ARTICLES
ARTICULES

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

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

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

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

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.

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

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.

1 INTRODUCTION

This study proposes a new approach for the discursive-semiotic analysis of TV news programs combined with computational techniques to automatically determine tension levels in these programs. The approach employs multimodal emotion recognition to determine the audiovisual characteristics of modulations in the audio signals of the video file as well as the facial expressions of the participants. It considers the effects of intensity or apprehension that occurs through the identification of the film shot of camera to which the faces are directed.

The style of these TV programs is generally built on redundancy in speech, a choice that aims to reflect a certain semantic meaning of the news to convey credibility. This unique style is employed by news presenters and reporters (GOFFMAN, 1981). The visual narrative also offers image elements that are used consistently and that have the capacity to convey emotional weight within the TV narrative using

camera framing, angles, shapes, colors, and movement (BLOCK, 2010). Thus, TV news programs are considered a type of institutional language that is ritualized in mutual relationships of authorization and legitimacy that are promoted by the subjects of discourse to convey information to the TV viewers (STAM, 1985).

The sequence of news and the articulation of TV program presenters have each been the object of several studies. Many efforts have been made to discover patterns that reveal the intended communication strategy and that are represented by news sequencing, camera framing plans in TV news, and the influence of the presenters in the construction of the discursive semiotic ethos of the programs in this genre (PEREIRA *et al.*, 2014; GUTMANN, 2012). Journalists inevitably interpret the news and thus impact and develop feelings in the TV viewer through body language. These objectives have long been established by TV broadcasters (GODOY-COTES, 2008; CHARAUDEAU, 2006).

A broad review of the literature and related works revealed the need to perform a discursive-semiotic analysis of TV news programs through audiovisual resources that are not first identified in a manual process of information validation. These resources and the corresponding extracted data represent the work to be undertaken by discourse analysts. The analyst must identify the communication strategy underlying the objects of study and extract relevant information regarding the use of certain camera framing plans at specific moments of intense emotion, which are common in the programs being studied. Thus, this report focuses on the implementation of an automatic process to extract these resources, particularly the intensity modulations in audio signals and facial expressions, to establish a multimodal recognition of emotions in videos. This data will support the discursive-semiotic analysis of tension levels in TV news programs.

2 CHARACTERIZATION OF THE PROBLEM

In the study of TV media, the problem of tension lies in the TV viewer's semantic perception of the facts, in other words, the moment of perception when facing the moment of production. In that manner, it is a subjective problem that is being modulated. For

instance, a close-up image suggests a higher tension for a certain facial expression, because one intends to draw the TV viewer closer; however, one cannot ensure that it actually happens. Therefore, the efforts of discourse analysis concentrate on the moments of production, on the search for patterns to characterize the TV genres, and on the identification of effects aiming to attract and retain viewers (CHARAUDEAU, 2006).

Therefore, the media is not an example of power, manipulating the individuals as much as they manipulate themselves, and not conveying what occurs in social reality. TV news programs, inserted in this context as vehicles of information are structurally guided by the media discourse, which leads to sensationalism of the fact. As they report an event, they tend to build a representation that replaces reality (CHARAUDEAU, 2006). To produce a TV news program in a way that retains the viewer, promotes the act of informing with legitimacy and the use of verbal and non-verbal visual resources that do not negatively affect the communicative intent of the program or the TV network itself, is a challenge for TV news program professionals and a valuable research source for discourse analysts. It is believed that determining the tension levels in TV news is a relevant task, since it serves as support for editorial teams in the construction of the program's image, in the production of more assertive news sequencing, and in finding patterns. It also helps researchers in their studies on Brazilian TV news programs.

Using the study patterns found in the literature, the presenters' facial expressions, speech modulations, and body language were analyzed along with the visual intensity of the TV news spaces and the camera framings. Based on these patterns, this manuscript proposes to automate processes, to extract audiovisual characteristics and to recognize modulations in audio signals and emotion interference on facial expressions. The process considers the intensity of these multimodal characteristics and the camera framing shots associated with image generation that will serve as an auxiliary data source for the work of the discourse analyst. The models should be robust in terms of the configuration of audiovisual resources of the information spaces that include more controlled environments, such as the program studio, and the external environments (which normally include the dynamic space of reports with diverse compositions of film shots of image coverage).

These spaces were closely evaluated. The film shots of

the camera framing are explored at length to analyze certain communication strategies employed by the TV news program. According to Hernandes (2006), the closer the Close-up framing, the larger the focus and intensity of the image. This technique highlights the presenter and dissolves the space. On the other hand, the closer the Long Shot, the more the space is highlighted and the more the presenter is dissolved. The film shot may be identified using measurements of face proportion in relation to video framing, and the proportion value may be used to enhance the intensity or extent of the emotion inferred from a facial expression.

For the assertive recognition of facial expressions, one must adequately trace the corresponding emotions associated with the granulating level to be applied. A crucial stage for the task of facial expression recognition is the detection of the individual faces in videos, as the individuals are always in motion in the scenes, even if the movements are inexpressive at times. We found several studies in the literature proposing groups of facial expressions around six basic emotions that are as follows: anger, fear, disgust, surprise, happiness, and sadness (BETTADAPURA, 2009; EKMAN and FRIESEN, 1978).

3 RELATED WORKS

In discourse analysis, some of the previous literature focuses on studying tension and the subject matters' sequence in the news presented in TV news programs during the broadcast of a block or the whole program, specifically in Braighi-Andrade (2013), Uribe and Gunter (2007), David-Silva (2005), and Mundorf *et al*. (1990). Furthermore, some studies examine the structural behavior of the TV news program or that of the presenter him/herself by studying non-verbal communication and intentional use of audiovisual resources on the information spaces that, apart from informing, characterize a dramatization that legitimates the feeling about that content according to Gutmann (2012), Pimentel (2009), Godoy-Cotes (2008), and Fechine (2008).

In the experiment performed by Mundorf *et al*. (1990), some of the interviewed individuals, of both genders, were exposed to some TV news program viewings. For each news show that was watched, whether with disturbing or with neutral content, the individuals studied or watched another news sequence. It was

observed that the individual's capacity to acquire information after viewing a disturbing news was poor for the immediate 3 min. This indicates that after watching the news with emotionally disturbing content, the interviewed individuals were unable to effectively understand the content of the next news segment for an average of 3 min. The reduction in information acquisition, processing, storage, and retrieval after emotion-laden news is discussed in terms of the Theory of Emotion.

In the study on news sequencing of subject matters in TV news programs, David-Silva (2005) presents a pattern for the tension level of the subject in question that can be used regardless of the subject matter under which the news report was classified. This study analyzed four TV news shows (two Brazilian and two French). The researchers found similarities between them in terms of their subject matters and news sequencing with a trend that went from a maximum point of tension, normally the news showing world disorder, to a certain feeling of lightness while approaching news with sport, leisure, and other content. With a variety of news studies that included several dates and different broadcasting times, the author modeled three tension levels based on the subject matter and the repercussion of the involved emotion: *Absence of Tension*, *Moderate Tension* and *High Tension*.

Uribe and Gunter's (2007) study analyzes whether sensationalized news reports are intrinsically more prone to provoking emotional responses from the audience than other TV news shows. The research analyzed a sample of British TV news to identify the presence of certain elements in the contents that the interviewed individuals indicated as the potential cause of emotion. The results showed that news on crime (the most frequent in sensationalized news categories) and, to a limited extent, political news reports (classic, non-sensationalized types) supply clear manifestations laden with high and low emotional tensions.

The research performed by Fechine (2008) reviews and elaborates studies that assume tension of news programs as strong indicators of the ethos of speech spaces or the posture of the presenter while informing to validate the legitimacy of the discourse regarding the truth conveyed about that content.

In the study by Godoy-Cotes (2008), the authors analyzed TV news presenters' performance in terms of interjection or non-verbal communication and considered the presenters' body language and minute facial expressions. Considering this, under a discursive semiotic

perspective, we may infer the communicative intentionality of the TV news program for that subject under the discursive positions, which may or may not be accompanied by the presenter's speech.

This type of analysis was discussed by Pimentel (2009), who researched TV news programs as a language ritual that is subject to flaws and analyzed the verbal-image conjunction during speech. This author verified the tense relationship between dispersion and coherence in supporting the news effect. The study corpus used comprised videos exhibited in four TV news on local broadcast in Brazil on the 13 November, 2006: *Jornal Nacional, SBT Brasil, Jornal da Band,* and *Jornal da Record*. The *corpus* was analyzed according to the subject theme sequencing about the construction of images of the Lula administration and was aimed at understanding TV news programs as a flawed language ritual. The news construction was observed from the speech places of the presenters, reporters, and commentators to verify the extent to which the destabilization of the informative effect was produced. This was performed through the analysis of deleted elements, silences, interruptions, and the visibility of communication of the subjects in the broadcast.

To contribute to the discussion of the audiovisual treatment that is applied to journalistic information, Gutmann (2012) analyzes the articulations between the use of camera framing in TV news presentations and the perceived meanings of present time and audience interest. The work identifies the choices in camera framings that are most recurrent in 15 TV news programs on Brazilian TV, translated into audiovisual forms of TV news programs. These factors are responsible for new types of space–time configurations and contribute to the study of that TV genre using these audiovisual devices as a communication strategy.

The work by Braighi-Andrade (2013) proposes a survey of tension levels considering the TV news discourse units, such as subject matters, readers, stand-up comedies, interviews, and weather forecasts. The news reports were classified as one of three tension coefficients: low, moderate, and high. Due to the complexity in determining the tension level of each unit (since each viewer may attribute a different semantic weight to news), the news reports were classified by their macro themes.

In terms of emotion recognition in videos, the expression of emotion is considered to mainly occur both in speech modulations and on human face (EKMAN & FRIESEN, 1978; AYADIA, KAMEL, & KARRAY,

2011). The following projects significantly contributed to research in the field and were based on the use of robust techniques for the recognition of emotional traces in speech and facial expressions.

Ekman and Friesen (1978) provided evidence that facial expressions may be inferred from rapid changes in facial signs. These signs are characterized by changes in appearance that last for mere seconds or fractions of a second, some of which are more visible than others. Hence, the authors formulated the model of basic emotions grounded on six facial expressions (anger, fear, disgust, surprise, happiness, and sadness) found in several cultures to be exhibited in the same way by everyone from children to the elderly. Singularity points were traced for each type of facial expression based on tests performed with a wide image bank, thus creating an important model used by diverse works.

Because facial expressions may be expressed differently by different people, inaccurate results are unavoidable. To solve this problem, Chang and Huang (2010) implemented a framework based on the individual visual characteristics of faces instead of representing facial expressions through generalized models, similar to previous works. A specific neural network was used for the stage of emotion classification on neutral, happy, angry, surprised, sad, afraid, and disgusted faces.

In the work by Ji and Idrissi (2012), the authors cite analysis machines for facial expressions as one of the most challenging problems in the field of Human-Machine Interaction. They report that facial expressions depend on subtle movements of facial muscles to show emotional states. The work includes a study on the relationships between basic expressions and corresponding models of facial changes, and it proposes two new methods to describe the transformation of the human face during facial expressions.

Regarding the recognition of speech modulations in audio signals, Florian *et al*. (2013) presented the development of openSMILE, a framework for the extraction of emotional characteristics in discourse, music, and sounds generally present in videos and audio signals. The video and audio descriptors may be processed together, in a single framework, allowing for the time synchronization of extraction parameters. The detection of voice activity and follow-up and face detection are also resources offered by the framework.

Many studies have attempted to develop systems of content identification in the study of emotional content in voice

signals. Ayadia, Kamel, and Karray (2011) study the classification of emotional expressions and consider three important aspects for a recognition system of emotions in speech: (i) the choice of adequate characteristics for voice representation, (ii) the conception of an adequate classification system, and (iii) the preparation of a database containing emotional discourse tracks to evaluate the system. Their study includes discussions on the performance of these kinds of recognition systems as well as their limitations in the field.

To the emotion recognition in videos, specifically TV news videos, robust modulation recognition techniques must be applied to speech and facial expressions for avoiding significant loss of the affective weight associated with the informative content of the news and its discursive semiotic significance. It is believed that the verbal–visual language adopted in TV news programs, including speech modulations, color choices, lighting, and camera framing, accompanies the emotional significance of the communicators, sometimes even in a subtle manner. In the process of news spectacularization for emerging and persuading the viewer on the mentioned content, that is, if the emotion while watching the TV news is real, the information is also considered to be true (FLAUSINO, 2003).

4 AUTOMATIC DETERMINATION OF TENSION LEVELS

This section presents the general methodology used to develop this study, including the conceptual level with elaboration of the proposed model and the implementation of the framework-processing prototypes.

The present study aimed to better understand the dynamics of discursive-semiotic analysis of TV news and the tension levels perceived in their information building units as well as to illustrate the potentiality of extraction of the developed computational system. To achieve this goal, this study was required to refer to a video stored in the database of the *Centro de Apoio a Pesquisas sobre Televisão* (CAPTE – Support Center for Television Research) of the CEFET-MG that allowed several studies, such as those by Pereira *et al*. (2015), Souza *et al*. (2014), Braighi-Andrade (2013), Jacob (2013) and Conceição (2013). This video from a news report by *Jornal Minas*, from the

broadcaster *Rede Minas*, This video presents a news report by *Jornal Minas* on August 8, 2011, in the broadcaster *Rede Minas*, about the increase in cases of violence against the elderly.

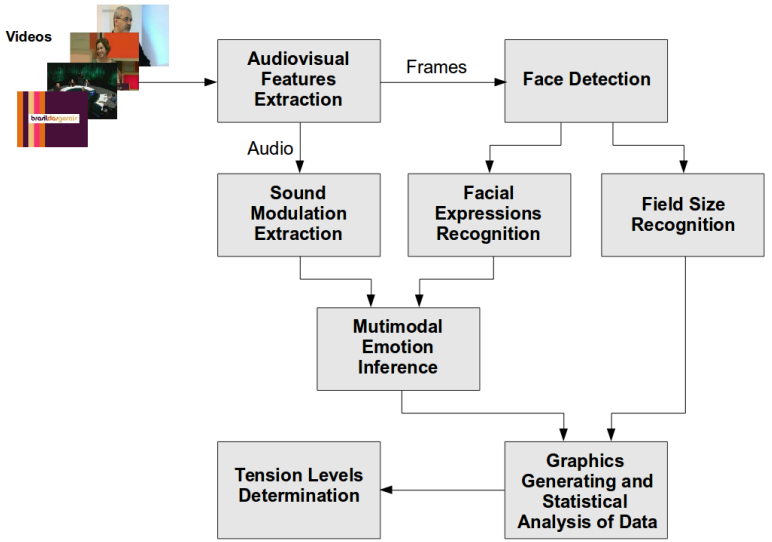
For the multimodal emotion recognition stage, this report begins with the recognition of emotions in audio signals, as proposed by Florian *et al*. (2013), and with facial expression recognition, according to the approaches described by Littlewort *et al*. (2004) and Bartlett *et al*. (2006). The report employs the Facial Action Codification System proposed by Ekman and Friesen (1978) in a detector based on the Haartraining algorithm by Viola & Jones (2001). In the present study, 93% accuracy was obtained. The discursive semiotic parameters proposed by David-Silva (2005) are used to classify the videos on the levels of *Absence of Tension*, *Moderate Tension* and *High Tension*, according to the emotional intensity inferred from facial expressions, the field size of the subject's face, and the intensity of sound values detected in the audio signals.

In the TV displays within the Absence of Tension category, one perceives that the speech direction regarding the subject matter sends the individual to the semantic field of "happiness," such as sports events, commemorations, cooking tips, provoking a kind of relief in the viewer. In the Moderate Tension displays, the news provokes a kind of pathos within the viewer, although it is considered as bearable because there is a certain distance from the daily life and place of the audience of the news piece, from the involved parties, and from the subject matter itself. The news pieces with High Tension refer to reports on conflict, violence, tragedy, and death (murders), revealing problems in the world that may produce a pathos response from the viewer, especially when the subjects discussed are more relevant to these viewers (DAVID-SILVA, 2005). Based on these concepts, we propose to classify the videos in which emotions such as fear, anger, and sadness occur as High Tension; the videos with predominance of facial expressions of disgust, surprise, and disgust as Moderate Tension; and the videos that present more emotions of happiness as Absence of Tension.

Figure 1 presents a general view of the proposed approach for multimodal emotion recognition in the automatic determination of tension levels in TV news videos. Algorithms for the extraction of audiovisual characteristics were applied to provide the resources necessary for facial detection. Once the faces were detected, we used the modules for facial expression recognition and film shot identifi-

cation, and regarding the video–audio signal, we used the module of speech modulation recognition. Multimodal emotion recognition combines the data from the recognized characteristics from the audio signal and facial expressions. Finally, the inferred emotions are presented in graphs that analyze the data to apply metrics for determining tension levels of the TV news videos and to support semiotics studies by discourse analysts.

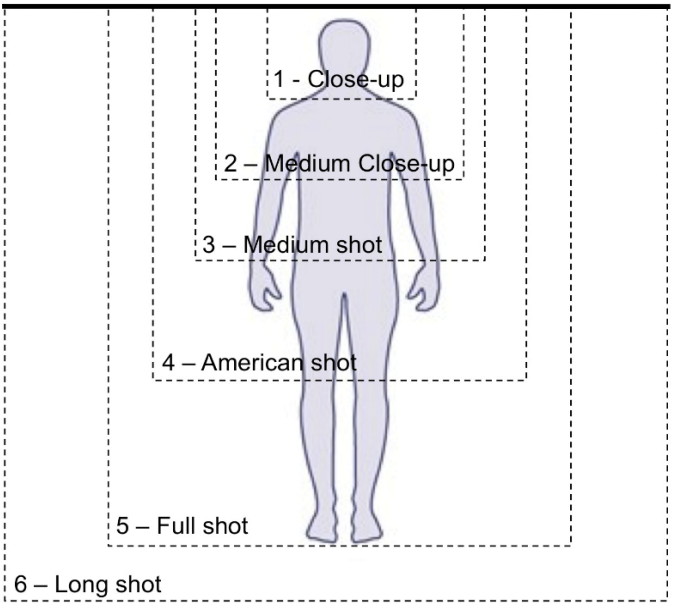
Figure 1 – Proposed model for automatic determination of tension levels in TV news.



Source: Elaborated by authors.

During the facial expression recognition process, the film shot that is represented by camera framing is identified using the approach offered by Conceição (2013) that calculates the proportion of the face in relation to the corresponding video frame in which it is shown. The influence of film shots on tension levels may be found in the intention to explore the intensity or apprehension about a certain facial expression. Figure 2 shows that plans are classified according to the size of the human figure in the frame. Therefore, plans can be understood as images captured and framed by a camera.

Figure 2 – Basic kinds of camera's field sizes.



Source: Elaborated by authors.

Tabela 1 – Face's proportion in the field sizes.

Code ID	Field Size	Proporção (α)	Acurácia
1	Long shot	$\alpha \leq 0.12$	0.88
2	Full shot	$0.12 < \alpha \leq 0.19$	0.84
3	American shot	$0.19 < \alpha \leq 0.22$	0.82
4	Medium shot	$0.22 < \alpha \leq 0.28$	0.60
5	Medium Close-up	$0.28 < \alpha \leq 0.40$	0.85
6	Close-up	$\alpha > 0.40$	0.95

Source: Adapted from Conceição (2013).

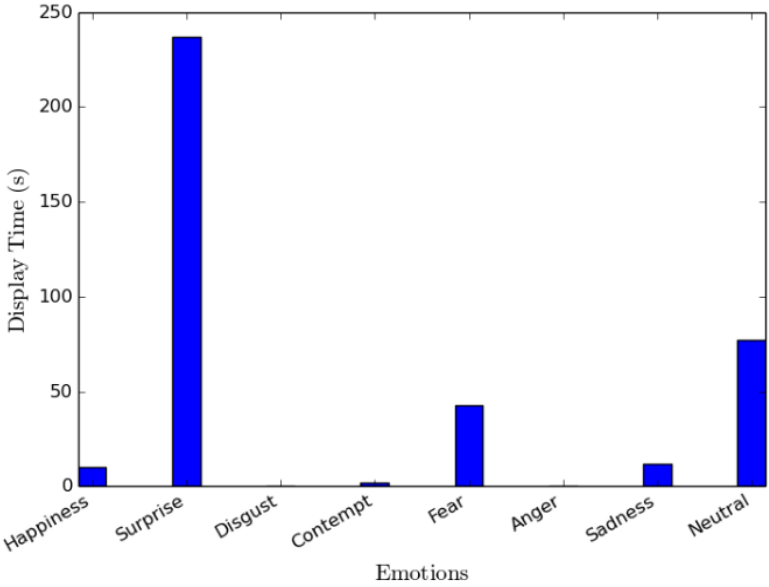
Table 1 presents the values of face proportion used in this study for the identification of camera framing film shots according to measurements used in the TV news program industry. Therefore, the automatic determination of reading points or points to mark control structures, such as the eyes, mouth, and nose, occurs immediately after the initial stage of automatic detection of face and other areas. As the model needs

a neutral facial expression to read the facial movements, we decided to mark the first face detected on the video as neutral that then strayed from the premise of emotive impartiality which the presenters should begin with and maintain.

For the process of extracting characteristics in audio signal modulations, we used the openSMILE framework proposed by Florian *et al*. (2013). In this stage of the model, the framework uses the spectrum of the video's audio component of a certain news report to extract the characteristics of sound intensity and the fundamental frequency of the modulations of those signals. The sound intensity reflects the rate of sound wave amplitude perception by the human ear, measured in decibels (dB). Fundamental frequency is defined as the primary harmony of a sound wave, and it is the most influential frequency in the perception of a certain sound. In the case of the human voice, these values confirm age and gender, with 85 to 180 Hz for men and 165 to 255 Hz for women. It is one of the main characterizing elements of voice (FLORIAN *et al*., 2013; PEREIRA *et al*., 2009; PEETERS, 2006). This stage is very important for the analysis of data on speech modulations in TV news programs that may influence the rhythm of news reception among viewers, including semantic aspects of emotional-verbal structures that should be tested for the efficacy of information transmission (MACHON, 2012).

The next stage is the simplified statistical analysis of the data obtained and the presentation of this data in the form of graphs to enable the application of certain metrics as well as to support research on the analysis of tension levels in TV news programs. In this report, the analysis of the data extracted begins with a histogram that shows the display time of emotions on the views or a graph model that portrays the frequency of occurrence of emotions on the video, reporting on the display time, in seconds, in which each emotion was inferred. Next, there are graphs of each audiovisual indicator, and the tension levels are identified.

Figure 3 – Graph for frequency of display time of emotions on the video.



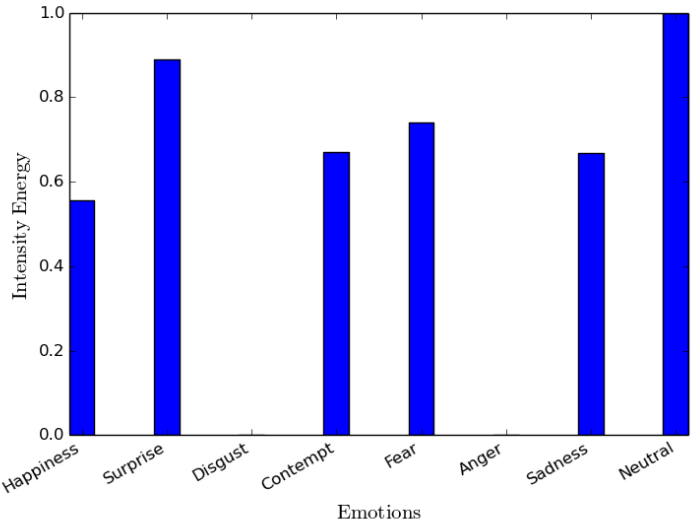
Source: Elaborated by authors.

Figure 3 illustrates the graph of occurrences of emotions per display time in the news report on the increase of cases of violence against the elderly. The X-axis represents the modeled emotions (Happiness, Surprise, Disgust, Contempt, Fear, Anger, Sadness, Neutral), and Y-axis shows the occurrences, in seconds, of each emotion. In a 372 s theme display, a significant trend of Moderate Tension was observed. In that, Happiness (Absence of Tension) was shown for 11 s, Moderate Tension for 236 s, and High Tension for 56 s. Following this to verify the trend, which may be refuted, it becomes necessary to analyze the influence of factors of the remaining audiovisual resources that, under the perspective of computational concepts but not restricted to these, compose the verbal–visual dynamics in the display of a news report.

Figure 4 presents the graphs of the values of visual intensity in emotion recognition during the news report and considers the level of expressiveness perception. Not considering the appearance of emotionally neutral faces, we observe that the process of recognition of the emotion Surprise has a considerably higher level of intensity compared with the other emotions, however, with an average level of expressiveness close to the occurrence of the face associated with Fear,

even if that emotion has been displayed for less time in the process of inference, suggesting a visual arrangement of heterogeneous narrative. In fact, this finding is expected, since the emotion of Fear possesses a more marked and intense expressiveness. The level of Surprise is shown from the 130-s point of the news report until the 210-s point. In that moment in the video, the space of information is internal (in the TV news studio), as a part of the interview with the professional Felipe Willer, president of the State Counsel of the Elderly, at which point a certain expressiveness of the participant occurs in a more energetic intention to clarify himself and to emphasize the rights of the elderly.

Figure 4 – Visual intensity in facial expression recognition.

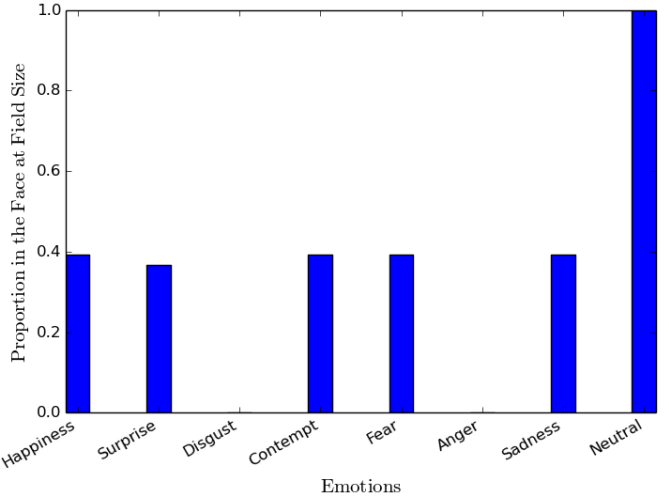


Source: Elaborated by authors.

Figure 5 presents graphs referring to visual intensity values of the proportions of a face detected in the video, the corresponding recognized emotion, and the type of film shot used for this face in the news report. According to codes described in Table 1, we perceive that most of the emotions were detected in the American and Full shot framings, even though most of the TV news programs occurred within Medium shot framing. According to the study by Braighi-Andrade (2013) on TV news from Minas Gerais, the *Jornal Minas* news show uses a more conservative format in terms of visual direction of camera shot resources. In the specific case of this news report, there were neutral facial expressions at times that should have re-

flected more emotional intensity. This unexpected situation is easily explained when the news report turned to an exterior space to record the statement of a young man who voiced his concerns over taking care of his mother, who suffered from Alzheimer's. The citizen interviewed was absent in the camera shot and therefore did not show any facial expression in the framing. There were Close Shot framings, almost close-ups, when the Alzheimer's patient was recorded, illustrating the attempt to invoke pathos in the audience for the story.

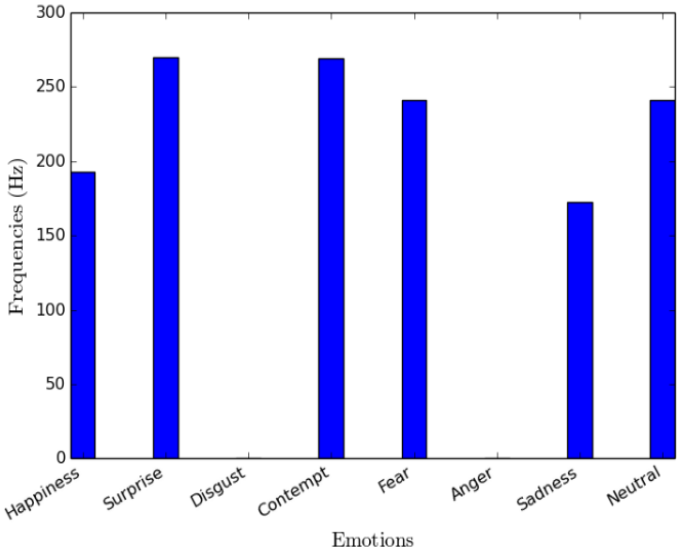
Figure 5 – Tendência da proporção das faces nos enquadramentos de câmera.



Source: Elaborated by authors.

Figure 6 presents the distribution of harmonic spectrum frequencies on the sound intensity in the speech process inside the TV news program. Even though not consistent with the corresponding facial expressions noted throughout the TV news broadcast, the applied computational method detected disparities in voice modulations that more intensely reflected some emotions without the same emotional correspondence in the visual process. An important aspect of the graph that needs to be considered and analyzed further involves the data regarding the emotion of Contempt. Even though on the time display graph, the emotion of contempt inferred from the facial expression was almost insignificant, considering the total program, it was more energetic than other emotions in terms of its vocal modulation.

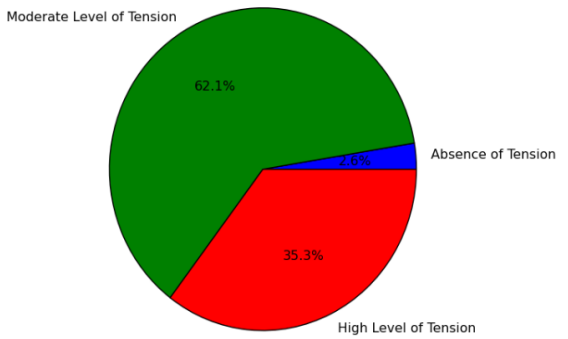
Figure 6 – Fundamental frequency in sound intensity of the detected speechs.



Source: Elaborated by authors.

With the treatment of the equivalent numeric values presented visually, a graph was created that assembled all data of the visual intensity of facial expression recognition, camera framing shots, modulation of the fundamental harmonic frequency of speech, and the sound intensity of that frequency in a weighted sum over display time of emotions inferred from the video.

Figure 7 – Analysis for the frequency of the display time of tension levels.



Source: Elaborated by authors.

Figure 7 presents the graph for frequency of display time of tension levels, and it considers the display times of all audiovisual resources from the corresponding video at a rate of 29 frames per seconds in the frames, and in the audio cuts in hundredths of a second. For the news report, we observed 62.1% of inference of emotions referring to Moderate Tension, 35.3% of High Tension, and only 2.6% of Absence of Tension that corresponds the inference distribution of emotions on the facial expressions recognized in the first graph.

When all the analyzed audiovisual resources are considered, it becomes clear that in terms of the statistical percentage of accuracy of each computational method that was applied, the news report in question possesses a predominantly regular tension level, that is, a Moderate Tension level according to the classification proposed by David-Silva (2005). The news report deals with the subject matter on increases in violence against the elderly from an educational perspective, and it presents means of help and support for people in the context. It is a subject that bothers viewers, but the level of world disorder is moderate.

5 CONCLUSION

This work presents a new approach that integrates computational methods for the extraction of audiovisual resources in the recognition of sound modulations in speech and emotions in facial expressions to identify and classify the tension levels in TV news program videos as a part of an interdisciplinary and complementary approach of a discursive-semiotic analysis of these TV programs.

Facial expressions are non-verbal ways to communicate that are constantly used in daily life. It is a topic that may promote promising studies on the use of ethos in TV news reports, particularly when this element is combined with other parameters of discourse analysis of TV media, such as camera framing film shots, enunciative modes, vision axis, and participant disposition. In the approach suggested in this work, we used the proportion of detected faces relative to the framing and in terms of pre-determined values in the study of film shots to determine the intensity of the emotion recognized in the frame. This combination may be a focus of more

specific studies that can determine, for example, whether the intensity of a certain emotion, depending on the tension level in which it was categorized, legitimately establishes the proximity of the TV news program with the audience. This status can be used as evidence with the presenter under a close-up. In addition, the increase and decrease in the physical proximity to the viewer, who looks for complicity toward the statement, are discursive modulators for the development of meaning before an emotion is intentionally directed in the communicative strategy.

Speech modulations and facial expressiveness of the reporters – particularly news anchors – may provide evidence on the discourse tension generated by statements in news reports and the sequencing pattern of the news arranged during the process of production based on the news items. In future, we expect to be able to apply the proposed method and to identify the tension levels of each news piece during the broadcast of several TV news programs. Such findings will support the studies on sequencing patterns of news in the identity of the program. Regarding verbal discourse, this work suggests the extraction of Closed Captioning from TV news to apply feeling analysis techniques (PANG & LEE, 2008) to better determine tension levels. By combining these techniques with a multimodal analysis approach, we may obtain the fluctuations of the polarities of feelings throughout each news piece and can therefore study the correlation of these feelings with the sound modulations of speech and audio signals to produce innovative research in semiotics and verbal–visual discourse analysis of TV news programs.

*This paper was translated by Ulatus.

REFERENCES

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 - M.Sc. degree in Mathematical and Computational Modeling from Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG). Assistant Professor at the Engineering and Technology Institute of Centro Universitário de Belo Horizonte (UNI-BH). moises.ramos@prof.unibh.br

Flávio Luís Cardeal Pádua - Ph.D degree in Computer Science from Universidade Federal de Minas Gerais (UFMG). Associate Professor at the Department of Computing of Centro Federal de Educação Tecnológica de Minas Gerais (CEFET -MG). cardeal@decom.cefetmg.br

Giani David Silva D.Sc. degree in Languages from Universidade Federal de Minas Gerais (UFMG). Associate Professor at the Department of Language and Technology of Centro Federal de Educação Tecnológica de Minas Gerais (CEFET -MG). gianids@deii.cefetmg.br

RECEIVED ON: 01/03/2015 | APPROVED ON: 26/08/2015