Fusão de Observações Afetivas em Cenários Realistas

Diego R. Cueva[1], Rafael A. M. Gonçalves[1], Fábio Cozman[1], Marcos R. Pereira-Barretto[1]

[1]Departamento de Engenharia Mecatrônica e Sistema Mecânicos
Escola Politécnica da Universidade de São Paulo (EPUSP)
Av. Prof. Melo Moraes 2231 – São Paulo – SP - Brasil

marcos.barretto@poli.usp.br

Abstract. *This paper demonstrates multimodal fusion of emotion sensory data in realistic scenarios of relatively long human-machine interactions. Fusion, combining voice and facial expressions, has been enhanced with semantic information retrieved from Internet social networks, resulting in more accurate determination of the conveyed emotion.*

Resumo. *Este artigo demonstra uma técnica para a determinação da emoção demonstrada por um interlocutor durante uma interação homem-máquina relativamente longa, utilizando a fusão multimodal de sinais provenientes das emoções na fala e na expressão facial, combinados à análise semântica do conteúdo da fala, realizada com base em informações recuperadas a partir de redes sociais na Internet.*

1. Introdução

Na contínua evolução da computação, o desenvolvimento de experiências de usuário amigáveis tem apresentado lenta inovação desde o surgimento das primeiras interfaces gráficas na década de 70. Ainda que o aprimoramento organizacional, gráfico e de entrada desses sistemas tenha sido de grande relevância, a conservação do paradigma é empecilho para que uma interação mais rica e inteligente entre ser humano e máquina possa prevalecer. O que de fato acaba por existir na atual realidade é a exigência de que o operador aprenda e adapte-se à maneira de operar de cada máquina, sempre de forma restritiva, delimitada pelos casos de uso fixados pelo desenvolvedor.

Emular a capacidade humana de contextualização de uma conversa e de reação flexível ao significado semântico seria a 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 ter-se máquinas capazes de compreender de forma mais adequada as necessidades e dificuldades do usuário, tornando mais próxima a possibilidade de uma interação genuinamente focada no indivíduo.

Motivados por essas questões, diversos trabalhos buscam elaborar algoritmos computacionais para detecção de emoções, sejam elas extraídas de elementos faciais ou

vocais. Em geral, tais trabalhos suportam-se sobre a conceituada e amplamente utilizada classificação facial elaborada por [Ekman e Friesen, 1977], a qual distingue seis blocos de expressões emocionais básicos: felicidade, tristeza, desagrado, medo, raiva e surpresa. 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 o senso comum disponível na rede.

Contudo, ainda que o campo em questão 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, de observação uma única fonte de emoção) ignora a relação intrínseca entre essas diversas entradas afetivas, cuja relevância é estudada em detalhes por [Scherer e Ellgring, 2007].

Tendo tal problemática como motivação, este artigo investiga a fusão dessas diversas componentes emocionais – ou "sensores" – para determinação da emoção demonstrada pelo interlocutor, em uma solução multimodal. Com esse objetivo, serão propostos dois classificadores para avaliação dos pesos e relações entre as diferentes entradas fornecidas, os quais serão comparados às soluções unimodais. O artigo está organizado de forma a apresentar os trabalhos relevantes nesse contexto na seção 2, descrever as ferramentas utilizadas na seção 3 e elaborar e apresentar experimentos nas seções 4 e 5.

2. Trabalhos Relacionados

O entendimento e modelagem formal do comportamento humano é assunto extensamente discutido pela psicologia e neurociência. Do determinismo comportamentalista às teorias cognitivas mais recentes, observa-se o escalonamento na complexidade e quantidade de variáveis necessárias à compreensão do indivíduo e de suas decisões. Entre os inúmeros trabalhos que discutem o tópico, a linha de estudo das Teorias Cognitivas (*Appraisal Theories*), discutida por [Roseman, 2001] e [Schorr, 2001], fornece um paradigma satisfatório para o escopo de diversos trabalhos da área da computação, 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 elicitação de emoções são comuns a todas as pessoas, mas o desenvolvimento desses processos varia individualmente, respeitando a experiência de vida de cada um. A quantificação da dinâmica dessas elicitações é também assunto frequentemente discutido, tal como em [Sander, 2005].

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. Na década seguinte, com o aumento da capacidade computacional, as abordagens baseadas em modelos tridimensionais da face e fluxo ótico em tempo real tornaram-se mais comuns, mostrando avanços em relação às anteriores.

Na análise de emoções na voz, pode-se citar os avanços recentes em [Eyben,2009] e [Rachuri, 2010]. Frequentemente, trabalhos discutem a classificação através da valência do tom do discurso (positiva, neutra ou negativa), enquanto outros tentam buscar comportamentos mais bem definidos; [Voigt e André, 2005] fazem um comparativo dessas alternativas sobre um mesmo classificador, discutindo sua eficiência.

Na utilização da Internet como banco de dados para compreensão emocional de palavras, algumas pesquisas recentes, tal como a de [Ptaszynski *et al*., 2009], começam a apresentar resultados animadores.

A análise multimodal, por sua vez, apresenta quantidade ainda pequena e recente de pesquisas e publicações. Além de [Scherer e Ellgring, 2007], [Campanella e Belin, 2007] realizam uma discussão atual dos estudos cognitivos, suportando a correlação entre voz e expressões faciais na demonstração de emoções. Do ponto de vista computacional, [Castellano *et al.*, 2007] apresentam a fusão de dados de voz, face e expressões corporais com mais de dez por cento de aprimoramento em comparação com a abordagem unimodal. [Chetty e Wagner, 2008], por sua vez, investigam o nível em que a fusão deve ocorrer, tomando face e voz como entradas.

3. Ferramentas Utilizadas

Para análise de expressões faciais, o *software* comercial eMotion [eMotion01], desenvolvido na Universidade de Amsterdam, Holanda, foi incorporado ao estudo. Sua escolha está relacionada ao bom comportamento do algoritmo de ajuste de malhas tridimensionais e a capacidade de avaliar os estados emocionais tradados neste artigo.

No processamento de voz, utilizou-se o pacote Emo-Voice [Vogt *et al*., 2008], desenvolvido pelo Instituto de Ciências da Computação da Universidade de Ausburgo. O Emo-Voice é disponibilizado em licença aberta e permite o treinamento personalizado de classificadores para detecção de emoções no discurso. O classificador SVM (*Support Vector Machine*) selecionado para voz foi treinado a partir de amostras do *corpus* (20 para cada emoção), de forma que o algoritmo adaptasse as características de gravação e da língua presentes no banco de dados.

Para a análise do significado emocional do discurso, foi desenvolvida uma ferramenta dedicada, batizada emoCrawler, a qual terá sua validade discutida nos experimentos apresentados neste artigo. A ferramenta utiliza os verbos, adjetivos e substantivos do discurso em buscas em diferentes bases de dados, observando a reação emocional apresentada nos textos.

3.1. O emoCrawler

Por ter sido desenvolvido no contexto deste trabalho, discute-se neste item o emoCrawler. O emoCrawler é um aplicativo modular em desenvolvimento, possuindo como objetivo a avaliação de contextos emocionais sem exigência de tratamentos semânticos complexos. Ignorando as questões abertas de processamento sintático em conversação, o programa realiza buscas em bases de dados *online* de redes sociais, processando grandes volumes de resultados e analisando frequências relativas de expressões de caráter emocional. Dessa forma, o programa conta com um dicionário

inspirado pelas propostas de Goleman (apud [Martines-Miranda,2005]) e [Laros, 2005], o qual mapeia expressões diversas às emoções utilizadas neste artigo.

As palavras consideradas significativas expressadas pelo usuário são extraídas do discurso e buscadas isoladamente na rede social. As mensagens relacionadas a essa consulta são então indexadas e verificadas quanto à existência de expressões de emoção contidas no dicionário. O algoritmo atua através de regras, buscando eliminar falsos positivos, por exemplo, considerando a presença de elementos de negação no tratamento das buscas. O fluxo de dados no aplicativo está ilustrado na Figura 1.

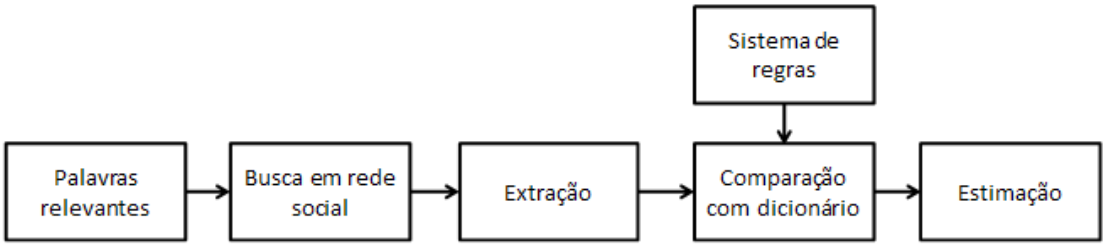

Figura 1. Fluxo de dados no emoCrawler

Dado o contexto modular do emoCrawler, diversas bases de dados podem ser examinadas. Para o presente trabalho, considerou-se uma amostra de consultas retirada da rede social Twitter [twitter01]. A escolha do Twitter está relacionada à abordagem concisa e altamente opinativa dos textos elaborados pelos seus usuários, de forma que mensagens frequentemente demonstram elementos de emoção, ao mesmo tempo em que não contêm construções complexas, dado a limitação de 140 caracteres. Dá-se preferência às consultas em Inglês, o que evita questões ligadas à flexão de verbos, substantivos e adjetivos em Português.

O sistema deve considerar também a questão temporal em sua análise. A busca de emoções relacionadas a um elemento pode apresentar diferentes resultados em diferentes momentos de pesquisa; a opinião popular instantânea pode divergir do passado. Todavia, dado o contexto predominantemente atemporal das frases utilizadas neste experimento, o modo de operação aplicado a este artigo busca apenas o senso comum presente relacionado a cada elemento de busca.

4. Metodologia experimental

Tendo como objetivo a análise multimodal, envolvendo expressões faciais, voz e contexto da fala, foi utilizado o *corpus eNTERFACE'05 Audio-Visual Emotion Database* [Martin *et al*., 2005], um banco de dados de cenas no qual indivíduos são convidados a expressar uma frase emocional da forma que mais desejarem (Figura 2).

Para a execução dos trabalhos, selecionaram-se três subconjuntos de amostras que pudessem ser prontamente qualificadas por observadores humanos, de forma que cenas com ambiguidades emocionais fossem descartadas. Um dos grupos foi utilizado no treinamento do Emo-Voice, enquanto outro foi utilizado para treinamento dos classificadores e, finalmente, o terceiro para validação dos resultados. Ainda que tenha

sido realizada uma seleção prévia do *corpus*, consideraram-se vídeos em condições não ideais para os diversos sistemas de classificação: ruído em áudio, iluminação não uniforme e movimentos abruptos da cabeça são alguns dos fatores que permaneceram intencionalmente no conjunto.

Figura 2. Exemplos de expressões do corpus

Como classificadores para a fusão foram utilizadas redes neurais, escolha comum para situações em que as fontes possuem ruídos diversos e nas quais o grau de confiabilidade relativa entre elas apresenta-se desconhecido. Além disso, tais redes acabam por avaliar também a significância entre os resultados de uma mesma fonte, o que é essencial no caso do emoCrawler.

A escolha das entradas analisadas levou em conta um subconjunto das emoções de Ekman, deixando de fora "surpresa", a qual é frequentemente compreendida como uma expressão sem valência, não abrangendo de fato um estado emocional (a surpresa pode estar atrelada a qualquer estado). Dessa forma, a rede neural possui entradas provindas das três diferentes fontes, uma por emoção de cada fonte. O classificador contém uma saída para cada emoção, de forma a realizar o caminho inverso de classificação nominal, ou seja, conversão para "nomes de emoções". Ainda que alguns trabalhos descrevam transições continuas para algumas emoções, neste contexto seria inviável mapear a saída para uma escala gradual, posto que as emoções trabalhadas não se apresentam ordenáveis em intensidade ou valência. A Figura 3 ilustra o processo.

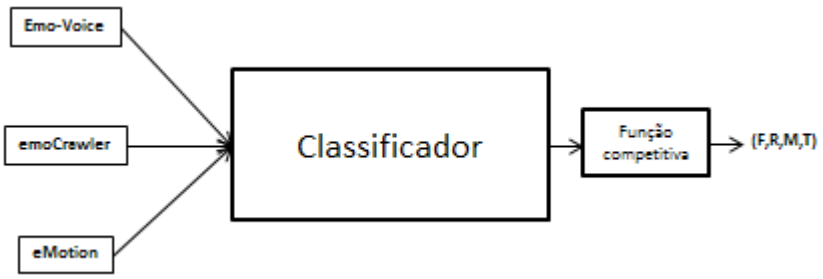

Figura 3. Fusão dos sensores

Para realização dos experimentos, duas abordagens de construção da topologia e algoritmo de treinamento foram consideradas: uma rede neural do tipo *Feed-Forward Back-Propagation* (FFBP) e uma rede neural probabilística (PNN). A rede FFBP é solução comum em problemas de classificação; possui uma camada oculta de nós, de forma a absorver não linearidades existentes e atua com diferentes funções de transferência nas camadas. A PNN, por sua vez, apresenta-se como solução alternativa relevante, dado seu tempo de treinamento ordens de grandeza mais rápido que a FFBP.

Entretanto, seu ajuste é por vezes de difícil realização em conjuntos de dados esparsos – como é o caso.

5. Experimento e Resultados

A primeira etapa de investigação consistiu na análise dos resultados provindos diretamente dos sensores, ou seja, da análise unimodal. A escolha preliminar de algumas amostras para estudo do ferramental sinalizou certos problemas na detecção. A Tabela 1 mostra os resultados de tais testes para um conjunto heterogêneo de amostras, indicando percentuais de acerto das ferramentas.

Tabela 1. Taxa percentual de acertos em análise unimodal - conjunto heterogêneo de amostras

Emoção	**Face**	**Voz**
Felicidade	12,5%	12,5%
Raiva	88,9%	11,1%
Desagrado	0%	0%
Medo	50,0%	50,0%
Tristeza	50,0%	100,0%

A observação da dificuldade dos sensores em captar as expressões de desagrado no conjunto preliminar sugeriu a eliminação delas do banco de entradas da rede neural. Além disso, o processo inicial de classificação desse tipo de expressão por humanos apresentou grande dificuldade de detecção, 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 as redes neurais, com números iguais de amostras para cada emoção. O treinamento da rede FFBP consistiu da aplicação do algoritmo *Robust Backpropagation* ao conjunto de treinamento. A definição da quantidade de neurônios na camada oculta foi realizada através do estudo dos índices de convergência da rede.

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 2 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 (não há motivo em realizar a

comparação do emoCrawler separadamente, posto sua eficiência estar inerentemente ligada à rede).

Tabela 2. Comparativo das medições individuais com a fusão multimodal: taxas percentuais de acerto para cada método

	Voz	**Face**	**Fusão FFBP (face/voz/semântica)**	**Fusão PNN (face/voz/semântica)**
Felicidade	20%	0%	60%	60%
Raiva	100%	0%	100%	100%
Medo	40%	20%	80%	60%
Tristeza	100%	60%	60%	60%
Acerto Médio	**65%**	**20%**	**75%**	**70%**

Para observação da relevância do emoCrawler sobre a fusão, a rede FFPB foi reconstruída e novamente treinada, agora com oito nós na camada de entrada. A reformulação também acarretou a eliminação de um nó na camada oculta. A Tabela 3 apresenta os resultados desse processo.

Tabela 3 – Avaliação da eficiência do emoCrawler sobre o grupo de teste para a FFBP: taxas percentuais de acerto em cada caso

	emoCrawler desabilitado	**emoCrawler habilitado**
Felicidade	20%	60%
Raiva	60%	100%
Medo	20%	80%
Tristeza	100%	60%
Acerto Médio	**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.

6. Conclusões e Trabalhos Futuros

Dados oriundos dos sistemas de medição utilizados são frequentemente sujeitos a ruídos sistemáticos. Observou-se essa característica, sobretudo, nos dados de face e semântica,

onde frequentemente houve tendência em classificar como mais intenso certo tipo de emoção, independentemente da entrada. As redes neurais foram bem sucedidas em observar essas tendências e realizar correções nos pesos das variáveis.

Verificaram-se indícios de melhoria no desempenho da fusão multimodal com a inclusão do emoCrawler . O grande volume de dados parece ter suprido a ausência de análise sintática profunda. Contudo, a melhoria dos resultados através do uso do emoCrawler deve ser ainda mais extensamente investigada. 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. Particularmente, observa-se no caso do *software* eMotion a influência da movimentação da cabeça e dos lábios durante o registro das emoções.

Como trabalhos futuros, planeja-se a utilização de outros classificadores e a análise comparativa em relação aos resultados descritos neste trabalho. Também se planeja a utilização da fusão multimodal em situações ainda mais longas de interação homem-máquina, buscando-se uma maior aproximação com a realidade.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), ao Departamento de Engenharia Mecatrônica da Escola Politécnica da Universidade de São Paulo e à FAPESP (por suporte através do processo 2008/03995-5) pela colaboração e financiamento das pesquisas.

Referências

[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.

[Campanella e Belin, 2007] Campanella, S., Belin, P. “Integrating face and voice in person perception”. Trends in Cognitive Sciences, 11, 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, Lisbon, September 2007.

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

[eMotion01] Visual Recognition. Disponível em: <http://www.visual-recognition.nl> Acesso em: 23 de março de 2011.

[Ekman e Friesen, 1977] Ekman, P., Friesen, W. “Facial Action Coding System”, Consulting Psychologist Press, 1977

[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.

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

[Pantic, 2000]Pantic, M.; Rothkrantz, L.J.M. “Automatic analysis of facial expressions: state of art”. IEEE Trans. On Pattern Analysis and Machine Intelligence, vol.22 no. 12, December, 2000.

[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 jont conference on Artifical intelligence. 2009.

[Rachuri et alli, 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, 2010, Copenhagen, Denmark.

[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, USA, 2001.

[Sander, 2005] Sander, D.; Grandjean, D.; Scherer, K. R. “A systems approach to appraisal mechanisms in emotion” Neural Networks vol. 18 pgs. 317-352, 2005.

[Scherer e Ellgring, 2007] Scherer, K. Ellgring, H. “Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns?”. American Psychological Association. 2007.

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

[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.

[Vogt *et al*., 2008] T. Vogt, E. André and N. Bee, "EmoVoice - A framework for online recognition of emotions from voice," in Proceedings of Workshop on Perception and Interactive Technologies for Speech-Based Systems, 2008.

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