Facial Expression Synthesis Using Self-Organizing Map

Alicia Alvarez Mon[1], María Elena Buemi[1,2], and Daniel Acevedo[1,2]

[1] Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Computación. alvarezmonaliciaa@gmail.com

[2] CONICET-UBA. Instituto de Investigación en Cs. de la Computación (ICC). Ciudad Autónoma de Buenos Aires, Argentina. {mebuemi,dacevedo}@dc.uba.ar

Resumen This model generates emotional facial expressions from a neutral image of a subject, based on a target emotion specified by the user. The emotional space guiding the synthesis process is represented through a structure based on Self-Organizing Maps (SOMs), trained in an unsupervised manner. These maps topologically group elements of the same class in proximity, using descriptors that capture information about facial shape, structure, geometry, and texture. The synthesis process selects the best representation of a subject from the map and applies a controlled deformation through linear interpolation and Gaussian functions, allowing the emotional intensity to be adjusted in a granular way. An external classifier provided by the OpenCV library was used to validate the synthesis results. The findings show that the images generated by the model exhibit classification behavior comparable to that of genuine expressions. However, synthesized images show a greater tendency to be misclassified under the "Happiness" and "Sadness" categories.

Keywords: Facial expression, Synthesis, Image processing, Features, Expression recognition

Síntesis de expresiones faciales mediante mapas Auto-organizados

Resumen Dada la imagen neutral de un sujeto, este modelo genera una imagen con una emoción objetivo la cuál es solicitada por el usuario. El espacio emocional que guía al proceso de síntesis se representa mediante una estructura basada en Mapas Auto-Organizados (SOM) que se entrenan de forma no supervisada y disponen topológicamente cerca aquellos elementos pertenecientes a una misma clase. Para dicho entrenamiento se emplean descriptores que capturan información relativa a la forma, la estructura, la geometría y la textura facial. La síntesis selecciona del Mapa la mejor representación de un sujeto y aplica una deformación controlada mediante interpolación lineal y funciones gaussianas para generar la imagen, pudiendo ajustar la intensidad de la emoción de forma granular. Se empleó un clasificador externo provisto por la librería OpenCV para validar

los resultados de síntesis, evidenciando que las imágenes generadas por el modelo presentan un comportamiento de clasificación comparable al de aquellas con expresiones genuinas; no obstante, se observa una mayor propensión de las imágenes sintetizadas a ser clasificadas erróneamente en las categorías “Happiness” y “Sadness”.

Keywords: Expresiones faciales, Síntesis, Procesamiento de imágenes, Features, Reconocimiento de Expresiones

1. Introducción

La síntesis de expresiones faciales se refiere al proceso de generar imágenes de rostros humanos que muestran una de las seis emociones básicas: Anger, Disgust, Fear, Happiness, Sadness, y Surprise (mantenemos sus nombres en inglés tal como son referenciadas en la comunidad). Esta tarea tiene aplicaciones significativas en la interacción humano-computadora, realidad virtual, animación y affective computing, donde las expresiones faciales realistas y controlables son esenciales para mejorar la experiencia del usuario y la comunicación.

En el proceso de síntesis, los datos de entrada incluyen: (a) una imagen de rostro neutral del sujeto, (b) la emoción objetivo a sintetizar, (c) la intensidad de emoción deseada. Nuestro enfoque se basa en el modelo de Mapa de Expresión (XM) introducido por (Agarwal & Mukherjee, 2019), el cual emplea Mapas Auto-Organizados (SOMs) para agrupar vectores de características extraídos de imágenes de entrenamiento. Las características incluyen forma, estructura, representaciones tipo caricatura y textura, permitiendo que la red codifique tanto las variaciones geométricas como de apariencia de las expresiones faciales. Cada neurona en el XM representa una estructura facial prototípica, la cual se utiliza posteriormente para guiar el proceso de síntesis. Dada una imagen de entrada neutral, el XM nos permite identificar la neurona cuyas características mejor coinciden con la estructura del rostro de entrada, permitiendo una transformación controlada hacia la emoción objetivo. La combinación de datos mediante interpolación lineal y mezcla de características ponderada por Gaussianas asegura que la expresión sintetizada preserve la identidad y la estructura general de la entrada, mientras refleja la intensidad emocional solicitada.

En este trabajo, analizamos el pipeline de síntesis. Además, evaluamos las imágenes generadas cuantitativamente utilizando el conjunto de datos CK+ y un clasificador FER externo. Los resultados demuestran la capacidad del método para producir expresiones realistas, particularmente para Happiness y Sadness, y resaltan limitaciones potenciales para emociones tales como Disgust y Fear.

2. Trabajo relacionado

La síntesis de expresiones faciales consiste en producir una emoción determinada en una persona objetivo para quien esa expresión en particular es desconocida. Este problema ha sido estudiado utilizando varios enfoques, que van desde modelos que requieren grandes conjuntos de datos hasta aquellos que utilizan datos limitados, y con diferentes

niveles de complejidad computacional. El tema ha sido ampliamente explorado con el objetivo de lograr animaciones faciales realistas.

En un trabajo previo se presenta un marco para generar una animación de una expresión objetivo, requiriendo un dataset de expresiones y conocer videos del sujeto para poder realizar el mapeo del rostro de la persona. Este enfoque calcula métricas que dan cuenta del movimiento a través de los cuadros de mapeo y aplica algoritmos de camino más corto para mantener la coherencia temporal (Li et al., 2012).

Otros métodos requieren solo una imagen neutral de la persona objetivo, y emplean núcleos bilineales, Regresión de Rango Reducido con Núcleo Bilineal (BKRRR) y factorización de dos factores basada en núcleos, respectivamente (Huang & De la Torre, 2010), (Zhou & Lin, 2005). Otro enfoque combina Imágenes de Relación de Expresión (ERI) y Modelos de Apariencia Activos (AAM) para aprovechar tanto los componentes de textura como los de forma para la síntesis (Xiong et al., 2007). Con Aprendizaje Profundo Conjunto, se combinan la síntesis y el reconocimiento de expresiones a través de una Red Adversaria Generativa de Síntesis de Expresiones Faciales (FESGAN) para el pre-entrenamiento y la generación del modelo (Yan et al., 2020). También hay técnicas que utilizan la deformación de la imagen neutral, utilizando datos que considera similares para guiarla (Dong et al., 2024). El trabajo (Luiz Testa et al., 2025) presenta un modelo GAN que genera videos de emociones desde rostros neutrales; el sistema usa descriptores para hallar un video de referencia y combina mapas de desplazamiento con datos temporales.

3. Síntesis de Emociones

En el proceso que usamos para sintetizar emociones utilizamos el dataset CK+ (Lucey et al., 2010), conformado por videos, cuyos frames tienen tamaño 640×490 en escalas de grises. Los videos corresponden a 123 sujetos distintos, que actúan 593 emociones en total. Seleccionamos los videos etiquetados con alguna de las 6 emociones básicas: Anger, Disgust, Fear, Happiness, Sadness y Surprise. Descartamos los videos que no están etiquetados y modificamos aquellos que consideramos mal etiquetados o sin emoción clara. En total quedan 317 videos, de los que tomamos el 70 % de los datos para entrenamiento y utilizamos el 30 % restante para testeo.

3.1. Obtención de características (*features*)

Para la representación de emociones en la que se basa la síntesis, se utiliza un modelo de entrenamiento que distingue *features* de forma, estructura y textura de las imágenes para entrenar un clasificador de emociones de manera no supervisada utilizando un Mapa de Expresiones (XM). Para el feature de forma se calcula la diferencia de los landmarks de cada frame de video con el frame inicial del mismo. Para el feature de estructura se utilizan 22 valores que representan medidas sobre los landmarks y sobre la textura, y se realiza una descomposición TV-L1 de la imagen para obtener los *features* de geometría y de textura. Este modelo se detalla en (Alvarez Mon et al., 2024).

Se utiliza Análisis de Componentes Principales (PCA) en cada feature para reducir su dimensión sin perder información relevante y cada imagen con la que entrenamos el

mapa de expresiones se representa con el vector resultante al concatenar las *features*. De cada video, utilizamos solamente el frame inicial y el frame final. El frame inicial de la serie expresa la emoción 'neutral', y el final expresa la emoción en su pico de intensidad (Apex). Podemos ver en la Figura 1 un ejemplo de un mapa entrenado donde se observa que cada expresión es asignada a grupos cercanos de neuronas.

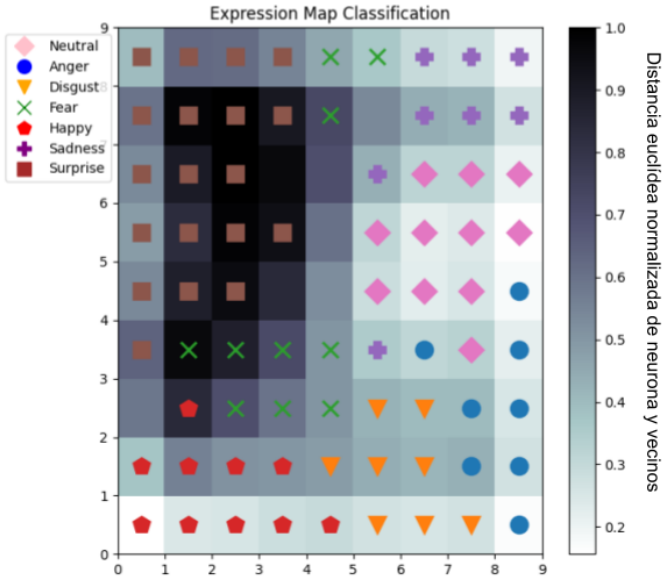


Figura 1. Mapa de Expresiones caracterizando el agrupamiento de diferentes expresiones para una grilla de 9×9 neuronas.

3.2. Generación de la imagen

Para sintetizar una emoción, recibimos como entrada una imagen neutral, que en este contexto llamamos *im*, la emoción *exp* buscada, y un valor δ entre 0 y 1 con el porcentaje de intensidad que se buscan para la expresión pedida. Dada una imagen de entrada *im*, se realizará una combinación de sus datos con los seleccionados del mapa de expresiones *XM* para obtener la imagen sintetizada final.

Cada neurona del *XM* está compuesta por:

1. Un vector de *features* $b = [b^s, b^e, b^u, b^v]$, donde cada componente de b representa información de forma, estructura, geometría y textura, respectivamente;
2. Una etiqueta de emoción asociada.

Esta estructura de cada neurona provee información que preserva la identidad del sujeto y que caracteriza una emoción facial específica. El componente estructural es el

que nos brinda más información sobre la identidad del sujeto, independientemente de la emoción que presente. En consecuencia, para generar la imagen requerida debemos seleccionar, entre las neuronas de XM etiquetadas con la emoción exp, aquella cuya componente estructural sea la más cercana a la estructura de im.

Sea S_{exp} el conjunto de vectores de *features* correspondientes a neuronas del XM etiquetadas con la emoción exp. Buscamos entonces el vector $b \in S_{exp}$ tal que su componente estructural b^e minimice la distancia al componente de estructura im^e de la imagen.

$$b_{exp} = \arg\min_{b \in S_{exp}} dist(im^e, b^e) \tag{1}$$

con $dist$ la distancia euclídea.

El vector b_{exp} representa una imagen que expresa la emoción exp con intensidad 1, según el mapa de expresiones entrenado. Nuestro objetivo es sintetizar la emoción con intensidad δ entre 0 y 1. Siendo $S_{neutral}$ el conjunto conformado por los vectores de las neuronas etiquetadas como *neutral*, de manera análoga a la Ecuación (1) se obtiene b_{neutral}.

Los vectores $b_{neutral}$ y b_{exp} fueron seleccionados en función de la similitud de su componente estructural con el de la imagen original, ya que dicho componente es el que nos proporciona información principalmente de la identidad del sujeto. Bajo la hipótesis de que la estructura facial no se modifica significativamente con la expresión emocional, este componente se mantiene fijo en el proceso de síntesis. En consecuencia, la síntesis se realiza modificando la forma, la geometría y la textura de la imagen de entrada.

Modelamos la intensidad de la emoción como la transición de los *features* de un estado neutral a uno emotivo. Para obtener una intensidad δ utilizamos combinaciones lineales según se describe en la Ecuación (2) dando lugar a los atributos $\tilde{b}^s, \tilde{b}^u, \tilde{b}^v$. Esto lo realizamos a partir de b^s_{neutral}, b^u_{neutral} y b^v_{neutral}, que se corresponden a las componentes de forma, geometría y textura de b_{neutral}, y de b^s_{exp}, b^u_{exp} , b^v_{exp} definidos análogamente para b_{exp}.

$$\begin{aligned} \tilde{b}^s &= \delta b^s_{exp} + (1-\delta) b^s_{neutral} \\ \tilde{b}^u &= \delta b^u_{exp} + (1-\delta) b^u_{neutral} \\ \tilde{b}^v &= \delta b^v_{exp} + (1-\delta) b^v_{neutral} \end{aligned} \tag{2}$$

Luego de obtenidos estos 3 atributos, usamos la proyección inversa de PCA para conseguir los parámetros en el espacio original que vamos a necesitar para la síntesis:

$$\begin{aligned} \tilde{d^s} &= \sigma^s \odot (P^s \tilde{b}^s) + \mu^s \\ \tilde{d^u} &= \sigma^u \odot (P^u \tilde{b}^u) + \mu^u \\ \tilde{v} &= \zeta(\sigma^v \odot (P^v \tilde{b}^v) + \mu^v) \end{aligned} \tag{3}$$

donde P^s, P^u y P^v son las matrices de transformación de PCA para cada componente, μ la media y σ el desvío estándar correspondiente, con $\odot$ el producto punto a punto. (En la Sección 3.1 mencionamos que redujimos la dimensión de los *features* con PCA y aquí mostramos cómo aplicamos P^s, P^u y P^v en las transformaciones inversas para recuperar una versión aproximada de los mismos.) Usamos el valor de escala ζ, estimado

empíricamente, con el objetivo que en la síntesis se genere menos ruido y un mejor resultado final. El valor de ζ usado se fija globalmente para todo el modelo. En la Figura 2 se muestra este procedimiento de re-escalado en un ejemplo.

El componente de forma $\tilde{d^s}$ representa la diferencia de landmarks respecto a los del frame inicial de su video y la componente geométrica $\tilde{d^u}$ representan las diferencias de la parte geométrica de la descomposición TV-L1 respecto al del frame inicial de su video. Por su parte im^s y im^u representan los landmarks y la parte geométrica de la descomposición TV-L1 (que descompone a la imagen en partes de geometría y de textura) de la imagen de entrada respectivamente. Luego, $\tilde{d^s}$ y $\tilde{d^u}$ son sumados a im^s y de im^u de la imagen de entrada de la siguiente manera:

$$\begin{aligned} \tilde{s} &= im^s + \tilde{d^s} \\ \tilde{u} &= im^u + \tilde{d^u} \end{aligned} \tag{4}$$

Sin embargo, esta síntesis aún no incorpora la información de textura de la imagen de entrada im^v. Queremos además integrar de manera más precisa el componente geométrico con el de la imagen original. Para ello, refinamos los componentes geométrico y de textura de modo de incorporar la emoción sin comprometer la identidad del sujeto.

Para determinar qué regiones deben recibir mayor influencia de la síntesis emocional, utilizamos el desplazamiento de los landmarks como indicador del movimiento facial inducido por la expresión. La hipótesis es que los píxeles cercanos a landmarks con mayor desplazamiento deberían reflejar con mayor intensidad los cambios geométricos y de textura asociados a la emoción.

Formalizamos esta idea mediante una función definida como la suma de gaussianas centradas en cada landmark. Cada gaussiana está ponderada por la magnitud del desplazamiento indicado en $\tilde{d^s}$, de modo que landmarks con mayor movimiento generan una contribución más significativa en su vecindad espacial. Para cada píxel en posición i, definimos:

$$h(i) = \sum_{j=1}^{l} \|\tilde{d_j^s}\| \exp(-dist(j,i)/2\beta^2) \tag{5}$$

donde $dist(j,i)$ es distancia euclídea entre el pixel i y el j, con $1 \leq j \leq l$ y l la cantidad total de landmarks. El parámetro β controla la extensión espacial de la influencia de cada landmark. Valores grandes producen una propagación más amplia y suave, mientras que valores pequeños restringen la influencia a regiones más localizadas. En nuestra experiencia hemos encontrado que para la mayoría de las emociones un valor de $\beta = 100$ funciona adecuadamente. En la emoción Surprise, donde la magnitud de los movimientos es considerablemente mayor, se necesita un ajuste más suave y se utiliza en cambio $\beta = 10$.

El componente geométrico final se define como una combinación ponderada entre el componente original y el sintetizado, donde el peso depende del grado de movimiento estimado en cada píxel. De este modo, si el desplazamiento es alto, predomina $\tilde{u}$; en caso contrario, se preserva principalmente im^u. El componente de textura final se obtiene de forma análoga. Sin embargo, el valor $h(i)$ puede presentar variaciones abruptas y no está acotado en el intervalo $[0,1]$. Para garantizar una transición suave y controlada entre

ambas contribuciones, utilizamos una función sigmoideal Ω, que permite mapear los valores de $h(i)$ a un rango acotado y producir una transición suave entre ambas fuentes de información. De esta forma, el componente geométrico $\tilde{u}_{final}$ y el componente de textura $\tilde{v}_{final}$ quedan definidos para cada pixel i como:

$$\begin{aligned} \tilde{u}_{final}(i) &= \tilde{u}(i)\,\Omega(h(i)) + im^{u}(i)(1-\Omega(h(i))) \\ \tilde{v}_{final}(i) &= \tilde{v}(i)\,\Omega(h(i)) + im^{v}(i)(1-\Omega(h(i))) \end{aligned} \tag{6}$$

donde la función sigmoideal se define como:

$$\Omega(h(i)) = \frac{1}{1+\exp(-\phi * (h(i)+\theta))} \tag{7}$$

donde la constante ϕ controla la pendiente y la constante θ la traslación.

Finalmente, combinamos $\tilde{u}_{final}$ y $\tilde{v}_{final}$ para conseguir la imagen objetivo im_{final}:

$$im_{final} = \tilde{u}_{final} + \tilde{v}_{final} \tag{8}$$

im_{final} expresa la emoción exp en la intensidad δ pedida. Las funciones antes definidas nos permiten integrar suavemente las contribuciones de geometría y textura definidas y obtener la imagen final, que es la síntesis de la emoción. En la Figura 3 mostramos ejemplos resultantes de utilizar este método y en la Figura 4 mostramos un diagrama que resume los pasos de la metodología explicada.

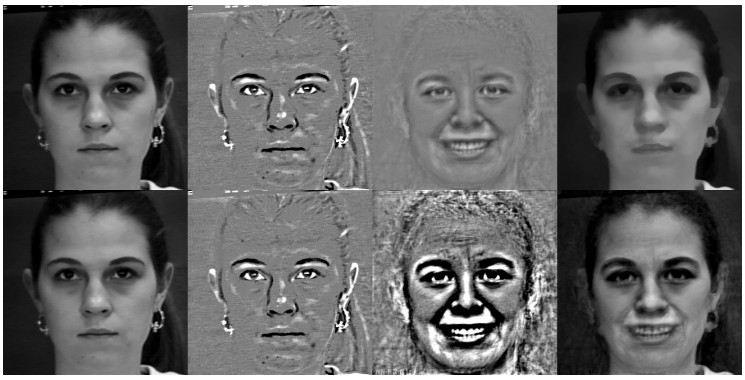


Figura 2. Ejemplo de síntesis de expresión Happy con $\zeta = 1$ (primera fila) y $\zeta = 7$ (segunda fila). De izquierda a derecha: imagen neutral de entrada, textura de la imagen de entrada, textura asociada del XM con parámetro de escalado ζ correspondiente, e imagen sintetizada con la emoción Happy.

4. Experimentación

En esta sección evaluamos nuestro modelo. Se analiza el desempeño del mismo al realizar síntesis de una emoción con distintos grados de intensidad. Además medimos

(a) Anger, Disgust & Fear (b) Happiness, Sadness, Surprise

Figura 3. Ejemplo de síntesis con cada una de las emociones básicas. En la subfigura a) se sintetiza en un sujeto las emociones Anger, Disgust y Fear. En la subfigura b) se sintetiza en otro sujeto las emociones Happiness, Sadness y Surprise. Cada columna muestra (de izquierda a derecha) la imagen neutral, la textura del XP usada para la síntesis, y por último la imagen resultado que expresa la emoción.

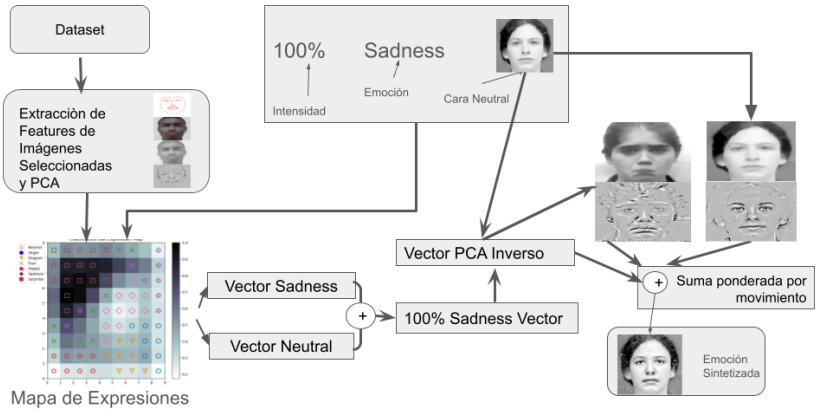


Figura 4. Esquema de Síntesis de Expresiones. El dataset se conforma por videos de sujetos expresando emociones. De cada video, se selecciona el primer y último frame y se obtiene su vector de features y se utiliza PCA para reducir su dimensión. Con los vectores obtenidos se entrena el Mapa de Expresiones. Cuando se quiere sintetizar una imagen, los datos de entrada son: la imagen neutral del sujeto, la etiqueta de emoción y la intensidad de la misma. A partir de estos datos se selecciona del Mapa de Expresiones un vector de *features* de la emoción pedida y de la emoción neutral, cuyos datos se combinan según la intensidad pedida. Se realiza la transformación inversa de PCA del vector resultante. Finalmente, los datos de geometría y textura se combinan con los de la imagen neutral para generar la imagen resultado con la emoción sintetizada.

cuantitativamente su rendimiento al sintetizar emociones utilizando un clasificador externo.

4.1. Síntesis de emoción con distintos grados de intensidad

Hasta este momento la síntesis emocional se ha realizado utilizando una intensidad $\delta = 1$. En este experimento analizamos cómo se comporta nuestro algoritmo cuando se varía este parámetro, evaluando su capacidad para generar emociones que se expresen mas débilmente. Para ello realizamos la progresión en la síntesis de una imagen con emoción Happy y diferentes valores de intensidad δ, con $\delta \in \{0, 0.1, 0.2, \ldots, 1\}$, como se muestra en la Figura 5.

Observando los resultados, se aprecia que incluso al sintetizar una expresión neutral ($\delta = 0$) la imagen generada no coincide exactamente con la original, aunque mantiene la neutralidad en la expresión y la identidad de la persona. A medida que δ aumenta, la emoción comienza a manifestarse de forma progresiva en las imágenes No obstante, para valores pequeños de δ, se dificulta evaluar si la intensidad emocional sintetizada corresponde realmente al valor especificado. Empíricamente, en nuestro modelo el cambio emocional comienza a ser perceptible a partir de intensidades comprendidas entre 0.5 y 0.6. A partir de ese umbral, la emoción se manifiesta de manera más evidente.

Por otro lado, en la Figura 5, la imagen generada con $\delta = 0.5$, presenta en la región de los labios simultáneamente rasgos de la posición neutral y de la sonrisa, produciendo una transición visual poco natural. Este comportamiento puede explicarse a partir del mecanismo interno del modelo. La síntesis se basa en una combinación ponderada entre la imagen neutral y una imagen emocional parcial construida a partir de los datos del XM. La proporción de cada componente se determina mediante el desplazamiento de los landmarks faciales: a mayor desplazamiento de estos, mayor aporte de la imagen emocional. Este procedimiento busca asegurar la expresión emocional y evitar la superposición de texturas en las regiones donde se produce movimiento respecto de la imagen original.

El vector empleado para generar la imagen emocional parcial se obtiene como combinación de los vectores neutral y emocional del XM, donde la intensidad δ controla la magnitud relativa de cada contribución. Para valores bajos de δ, el vector resultante conserva una mayor componente neutral, siendo entonces la imagen parcial más neutral y los movimientos de los landmarks más chicos. En estas condiciones, la fusión entre la imagen neutral y la modificada puede no compensar completamente las proporciones para que no se superpongan las texturas, lo que da lugar a artefactos visuales.

4.2. Clasificación de imágenes sintetizadas usando el clasificador de openCV

Queremos evaluar cuantitativamente la calidad de las imágenes sintetizadas por nuestro modelo. Introducimos un clasificador automático de emociones externo, que determine en qué medida las expresiones generadas por nuestro modelo son reconocidas correctamente por un sistema independiente.

Utilizamos el clasificador de emociones previsto por la librería openCV2 (denominado de aquí en más como “FER”). Este modelo predice la emoción predominante en una imagen facial, y este resultado nos permite comparar la emoción solicitada de la

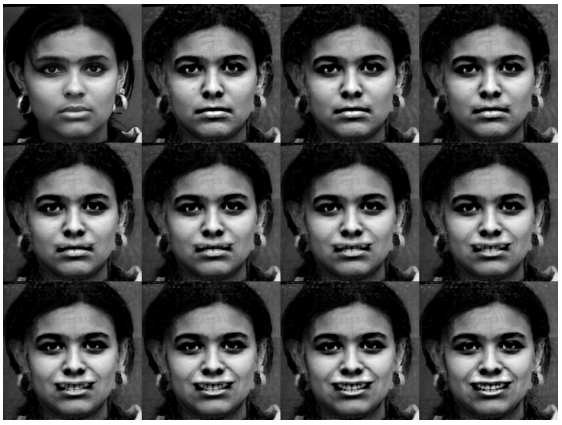


Figura 5. Síntesis de la emoción Happy con distintas intensidades de emoción. Se muestra la imagen de entrada, y el resto de las imágenes son producto de la síntesis con intensidad δ. El valor de δ aumenta 0.1 en cada imagen, ordenadas de izquierda a derecha, y de arriba a abajo. En la esquina superior izquierda se encuentra la imagen de entrada, a su derecha la imagen generada con $\delta = 0$, y sigue el patrón descripto hasta que termina en la imagen de la esquina inferior derecha, con $\delta = 1$.

síntesis con la que efectivamente percibe un clasificador externo. FER está entrenado con su propio dataset: FER2013 (Zahara et al., 2020). El dataset FER2013 contiene aproximadamente 30.000 imágenes, tomadas de internet, centradas y mirando hacia el frente, con tamaño normalizado a 48×48. El clasificador automáticamente recorta el fondo, convierte a escala de grises y reescala a 48×48 las imágenes de entrada que recibe para clasificar.

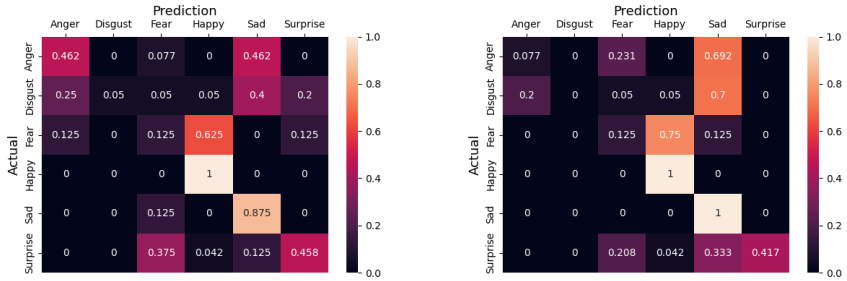
Emociones	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Precision	0.28	0.05	0.125	1	0.43	0.45
Recall	0.5	1	0.077	0.75	0.29	0.34
F1-Score	0.36	0.095	0.095	0.85	0.35	0.39

Cuadro 1. Métricas para imágenes de Testeo.

Emociones	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Precision	0.076	0	0.125	1	1	0.41
Recall	0.2	0	0.1	0.72	0.2	1
F1-Score	0.11	0	0.11	0.84	0.33	0.58

Cuadro 2. Métricas para Imágenes Sintetizadas.

Para ver cómo se comporta este clasificador con imágenes del dataset CK+, utilizamos primero el clasificador FER para clasificar las imágenes (en su apex de expresión) de los videos de testeo. Al conocer la emoción del video original, la predicción sobre



(a) Resultados sobre Imágenes de Testeo (b) Resultados sobre Imágenes Sintetizadas

Figura 6. Matrices de confusión obtenidas al utilizar el clasificador FER con imágenes del dataset CK+. En a) se muestran los resultados de clasificar las imágenes apex de los videos de testeo. En b) se muestran el resultado de clasificar las imágenes sintetizadas con nuestro modelo. Los datos de entrada para la síntesis son las imágenes neutrales de los videos de testeo, intensidad con valor $\delta = 1$ y la emoción etiquetada del video.

estas imágenes nos da una medida clara del desempeño del clasificador con este dataset. Utilizamos la versión de FER que usa la red MTCNN para detectar caras. En la matriz izquierda de la Figura 6 se muestra la matriz de confusión. En el Cuadro 1 se muestran las métricas de Precisión, Recall y F1-Score de cada emoción. El porcentaje de acierto general es 64.89 %.

Estos resultados nos dan un marco de referencia necesario para poder evaluar nuestras imágenes sintetizadas con el clasificador FER. Esperamos que el clasificador presente un comportamiento similar con nuestras imágenes generadas como una manera cuantitativa de evaluar nuestro proceso de síntesis.

Para comparar de la manera más igualitaria posible, las imágenes que clasificamos son sintetizadas usando como datos de entrada del algoritmo las imágenes neutrales de las series de testeo, la emoción con la que están etiquetadas en cada caso, y valor de intensidad 1. Analizamos los resultados de la clasificación sin considerar la etiqueta 'Neutral' como respuesta de FER, ya que al estar incluida la mayor parte de las imágenes eran clasificadas como neutrales. De esta manera el porcentaje de acierto general es de 43.61 %, lo que se observa en la matriz de confusión de la derecha en la Figura 6. En el Cuadro 2 se muestran las métricas de Precisión, Recall y F1-Score de cada emoción de estas imágenes sintetizadas.

Analizando los resultados, en la matriz izquierda observamos que para las emociones de Happy y Sad, la tasa de aciertos es superior a 80 % en ambos casos. Las emociones Surprise y Anger tienen una tasa de acierto menor, confundiendose Anger con Sad, y Surprise con Fear. Las emociones Disgust y Fear, tienen una tasa de aciertos menor al 15 %; se observa que en mayor parte son mal clasificadas como Sad y Happy respectivamente.

Al comparar con la matriz derecha, el comportamiento del clasificador FER es similar, es decir, clasifica correctamente la mayor parte de imágenes con emociones de Happiness y Sadness, tiene un desempeño intermedio con Surprise y Anger y las emociones de Disgust o Fear son con las que tiene más dificultades al clasificar. El mayor cambio observado en esta matriz es que mayor proporción de imágenes de Anger y Disgust son clasificadas falsamente como Sadness. Aumenta además la cantidad de imágenes Surprise que se confunden con Sadness y Fear con Happy.

5. Conclusiones

En este trabajo logramos realizar síntesis que muestran expresiones faciales a partir de una imagen neutral, con un costo operacional bajo. Exploramos las dificultades encontradas al utilizar el método, tanto en la necesidad de ajustar los valores como en el desempeño del mismo con una menor intensidad de emoción. Utilizando el clasificador FER, comparamos las imágenes de una persona expresando emoción real con la generada a partir de su imagen neutral. Este experimento presenta un resultado que muestra un desempeño similar en las emociones, y un mayor cantidad de falsos positivos en Happy y Sad en las sintetizadas, mostrando un posible bias del algoritmo al sintetizar esas emociones o del clasificador. La identificación del bias y sus causas, así como el ajuste requerido para minimizarlo sería un buen aporte para una posible mejora a futuro.

Referencias

Agarwal, S., & Mukherjee, D. P. (2019). Synthesis of Realistic Facial Expressions Using Expression Map. *IEEE Transactions on Multimedia*, *21*(4), 902-914.

Alvarez Mon, A., Buemi, M. E., & Acevedo, D. (2024). Mapas de expresiones para la clasificación de expresiones faciales basadas en features. *JAIIO, Jornadas Argentinas de Informática*, *10*(11), 80-88.

Dong, X., Ning, X., Xu, J., Yu, L., Li, W., & Zhang, L. (2024). A Recognizable Expression Line Portrait Synthesis Method in Portrait Rendering Robot. *IEEE Transactions on Computational Social Systems*, *11*(1), 1440-1450. https://doi.org/10.1109/TCSS.2023.3241003

Huang, D., & De la Torre, F. (2010). Bilinear kernel reduced rank regression for facial expression synthesis. *Proceedings of the 11th European Conference on Computer Vision: Part II*, 364-377.

Li, K., Xu, F., Wang, J., Dai, Q., & Liu, Y. (2012). A data-driven approach for facial expression synthesis in video. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 57-64. https://doi.org/10.1109/CVPR.2012.6247658

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 94-101.

Luiz Testa, R., Machado-Lima, A., & Nunes, F. L. S. (2025). Enhancing Temporal Coherence in Image-to-Video Facial Expression Synthesis: A Dual-Loss Framework for Smoother Generation. *IEEE Access*, *13*, 170876-170894. https://doi.org/10.1109/ACCESS.2025.3612820

Xiong, L., Zheng, N., You, Q., & Liu, J. (2007). Facial Expression Sequence Synthesis Based on Shape and Texture Fusion Model. *2007 IEEE International Conference on Image Processing*, *4*, IV - 473-IV -476. https://doi.org/10.1109/ICIP.2007.4380057

Yan, Y., Huang, Y., Chen, S., Shen, C., & Wang, H. (2020). Joint Deep Learning of Facial Expression Synthesis and Recognition. *IEEE Transactions on Multimedia*, *22*(11), 2792-2807. https://doi.org/10.1109/TMM.2019.2962317

Zahara, L., Musa, P., Prasetyo Wibowo, E., Karim, I., & Bahri Musa, S. (2020). The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi. *2020 Fifth International Conference on Informatics and Computing (ICIC)*, 1-9. https://doi.org/10.1109/ICIC50835.2020.9288560

Zhou, C., & Lin, X. (2005). Facial expressional image synthesis controlled by emotional parameters. *Pattern Recognition Letters*, *26*(16), 2611-2627. https://doi.org/https://doi.org/10.1016/j.patrec.2005.06.007