ARTIGO

ABORDAGEM MULTIMODAL PARA RECONHECIMENTO AUTOMÁTICO DE EMOÇÕES APLICADA AO ESTUDO DE NÍVEIS DE TENSÃO EM TELEJORNAIS

MOISÉS HENRIQUE RAMOS PEREIRA
Centro Federal de Educação Tecnológica de Minas Gerais, Brasil

FLÁVIO LUIS CARDEAL PÁDUA
Centro Federal de Educação Tecnológica de Minas Gerais, Brasil

GIANI DAVID SILVA
Centro Federal de Educação Tecnológica de Minas Gerais, Brasil

RESUMO - Este artigo apresenta uma abordagem multimodal para reconhecimento automático de emoções em participantes de telejornais (apresentadores, repórteres, comentaristas, entre outros) capaz de auxiliar o estudo de níveis de tensão em narrativas de acontecimentos neste gênero televisivo. A metodologia faz uso de métodos computacionais do estado da arte para processamento e análise de expressões faciais, bem como modulações sonoras de falas. A abordagem proposta contribui para o estudo semiodiscursivo de telejornais e suas práxis enunciativas, auxiliando, por exemplo, na identificação das estratégias de comunicação desses programas. Para avaliar a aplicabilidade da abordagem proposta, utilizou-se uma amostra de vídeo referente a uma reportagem exibida em um telejornal brasileiro de grande popularidade no estado de Minas Gerais. Os resultados experimentais são promissores quanto ao reconhecimento de emoções sobre as expressões faciais dos telejornalistas e à distribuição dos indicadores audiovisuais extraídos ao longo de um telejornal, demonstrando o potencial da abordagem para apoiar a análise do discurso telejornalístico.
Palavras-chave: Telejornalismo. Níveis de Tensão. Reconhecimento de Emoções. Fala. Expressões Faciais.

ENFOQUE MULTIMODAL DE RECONOCIMIENTO AUTOMÁTICO DE EMOCIONES APLICADA AL ESTUDIO DE NIVELES DE TENSIÓN EN TELEDIARIO

RESUMEN - EEste artículo presenta un enfoque multimodal para el reconocimiento automático de las emociones en los participantes de los programas de noticias (presentadores, reporteros, comentaristas, etc.) capaz de ayudar al estudio de los niveles de estrés en los relatos de los sucesos en este género televisivo. La metodología ha utilizado métodos del estado

del arte para el procesamiento y análisis de las expresiones faciales y modulaciones sonoras del habla. El enfoque propuesto contribuye para las investigaciones semiodiscursivas de telediarios y sus praxis enunciativas, contribuyendo, por ejemplo, a la identificación de las estrategias de comunicación de estos programas. Para evaluar la aplicabilidad del enfoque propuesto, se utilizó una muestra de vídeo que se refiere a un reportaje presentado en un telediario brasileño de gran popularidad en Minas Gerais. Los resultados experimentales son prometedores para el reconocimiento de las emociones en las expresiones faciales de los periodistas y la distribución de los indicadores audiovisuales extraídos en un programa de noticias de televisión, lo que demuestra el potencial del enfoque para apoyar el análisis del discurso periodístico.
Palabras clave: Telediarios. Niveles de Tensión. Reconocimiento de Emociones. Discurso. Expresiones Faciales.

MULTIMODAL APPROACH FOR AUTOMATIC EMOTION RECOGNITION APPLIED TO THE TENSION LEVELS STUDY IN TV NEWSCASTS

ABSTRACT - This article addresses a multimodal approach to automatic emotion recognition in participants of TV newscasts (presenters, reporters, commentators and others) able to assist the tension levels study in narratives of events in this television genre. The methodology applies state-of-the-art computational methods to process and analyze facial expressions, as well as speech signals. The proposed approach contributes to semiodiscoursive study of TV newscasts and their enunciative praxis, assisting, for example, the identification of the communication strategy of these programs. To evaluate the effectiveness of the proposed approach was applied it in a video related to a report displayed on a Brazilian TV newscast great popularity in the state of Minas Gerais. The experimental results are promising on the recognition of emotions on the facial expressions of tele journalists and are in accordance with the distribution of audiovisual indicators extracted over a TV newscast, demonstrating the potential of the approach to support the TV journalistic discourse analysis.
Key words: TV Newscasts. Tension Levels. Emotion Recognition. Speech. Facial Expressions.

1 INTRODUÇÃO

Este artigo propõe uma nova abordagem para a análise semiodiscursiva de telejornais ao contribuir com técnicas computacionais para a determinação automática de níveis de tensão nos vídeos desses programas, por meio do reconhecimento multimodal de emoções sobre as características audiovisuais das modulações nos sinais de áudio dos arquivos de vídeo e nas expressões faciais de seus participantes, considerando os efeitos de intensidade ou apreensão promovidos pela identificação do plano fílmico da câmera em que as faces estão submetidas.

O estilo dos programas televisivos é construído, geralmente, pela redundância na enunciação, refletindo certa significância

semântica das notícias a fim de transmitir credibilidade e isso é pautado e conduzido, sob um estilo próprio, pelos telejornalistas apresentadores e pelos repórteres (GOFFMAN, 1981). Além disso, a narrativa visual oferece elementos imagéticos que podem ser usados de forma padrão e possuem a capacidade de transmitir uma carga emocional à narrativa televisiva por meio do enquadramento, ângulo, forma, cor e movimento da câmera (BLOCK, 2010). Neste contexto, o telejornalismo é considerando um tipo de linguagem institucional que se ritualiza nas relações mútuas de autorização e de legitimidade promovidas pelos sujeitos do discurso a fim de transmitir a informação para os telespectadores (STAM, 1985).

A sequência das notícias e a articulação dos apresentadores de programas televisivos vêm sendo objetos de estudo. Existem diversos esforços em descobrir padrões que revelem a estratégia de comunicabilidade pretendida por meio do sequenciamento de notícias, dos planos de enquadramento de câmera em telejornais e da influência dos apresentadores na construção do *ethos* semiodiscursivo de programas desse gênero televisivo (PEREIRA *et al.* 2014; GUTMANN, 2012). De forma natural, os jornalistas devem causar impacto, interpretar e despertar sentimentos diversos no telespectador por meio da expressividade corporal, e esses objetivos são pautados pelas emissoras de televisão (GODOY-COTES, 2008; CHARAUDEAU, 2006).

Por meio de uma ampla revisão da respectiva literatura e dos trabalhos relacionados, observa-se a demanda em se realizar a análise semiodiscursiva de telejornais por meio dos recursos audiovisuais que são identificados a priori em um processo manual de cruzamento de informações. Sobre esses recursos e os respectivos dados extraídos, ocorre o trabalho dos analistas do discurso em identificar a estratégica de comunicabilidade subentendida naqueles objetos de estudo e extrair informações relevantes sobre o uso de certos planos de enquadramento da câmera em momentos específicos de intensidade de determinadas emoções que predominam sobre a temática abordada. Neste contexto, este artigo se debruça na implementação de um processo automático da extração desses recursos, em especial, as configurações referentes às modulações de intensidade em sinais de áudio e as expressões faciais, para o reconhecimento multimodal de emoções em vídeos a fim de subsidiar a análise semiodiscursiva dos níveis de tensão em programas de telejornais.

2 CARACTERIZAÇÃO DO PROBLEMA

No estudo sobre a mídia televisiva, percebe-se que o problema da tensão está na percepção semântica do telespectador sobre o fato, ou seja, na instância de percepção frente à instância de produção e, dessa forma, trata-se de um problema subjetivo a ser modelado. Uma imagem em *Close-up*, por exemplo, sugere uma tensão maior para uma determinada expressão facial, porque pretende-se envolver mais o espectador, mas não se pode afirmar se isso realmente ocorreu. Dessa forma, os esforços da Análise do Discurso concentram-se sobre as instâncias de produção, na busca por padrões que caracterizem os gêneros televisivos, assim como a identificação de efeitos que visam captar e fidelizar o seu interlocutor (CHARAUDEAU, 2006).

Dessa forma, as mídias acabam não sendo uma instância de poder, e que elas manipulam os indivíduos tanto quanto manipulam a si próprias, não transmitindo o que ocorre na realidade social. Os telejornais, inseridos nesse contexto como veículos de informação pautados, estruturalmente, sob o discurso midiático que leva a um processo de espetacularização do fato, ao relatarem um acontecimento, tendem a construir uma representação que toma lugar da realidade (CHARAUDEAU, 2006). Produzir o telejornal a fim de fidelizar o telespectador, promover o ato de informar o fato com legitimidade e, ainda, conduzir o uso de recursos não-verbais e verbo-visuais que não prejudiquem a intencionalidade comunicativa do programa, bem como da própria rede de televisão, é um desafio para os profissionais do telejornalismo e uma valiosa fonte de pesquisa para os analistas do discurso. Acredita-se que determinar os níveis de tensão em telejornais seja uma contribuição relevante para essas questões, servindo de suporte às equipes de redação na construção da imagem do programa, no sequenciamento assertivo das notícias e na descoberta de padrões, bem como aos pesquisadores em seus estudos sobre o telejornalismo brasileiro.

Dentre os padrões de estudos encontrados na literatura, analisam-se as expressões faciais dos apresentadores, as modulações da fala, os movimentos gestuais, a intensidade visual nos espaços do telejornal e os enquadramentos da câmera. Desses padrões, este artigo propõe-se a automatizar os processos de extração de características audiovisuais, de reconhecimento de modulações em sinais de áudio e de inferência de emoções sobre expressões faciais,

considerando a intensidade da ocorrência dessas características multimodais e o plano de enquadramento da câmera associado para a geração de gráficos que servirão como uma fonte de dados auxiliar para o trabalho do analista do discurso. Os modelos devem ser robustos em relação à configuração dos recursos audiovisuais dos espaços de informação que incluem os ambientes mais controlados, tais como o estúdio do programa, e os ambientes externos que, geralmente, formam o espaço dinâmico das reportagens com diversas composições de planos fílmicos na cobertura das imagens.

Sobre esses espaços, os planos fílmicos referentes ao enquadramento da câmera são bastante explorados com o intuito de conduzir alguma estratégia de comunicação por parte do telejornal. Segundo Hernandes (2006), quanto mais próximo do enquadramento em *Close-up*, maior foco e intensidade são dados àquela imagem, ressaltando o locutor e dissolvendo o espaço. Em contrapartida, quando mais próximo do Plano Geral, o espaço é ressaltado e dissolve-se o locutor. A identificação do plano fílmico pode ser elaborada por meio da proporção da face em relação ao quadro corrente do vídeo e esse valor de proporção pode ser usado para refinar a intensidade ou extensidade da emoção inferida por meio de uma expressão facial.

Para um processo de reconhecimento de expressões faciais assertivo, deve-se mapear devidamente as respectivas emoções associadas e o nível de granularidade a ser empregado. Uma etapa crucial para a tarefa de reconhecimento de expressões faciais é a detecção de faces dos indivíduos nos vídeos, pois os indivíduos se movem o tempo todo nas cenas, mesmo que os movimentos, em alguns momentos, sejam inexpressivos. Foram encontrados diversos trabalhos na literatura que propunham grupos de expressões faciais em torno de seis emoções básicas, quais sejam: raiva, medo, aversão, surpresa, alegria e tristeza (BETTADAPURA, 2009; EKMAN; FRIESEN, 1978).

3 TRABALHOS RELACIONADOS

No âmbito da Análise do Discurso, tem-se alguns trabalhos na literatura que se debruçaram em estudar a tensão e a sequência das temáticas nas notícias apresentadas em telejornais durante o tempo de exibição de um bloco ou durante todo o programa, especificamente em Braighi-Andrade (2013), Uribe e Gunter (2007), David-Silva (2005) e Mundorf *et al*. (1990), bem como o comportamento estrutural do

telejornal ou do próprio apresentador perante as notícias por meio do estudo de sua comunicabilidade não-verbal e o uso intencional dos recursos audiovisuais sobre os espaços de informação que, além do ato de informar, caracteriza uma dramatização que legitima o sentimento sobre aquele conteúdo, conforme Gutmann (2012), Pimentel (2009), Godoy-Cotes (2008) e Fechine (2008).

No experimento elaborado por Mundorf *et al*. (1990), alguns entrevistados, de ambos os sexos, foram expostos a algumas exibições de um noticiário de televisão. Para cada notícia assistida, ora com conteúdo emocional perturbador, ora com afetividade neutra, os entrevistados assistiram a outra sequência de notícias. Foi observado que a capacidade da pessoa em adquirir informação após uma notícia perturbadora ficou pobre durante 3 minutos, ou seja, depois de assistir a uma notícia com conteúdo emocional perturbador, os entrevistados ficaram, em média, 3 minutos sem entender, efetivamente, o conteúdo da próxima notícia. O prejuízo aparente da aquisição de informação, processamento, armazenamento e recuperação após notícias carregadas de emoção é discutido em termos da Teoria da Emoção.

No estudo sobre a ordenação das notícias em telejornais quanto às temáticas, David-Silva (2005) apresenta certo padrão em relação ao nível de tensão do assunto tratado, independente da temática em que o assunto foi classificado. Este trabalho analisou quatro telejornais, dois brasileiros e dois franceses, encontrando semelhanças entre eles quanto às temáticas abordadas e ao sequenciamento das notícias, tendo-se a tendência de ir de um ponto máximo de tensão, normalmente sobre notícias que demonstram a desordem do mundo, para certa sensação de leveza ao abordar notícias de conteúdo esportivo, de lazer, dentre outros assuntos. Com vasto levantamento das notícias, em diversas datas e extensos tempos de exibição, a autora modelou três níveis de tensão com base no conteúdo temático e de repercussão do sentimento envolvido: *Distenção*, *Tensão Moderada* e *Alta Tensão*.

O artigo de Uribe e Gunter (2007) analisa se as notícias sensacionalistas são intrinsecamente mais propensas a provocar respostas emocionais no público do que outras notícias de televisão. A pesquisa analisa uma amostra de notícias de telejornais britânicos para identificar nos respectivos conteúdos a presença de determinados elementos sobre os quais o público entrevistado apontou possuírem a capacidade de provocar emoção. Os resultados

mostraram que as notícias relacionadas a crimes (as mais frequentes em categorias de notícias sensacionalistas) e, de forma limitada, as notícias de temáticas políticas (tipo clássico não-sensacionalista) fornecem claras manifestações da presença de atributos carregados de alta e baixa tensões emocionais.

A pesquisa realizada por Fechine (2008) revisa e elabora estudos que tomem a tensão dos programas jornalísticos como fortes indicadores dos *ethos* dos espaços de enunciação, ou seja, qual a postura do apresentador ao informar com o intuito de fazer valer a legitimidade do seu discurso sobre a verdade dita daquele conteúdo.

No estudo realizado em Godoy-Cotes (2008), os autores dedicaram-se à análise do desempenho dos apresentadores nos telejornais por meio da interjeição ou comunicação não-verbal, referindo-se à análise gestual e às rápidas expressões faciais dos jornalistas. Com isso, pode-se inferir a intencionalidade comunicativa do telejornal para aquele assunto sob os posicionamentos discursivos, que podem estar acompanhados ou não da fala do apresentador.

Esse tipo de análise foi discutido por Pimentel (2009), que investigou o telejornalismo como um ritual de linguagem sujeito a falhas, analisando a conjunção verbal-imagem durante a enunciação dos sujeitos, observando-se a tensa relação entre dispersão e coerência na sustentação do efeito da notícia. O *corpus* de estudo utilizado foi composto por vídeos da exibição de quatro telejornais veiculados na TV aberta no dia 13 de novembro de 2006: Jornal Nacional, SBT Brasil, Jornal da Band e Jornal da Record. O *corpus* foi analisado conforme o ordenamento das temáticas sobre a construção das imagens do governo Lula, tendo-se como foco compreender o telejornalismo como um ritual de linguagem em que algo falha. Foi observada a construção da notícia a partir dos lugares enunciativos dos apresentadores, repórteres e comentaristas para verificar como se produzia a desestabilização ou não do efeito informacional pela análise da não-coerência em relação aos elementos de apagamentos, silenciamentos, interdições e visibilidades da comunicação dos sujeitos nas exibições.

De modo a contribuir para a discussão sobre o tratamento audiovisual que a informação jornalística recebe, Gutmann (2012) analisa as articulações entre o uso de enquadramentos de câmera na apresentação do telejornal e os sentidos percebidos de tempo presente e de interesse do público. O trabalho identifica as apropriações de enquadramentos de câmera recorrentes em 15

telejornais do sistema brasileiro de televisão que respondem por formas audiovisuais contemporâneas do telejornalismo responsáveis por novos tipos de configurações espaço-temporais, contribuindo para o estudo desse gênero televisivo ao conceber os usos desses dispositivos audiovisuais enquanto estratégia de comunicabilidade do telejornal.

O trabalho realizado por Braighi-Andrade (2013) propõe o levantamento dos níveis de tensão considerando as unidades de discurso dos telejornais, tais como matérias, notas cobertas, notas simples, *stand ups*, entrevistas e previsões do tempo. As notícias foram classificadas por meio de três coeficientes de tensão: baixa, moderada e alta. Devido à complexidade em determinar o nível de tensão de cada unidade, visto que cada telespectador pode atribuir um peso semântico diferente às notícias, procedeu-se em classificar as notícias pela sua macrotemática.

No âmbito do reconhecimento de emoções em vídeos, considera-se que a expressão de emoções, além das modulações na fala, ocorre principalmente na face humana (EKMAN e FRIESEN, 1978; AYADIA, KAMEL e KARRAY, 2011). Os projetos a seguir contribuíram de forma significativa para as pesquisas na área e se basearam no uso de técnicas robustas para o reconhecimento de vestígios emocionais na fala e nas expressões faciais.

Em Ekman e Friesen (1978), os autores demonstraram evidências de que as expressões faciais de emoções podem ser inferidas por sinais rápidos da face. Estes sinais são caracterizados por mudanças na aparência da face que duram segundos ou frações de segundo, uns mais visíveis que outros. Com isso, os autores formularam o modelo de emoções básicas, fundamentado sobre seis expressões faciais (raiva, medo, repulsa, surpresa, alegria e tristeza) que são encontradas em diversas culturas e são exibidas da mesma forma, desde crianças até idosos. Foram mapeados os pontos de singularidade de cada tipo de expressão facial com base em testes realizados sobre um vasto banco de imagens, gerando um importante modelo usado por diversos trabalhos.

Como as expressões faciais podem ser expressas de forma diferente por pessoas diferentes, resultados imprecisos são inevitáveis. Para evitar esses problemas, Chang e Huang (2010) implementaram um arcabouço baseado nas características visuais individuais das faces em vez de representar as expressões faciais por meio de modelos generalizados como feito em outros trabalhos. Foi

utilizada uma rede neural específica para a etapa de classificação de emoções em rostos neutros, felizes, irritados, surpresos, tristes, com medo e com nojo.

No trabalho de Ji e Idrissi (2012), os autores abordam as máquinas de análise de expressões faciais como um dos problemas mais desafiadores na área de Interação Homem-Máquina (IHM) e que as expressões faciais dependem de movimentos sutis dos músculos faciais para mostrar estados emocionais. O trabalho realiza um estudo sobre as relações entre as expressões básicas e os modelos correspondentes de deformação facial, propondo dois novos métodos para descrever a transformação do rosto humano durante expressões faciais.

No âmbito do reconhecimento de modulações da fala em sinais de áudio, os autores de Florian *et al*. (2013) apresentaram o desenvolvimento do openSMILE, um arcabouço para extração de características emocionais em discurso, música e sons em geral existentes em vídeos e em sinais de áudio. Os descritores de vídeo e de áudio podem ser processados em conjunto, em um único quadro, permitindo a sincronização de tempo dos parâmetros de extração. A detecção de atividade de voz e a detecção de face também são recursos oferecidos pelo arcabouço.

Em um contexto com diversos esforços no desenvolvimento de sistemas para identificar o conteúdo emocional de sinais de voz, o trabalho de Ayadia, Kamel e Karray (2011) realiza uma pesquisa sobre a classificação de expressões emocionais abordando três aspectos importantes para um sistema de reconhecimento de emoção na fala: (i) a escolha de características adequadas para a representação de voz, (ii) a concepção de um sistema adequado de classificação e (iii) e a preparação de um banco de dados contendo faixas de discurso emocionado para avaliar o sistema, contribuindo com discussões sobre o desempenho de sistemas de reconhecimento desse tipo, bem como as suas limitações no campo de processamento.

Para o reconhecimento de emoções em vídeos, especificamente vídeos de telejornais, é de extrema importância a aplicação de técnicas robustas de reconhecimento de modulações na fala e de expressões faciais a fim de evitar perdas significativas da carga afetiva associada ao conteúdo informacional da notícia e da sua significância semiodiscursiva. Acredita-se que a linguagem verbo-visual adotada na esfera telejornalística, incluindo as modulações

da fala, escolhas de cores, iluminação e enquadramento da câmera, acompanham a expressividade emocional dos comunicadores, mesmo que suave, no processo de espetacularização da notícia para emergir e persuadir o telespectador sobre o conteúdo enunciado, ou seja, se a emoção ao assistir o telejornal for verdadeira, então a informação é verdadeira (FLAUSINO, 2003).

4 DETERMINAÇÃO AUTOMÁTICA DE NÍVEIS DE TENSÃO

Esta seção apresenta a metodologia geral usada ao longo do desenvolvimento deste trabalho, incluindo o nível conceitual com a elaboração do modelo proposto e a implementação dos protótipos de processamento do arcabouço.

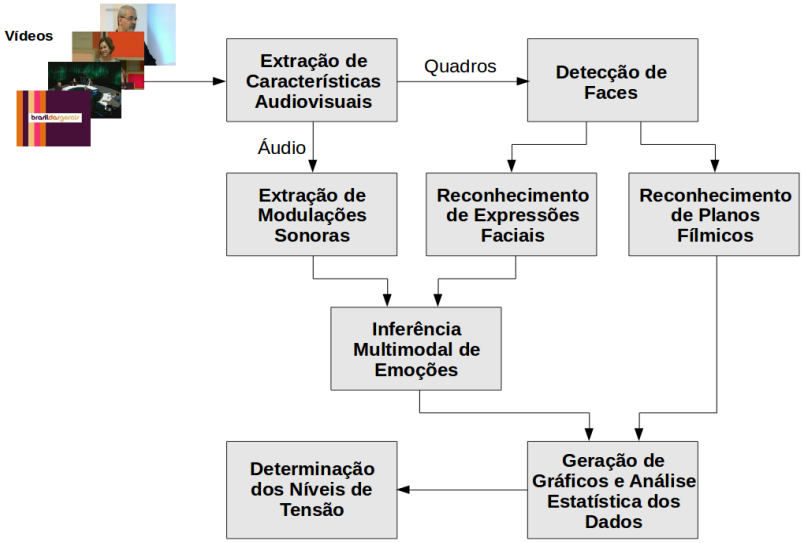
Com o intuito de compreender de forma precisa as dinâmicas de análise semiodiscursiva dos telejornais e dos níveis de tensão a serem percebidos em suas unidades constitutivas de informação, bem como para ilustrar a potencialidade de extração do sistema computacional desenvolvido, detectou-se a necessidade de se realizar este trabalho utilizando-se, como referencial, um vídeo armazenado na base de dados do Centro de Apoio a Pesquisas sobre Televisão (CAPTE) do CEFET-MG onde diversas pesquisas foram realizadas, tais como os trabalhos de Pereira *et al*. (2015), Souza *et al*. (2014), Braighi-Andrade (2013), Jacob (2013) e Conceição (2012). Esse vídeo apresenta uma reportagem exibida no Jornal Minas em 08 de Agosto de 2011, na emissora Rede Minas, sobre o aumento dos casos de violência contra o idoso.

Para a etapa de reconhecimento multimodal de emoções, este artigo se baseia no reconhecimento de emoções em sinais de áudio proposto por Florian *et al*. (2013) e no reconhecimento de expressões faciais, conforme as abordagens descritas em Littlewort *et al*. (2004) e Bartlett *et al*. (2006), utilizando-se o Sistema de Codificação de Ação Facial (FACS) proposto por Ekman & Friesen (1978) em um detector baseado no algoritmo Haartraining de Viola & Jones (2001), alcançando-se uma acurácia de 93%. Os parâmetros semiodiscursivos propostos no trabalho de David-Silva (2005) são utilizados para classificar os vídeos nos níveis de *Distensão*, *Tensão Moderada* e *Alta Tensão*, conforme a intensidade da emoção inferida por meio das expressões faciais, o enquadramento da câmera em que a respectiva face está submetida

e os valores de intensidade sonora detectados nos sinais de áudio. Nas exibições televisivas sob *Distensão* (DT), tem-se a percepção de que a condução do discurso em relação à temática nos remete ao campo semântico da "alegria", tais como eventos esportivos, comemorações, dicas de culinária, dentre outros, provocando uma espécie de alívio no telespectador. Nas exibições sob *Tensão Moderada* (TM), as notícias promovem uma implicação patêmica no telespectador, ainda que suportável por existir certa distância do cotidiano do público do local da notícia, dos envolvidos ou da própria temática abordada. As matérias com níveis de *Alta Tensão* (AT) referem-se a conteúdos que nos remete para uma categoria de conflito, de violência, de tragédia e de morte (homicídios), revelando problemas do mundo que podem produzir uma resposta patêmica no telespectador, principalmente quando os assuntos abordados são mais próximos do público ao qual se endereçam (DAVID-SILVA, 2005). Por esses conceitos, propõe-se classificar nos níveis de *Alta Tensão* os vídeos em que mais ocorrerem emoções de medo, de raiva e de tristeza; de *Tensão Moderada* os vídeos em que predominarem as expressões faciais de aversão, surpresa e desgosto; e de *Distensão* os vídeos que apresentarem mais emoções de alegria.

A Figura 1 apresenta uma visão geral da abordagem proposta para o reconhecimento multimodal de emoções na determinação automática dos níveis de tensão em vídeos de telejornais. Algoritmos para extração de características audiovisuais foram aplicados a fim de fornecer os recursos necessários para a detecção das faces dos indivíduos atuantes nos vídeos. Sobre as faces detectadas, atuam os módulos para o reconhecimento de expressões faciais e a identificação de planos fílmicos, e sobre o sinal de áudio do vídeo, atua o módulo de reconhecimento de modulações da fala. O reconhecimento multimodal de emoções combina os dados das características reconhecidas sobre o sinal de áudio e as expressões faciais. Enfim, as emoções inferidas são apresentadas em gráficos que possibilitam analisar os dados associados a fim de aplicar métricas para determinar os níveis de tensão dos vídeos de telejornais e apoiar as pesquisas em semiótica do analista do discurso.

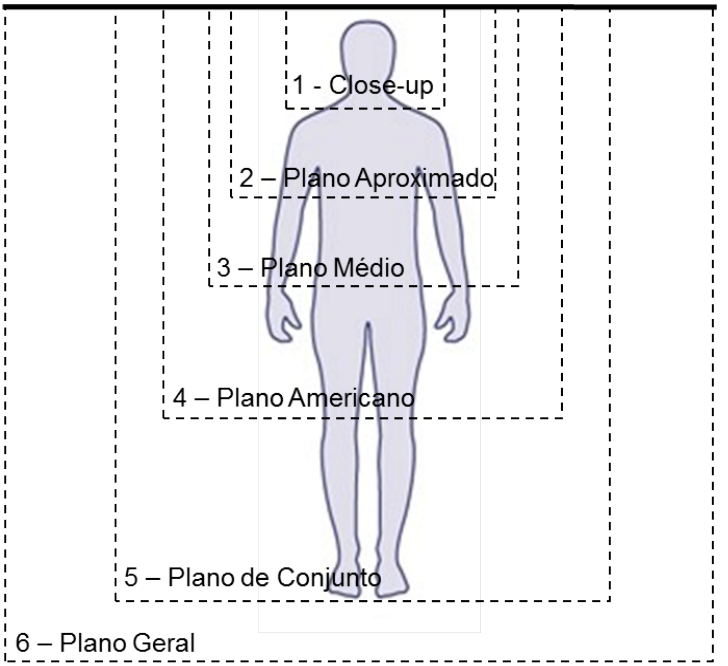
Figura 1 – Modelo proposto para determinação automática dos níveis de tensão em telejornais.



Fonte: Elaborado pelos autores.

Durante o processo de reconhecimento das expressões faciais, o plano fílmico referente ao enquadramento da câmera é identificado conforme a abordagem de Conceição (2013), calculando-se a proporção da área da face em relação à dimensão do respectivo quadro do vídeo em que ela aparece. A influência dos planos fílmicos nos níveis de tensão pode ser atribuída à intenção de se explorar a intensidade ou apreensão sobre uma determinada expressão facial. Na Figura 2, percebe-se que os planos são classificados conforme o tamanho da figura humana dentro do quadro. Assim, os planos podem ser compreendidos como imagem capturada e enquadrada por uma câmera.

Figura 2 – Planos de enquadramento da câmera.



Fonte: Elaborado pelos autores.

Tabela 1 – Proporções da face nos planos fílmicos.

Código	Plano Fílmico	Proporção (α)	Acurácia
1	Geral	$\alpha \leq 0.12$	0.88
2	Conjunto	$0.12 < \alpha \leq 0.19$	0.84
3	Americano	$0.19 < \alpha \leq 0.22$	0.82
4	Médio	$0.22 < \alpha \leq 0.28$	0.60
5	Aproximado	$0.28 < \alpha \leq 0.40$	0.85
6	Close-up	$\alpha > 0.40$	0.95

Fonte: Adaptado de Conceição (2013).

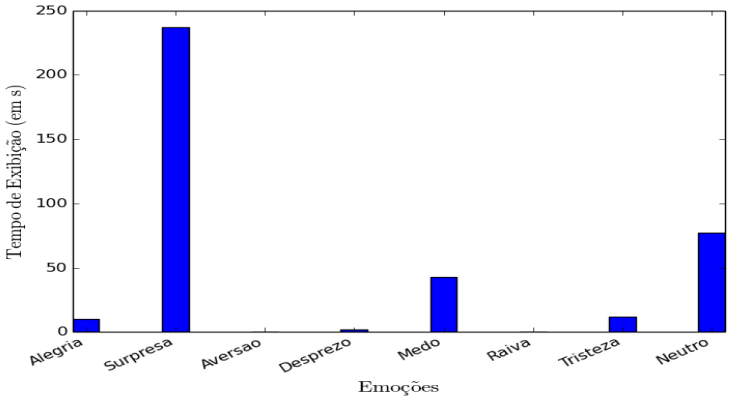
A Tabela 1 apresenta os valores de proporção da face utilizados neste trabalho para a identificação dos planos fílmicos de enquadramento da câmera seguindo as medições realizadas para o universo telejornalístico. Assim, a determinação automática dos pontos de leitura, isto é, pontos de marcação de estruturas de

controle como olhos, boca e nariz, ocorre logo após a etapa inicial de detecção automática da face e das regiões de interesse. Como o modelo necessita de uma face neutra de expressões faciais para que, sobre ela, os movimentos faciais possam ser lidos, definiu-se neste trabalho marcar como neutra a primeira face detectada no vídeo, partindo-se da premissa de imparcialidade emotiva em que os apresentadores devem iniciar e se manter.

Para o processo de extração de características de modulações em sinais de áudio, foi utilizado o arcabouço openSMILE proposto por Florian *et al*. (2013). Nessa etapa do modelo, o arcabouço utiliza o espectro do componente de áudio do vídeo de uma determinada reportagem para extrair as características de intensidade sonora e a frequência fundamental das modulações desses sinais. A intensidade sonora reflete o percentual de percepção da amplitude da onda sonora pelo ouvido humano medida em decibéis (dB). Já a frequência fundamental corresponde ao primeiro harmónico de uma onda sonora, sendo a frequência mais influente na percepção de um determinado som. No caso da voz humana, tais valores variam conforme a idade e o sexo, estando entre 85 e 180 Hz para os homens, e entre 165 e 255 Hz para as mulheres. É um dos principais elementos caracterizadores da voz (FLORIAN *et al.*, 2013; PEREIRA *et al*., 2009; PEETERS, 2006). Essa etapa é muito importante para a análise dos dados referentes às modulações no discurso do telejornal que podem influenciar o ritmo na locução de notícias, incluindo os aspectos semânticos das estruturas emotivo-verbais que devem ser testadas quanto à eficácia da transmissão de informação (MACHON, 2012).

A próxima etapa é a análise estatística simplificada dos dados obtidos e a apresentação desses dados em gráficos para possibilitar a aplicação de alguma métrica, bem como auxiliar as pesquisas em análise dos níveis de tensão dos programas telejornalísticos. Neste trabalho, a visualização dos dados extraídos inicia-se pelo histograma referente ao tempo de exibição das emoções no vídeo, isto é, um modelo de gráfico que retrata a frequência de ocorrência de emoções no vídeo, informando o tempo de exibição, em segundos, em que cada emoção foi inferida. Em seguida, tem-se os gráficos de cada um dos indicadores audiovisuais e os níveis de tensão identificados.

Figura 3 – Frequência do tempo de exibição das emoções no vídeo.



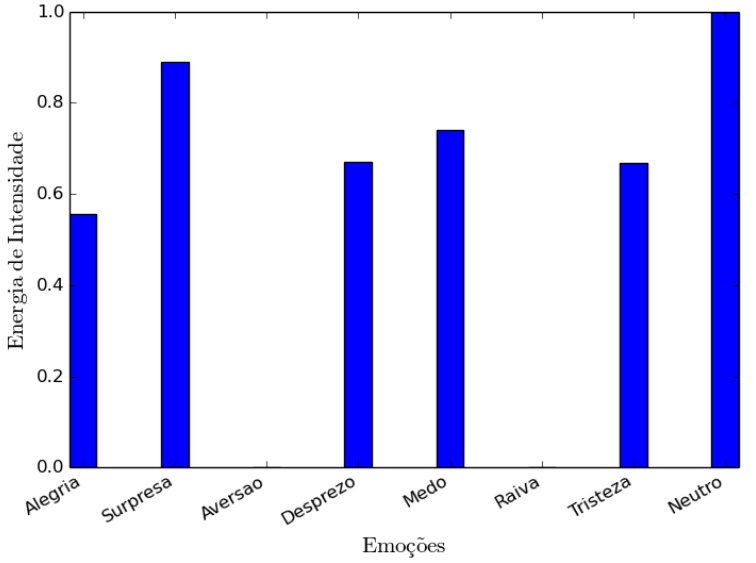
Fonte: Elaborado pelos autores.

A Figura 3 ilustra o gráfico de ocorrências de emoções pelo tempo de exibição para a reportagem sobre o aumento dos casos de violência contra idosos. O eixo x contempla as emoções modeladas (Alegria, Surpresa, Aversão, Desprezo, Medo, Raiva, Tristeza, Neutro) e no eixo y as ocorrências, em segundos, de cada emoção. Tem-se, assim, 11 segundos referentes à emoção de Alegria (*Distensão*), 236 segundos de *Tensão Moderada* e 56 segundos de *Alta Tensão* 372 segundos da exibição desse tema, configurando uma tendência significativa de ser um vídeo de *Tensão Moderada*. Dessa forma, a fim de verificar essa tendência, que pode ou não ser refutada, faz-se necessário analisar a influência dos fatores dos demais recursos audiovisuais que, sob um olhar dos conceitos computacionais, porém não restrito a eles, compõem a dinâmica verbal-visual na exibição de uma reportagem.

A Figura 4 apresenta os gráficos referentes aos valores de intensidade visual no reconhecimento das emoções ao longo da reportagem referente ao nível de percepção da expressividade. Desconsiderando-se as aparições de faces neutras de emoções, observa-se que o processo de reconhecimento da emoção *Surpresa* tem um nível de intensidade considerável em relação às outras emoções, mas com um nível médio de expressividade próximo à ocorrência da face associada ao medo, mesmo que essa emoção tenha tido pouco tempo de exibição no processo de inferência, sugerindo um arranjo visual de narrativa heterogênea.

Isso é esperado, de fato, visto que a emoção de medo possui uma expressividade mais marcante e intensa. O nível de surpresa se concentra por volta de 130 segundos da reportagem até cerca de 210 segundos. Nesse momento do vídeo, o espaço de informação é interno, no estúdio do telejornal, para exibição da entrevista com o profissional Felipe Willer, presidente do Conselho Estadual do Idoso, em que ocorre certa expressividade desse participante em uma intenção mais enérgica de se fazer entender e em esclarecer os direitos do idoso.

Figura 4 – Intensidade visual no reconhecimento das expressões faciais.

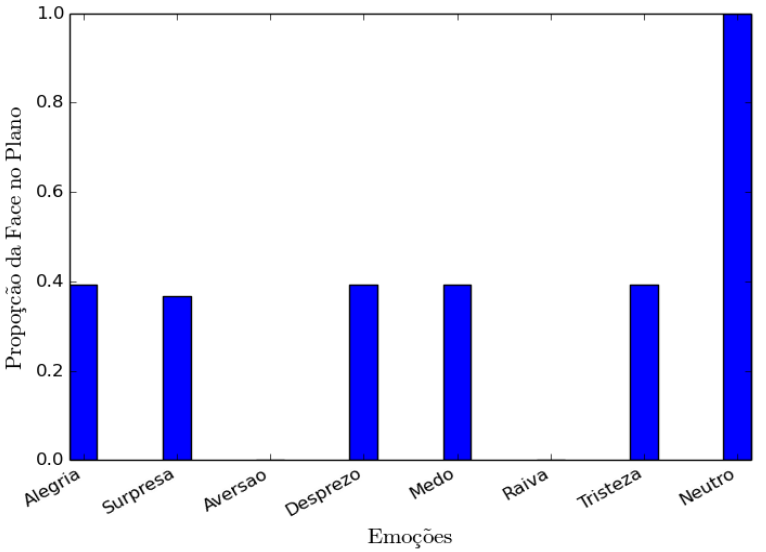


Fonte: Elaborado pelos autores.

A Figura 5 apresenta os gráficos referentes aos valores de intensidade visual das proporções de uma face detectada no vídeo, a emoção correspondente reconhecida e o tipo de plano fílmico a qual essa face estava submetida na reportagem. De acordo com os códigos descritos na Tabela 1, percebe-se que a maioria das emoções foi detectada nos enquadramentos do plano Americano e de Conjunto, mesmo que boa parte do telejornal tenha ocorrido mais o enquadramento Médio. Conforme os estudos

de Braighi-Andrade (2013) sobre os telejornais mineiros, o Jornal Minas possui um formato mais conservador no que se refere à condução visual dos recursos de planos da câmera. Para o caso específico dessa reportagem, ocorreu neutralidade de expressões faciais em momentos que deveriam refletir mais intensidade emocional. Essa situação inesperada é facilmente explicada quando a reportagem se voltou para um espaço externo para recolher o depoimento de um rapaz sobre as preocupações em cuidar de sua mãe que possui o Mal de Alzheimer e, dessa forma, parecia alheia à câmera e não esboçou expressividade facial nesse enquadramento. Ocorreram enquadramentos do Plano Aproximado, quase ocorrendo o Plano *Close-up*, ao filmar a senhora doente, ilustrando uma tentativa de patemização do público a se sensibilizar com a história.

Figura 5 – Tendência da proporção das faces nos enquadramentos de câmera.



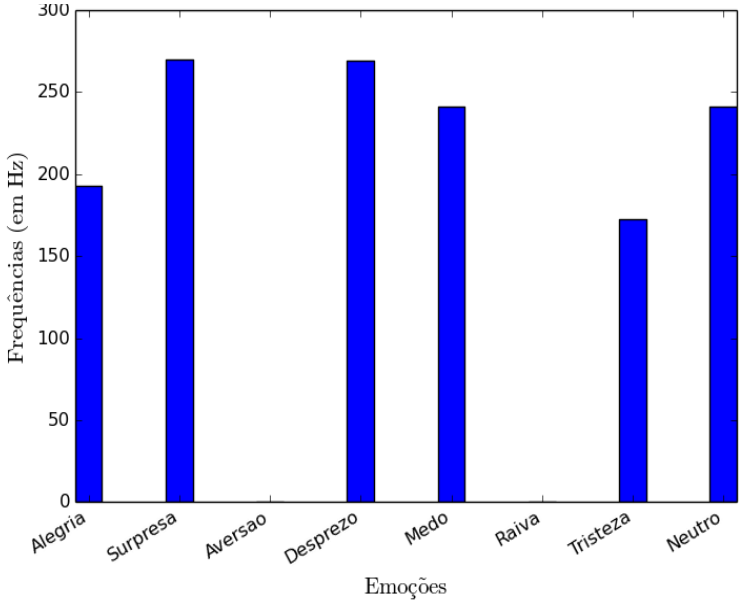
Fonte: Elaborado pelos autores.

A Figura 6 apresenta a distribuição de frequências dos espectros harmônicos sobre a intensidade sonora no processo da fala dentro do telejornal. Mesmo não condizentes com as respectivas expressões faciais que foram anotadas durante a exibição

do telejornal, o método computacional empregado detectou disparidades da modulação das vozes que refletiram de forma mais intensa algumas das emoções que não tiveram a mesma correspondência emotiva no processo visual. O importante desse gráfico é a contribuição considerável que promove sobre os outros recursos audiovisuais um ajuste de ponderação importante, pois no gráfico de tempo de exibição, a emoção de desprezo inferida pela expressão facial quase não apareceu ao se considerar todo o tempo do telejornal, porém foi mais intensa em sua respectiva modulação vocal.

Com o tratamento de todos os valores numéricos de equivalência apresentados visualmente, foi gerado um gráfico agrupando todos os dados de intensidade visual do reconhecimento de expressões faciais, planos fílmicos de enquadramento da câmera, modulação da frequência fundamental harmônica da fala e a intensidade sonora dessa frequência em uma soma ponderada sobre o tempo de exibição das emoções inferidas ao longo do vídeo.

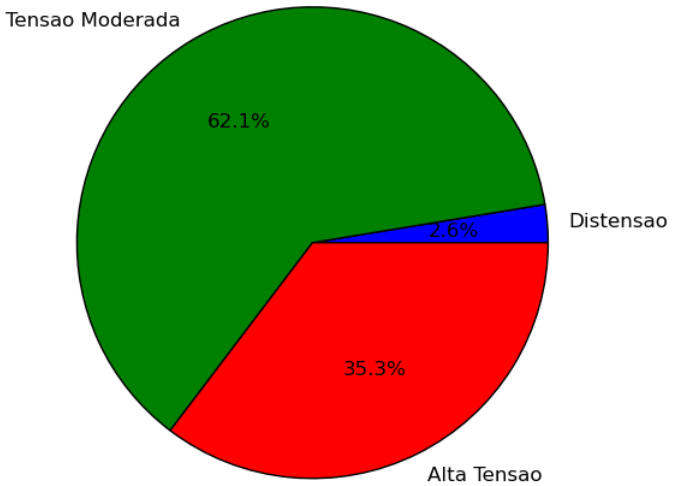
Figura 6 – Frequência fundamental na intensidade sonora das vozes detectadas.



Fonte: Elaborado pelos autores.

A Figura 7 apresenta o gráfico para a frequência do tempo de exibição dos níveis de tensão, considerando-se os tempos de exibição de todos os recursos audiovisuais do respectivo vídeo, apresentados pelos quadros, sob uma taxa de 29 quadros por segundo, e pelos cortes de áudio em centésimos de segundo. Para essa reportagem, tem-se 62,1% de inferência de emoções relativas à *Tensão Moderada*, 35,3% de *Alta Tensão* e apenas 2,6% de *Distensão*, condizente com a distribuição da inferência de emoções sobre as expressões faciais reconhecidas no primeiro gráfico.

Figura 7 – Gráfico para a frequência do tempo de exibição dos níveis de tensão.



Fonte: Elaborado pelos autores.

Na conjunção de todos os recursos audiovisuais analisados, pode-se afirmar que, considerando-se a porcentagem estatística das acurácias de cada método computacional empregado, a reportagem em questão possui nível de tensão predominantemente regular, ou seja, de Tensão Moderada, refletindo com a classificação proposta por David-Silva (2005), visto que a reportagem trata a temática sobre o aumento dos casos de violência contra idosos de forma educativa, apresentando os meios

de suporte e apoio para as pessoas que se encontram nesse contexto de história, ou seja, é um assunto que incomoda, mas o nível de desordem do mundo é moderado.

5 CONCLUSÃO

Este trabalho apresenta uma nova abordagem que integra métodos computacionais para a extração de recursos audiovisuais no reconhecimento de modulações sonoras da fala e de emoções em expressões faciais a fim de identificar e classificar os níveis de tensão em vídeos de telejornais como abordagem interdisciplinar e complementar à análise semiodiscursiva desses programas televisivos.

As expressões faciais são formas de comunicação não-verbal amplamente utilizadas em nosso cotidiano e podem promover estudos promissores sobre o universo da construção do *ethos* dos telejornais quando associadas a outros parâmetros da análise do discurso das mídias televisivas, tais como os planos fílmicos de enquadramento da câmera, os modos enunciativos, os eixos de visão, a disposição dos participantes, dentre outros. Na abordagem proposta neste trabalho, utilizou-se a proporção da face detectada em relação ao quadro, sob os valores pré-determinados no estudo de planos fílmicos, como um dos indicadores de ponderação sobre a intensidade da emoção reconhecida naquele quadro. Essa combinação pode ser foco de estudos mais específicos para determinar, por exemplo, se a intensidade de uma determinada emoção, dependendo do nível de tensão em que ela foi categorizada, estabelece, com legitimidade, a proximidade do telejornal em relação à audiência, que pode ser evidenciada pela postura do apresentador sob o plano *Close-up*. Além disso, os movimentos de aproximação ou distanciamento físicos com o interlocutor que busca cumplicidade em relação ao enunciado são modalizadores discursivos para a produção de sentido perante uma emoção intencionalmente pautada na estratégia comunicacional.

As modulações da fala e a expressividade facial dos repórteres e locutores, principalmente dos apresentadores, podem fornecer indícios sobre a tensão do discurso gerado

pela enunciação da respectiva reportagem e o padrão no sequenciamento das notícias informadas nas instâncias de produção desses objetos informacionais. Para trabalhos futuros, pretende-se aplicar o modelo proposto e identificar os níveis de tensão de cada notícia durante a exibição de diversos telejornais, auxiliando nos estudos sobre o padrão do sequenciamento das notícias na identidade do programa. Sobre o discurso verbal, este trabalho sugere a extração do *Closed Caption* dos vídeos de telejornais para aplicar técnicas de análise de sentimentos (PANG and LEE, 2008) com o propósito de alcançar maior assertividade na determinação dos níveis de tensão. Ao combinar essas técnicas em uma abordagem de análise multimodal, pode-se obter as flutuações das polaridades de sentimentos ao longo da enunciação de cada notícia e estudar a correlação dessas com as modulações sonoras da fala nos sinais de áudio para subsidiar pesquisas inovadoras na semiótica e análise do discurso verbo-visual de telejornais.

REFERÊNCIAS

AYADIA, M. E.; KAMEL, M. S.; KARRAY, F.. Survey On Speech Emotion Recognition: Features, Classification Schemes and Databases. **Pattern Recognition**, v. 44, n. 3, 2011, p. 572-587. Disponível em: http://www.sciencedirect.com/science/article/pii/S0031320310004619. Acesso em: 19 fev. 2015.

BARTLETT, M. S., et al. Fully Automatic Facial Action Recognition in Spontaneous Behavior. 7th International Conference on Automatic Face and Gesture Recognition - FGR 2006, p. 223-230, **IEEE**, 2006.

BETTADAPURA, V. Face Expression Recognition and Analysis: The State of the Art. **ArXiv e-prints**, Março 2009, p. 1-27. Disponível em: http://arxiv.org/pdf/1203.6722.pdf. Acesso em: 05 jan. 2015.

BRAIGHI-ANDRADE, A. A. **Análise de Telejornais:** um Modelo de Exame da Apresentação e Estrutura de Noticiários Televisivos. Rio de Janeiro: E-Papers, 2013.

CHARAUDEAU, P. **Discurso das Mídias**. São Paulo: Contexto, 2006.

CHARAUDEAU, P.; GHIGLIONE, R.. **A Palavra Confiscada:** um Gênero Televisivo: o Talk Show. Instituto Piaget, Lisboa, 1997.

CONCEICAO, F. L. A. **Metodologia Baseada em Mineração de Dados**

para Apoio à Análise do Discurso de Telejornais. Dissertação (Mestrado em Modelagem Matemática e Computacional) - Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, Agosto 2013.

DAVID-SILVA, G. A **Informação Televisiva: uma Encenação da Realidade (Comparação entre Telejornais Brasileiros e Franceses)**. Tese (Doutorado em Estudos Linguísticos) - Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

EKMAN, P.; FRIESEN, W. **Facial Action Coding System (FACS): Manual**. Consulting Psychologists Press, Palo Alto, 1978.

FECHINE, Y. Performance dos Apresentadores dos Telejornais: a Construção do Ethos. **Revista FAMECOS - Mídia, Cultura e Tecnologia**, Porto Alegre, v. 1, n. 36, 2008, p. 69-76. Disponível em: http://revistaseletronicas.pucrs.br/ojs/index.php/revistafamecos/article/view/4417/3317. Acesso em: 03 dec. 2014.

FLAUSINO, C. V. Choro Gratuito: a Violência no Telejornalismo Brasileiro. **CBCC'03**. Anais do XXVI Congresso Brasileiro de Ciências da Comunicação. São Paulo, 2003

FLORIAN E., et al. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In: **ACM Multimedia (MM)**, Barcelona, October 2013. Proceedings of the 21st ACM international conference on Multimedia, p. 835-838.

GAGE, D. L.; MEYER, C. **O Filme Publicitário**. 2. ed. São Paulo: Atlas, 1991.

GODOY-COTES, C. S. **O Estudo dos Gestos Vocais e Corporais no Telejornalismo Brasileiro**. Tese - Pontifícia Universidade Católica de São Paulo, Sao Paulo, 2008.

GOFFMAN, E. The lecture. **Forms of talk**. Pennsylvania, University of Pennsylvania Press, 1981, p. 162-195.

GUTMANN, J. F. O Que Dizem os Enquadramentos de Câmera no Telejornal? Um Olhar sobre Formas Audiovisuais Contemporâneas do Jornalismo. **Brazilian Journalism Research**, v. 8, 2012, p. 64-79. Disponível em: http://bjr.sbpjor.org.br/bjr/article/view/422. Acesso em: 20/02/2015.

HERNANDES, N.. **A Mídia e seus Truques:** o que Jornal, Revista, TV, Rádio e Internet Fazem para Captar e Manter a Atenção do Público. 1. ed. São Paulo: Contexto, 2006.

JACOB, H. D. **Desenvolvimento de um Modelo de Atenção Visual para Sumarização Automática de Vídeos de Programas Televisivos**. Dissertação (Mestrado em Modelagem Matemática e Computacional) - Centro Federal de Educação Tecnológica de Minas

Gerais, Belo Horizonte, Agosto 2013.

LITTLEWORT, G, et al. Dynamics of Facial Expression Extracted Automatically from Video. **CVPRW'04**. Conference on Computer Vision and Pattern Recognition Workshop, IEEE, 2004.

MACHON, L. M.. Estrutura Rítmica na Locução de Notícias. **Brazilian Journalism Research**, v. 8, p. 8-27, 2012. Disponível em: http://bjr.sbpjor.org.br/bjr/article/view/487. Acesso em: 07 jan. 2015.

PANG, B.; LEE, L. Opinion Mining and Sentiment Analysis. **Foundations and Trends in Information Retrieval**. Hanover, January, 2008. v. 2, n. 1-2, p. 1-135.

PEETERS, G. Chroma-Based Estimation of Musical Key from Audio-Signal Analysis. **ISMIR**. International Symposium for Music Information Retrieval. Victoria, Canada, 2006.

PEREIRA, F. et al. **Comunicações Audiovisuais:** Tecnologias, Normas e Aplicações. IST Press, 1. ed. 2009.

PEREIRA, M. H. R., et al. **SAPTE: A Multimedia Information System to Support the Discourse Analysis and Information Retrieval of Television Programs**. Journal Multimedia Tools and Applications, v. 74, n. 2, 2015.

PIMENTEL, R. M. L. Memória e Apagamento no Imaginário dos Telejornais. **Discursos Fotográficos**, Londrina, v. 5, n. 6, Junho 2009, p. 13-33.

SABINO, J. L. F.; DAVID-SILVA, G.; PÁDUA, F. L. C.. AD e Eventos da Mídia: Uma Análise da Espetacularização do Conflito Verbal. **Acta Semiótica et Linguística**, Paraíba, v. 19, p. 1-15, 2014.

SOUZA, C. L. et al. A Unified Approach to Content-Based Indexing and Retrieval of Digital Videos from Television Archives. **Artificial Intelligence Research**, v. 3, p. 49-61, 2014. Disponível em: http://www.sciedu.ca/journal/index.php/air/article/view/5251. Acesso em: 15 fev. 2015.

VIOLA, P.; JONES, M. J. Rapid Object Detection using a Boosted Cascade of Simple Features. **IEEE Computer Society**. Conference on Computer Vision and Pattern Recognition, v. 1, 2001, p. 511–518. Disponível em: https://www.cs.cmu.edu/~efros/courses/LBMV07/Papers/viola-cvpr-01.pdf. Acesso em: 05 dec. 2014.

Moisés Henrique Ramos Pereira - Mestre em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG). Professor do Instituto de Engenharia e Tecnologia do Centro Universitário de Belo Horizonte (UNI-BH). moises.ramos@prof.unibh.br

Flávio Luís Cardeal Pádua - Doutor em Ciência da Computação pela Universidade Federal de Minas Gerais (UFMG). Professor do Departamento de Computação do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET -MG). cardeal@decom.cefetmg.br

Giani David Silva - Doutora em Estudos Linguísticos pela Universidade Federal de Minas Gerais (UFMG). Professora do Departamento de Linguagem e Tecnologia do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET -MG). gianids@deii.cefetmg.br

RECEBIDO EM: 01/03/2015 | ACEITO EM: 26/08/2015