Redes neuronales aplicadas al reconocimiento básico de lenguaje de señas en fotografías

Geovanny Cordero[1], Juan Campos[2], Leonardo Jiménez[1]
CI2651 Inteligencia Artificial
Escuela de Ciencias de la Computación e Informática
Universidad de Costa Rica
[1]*geovanny.corderovalverde@ucr.ac.cr*, [2]*juan.campos@ucr.ac.cr*, [3]*leonardo.jimenez@ucr.ac.cr*

13 de diciembre de 2019

Resumen

La comunicación asertiva es de gran importancia para la adecuada convivencia entre personas y nadie, independientemente de si posee algún tipo de discapacidad, se debe ver privada de esta posibilidad. Para esto, se pretende construir un sistema que identifique conjuntamente una seña y un gesto facial; esto, con el objetivo de dar contexto a lo que se pretende transmitir al realizar dicha seña. Para esto, se utiliza un método de agrupamiento para la clasificación de imágenes de rostros en diferentes emociones predeterminadas, esto con el objetivo de identificar el significado de una seña en el lenguaje LESCO. Se obtuvieron resultados parcialmente bueno, es necesario completar la funcionalidad el proyecto al unir todas sus partes.
Palabras clave: agrupamiento, inteligencia, lesco, señas, lenguaje.

1. Introducción

Existe una gran brecha entre las personas que se logran comunicar mediante un lenguaje hablado y las que lo hacen por medio de un lenguaje de señas, esto según Muskan Dhiman (2017). El lenguaje de señas ha venido a crear una gran oportunidad para disminuir esta diferencia, la cual sin embargo continúa siendo muy grande, debido a que no todas las personas dominan este lenguaje. Teniendo a mano una herramienta que permita la traducción de el lenguaje de señas, ya sea a un medio escrito o hablado, facilitaría en gran medida la comunicación entre personas no escuchas y aquellas que sí poseen este sentido pero que desconocen el lenguaje de señas. Se han utilizado diferentes métodos para intentar lograr lo anterior descrito, sin embargo aún no se ha implementado uno idóneo y que sea bien acogido por las posibles personas usuarias.

Ya se han desarrollado soluciones para el reconocimiento de diferentes señas en imágenes y videos, pero según lo investigado, muy pocas de estos acercamientos le dan importancia a el tono con el que el gesto se genera. Por ejemplo, es muy diferente si una persona expresa la misma oración triste o feliz, en una situación de presión o mientras mantiene una conversación tranquila en un café. Lo que se pretende es determinar el sentimiento de una determinada seña a partir de la expresión facial, esto con el principal objetivo de darle un contexto a lo que la persona que se está comunicando por medio de señas quiere expresar.

2. Pregunta de investigación

¿Es posible realizar una asociación representativa entre una expresión facial que se da al mismo tiempo

en el cual se muestra una seña en Lengua de Señas Costarricense (LESCO)?

3. Problema

Crear una aplicación que interprete una seña en LESCO junto con una expresión facial, esto con el objetivo de inferir la intención de la persona al realizar dichos gestos.

4. Antecedentes

Esta sección incluye una descripción del estado del arte y el marco teórico. Se mencionan diferentes trabajos los cuáles se revisaron y se determinó que tienen relación con el tema propuesto.

4.1. Uso de la expresión facial en el lenguaje de señas

El lenguaje de señas no es solo el movimiento de las manos, sino que este involucra movimiento del cuerpo, influye en gran medida la expresión facial y también la vocalización que se le pueda dar a la seña al momento de realizarla, aspectos que revisten de suma importancia para el significado de dicho gesto realizado con las manos.

Para distinguir el significado de una seña a otro, es trascendental la vocalización y la expresión de la cara. Un buen ejemplo: dulce y dolor se expresan con la misma seña, pero la diferencia radica en la expresión facial que se realice.

4.2. Reconocimiento de emociones en la expresión facial

Según Gupta (2018), existen ocho emociones faciales universales de las cuales se posee gran evidencia: felicidad neutral, tristeza, ira, desprecio, asco, miedo y sorpresa. En este caso utilizan algoritmos de visión artificial y aprendizaje automático para realizar la clasificación de estas ocho emociones diferentes. El mejor resultado se dio con las máquinas de soporte vectorial, con una precisión de alrededor del 94,1 %.

En el trabajo de Senthilkumar et al. (2017), se propone y utiliza un modelo basado en agrupamiento automático a través de segmentación morfológica, el cual dividen en cuatro módulos principales: segmentación facial, segmentación de las diferentes partes del rostro (ojo, brecha ocular, frente, nariz y boca), extracción de características de cada una de las diferentes partes de la cara y la clasificación basada en reglas. Algoritmos de segmentación automáticos (Ostu y agrupamiento automático K-medias) se desarrollaron para eliminar el fondo y segmentar las diferentes partes de la cara. Luego, se calcularon diferentes formas y patrones de bordes para definir cuantitativamente dichas regiones segmentadas. Por último, se implementó un clasificador basado en reglas de dos niveles para clasificar las diferentes emociones. Todo en conjunto dio muy buenos resultados en la clasificación de emociones faciales.

4.3. Métodos de agrupamiento

Hay muchas formas de clustering, hay que hablar de un par importantes y buscarse los whitepapers de la vara, o ver que se ha usado antes en conjunto con reconocimiento de emociones

4.3.1. Particionamiento en clúster

Los algoritmos de particionamiento en clúster subdividen conjuntos de datos en un conjunto de k grupos, donde k es un número preestablecido de grupos para realizar algún tipo de análisis MacQueen (1967). El método de particionameiento en clúster más popular según Satapathy et al. (2015) es el de k-medias. En este, cada grupo se supone que contiene öbjetosl̈o más semejantes posible (esto se determina por el la media de los puntos de los datos que pertenecen al grupo).

Se han realizado trabajos comparando diferentes técnicas basadas en la evolución, como el realizado por David and Kosala (2019), en el cual se comparan algoritmos genéticos y el método de optimización basada en el aprendizaje docente (TLBO, por sus siglas en inglés). Este dio como resultado un mejor rendimiento por parte de el método TLBO. En la imagen 1 se puede ver una comparación entre dichos méto-

dos. En el eje Y se indican los valores de la suma del error cuadrado, mientras que en el eje X el número de clústers utilizados. En esta se puede ver que la diferencia es significativa en algunos puntos.

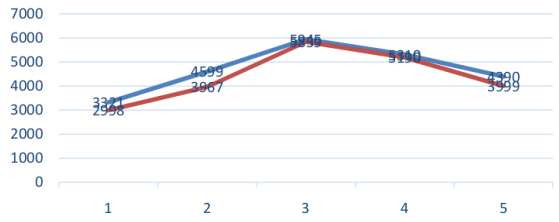

Figura 1: Comparación del error con diferente número de clústers

4.3.2. Agrupamiento difuso

El método *fuzzy clustering* o agrupamiento difuso determina la probabilidad de que un elemento esté en un grupo, de manera que este puede estar en más de una partición a la vez. Cada elemento tiene un conjunto de coeficientes de membresía, correspondientes al grado en que pertenece a un grupo dado. Los grados van desde 0 hasta 1, siendo más altos en cuanto estén más cerca del centro del cluster y más bajos cuando están más cerca del borde.

El algoritmo *fuzzy c-means*(FCM) es uno de los algoritmos de agrupación difusa más utilizados. El centroide de un grupo se calcula como la media de todos los puntos, ponderado por su grado de pertenencia al grupo.

Las expresiones humanas son naturalmente ambiguas, por lo que un algoritmo de agrupamiento difuso puede dar muy buenos resultados. Dewi Yanti Liliana (2016) utiliza *fuzzy c-means* para medir de forma adaptativa la distancia entre imágenes de caras arbitrarias al centroide de un grupo pre-entrenado de cada clase de emoción.

5. Objetivos

5.1. Objetivo General

Construir un sistema que identifique de manera precisa el significado de una seña en LESCO por medio del análisis del rostro al momento de realizar dicho gesto mediante el análisis de la expresión facial.

5.2. Objetivos Específicos

1. Identificar un método apropiado que genere un agrupamiento adecuado según las emociones seleccionadas para el proyecto, de manera que se puedan validar los resultados contra la opinión de un experto.

2. Utilizar el sistema creado para clasificar el significado que se pretende transmitir al realizar una seña en LESCO por medio del análisis de imágenes de rostros.

3. Utilizar tecnologías de fácil acceso y libres, de manera que el trabajo pueda ser ya sea, replicado, modificado, extendido y complementado con las menores restricciones posibles.

6. Metodología

A pesar de toda la investigación que se realice, resulta muy diferente y mucho más enriquecedor realizar la consulta directamente a una persona que se comunique con frecuencia mediante el lenguaje de LESCO. Es por eso que decidimos comunicarnos con una persona con estas características, que nos pudiera brindar información acerca del tema. Este nos habló de su experiencia y de como influyen diferentes factores a la hora de comunicarse por medio del lenguaje de señas.

Se debe seleccionar un conjunto de datos de imágenes de caras tomadas desde el frente, del cual se seleccionara un 10 % para pruebas y el 90 % restante para el ëntrenamiento" correspondiente. Ya existen diferentes conjuntos de datos estándares que han sigo utilizados en gran cantidad de investigaciones, como por ejemplo CK y CK+ Gupta (2018).

Se debe seleccionar un método de agrupamiento adecuado para el trabajo que se pretende realizar. Según Satapathy et al. (2015) k-medias es uno de los algoritmos de particionamiento más popular y en su trabajo obtuvieron resultados óptimos. También, en el trabajo realizado por Gupta (2018), se obtuvo el tercer mejor resultado al comparar varios métodos de agrupamiento. Dada esta revisión, identificamos que el método k-medias ha dado buenos resultados en trabajos anteriores y es recomendado para llevar a cabo el tipo de trabajo que se pretende con esta investigación, por lo que se implementa la solución con el uso de este.

Luego, se debe procesar el el conjunto de datos seleccionado. Esto se realizará utilizando OpenCV. Según Beltrán Prieto and Komínkova-Oplatková (2017) este tiene mejor rendimiento que el servicio de Microsoft Azure llamado *Cognitive Services*.

Al resultado de ese procesamiento se le debe aplicar el agrupamiento para determinar a cuál conjunto pertenece, en este caso cada conjunto representa una emoción diferente.

Luego, se debe probar el sistema con el 10 % de el conjunto de datos que se seleccionó para este propósito.

Por último, se deben analizar los resultados para generar conclusiones y si es posible añadir recomendaciones para futuros trabajos relacionados.

Se pretende utilizar únicamente tres señas diferentes: días de la semana, el gustar algo y poder hacer algún tipo de actividad.

7. Resultados y Discusiones

Debido a la poca información disponible de datos que se obtuvo, para probar nuestro trabajo se utilizaron conjuntos de datos genéricos, se describirá cada uno de ellos en las secciones posteriores. Sin embargo, una vez que se obtenga una cantidad de datos suficientes, en este caso de personas realizando señas de LESCO, en donde también se aprecie su rostro, se podrían aplicar la misma metodología y lograr resultados similares. De igual forma, para la detección de rostros en imágenes de mayor tamaño o que contienen más elementos, se utilizaron datos genéricos.

7.1. ¿Por qué LESCO?

El proyecto está enfocado en disminuir la brecha con la comunidad sorda de Costa Rica, donde se utiliza el lenguaje de señas LESCO. Este lenguaje ha sido abordado pobremente en cuanto a sistemas de reconocimiento, por lo que hay muchas necesidades por cubrir. No existen conjuntos de datos para entrenamiento o por lo menos no son de acceso libre, además las investigaciones se han enfocado solo en las señas y no en el resto del lenguaje corporal.

7.2. Uso de OpenCV

OpenCV es una biblioteca de código abierto enfocada a ofrecer herramientas para resolver problemas de visión artificial o visión por computadora. Esta está publicada bajo una licencia de código abierto BSD, lo que implica que se puede integrar en el proyecto de manera irrestricta.

Ofrece múltiples funciones de procesamiento de imágenes de bajo nivel y algoritmos de alto nivel como detección de peatones, coincidencia de características, seguimiento o, lo que es de interés para este proyecto, detección de rostros.

OpenCV no solo soporta equipos de escritorio, si no también móviles con sistemas operativos Android y iOS, lo que significa que eventualmente el conocimiento generado se podría aplicar en una app para *smartphone*. Kari Pulli (2012)

7.3. Agrupamiento

La idea original era utilizar el agrupamiento de datos con las imágenes como entrada y no utilizar redes neuronales para clasificar estas. En este momento, se utiliza el agrupamiento pero aplicado a cada imagen, es decir sobre los colores de esta. Esto con el objetivo de simplificar un poco la imagen para enviarla como entrada a cada una de las redes neuronales. Se utiliza k-means con cinco clústers como número predeterminado y la ubicación de los centroides de manera aleatoria. Esto nos genera una imagen únicamente con cinco colores, lo que permite que se mantengan

los rasgos más importantes de esta para su análisis. Una vez que se aplica el agrupamiento en cada imagen, esta se guarda ahora con su nueva forma para el siguiente paso, el entrenamiento de la red neuronal correspondiente o su clasificación en la red ya entrenada.

7.4. Redes neuronales

Para el análisis de los datos de rostros y señas se utilizaron dos redes neuronales convolucionales, ambas con una configuración similar. Aparte del preprocesamiento anteriormente mencionado, se convierte cada imagen a una escala de colores *RGB* y se les da un tamaño de $100x100$. Luego, se pasa a un vector el cuál se guarda en una estructura de datos para posteriormente ya sea, entrenar la red neuronal o probarla. De igual forma, es necesario la extracción de las etiquetas de cada una de las imágenes para entrenar y probar la red.

En ambas redes se utilizan una arquitectura secuencial, con cuatro capas: la de entrada que recibe el vector de cada imagen, una que aplana la entrada, una de activación que aplica la función *relu* (unidad lineal rectificada, elimina las entradas negativas de sus entradas) con ciento veintiocho nodos y la de salida que aplica la transformación *softmax* (que retorna como salida un tensor) con diez nodos para el reconocimiento de gestos y siete para la red que reconoce sentimientos, en cada uno de los cuales se obtiene la probabilidad de que dicha entrada pertenezca a una categoría. Dichas categorías y su orden se definen al principio del programa en un vector que contiene las etiquetas y uno que asocia una etiqueta a un índice. Este último es el que se utiliza para realizar la verificación de las salidas de la red.

7.4.1. Clasificación de emociones en rostros

Para el entrenamiento de la red neural y como parte de las pruebas realizadas en esta, se utilizó en conjunto de datos *Cohn-Kanade (CK+)*, que se encuentran en su página oficial. Se utilizaron las siete emociones diferentes presentes en este conjunto de datos: ira, disgusto, feliz, sorpresa, desprecio, miedo, tristeza. En este caso, la red entrenada queda lista para su posterior utilización con datos reales, ya que estos datos son bastante estándares y diversos, únicamente habría que realizar un pre-procesamiento a las imágenes (de tamaño y color principalmente) que es lo que se pretende al utilizar las herramientas de agrupamiento y OpenCV descritas anteriormente.

En este caso, se entrenó la red con 949 imágenes y se probó con 32 ejemplos, con lo que se obtuvo una precisión del 25,18 % en el reconocimiento de sentimientos.

7.4.2. Clasificación de señas

Como se indicó anteriormente, debido la poca, casi nula cantidad de datos disponibles sobre LESCO, para el entrenamiento de la red neuronal que reconoce señas, se utilizó el conjunto de datos *Hand Gesture of the Colombian Sign Language*, debido a que son señas comunes y en español. De este conjunto de datos se utilizan las vocales (a, e, i, o, u) y los dígitos 1, 2, 3, 4, 5 (se excluye el cero debido a que es similar a la o y para nuestros efectos no es significativa la cantidad).

Para efectos de este trabajo, se entrenó la red con 2738 imágenes y se probó con 300, obteniendo una precisión en la predicción de 10.56 % únicamente.

Es necesario indicar que no se obtuvieron resultados muy precisos en ninguna de las dos redes neuronales utilizadas, en ambas no se obtuvieron valores al 1s % de exactitud. Se considera que de debe a que la cantidad de datos no es suficiente para el entrenamiento de dichas redes o que la configuración de está no es la adecuada, queda pendiente realizar dicha revisión para trabajos posteriores.

En el repositorio *http://bit.ly/38z5k61* se puede consultar toda la información referente al proyecti, tanto el código como las fuentes de los conjuntos de datos utilizados.

7.5. Trabajo por Realizar

Actualmente el trabajo cuenta con la detección de rostros en imágenes grandes, que contienen muchos otros elementos, lo que generará las entradas para la red que reconoce las emociones en rostros. Se crearon dos redes neuronales como se describe en este

documento, que por el momento funcionan con datos genéricos, pero que son capaces de funcionar con datos reales.

Para que el proyecto se de por completado, es necesario unir todas sus partes, de manera que se pueda pasar una imagen completa de un individuo realizando una seña y que se pueda dar el resultado esperado: generar una aproximación a lo que el individuo quiso expresar al realizar dicha seña, esto al analizar el sentimiento expresado en su cara.

Se desea agregar nuevas emociones para que el análisis sea más específico y pueda generar más valor.

Se espera que en un futuro se inviertan recursos económicos y de tiempo en la generación de un conjunto de datos lo suficientemente grande de señas en LESCO que permita realizar trabajos similares al aquí descrito.

Referencias

Luis Antonio Beltrán Prieto and Zuzana Komínkova-Oplatková. A performance comparison of two emotion-recognition implementations using OpenCV and Cognitive Services API. *MATEC Web of Conferences*, 125:1–5, 2017. ISSN 2261236X. doi: 10.1051/matecconf/201712502067.

David and Raymondus Raymond Kosala. Clustering Algorithm Comparison of Search Results Documents. *2018 6th International Conference on Cyber and IT Service Management, CITSM 2018*, (Citsm):1–6, 2019. doi: 10.1109/CITSM.2018.8674246.

T. Basaruddin Dewi Yanti Liliana, M. Rahmat Widyanto. Human emotion recognition based on active appearance model and semi-supervised fuzzy C-means. *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2016. doi: 10.1109/ICACSIS.2016.7872744.

Shivam Gupta. Facial emotion recognition in real-time and static images. *Proceedings of the 2nd International Conference on Inventive Systems and Control, ICISC 2018*, (Icisc):553–560, 2018. doi: 10.1109/ICISC.2018.8398861.

Kirill Kornyakov Victor Eruhimov Kari Pulli, Anatoly Baksheev. Realtime computer vision with opencv. *Magazine Queue*, 10, 2012. doi: 10.1145/2181796.2206309.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press. URL `https://projecteuclid.org/euclid.bsmsp/1200512992`.

Muskan Dhiman. Sign Language Recognition. *2017 Summer Research Fellowship Programme of India's Science Academies*, 2017. URL `http://bit.ly/33liOPH`.

Suresh Chandra Satapathy, A. Govardhan, K. Srujan Raju, and J. K. Mandal. Partition based clustering using genetic algorithm and teaching learning based optimization: Performance analysis. *Advances in Intelligent Systems and Computing*, 338:V–VI, 2015. ISSN 21945357. doi: 10.1007/978-3-319-13731-5.

T. K. Senthilkumar, S. Rajalingam, S. Manimegalai, and V. Giridhar Srinivasan. Human facial emotion recognition through automatic clustering based morphological segmentation and shape/orientation feature analysis. *2016 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2016*, pages 0–4, 2017. doi: 10.1109/ICCIC.2016.7919663.