Desarrollo de un sistema para simulación y análisis de estrés en exposiciones orales

H. Tonutti[1], E. Schmidt[1], C. Martínez[1,2], E. Albornoz[1]

[1] Instituto de investigación en señales, sistemas e inteligencia computacional, sinc(i) UNL-CONICET, Ciudad Universitaria, Ruta Nacional Nº 168, km 472.4, (3000) Santa Fe.

[2]Laboratorio de Cibernética, Facultad de Ingeniería, UNER

(*) emalbornoz@sinc.unl.edu.ar - http://sinc.unl.edu.ar

Resumen. El estudio del estrés y los trastornos producidos por la ansiedad han despertado mucho interés multidisciplinar en los últimos tiempos, por su progresiva presencia y los efectos adversos que pueden producir. Dosis altas de estrés pueden desencadenar una situación de ansiedad, impulsando decisiones erróneas o una pérdida de rendimiento. Incluso si el estrés persiste, podría ser perjudicial para la salud, causando una amplia gama de enfermedades como ser diabetes, trastornos cardiovasculares o irregularidades inmunitarias. En este trabajo se presenta el desarrollo de una base de datos y algunas aproximaciones para detectar automáticamente manifestaciones del estrés en exposiciones orales a través del procesamiento de video.

Keywords: estrés - procesamiento de video - procesamiento de audio - computación afectiva

1 Introducción

Usualmente no es trivial diferenciar el estrés y la ansiedad, ya que son palabras utilizadas en contextos similares. El estrés es un proceso que se origina cuando las demandas del ambiente superan la capacidad adaptativa de un organismo, en otras palabras, es un sentimiento generado por el cuerpo humano para afrontar ciertos desafíos [1]. Esto ocurre mediante la activación del sistema nervioso y las hormonas, existiendo la posibilidad de un incremento en la frecuencia respiratoria, en la presión sanguínea, en el metabolismo, una disminución del diámetro de la pupila, tensión en los músculos, entre otros. Por otro lado, la ansiedad, además de ser una respuesta emocional al estrés, puede ser una reacción emocional de alerta ante una amenaza que puede originarse sin agentes estresantes [2]. Algunas veces, pequeñas dosis de estrés ayudan a estar listo para superar una situación adversa [3], pero si el estrés aumenta demasiado, puede desencadenar una situación de ansiedad, produciendo decisiones erróneas o una pérdida de rendimiento. Incluso si el estrés persiste, podría ser perjudicial para la salud, causando una amplia gama de enfermedades como ser

diabetes, trastornos cardiovasculares o irregularidades inmunitarias [4]. Las manifestaciones físicas del estrés también incluyen un componente conductual que acompaña a los síntomas (apretar el puño, rigidez del cuerpo, cruzar los brazos, gestos faciales y otras conductas [5]). Reconocer cómo el estrés afecta y cuándo ocurre, se ha convertido en un reto de la salud médica actual [2].

Gran parte de las personas manifiestan distintos niveles de ansiedad o estrés cuando deben expresarse en público, lo cual puede ser leve y pasar desapercibido o bien, puede tratarse de un trastorno al hablar públicamente [6]. Un tratamiento para reducir el nivel de ansiedad es la realización de prácticas en exposiciones de prueba en un ambiente natural, siendo eficaz para disminuir las manifestaciones producidas por la ansiedad y aumentar las habilidades para hablar en público [7]. Las habilidades para hablar o exponer en público son esenciales en prácticamente todas las profesiones. Sin embargo, pocas personas se sienten naturalmente cómodas hablando frente a una audiencia [8].

En todo acto comunicativo, y por lo tanto en las exposiciones públicas que aquí nos competen, las personas usan una variedad de vías comunicativas para expresar mensajes y emociones. Entre estas vías se encuentra el habla (que porta la información en el significado de las palabras y de la "manera" de hablar), y otras señales no verbales como expresiones faciales, gestos, posturas y el lenguaje corporal [9]. Algunos estudios indican que realizar acciones como rascarse, tocarse la cara y morderse los labios está asociado a experiencias estresantes y pueden proporcionar información valiosa sobre el estado emocional de los sujetos, más allá de las propias declaraciones y expresiones verbales [10]. Existe interés en cómo se utilizan las señales verbales y no verbales para transmitir el estrés y cómo evaluarlo automáticamente de forma objetiva a través de indicadores asociados a los cambios físicos, incluyendo un análisis de qué señales permiten su identificación [11].

La implementación de tecnologías de detección automática, que permitan la retroalimentación inmediata, mientras que los individuos desarrollan el acto de comunicación en diversos escenarios sociales, tiene un gran campo de aplicación. Entre los fines potenciales de este tipo de tecnología, se pueden mencionar la posibilidad de mejorar la capacidad de hablar en público, mejorar la dicción, tratar la ansiedad social o algún otro tipo de trastorno. Los resultados de este proceso podrían incidir favorablemente en entrevistas de trabajo, presentaciones de ventas, adquisición de nuevos idiomas, comunicación intercultural, atención al cliente, comprensión médico-paciente, entre otros. En el contexto de la pandemia que estamos atravesando, donde la mayoría de las actividades se realizan de forma virtual, este tipo de tecnologías cobra mucha importancia. Además, muchos investigadores están prestando atención a los riesgos y consecuencias que puede traer este evento en la salud mental de las personas [12].

En este trabajo se presenta el diseño y desarrollo de un sistema que permite el análisis automático del estrés que presentan las personas en exposiciones orales. A partir de una grabación audiovisual, el sistema brinda información acerca de los niveles de estrés que presentan las personas en los distintos momentos. Los resultados

contemplan la dinámica temporal en la detección del nivel de estrés, y permiten relacionarlos con lo que sucede en cada intervalo de tiempo.

2 Materiales y Métodos

Para este trabajo fue necesario realizar la registración de una base de datos propia, estableciendo un protocolo de registración, etiquetado e inducción del estrés en los voluntarios, con el asesoramiento de una estudiante avanzada de licenciatura en psicología.

2.1 Protocolo de registración

La metodología empleada para la registración de la base de datos incluyó entrevistas en las que se pedía al voluntario que realice la lectura de dos textos, mientras se lo registraba. Los voluntarios fueron citados con el motivo de realizarles una entrevista para evaluar la comprensión de texto, sin explicar el verdadero motivo del experimento hasta no haber concluído los registros. Cada entrevista se realizó en dos etapas: la lectura de un texto y luego, responder una serie de preguntas.

Configuración de la escena y registro Los voluntarios se encontraban ubicados frente a dos entrevistadores. El registro se realizó utilizando una webcam y un grabador de voz posicionados entre el entrevistador y el entrevistado. En la Figura 1 puede apreciarse la escena en la cual se realizaron los registros. Uno de los entrevistadores, estaba frente a la laptop con la webcam incorporada en la parte superior, el grabador de voz a la derecha, los textos impresos a la izquierda y junto a este el soporte donde se ubicó el temporizador.

La prueba comienza con la lectura de un fragmento del texto *"Momentos constitucionales en el gobierno de la ciencia y la tecnología*" [13]. El voluntario tiene 4 minutos para realizar la lectura. El tiempo transcurrido se indica mediante un temporizador que se encuentra a la vista del entrevistado con un sonido de "*tic tac*" marcando los segundos. Al finalizar el tiempo, el voluntario debe dejar de leer aunque no haya finalizado el texto. Luego, uno de los entrevistadores le efectúa 7 preguntas relacionadas al texto. Finalizado este intercambio, se le otorga un descanso de 1 a 3 minutos, mientras se le ofrece agua al voluntario y se entablaba una conversación casual. La segunda etapa es realizada por el otro entrevistador. En esta parte se solicita al voluntario que lea "*El buscador*" [13], para lo cual no se establece ningún límite de tiempo y, por lo tanto, no se incluye el temporizador en la escena. El entrevistado da aviso cuando finaliza la lectura y, acto seguido, se le realizan 6 preguntas acerca del texto.

Textos elegidos El primero fue elegido porque presenta una mayor complejidad respecto del segundo en cuestiones de vocabulario, esquema y temática. En cambio, el segundo texto presenta una lectura sencilla, narrativa lineal, y además una enseñanza

o moraleja que capta más fácilmente la atención del lector. A continuación se presentan fragmentos de los textos elegidos.

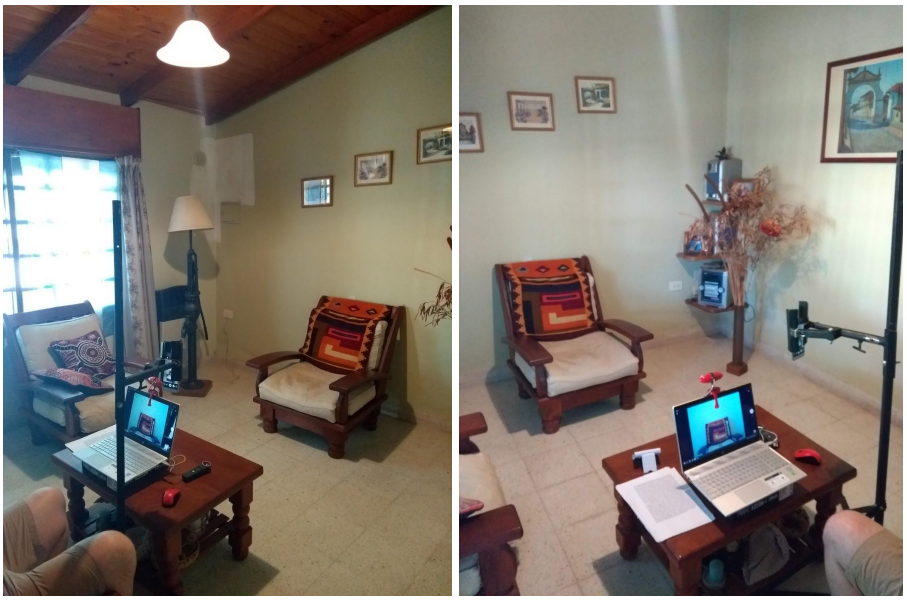

Fig. 1. Vistas de la escena donde se realizaron los registros.

Jasanoff (2011):

> La necesidad de involucrar con más intensidad a un mayor número de personas en el gobierno de sí mismas adquiere una urgencia creciente cada año, a medida que las distancias se acortan y las conexiones se vuelven más intrincadas entre las culturas del mundo insistentemente auto-afirmativas. [...] En particular, ¿pueden las trayectorias de la innovación tecnológica responder efectivamente al incremento de la comprensión de sistemas complejos, los cuales se derivan a menudo de campos interdisciplinarios como los ESCT, así como a las necesidades y preferencias de las multitudes globales, cuando las instituciones para discernir tales necesidades y articular tales preferencias son tristemente parroquiales y miopes? (pág. 18)

Bucay (2012):

> El Buscador traspasó el portal y empezó a caminar lentamente entre las piedras blancas que estaban distribuidas como por azar entre los árboles. Dejó que sus ojos, que eran los de un buscador, pasearan por el lugar […] Cuando un joven cumple quince años, sus padres le regalan una libreta, como ésta que tengo aquí, colgando del cuello, y es tradición entre nosotros que, a

> partir de entonces, cada vez que uno disfruta intensamente de algo, abra la libreta y anote en ella: a la izquierda, qué fue lo disfrutado, a la derecha, cuánto tiempo duró ese gozo. [...] Y cuando alguien se muere, es nuestra costumbre abrir su libreta y sumar el tiempo de lo disfrutado, para escribirlo sobre su tumba. Porque ése es, para nosotros, el único y verdadero tiempo vivido.

En resumen, en la primera etapa se buscó generar la situación estresante, inducida por el texto elegido, la imposición de un tiempo máximo de lectura y la distracción del sonido del temporizador, además el entrevistador se limitaba a realizar las preguntas sin efectuar ningún gesto o empatía al final de cada respuesta. Luego, la transición de descanso se implementó con el objetivo de brindar tranquilidad y contar con un lapso entre la etapa estresante y la neutra. Por último, en la segunda etapa el entrevistador realizó las preguntas de una forma amigable, donde las mismas fueron más simples y generales, y el entrevistador asintió luego de las respuestas de los voluntarios, con el objetivo de brindar más confianza y soltura en el entrevistado.

2.2 Base de datos y etiquetado

La base de datos se realizó en 21 sesiones, donde los voluntarios presentan edades entre 16 y 55 años, y se distribuyen en 10 hombres y 11 mujeres.

Con respecto a los datos técnicos, la cámara presenta una resolución máxima de 1024x768 píxeles, y frecuencias de cuadro de 8 fps. El grabador de voz se utilizó en una calidad de 192 kbps. El audio y video se sincronizó de forma manual, con la utilización de la herramienta de código abierto OpenShot Video Editor[1]. Finalmente, para la iluminación se empleó una lámpara de luz cálida en la parte superior central de la habitación, además de una ventana que iluminó la escena lateralmente con luz natural.

Para el proceso de etiquetado se usaron cuatro etiquetas posibles. Una etiqueta corresponde al estado neutral y las restantes refieren a distintos niveles de estrés: *bajo, medio y alto.* Al intervalo desde que el entrevistador comienza la pregunta hasta que el entrevistado finaliza su respuesta se le asigna una de estas etiquetas. De la primer etapa se obtuvieron 7156 fragmentos etiquetados como estado neutral, 11638 con estado de estrés bajo, 5499 a un estado de estrés medio y 2829 a un estado de estrés alto. Con respecto a la segunda etapa, 12425 fragmentos corresponden a un estado neutral, 4241 a un estado de estrés bajo, 144 a un estado de estrés medio y 0 a un estado de estrés alto.

En la fase de segmentación se efectuó el recorte de las respuestas. Cada video fue analizado manualmente y se extrajeron los tiempos correspondientes a los intervalos de respuesta. Se empleó la aplicación Online Video Cutter[2] para hacer este proceso.

[1] https://www.openshot.org/
[2] https://online-video-cutter.com/

En los casos donde las respuestas contienen interrupciones de los entrevistadores, se cortaron las mismas en múltiples partes, para luego ser unidas en un único video. Para esta última tarea de concatenación se empleó Aconvert[3]. Para este trabajo se utilizaron solamente los registros de video.

2.3 Extracción de características

Una gran variedad de métodos de análisis y descripción de imágenes son y pueden ser utilizados para el análisis de emociones y expresiones faciales [15, 16]. Aquí se utilizan: el método de patrones binarios locales (LBP, del inglés *Local Binary Pattern*), muy interesante por su tolerancia a los cambios de iluminación, eficiencia y simplicidad de cálculo [16]; el histograma de gradiente orientado (HOG, del inglés *Histogram of Oriented Gradient*), que demuestra un buen rendimiento en la detección de objetos dado que caracteriza eficientemente los rostros humanos por su apariencia y forma locales; el histograma de fase orientado (HOP, del inglés *Histogram of Oriented Phase*), un descriptor novedoso que ha presentado buenos resultados en la detección de pedestres y objetos en imágenes aéreas [17]. Además, se consideraron las unidades de acción (AU, del inglés *Action Unit*), que son puntos específicos del rostro cuya dinámica está dirigida por determinados músculos, capaces de representar casi todos los posibles movimientos de los músculos faciales y guardan una relación estrecha con las expresiones faciales emocionales [15].

Para los métodos de LBP y HOP, se emplearon implementaciones descritas en [18] y [17], respectivamente. Por otro lado, HOG se obtuvo utilizando rutinas de OpenCV, implementado según [19]. Para las AUs se utilizó la biblioteca OpenFace 2.0 [20].

Para reducir las dimensiones de las características y extraer las características más relevantes, se proponen diferentes métodos de selección del estado del arte. Se utilizó uno de los métodos más clásicos, de análisis de componentes principales (PCA, del inglés *Principal Analysis Component*s); y el método de subconjunto de funciones basado en correlación (CFS, del inglés *Correlation-based Feature Subset*). Para evaluar las características de CFS se utilizaron los métodos de búsqueda: optimización por enjambre de partículas (PSO, del inglés *Particle Swarm Optimization*) y Best Firsts (BF).

2.4 Métodos de clasificación

Existe un gran variedad de métodos empleados en el ámbito del reconocimiento de emociones y expresiones faciales [9, 15, 21, 22], y aquí se exploraron varios de distinta naturaleza. Se utilizó un clasificador basado en redes neuronales: el perceptrón multicapa (MLP, del inglés *MultiLayer Perceptron*); dos clasificadores basados en árboles de decisión: bosques aleatorios (RF, del inglés *Random Forest*) y C4.5 (también conocido como J48); y máquinas de vectores de soporte (SVM, del

[3] https://www.aconvert.com/

inglés *Support Vector Machine*). Los experimentos se realizaron utilizando las implementaciones disponibles en WEKA [23].

3 Experimentos y resultados

La extracción de características se realizó en cada cuadro de video. A diferencia de la mayoría de los trabajos que utilizan LBP y HOG sobre toda la imagen facial, aquí se decidió que fueran aplicados sólo en aquellas regiones que presentan una mayor variabilidad durante la realización de gestos faciales, es decir, las zonas de las cejas, los ojos, la nariz y la boca. La ubicación y delimitación de estas zonas fueron calculadas a partir de los puntos faciales obtenidos a través de OpenFace 2.0. Estas regiones son redimensionadas a un tamaño estándar por área. Luego de obtener las características para cada área facial, estas se concatenan en un solo vector junto a la intensidad de las AUs.

La selección de características comienza con una selección inicial según la correlación de Pearson entre cada característica y las clases. Sobre las características elegidas se realiza un nueva selección, utilizando cada uno de los 3 métodos que se presentaron previamente. Cada uno de estos tres conjuntos de características, fueron utilizados con los 4 clasificadores propuestos, obteniendo 12 configuraciones diferentes (Tabla 1).

En el presente experimento se empleó el método de validación cruzada con leave-3-out [24]. Es decir que, en cada fold se utilizaron 18 sesiones/individuos para el entrenamiento y las 3 restantes para la validación. Las características son normalizadas automáticamente por los clasificadores según lo requieran. En cada tarea, se calculó el Accuracy y el Unweighted Average Recall (UAR). Este último es importante ya que las clases están desbalanceadas.

En la Tabla 1 se pueden apreciar los resultados de clasificación para cada configuración (promedio de la validación cruzada). Las filas representan cada una de la combinaciones entre métodos de selección de características y de clasificación, mientras que las columnas indican las dos métricas empleadas: Accuracy y UAR. Como era previsible, los resultados muestran un valor de UAR por debajo del Accuracy, dado el desbalance de clases. Es interesante ver el buen rendimiento de RF respecto de los otros clasificadores para los distintos grupos de características.

En la Tabla 2 se presenta la matriz de confusión para la configuración que presentó mejores resultados, CFS combinado con PSO y RF. En esta matriz, las filas indican las etiquetas reales y las columnas cuales fueron las predicciones realizadas por la configuración mencionada. Puede observarse como la mayoría de predicciones quedan incluidas dentro de las etiquetas *neutral* o *estrés bajo,* resaltando nuevamente el desbalance de clases, y en este caso, la falta de características o muestras capaces de representar correctamente las demás etiquetas.

Tabla 1. Resultados de las distintas configuraciones de selección y clasificación [%].

	Accuracy	**UAR**
PCA + RF	42.15	29.60
PCA + SVM	44.00	27.39
PCA + J48	37.13	26.05
PCA + MLP	38.69	27.18
CFS-BF + RF	44.56	30.72
CFS-BF + SVM	36.13	27.34
CFS-BF + J48	38.22	26.19
CFS-BF + MLP	41.68	28.55
CFS-PSO + RF	**45.01**	**30.80**
CFS-PSO + SVM	36.12	27.33
CFS-PSO + J48	36.98	26.25
CFS-PSO + MLP	41.72	27.66

Tabla 2. Matriz de confusión para la configuración CFS-PSO + RF.

	Neutral	Estrés bajo	Estrés medio	Estrés alto
Neutral	13959	5529	90	3
Estrés bajo	9961	5769	142	7
Estrés medio	2917	2528	193	5
Estrés alto	1214	1400	213	2

Estos resultados preliminares indicarían que el desbalance de clases no permite el correcto entrenamiento del clasificador. Además, respecto a las clases asignadas a los fragmentos de las sesiones, es posible que en el proceso de etiquetado las etiquetas

sean relativas a lo observado en la sesión, es decir, lo que parece un estrés bajo en una sesión puede ser un estrés medio en la otra. Parece necesario hacer las pruebas con las sesiones de forma independiente.

4 Conclusiones y trabajos futuros

En este trabajo se ha avanzado en la definición de un protocolo para la recolección de una base de datos audiovisual, generando situaciones donde las personas manifiestan distintos niveles de estrés. El método desarrollado cumple con los objetivos e hipótesis iniciales acerca de las forma de simular una situación real mientras se inducen situaciones que generan más o menos estrés.

Un primer experimento de clasificación arroja resultados aceptables y nos da pautas hacia dónde debemos orientar las futuras pruebas, por ejemplo, utilizando técnicas de balance de clases o simplificando las clases a estresado y no estresado.

En los próximos trabajos se incorporará el análisis del audio junto al del video, obteniendo un experimento multimodal. Dentro de este enfoque se tendrán en cuenta dos propuestas: una selección de características y clasificación conjunta, y una selección y clasificación de manera individual por modalidad, fusionando luego los resultados de los clasificadores para obtener una predicción final. Además también se aplicará un esquema de fusión para combinar las distintas configuraciones presentadas en este trabajo.

Agradecimientos

Los autores desean agradecer a Pamela Wuignier (asesoría en psicología), al instituto sinc(i) y al CIMEC [institutos UNL-CONICET pertenecientes al CCT CONICET Santa Fe] por el acceso al Cluster Pirayu, a la UNL (CAI+D 50420150100098LI y CAID-PJ-50020150100055LI) y a la ANPCyT (PICT-2016-0651), por su apoyo.

Referencias

1. P. Benito Lahuerta, J. Simon, A. Sánchez Moreno and M. Matachana Falagán, *Promoción de la salud y apoyo psicológico al paciente*. Edición 2011, España: McGraw-Hill, 2011.
2. J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156-166, June 2005.
3. B. McEwen and T. Seeman, “Stress and affect: Applicability of the concepts of allostasis and allostatic load, ” Ed. R. J. Davidson, K. R. Scherer, & H. H. Goldsmith, *Series in affective science. Handbook of affective science*s, pp. 1117-1137. Oxford University Press, New York, United States.

4. G. Armaiz-Pena, S. Lutgendorf, S. Cole, and A. Sood, “Neuroendocrine modulation of cancer progression,” *Brain, behavior, and immunity*, vol. 23, no. 1, pp. 10-15, January 2009.
5. S. Greene, H. Thapliyal and A. Caban-Holt, "A Survey of Affective Computing for Stress Detection: Evaluating technologies in stress detection for better health," *IEEE Consumer Electronics Magazine* , vol. 5, no. 4, pp. 44-56, Oct. 2016.
6. Trastornos de ansiedad. Instituto de Neurología Cognitiva. Web: https://www.ineco.org.ar/patologias/trastornos-de-ansiedad. (Accedido el 15/05/2019).
7. A. Bados López, “Miedo a hablar en público: naturaleza, evaluación y tratamiento”, Universitat de Barcelona, 2015. http://hdl.handle.net/2445/65625. (Accedido el 18/05/2019).
8. M. E. Hoque and R. W. Picard, "Rich Nonverbal Sensing Technology for Automated Social Skills Training," *Computer* , vol. 47, no. 4, pp. 28-35, Apr. 2014.
9. I. Lefter, G. J. Burghouts and L. J. M. Rothkrantz, "Recognizing Stress Using Semantics and Modulation of Speech and Gestures," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 162-175, April 2016.
10. Alfonso Troisi. Displacement Activities as a Behavioral Measure of Stress in Nonhuman Primates and Human Subjects. Stress , 5(1):47–54, July 2009.
11. A. R. Subhani, W. Mumtaz, M. N. B. M. Saad, N. Kamel and A. S. Malik, "Machine Learning Framework for the Detection of Mental Stress at Multiple Levels," *IEEE Access*, vol. 5, pp. 13545-13556, July 2017.
12. R. P. Rajkumar, “COVID-19 and mental health: A review of the existing literature,” *Asian Journal of Psychiatry,* vol. 52, pp. 102066, August 2020.
13. C. Aguirre, T. Arboleda, S. Casiani, S. Daza, J. Guivant, D. Hermelín, S. Hilgartner, S. Jasanoff, M. Lozano Borda, M. Lozano Hincapié, O. Maldonado, J. Metcalfe, L. Olivé, T. Pérez Bustos, C. Raigoso, M. Sequera and J. Sutz. *Ciencia, Tecnología y Democracia: reflexiones en torno a la Apropiación Social del Conocimiento*. Memorias del Foro-Taller de Apropiación Social de la Ciencia, la Tecnología y la Innovación. T. Pérez-Bustos & M. Lozano-Borda, Eds. Medellín: Colciencias, Universidad EAFIT, 2011.
14. J. Bucay. *Cuentos para pensar* . Edition 2012. Buenos Aires: RBA Libros, 2012.
15. D. Y. Liliana and T. Basaruddin, “Review of Automatic Emotion Recognition Through Facial Expression Analysis," 2018 *International Conference on Electrical Engineering and Computer Science (ICECOS)*, Bangka-Belitung, Indonesia, pp. 231-236, October 2018.
16. S. Nigam, R. K. Singh and A. Misra, “A Review of Computational Approaches for Human Behavior Detection," *Archives of Computational Methods in Engineering*, vol. 26, no. 4, pp. 831-863, September 2019.
17. H. Ragb and V. Asari, “Histogram of oriented phase (HOP): a new descriptor based on phase congruency," *SPIE Proceedings*, vol. 9869, p. 98690V, May 2016.
18. T. Ahonen, A. Hadid and M. Pietikäinen, “Face Recognition with Local Binary Patterns,” *European Conference on Computer Vision (ECCV)*, Prague, Czech Republic, pp. 469-481, May 2004.
19. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 2005, pp. 886-893 vol. 1.
20. T. Baltrusaitis, A. Zadeh, Y. C. Lim and L. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, Xi'an, 2018, pp. 59-66.
21. T. Pfister and P. Robinson, "Real-Time Recognition of Affective States from Nonverbal Features of Speech and Its Application for Public Speaking Skill Analysis," *IEEE Transactions on Affective Computin*g, vol. 2, no. 2, pp. 66-78, April 2011.
22. R. Basharirad and M. Moradhaseli, “Speech emotion recognition methods: A literature review," *AIP Conference Proceedings*, vol. 1891, no. 1, p. 020105, October 2017.
23. E. Frank, M. Hall and I. H. Witten, *Data Mining: Practical Machine Learning Tools and Techniques*. 4th ed. San Francisco: Morgan Kaufmann, 2016.

24. C. M Bishop, *Neural networks for pattern recognition*. 1st ed. New York: Oxford university press, 1995..