Base de Imágenes Latinoamericana para Reconocimiento Facial

Camila Di Ielsi[1][0000-0002-3734-5550] and Pablo Negri[1,2][0000-0003-0250-5208]

[1] Departamento de Computación, FCyN (Universidad de Buenos Aires), Argentina
[2] Instituto de Investigación en Ciencias de la Computación (UBA-CONICET)
cdiielsi [at] dc.uba.ar, pnegri [at] dc.uba.ar

Abstract. Considerando la relevancia e impacto de los crecientes data-sets en el área de reconocimiento facial, introducimos una base de datos conformada exclusivamente por sujetos nacidos en América Latina con el objetivo de que sea útil para el entrenamiento y evaluación de modelos de reconocimiento facial dedicados a utilizarse en proyectos dentro de esta región.

Keywords: Reconocimiento Facial · Data Set · Latinoamérica.

1 Introducción

El éxito de los sistemas de Deep Learning esta fundado en los siguientes pilares: modernas arquitecturas de redes neuronales y estrategias de entrenamiento, dispositivos de hardware para procesamiento en paralelo, y extensas bases de datos. En el caso de Reconocimiento Facial, la disponibilidad de bases de imágenes permitió el rápido desarrollo de sistemas robustos. Sin embargo, estas bases suelen tener un sesgo muy importante relativo a la fuente de donde son extraídas: generalmente provienen de sitios disponibles en internet relacionados con celebridades del mundo del espectáculo. Las celebridades que poseen más fotografías en estas plataformas son generalmente las preferidas en la confección de las bases. Y en su mayoría, estos casos se dan con celebridades que trabajan en Estados Unidos o Europa.

El entrenamiento y evaluación de modelos de reconocimiento facial para ser utilizados en latinoamérica crea la necesidad de generar una base de datos de personas de origen latino. En este trabajo se presenta una base de imágenes confeccionada con este objetivo, conteniendo fotografías de celebridades tomadas del sitio web denominado Internet Movie Database (IMDb) [1]. IMDb es una base de datos en línea pública que almacena información relacionada con la industria del cine y la televisión. Particularmente cuenta con perfiles de actores, actrices y otros trabajadores en la industria con sus respectivas galerías de imágenes, además de información personal. A partir de este sitio web, constituimos el data-set mediante la búsqueda de sujetos por país de nacimiento. Sabemos que esta base de imágenes no representaría en toda su variedad a la demográfica latinoamericana, pero consideramos que no deja de ser útil contar con un data-set

que no incluya rostros de etnias europeas y estadounidense, las cuales siempre son predominantes en todas las bases públicas disponibles, para así poder empezar a plantearnos implementar y desarrollar modelos de reconocimiento facial dedicados exclusivamente a nuestra región.

Este paper se organiza de la siguiente forma: En la sección 2 se mencionan otras bases de datos conformadas con imágenes de celebridades y se introduce nuestra base de datos. En la sección 3 se detalla la metodología para la descarga de las galerías de fotos de las celebridades y las estrategias de limpieza de los parches correspondientes a los rostros. En la sección 4 se describe el data-set a partir de su estudio estadístico según la distribución de género, edad y por país. Por último, en la sección 5 concluimos enunciando posibles uso para esta base.

2 Trabajos Relacionados

IMDb ha sido una fuente importante para la confección de bases de imágenes utilizadas para entrenamiento y evaluación de modelos de reconocimiento facial. Además de las fotografías, IMDb aporta importantes metadatos de cada persona, como la fecha y lugar de nacimiento, fecha de captura de la fotografía, etc., aunque estos datos no son completos para todas las identidades. La base CACD2000 [3] consiste en unas 163 mil fotografías de aproximadamente 2000 celebridades con una estimación de la edad que va de los 16 a los 62 años. En forma similar, en [8] se comparte una base que contiene más de 500 mil fotografías de IMDb y Wikipedia, correspondiente a unas 20 mil identidades con información de edad y género. Un dataset ampliamente utilizado para entrenar modelos es el MS1M [5] que contiene unas 90K identidades y más de 4M de fotografías. IMDb-Face [10] es también una base de imágenes que se ocupó de tratar de no utilizar fotografías de MS1M para poder luego realizar validaciones. Esta base contiene 59K identidades y más de 1.7M de imágenes. Recientemente, se publicó un dataset denominado WebFace260M [12] que ha descargado utilizando diversos buscadores, unas 260M de fotografías. Una primera versión limpia de WebFace260M contendría unas 40M de fotos pertenecientes a 2M de identidades.

Este trabajo presenta una base de imágenes de los rostros de celebridades registradas en IMDb, con la particularidad de que se utilizaron filtros por lugar de nacimiento, guardando además esa información como metadatos [2]. Las celebridades de nuestro dataset pertenecen a una lista de 20 países latinoamericanos. De cada persona se registra la fecha de nacimiento y, de estar disponible, la fecha en que fue sacada la fotografía, para poder hacer un cálculo de su edad.

3 Construcción de la base de datos

El dataset es recopilado a partir de lanzar en IMDb una búsqueda de celebridades por lugar de nacimiento. Utilizando un parser html, se accede a los metadatos de los sujetos listados, recuperando su fecha de nacimiento y accediendo a su galería de imágenes (si no está vacía). Cuando las fotografías contienen información en el título u otro texto asociado, es posible calcular la edad de la celebridad en el

Fig. 1. Pasos para la descarga e identificación de las fotografías de cada celebridad.

momento de la captura. El enlace de la imagen es guardado para su posterior descarga y procesamiento. Cada celebridad es identificada con un código único que posee IMDb, y que lo identifica en esta base de datos.

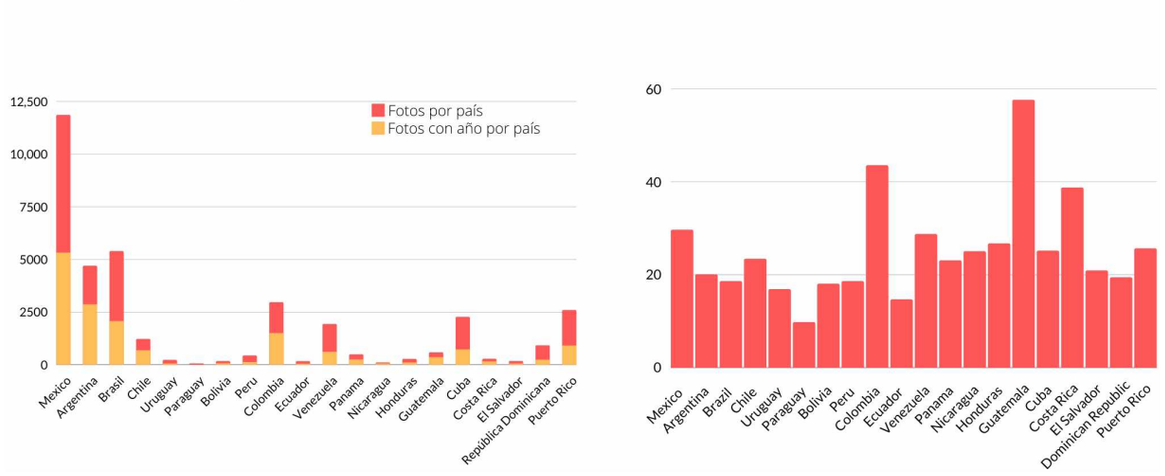
Concretamente el protocolo de búsqueda, descarga y procesamiento se divide en 4 etapas:

1. **Búsqueda:** Ids de celebridades nacidas en un país latinoamericano.
2. **Descarga de la galería:** Se descargan las imágenes por cada celebridad de la página de IMDb.
3. **Detección y Alineación de Rostros:** Detección de todos los distintos rostros en cada foto utilizando MTCNN [11], una CNN multi tarea en cascada. Los rostros que se encuentren de perfil son descartados. Una vez ubicados los rostros los recortamos a un tamaño de 128×128 píxeles y alineamos.
4. **Clustering:** Se realiza un curado automático de los rostros para eliminar los intrusos. Para ello se aplica DBSCAN (Density-based spatial clustering of applications with noise) con $\varepsilon = 1.0$, 5 cómo el número mínimo de fotos requeridas para que se considere un cluster y según la métrica Euclidea, que realiza el agrupamiento de los rostros a partir de los embeddings generados aplicando Arcface [4], una función de pérdida que mejora el poder discriminativo de los embeddings aprendidos.

4 Estadísticas del data-set

El dataset contiene en total 20 países de latinoamérica, cuyo detalle puede observarse en la fig. 2. El total de identidades diferentes es de 1.469 y de 36.658 fotografías de rostros, dando un promedio general de aproximadamente 25 fotos por persona. Mas detalles sobre la constitución de la base pueden encontrarse en el repositorio del proyecto [2].

La fig. 2 (a) presenta la cantidad de fotografías con información de edad, con respecto a la cantidad de fotos disponibles por identidad y por país. Este gráfico de barras también permite apreciar que tanto Argentina, México y Brasil poseen la mayor cantidad de identidades en el dataset.



(a) (b)

(c) (d)

Fig. 2. (a) Distribución de fotos con año sobre el total para cada país. (b) Promedio de fotos por persona por país. (c) Se muestra la proporción de celebridades por género. (d) Es un histograma de la cantidad de fotografías por edad.

Se utiliza un predictor de género para analizar el balance de sujetos femeninos y masculinos. El clasificador posee un backbone con arquitectura EfficientNet-b0 [9] entrenado con ID-DATASET [7], una base que cuenta con 2.527.079 imágenes de 793.280 individuos, distrubuída de forma casi equitativa en cantidad de hombres y mujeres. En la fig. 2 (c) se muestra esta distribución, donde puede apreciarse que hay una mayoría de sujetos masculinos (56.4%). Esto representa un importante desbalance que muestra la necesidad de seguir trabajando en el dataset para poder equiparar el numero de identidades por género. La fig. 2 (d) muestra el histograma de la cantidad de rostros por edades del total de fotos con edad registrada, mostrando una distribución muy concentrada entre los 20 y 40 años.

5 Conclusión

Este trabajo introduce una base de imágenes de rostros de celebridades de origen latinoamericano, obtenida a partir de IMDb. Esta base puede utilizarse para entrenamiento de modelos de reconocimiento facial, ya sea desde scratch o haciendo finetunning de modelos preentrenados. Además, puede emplearse para hacer evaluación de modelos de reconocimiento facial, armando test de pares como LFW [6], y estudiar la existencia de sesgos con personas latinoamericanas. En este sentido nos parece relevante distinguir el país de origen de cada individuo, en caso de que se quieran hacer estudios específicos. Esta es una característica en la que nuestra base se distingue de las demás mencionadas, el país de origen es parte de la metadata. Por otro lado, si bien el porcentaje de fotos con referencia

de edad es bajo considerando el total de fotos, nos parecía pertinente incluir la metadata disponible, sobretodo considerando su posible uso para finetunning de un modelo que ya trabaja con esa metadata, replicando las características de las bases mencionadas el la sección 2 (distinción por género y por rango de edad). Esta base puede considerarse aún en desarrollo, ya que se va a buscar equilibrar y balancear la misma con respecto a género y a grupos étnicos locales.

References

1. IMDb Homepage, http://www.imdb.com.
2. Dataset de Celebridades Latinoamericanas, https://git.exactas.uba.ar/pnegri/imdb-latam-facedataset.
3. Chen, B. C., Chen, C. S., and Hsu, W. H. Cross-age reference coding for age-invariant face recognition and retrieval. In European Conference on Computer Vision, pp. 768-783. Springer, Cham, (2014).
4. Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 4690-4699, (2019)..
5. Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In European conference on computer vision, pp. 87-102. Springer, Cham, (2016).
6. Huang, G. B., Ramesh, M. Berg, T., and Learned-Miller, E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, (2007).
7. Negri, P., Cumani, S., and Bottino, A. Tackling Age-Invariant Face Recognition With Non-Linear PLDA and Pairwise SVM. IEEE Access, 9, 40649-40664, (2021).
8. Rothe, R., Timofte, R., and Van Gool, L.. Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision, 126(2), p. 144-157, (2018).
9. Tan, M., and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning, p. 6105-6114, (2019).
10. Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C., and Loy, C. C. The devil of face recognition is in the noise. In Proceedings of the European Conference on Computer Vision (ECCV), p. 765-780, (2018)
11. Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10), p. 1499-1503, (2016).
12. Zhu, Z., Huang, G., Deng, J., Ye, Y., Huang, J., Chen, X., and Zhou, J. WebFace260M: A Benchmark Unveiling the Power of Million-Scale Deep Face Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10492-10502, (2021).