastante eficiente, apesar de intensivo computacionalmente.

O problema do treinamento levou à proposta de várias outras modalidades de redes neurais, resultando, por exemplo, nas redes neurais probabilísticas, ou PNNs (*Probabilistic Neural Networks*), eficientes em processos de classificação. Em uma PNN há sempre três camadas. Os dados de entrada são inicialmente comparados ao conjunto de treinamento fornecido previamente e a distância entre eles é calculada. Essa métrica de proximidade é então utilizada para a geração de um vetor de probabilidades que toma em conta a relevância de cada classe de resultados. Por fim, uma função competitiva escolhe a classificação com maior probabilidade. A partir dessa abordagem, o treinamento realizado é consideravelmente mais rápido do que o exigido através de algoritmos como *backpropagation*. A Figura 2.10 ilustra o processo de classificação descrito.

Figura 2.10. Estrutura de Rede Neural Probabilística. Adaptado de [MathWorks01].

Este trabalho utiliza as duas diferentes abordagens de aprendizado descritas: BPNNs (*Backpropagation Neural Networks*) e PNNs.

2.7.2 Naive Bayes

O Teorema de Bayes tem papel fundamental no desenvolvimento da estatística e probabilidade, descrevendo como uma nova evidência (B) altera o grau de crença em um acontecimento (A).

$$P(A/B) = \frac{P(B/A).P(A)}{P(B)} \quad (2.1)$$

A Equação (2.1) ilustra essa relação simples, através da qual a probabilidade de A acontecer $P(A)$ é reavaliada pelo conhecimento da probabilidade isolada do evento B acontecer.

A técnica de classificação conhecida como *Naive Bayes* suporta-se da aplicação do Teorema de Bayes para determinar a probabilidade de uma amostra de testes pertencer a uma classe, utilizando a base de treino como informação para a probabilidade *a priori* (evidência) [Russel & Norvig, 2003].

Seja o processo de classificação de um elemento I dependente de ocorrências $E_1...E_n$, através do Teorema de Bayes dado por:

$$P(I / E_1...E_n) = \frac{P(E_1...E_n / I).P(I)}{P(E_1...E_n)}. \quad (2.2)$$

O termo numerador à direita da equação refere-se ao modelo de probabilidade conjunta o qual, considerando hipótese de independência nas variáveis (de onde provem o termo “Naive”), pode ser reescrito como:

$$P(E_1...E_n / I).P(I) = P(I, E_1, ..., E_n) = P(I) \cdot \prod_{i=1}^n P(E_i / I). \quad (2.3)$$

Desta forma, a equação fundamental de estimação toma a forma de:

$$P(I / E_1...E_n) = \frac{1}{kb} P(I) \cdot \prod_{i=1}^n P(E_i / I), \quad (2.4)$$

onde kb é o fator $P(E_1...E_n)$, constante referente às evidências. As probabilidades necessárias para o cálculo do classificador são então obtidas através das frequências do banco de treinamento e escolha de distribuições.

Por fim, o classificador estruturado utiliza a hipótese resultante mais provável (ou alguma outra regra de decisão) para separação dos elementos da base de testes.

Apesar da hipótese de independência ser necessária para a formulação do algoritmo, o *Naive Bayes* opera de forma adequada em diversas aplicações nas quais tal hipótese não pode ser verificada, sendo seu sucesso em parte consequência da versatilidade. À medida que a hipótese de independência perde força em um dado modelo, o ajuste a distribuição assumida piora. Entretanto, se ambas as distribuições estimada e real concordarem na classe mais provável, o classificador ainda terá bom comportamento [Rish, 2001].

2.7.3 Support Vector Machine

As *Support Vector Machines* (Máquinas de Vetor de Suporte ou simplesmente SVMs) foram inicialmente introduzidas por Boser, Guyon e Vapnik [Boser et al, 1992] e tornaram-se posteriormente populares com aplicações computacionais bem sucedidas em identificação de caligrafia, categorização de textos, detecção de padrões em som e imagem [Burgess, 1998].

Em um classificador do tipo SVM, o problema de separação consiste em encontrar um hiperplano que separe dois conjuntos de elementos, posto que esse hiperplano possua a maior margem possível o distanciando dos conjuntos a serem classificados, como ilustra a Figura 2.11:

Figura 2.11. Grupos separados por SVM. Os elementos sombreados são os vetores de suporte.

Assim, seja um hiperplano do tipo:

$$W^T X + u = 0. \quad (2.5)$$

Considerando que todos os dados de treinamento estejam pelo menos a uma unidade de distância do hiperplano, a margem em torno do hiperplano será dada por:

$$\rho = 2d = 2 \frac{W^T X + u}{\|W\|} = \frac{2}{\|W\|}, \quad (2.6)$$

onde d é a distância do hiperplano até os vetores de suporte. Dessa forma, o classificador torna-se um problema de otimização, na qual a margem deve ser maximizada (ou $\|W\|$, fator denominador, ser minimizado).

Outra característica relevante das SVM é a capacidade de classificação de conjuntos de dados com separações complexas, como no caso da Figura 2.12 (a). Para que isso seja realizado, o mapeamento do problema para um espaço diferente permite que um hiperplano realize a separação de classes (b). Para que a transformação seja simplificada e computacionalmente menos custosa, a introdução do que é conhecido

como “kernel trick” [Hofmann, 2006] é realizada, deixando a cargo do usuário a escolha de algumas funções kernel de transformação.

Figura 2.12. O fundamento por trás das função kernel. A transformação para um espaço de dimensão maior pode tornar o problema original (a) separável por um hiperplano (b) .

Postas as considerações anteriores, o classificador SVM apresenta uma limitação intrínseca a sua fundamentação – a segmentação binária. Não é possível, através da metodologia usual, realizar a classificação de mais do que dois tipos de elementos em um conjunto heterogêneo. Posto que este trabalho trata da seleção de diversos aspectos afetivos, a solução encontrada apresenta-se no cascadeamento de processos de classificações sucessivas.

Nessa abordagem, conhecida como “SVM multiclasse”, os elementos a serem classificados podem ser, a cada iteração da SVM, removidos um a um do grupo original, como ilustra a Figura 2.13. Nesse processo de confronto do tipo “um contra todos”, a SVM seleciona um atributo relevante a cada iteração, fornecendo ao final, um resultado qualitativamente semelhante a outros classificadores multiclasse.

Figura 2.13. Processo de classificação SVM multiclasse tipo “um contra todos”.

3 Abordagem Proposta

Neste capítulo, apresentam-se as abordagens propostas para a solução do problema de fusão e de sua entrada emocional contextual, na forma da ferramenta emoCrwaler. Após uma breve discussão introdutória na seção 3.1, os dois tópicos centrais de desenvolvimento são tratados separadamente, através de etapas de motivação, delimitação do problema e escolha de solução.

3.1 Visão Geral

Este trabalho lida com a questão da coleta de dados emocionais por diferentes sensores e seu processamento para detecção confiável da relação entre um usuário e a máquina operada.

Para que haja riqueza na incorporação de dados, busca-se aquisição de informações de três fontes básicas de detecção: voz, face e semântica. Os dados coletados são então filtrados, caso necessário, e alimentados a um sistema de fusão, o qual deve ser capaz de reduzir imperfeições de medição de cada uma das entradas, através da análise conjunta das informações coletadas. Após o processo de fusão, uma etapa de pós-processamento é realizada, para decisão final sobre a expressão emocional detectada.

Cada entrada emocional, ou “sensor”, utilizada possui suas próprias características de confiabilidade ante diferentes emoções, de forma que a análise combinada de suas observações através de algoritmos de fusão sensorial deve poder obter melhores resultados do que a abordagem unimodal.

A Figura 3.1 apresenta o fluxo de informações no sistema proposto. Os quadros hachurados, indicam blocos desenvolvidos no escopo deste trabalho, enquanto os em cinza apresentam processos de filtragem e ferramentas de *hardware* ou *software* por terceiros (as ferramentas aplicadas no processamento unimodal são descritas na seção 4.1).

Figura 3.1. Área de pesquisa. A etapa de pré-processamento e os itens hachurados foram desenvolvidos no escopo do trabalho.

Como cada módulo unimodal tem autonomia e é aplicado independentemente, tem-se, portanto, um sistema global com foco em nível de tomada de decisão, assim como visto na seção 2.6. Tal abordagem ignora os processos internos de cada módulo, mas reduz a susceptibilidade do fusor a ruídos e põe foco nos pontos fracos do sistema de sensoriamento.

Posto que trabalhos complementares na área ocorrem na Escola Politécnica da Universidade de São Paulo durante o desenvolvimento desta dissertação, uma série de discussões com outros pesquisadores sobre interoperabilidade foi realizada para padronização de entradas e saídas, escalas e mensagens. É no contexto dessas restrições requeridas pelo esforço conjunto que o desenvolvimento dos módulos ocorre. Dessa forma, todos os subsistemas construídos para este projeto têm como meta poderem ser reutilizados em pesquisas subsequentes.

3.2 Fusão por Classificação

Sejam Sk ($k=1,...,n$) subsistemas de aquisição de expressões emocionais (sensores) em um sistema de detecção afetiva e Er ($r=1,...,m$) as diversas expressões detectadas discretamente no conjunto Sk .

A probabilidade Bayesiana Pkr de um sensor Sk determinar corretamente uma expressão emocional Er será descrita por:

$$Pkr = P(Sk = Er | Er). \quad (3.1)$$

Sabe-se que para cada Sk ter-se-á um valor de Pkr para cada emoção descrita em um dado conjunto, fruto de confiabilidades e vieses diversos dos sistemas (um sistema pode ter melhor frequência de acertos em certo tipo de emoção, ou ter tendência de confusão em relação à detecção de algum par de expressões).

Espera-se que, sendo observadas as diferentes fontes emocionais, seja possível realizar a aplicação de uma função de transferência f (estática ou adaptativa) de tal forma que:

$$\Pr(f(P11, P12, \dots, P21, \dots, Pnm) = Er \mid Er) > \max(Pkr), \quad (3.2)$$

onde Pr é a probabilidade de acerto de uma certa expressão dada pela fusão dos diversos sensores pela função de transferência.

A função de transferência f (fusor) pode ser obtida através da observação dos resultados dos sensores de forma independente sobre um banco de treinamento e da identificação da real emoção por um especialista. A alternativa considerada para a solução deste problema no presente trabalho consiste da aplicação de algoritmos de classificação de padrões para a realização da fusão. Nessa abordagem, o sistema em questão deverá receber como entrada uma matriz de valores correspondentes à descrição quantitativa das expressões emocionais concluídas pelo conjunto de sensores, processá-la e identificar padrões que surjam com maior frequência para cada emoção nos sensores. Através do treinamento supervisionado, espera-se por fim que o reconhecimento desses padrões leve à identificação da emoção com maior precisão do que com cada sensor individualmente.

A escolha de técnicas de classificação traz pontos positivos relacionados à facilidade de modelagem das imperfeições dos diversos sensores – pouco sobre as falhas em detecção precisa ser conhecido. Contudo, há o sacrifício da observação do comportamento temporal da interação usuário-máquina, posto que todos os processos de classificação considerados são estáticos. Para contornar tal problema, as aplicações foram apoiadas na avaliação de trechos curtos de interação, nos quais há a prevalência de uma única emoção. Desta forma, os dados dos sensores são tratados unicamente por sua média temporal para serem em seguida alimentados ao sistema. Maiores considerações sobre esse aspecto da solução são realizadas no Capítulo 4.

Técnicas de Fusão Aplicadas

Em um primeiro momento, escolheu-se a realização da fusão dos sensores através de redes neurais, abordagem adequada para situações em que as fontes possuem ruídos diversos e nas quais o grau de confiabilidade relativa entre elas apresenta-se desconhecido.

As redes neurais são aproximadores universais com grande capacidade de generalização. Elas permitem que, através de aprendizado (neste caso supervisionado), seja possível avaliar as diversas relações e vícios entre os sistemas de detecção, de forma a encontrar uma confiabilidade conjunta do sistema ante cada tipo de expressão emocional. As redes possuem também grande flexibilidade em relação às características dos sinais de entrada fornecidos pelos diversos sensores, não sendo necessários grandes tratamentos prévios dos dados alimentados ao sistema além de alguma normalização para evitar saturação dos neurônios.

Dentre as diversas abordagens de treinamento disponíveis para redes neurais, escolheu-se uma alternativa acíclica clássica – a *Feedforward Backpropagation Network*, ou FFBN – e uma alternativa probabilística – a *Probabilistic Neural Network*, ou PNN. A rede FFBN é solução comum em problemas de classificação; no escopo deste trabalho possui uma camada oculta de nós, de forma a absorver não linearidades existentes e atua com diferentes funções de transferência nas camadas. A rede probabilística, por sua vez, apresenta-se como solução alternativa relevante, dado seu tempo de treinamento ordens de grandeza mais rápido que a FFBN. Entretanto, seu ajuste é por vezes de difícil realização em conjuntos de dados esparsos.

Figura 3.2. FFBPN utilizada para fusão de expressões emocionais (12, 7 e 4 nós respectivamente em cada camada).

A Figura 3.2 ilustra a rede neural construída para experimentos realizados com *Feedforward Backpropagation* (ver Capítulo 4 - Experimentos e Resultados), correspondendo cada círculo a um neurônio distinto. Observa-se que os dados dos diferentes sensores são mesclados na camada intermediária, a qual busca levar em conta quaisquer não-linearidades possíveis na detecção, sendo que sua quantidade de neurônios foi definida a partir de testes com diferentes topologias. Nota-se também que a rede possuirá tantas saídas quantas emoções forem observadas. Esse processo gera um vetor que indica o grau de significância de cada expressão e pode ser avaliado através de uma função de exclusão competitiva dentre as saídas de diferentes valores.

Após o estudo das técnicas neurais, lançou-se mão do uso de um classificador baseado no teorema de Bayes conhecido como “*Naive Bayes*” em abordagem alternativa ao problema de fusão multimodal. Tal família de classificadores possui poucos parâmetros de ajuste e abordagens de aprendizado, sendo de aplicação direta e simples. A configuração do algoritmo precisou se focar apenas na escolha da distribuição de modelagem, neste caso do tipo *kernel*. Além disso, observou-se que as fortes suposições

de independência de variáveis exigidas pelo *Naive Bayes* parecem não interferir com o processo de detecção, de forma que não houve retrabalho na formulação do problema.

Por fim, uma terceira técnica de classificação foi avaliada, na forma da *Support Vector Machine* (SVM) para classes múltiplas. A escolha do SVM foi motivada pela sua propriedade de apresentar bom comportamento mesmo ante conjuntos de treinamento de tamanho limitado, como é o caso encontrado neste trabalho. Como o classificador SVM é por natureza um seletor binário (isto é, permite a separação de elementos em apenas duas classes), uma implementação multiclasse foi necessária, de forma que todas as emoções fossem devidamente catalogadas. A função *kernel* utilizada para esse classificador (elaborada na seção 2.7.3) foi do tipo linear, isto é, não aplicou-se qualquer transformação dimensional ao conjunto de dados.

Todos os processos de classificação em fusão realizados neste trabalho foram construídos com base em bibliotecas padrão Matlab 2010 [MathWorks02], sendo que os trabalhos em SVM implicaram a utilização do método aberto multiclasse MULTISVM [Mishra, A., 2011].

3.3 emoCrawler: Detecção Semântica de Emoções

Todos os dias, milhões de pessoas geram conteúdo opinativo na Internet. *Blogs*, *microblogs*, fóruns, redes sociais⁴ – todos são carregados por uma quantidade massiva de carga emocional sobre os mais diversos assuntos.

É possível admitir que essa massa enorme de dados, em sua média, corresponda ao senso comum dos autores sobre um determinado aspecto, o que fornece uma base de dados extremamente rica e acessível sobre reações afetivas.

A ferramenta emoCrawler – um *software* desenvolvido na linguagem de programação C# no contexto deste trabalho - utiliza-se dessa premissa para a detecção de emoções em palavras e expressões. Seu funcionamento é descrito a seguir.

⁴ Estudos são escassos, mas trabalhos indicam que, somente via blogs, cerca de um milhão e meio de peças de conteúdo são criadas a cada dia na Internet [Goulding et al, 2012].

3.3.1 Motivação

Suponha-se que o usuário do sistema de detecção de emoções expresse a seguinte frase:

“Estou feliz, viajarei nas férias.”
[v] [adj] [v] [p] [s]

Há um indicativo objetivo de estado emocional do interlocutor, dado pela primeira oração do texto – “estou feliz”.

Considere-se agora um segundo exemplo:

“Viajarei nas férias.”
[v] [p] [s]

É impossível definir com certeza o estado emocional do indivíduo neste caso, posto que não há qualificadores objetivos de estado emocional. Contudo, é possível realizar a hipótese razoável de que a pessoa deva estar contente ao proclamar tal afirmação: é senso comum que férias são um momento de relaxamento desejado e a possibilidade de realizar uma viagem durante esse momento de descanso poderá trazer gratificação adicional.

Não é possível descartar que o indivíduo esteja infeliz (poderia estar sendo motivado a viajar contra seu interesse durante seu período de férias, por exemplo), contudo, na média, é razoável acreditar que esteja satisfeito com sua viagem.

Para avaliar tal hipótese, um procedimento objetivo seria observar com que frequência expressões de felicidade surgem quando as palavras “viajarei” e “férias” são buscadas em uma base de dados de textos opinativos. É a partir desta ideia que o método utilizado pelo emoCrawler é elaborado.

3.3.2 A Problemática da Semântica e o Uso do Twitter

Ainda que textos opinativos publicados na Internet frequentemente possuam um único tema central, normalmente identificado no título, primeiro ou último parágrafos [Li & Ren, 2008], a extração do tema e a identificação das emoções atrelada podem ser tarefas complexas. A estrutura sintática das frases individuais pode variar, assuntos podem ser retomados e a organização das ideias pode ser não linear, de forma que a

ausência de um motor de processamento altamente complexo e robusto implicaria a identificação incorreta de emoções.

A alternativa proposta para que tal problema seja evitado está no processamento apenas de textos curtos de *microblogs*, tais como o Twitter [twitter01]. Postagens de Twitter possuem uma abordagem concisa e altamente opinativa, de forma que mensagens frequentemente demonstram elementos de emoção, ao mesmo tempo em que não contêm construções complexas, dado a limitação de 140 caracteres. Sugere-se, portanto, que a análise de construções simples em grande volume eliminarão a necessidade de análise sintática complexa.

Dá-se também preferência às consultas em Inglês, o que evita questões ligadas à flexão de verbos, substantivos e adjetivos em Português.

3.3.3 Arquitetura

Dado um discurso curto apresentado pelo usuário, o emoCrawler inicialmente realiza a extração de substantivos, adjetivos e verbos – palavras consideradas relevantes para a busca de emoções.

Em seguida, tais elementos são buscados isoladamente em mensagens da rede social avaliada, retornando um volume considerável de textos que os contêm. Os textos retornados são então processados, formatados e filtrados, sendo descartadas do banco orações complexas que possam influir negativamente na detecção, tais como as que apresentam termos de negação (-como “não”, “nunca”, entre outros).

A partir da base de dados gerada, busca-se em cada texto por expressões literais de emoção, definidas pelo dicionário descrito na seção 3.3.4. A análise da frequência relativa dessas expressões para cada uma das palavras relevantes do discurso é combinada, de forma que um perfil de emoção do usuário é criado.

A Figura 3.3 ilustra o fluxo de informações no emoCrawler:

Figura 3.3. Fluxo de Informações no emoCrawler. O processo de estimativa corresponde à média dos valores gerados para diferentes expressões.

A escolha de uma base de mensagens adequada para busca de emoções impõe restrições de compromisso em relação a seu tamanho. Se por um lado a busca de grande número de textos aumenta a generalidade da conclusão gerada pelo algoritmo, por outro cria dificuldades de *download* e extração local de emoções, tornando o processo lento, distanciando-se da aplicação prática.

Como experimento para avaliação da questão, escolheram-se algumas palavras consideradas neutras – isto é, que devam oscilar pouco em seu conteúdo emocional com o passar do tempo – e avaliou-se a saída emocional não tratada gerada por elas a partir do emoCrawler para diferentes números de amostras (entre aproximadamente 100 a 1500). Os resultados desse ensaio para três diferentes palavras chave podem ser observados na Figura 3.4.

Figura 3.4. Estabilidade do emoCrawler em função do número de amostras (N) – resultados sem normalização dos pesos.

Da figura, observa-se que a partir de 500 amostras o sistema começa a apresentar menor variação, dirigindo-se ao senso comum sobre determinado assunto. É a partir de tais observações que se escolhe tal quantidade de amostras para realização dos experimentos do Capítulo 4.

A Figura 3.4 também apresenta uma observação de relevância sobre o motor de conclusões do emoCrawler: emoções em texto não podem ser definidas de forma absoluta sobre sua medição normalizada (como apresentado no eixo das ordenadas dos gráficos da figura). O que se observa, na realidade, é que expressões de felicidade, por exemplo, são sempre abundantes nas mensagens postadas pelos usuários, ainda que em assuntos considerados frequentemente tristes. Ainda que o motivo psicológico de tal

comportamento não seja escopo deste texto, é fundamental extrair a conclusão de que a tomada de decisão sobre qual emoção o usuário está exprimindo a partir dos resultados fornecidos pelo emoCrawler exige o treinamento de um classificador, corrigindo o peso concedido a cada tipo de emoção extraída. Esta questão é explorada com maior profundidade na seção 4.3.5.

3.3.4 O Dicionário Emocional

A verificação da existência de expressões de emoção nos textos exige a construção de um dicionário de emoções, o qual apresente uma listagem abrangente de palavras consideradas ligadas à determinada expressão emocional. Para sua construção, tomou-se como referência inicial as propostas de Goleman (apud [Martinez-Miranda & Aldea, 2005]) e [Laros, 2005], os quais realizaram trabalhos de classificação afetiva. Em seguida, termos adicionais selecionados ao longo do projeto de pesquisa foram acrescentados.

Além da correspondência emocional, propôs-se a criação de um peso (pE), qualificador da intensidade e relevância da expressão na designação de uma dada emoção, o qual pode ser tratado como parâmetro de modulação dos elementos. A Tabela 3.1 lista as palavras presentes no dicionário:

Tabela 3.1. Listagem de expressões do dicionário emocional. Diversas interjeições e *emoticons* estão presentes no dicionário, mas foram omitidos nesta listagem.

Emoção	Expressões
Felicidade	"joy", "joyful", "enjoy", "enjoyed", "happiness", "happy", "glad", "excellent", "wonderful", "awesome", "nice", "great", "relieve", "relieved", "excite", "exciting", "excited", "euphoria", "euphoric", "euphorically", "emotion", "emotional", "thrill", "thrilled", "contentment", "contented", "pleasure", "gratification", "ecstasy", "bliss", "delight", "delighted", "delighting", "satisfaction", "satisfied", "satisfying"
Tristeza	"sadness", "sad", "pity", "pitied", "pitying", "unhappy", "unhappiness", "melancholia", "melancholy", "melancholic", "depression", "depressed", "lament", "lamented", "lamenting", "regret", "regretted", "regretting", "grief", "grieved", "sorrow", "sorrowed", "loneliness", "despair", "despaired", "hopeless", "hopelessness", "hopelessly"
Raiva	"anger", "angered", "angering", "angers", "mad", "disappoint", "disappointed", "disappointing", "disappoints", "disappointingly", "annoyance", "annoyed", "annoying", "annoys", "fury", "furious", "furiously", "indignation",

	<i>"hatred", "hate", "hated", "hating", "hates", "outrage", "outraged", "outraging", "outrages", "rage", "wrath", "irritation", "irritated", "irritating", "irritates"</i>
Desagrado	<i>"disgust", "disgusted", "disgusting", "disgusts", "contempt", "aversion", "disdain", "scorn", "scorned", "scorning", "scorns", "distaste", "distasted", "distasting", "distastes", "repulse", "repulsed", "repulsing", "repulses", "repugnance", "repugnant", "revulsion", "revulsive", "despise", "despised", "despising", "despises"</i>
Medo	<i>"fear", "feared", "fearing", "fears", "fearful", "fearfully", "fearfulness", "afraid", "scare", "scared", "scaring", "scares", "fright", "frightened", "frightening", "frights", "frighten", "terror", "terrify", "terrified", "terrifying", "terrifies", "terrorize", "anxiety", "anxieties", "anxious", "apprehension", "apprehensive", "apprehensively", "apprehensiveness", "phobia", "panic", "panicked", "panicking", "panics", "dread", "dreaded", "dreading", "dreads"</i>

3.3.5 O Contexto Temporal

O banco de dados de consulta nas redes sociais é um objeto em constante atualização. A cada dia, as avaliações afetivas atreladas a um determinado conceito são reavaliadas de acordo com acontecimentos sociais em andamento. Desta forma, a emoção que certa palavra suscita na sociedade muda todo tempo, função do contexto temporal.

Para observação dessa estabilidade temporal, realizou-se durante abril de 2012 um pequeno experimento, avaliando a variação do julgamento de algumas palavras-chave após intervalos de tempo no Twitter. Nesse sentido, a Figura 3.5 ilustra, para algumas palavras-chave, sua estabilidade com relação à dinâmica afetiva no tempo.

Figura 3.5. Avaliação emocional de diferentes palavras-chave em função do tempo decorrido (em minutos) – resultados sem normalização por pesos.

Nas palavras testadas, observa-se que oscilações nas avaliações foram constatadas, ainda que sem grandes alterações. Contudo, buscas realizadas a partir de tópicos em grande exposição na mídia, por exemplo, poderiam ilustrar mudanças mais agudas. Um potencial exemplo disso pode ser visto em $t = 1000 \text{ min.}$ para “travel”, onde a diminuição da frequência de termos associados à felicidade é acompanhada pelo aumento de medo, raiva e surpresa. Um acidente aéreo ou rodoviário largamente anunciado, por exemplo, poderia desencadear esse tipo de reação.

Ainda que a questão temporal esteja fixada no instante da pesquisa para os testes realizados neste trabalho, é objetivo no desenvolvimento do emoCrawler a

caracterização do escopo histórico das buscas, de forma a permitir que o sistema tenha acesso a buscas de longo e curto prazos.

3.3.6 A Questão do Janelamento

Uma asserção com conteúdo hermético, contido, pode exigir um conjunto variável de sentenças para ser expressa. Sem análise semântica, é impreciso realizar o encapsulamento dos elementos semânticos completos que devem ser processados em conjunto para que a extração emocional funcione da mesma forma que em um ser humano. Em outras palavras, é complexo decidir em que momento o usuário está deslocando-se para um novo estado emocional.

Desta forma, o algoritmo do emoCrawler deve ser capaz de se apoiar em pistas não semânticas para definir as janelas de observação da conversa para a busca de emoções. Como os trechos audiovisuais usados nos experimentos deste trabalho expõem uma única frase e uma única emoção por vídeo, em sua atual versão, o algoritmo proposto apenas avalia a média dos valores obtidos por cada palavra relevante:

$$EstFrase = \frac{\sum_k EstPalavra(k)}{Norm(EstPalavra(1), EstPalavra(2), \dots, EstPalavra(k))}, \quad (3.3)$$

onde $EstFrase$ corresponde a estimativa conjunta desejada, $EstPalavra(i)$ corresponde ao vetor de detecção de expressões da palavra k da frase pelo emoCrawler:

$$EstPalavra(i) = (Felicidade, Raiva, Nojo, Medo, Tristeza), \quad (3.4)$$

e $Norm(x)$ é alguma função de normalização de soma para evitar crescimento descontrolado do resultado.

Contudo, apesar dessa abordagem ser suficiente na pesquisa em questão, visto a estimativa para a frase ser equivalente a estimativa para o vídeo por completo, versões posteriores deverão considerar as pausas e a evolução da transição entre emoções para a

escolha de um janelamento dinâmico na definição das palavras que devem ser consideradas conjuntamente.

4 Experimentos e Resultados

Este capítulo introduz as ferramentas produzidas por terceiros e as adaptações realizadas para aplicação neste trabalho, considerando o *corpus* aplicado. Em seguida, a metodologia de experimentação é formalizada e posta em prática para os diferentes algoritmos de classificação.

4.1 Ferramental

Descreve-se a seguir as contribuições e ferramental desenvolvidos por fontes diversas do meio acadêmico os quais foram incorporados nos diversos experimentos descritos e realizados neste trabalho.

4.1.1 Corpus

Há diversos trabalhos de registro de emoções em seres humanos para estudos acadêmicos, os quais fornecem grandes bancos de dados de emoções espontâneas ou encenadas. Contudo, tais *corpora* frequentemente apresentam-se unimodais, comumente com observação de emoções somente na face [Kanade et al, 2000], [Martinez & Benavente, 1998] ou voz [Amir et al, 2000].

Para o contexto deste trabalho, buscou-se um *corpus* com caráter especificamente multimodal que pudesse fornecer gravações audiovisuais de conversas cujas verbalizações realmente contivessem um argumento semântico. Dentre as escassas compilações disponíveis, escolheu-se o eNTERFACE'05 Audio-Visual Emotion Database [Martin et al., 2005], um banco de dados de cenas de vídeo no qual diversos indivíduos são convidados a expressar uma frase emocional da forma que mais desejarem (Figura 4.1).

Figura 4.1. Exemplos de amostras do eNTERFACE.

O eNTERFACE foi desenvolvido para o uso em tarefas de processamento de emoções e fornece uma média de 194 gravações de cada emoção selecionada, geradas por 42 indivíduos diferentes. As gravações são curtas, com apenas alguns segundos de duração, de forma que buscam transparecer um conteúdo emocional único. As frases declamadas durante a encenação encaixam-se em uma ficha de contexto previamente apresentada ao autor. Alguns exemplos encontram-se na Tabela 4.1:

Tabela 4.1 Algumas expressões semânticas contidas no banco eNTERFACE 2005. Os itens em **negrito** nas frases representam palavras processadas pelo emoCrawler.

Emoção	Contexto	Frase
Felicidade	Ganhou na loteria.	<i>"That's great, I'm rich now."</i>
Raiva	Descontentamento com um atendente bancário.	<i>"I can have you fired right now, you know?"</i>
Desagrado	Um inseto é encontrado em uma refeição.	<i>"That's horrible, I'll never eat noodles again."</i>
Medo	Um barulho desconhecido o faz acordar.	<i>"My God, there's someone in the house."</i>
Tristeza	O término de uma relação com uma pessoa querida.	<i>"I still love her."</i>

O *corpus* dessa forma apresenta-se completo para a fusão multimodal desejada.

4.1.2 Detecção de Emoções na Face: eMotion

Para a análise de expressões faciais, optou-se pela utilização de uma solução já desenvolvida e com nível de maturação adequado. Dentre as diversas soluções comerciais disponíveis, escolheu-se o aplicativo eMotion [eMotion01], produzido na Universidade de Amsterdam, Holanda.

O eMotion (Figura 4.2) utiliza técnicas de ajuste de malhas tridimensionais sobre a face, processo realizado de forma parcialmente manual para o primeiro quadro de vídeo. Em seguida, o algoritmo infere a deformação da malha durante os quadros sucessivos, extraindo dessas deformações parâmetros relacionados a expressões faciais

conhecidas. O eMotion propõe-se à detectar surpresa, felicidade, tristeza, desagrado, raiva e medo, em diferentes intensidades e composições.

Figura 4.2. Aplicativo eMotion em funcionamento.

Para os experimentos descritos neste trabalho, considerou-se como premissa a existência de apenas um contexto emocional em cada amostra de vídeo, o que é suportado pelo caráter de construtivo do banco de dados eNTERFACE. Desta forma, para alimentação das redes neurais, avaliou-se a média temporal das emoções presentes nos quadros de cada filme, considerando-se um valor normalizado para cada emoção. Ainda que melhor filtragem pudesse fornecer resultados mais satisfatórios, o experimento propôs-se intencionalmente a considerar a situação ideal de observação a fim de avaliar a eficiência não adulterada do sistema eMotion (isto é, se apenas uma emoção é expressa, ela será detectada na média).

A Figura 4.3 e a Figura 4.4 mostram a variação dos sinais detectados pelo *software* e as médias em dois vídeo-exemplos do *corpus*:

Figura 4.3. Vídeo classificado por humano como "Felicidade", mas detectado incorretamente pelo *software* eMotion. A linha laranja representa o valor médio de "Felicidade" para todo o vídeo. As ordenadas representam o nível de detecção, enquanto as abscissas indicam a progressão de quadros de vídeo (N).

Figura 4.4. Vídeo classificado por humano como "Felicidade" e detectado corretamente pelo *software* eMotion. A linha laranja representa o valor médio de "Felicidade" para todo o vídeo. As ordenadas representam o nível de detecção, enquanto as abscissas indicam a progressão de quadros de vídeo (N).

4.1.3 Detecção de Emoções na Voz: EmoVoice

Assim como no caso da detecção de emoções na face, optou-se neste trabalho pelo uso de uma solução existente e testada para a análise de expressões emocionais na voz. Para isso, escolheu-se o pacote EmoVoice [Vogt et al., 2008], desenvolvido pelo Instituto de Ciências da Computação da Universidade de Ausburgo.

O EmoVoice é disponibilizado em licença aberta e, apesar de não oferecer uma solução pronta para detecção, permite a flexibilidade do treinamento personalizado de classificadores para identificação de emoções no discurso. Dessa forma, foi possível

realizar a criação de um classificador dedicado que distinguísse todas as emoções trabalhadas neste trabalho e na língua do *corpus* utilizado (inglês).

Avaliações preliminares com diferentes algoritmos de seleção em áudio apontaram para o uso do SVM (*Support Vector Machine*) o qual foi treinado a partir de amostras do *corpus* descartadas nos experimentos posteriores. Para isso, realizou-se a extração do áudio de cada amostra de vídeo e posterior tratamento para eliminação de ruídos. Em seguida, as amostras foram concatenadas e anotadas manualmente em um fluxo sonoro de acordo com as etiquetas fornecidas pelo *corpus*. A matriz de confusão resultante do grupo de treinamento com tal algoritmo apresentou-se puramente diagonal.

Em relação ao intervalo de amostragem na detecção, o EmoVoice fornece o processamento de emoção a cada pausa longa realizada pelo locutor. No contexto do *corpus* eNTERFACE, isso corresponde a cada vídeo por completo, dado que os atores realizam a declamação de uma única frase. Assim, de forma média, a saída do aplicativo eMotion pode ser comparada com o resultado fornecido pelo EmoVoice, desde que ambas as escalas sejam normalizadas.

4.2 Experimentos

O estudo do comportamento do sistema de fusão partiu da seleção manual de vídeos do *corpus* eNTERFACE 2005 que pudessem ser prontamente classificados por observadores humanos, de forma a evitar ambiguidades emocionais fruto do caráter amador dos atores (em geral professores e alunos de ensino superior). Para o sistema de fusão, por exemplo, foram considerados apenas cerca de 5% dos vídeos do banco de dados referentes às emoções observadas. Contudo, ainda que tenha sido realizada uma seleção prévia do *corpus*, consideraram-se vídeos em condições não ideais para os diversos sistemas de classificação: áudio de baixa qualidade, iluminação não uniforme e movimentos abruptos da cabeça são alguns dos fatores que permaneceram intencionalmente no conjunto. A motivação por trás da inclusão de amostras nessas condições encontra-se em aplicar a detecção em sistemas mais próximos da realidade, apoiando-se na existência dos três mecanismos de detecção afetiva.

Os vídeos considerados adequados no processo de seleção foram subdivididos em três conjuntos isolados:

- Conjunto de treinamento SVM EmoVoice (20 amostras por emoção);
- Conjunto de treinamento do Sistema de Fusão (5 amostras por emoção);
- Conjunto de validação do Sistema de Fusão (5 amostras por emoção).

Além disso, considerou-se, para o sistema de fusão, a criação de uma amostra artificial por expressão emocional, baseada em casos previamente elaborados como ideais. Essa amostra consiste em aplicar o valor referência “1” para a entrada desejada e “0” para as outras.

Para a escolha do tipo de classificação usada, há diversos trabalhos que buscam definir um conjunto de emoções básicas, os quais poderiam ser utilizados como referência nos experimentos, como discutido no Capítulo 2. Dentre as diversas alternativas, escolheu-se trabalhar com um subconjunto das emoções de Ekman, deixando de fora “surpresa”, a qual não é considerada de fato um estado emocional [Ortony et al, 1988], mas sim um evento de transição abrupta entre diferentes estados (isto é, a surpresa pode estar atrelada a qualquer estado). A abordagem de Ekman é com frequência vista em pesquisas da área, de forma que a escolha facilita a comparação dos resultados com outros trabalhos.

Apresentadas as considerações anteriores, os experimentos foram realizados através do processamento dos diversos vídeos curtos pelos diferentes sensores. Para cada vídeo selecionado e pré-processado para remoção de ruídos, os seguintes passos foram realizados, em ordem cronológica:

- Extração dos *strings* relevantes ao emoCrawler;
- Processamento dos *strings* pelo emoCrawler (análise de aproximadamente 500 amostras por palavra buscada);
- Obtenção de média temporal do emoCrawler;
- Processamento do áudio pelo EmoVoice;
- Processamento do vídeo pelo eMotion;
- Obtenção da média temporal do eMotion;
- Normalização das saídas dos sensores para não saturação dos classificadores;
- Alimentação da rede neural com os dados normalizados;
- Alimentação do classificador *Naive Bayes* com os dados normalizados;
- Alimentação do classificador SVM multiclasse com os dados normalizados.

Em seguida, para os conjuntos de dados realizou-se:

- Execução das redes neurais;
- Execução do *Naive Bayes*;
- Execução do SVM multiclasse;
- Pós-processamento para análise competitiva das saídas.

4.3 Resultados

A primeira etapa de investigação consistiu no estudo dos resultados providos diretamente dos sensores, ou seja, da análise unimodal, com intuito de avaliar sua capacidade individual de acerto, com obtenção de compilação de resultados como os da Tabela 4.2.

Tabela 4.2. Exemplo de ficha de resultados.

ID Vídeo:	s43ha1					
Classificação humana:	Felicidade					
Saída	Resultado	Felicidade	Raiva	Nojo	Medo	Tristeza
Emo-Voice:	Tristeza	n.a	n.a	n.a	n.a	n.a
eMotion:	Felicidade	0,7026	0,0235	0,0498	0,1425	0,0475
emoCrawler	Felicidade	0,9705	0,0275	0,0000	0,0000	0,0020

Contudo, a escolha preliminar de algumas amostras para estudo do ferramental sinalizou certos problemas na correta detecção de emoções. A Tabela 4.3 ilustra alguns resultados de tais testes para um conjunto heterogêneo de amostras, indicando percentuais de acerto das ferramentas eMotion (face) e Emo-Voice (voz).

Tabela 4.3. Taxa percentual de acertos em análise unimodal – conjuntos com número de amostras variável.

<i>Emoção</i>	Face	Voz
<i>Felicidade</i>	12,5%	12,5%
<i>Raiva</i>	88,9%	11,1%
<i>Desagrado</i>	0%	0%
<i>Medo</i>	50,0%	50,0%

<i>Tristeza</i>	50,0%	100,0%
-----------------	-------	--------

A partir da observação da dificuldade dos sensores em captar as expressões de desagrado no conjunto preliminar, sugeriu-se a eliminação delas do banco de entradas dos sistemas de fusão. De fato, notou-se que a complexidade da expressão tanto do ponto de vista facial quanto vocal trouxe intensa confusão na detecção. Além disso, o processo inicial de classificação desse tipo de expressão por humanos apresentou grande dificuldade, dado o perfil amador dos atores. Isso tornou difícil a obtenção de um grupo razoável de amostras de desagrado para treinamento e validação das redes. A exclusão evitou, portanto, problemas de contaminação na classificação das outras quatro emoções.

Em um segundo momento, os dados das quatro emoções restantes foram separados em conjuntos de treinamento e testes para os algoritmos de classificação supervisionados, com números iguais de amostras para cada emoção.

4.3.1 Fusão com Redes Neurais

O treinamento da rede FFBP consistiu da aplicação do algoritmo *Resilient Backpropagation* ao conjunto de treinamento. Como método de definição da quantidade de neurônios na camada oculta, realizou-se estudo dos índices de convergência da rede (ver Figura 3.2).

O ajuste da rede probabilística, por sua vez, depende fortemente do parâmetro de espalhamento, um escalar positivo relacionado à distância entre os vetores de treinamento. A metodologia para a escolha do parâmetro consistiu em iniciá-lo com um valor alto (generalista) e realizar passos de redução até garantir total aderência dos dados de treinamento em simulação. Obteve-se através dessa operação um fator de espalhamento de 0,17.

A Tabela 4.4 apresenta, após o treinamento supervisionado da rede, os resultados do conjunto final de testes, comparando-se a avaliação isolada das emoções na face e na voz com os dados provindos da fusão multimodal.

Tabela 4.4. Comparativo das medições individuais com a fusão multimodal em redes neurais: taxas percentuais de acerto para cada método.

	Voz	Face	Fusão FFBP (face/voz/semântica)	Fusão PNN (face/voz/semântica)
<i>Felicidade</i>	20%	0%	60%	60%
<i>Raiva</i>	100%	0%	100%	100%
<i>Medo</i>	40%	20%	80%	60%
<i>Tristeza</i>	100%	60%	60%	60%
Acerto Médio	65%	20%	75%	70%
Desvio Padrão	41%	28%	19%	20%

Ainda que nos experimentos com fusão considere-se que a taxa de acerto aleatório seja de 25%, é importante ressaltar que para os algoritmos unimodais para voz e face, apresentados nas primeiras duas colunas, acertos acima aproximadamente 16,7% são considerados melhores do que a chance, posto classificarem mais elementos dos que constantes nesta análise.

4.3.2 Fusão com *Naive Bayes*

A avaliação da capacidade de classificação com *Naive Bayes* tomou como referência os mesmos conjuntos de treinamento e testes utilizados no caso das redes neurais. Os resultados obtidos através da avaliação probabilística no banco de testes podem ser observados na Tabela 4.5:

Tabela 4.5. Comparativo das medições individuais com a fusão multimodal em *Naive Bayes*: taxas percentuais de acerto para cada método.

	Voz	Face	Fusão <i>Naive Bayes</i> (face/voz/semântica)
<i>Felicidade</i>	20%	0%	80%
<i>Raiva</i>	100%	0%	60%
<i>Medo</i>	40%	20%	20%
<i>Tristeza</i>	100%	60%	40%
Acerto Médio	65%	20%	50%
Desvio Padrão	41%	28%	26%

O algoritmo apresentou a taxa de acerto mais baixa dentre os três classificadores testados, além do maior desvio padrão em relação à capacidade de acerto do conjunto total de expressões emocionais.

4.3.3 Fusão com *Support Vector Machine*

Os resultados obtidos com a fusão SVM multiclasse a partir dos bancos de treinamento e testes são ilustrados na Tabela 4.6:

Tabela 4.6. Comparativo das medições individuais com a fusão multimodal em SVM multiclasse: taxas percentuais de acerto para cada método.

	Voz	Face	Fusão SVM (face/voz/semântica)
<i>Felicidade</i>	20%	0%	80%
<i>Raiva</i>	100%	0%	80%
<i>Medo</i>	40%	20%	60%
<i>Tristeza</i>	100%	60%	100%
<i>Acerto Médio</i>	65%	20%	80%
<i>Desvio Padrão</i>	41%	28%	16%

Nota-se que o classificador não apenas apresenta a maior taxa de acerto dentre as três modalidades de seleção, como também a maior uniformidade de resposta, indicada pelo baixo desvio padrão.

Esse bom comportamento pode ser observado pela alta aderência dos itens testados ao banco de treinamento, com baixo índice de confusão. Nesse sentido, a Tabela 4.7 apresenta a matriz de confusão para o classificador SVM.

Tabela 4.7. Matriz de confusão para a técnica SVM multiclasse. Nas linhas encontram-se os valores previstos pelo classificador.

	<i>Felicidade</i>	<i>Raiva</i>	<i>Medo</i>	<i>Tristeza</i>
<i>Felicidade</i>	80%	0%	0%	0%
<i>Raiva</i>	0%	80%	20%	0%
<i>Medo</i>	20%	20%	60%	0%
<i>Tristeza</i>	0%	0%	20%	100%

Observa-se que a tendência de erro em SVM encontra-se localizada totalmente na decisão sobre a expressão emocional “Medo”: não apenas a detecção de outras emoções foi contaminada por ela, mas também sua validação apresentou a taxa de acerto mais baixa dentre as expressões (60%).

4.3.4 Tempo de Execução em Classificação

Além da eficiência de classificação dos algoritmos, é importante a avaliação do tempo de computação necessário para análise de cada conjunto de dados, sobretudo considerando-se que em condições normais de uso o sistema deverá realizar o processamento em um tempo mais rápido do que a duração da expressão que pretende detectar.

Quando se trata estritamente do tempo de execução de um classificador treinado, o custo para qualquer algoritmo é baixo e semelhante, relacionado à aplicação de uma função de transferência $O(M \times N)$ sobre as entradas, sendo M a quantidade de emoções e N o número de entradas. Entretanto, quando se observa também o tempo gasto em treinamento, os períodos serão mais longos. Posto que em um caso cotidiano o treinamento poderá ser reajustado durante a execução do sistema, é relevante avaliar o tempo total gasto para a tarefa de treinamento e subsequente execução. Nesse sentido, a Tabela 4.8 ilustra, para os diferentes algoritmos, o tempo médio de computação exigido:

Tabela 4.8. Tempo médio para treinamento e posterior execução de cada classificador, em segundos.

Classificador	Tempo de Execução (s)
<i>FFBPN</i>	1,00
<i>PNN</i>	(desprezível)
<i>Naive Bayes</i>	1,82
<i>SVM</i>	2,44

Observa-se que o classificador *PNN* obteve melhor desempenho em tempo total dentre os algoritmos testados, enquanto a *Support Vector Machine* teve o maior custo computacional.

4.3.5 Influência do emoCrawler

Antes da avaliação do uso do emoCrawler na base de dados considerada, um estudo adicional foi realizado com a aplicação direta do algoritmo sobre palavras diversas. A Tabela 4.9 ilustra um subconjunto desse estudo. Nela, observa-se os valores resultantes do emoCrawler, além da explicitação da emoção mais frequentemente detectada (maior valor) e da primeira emoção que classificadores humanos associaram a palavra.

Tabela 4.9 – Resultados da execução do emoCrawler para palavras diversas (não contidas no *corpus*), sem qualquer etapa de classificação. Aproximadamente 1500 buscas foram realizadas por palavra. Itens com (*) indicam expressões contidas no dicionário emocional e que, por consequência, são prontamente identificadas.

	Cl. humana	Cl. emoCrawler					
	1a. Associação	1a. Associação	Felicidade	Raiva	Medo	Tristeza	
Negativas	SCARY	Medo	Felicidade	0,111	0,027	0,05	0,008
	MONSTER	Medo	Felicidade	0,112	0,014	0,007	0,008
	HURT	Medo	Felicidade	0,145	0,105	0,023	0,07
	NIGHTMARE	Medo	Felicidade	0,082	0,013	0,014	0,021
	UNEASY	Medo	Felicidade	0,072	0,048	0,03	0,019
	TERROR*	Medo	Medo	0,042	0,005	1,42	0,005
	TORMENT	Raiva	Medo	0,028	0,008	0,037	0,006
	CRIME	Raiva	Felicidade	0,074	0,043	0,004	0,014
	MURDER	Raiva	Raiva	0,06	0,066	0,006	0,004
	ANNOYANCE*	Raiva	Raiva	0,111	1,742	0,009	0,014
	FIGHT	Raiva	Felicidade	0,087	0,038	0,008	0,017
	BRUTAL	Raiva	Felicidade	0,186	0,004	0,005	0,011
	SAD*	Tristeza	Tristeza	0,493	0,214	0,005	1,859
	ANGUISH	Tristeza	Felicidade	0,185	0,048	0,043	0,143
	DEATH	Tristeza	Felicidade	0,062	0,023	0,02	0,015
	SICKNESS	Tristeza	Tristeza	0,116	0,056	0,018	0,135

Positivas	SUFFERING	Tristeza	Felicidade	0,14	0,122	0,019	0,051
	GLOOMY	Tristeza	Felicidade	0,099	0,026	0	0,024
	INSPIRE	Felicidade	Felicidade	0,148	0,003	0,003	0,003
	MARVELOUS	Felicidade	Felicidade	0,142	0,014	0,007	0,005
	ROMANTIC	Felicidade	Felicidade	0,126	0,005	0	0,023
	HILARIOUS	Felicidade	Felicidade	0,127	0,025	0,003	0,004
	LOVE	Felicidade	Felicidade	0,157	0,054	0,001	0,01
	MUSIC	Felicidade	Felicidade	0,111	0,019	0	0,012
	MAGIC	Felicidade	Felicidade	0,09	0,004	0,001	0,009
	LAUGH	Felicidade	Felicidade	0,072	0,044	0,02	0,01
	WIN	Felicidade	Felicidade	0,134	0,0114	0,003	0,009
	SUCCESS	Felicidade	Felicidade	0,215	0,017	0,016	0,001
	FESTIVE	Felicidade	Felicidade	0,19	0,004	0,001	0,004
	WONDERFUL*	Felicidade	Felicidade	1,872	0,008	0,003	0,025

Tais resultados confirmaram a hipótese de projeto de que os dados gerados pelo emoCrawler, sem qualquer pós processamento, podem apresentar vieses, fruto de tendências culturais nas expressões apresentadas no Twitter. O sarcasmo e a ironia, muito frequentes nas postagens observadas, podem, por exemplo, ser tratados de forma equivocada como aspectos positivos em um texto analisado pelo algoritmo. Postagens tristes de um usuário por vezes são maquiadas como raiva para observação do grupo de amigos e familiares, evitando a exposição de fraquezas ou potenciais situações embaraçosas. Trabalhos em psicologia [Reimer, 2008] discutem o fenômeno da filtragem emocional e encenação para apresentação no círculo social e o que se nota é que as mídias digitais operam da mesma maneira, de forma que os dados recebidos pelo emoCrawler são passíveis de ruído de sua própria geração.

Portanto, para que o sistema de fusão tenha resultados relevantes, torna-se necessária a classificação ponderada das saídas de detecção semântica. Ainda que uma primeira aproximação manual e simplificada de classificação – a diminuição da influência da medição “felicidade” através de um multiplicador único - já trouxesse grande benefício ao sistema, optou-se por permitir que o próprio classificador de fusão realizasse a ponderação, como observado nos resultados dos itens 4.3.1 a 4.3.3.

Contudo, nota-se que os resultados apresentados na ocasião não realizam um comparativo com o uso unimodal do emoCrawler como fonte de dados. De fato, não é interessante realizar a comparação do emoCrawler separadamente, posto sua eficiência estar inerentemente ligada a técnica de classificação selecionada e os pesos associados aos termos de dicionário serem ajustados na fusão.

Dessa forma, como ferramenta comparativa para melhor observação da relevância do emoCrawler sobre a fusão, a rede neural tipo FFBP foi reconstruída e novamente treinada, agora com oito nós na camada de entrada, contudo desta vez apenas com alimentação dos dados de face e voz. A reformulação também acarretou a eliminação de um nó na camada oculta. A Tabela 4.10 apresenta os resultados desse processo:

Tabela 4.10 – Avaliação da eficiência do emoCrawler sobre o grupo de teste para a FFBP: taxas percentuais de acerto em cada caso.

	emoCrawler desabilitado	emoCrawler habilitado
<i>Felicidade</i>	20%	60%
<i>Raiva</i>	60%	100%
<i>Medo</i>	20%	80%
<i>Tristeza</i>	100%	60%
<i>Acerto Médio</i>	50%	75%

Nota-se o melhor comportamento da fusão quando da incorporação dos elementos de compreensão semântica, particularmente em emoções nas quais os sistemas de face e voz tiveram desempenho ruim. Apesar do resultado coerente, observa-se a diminuição do acerto individual no caso da tristeza, consequência provável da existência de grandes ruídos na base de treinamento do emoCrawler, os quais geraram confusão no tratamento dos dados.

4.3.6 Tempo de Execução no emoCrawler

Assim como no caso dos sistemas de classificação, considerações sobre a velocidade de execução do emoCrawler são relevantes para a avaliação da viabilidade de aplicação no caso em tempo real.

Na máquina utilizada durante a pesquisa, com conexão à Internet de aproximadamente 2MB/s, a avaliação de uma única postagem, incluindo sua busca, extração e comparação com dicionário, tomou aproximadamente 17 milissegundos. Considerando-se que as simulações deste trabalho amostram entre 500 e 1500 postagens por palavra de vídeo relevante, cada avaliação completa pode levar algumas dezenas de segundos.

Desse período, o gargalo de processo parece estar presente no subprocesso de busca e coleta dos textos do Twitter, via conexão com a Internet. De fato, mesmo com o uso da API dedicada para interface com o banco de dados da rede social, as buscas mostraram-se lentas e pouco confiáveis, por vezes retornando erros de conexão. Além do uso de banda de maior qualidade como solução para o problema, uma alternativa mais robusta para contornar essa questão é apresentada na seção 5.2.

5 Comentários Finais

Os resultados experimentais deste trabalho indicam que a fusão pode de fato fornecer dados melhores que abordagem unimodal, uma conclusão já esperada. Contudo, tal resultado superior apresentou-se verdadeiro mesmo com um pequeno banco de treinamento e algoritmos de classificação sem qualquer base probabilística com conhecimento *a priori* a respeito da eficiência dos sensores, o que fornece pistas para um aumento ainda maior da confiabilidade caso mecanismos de seleção dedicados a esse fim fossem elaborados.

Tal conclusão comprovou-se mais acentuada nas modalidades de fusão baseadas em Rede Neural do tipo *Feedforward Backpropagation* e em *Support Vector Machine*. Particularmente, o classificador apoiado em SVM apresentou-se com a melhor relação de benefício dentre as abordagens avaliadas, posto sua facilidade de ajuste, tempo de execução adequado e alta taxa de acertos para o pequeno conjunto amostral.

Verificaram-se também indícios de melhoria no desempenho da fusão multimodal com a inclusão do sistema de detecção semântica, o emoCrawler. Em sua aplicação, o grande volume de dados parece ter suprido a ausência de análise sintática profunda, colaborando na escolha da expressão emocional correta. Todavia, a melhoria dos resultados através do uso do emoCrawler deve ser ainda mais extensamente investigada, particularmente em casos de interações mais longas. A quantidade relativamente pequena de expressões linguísticas presentes no corpus pode ter afetado os resultados de forma a facilitar a identificação no grupo de testes.

Apesar dos resultados positivos, o trabalho sofreu consideravelmente as implicações de sua premissa de tratar dados fora de condições ideais de operação dos sistemas de detecção providos por terceiros, voltando-se a casos mais naturais e corriqueiros de interação. Dentre os diversos aspectos que contribuíram negativamente nessa abordagem, observa-se que, particularmente no caso do software eMotion, a influência da movimentação da cabeça e dos lábios durante a fala influenciou na geração de falsos positivos. Um sistema que tratasse essas variações seria consideravelmente mais preciso.

De qualquer forma, apesar das diversas limitações existentes, observa-se que a computação encontra-se em um momento de intensa humanização de seus produtos e os resultados apresentados neste trabalho, assim como em outras pesquisas da área, apontam para a viabilidade de um sistema de captura sensorial que forneça subsídios para a inteligência artificial lidar com o aspecto emocional do usuário. Ainda que quesitos graves relacionados à robustez em detecção precisem ser resolvidos, o maior desafio será encontrado na capacidade dos interessados em utilizar esse novo arcabouço de inteligência afetiva de maneira rica, inovadora e de forma harmoniosa com a experiência com o usuário.

5.1 Publicações

Os diversos resultados em pesquisa discutidos neste trabalho foram enviados para apreciação e aceitos em congressos e *proceedings* no campo da inteligência artificial, incorrendo nas seguintes produções:

- “*Crawling to Improve Multimodal Emotion Detection*”, em Springer Lecture Notes in Artificial Intelligence (LNAI), Mexican International Conference on Artificial Intelligence 2011 [Cueva et al 01, 2011];
- “*Fusão de Observações Afetivas em Cenários Realistas*”, em VIII Encontro Nacional de Inteligência Artificial do XXXI Congresso da Sociedade Brasileira de Computação [Cueva et al 02, 2011].

Além disso, a compreensão dos resultados e a metodologia de testes foram de cunho fundamental para a publicação de outros três artigos, relacionados ao estudo temporal de expressões faciais, nos quais o pesquisador é coautor:

- “*A Dynamic Model for Identification of Emotional Expressions*” em Mexican International Conference on Artificial Intelligence 2011[Gonçalves et al 01, 2011];

- “*Determinação da Emoção Demonstrada pelo Interlocutor*” em VIII Encontro Nacional de Inteligência Artificial do XXXI Congresso da Sociedade Brasileira de Computação [Gonçalves et al 02, 2011];
- “*A Model for Inference of Emotional States Based on Facial Expressions*” em Journal of the Brazilian Computer Society [Gonçalves et al 03, 2012].

5.2 Trabalhos Futuros

Os bons resultados apresentados pelas metodologias de fusão multimodal abrem caminho para o aperfeiçoamento da técnica. Para trabalhos futuros, é fundamental a busca de um *corpus* mais rico para treinamento e testes do sistema, fornecendo maior número de registros por emoção. Com isso, espera-se eliminar diversos vícios de atuação amadora e criar uma base de treinamento ainda mais sólida e relevante dos que a produzida neste trabalho.

Em relação ao sistema de fusão, um caminho proposto pretende incorporar ao projeto redes *Bayesianas* no módulo de fusão e realizar a comparação dos resultados obtidos com os dados fornecidos pelos classificadores aplicados. Apesar de trazerem consigo dificuldades na obtenção de uma topologia satisfatória, algo ausente nos classificadores aplicados neste trabalho, as redes bayesianas tem a possibilidade de estruturar relações e conhecimento mais relevantes sobre a capacidade de cada sensor disponível.

Por fim, o autor acredita que os resultados satisfatórios encontrados na utilização do emoCrawler permitem a continuação de seu desenvolvimento, com intuito de buscas semânticas mais extensas e melhor parametrizadas. A questão relacionada ao longo tempo de processamento do emoCrawler e a dependência de uma conexão constante com a Internet poderia ser abordada através da criação de um banco de dados local com atualização assíncrona.

Nesse sentido, um processo paralelo ao emoCrawler poderia buscar constantemente indexar localmente postagens do Twitter para posterior fácil acesso, mantendo o banco sempre que possível atualizado em relação ao que está sendo produzido na rede social em um dado momento. Como abordagem alternativa, tal processo poderia já realizar um pré-processamento das postagens e estruturar um grafo

de relações entre palavras indiretas de emoção com as do dicionário emocional, o que aceleraria ainda mais o processo e permitiria conclusões mais profundas sobre os textos da rede social.

6 Referências Bibliográficas

- [Alami et al, 1998] Alami, R., Chatila, R., Fleury, S., Ghallab, M., Ingrand, F. "An architecture for autonomy." *The International Journal of Robotics Research* 17.4, pgs. 315-337, 1998.
- [Amir et al, 2000] Amir, N., Ron, S., Laor, N. "Analysis of an emotional speech corpus in Hebrew based on objective criteria". *Proceedings of the ISCA Workshop on Speech and Emotion* (pp. 29–33), 2000.
- [Aristóteles, s.d.] Aristóteles. "Retórica". *Biblioteca de Autores Clássicos*. Imprensa Nacional – Casa da Moeda. Portugal, 2005.
- [Averill, 2001] Averill, J. "The Rhetoric of Emotion, With a Note on What Makes Great Literature Great". *Empirical Studies of the Arts*, Vol 19(1) 5-26, 2001.
- [Azcarate et al, 2005] Azcarate, A., Hageloh, F., van de Sande, K., & Valenti, R. "Automatic facial emotion recognition". *Universiteit van Amsterdam*. June, 2005.
- [Bartlett et al., 1999] Bartlett MS, Hager JC, Ekman P, Sejnowski TJ. "Measuring facial expressions by computer image analysis". *Department of Cognitive Science, University of California, San Diego, USA*, 1999.
- [Bedworth & O'Brien, 1999] Bedworth, M. D., O'Brien, J. "The Omnibus Model: A New Architecture for Data Fusion?" *Proceedings of the 2nd International Conference on Information Fusion (FUSION'99)*, Helsinki, Finlândia, Julho de 1999.
- [Bettadapura, 2012] Bettadapura, V. "Face expression recognition and analysis: The state of the art." *arXiv preprint arXiv: 1203.6722*, 2012.
- [Birdwhistell, 1970] Birdwhistell, R. "Kinesics and Context". *University of Pennsylvania Press, Philadelphia*, 1970.
- [Boser et al, 1992] Boser, B. E., Guyon, I. M., Vapnik, V. N. "A training algorithm for optimal margin classifiers". *COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. New York, NY, USA: ACM Press, Pgs. 144–152, 1992.
- [Bosse et al, 2008] Bosse, T.; Jonker, C.M.; Treur, J. "Formalisation of Damasio's theory of emotion, feeling and core consciousness". *Consciousness and Cognition* vol.17 Pgs. 94-113, 2008.
- [Brachman, 2004] Brachman, R. "Knowledge Representation and Reasoning". *Morgan Kaufmann, USA*, 2004.

[Broek & Westerink, 2009] Broek, E., Westerink, J., “Considerations for emotion-aware consumer products”. *Applied Ergonomics*, Volume 40, Issue 6, Pgs. 1055–1064, November, 2009.

[Burges, 1998] Burges, C. J. C. “A Tutorial on *Support Vector Machines* for Pattern Recognition”. *Data Mining and Knowledge Discovery*, vol. 2, Pgs. 121-167. 1998.

[Campanella e Belin, 2007] Campanella, S., Belin, P. “Integrating face and voice in person perception”. *Trends in Cognitive Sciences*, 11, Pgs. 535–543. 2007.

[Castellano et al., 2007] Castellano, G., Kessous, L. Caridakis, G. “Multimodal emotion recognition from expressive faces, body gestures and speech”. In Fiorella de Rosis, Roddy Cowie (Ed.), *Proc. of the Doctoral Consortium of 2nd International Conference on Affective Computing and Intelligent Interaction*, Pgs. 375-388, Lisbon, September 2007.

[Charniak, 1993] Charniak, E. “Statistical Language Learning (Language, Speech and Communication”. The MIT Press, USA, 1993.

[Chen & Huang, 2000] Chen, L. S., Huang, T. S. “Emotional expressions in audiovisual human computer interaction,” in *Proc. ICME*, Pgs. 423–426, 2000.

[Chetty & Wagner, 2008] Chetty, G. Wagner, M. “A Multilevel Fusion Approach for Audiovisual Emotion Recognition”. *International Conference on Auditory-Visual Speech Processing* 2008.

[Cueva et al 01, 2011] Cueva, D., Gonçalves, R., Cozman, F., Pereira-Barretto, M. “Crawling to Improve Multimodal Emotion Detection”. *Mexican International Conference on Artificial Intelligence*. Springer LNAI. Pgs. 343-350, 2011.

[Cueva et al 02, 2011] Cueva, D., Gonçalves, R., Pereira-Barretto, M., Cozman, F. “Fusão de Observações Afetivas em Cenários Realistas”. *VIII Encontro Nacional de Inteligência Artificial do XXXI Congresso da Sociedade Brasileira de Computação*. Pgs. 833-842. Natal, 2011.

[Darwin, 1872] Darwin, C. “The Expression of the Emotions in Man and Animals”. John Murray, London, 1872.

[De Silva & Ng, 2000] De Silva, L. C., Ng, P. C. “Bimodal emotion recognition,” in *Proc. FG*, pp. 332–335. 2000.

[De Silva et al, 1997] De Silva, Liyanage C., Tsutomu Miyasato, and Ryohei Nakatsu. "Facial emotion recognition using multi-modal information." *Information, Communications and Signal Processing*, 1997. ICICS., *Proceedings of 1997 International Conference on*. Vol. 1. IEEE, 1997.

[Ekman & Friesen, 1978] Ekman, P., Friesen, W. “Facial Action Coding System: A Technique for the Measurement of Facial Movement”. Consulting Psychologists Press, Palo Alto, 1978.

[Ekman & Oster, 1979] Ekman, P., Oster, H. “Facial Expressions On Emotion”. Annual Review of Psychology, 30, 527-554, 1979.

[Ekman, 1972] Ekman, P. “Universals And Cultural Differences In Facial Expressions Of Emotions”. In J. Cole (ed.), NEBRASKA SYMPOSIUM ON MOTIVATION, 1971. Lincoln, Neb.: University of Nebraska Press, 1972. Pp. 207- 283.

[Ekman, 2001] Ekman, P. “Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage”. W.W. Norton & Co., USA, 2001.

[eMotion01] ISLA, University of Amsterdam. “eMotion Visual Recognition” (software). Disponível em: <<http://www.visual-recognition.nl>> Acesso em: 23 de março de 2011.

[Essa & Pentland, 1997] Essa, I., Pentland, P. “Coding, analysis, interpretation, and recognition of facial expressions”. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7), 757-763. 1997.

[Eyben et al, 2009] Eyben, F., Wollmer, M., Schuller, B. “openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit”. In: Proc. ACII, Amsterdam, Netherlands, pp. 576–581. 2009.

[Feldman & Rimé, 1991] Feldman, R. S., Rimé, B. (Eds.). (1991). “Fundamentals of nonverbal behavior”. Cambridge, England: Cambridge University Press.

[Foreman, 2003] Foreman, J. “The 43 Facial Muscles That Reveal Even the Most Fleeting Emotions. A Conversation with Paul Ekman” New York Times, Ed. 5 de agosto de 2003.

[Freud, apud (Watson & Rayner, 1920)] Watson, J. B., & Rayner, R. “Little Emotional Albert. Conditioned emotional responses”. Journal of Experimental Psychology, 3, 1- 14. 1920.

[Gonçalves et al 01, 2011] Gonçalves, R., Cueva, D., Pereira-Barretto, M., Cozman, F. “A Dynamic Model for Identification of Emotional Expressions”. Mexican International Conference on Artificial Intelligence. 2011.

[Gonçalves et al 02, 2011] Gonçalves, R., Cueva, D., Cozman, F., Pereira-Barretto, M. “Determinação da Emoção Demonstrada pelo Interlocutor”. VIII Encontro Nacional de Inteligência Artificial do XXXI Congresso da Sociedade Brasileira de Computação. Pgs. 737-748, Natal, 2011.

[Goulding et al, 2012] Goulding, J., Smith, G., Barrack, D. “Neo-demographics and Distributions in the Digital Shadow” Third Annual Digital Economy All Hands Conference, 2012.

- [Hall, 1992] Hall, D. L. "Mathematical Techniques in Multi-Sensor Data Fusion". Artech House, Norwood, Massachusetts, 1992.
- [Hellweg & Carroll, 2011] Aguilera-Hellweg, M., Carroll, C. "Us. And them. Robots are being created that can think, act, and relate to humans. Are we ready?" National Geographic Magazine. Ago 2011.
- [Hofmann, 2006] Hofmann, M. "Support Vector Machines – Kernels and the Kernel Trick". Reading Club: *Support Vector Machines*. 2006.
- [Höök, 2012] Höök, Kristina. "Affective Computing: Affective Computing, Affective Interaction and Technology as Experience". Soegaard, Mads and Dam, Rikke Friis (eds.). "Encyclopedia of Human-Computer Interaction". Aarhus, Denmark: The Interaction Design Foundation. Disponível em: <http://www.interaction-design.org/encyclopedia/affective_computing.html>. Acesso em: 20 de novembro de 2012.
- [James, 1884] James, W. "What is an Emotion?" Mind, Vol. 9, Pgs. 188-205, 1884.
- [Kanade et al, 2000] Kanade, T., Cohn, J. F., Tian, Y. L. "Comprehensive database for facial expression analysis". Proc. 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), pgs 46--53, 2000.
- [Kim et al, 2007] Kim, S., Georgiou, P.G., Sungbok Lee, Narayanan, S. "Real-time Emotion Detection System using Speech: Multi-modal Fusion of Different Timescale Features" IEEE 9th Workshop on Multimedia Signal Processing, 2007. MMSP. 2007.
- [King, 2010] King, P. "Emotions in Medieval Thought", The Oxford Handbook of the Emotions, edited by Peter Goldie. Oxford University Press 2010, 167-188.
- [Kleinginna & Kleinginna, 1981] Kleinginna Jr., P., Kleinginna, A. "A Categorized List of Emotion Definitions, with Suggestions for a Consensual Definition". Georgia South College. Motivation and Emotion, Vol. 5, No. 4, Pgs. 345-379, 1981.
- [Laros, 2005] Laros, F.J.M.; Steenkamp, J.E.M. "Emotions in consumer behavior: a hierarchical approach". Journal of Business Research vol.58 pgs.1437-1445, 2005.
- [LI & REN, 2008] LI, J., & REN, F. "Emotion Recognition from Blog Articles". In : International Conference on Natural Language Processing and Knowledge Engineering. pp. 1-8. NLP-KE '08, Beijing, 2008.
- [Markin et al, 1997 (apud Esteban et al, 2005)] Esteban, J., Starr, A., Willetts, R., Hannah, P., & Bryanston-Cross, P. "A review of data fusion models and architectures: towards engineering guidelines." Neural Computing & Applications 14.4: 273-281, 2005.

[**Martin et al., 2005**] Martin, O. Kotsia, I. Macq, B. Pitas, I. “The eNTERFACE’05 Audio-Visual Emotion Database”. Université Catholique de Louvain; Aristotle University of Thessaloniki, 2005.

[**Martinez & Benavente, 1998**] Martinez, A. M., Benavente, R. "The AR Face Database," tech. rep., CVC #24, 1998.

[**Martinez-Miranda & Aldea, 2005**] Martinez-Miranda, J.; Aldea, A. “Emotions in human and artificial intelligence”. Computers in Human Behavior Vol.21 Pgs.323-341, 2005.

[**MathWorks01**] Probabilistic Neural Networks - MathWorks Matlab R2012b Documentation Center. Disponível em: <
<http://www.mathworks.com/help/nnet/ug/probabilistic-neural-networks.html>>. Acesso em: 20 de novembro de 2012.

[**MathWorks02**] The Mathworks. “Matlab R2010a” (software). Versão 7.10.0.499. [S.l.], 2010.

[**McCulloch & Pitts, 1943**] McCulloch, W. S. and Pitts, W. H. “A logical calculus of the ideas immanent in nervous activity”. Bulletin of Mathematical Biophysics, 5:115-133, 1943.

[**Mishra, A., 2011**] Mishra, A. “Multi Class *Support Vector Machine*”. Disponível em: <
<http://www.mathworks.com/matlabcentral/fileexchange/33170-multi-class-support-vector-machine>>. Acesso em: 7 de setembro de 2012.

[**Muramatsu & Hanoch, 2005**] Muramatsu, R.; Hanoch, Y. “Emotions as a mechanism for boundedly rational agents: the fast and frugal way”. Journal of Economic Psychology vol.26 pgs.201-221, 2005.

[**Naab & Russel, 2007**] Naab, P. Russel, J. “Judgments of Emotion From Spontaneous Facial Expressions of New Guineans”. Emotion, 2007, Vol. 7, No. 4, pgs. 736 –744.

[**Ortony et al, 1988**] Ortony, A., Clore, G. C., Collins, A. “The Cognitive Structure of Emotion”. Cambridge University Press. 1988.

[**Oudeyer, 2002**] Oudeyer, P. “The Production and recognition of emotion in speech: features and algorithms” Human-Computer Studies, UK, 2002.

[**Pantic & Rothkrantz, 2003**] M. Pantic, L.J.M. Rothkrantz, “Toward an affect-sensitive multi-modal human–computer interaction”, Proceedings of the IEEE 91 (9) 1370–1390. 2003.

[**Picard, 1995**] Picard, R. "Affective Computing". M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321, 1995.

[**Planalp, 1999**] Planalp, S. “Communicating Emotion: Social, Moral, and Cultural Processes.” Pg. 11. Cambridge: Cambridge UP, 1999.

- [Ptaszynski et al., 2009] Ptaszynski, M., Dybala, P., Shi, W., Rzepka, R., Araki, K. "Towards Context Aware Emotional Intelligence in Machines: Computing Contextual Appropriateness of Affective States". IJCAI'09 Proceedings of the 21st international joint conference on Artificial intelligence. 2009.
- [Rachuri et al, 2010] Rachuri, K.K.; Musolesi, M.; Mascolo, C.; Rentfrow, P.; Longworth, C.; Aucinas, A. "EmotionSense: a mobile phone based adaptive platform for experimental social psychology research". UbiComp '10, Sep 26-Sep 29, Copenhagen, Denmark, 2010.
- [Rao, 2004] Rao, Nageswara S. V. "A Generic Sensor Fusion Problem: Classification and Function Estimation". In Proceedings of the Multiple Classifier Systems: 5th International Workshop. Springer, 2004.
- [Reimer, 2008] Reimer, S. G. "The Hidden Cost of Hiding Feelings: Emotion Suppression and Inauthenticity in Social Anxiety". Dissertação. University of Waterloo, 2008.
- [Rish, 2001] Rish, I. "An empirical study of the *Naive Bayes* classifier." IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. 2001.
- [Rojas, 1996] Rojas, R. "Neural Networks - A Systematic Introduction". Springer-Verlag, Berlin, New-York, 1996.
- [Rosa, 1946] Guimarães Rosa, J. "A Hora e a Vez de Augusto Matraga". Sagarana. São Paulo: Nova Fronteira, 2001.
- [Roseman, 2001] Roseman, I.J.; Smith, C.A. "Appraisal Theory - Overview, Assumptions, Varieties, Controversies". In "Appraisal Processes in Emotion – Theory, Methods, Research" editado por Scherer, K; Schorr, A.; Johnstone, T. Oxford University Press, 2001.
- [Rumelhart et al, 1986] Rumelhart, D. E., Hinton, G. E., Williams, R. J. "Learning internal representations by error propagation". In Parallel distributed processing: explorations in the microstructure of cognition, vol. 1, MIT Press, Cambridge, MA, EUA, Pgs. 318-362, 1986.
- [Russel & Norvig, 2003] Russell, S. J., Norvig, P. "Artificial Intelligence: A Modern Approach". ed. 2, Upper Saddle River, New Jersey: Prentice Hall, 2003.
- [Sander et al, 2005] Sander, D., Grandjean, D., Scherer, K. A systems approach to appraisal mechanisms in emotion. Neural Networks., 18, Pgs. 317–352, 2005.
- [Scherer & Ellgring, 2007] Scherer, K. Ellgring, H. "Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns?". Emotion; Emotion 7.1, Pgs. 158-171, 2007.

[**Scherer, 1981**] Scherer, K. "Speech and Emotional States". In Darby, JK, Speech evaluation in psychiatry, 1981.

[**Scherer, 1995**] Scherer, K. "Expression of Emotion in Voice and Music". Journal of Voice, 9(3), Pgs. 235–248, 1995.

[**Scherer, 2005**] Scherer, K. "What are emotions? And how can they be measured?" Social Science Information December 2005, vol. 44, no.4, pgs. 695-729.

[**Scherer, 2010**] Scherer, K. "A blueprint for affective computing: a sourcebook". Oxford : Oxford University Press, 2010.

[**Schorr, 2001**] Schorr, A. "Appraisal – The Evolution of an Idea". In "Appraisal Processes in Emotion – Theory, Methods, Research" editado por Scherer, K; Schorr, A.; Johnstone, T. Oxford University Press, 2001.

[**Schuller et al, 2011**] Schuller, B., Batliner, A., Steidl, S., Seppi, D. "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge". Speech Communication. Volume: 53, Issue: February, Publisher: Elsevier, Pgs. 1062-1087, 2011.

[**SOAR, 2006**] Lehman, J., Laird, J., Rosenbloom, P. "A Gentle Introduction to SOAR, an Architecture for Human Cognition: 2006 Update". SOAR Team, 2006.

[**St. Agostinho, s.d.**] Santo Agostinho. "The City of God". Trad. Gerald Groveland Walsh. Garden City. Image, NY, 1958.

[**Strapparava et al, 2006**] Strapparava, C., Valitutti, A., Stock, O. "The affective weight of lexicon". In Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Pgs. 423-426, Italy, May 2006.

[**Sun & Yin, 2008**] Sun, Yi, Yin, Lijun. "Facial Expression Recognition Based on 3D Dynamic Range Model Sequences". Computer Vision ECCV 2008, pgs. 58-71, 2008.

[**THOCP, 2007**] "Xerox Star". The History of Computing Project. Disponível em: < http://www.thocp.net/hardware/xerox_star.htm>. Acesso em: 19 de janeiro de 2013.

[**Thomson, 2009**] Thomson, H. "Emotional speech leaves 'signature' on the brain". New Scientist Health. Disponível em: < <http://www.newscientist.com/article/dn17133-emotional-speech-leaves-signature-on-the-brain.html>>. Acesso em: 20 de setembro de 2011.

[**Tokuhisa et al., 2008**] Ryoko Tokuhisa, Kentaro Inui, Yuji Matsumoto. Emotion Classification Using Massive Examples Extracted from the Web, In Proc. of Coling 2008, pgs. 881-888, 2008.

[**Turing, 1950**] Turing, A. "Computing Machinery and Intelligence", Mind V. 59, Pgs. 433-460, Out., 1950.

[**twitter01**] Twitter – The best way to discover what’s new in your world. Disponível em: <<http://www.twitter.com>>. Acesso em: 30 de março de 2011.

[**Valstar et al, 2007**] Michel F. Valstar, Hatice Gunes, and Maja Pantic. 2007. “How to distinguish posed from spontaneous smiles using geometric features”. Proceedings of the 9th international conference on Multimodal interfaces (ICMI '07). ACM, New York, NY, USA, Pgs. 38-45, 2007.

[**Vogt & André, 2005**] Vogt, T., André, E., "Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition," IEEE International Conference on Multimedia & Expo (ICME 2005), Pgs. 474-477, 2005.

[**Vogt et al., 2008**] Vogt, T., André, E., Bee, N. "EmoVoice – A framework for online recognition of emotions from voice". Perception in Multimodal Dialogue Systems, Pgs. 188-199, 2008.

[**Zajonc, 1980**] Zajonc, R. “Feeling and Thinking: Preferences Need No Inferences”. American Psychologist, Vol. 35, No. 2. (February 1980), Pgs. 151-175, 1980.