# Análisis semántico en rostros utilizando redes neuronales profundas

Nicolás F. Pellejero
Facultad de Ciencias Exactas, Ingeniería y Agrimensura.
Universidad Nacional de Rosario.
Email: pellejero.nicolas@gmail.com

Guillermo L. Grinblat
CIFASIS-CONICET
Rosario, Argentina
Email: grinblat@cifasis-conicet.gov.ar

Lucas C. Uzal
CIFASIS-CONICET
Rosario, Argentina
Email: uzal@cifasis-conicet.gov.ar

***Resumen*—En este trabajo se aborda el problema de reconocimiento y clasificación de Expresiones Faciales a partir de video. Actualmente existen excelentes resultados enfocados en entornos controlados, donde se encuentran expresiones faciales artificiales. En cambio, queda mucho por mejorar cuando se trata de entornos no controlados, en los cuales las variaciones de iluminación, ángulo a la cámara, encuadre del rostro, hacen que la poca cantidad de datos etiquetados disponibles sea un impedimento a la hora de entrenar modelos de aprendizaje automatizado.**
**Para atacar esta dificultad se utilizó de forma innovadora la técnica Generative Adversarial Networks, que permite utilizar un gran cúmulo de imágenes no etiquetadas con un estilo de entrenamiento semi supervisado.**

## I. Introducción

En los últimos años han surgido diversas tecnologías relacionadas en mayor o menor medida con la Inteligencia Artificial. Entre ellas podemos nombrar Internet of Things, la Robótica en nuestra vida cotidiana, Drones, vehículos no tripulados, etc.
Estas tecnologías podrían salir al mercado masivo en los próximos años, impactando de forma positiva en la sociedad. Para que esto suceda es fundamental que posean una interfaz para interactuar con los usuarios la cual permita maximizar la facilidad de uso y las funcionalidades que nos puedan brindar[1].
En muchos casos es beneficioso o hasta necesario que los dispositivos inteligentes, ya sea un robot, un sistema de domótica, un televisor, puedan detectar las emociones predominantes de sus usuarios. Así, se abre una gama amplia de posibilidades y aplicaciones, que van desde la medicina robótica hasta desarrollos en e-learning.
Para que esto sea posible, son necesarios desarrollos robustos a variaciones que fácilmente se pueden dar fuera del entorno del laboratorio, como pueden ser variaciones lumínicas, imágenes con ruido, variaciones de traslación, etc.
También es importante diferenciar entre las expresiones faciales comúnmente llamadas artificiales que suceden cuando la persona es especialmente llamada a realizar cierta expresión y por lo tanto es artificial, y las llamadas espontáneas que suceden cuando la persona realmente esta sintiendo tal o cual emoción, o al menos la misma es actuada por un actor profesional.
Actualmente existen excelentes resultados en entornos controlados, enfocados en expresiones faciales artificiales. En cambio, queda mucho por mejorar cuando se trata de entornos no controlados.
La técnica de GAN se presenta como una alternativa que permite realizar transferencia de conocimiento encapsulado en los pesos sinápticos aprendidos durante la parte del entrenamiento no supervisado. A medida que el entrenamiento avanza, se reutiliza este conocimiento realizando también un entrenamiento supervisado con un conjunto de datos etiquetados con la emoción predominante del rostro en la imagen.
A continuación se comentarán los trabajos más relevantes del estado del arte en el reconocimiento de emociones. En la Sección 3 se explicará brevemente el concepto de Generative Adversarial Network, fundamental para el presente trabajo. La Sección 4 se centrará en las metodologías de preprocesamiento desarrolladas. En la sección 5 se enumeran los conjuntos de datos, sus características y el objetivo con el cual cada uno fue usado. La sección 6, resumirá las pruebas realizadas tanto utilizando técnicas de transferencia de conocimiento convencionales como utilizando Generative Adversarial Networks. Por último se presentarán las conclusiones en la Sección 7.

## II. Estado del Arte

### A. Reconocimiento de Emociones

La metodología utilizada para el reconocimiento de emociones en imágenes de rostros puede ser dividida en dos grupos. Como primer grupo tenemos los trabajos basados en el marco teórico “Facial Action Coding System” (FACS) de Paul Ekman. Este investigador fue el que sentó las bases del reconocimiento de emociones en la década del 70. Su teoría busca dividir al rostro en un conjunto de movimientos musculares o Action Units (AU) y luego realizar un análisis basado en reglas teniendo en cuenta las AU activas en cada instante. Estas reglas fueron formuladas originalmente en el manual de FACS [2]. Además, en esta serie de investigaciones fue donde Ekman buscó un conjunto de emociones básicas, pan culturales, que luego fueron adoptadas por gran parte de la comunidad científica y representaron el esquema hegemónico hasta los 90, cuando el mismo Eckman comenzó a agregar otras emociones para extender su trabajo. En adelante se

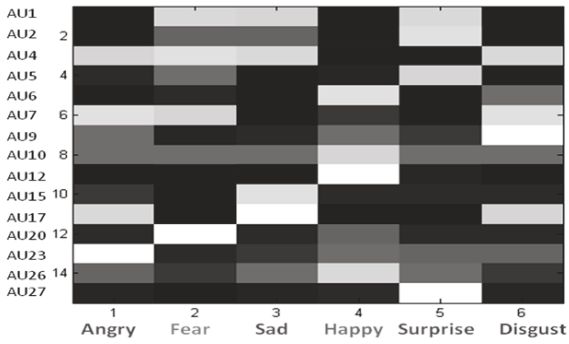


Figura 1. Ejemplo de relación de peso de AU con cada emoción básica. Aquí los diferentes tonos de gris representan pesos entre 0 y 1, indicando qué tan importante es la presencia de cada AU para definir cada emoción

utilizará este conjunto de emociones básicas planteadas por el investigador, refiriéndose a ellas como 'las 6 emociones básicas' o 'las 6 emociones básicas de Paul Ekman'.

Ciertos trabajos de este tipo fijan como objetivo final la detección de AU, suponiendo que luego esto servirá de apoyo a un codificador humano o a algún otro sistema [3]. Otros hacen uso de sistemas expertos para detectar las emociones predominantes [4]. Esto es especialmente ventajoso ya que puede servir de apoyo para otras aplicaciones como por ejemplo a partir de las AU detectadas, inferir si la persona está mintiendo o no.

También existen casos donde las reglas son inferidas estadísticamente, usando conjuntos de datos etiquetados tanto con información sobre las AU presentes como con las emociones predominantes [5]. En la Figura 1 se puede apreciar un ejemplo de la relación de peso entre las diferentes AU y las seis emociones básicas, inferidas estadísticamente del conjunto Cohn-Kanade [4]. Al inferir las reglas y usar una metodología que busque ser robusta a errores en las etiquetas y en la detección de las AU, se busca que el sistema final sea tolerante a algunos tipos de errores frecuentes.

Algunos de los trabajos fundacionales en este sentido fueron los realizados por Jeffrey Cohn y Takeo Kanade [6]. Ellos confeccionaron uno de los primeros conjuntos de datos etiquetados tanto por AU como por emociones [4]. Estos trabajos sirvieron como referencia para posteriores investigaciones.

Cabe destacar que el conjunto de datos de Cohn-Kanade esta realizado en un ambiente sumamente controlado con transiciones suaves entre la emoción neutral y el pico de intensidad de la expresión facial. Estas condiciones actualmente están superadas con lo cual comenzaron a realizarse conjuntos de datos en entornos no controlados, con expresiones faciales que buscan ser naturales o estar motivadas por ciertos disparadores, como puede ser ver algún tráiler de película o publicidad.

El segundo grupo lo forman los trabajos por fuera de FACS. Estos, en lugar de tomar como características principales las AU, toman otro tipo de características para tomar como base de la clasificación.

Hay actualmente conjuntos de datos etiquetados según las 6 emociones básicas de Ekman, lo cual es más sencillo y permite mayor cantidad de imágenes etiquetadas con menor esfuerzo humano. Inclusive se han realizado métodos semi automáticos para generar estos conjuntos a partir de imágenes de películas o extraídas de la web [7].

Estos conjuntos de datos pueden ser usados como entradas de Redes Neuronales Profundas, las cuales toman generalmente imágenes crudas como entrada, y usan las etiquetas de emociones para calcular su función de costo. Además, con el éxito de deep learning en Imagenet [9] comenzaron a probarse técnicas de transferencia de conocimiento, en este caso para detección de emociones [12].

También pueden usarse modelos convolucionales para aprender características que hagan buenas representaciones de rostros y luego usar estas características como entrada de otros métodos, como una SVM o un Random Forest [10]. Estas técnicas de reuso de detectores de ciertos patrones que son compartidos por diferentes objetos es muy útil cuando se tiene conjuntos de datos pequeños, ya que al comenzar con una red pre entrenada se aprovecha la experiencia ya guardada en las primeras capas y se espera que sólo las últimas capas den un salto de aprendizaje.

Además se han estudiado otros conjuntos de características diseñados especialmente para la tarea de reconocimiento de emociones, que no son las AU pero poseen su misma idea central, ser elementos atómicos de las expresiones faciales, fácilmente reconocibles y diferenciables [11].

### *B. SFEW y reconocimiento de emociones 'In the wild'*

En los últimos años, han surgido trabajos de investigación en conjuntos de datos de características artificiales, con errores de clasificación en test menores al 2% [8]. Estos trabajos utilizan datos realizados en laboratorio, con condiciones controladas de luz, ángulo de la cámara, expresiones que no son producto de un estímulo, es decir que son forzadas o actuadas, etc. Por el contrario, la comunidad científica es consciente que en muchas posibles aplicaciones para el reconocimiento de emociones por medio de video, el entorno es no controlado, teniendo variabilidad en la luz, el brillo, el contraste en la imagen.

Por otro lado, en muchas de estas aplicaciones las transiciones entre la expresión facial neutral y el pico de expresividad no son suaves sino que se dan en el transcurso de unos pocas imágenes.

Es por esto que se ha puesto esfuerzo en construir nuevos conjuntos de datos, donde las condiciones sean más cercanas a las que se puedan dar fuera del laboratorio, y representen un desafío tecnológico.

Un ejemplo de este tipo de conjuntos de datos es el recolectado en el marco del concurso anual EMOTIW. Este concurso busca motivar el avance del reconocimiento de emociones tanto en video como en imágenes. Con esta idea, se cuenta con un conjunto de videos cortos recortados de películas, seleccionados de forma semi automática y clasificados por emoción según las 6 emociones básicas que contempla Ekman. Esta clasificación está hecha en 2 pasos. Un primer paso

consiste en utilizar técnicas de análisis de sentimiento en los subtítulos de las películas, y así extraer y clasificar pequeños clips en donde se tenga una estimación de la emoción y sea probable que aparezca uno o mas rostros. Luego se procede a descartar los videos en donde no aparezcan rostros, utilizando un algoritmo de reconocimiento de rostros. Por último se hace una limpieza manual de los datos detectando los últimos errores que puedan haber quedado.
Este conjunto de videos cortos se usa como material para el concurso EMOTIW. Además, se selecciona un subconjunto de aproximadamente 2000 imágenes (50% aproximadamente para entrenamiento, 25% para validación y 25% para testeo) para el concurso SFEW, análogo a EMOTIW pero en imágenes en vez de video.
El conjunto SFEW se ha convertido en uno de los conjuntos de datos más representativos cuando se trata de clasificación de emociones espontáneas. Será usado en el presente trabajo, además del conjunto confeccionado por Cohn-Kanade en el laboratorio.
En el presente se tendrán especialmente en cuanta varios trabajos salidos de las últimas ediciones del concurso EMOTIW. En algunos casos, estos trabajos se enfocan en cómo sortear lo mejor posible el problema del sobre ajuste al trabajar con conjuntos de datos pequeños. También es de interés ver en cuántas etapas es conveniente dividir el entrenamiento, es decir, si para esta tarea en particular es beneficioso realizar varias etapas de transferencia de conocimiento y no sólo una [13][14].

## III. Generative Adversarial Networks

En esta sección explicaremos en qué consiste el método Generative Adversarial Networks, cómo fue evolucionando en el último tiempo, cuáles son algunas de sus ventajas y cómo nos permitió hacer frente al problema de entrenar un modelo profundo disponiendo de poca cantidad de datos etiquetados para la tarea en cuestión.
El modelo GAN surge en 2014 [15] como una alternativa de modelo generativo que busca obtener un buena representación de un cierto conjunto de datos. Para esto se cuenta con dos redes neuronales. Por un lado un modelo $generativo\ G$, que busca capturar la distribución del conjunto de datos, y por el otro un modelo $discriminador\ D$, que estima la probabilidad de que un ejemplo venga del conjunto de entrenamiento y no de $G$.
Así se establece un juego minimax de dos jugadores, donde se entrena $D$ para maximizar la probabilidad de etiquetar correctamente tanto a los datos de entrenamiento como a los datos generados por $G$, y al mismo tiempo se entrena $G$ para engañar al discriminador $D$[15].
Mediante este mecanismo competitivo se busca que cada modelo se enriquezca del aprendizaje de su adversario, y en particular que el discriminador aprenda los patrones propios del objeto que se está estudiando, en este caso el rostro humano.
Si bien la metodología es muy reciente, existen evidencias en múltiples conjuntos de datos de que el modelo logra aprender buenas representaciones de los datos. [16].
Desde el trabajo original de 2014 hasta la fecha, esta técnica ha ido evolucionando. Por un lado, han surgido una familia de arquitecturas profundas que han probado ser especialmente estables y tener buena velocidad de convergencia con relativamente pocos datos. Esta familia de arquitecturas fue llamada Deep Convolutional GAN o simplemente DCGAN [16].
El trabajo de DCGAN contribuye de varias formas al estado del arte:

- Define las características de las DCGAN, explicitando varias restricciones en su arquitectura, las cuales probaron empíricamente dar estabilidad al proceso.
- Usa la representación aprendida en varios conjuntos de datos para entrenar modelos de clasificación, llegando a resultados prometedores, comparables a otros algoritmos no-supervisados.
- Hace por primera vez un análisis visual tanto de los datos generados por G, como de los filtros de activación. Encontrando la propiedad de que algunos filtros en particular habían aprendido a generar imágenes de objetos comúnmente presentes en el conjunto de datos.
- Hallan propiedades aritméticas en los generadores, que les permiten manipular fácilmente algunas características de los ejemplos generados.

Otro sentido en el cual esta técnica evolucionó, es en cuanto a la metodología que se usa para transferir el conocimiento para resolver tareas de clasificación.
A mediados de 2016 Radford y Goodfellow dieron a conocer en conjunto varias mejoras que idearon para GAN, una de las cuales consiste en dar periódicamente a la red D como entrada ejemplos etiquetados y minimizar un error de clasificación convencional [17].
Esta estrategia de entrenamiento tiene la gran ventaja de que permite utilizar un enorme cúmulo de imágenes de rostros sin etiquetar, o etiquetadas para otra tarea, que existe de forma pública en la web. De esta forma se logra sobrellevar el problema de tener una cantidad sumamente reducida de datos correctamente etiquetados con la emoción predominante del rostro en la imagen.
Recientemente se ha explorado la utilidad de la metodología en variadas aplicaciones. Se aprovechan tanto la posibilidad de generar nuevos datos similares a los del conjunto de entrenamiento no supervisado, como la opción de reutilizar la representación del modelo discriminador de los datos para tareas de clasificación[31]. No se tiene conocimiento de que se hayan aplicado estas técnicas para clasificación de emociones. En el presente trabajo se usa un procedimiento similar, desarrollado de forma independiente, y se lo aplica a los conjuntos de datos CK+: Cohn Kanade Extended y SFEW 2.0.

## IV. Conjuntos de Datos

Durante el proceso de desarrollo se utilizaron varios conjuntos de datos, con características muy distintas entre

sí. Cada uno respondió a una necesidad y fue utilizado con cierto objetivo. A continuación se presentará en detalle cada uno.

### A. FER 2013

Este conjunto fue confeccionado para el concurso 'Facial Expresion Recognition 2013' (FER 2013) [26] organizado por Kaggle [1]. Consta de 35887 imágenes, recolectadas de forma semi automática, mediante una metodología basada en la API del motor de búsqueda de Google.
Se hicieron cadenas de palabras combinando conceptos relacionados al género, a diferentes edades y etnias, con 181 palabras claves asociadas con estados emocionales, como por ejemplo "odio" o "dichoso".
Luego se ejecutó el algoritmo de detección de rostros de openCV obteniendo regiones de interés en cada imagen. Por último, se terminaron de corregir los recortes y etiquetar correctamente el conjunto de forma manual.
Mapeando las 181 palabras claves hacia las 6 emociones básicas de Paul Ekman, y la expresión neutral. Se obtuvieron 7 conjuntos de imágenes separadas por clase. Las imágenes son de 48x48 píxeles, con lo cual el conjunto es muy liviano. Así los organizadores terminaron por confeccionar un conjunto de datos donde el rostro está en la zona central de la imagen y se puede asumir que hay exactamente un rostro en cada una de ellas. Al igual que GENKI-4K, FER 2013 posee enorme cantidad de sujetos diferentes, además de mucha varianza en el brillo y la posición de los rostros en la imagen.
FER 2013 fue utilizado como conjunto de apoyo para realizar transferencia de conocimiento al momento de hacer reconocimiento de emociones directamente con etiquetas de las 6 emociones básicas.

### B. CASIA

Este conjunto de datos fue creado para el estudio del reconocimiento de la identidad en rostros, con el objetivo de que fuera público y de un tamaño mucho mayor a cualquier otro conjunto disponible a la comunidad académica [27].
Al igual que FER, se realizó mediante un método semi automático, basado en recolectar imágenes de rostros de celebridades de la página IMdB [2]. Las imágenes fueron luego etiquetadas usando los metadatos de la página, entre los que se encontraba el nombre de cada celebridad.
Así se obtuvo un conjunto de datos de casi 500000 imágenes, con mas de 10000 sujetos diferentes. El conjunto fue diseñado y realizado para que pueda ser compatible con 'labeled Faces in the Wild' otro conjunto para reconocimiento de identidad con características similares. Actualmente se suelen distribuir juntos y se les ha realizado algunas operaciones para el aumento de la cantidad de datos, como por ejemplo el espejado. De esta forma se ha conseguido obtener un conjunto de mas de un millón de fotografías.

[1] https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge
[2] http://www.imdb.com/

Si bien el conjunto fue pensado para el entrenamiento de la tarea de reconocimiento de identidad, en el presente trabajo es usado como un conjunto de datos auxiliar, que permite a los modelos aprender características y patrones típicos del rostro humano.
Se usó aproximadamente el 10 % del conjunto, que de todas formas es mas de 3 veces mayor al tamaño de FER2013, el segundo conjunto de mayor tamaño con el cual se trabajó. Esto fue así por razones de capacidad computacional y tiempo.

### C. CK+

Es el conjunto de datos hecho por Jeffrey Cohn y Takeo Kanade [4]. Es extremadamente usado ya que fue uno de los primeros dedicado a la clasificación de emociones y fue tomado como referencia por la comunidad académica. Fue realizado y etiquetado de forma totalmente manual, y es de los conjuntos mas pequeños con el cual se trabajó.
Tiene características de laboratorio, con expresiones faciales actuadas, no motivadas por ningún agente externo. Todas las imágenes fueron tomadas con el mismo fondo, la misma iluminación, la misma cámara.
Cuenta con 593 secuencias de 107 sujetos, cada secuencia de entre 5 y 60 imágenes, las cuales van de forma progresiva desde la expresión neutral hasta el pico de una determinada emoción. Sólo la última imagen de cada secuencia está codificada con FACS, y la mayoría de los investigadores usan solo esta imagen para entrenar sus modelos, o el último 20% de las imágenes de cada secuencia. Las mismas tienen asociada una emoción predominante, con lo cual el conjunto puede ser dividido y etiquetado según las 7 emociones básicas.
En el presente trabajo este conjunto fue usado como uno de los conjuntos objetivo, a la hora de hacer reconocimiento de emociones de forma directa, sin la etapa intermedia de la detección de AU activas. Se eligió usar el último 20 % de las imágenes, para respetar el protocolo que utilizan otros trabajos y a la vez no caer en el uso de un conjunto con imágenes demasiado 'artificiales' y de expresiones exageradas.

Figura 2. Ejemplos del conjunto de datos CK+.Se muestran 8 imágenes de una secuencia de 11. La última imagen corresponde al pico de la expresión facial, la primera a la expresión neutral.

### D. SFEW 2015

SFEW 2015 es otro conjunto de datos confeccionado especialmente para un concurso, que luego fue evolucionando y tomado como referencia [7]. Es un conjunto de datos

que se suele tomar como parámetro cuando se habla de reconocimiento de emociones 'in the wild'.
Fue realizado de forma semi automática, mediante un recomendador basado en subtítulos. Se comenzó desde un total de 54 películas en DVD. Se extrajeron los subtítulos y los subtítulos especiales para personas con capacidades diferentes, los cuales vienen acompañados de palabras claves sobre las emociones de los personajes ([SORPRENDIDO], [TRISTE], [AVERGONZADO], etc.). Luego se hizo un sistema que aceptaba búsquedas por palabras claves y recomendaba videos relacionados con dichas palabras. Con este procedimiento se eligieron aproximadamente 1000 clips de entre 1 y 5 segundos. Apoyados por las palabras claves de la búsqueda y confirmados por los etiquetadores, se separaron los videos en 7 grupos según la emoción predominante del personaje principal del videoclip.
El conjunto de datos fue sufriendo pequeñas modificaciones en cada edición del concurso, y además se confeccionó un subconjunto de imágenes llamadas Static Facial Expresions in the Wild (SFEW). En el presente trabajo usaremos la segunda versión de SFEW como conjunto 'objetivo', al hacer las pruebas de reconocimiento de emociones 'in the wild' de forma directa, sin pasar por el estadio intermedio de las AU.

Figura 3. Ejemplos del conjunto de datos SFEW. Se puede apreciar la gran variabilidad en los datos, luz, posición del rostro, expresiones faciales que proponen ser mas naturales que las actuadas en el laboratorio.

## V. Preprocesamiento

En esta sección se detallarán las diferentes metodologías de preprocesamiento de los datos, efectuados antes de comenzar con los entrenamientos de modelos profundos.
Se desarrollaron dos metodologías, respondiendo al hecho de que los diferentes conjuntos de datos ya poseían diferentes tratamientos de base. Además, las imágenes en algunos conjuntos eran demasiado pequeñas (por ejemplo FER2013 posee imágenes de 48x48 píxeles) para realizar sobre ellas algunas operaciones, como ser la detección de pupilas para una posterior alineación del rostro. De esta forma, se aplicó a cada conjunto una, otra, o ambas metodologías, según fue más adecuado.
Las características principales que se pretendieron normalizar fueron:

- Todos los conjuntos in the wild, poseían una amplísima variabilidad en la iluminación. Para disminuir esto, se utilizó la técnica de ecualización adaptativa del histograma que provee openCV.
- Por otro lado, se intentó detectar y recortar el rostro presente en la imagen. De esta forma pueden descartarse imágenes donde ningún rostro aparezca, y si por el contrario aparecen varios, se puede detectar detectar el principal basándose en heurísticas, por ejemplo mayor área, posición central en la imagen, y así recortar sólo la región de interés.
- Se trabajó durante todo el proyecto con imágenes en escala de grises, esto fue así ya que había algunos conjuntos que ya estaban presentados de esta forma.
- Se advirtió que era importante que los rostros estuviesen lo más centrados posibles en la imagen. Con lo cual se uso una metodología de detección de pupilas y luego alineación en base a la posición de estos puntos claves. Además esto permitió hacer un recorte mucho mas fino del rostro a clasificar, dando la ventaja adicional de colocar todos los rostros a la misma escala. Esto fue beneficioso ya que en los diferentes conjuntos hay rostros mas lejanos a la cámara que otros, lo que ocasiona diferencias de tamaño y hasta de proporción de un rostro al siguiente.

### A. *Primer Metodología de preprocesamiento*

Esta primer parte se realizó para los conjuntos en donde podía haber una cantidad variable de rostros, y una gran porción de la imagen pertenecía al fondo. Se buscó implementar un primer filtro que quite las imágenes sin ningún rostro o con varios de ellos. Al mismo tiempo se recortó la mayor parte del fondo, dejando la imagen lista para ser procesada por la segunda metodología. Además, se mejoró la iluminación.
Los diferentes pasos que se siguieron en esta metodología fueron:

- Pasar la imagen inicial a escala de grises, de ser necesario.
- Ecualizar su histograma de forma adaptativa, por sectores, para mejorar brillo y contraste.
- Detectar la zona de interés, es decir el rostro principal sobre el cual se hará luego la clasificación de emociones. También fue necesario detectar cuando no hay rostro presente en la imagen y así descartar la misma.

- Una vez detectado el rostro se recorta le región de interés y se escala la imagen a un tamaño uniforme.

En la figura 4 se pueden apreciar varios ejemplos de imágenes procesadas con este primer paso.

Figura 4. Ejemplos de imágenes procesadas con la primer metodología desarrollada.

Figura 5. Ejemplos del conjunto SFEW procesados con la segunda metodología desarrollada. Se incluyen también 4 ejemplos de la figura 4 para resaltar las diferencias entre los resultados de ambas metodologías. Nótese particularmente como en el segundo caso se logró quitar el fondo sobrante y alinear mejor el rostro en la imagen.

### *B. Segunda Metodología de preprocesamiento*

El objetivo de la segunda fase de preprocesamiento fue terminar de recortar el rostro de forma mas precisa y además centrar el rostro en la imagen. De esta forma el fondo quedaría prácticamente descartado y esto le permitiría a los modelos centrarse en aprender de la información relevante para la tarea. Esto se logró aplicando el algoritmo de detección de pupilas proveído por openCV. El mismo es una implementación del método planteado por Viola y Jones en 2001[23]. Es un método de boosting que usa como modelo básico árboles de decisión, que toman como entradas miles de características calculadas mediante filtros previamente aprendidos en una etapa de entrenamiento, que son aplicados a una cierta región o parche en la imagen. Este procedimiento se repite moviendo el parche por toda la imagen, y a diferentes escalas. Así, el algoritmo devolverá las zonas donde la probabilidad de que allí esté el objeto buscado sea mayor que un cierto límite previamente establecido.

Luego de detectar las pupilas, se aplicó una heurística para descartar falsos positivos y se tomó la distancia entre ellas. En base a esta medida se recortó el rostro de forma más precisa para descartar el fondo que pudiera haber quedado en el paso anterior.

Por último se realizaron operaciones de traslación rotación y escalado en base a la posición de las pupilas para dejarlas posicionadas en el mismo lugar en todas las imágenes. Aquí se introduce un cierto error producido por imperfecciones en la posición de las pupilas, pero luego de algunas pruebas se concluyó que el mismo es aceptable. En la Figura 5 se pueden ver ejemplos de imágenes del conjunto de datos SFEW procesados con la esta segunda metodología.

## VI. Experimentación

La experimentación con clases de emociones se hizo en dos etapas, con el objetivo de estudiar dos metodologías de transferencia de conocimiento y poder compararlas.

La primer estrategia fue el enfoque mas tradicional, donde todos los pesos de los modelos entrenados para la tarea $t$ son transferidos a la tarea $t'$. La granularidad más fina en este enfoque está dada por transferir ciertas capas y otras no, por lo general se suelen pasar las primeras capas, las mas cercanas a la entrada, ya que se espera que aprendan información más general, aplicable a varios tipos de tareas.

En el presente trabajo se transfirieron todas las capas salvo la última, la cual toma la decisión sobre la clasificación final, esto fue así ya que los conjuntos de datos de apoyo estaban pensados para la misma tarea, el reconocimiento de emociones con lo cual los características de alto contenido semántico aprendidos por las capas mas cercanas a la salida de los modelos también podían ser transferidos.

La segunda estrategia fue utilizar el modelo Generative Adversarial Networks, modificado para aceptar un conjunto de datos no etiquetado, y otro etiquetado. De esta forma, se agrega a la función a minimizar un término correspondiente al error de clasificación. Así, el modelo puede aceptar mini-batchs de datos con o sin etiquetas y en cada caso la función de error se adaptará.

### *A. Transferencia de Conocimiento convencional*

En la etapa de Transferencia de Conocimiento se utilizaron los conjuntos de datos CK+ y SFEW, de características muy disímiles entre sí, como conjuntos de datos objetivos, y se utilizó el conjunto FER2013 como conjunto de apoyo. FER2013 está etiquetado por emoción y consta de más de 35000 imágenes, un orden de magnitud más que los otros dos conjuntos.

Los modelos utilizados fueron:

- **VGG16:** Un modelo de 16 capas, muy profundo, desarrollado por le grupo de visión por computadora de la universidad de Oxford [28].
- **VGG-N-2048:** Desarrollado por el mismo grupo, pero de aproximadamente la mitad del tamaño. En el trabajo donde desarrollan este modelo, también se estudian varias características de implementación de modelos profundos, que luego tendrían repercusión e inspiraron otros modelos [29].
- **SqueezeNet:** Este es el modelo mas pequeño con el cual se hicieron las pruebas, y el más rápido de entrenar. Mas adelante se verá que cuando los otros dos modelos tuvieron problemas de convergencia por ser demasiado grandes para la cantidad de datos disponibles, este fue el único que convergió hacia un mínimo la función de Loss [30].

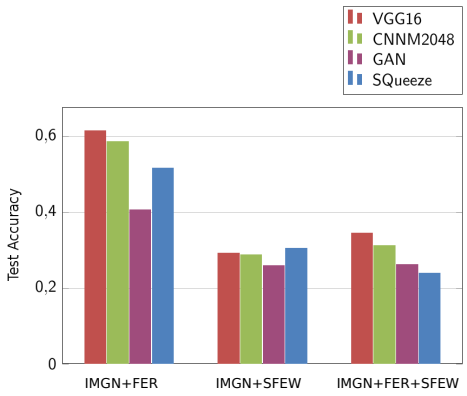


Figura 6. Resultados de las pruebas con los conjuntos de datos FER2013 y SFEW. En la primer columna vemos los resultados de cada modelo con FER2013 comenzando de pesos pre entrenados con Imagenet. En la segunda vemos los resultados de SFEW pre entrenado con imagenet. En la tercer columna vemos la precisión sobre SFEW luego de haber comenzado con los pesos resultantes del entrenamiento de FER2013.

El procedimiento en ese caso fue el siguiente:

- Entrenar 4 modelos de características diferentes con el conjunto de datos SFEW, partiendo de pesos pre-entrenados con el conjunto imagenet. Imagenet no posee rostros humanos dentro de sus categorías, sin embargo, los filtros aprendidos en la primeras capas, generalmente asociados a filtros de Gabor y en general de detección de bordes, pueden ser útiles para esta tarea.
- Realizar otro entrenamiento con el conjunto de apoyo, FER2013, también partiendo de los modelos entrenados con imagenet.
- Hacer un Fine-tuning con SFEW partiendo de los pesos previamente obtenidos de FER2013, y comparar los resultados de las pruebas con y sin este último refinamiento.
- Repetir el procedimiento anterior con el conjunto CK+.

De esta forma se obtienen resultados de reconocimiento de emociones tanto en un conjunto de datos confeccionado en laboratorios como en otro conjunto con características 'in the wild'.
En la figura 6 se pueden apreciar los resultados de esta etapa con el conjunto SFEW y en 7 se encuentran los resultados con CK+. Cabe destacar que en este ocasión se agregó al conjunto de modelos utilizado en las pruebas anteriores, la arquitectura correspondiente al discriminador de la metodología GAN. Esto fue así ya que existe evidencia de que el modelo logra aprender características útiles en objetos complejos en la imagen como era en este caso el rostro humano. Sin embargo en esta etapa sólo se experimentó con la arquitectura del discriminador de forma independiente. En la etapa siguiente veremos como se utilizó el sistema GAN completo para realizar un entrenamiento semi-supervisado.
En este caso las conclusiones fueron las siguientes:

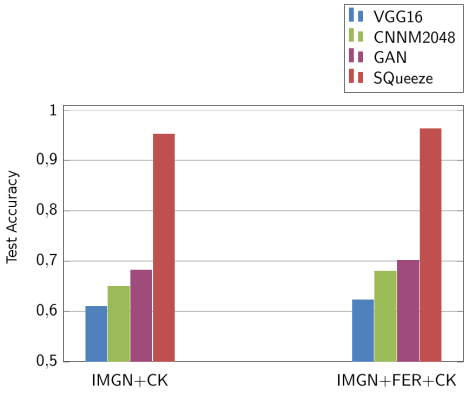


Figura 7. Resultados de las pruebas con el conjunto Cohn Kanade Extended. En la primer columna pueden verse la precisión de cada modelo con pesos pre entrenados con el conjunto Imagenet. En la segunda se ven los resultados de los entrenamientos comenzando por los pesos pre-entrenados con FER2013

- En cuanto al conjunto FER2013, los resultados fueron en general satisfactorios, comparables a otros trabajos realizados sobre este conjunto [19][20][21]. Esto se debe a lo siguiente, los modelos logran hacer un buena generalización de los datos a partir del entrenamiento, sin tender al sobre ajuste, probablemente porque FER2013 posee una enorme variedad de sujetos diferentes, y en la imagen el rostro ocupa la mayor parte de la superficie.
- Observando los datos de SFEW podemos ver un notorio cambio en cada una de las arquitecturas. Por un lado, las arquitecturas mas profundas tuvieron un incremento importante de aproximadamente un 5 %, sobre todo VGG16. Por otro lado las arquitecturas menos profundas, SqueezeNet y el discriminador de GAN sufrieron un decremento de la precisión.
- En cuanto a los resultados de CK+ podemos decir que mantuvieron la correspondencia entre los diferentes modelos, antes y después del proceso de transferencia de conocimiento. Se observa un incremento parejo en

cada una de las arquitecturas. Además, es importante recordar que el estado del arte en este conjunto esta sumamente avanzado [8] con lo cual el único resultado obtenido comparable es el de la arquitectura SqueezeNet. Es esperable que esta arquitectura sea la que mejor se comporte en un conjunto como CK+, ya que es sabido que los modelos demasiado profundos tienen problemas para aprender de conjuntos de datos pequeños. [22].

### *B. Transferenca de Conocimiento utilizando GAN*

Se utilizó el modelo GAN, el cual se modificó para aceptar de forma periódica mini-batch de datos etiquetados, en adición a los datos no etiquetados que toma para realizar el entrenamiento competitivo, no supervisado.

Se usó CASIA como conjunto de datos no etiquetados, el cual se prefirió ante otros conjuntos disponibles sin etiquetar por ser varias veces más grande, poseer enorme cantidad de sujetos y características 'in the wild'. CASIA es un conjunto de datos abierto que fue construido a partir de imágenes de rostros de figuras públicas reconocidas y se ideó originalmente para el reconocimiento de identidad.

Se utilizaron SFEW y CK+ como conjuntos de datos etiquetados. El primero es el conjunto de datos que se utiliza en la competencia anual EMOTIW, que justamente busca que cada equipo participante prediga las emociones predominantes utilizando dicho conjunto para entrenamiento y testeo. En el presente trabajo se utilizó para realizar pruebas 'in the wild', con expresiones más naturales, espontáneas, y se aprovechó la oportunidad para compararse con los resultados de la edición 2016 del concurso.

CK+ es el conjunto de datos más ampliamente utilizado para hacer pruebas de detección de emociones en ambientes controlados, con lo cual en el presente trabajo se utilizó para testear el procedimiento en un ambiente de laboratorio, con expresiones faciales espontáneas.

Cabe destacar que el conjunto de datos no etiquetados consta de unas 100.000 imágenes, con la posibilidad de ser extendido, en contraste a los conjuntos de datos etiquetados, que cuentan con aproximadamente 1.500 imágenes cada uno. Aquí queda en evidencia la gran ventaja que significa tener disponible el conocimiento aprendido por el discriminador luego de haber inspeccionado varios miles de imágenes de rostros y usarlo de forma **online** para entrenar de manera supervisada.

Hasta donde se sabe, este aspecto es original ya que los trabajos donde se utilizó GAN como apoyo a tareas de clasificación lo hicieron de forma offline, es decir en un paso posterior al entrenamiento no supervisado. Además, lo que se usa en esos casos para el entrenamiento supervisado son las características que surgen de la última capa del discriminador, no de todo el modelo.

Figura 8. Ejemplos de rostros generados por el modelo GAN en la primera iteración de la prueba con SFEW.

El procedimiento realizado fue el siguiente:

- Los tres conjuntos utilizados, CK+, SFEW y CASIA fueron escalados a 64x64. Además, los tres fueron previamente procesados con metodologías de procesamiento de imágenes desarrolladas con el foco en normalizar luz, centrar el encuadre del rostro en la imagen y recortar el fondo.
- Para SFEW, se utilizaron los conjuntos de validación y testeo proveídos por la organización del concurso.
- Para las pruebas con el conjunto CK+ se hizo un K-Fold con K = 5. Cada uno de los 5 conjuntos resultantes contaba con 530 imágenes.
- Los resultados fueron comparados con un procedimiento análogo aplicado sobre CK+ y SFEW, pero utilizando la metodología mas tradicional de transferencia de conocimiento, esto es realizar varias fases de entrenamiento puramente supervisado y reutilizar los pesos sinápticos de las primeras N capas.

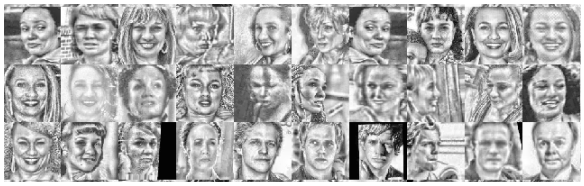

Figura 9. Ejemplos de imágenes preprocesadas del conjunto CASIA, nótese la gran similaridad de estos ejemplos con las imágenes generadas en la última iteración de GAN.

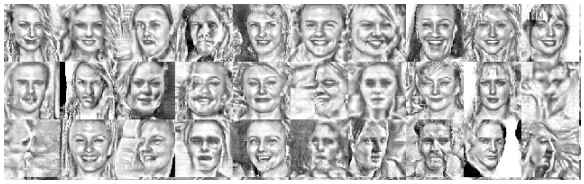

Figura 10. Ejemplos de rostros generados por el modelo GAN en la última iteración de la prueba con SFEW.

En las Figuras 8 y 10 se pueden apreciar ejemplos de salidas del modelo generador luego de la primera y última iteración de la prueba realizada con SFEW. Se consiguió una precisión de **44,97 %** en el conjunto de test, en las pruebas con SFEW. En el 5-Fold sobre el conjunto CK+ se alcanzó el **95,66 %** con **(1,02)** de desvío estándar en test.

## VII. Conclusión

Luego de haber realizado una revisión del estado del arte, desarrollado metodologías de procesamiento de imágenes para

normalizar ciertos aspectos y poder trabajar 'in the wild', y realizar por último las pruebas mencionadas en la sección anterior, podemos concluir:

- En cuanto a las pruebas en CK+, fueron satisfactorias y son comparables con otros trabajos realizados sobre este conjunto [3][6]. Además, es positivo el hecho de haber logrado un porcentaje de exactitud comparable, mediante métodos que usan características de la imagen aprendidas de forma automática, las cuales pueden ser reusadas o pueden aportar avances para solucionar otro tipo de tareas. Esto es en contraste con otro tipo de características desarrolladas de forma artesanal, que en líneas generales pueden dar excelentes resultados pero son de uso acotado a la tarea que pretenden resolver.
- Por el lado de las pruebas realizadas en el conjunto SFEW, el resultado de las pruebas también es sumamente satisfactorio. Si bien no está entre los primeros lugares comparado con la edición 2015 del concurso, la mayoría de los resultados finales del concurso pertenecen no a un modelo único sino a un conjunto de ellos. En muchos casos, los resultados del modelo original a partir del cual se crean estos conjuntos son similares o inferiores a los presentados en este trabajo[14] [18].
- Se utilizó de manera novedosa la metodología de Generative Adversarial Networks, para reutilizar los parámetros entrenados de forma no supervisada en una tarea de clasificación de forma online. Por otro lado, se lograron generar de forma artificial imágenes visualmente muy similares a las del conjunto de entrenamiento sin etiquetar (CASIA).

## Referencias


[1] Blackmore, Simon. Farming with robots 2050. Presentation delivered at Oxford Food Security Conference. 2014.

[2] Observer-Based Measurement of Facial Expression With the Facial Action Coding System, Ekman, Paul and Friesen, Wallace and Hager, John ,Facial Action Coding System: Research Nexus. Network Research Information, Salt Lake City, UT, USA, 2002

[3] Continuous au intensity estimation using localized, sparse facial feature space, Jeni, László A and Girard, Jeffrey M and Cohn, Jeffrey F and De La Torre, Fernando, Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, pages 1–7, 2013

[4] The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, Lucey, Patrick and Cohn, Jeffrey F and Kanade, Takeo and Saragih, Jason and Ambadar, Zara and Matthews, Iain, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pages 94–101, 2010

[5] A method to infer emotions from facial action units, Velusamy, Sudha and Kannan, Hariprasad and Anand, Balasubramanian and Sharma, Anshul and Navathe, Bilva, 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2028–2031, 2011,

[6] Foundations of human computing: facial expression and emotion, Cohn, Jeffrey F, Proceedings of the 8th international conference on Multimodal interfaces, pages 233–238, 2006,

[7] Collecting large, richly annotated facial-expression databases from movies, Dhall, Abhinav and others, 2012,

[8] Emotional expression classification using time-series kernels, Lorincz, Andras and Jeni, Laszlo and Szabo, Zoltan and Cohn, Jeffrey and Kanade, Takeo, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 889–895

[9] Imagenet classification with deep convolutional neural networks, Krizhevsky, Alex and Sutskever, Ilya and Hinton, Geoffrey E, Advances in neural information processing systems, pages 1097–1105,

[10] Real-time emotion recognition for gaming using deep convolutional network features, Ouellet, Sébastien, arXiv preprint arXiv:1408.3750, 2014

[11] Evaluation of vision-based real-time measures for emotions discrimination under uncontrolled conditions, Gómez Jáuregui, David Antonio and Martin, Jean-Claude, Proceedings of the 2013 on Emotion recognition in the wild challenge and workshop, pages 17–22, 2013,

[12] A Deep Learning Approach for Subject Independent Emotion Recognition from Facial Expressions, Neagoe, Victor-Emil and Andrei-Petru, Brar and Sebe, Nicu and Robitu, Paul, Recent Advances in Image, Audio and Signal Processing, pages 93–98, 2013

[13] Deep learning for emotion recognition on small datasets using transfer learning, Ng, Hong-Wei and Nguyen, Viet Dung and Vonikakis, Vassilios and Winkler, Stefan, Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pages 443–449, 2015,

[14] Emotion recognition in the wild via convolutional neural networks and mapped binary patterns, Levi, Gil and Hassner, Tal, Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pages 503–510, 2015

[15] Generative adversarial nets, Goodfellow, Ian and Pouget-Abadie, Jean and Mirza, Mehdi and Xu, Bing and Warde-Farley, David and Ozair, Sherjil and Courville, Aaron and Bengio, Yoshua, Advances in Neural Information Processing Systems, pages 2672–2680, 2014

[16] Unsupervised representation learning with deep convolutional generative adversarial networks, Radford, Alec and Metz, Luke and Chintala, Soumith, arXiv preprint arXiv:1511.06434, 2015

[17] Improved techniques for training gans, Salimans, Tim and Goodfellow, Ian and Zaremba, Wojciech and Cheung, Vicki and Radford, Alec and Chen, Xi, arXiv preprint arXiv:1606.03498, 2016

[18] Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition, Kim, Bo-Kyeong and Lee, Hwaran and Roh, Jihyeon and Lee, Soo-Young, Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pages 427–434, 2015

[19] Fast and robust smile intensity estimation by cascaded support vector machines, Shimada, Keiji and Noguchi, Yoshihiro and Kuria, Takio, International Journal of Computer Theory and Engineering, 2013

[20] Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, Hinton, Geoffrey and Deng, Li and Yu, Dong and Dahl, George E and Mohamed, Abdel-rahman and Jaitly, Navdeep and Senior, Andrew and Vanhoucke, Vincent and Nguyen, Patrick and Sainath, Tara N and others, IEEE Signal Processing Magazine, pages 82–97, 2012

[21] Facial expression analysis based on high dimensional binary features, Kahou, Samira Ebrahimi and Froumenty, Pierre and Pal, Christopher, European Conference on Computer Vision, pages 135–147, 2014

[22] Understanding the difficulty of training deep feedforward neural networks, Glorot, Xavier and Bengio, Yoshua, Aistats, volume 9, pages 249–256, 2010

[23] Rapid object detection using a boosted cascade of simple features, Viola, Paul and Jones, Michael, Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, volume 1, I–511, 2001

[24] Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected, McDuff, Daniel and Kaliouby, Rana and Senechal, Thibaud and Amr, May and Cohn, Jeffrey and Picard, Rosalind, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 881–888, 2013

[25] Toward practical smile detection, Whitehill, Jacob and Littlewort, Gwen and Fasel, Ian and Bartlett, Marian and Movellan, Javier, IEEE transactions on pattern analysis and machine intelligence, volume 31, pages 2106–2111, 2009

[26] Challenges in representation learning: A report on three machine learning contests, Goodfellow, Ian J and Erhan, Dumitru and Carrier, Pierre Luc and Courville, Aaron and Mirza, Mehdi and Hamner, Ben and Cukierski, Will and Tang, Yichuan and Thaler, David and Lee, Dong-Hyun and others, International Conference on Neural Information Processing, pages 117–124, 2013

[27] Learning face representation from scratch, Yi, Dong and Lei, Zhen and Liao, Shengcai and Li, Stan Z, arXiv preprint arXiv:1411.7923, 2014

[28] Very deep convolutional networks for large-scale image recognition, Simonyan, Karen and Zisserman, Andrew, arXiv preprint arXiv:1409.1556, 2014

[29] Return of the devil in the details: Delving deep into convolutional nets, Chatfield, Ken and Simonyan, Karen and Vedaldi, Andrea and Zisserman, Andrew, arXiv preprint arXiv:1405.3531, 2014

[30] Forrest N. Iandola and Matthew W. Moskewicz and Khalid Ashraf and Song Han and William J. Dally and Kurt Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size, arXiv:1602.07360, 2016

[31] Learning Representations of Emotional Speech with Deep Convolutional Generative Adversarial Networks, Chang, Jonathan and Scherer, Stefan, arXiv preprint arXiv:1705.02394, 2017