.UBAfadu
FACULTAD DE ARQUITECTURA
DISEÑO Y URBANISMO

Universidad de Buenos Aires
Facultad de Arquitectura, Diseño y Urbanismo
F.A.D.U.

Posgrado
Maestría en
Diseño Interactivo

Defensa de Tesis

TEMA: La región sin nosotros. Estructuras de sentido automáticas reconocidas en el consumo audiovisual contemporáneo interpretadas como vectores de generación de nuevas formas de montaje y narración con imágenes y sonidos.

Alumno/a: .Mariano Ramis
Director de tesis: Anabella Speziale

Buenos Aires, 11 de Agosto de 2024

La región sin nosotros

Edición de video mediante redes neuronales artificiales

Mariano Ramis

Tésis MAEDI - Tutora Anabella Speziale

Agradecimientos

Deseo agradecer a mi directora de tesis, la Dra. Anabella Speziale por sus consejos, enorme dedicación y acompañamiento en el desarrollo y escritura de este texto.

Al Dr. Martín Groisman y al D.G. Alejandro Papa como gestores de la maestría MAEDI, por representar una oferta de posgrado indispensable para la actualización y sofisticación del pensamiento de docentes y ex estudiantes de diseño, particularmente para quienes provenimos de FADU, UBA.

Al Dr. Ricardo Dal Farra por siempre abrir puertas y por su inmejorable tarea como tutor de la beca ELAP, a la Concordia University de Montreal por recibirme, y el Canadian Bureau of International Education, por haberme asignado la beca.

A mi familia, amigos y colegas.

Índice

Abstract

La presente investigación estudia el desarrollo y las posibilidades del montaje de video mediante el uso de Inteligencia Artificial, y el contexto de proliferación de sistemas de moderación y distribución de contenidos en video, fundamentalmente en la plataforma YouTube. El objetivo de esta investigación es reconocer y proyectar algunas de las capacidades expresivas surgidas a partir de la informatización del vasto contenido en video que circula en la web, y al mismo tiempo revelar una porción de la lógica de funcionamiento y la tecnología que da orden al material audiovisual disponible, asociando esta lógica de orden a la del montaje cinematográfico.

La investigación está enmarcada en la comprensión técnica de los algoritmos y redes neuronales artificiales que motorizan la descripción informática de videos, por lo que se hace foco en bibliografía técnica durante la tesis. Por otro lado obras de cine de found footage y cine estructural son también contrastadas con algunas de las posibilidades del montaje de video automatizado, y se recurre a textos de autores como Sergei Eisenstein, cuyos estudios sobre montaje entran en tensión con los mencionados géneros de cine de corte experimental.

Dentro de los recursos metodológicos se generan cruces históricos respecto al orden y yuxtaposición de imágenes, asociados a las genealogías de medios, incluyendo en el estudio dispositivos del siglo XIX, asociándolos con otros desarrollos actuales como el reciente texto a imagen (text2image). Al mismo tiempo se ponen en práctica y estudian aplicaciones de montaje automatizado.

Entre los hallazgos de la investigación se encuentra el hecho de que las herramientas de etiquetado y moderación desarrolladas para organizar y distribuir la descomunal cantidad de videos que circula en la red, (cantidad cuyo volumen corresponde a las magnitudes denominadas Big Data), son adaptadas como parte de la maquinaria que se utiliza para comprender y montar secuencias de sentido en la aplicaciones de montaje de video automatizado mediante I.A. utilizadas por millones de usuarios.

Es sensato concluir que el corte generalista que introducen estos modelos entrenados en la web, utilizados fundamentalmente para sugerir compras y censurar contenido inapropiado, poco aporte pueden hacer si se los pone a funcionar como montajistas, en el sentido y utilidad clásicos del término, es decir el que implica emular la sensibilidad humana.

Se plantea entonces otra forma de montaje para esas mismas herramientas, desarrollando otros tipos de algoritmos, motorizados por estos mismos modelos y redes neuronales, pero realizando búsquedas libres en la marea audiovisual que nutre la red, en una suerte de minería audiovisual. Estos algoritmos podrían montar secuencias de video, antropológicamente valiosas, abriendo al conocimiento general las rutinas y actitudes humanas registradas en video, atravesando todas las culturas y etnias, siendo un espejo inédito de lo que las personas producen en distintos puntos del planeta, y al mismo tiempo poniendo a la informática en una relación de complementación con la de los seres humanos, sin reemplazar a las personas en el dominio y comprensión sensible del mundo.

Abstract

This research studies the development and possibilities of video editing through the use of Artificial Intelligence, in the context of systems for moderating and distributing video content, mainly on the YouTube platform. The goal of this research is to recognize and project some of the expressive capabilities that have arisen from the computerization of the vast amount of video content circulating on the web, and at the same time to reveal a portion of the logic that operates the technology that gives order to the available audiovisual material, associating this logic of order with that of cinematographic editing/montage. This research is framed in the technical understanding of the algorithms and artificial neural networks that drive the computer description of videos, so the focus is on technical bibliography during the thesis.

On tthe other hand, works of found footage cinema and structural cinema are also contrasted with some of the possibilities of automated video editing, and also texts by authors such as Sergei Eisenstein are used, whose studies on editing come into tension with the aforementioned genres of experimental cinema.

Within the methodological resources, historical intersections are generated regarding the order and juxtaposition of images, associated with media genealogies, including devices from the 19th century in the study, comparing them with other current developments such as recent text to image (text2image) technique and automated video editing applications (neural edit)

Among the findings of the research is the fact that the labeling and moderation tools developed to organize and distribute the huge amount of videos circulating on the network (a quantity whose volume corresponds to the magnitudes called Big Data), are adapted as part of the machinery used to understand and assemble sequences in the automated video assembly applications using AI used by millions of users.

It is reasonable to conclude that the generalist cut introduced by these models trained on the web, used primarily to suggest purchases and censor inappropriate content, can make little contribution if they are put to work as editors, in the classic sense and utility of the term, that is, the one that implies emulating human sensitivity, creating sense by juxtaposing images and sounds.

So, another form of editing is then proposed for these same tools, developing other types of algorithms, driven by these same models and neural networks, but carrying out free searches in the audiovisual universe, in a kind of audiovisual mining. These algorithms could

edit video sequences, anthropologically valuable, opening up to general knowledge the human routines and attitudes recorded on video, crossing all cultures and ethnicities, being an unprecedented mirror of what people produce in different parts of the planet, and at the same time putting AI, and computing in a complementary relationship with that of human beings, without replacing people in their sensitive domain, but helping them to better understand the world.

Introducción

La presente investigación estudia las posibilidades y alternativas del montaje de video automatizado, mediante el uso de Inteligencia Artificial, para comprender la magnitud de los cambios que esto puede traer en la disciplina del diseño audiovisual y al mismo tiempo proyectar mediante los conocimientos obtenidos usos y utilidades innovadoras.

La propuesta de esta investigación analítica y descriptiva, es entonces estudiar prácticas y tecnologías cuya confluencia da lugar a este nuevo campo de posibilidades comunicacionales y expresivas, configurado por la informática, fundamentalmente por algunos tipos y funciones específicas de Redes Neuronales, y su vínculo con el montaje y la yuxtaposición de secuencias de imágenes, elemento distintivo del cine y las artes audiovisuales.

Gran parte del insumo de análisis teórico y práctico de esta investigación proviene de repositorios de video masivos que existen en la web, que han sido el escenario de la puesta en práctica de herramientas para manipular contenidos, y sistemas de moderación y distribución, particularmente en el caso de la plataforma YouTube que desde 2017 se volvió mucho más intervencionista con su contenido, en base el extensivo e intensivo desarrollo y aplicación de estas tecnologías (Burgess, J. y Green, J., 2018).

Entre los objetivos se destaca la tarea de evaluar la importancia y posibles proyecciones de este reciente vínculo entre informática, representada por la I.A. y el montaje. Por un lado analizando el alcance en la historia reciente de ambas cuestiones y también estudiando experiencias de piezas de vídeo editadas automáticamente mediante las cuales será posible ilustrar algunas de estas ideas y técnicas de montaje automático surgidas durante mediados de la década de 2010. Otro objetivo es generar concepciones didácticas para transmitir los elementos estudiados en espacios de formación, siendo que al día de culminación de este texto, no abundan estudios que aporten reflexiones teórico estéticas relacionadas al montaje automatizado de video, aún cuando el mismo es sumamente difundido y utilizado en redes sociales y aplicaciones para móviles.

Para introducir el estado actual de este contacto entre disciplinas y saberes, es necesario hacer una breve genealogía de medios, detectando algunos puntos nodales, dentro de las innovaciones en el campo de la informática y otro tanto de reflexiones y experiencias vinculadas al montaje, y dentro de estas dos vertientes discriminar cuales serán

fundamentales, ya que desde luego no todos los avances sobre Inteligencia Artificial, o informática competen al presente trabajo, ni tampoco todo lo que se ha hecho o escrito sobre montaje en audiovisual. Para discriminar esos puntos nodales que hacen al interés del presente texto, es necesario en principio aislar ambas disciplinas por un momento, la de la informática, computación y el estudio de datos, y la del montaje, con sus criterios narrativos y estéticos a lo largo del tiempo.

En el terreno informático se tomarán como puntos a analizar los desarrollos vinculados a la automatización de procesos a través de capacidades de juicio autónomas, estas facultades computacionales de decidir sin la supervisión de un humano, son las que implican a grandes rasgos la inteligencia artificial. Como fue mencionado con anterioridad, no es toda la inteligencia artificial la que importa aquí, sino la que se utiliza para reconocer el contenido de una imagen (*feature recognition)* y un tipo de red neuronal, las *redes neuronales convolucionales*, y las redes neuronales que se utilizan para organizar secuencialmente la sintaxis de texto, *las redes neuronales recurrentes*.

En el desarrollo del texto estos ítems serán visitados con frecuencia, ya que constituyen las herramientas más importantes que se ponen en práctica para que una computadora perciba y desglose lo que ve cuando se le presenta una imagen.

Otros puntos esenciales que serán estudiados como antecedentes dentro del ámbito informático son el Perceptron Mark I de Rosemblatt, primer sistema computacional creado para reconocer el contenido de una fotografía, el deep learning o aprendizaje profundo, como área de evolución de las capacidades cognitivas de la I.A., y la aplicación inmediata de esos avances explicada en una breve línea de tiempo sobre los 18 años historia de YouTube, y como la puesta en práctica de algoritmos, munidos de herramientas desarrolladas via IA, han permitido y permiten moderar cantidades de contenido audiovisual de magnitudes sin duda inéditas.

El estudio de YouTube como plataforma, gestora de métodos automatizados de distribución, moderación y edición audiovisual, será otro de los elementos a estudiar en este texto, con el objetivo de comprender y luego socializar las estructuras que motorizan su funcionamiento y la trascendencia de los mismos a otras aplicaciones.

Dentro de este terreno se estudiarán también los denominados *sistemas de recomendación*, para verificar la similitud de los mismos con el montaje, siendo estos sistemas estructuradores de sentido que yuxtaponen información. Esto se hará mediante en estudio de papers científicos, y cursado de clases sobre motores de recomendación.

Sitios web y aplicaciones de montaje automatizado serán puestas en práctica y analizadas, comparando lo que ofrecen como producto, el resultado que es posible extraer de las mismas, apelando a la comprensión del algoritmo que determina sus funcionamientos.

En el campo del montaje, algunos de los nodos principales se analizarán, serán tomados del estudio genealógico de medios, dispositivos pre cinematográficos como el myriorama o el panorama, y del cine de los primeros tiempos, surgidos de las vanguardias y sus reflexiones sobre el cine, como nuevo medio en desarrollo. Pero fundamentalmente el análisis se centrará en ejemplos de obras del denominado Cine Estructural de la década de 1960 y las experiencias de cine de montaje encontrado (*found footage*), sus autores más destacados, cuyos proyectos y textos de disertación suelen tener una interpretación novedosa del montaje como articulador de sentido.

Es sabido que la capacidad intelectual de crear sentido uniendo una imagen con la siguiente ha sido motivo de ensayos y argumento de análisis de innumerables obras. La posibilidad de que la autoría intelectual de la edición de una secuencia audiovisual esté en manos una red neuronal artificial (una computadora), permite advertir un posible cambio histórico.

Siendo el planteo de este estudio de perfil exploratorio, no se proponen preguntas específicas a responder, sino más bien la examinación de una serie de ítems que mecanizan la comprensión y posterior montaje de video automatizados . Los mismos serán estudiados para extraer reflexiones y para intentar comprender la magnitud de los posibles cambios que traen a la práctica del montaje audiovisual, a partir de reconocer el hecho de que representan un contexto novedoso y de inquietante potencial.

En esta investigación se pretenderá reunir los datos para saber si existen maneras de que esas estrategias y técnicas, puedan estar al servicio también de otros fines y utilidades concretas, fundamentalmente estudiando sus cualidades como recursos creativos y artísticos, manipulando volúmenes de información accesibles, desligando su uso de las necesidades de marketing y los modelos de negocios para los que fueron desarrolladas, proyectando otras

funciones en el campo del arte y el diseño, particularmente examinando cuán capaces son de construir secuencias de sentido yuxtaponiendo videos.

Capítulo 1 Las herramientas artificiales

Para comprender los aportes teórico estéticos relacionados al montaje automatizado de video, se propone un recorrido por el estado del arte en cuanto a tecnologías que hacen posible que en las redes y en aplicaciones móviles exista montaje automatizado de secuencias de video. Ese recorrido a priori incluye, la meteórica historia de YouTube y otros repositorios similares, como escenarios de la automatización en la organización de información audiovisual, y el uso que han hecho estas plataforma de los avances en aprendizaje automatizado (machine learning) y visión por computadora (computer vision) para el desarrollo de su algoritmo propietario (Burgess y Green, 2018). También algunas observaciones sobre las redes neuronales profundas y su origen, principalmente en la detección y discriminación de elementos en videos, y el campo de estudio general las decisiones tomadas a partir del estudio de datos, parte fundamental de los sistemas que utilizan inteligencia artificial.

Reconocimiento de elementos en imágenes, la visión por computadora, orígenes y aplicaciones

La visión por computadora o visión artificial es una disciplina que desarrolla métodos para procesar, analizar, discriminar y finalmente comprender el contenido de información capturada desde el mundo ¨real¨ con el fin de generar información numérica o simbólica que pueda ser comprendida por una computadora. Así como los humanos utilizan sus sentidos para interactuar con el mundo que les rodea, la visión por computadora trata de emular estos procesos para que una computadora pueda percibir y comprender el contenido de una fotografía, o las acciones desarrolladas en una secuencia de fotogramas o de unas señal de video en tiempo real.

Aunque el posible origen de la denominada visión por computadora cuenta con al menos 70 años, mediado de la década del 1950, es un hecho que su utilización a gran escala es reciente, desde finales de la década de 1990 con la proliferación de internet, y gracias a las capacidades de cómputo disponibles en la actualidad, y a un ecosistema de utilización masivo que demanda soluciones de moderación y discriminación automática y sistemas autónomos de todo tipo. “La computación representa la convergencia entre tecnologías de cálculo, almacenamiento de información, y automatización” (Ceruzzi, P, 2012, p11).

En los orígenes, a un sistema de visión artificial se asignaban tareas de reconocimiento de carácter binario, es decir, cuyo resultado tenía sólo dos alternativas, si, o no. La primera computadora desarrollada para *ver* y discriminar elementos en fotografías se llamó Mark I Perceptron, y se puso a funcionamiento a mediados de la década de 1950. El Perceptron, desarrollado por el psicólogo norteamericano Frank Rosenblatt, tenía como objetivo clasificar imágenes de manera automatizada, a partir de su sistema de visión.

El ingreso de información visual en el sistema, se realizaba a través de un sensor que funcionaba como cámara o escáner, capaz de analizar fotografías mediante una matriz de 400 pixeles de resolución (20×20), siempre usando como insumo de análisis fotografias en escala de grises de un mismo tamaño, de aproximadamente 4x4 centímetros, con rostros, similares a la conocidas como foto-carnet. A partir del input de 400 píxeles el sistema era capaz, con relativa eficacia, de reconocer el género de la persona analizando cada fotografía, a partir del desglose en partes (puntos) de cada píxel obtenido en la captura visual, la respuesta final al análisis de cada foto podía positiva o negativa para el género varón, o positiva o negativa para el género mujer. (Kelleher, J. 2019)

Es importante destacar el pequeño tamaño de las fotografías utilizadas (4x4cm), el hecho de que en las fotos siempre había un rostro humano y que las fotos estaban impresas en escala de grises, estas tres particularidades son esenciales para el funcionamiento del sistema.

Para que el sistema aprenda a reconocer ambos géneros, una extensa etapa previa de entrenamiento fue necesaria, donde se le presentaron al Perceptrón miles de casos en fotografías de rostros humanos, y manualmente, con un ser humano supervisando la tarea, en el entrenamiento[1] se fueron etiquetando los resultados de cada observación, asignando el dato correcto binario, Mujer u Hombre como etiquetas a cada ejemplo de fotografía analizada. Este proceso describe lo que actualmente se conoce como Machine Learning (ML) aprendizaje automatizado en castellano, esta etapa de entrenamiento, a grandes rasgos, sigue vigente en la actualidad, claro que con otras magnitudes, infinitamente superiores y diversas.

En el caso del Perceptron la etapa de entrenamiento es central y distintiva en el proceso, como lo es en la actualidad en sistemas mucho más complejos, y se utiliza para alimentar y sofisticar la capacidad de sistemas que utilizan neuronas artificiales, para permitirles luego

[1] Se conoce como entrenamiento en el ámbito de la Inteligencia artificial al proceso de aprendizaje mediante el estudio de ejemplos

evaluar a partir de la información latente, obtenida a partir de la suma de todas las fotos de rostros vistos y etiquetados, el resultado del análisis de una nueva foto, de similares características, pero que el sistema nunca vió antes. Patrones de luz emitidos por las fotografías desde todas sus regiones, terminan por generar información visual latente, particular y distinta en el caso de rostros de hombres, y en el de mujeres, esa información abstracta, obtenida luego del entrenamiento, puede ser manipulada y utilizada como parámetro análisis.

Esta información latente se conoce como modelo[2], y representa la experiencia previa que utilizara el sistema para contrastar cada caso nuevo con el que se encuentre y luego dar un veredicto, con mayor o menor grado de fiabilidad.

Las redes neuronales, capaces de evaluar información, sus estructuras y complejidad, junto con los avances en potencias computacionales, se han hecho más eficientes, capaces de hacer cosas nunca imaginadas y muy probablemente cosas que aún puedan sorprendernos. Dentro de esas capacidades la de distinguir elementos en una fotografía o imagen, se convirtió en un área de aplicación concreta conocida como feature recognition, que implica el reconocimiento de figuras, elementos o cosas que puedan ser discriminadas como algo reconocible o figurado, por ejemplo una plancha, una auto, su modelo marca y año de fabricación, una persona sentada, o una especie particular de árbol.

Es llamativo el dilema que se presenta al intentar reconocer cuántas cosas hay en el universo que podemos identificar, se trata de un problema de una escala mayúscula, porque cada región del planeta, cada cultura y cada persona, por nombrar sólo uno parte mínima de los aspectos del problema, tiene un universo de cosas que puede reconocer visualmente y les son familiares, así como son exóticas, para otras.

La magnitud del problema implica una análoga magnitud para su solución, que en este caso implica la inmensa cantidad de información visual con la que redes neuronales son entrenadas a diario, por las personas que navegan en internet. En muchos casos sin saberlo el tráfico humano de la web colabora para hacer que estas redes neuronales funcionan mejor, participando en su entrenamiento, etiquetando las fotografías que compartimos, o respondiendo a *captchas* visuales para descargar un archivo, así se aumenta la capacidad de

[2] El modelo es la formalización de parámetros, resultado del proceso de entrenamiento, que permitirán al sistema autónomo llevar a cabo sus tareas.

los sistemas para distinguir el aspecto visual de cosas. Gracias a esto, las redes se van volviendo mejores en función de la calidad y cantidad de imágenes que clasifican con supervisión de seres humanos, cosa que sucede de manera casi automática, gracias al caudaloso flujo de fotos y videos acompañadas con un texto que describe, o da contexto su contenido, que las personas suben la red, o comparten a diario.

Es esta capacidad de comprensión visual artificial, que suele ser utilizada por grandes compañías para desarrollar estrategias de marketing que surcan la marea de la Big data, es la que me interesa direccionar hacia un campo expresivo y filosófico en la presente investigación.

YouTube, las redes neuronales y los sistemas de recomendación

El presente texto e investigación se propone reconocer la influencia de muchas de las carácteristicas de YouTube como repositorio de archivo y consumo audiovisual, en la modelación de conductas sociales e intelectuales de quienes a diario interactúan con la plataforma, sin dejar de lado las posibles derivaciones artísticas que hacen uso de las tecnologías desarrolladas y luego puestas en práctica en la compleja tarea de catalogación y distribución de videos disponibles en la plataforma.

En cuanto a las tecnologías desarrolladas para superar inconvenientes surgidos de la inédita masividad del repositorio audiovisual de YouTube, es importante detallar el problema que enfrentó la plataforma en 2007, luego de haber sido adquirida por Google, y que requirió del uso de la inteligencia Artificial (IA) y las ya antes mencionadas redes neuronales, las mismas serán descritas en detalle un poco más adelante.

La consigna original con la que YouTube se presentó como medio fue “Broadcast yourself” (Burgess, y Green, J, 2018, p7) alentando la creación de videos hogareños, como gente visitando el zoológico o celebrando un cumpleaños. Pero muchos usuarios comenzaron a modificar ese uso sugerido por la plataforma, comprobando que también era posible hacer otras cosas, puntualmente subir capítulos de series de TV, videos musicales y fragmentos de películas, es decir, contenidos que no les pertenecían y estaban resguardados por derechos de autor.

A medida que la cantidad de usuarios aumentó, y los videos mainstream subidos por usuarios se volvieron más populares, es decir sumaban más vistas, más preocupante se volvió la situación para las compañías propietarias de esos derechos, que veían como delante de sus narices sus producciones eran distribuidas sin recibir un centavo. Disney, Warner, MTV, CBS, SONY otros gigantes del entretenimiento tomaron la decisión de demandar a YouTube, lo particular de la situación nunca había existido algo como YouTube, antes de YouTube, ningún modelo de consumo, producción y distribución semejante, ni caso de referencia directa con el cual contrastar el nuevo escenario dado.

YouTube salió airoso de las múltiples litigios por copyright, o que al menos no las perdió, pero que al final de cuentas se tuvo que comprometer en generar un sistema autónomo de detección de contenido audiovisual protegido por derechos de autor, para prohibir su subida y publicación dentro de la plataforma, o al menos para direccionar los ingresos de su visualización a los propietarios originales (Burgess y Green, 2018). Fue en este contexto en el que las redes neuronales artificiales y su capacidad para evaluar autónomamente cuestiones relativas al contenido de un video, o un fotograma del mismo, para generar en consecuencia una acción de moderación o alerta, entran en acción. En gran parte, gracias a las mismas, es que YouTube pudo resolver con relativa eficiencia este problema y contener la flagrante utilización de su plataforma para fines ilícitos y aún así mantener la rentabilidad de su negocio. La explicación implica comprender que a esta altura, entre los años 2007 y 2008, el contenido de la plataforma y la cantidad de usuarios subiendo vídeos, hacía impracticable una solución humana, YouTube rápidamente duplicó a su principal competidor (AOL video) en cantidad de videos y usuarios. Moderar y distribuir ese volumen de datos requeriría de una enorme cantidad de empleados pasando enorme cantidad de tiempo visualizando el contenido subido por usuarios, e interviniendo en consecuencia de cada infracción (Cool, K, Seitz, M., Mestrits, J., Bajaria, S, Yadati, U., 2017).

En este punto cobra sentido la explicación sobre el funcionamiento del Perceptron de Rosenblatt mencionado un poco antes, ya que de manera análoga fue requerido un proceso de entrenamiento de redes neuronales, para hacerlas capaces de reconocer y discriminar contenidos audiovisuales, que en este caso no se tratan sólo de saber si hay un hombre o una mujer en una foto, sino que se trata de conocer el contenido audiovisual realizado por las compañías querellantes ,puntualmente gigantescos volúmenes de horas de producción audiovisual, algo desde luego humanamente inabarcable.

La primera solución fue entonces entrenar redes neuronales para reconocer coincidencias sólo en la información sonora, que en términos de volumen de información es menor al del contenido visual, y por tanto requiere menor tiempo y menor tarea computacional para ser realizado.

La estrategia logró calmar a las compañías en litigio, pero el sistema era al principio muy vulnerable y comenzó a generar nuevas conductas y estrategias en los usuarios para hackear y evadir, no sólo la prohibición, sino la tecnología que se utiliza para ejecutar el acatamiento de las normas. Ejemplo de esto es la práctica de algunos usuarios que comprendieron la forma en que la tecnología estaba funcionando, y comenzaron a subir videos con copyright, pero modificando la información sonora, lo que se comenzó a conocer como ¨shreds ¨ o videos desmenuzados, en los que se satirizaba el contenido de audio original reemplazandolo por uno cómico o paródico. “Esta técnica significó una parodia cultural al mismo tiempo que una eficiente manera de evadir al algoritmo de control de contenido” (Burgess y Green , 2018, p51).

A lo largo de la década del 2010 al 2020, el volumen del tráfico de YouTube se volvió exponencialmente mayor, billones de usuarios, y miles de horas de contenido subidos por minuto, el problema para controlar y moderar el contenido audiovisual se volvió aún más complejo (Zappin, A., Malika, H., Dampiera,D. A., Shakshukib, E.N. , 2022). Otra de las situaciones habituales que la plataforma también debe resolver, implica el uso del reconocimiento de elementos con contenido explícito, violento o pornográfico, para censurar, bajar o catalogarlo debidamente para que no llegue al público infantil que utiliza la plataforma.

> El crecimiento exponencial de los vídeos en YouTube ha atraído a miles de millones de espectadores entre los que la mayoría pertenece a un grupo demográfico joven. Los usuarios maliciosos también encuentran esta plataforma como una oportunidad para difundir contenido visual indebido, o inapropiado con niños. Por lo tanto, se recomienda un mecanismo automático de filtrado de contenido de vídeo en tiempo real (Yousaf. K ,Nawaz,T. 2022, p 16283).

Esto representa un complejo nudo de responsabilidades entre YouTube, cada usuario que sube un video y cada espectador. Definitivamente, la revisión de todo lo que se sube a la plataforma debe ser moderado, no sólo por su descripción y etiquetado, sino

fundamentalmente por su contenido visual y sonoro, es decir la morfología de cada cuadro y los elementos y acciones que allí aparecen.

Este contexto da lugar al desarrollo y avance de las capacidades de reconocimiento de objetos las Redes Neuronales Convolucionales (CNN) a lo largo de la última década y lo que termina siendo consolidado en algoritmos munidos de redes neuronales con ajustes específicos y filtros, que controlan el flujo de la información en todos los sentidos, dentro de la órbita de YouTube.

El último dato estimativo tomado en noviembre de 2021, dice que "alrededor de 4000 horas de video, son subidas a YouTube en cada minuto lo que suma 65 años de video por día" (Zappin, A., Malika, H., Dampiera, D. A., Shakshukib, E.N. 2022, p 24). Para lidiar con la moderación de esa descomunal cantidad de contenido digital, YouTube analiza los videos mediante un algoritmo, del que es único propietario, que observa el contenido visual y toda su meta data, es decir, la actividad del usuario, comentarios, cantidad de vistas, etc para entre otras cosas censurar, o recomendar vídeos similares. Este algoritmo es sumamente valioso, y es mantenido y actualizado a diario por cientos de ingenieros de software, investigadores y moderadores humanos. YouTube, que pertenece a Google, no permite que la información de cómo este algoritmo funciona y como es actualizado sea pública. YouTube siquiera comparte los resultados del algoritmo de censura entrenado, es decir, datos agregados que revelan qué videos de YouTube son fuertemente promocionados por el algoritmo o cuántas visitas reciben los vídeos individuales de la sugerencia *a continuación*. Revelar esos datos sería permitir que las instituciones académicas, los verificadores de hechos y los reguladores, así como periodistas, evalúen el tipo de contenido que es más probable que YouTube promueva (Zappin, A., Malika, H., Dampiera, D. A., Shakshukib, E.N. 2022).

Uno de los principales usos que se le da al estudio de los flujos de información y consumo de contenidos, alimenta lo que se conoce como sistema de recomendaciones. El sistema de recomendación, o motor de recomendaciones es simple de describir pero complejo de desarrollar, el objetivo del mismo es recomendar contenido que satisfagan los intereses de cada usuario, que haga click en los mismos, los vea, y se sienta identificado y a gusto con lo que un sistema automático le recomienda. En principio el sistema de recomendación debe conocer al usuario, en qué cosas hace click, qué cosas compra y ve, dónde vive y qué edad tiene. Este perfil es esencial, para que de las recomendaciones dadas, sean en su mayoría bien

recibidas. El estudio de los comportamientos es casi un arte, tiene algo de detectivismo, algo de psicoanálisis y mucho de marketing, porque en definitiva, de los parámetros evaluados, al momento de reconocer la potencia de un sistema de recomendaciones, lo que mayor peso tendrá en cuanta ganancia ha generado la empresa que lo utiliza.

Se percibe en esta instancia que en cada recomendación del sistema de recomendaciones, hay un punto en la secuencia de montaje en términos cinematográficos, lo que viene luego y la estructura que lo contiene, configura un relato, y ese relato que guarda la llave del éxito comercial de YouTube o Netflix, es resguardado por las compañías y representa una influencia determinante en la vida de las personas, en cómo piensan y que consumen.

Los sistemas de recomendación y sus similitudes con el montaje serán tratados en detalle más adelante en este texto, pero como dato que ilustra la importancia de la eficiencia del sistema de recomendaciones para las empresas, es ilustrativo mencionar la siguiente situación que sucedió con el sistema de recomendaciones de la plataforma audiovisual Netflix.

En el año 2009, Netflix creó un concurso abierto para que desarrolladores creen el mejor sistema de recomendaciones posible para su plataforma, basado en satisfacción de usuarios, y baja tasa de recomendaciones no útiles, o clickeadas (Van Buskirk, 2009).

El grupo ganador llamado Belkor se hizo de la suma de 1 millón de dólares gracias a Pragmatic Chaos, el algoritmo de recomendaciones que presentó al concurso (**Figura 1**).

La compañía, no obstante, nunca puso en práctica el algoritmo, ya que reconoció rápidamente que la eficiencia del mismo no era la que mejor se ajustaba a su modelo de negocios. Es decir, un sistema de recomendaciones debe lograr un equilibrio que el sistema ganador del concurso no tenía, que constan en ser útil para que cada usuario vea rápidamente lo que le gusta dentro del menú audiovisual disponible en la plataforma, pero sin que esto atente contra la rentabilidad de la plataforma.

Figura 1: *El equipo de Belkor cobrando su cheque de 1 millón. Autor y/o fuente de procedencia.*

YouTube como medio, epicentro de desarrollos tecnológicos y modelador de consumo

La suma de las herramientas mencionadas que caracterizan al ecosistema de YouTube, desde luego no terminan allí, sino que se expanden e influencian a la sociedad global que interactúa con la plataforma, nutriendo de videastas, especialistas y fenómenos de popularidad en tiempo record, lo que es desde luego muchas veces beneficioso. Pero no son esos efectos colaterales los que interesan a esta investigación, sino otros que tal vez impacten en un plano inconsciente, de quien observa y de quien concibe, que es lo que hay para ver y cómo se despliega este menú.

La técnica de recomendaciones, característica de YouTube pero que también domina la web, no deja de ser una estrategia de montaje. En estrictos términos cinematográficos se trata de pegar una escena, un plano, o una acción con la siguiente, esa decisión central de unión queda librada a los algoritmos, si quien observa no interviene, y aún interviniendo, cambiando a una

nueva búsqueda, lo que sucederá a continuación está prioritariamente pautado por el algoritmo, el del sistema de recomendación.

Este sistema termina generando una sensación de saciedad e identificación en que utiliza la plataforma, porque relaciona una cosa con otra, entendiendo que la continuidad de lo que se parece y conocemos, es lo que preferimos ver, lo cual tiene sentido, pero al mismo tiempo parece explorar un mínimo porcentaje de las continuidades posibles.

Es sumamente improbable que esto no genere una huella en quien consume lo disponible en la plataforma, una huella de valor intelectual, e identitario.

A continuación un raconto de las herramientas que concretamente se utilizan dentro de los algoritmos en sistema de recomendación, para moderar contenidos en videos e imágenes, a partir de comprenderlos artificialmente, es decir la maquinaria que realiza las tareas de montaje sobre enormes cantidades de datos audiovisuales y de alguna forma modera su lógica de funcionamiento.

El estado actual de la detección de elementos en videos, redes CNN, RNN y datasets

Redes neuronales convolucionales (CNN)

Como fue mencionado con anterioridad, existen muchas aplicaciones para la IA, y luego también muchos tipos de redes neuronales, abocadas a distintas funciones. Los diseños en la arquitectura de las redes neuronales son casi tan variados como sus usos y funciones, pero en el caso del feature recognition, las redes que se utilizan mayoritariamente, son las llamadas redes neuronales convolucionales. Su utilización a gran escala es reciente, gracias a las capacidades de cómputo disponibles en la actualidad, y a un ecosistema de utilización masivo de datos que demanda soluciones de moderación, discriminación y distribución automatizada (Burgess and Green, 2018).

En términos generales, las redes convolucionales tienen la capacidad de contrastar algo que se les presenta como información visual, con el contenido abstracto de toda la información visual con la que fueron entrenadas, para luego a medida que avanza el proceso de análisis devolver porcentajes de coincidencia de clasificación. Por ejemplo al ver un fotograma con un sofá, responder con grados de aproximación a elementos que conoce gracias a entrenamiento, por ej. 79% sillón, 65% oso, 23% canguro.

Las redes convolucionales cuentan con distintos layers o capas, todas ellas asociadas a un modelo y este modelo a su previo entrenamiento. Las capas de la red más superficiales son más abstractas y se encargan de contrastar las texturas, y las más profundas contienen información con detalles visuales más específicos de una clase y categoría de objetos o cosas, la profundidad de las redes está de alguna manera vinculada con la sofisticación de sus funciones. Sin entrar en detalles que no hacen a la intención de esta introducción, es importante comprender que el reconocimiento de elementos gráficos simples, como letras o números, o señales de tránsito, implica un proceso de entrenamiento menos complejo que el necesario para reconocer elementos tan diversos como un automóvil familiar fabricado en Francia, o una liebre dando saltos en un una pradera nevada.

La analogía que comúnmente se utiliza para describir cómo funciona el análisis para reconocer los contenidos de una imagen, es la de una linterna alumbrando una habitación oscura. Si un espacio a oscuras es barrido con la luz de la linterna, los fragmentos que se van haciendo visibles, suman argumentos a favor de vislumbrar lo que hay dentro de esa habitación. En la redes convolucionales el ¨haz de luz de la linterna¨ se llama máscara convolucional, esta máscara estudia la estructura pixelar de cada fragmento de la imagen, y sus gradientes pixelares, para obtener similitudes porcentuales que le permitan saber qué cosas están presentes en la imagen y que cosas no, siempre en términos porcentuales, basándose en el grado de coincidencia.. Este proceso comienza por las coincidencias generales, y culmina por las más particulares, sopesando opciones para determinar con mayor certeza el resultado del análisis (Kelleher, J. , 2019).

Datasets

Para entrenar redes convolucionales, y hacerlas eficientes para reconocer elementos, hace falta que estas redes pasen por un proceso de entrenamiento en el que vean muchos ejemplos de esos elementos, para ese fin se suelen utilizar colecciones de datos conocidos como datasets. Estos datasets contienen colecciones de elementos, en este caso visuales, como fotografías, grafismos, símbolos, dibujos etc, etiquetados con nombres y referencias. A continuación algunos de los datasets más populares para entrenar redes convolucionales y algunos ejemplos de aplicación en redes neuronales entrenadas mediante los mismos.

Cifar

(Canadian Institute For Advanced Research) es una de esas colecciones de imágenes utilizadas para entrenar redes, el dataset versión 10 publicado en 2010, contiene 600 mil imágenes catalogadas de perros, gatos, aves, aviones, ranas, caballos, ovejas, camiones, autos, y ciervos, en una resolución de 32x32 píxeles. La versión Cifar 100, publicada en 2020 tiene 80 millones de imágenes, también en 32x32 píxeles de resolución, las mismas fueron descargadas de la web utilizando búsquedas automáticas por sustantivos encontrados en la base de datos wordnet. La baja resolución de cada imagen da cuenta de la simpleza geométrica a la que se reduce cada elemento, para convertirlo en una suerte de patrón visual, a partir del cual, entrenamiento mediante, se genera un modelo. (Koonce, B. 2021)

ImageNet

Es una colección de datos de imágenes organizada según la jerarquía de WordNet, solo sustantivos, 1.281.167 imágenes para el entrenamiento y 50.000 imágenes para la validación, es decir para verificar que el proceso de entrenamiento sea exitoso, organizadas en 1.000 categorías. Al igual que Cifar, ImageNet es de descarga libre y gratuita para que investigadores las utilicen para entrenar sus propias redes. (Koonce, B. 2021)

Efficienet

es una red neuronal puesta en práctica y entrenada con los datasets: CIFAR-10, CIFAR-100, Oxford 102 Flowers, Stanford Cars, Food-101, FGVC-Aircraft, Oxford-IIIT Pets, Birdsnap y constituye la herramienta más avanzada para reconocer elementos en imágenes en el presente. Efficienet, alcanza una precisión del 84,3% utilizando el dataset ImageNet, siendo

8,4 veces más pequeña y 6,1 veces más rápida en la inferencia que la mejor red convolucional existente. También logra una precisión de vanguardia en CIFAR-100 (91,7%), Flowers (98,8%), simplificando la mayoría de los parámetros de otras neuronales convolucionales (Koonce, B. 2021).

Los últimos 20 años han representado un avance notorio en el área del reconocimiento de elementos en imágenes, y hoy es posible contar con que mediante las técnicas y tecnología mencionadas es posible catalogar y clasificar contenidos visuales sin presencia de un ser humano y con un muy alto grado de precisión que tiende claramente a aumentar. Al mismo tiempo una enorme cantidad de incógnitas quedan pendientes, con respecto a la fiabilidad y universalidad de los datasets que se usan para entrenar redes, las capacidades superhumanas que desarrollen y los múltiples potenciales usos de las herramientas desarrolladas con los mismos, por sólo mencionar algunas.

Redes neuronales recursivas

Existen diversos tipos de redes neuronales, y en principio se podría decir que la estructura de las mismas, responde a la función para la que fueron diseñadas, no es el propósito de esta investigación describir todas la redes neuronales existentes, sino sólo las mismas que puedan responder la pregunta central de esta investigación, es decir, ¿cuán posible es editar videos mediante el uso de herramientas de inteligencia artificial y qué nuevas formas narrativas y usos pueden proyectar?

Se ha tratado hasta aquí el problema de la detección de elementos en imágenes de manera estática, pero tratándose de video es necesario comprender cómo también el movimiento aporta a la comprensión del contenido. Es necesario entonces tratar otros tipos de redes, con las que estamos muy familiarizados, porque están cada vez más presentes en la vida cotidiana, sugiriendo modificaciones a textos, ayudando a traducir oraciones y haciendo más veloz la redacción de un mensaje.

Estas son las Redes Neuronales Recurrentes (RNN) y la subcategoría de la redes de memoria a corto plazo (LSTM). A diferencia de las redes neuronales de una sola dirección, de entrada y salida estándar, las LSTM tienen conexiones de retroalimentación. Puede procesar no solo

puntos de datos individuales (como imágenes), sino también secuencias completas de datos como por ejemplo series de palabras en una oración, o secuencias de frames de video.

Cuando un mensaje es escrito en un procesador de texto o una red social, suele suceder el modo predictivo se anticipa a la siguiente palabra y la sugiere antes de sea efectivamente escrita, si modificamos el texto previo a esa palabra latente, muy probablemente la palabra que nos sugiere el procesador de texto también cambie. Esto es la evidencia que revela que el sistema de recomendaciones está rastreando continuamente en la secuencia, iterando y analizando nuevamente tratando de comprender el sentido del orden. Esto es posible gracias a las conexiones de retroalimentación (recurrentes) en este tipo de redes neuronales, que son eficaces para estos tipos de problemas, los temporales y de secuencia.

Las RNN también son originalmente entrenadas mediante secuencias válidas o correctas de texto, y en ese entrenamiento se basa su capacidad para aprender a colaborar con la escritura en tiempo real, a partir de un modelo surgido del entrenamiento con millones de ejemplos de textos validados como correctos.

Como se mencionó anteriormente, la capacidad de estas redes es la de ser conscientes de una secuencia temporal o de orden, a diferencia de las redes convolucionales, que son utilizadas para escrutar una imagen y determinar los elementos dentro de la misma, las redes recursivas son capaces de ordenar datos uno detrás el otro, pudiendo verificar a cada paso el sentido general de la secuencia que se va conformando.

A partir de lo que se ha detallado previamente es posible imaginar algunos usos dentro del ámbito del video para estas redes. Como hemos visto, las redes CNN son capaces de reconocer elementos en imágenes, y reconocer que hay en una imagen significa, entre otras cosas, que una imagen o un fotograma pueden convertirse en una etiqueta, es decir en una palabra. Algunos de lo diseños de comprensión de secuencia de video proceden del siguiente modo, extrayendo fotogramas para analizarlos con una red convolucional cada determinada cantidad de tiempo, a una tasa promedio de 5 fotogramas por cada segundo, esto permite tener las palabras que detallan el contenido de cada imagen en una secuencia temporal.

Es decir, lo que finalmente resulta en palabras sueltas, que de algún modo representan a las escenas visuales de cada fotograma en secuencia, esto potencialmente significa que si accedemos a ordenar las palabras utilizando redes RNN, que ayuden a secuenciar las esas

palabras en un sentido correcto, válido o comprensible, será posible contar con la posibilidad de hacer lo mismo con los videos que les anteceden, editando video a partir de etiquetas, sin nunca haber manipulado el video original.

No existe un nombre académicamente unificado para esta técnica de edición, ya que su uso es muy reciente y no aún popular, pero se ha comercializado para aplicaciones móviles, como es el caso de FLO, bajo el nombre de Neural edit, edición neuronal.

Esta clase de aplicaciones se promocionan como capaces de sintetizar en pocos minutos, los más interesante del contenido de los videos en tu teléfono móvil, utilizando sólo Inteligencia artificial, pero como será detallado más adelante, este slogan no solo es ambiguo, sino también, en alguna medida, jactancioso.

Otra estrategia implica una aproximación muy distinta, y está vinculada a reconocimiento de movimiento en sí, (human action recognition) más que en el contenido de elementos.Una analogía podría ser que que se busca el verbo de lo que está sucediendo en una escena y a partir de comprender el flujo pixelar en diferentes fotogramas analizados, reconocer un patrón de movimiento característico, asimilado en entrenamientos previos (Orozco C. I., Xamena, E. Buemi, M. E. Berlles, J.J. , 2020). (**Figura 2**)

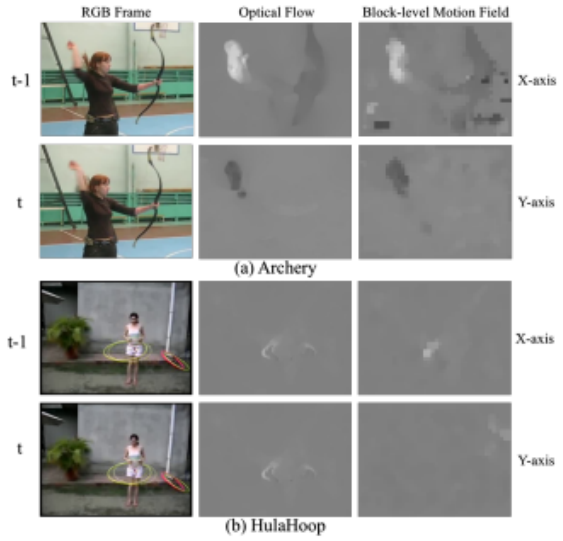

Figura 2: *Detección de flujos de movimiento para reconocer movimientos humanos. Wukui, Y, Shan, G, Wenran, L, Xiangyang, ji, (2018)*

El estudio de este flujo pixelar, requiere de un tipo de red que sea consciente del tiempo y el orden, dentro de las rede RNN hay algunas específicamente denominadas De largo corto tiempo de memoria (LSTM) que pueden procesar no sólo puntos individuales de datos,sino también secuencias, como habla humana, o video, para reconocimiento de tareas no segmentadas. Sin profundizar en aspectos técnicos complejos de esta otra aproximación al problema de la comprensión artificial del contenido de un video, es posible decir que su desarrollo está muy emparentada a las técnicas de compresión de video por macrobloques pixelares.

El crecimiento del volumen de contenido audiovisual

En los últimos 20 junto con el crecimiento de la web han proliferado y evolucionado con creciente velocidad, técnicas de comprensión artificial, muchas surgidas de un dato difícil de eludir, es humanamente imposible moderar el contenido audiovisual de lo que se produce y se sube al red, hace falta que sistemas autónomos se encarguen de controlar y distribuir el enorme flujo audiovisual que a tasas gigantescas se sube a plataformas como YouTube a diario, y para esa función estos sistemas autónomos deben relativamente comprender lo que ordenan. En este capítulo se pretendió discriminar y describir algunas de las técnicas surgidas o potenciadas en estos procesos, mediante las cuales, como en un una fábrica sin seres humanos, las imágenes van de un lado al otro del mundo, siendo catalogadas no solo por las palabras que las describen, sino también por sus contenidos, siempre evaluados por el ojo de un algoritmo.

Tal como se menciona en la introducción en esta investigación se pretenderá reunir los datos para saber si existen maneras de que esas estrategias y técnicas, puedan estar al servicio de otros fines y utilidades concretas, fundamentalmente estudiando sus cualidades como recursos creativos y artísticos, manipulando volúmenes de información accesibles, desligando su uso de las necesidades de marketing y los modelos de negocios para los que fueron desarrolladas, proyectando otras funciones en el campo del arte y el diseño, particularmente viendo cuán capaces son de construir secuencias de sentido yuxtaponiendo videos.

Capítulo 2 Variaciones sobre el montaje

Introducción al capítulo 2

En el primer capítulo se establecieron datos sobre los avances y estado actual de la informática asociada a tareas de comprensión artificial de contenidos audiovisuales, entendiendo que la comprensión del contenido de un video, es un paso previo y necesario, al montaje de un video con otro, con fines narrativos, o de conformación de sentido.

En este segundo capítulo, Variaciones sobre el montaje, se ejemplifican y estudian varias concepciones de montaje dentro de las artes audiovisuales, esencialmente como facilitador de relatos y articulador de sentido. Abrir esta pregunta a ejemplos menos convencionales es importante, intentando reconocer las capacidades de la I.A. como montajista, es necesario también presentar definiciones de montaje diversas sobre las que sea posible contrastar esas capacidades.

La mayoría de las concepciones que serán estudiadas en este capítulo surgieron previamente a la convergencia tecnológica entre informática y montaje, lo que intenta validar la posterior interrogante respecto a las posibilidades reales de si un sistema autónomo es capaz de confeccionar sentido uniendo secuencias audiovisuales.

No obstante eso, algunas de estas concepciones y prácticas de montaje que serán detalladas, en algunos casos consideradas alternativas, guardan a priori, algunas características en común con las posibilidades que presenta este nuevo contexto.

El montaje como estructura, la estructura como pensamiento

El montaje, puede decirse, es uno de los recursos más distintivos dentro del campo audiovisual y los lenguajes cinematográficos, como herramienta constructora, por su poder como hilvanador de sentido y su presencia es tan definitoria como invisible, no hace falta argumentar demasiado para decir que en las bases fundacionales del cine, ya sea para generar

la ilusión de movimiento, como para estructurar sentido, la yuxtaposición de fotogramas, planos, o escenas, es esencial.

Cada escuela o vanguardia cinematográfica ha desarrollado una teoría y técnica del montaje, no es el objetivo de este estudio generar definiciones sobre esto, pero valen pocos ejemplos, como el efecto Kuleschov, al montaje de atracciones de Einsestein, al jumpcut del cine moderno, el collage posmoderno, el decollage fluxus, la matemática del montaje en el cine estructural, etc, para entender que el montaje no es solo un gesto, sino una herramienta estética que expone una mirada y posición sobre qué implica contar en imágenes y sonidos.

En los siguientes párrafos serán expuestos también ejemplos previos al cine, y otros de ámbitos experimentales, y las prácticas audiovisuales contemporáneas, no para estudiarlas en extenso, sino con el objetivo de poner en evidencia el parentesco que existe entre los sistemas de recomendación y las prácticas de montaje, como estructuradoras de sentido, surgidas algunas en la génesis del cine, y otras basadas en técnicas de reciclado de material audiovisual, en la informática y en la interacción.

El montaje no lineal, antecedentes

> *Cuando viejas construcciones son sumadas dentro de nuevas, elementos individuales de lo antiguo, son preservadas en lo nuevo* (Zielinski, S., 1999)

Panorama y montaje

Es lógica la mención del Panorama, popular luego de fines de la Revolución industrial de fines del siglo XVIII, como antecedente del montaje cinematográfico. Dos de los denominadores comunes de la época son sin dudas la velocidad y la eficiencia, ambos encarnados por la locomotora a vapor como símbolo y cúspide de ese periódo. La introducción de este nuevo medio de transporte en el imaginario popular, habilitó probablemente muchas fantasías, una de ellas, la de recorrer cientos de kilómetros a gran velocidad uniendo sitios distantes en un abrir y cerrar de ojos.

En ese contexto surgió este tipo de espectáculo que parece remediar la ansiedad por experimentar en primera persona la sensación del viaje, es este dispositivo para viajar mediante imágenes y algunos sonidos, se hizo popular con el nombre de Panorama.

La primera característica de los panoramas es que requieren de una arquitectura específica donde ser visitados, deben ser grandes espacios, para que el espectador recorra y se deje envolver por las imágenes que ilustran las paredes de la sala, la idea es sentirse dentro del paisaje. Los Panoramas no fueron sólo populares en Europa y Norteamérica, también Argentina hubo varias representaciones de batallas y pasajes a modo de panorama.

La arquitectura debía conceder lugar para cubrir 360 grados de continuidad visual, algunas de estas estructuras se conocen como rotundas o rotondas, ya que son edificios de forma circular que permiten la circulación del público acompañando la curvatura de las paredes. En algunos casos estas edificaciones contaban con varios pisos, segmentando el recorrido del paisaje o presentando varios recorridos en una sola visita.

El sentido por el que resulta oportuno mencionar el Panorama, es por que cumple con varios principios básicos de lo que luego se instituyó como montaje, fundamentalmente para poder sintetizar en pocos metros el contenido de varios kilómetros de paisaje, lo que significó la implementación de distintas estrategias de discontinuidad. “Algunos panoramas no hicieron ningún esfuerzo por ocultar los espacios entre las vistas. No se trata, propiamente hablando, de un panorama… sino más bien de una serie de imágenes” (Huhtamo, E., 2013, p 255).

Había quienes intentaban construir los extensos paisajes generando formas sutiles que se encadenan las unas con las otras, lagos con bosques y montañas con ciudades, dando siempre sensación de continuidad, y por el contrario había artistas que generaban notorias interrupciones o cortes sobre el plano, sobre la morfología continua del paisaje. Estas posturas, la de ocultar, o evidenciar los cortes, fueron, de algún modo, heredadas por el cine, y como será expuesto más adelante, también se emparentan con estrategias contemporáneas en plataformas de consumo audiovisual.

El caso del Myriorama

Junto a la aparición de los Panoramas, también se difundió el uso de un dispositivo relevante desde el punto de vista del estudio del montaje, que por otro lado también es motor de pensamiento principal que dió origen a la presente investigación, el Myriorama.

El Myriorama, cuyo origen etimológico lo emparenta con la posibilidad técnica de ver una miríada de imágenes, es decir una decena de miles, o una gran cantidad de imágenes distintas, es una versión de bolsillo del panorama, que permite dividir un paisaje dibujado en fracciones de pequeñas tarjetas. Una de las cualidades del Myriorama, es que cualquiera sea el orden en el que se colocan las tarjetas, una al lado de la otra, siempre se obtendrá la sensación de continuidad.

El truco para que esto suceda es que las breves fracciones de paisaje presentadas en cada tarjeta, tienen siempre el horizonte dibujado a la misma altura, a la izquierda y a la derecha, configurando un punto de transición que permite interactuar colocándolas en cualquier orden. Todas las tarjetas pueden ser consecutivas entre sí (Hyde, R. 2007).

La función de los mirioramas, aparte de despertar la curiosidad mágica antes descrita, es la de proveer a su dueño de un sistema no lineal de relatos, casi infinito. Las tarjetas se pueden mezclar como naipes, para luego darlas vuelta y ordenarlas prolijamente una al lado de la otra, y así comenzar a improvisar un relato oral que sea consecuente con el orden asignado por el azar y permitido por el ingenio del horizonte continuo.

Hay, al menos, dos características principales que hacen que este dispositivo técnico narrativo del siglo XIX sea importante para la presente investigación. La primera es la presencia de la no linealidad, y la interacción que permite este simple juego de 200 años de antigüedad, que es previo al cine, y que explora con sofisticación las posibilidades de lo aleatorio y los relatos múltiples, que el cine no se permitió durante muchos años, tal vez por estar atado a tecnologías que no propiciaban esa versatilidad.

Por otro lado, el Myriorama es un notable artilugio para enseñar a componer un relato. Cada tarjeta tiene su núcleo de interés, en cada tarjeta sucede algo, un accidente geográfico, una dama que se perdió en el bosque, un joven que se quedó dormido, ganado suelto en un bañado, un grupo de músicos ejecutando sus instrumentos, acciones que permiten anclar

relatos. Las tarjetas permiten según su casi infinito orden generar pausas, desencuentros, distancias infranqueables o reuniones, hay lo suficiente allí como para entretener a quien pone en práctica el juego, y todo gracias al perfecto punto de transición en el horizonte.

De alguna forma, esta exploración transversal que permiten las tarjetas revela que con algo de ingenio podríamos contar historias sin pensar en la secuencia que se conforma, sino en las características de cada una de sus partes aisladas que luego la compondrán.

En ese, entre otros, sentidos es que el contenido audiovisual casi infinito y apenas conocido de YouTube parece propicio para transponer el funcionamiento del myriorama, porque representa una materia prima inédita, de las acciones y el aspecto de las personas de nuestro tiempo. Configurar historias uniendo partes de videos ignotos, nos permitiría visualizar una suerte de estado de las cosas global, junto con sus múltiples paisajes y seres.

Breve historia del montaje interactivo en relatos cinematográficos

"El cine constituía una forma narrativa tan nueva e insólita que la inmensa mayoría del público no acertaba en comprender lo que veía en la pantalla, ni establecer una relación entre los hechos" (Buñuel, L. 1982).

La experiencias del cine interactivo, o que implica alguna forma de participación del público, son casi tan antiguas como el espectáculo cinematográfico en sí. Probablemente, porque el público carecía de formación, el modelo de cómo ver y comportarse en el cine tomó unos años en gestarse, y las experiencias previas en espectáculos populares eran más interactuables y variaban de vez en vez. En tiempos primitivos del cine se sabe que el público compenetrado en los relatos, arrojaba cosas a la pantalla enardecido por la presencia de un villano, o pedía ver nuevamente una escena que le causaba gracia, lo que de algún modo representa la presencia de la interacción. También a principios del siglo 20 existieron experiencias interactivas que incluían proyecciones cinematográficas, por ejemplo la insólita "caza en el cine", dónde grupos de personas armadas asistían a practicar caza apuntando (y disparando) contra la pantalla. Cabe decir que por más disparos que recibía la tela blanca donde se proyectaba, el resultado de la proyección era el mismo, los ciervos y venados que correteaban

en la película no sufrían ni un rasguño, por el contrario la pantalla si debía ser reemplazada por una nueva al finalizar el espectáculo.

Más adelante, durante la década de 1960, se comienzan a desarrollar experiencias más sofisticadas, incluyendo modificaciones en el trayecto del relato audiovisual y desde luego también en el final, mediante la interacción del público. Proyectos como Mr. Sardonicus de 1961, donde el público bajaba o subía un pulgar, como en YouTube, pero en este caso para modificar el final de la película. Otro caso es el denominado Kino Automat, estrenado en 1967 son ejemplo de esas inquietudes. En este último espectáculo se presentaban puntos de transición nodales, seleccionados por quien creaba la obra, donde el público debía votar por distintas alternativas, para luego proseguir con el relato. La experiencia y trayectoria dentro del relato terminaba siendo particular para cada persona que asistía a una proyección (Koenitz H., Ferri G., Haahr M., Sezen D., Sezen T. I., 2015).

La década de 1980 con el advenimiento de la digitalidad a gran escala y la cultura emergente de los videojuegos se produjeron cientos de piezas que se parecen al cine, pero que al mismo tiempo dependen de la interacción del público, cuya categoría también entró en discusión, apareciendo términos como *usuario, participante, viewser* (Koenitz H., Ferri G., Haahr M., Sezen D. Sezen T. I., 2015), actores de la interacción para cerrar la trama, a partir de la destreza o del azar.

Mediante esta breve enumeración, es posible hablar de un vínculo tal vez no tan antiguo, pero sí maduro, entre la narración audiovisual y la interacción, que parece terminar de establecer nunca un estándar dentro del cine. Las películas nunca se volvieron interactivas, el cine conservó y conserva como espectáculo tradicional su linealidad, al margen de experiencias aisladas de diverso éxito y notoriedad.

Otros ámbitos se nutrieron más profundamente entonces de esas experiencias de interacción, la televisión, y los videojuegos fundamentalmente, son familiares más cercanos de ese matrimonio entre interacción y relato que el cine canónico nunca terminó de conciliar.

Regresando a la cuestión fundamental de este texto, el montaje, la consecución de partes audiovisuales que constituye una herramienta de pensamiento, transmisión de ideas y generación de sentido, para la cual es difícil encontrar un ejemplo más determinante de la utilización de ese potencial, que el que despliegan plataformas como las que se han

mencionado previamente, fundamentalmente YouTube, dentro del espectro más amplio de las plataformas que utilizan sistemas basados en recomendaciones.

Es cierto que algo semejante sucede con la televisión a demanda (on demand), el caso de Netflix, donde también hay un motor de recomendación, que sugiere que pelicula o serie ver luego, en función al minucioso estudio de las acciones previas del usuario y las conductas generales de su contexto socioeconómico y geográfico al que pertenece. Pero hay diferencias fundamentales aquí, y son las que vuelcan el interés de esta investigación hacia el caso de YouTube.

La primera es la naturaleza diversa del contenido audiovisual disponible en YouTube, allí nos podemos cruzar con videos para reparar licuadoras,ensayos de obras teatrales en un jardín de infantes en Mongolia, o breves videos de un exótico y colorido insecto. Más del 99% de lo que está subido en YouTube, podría catalogarse bajo lo que en la jerga se denomina *crudo*, mientras que en el caso de Netflix, todo está empaquetado, cumpliendo con las normas de Modo de Representación Institucional (MRI).

Con respecto a esto hay otro dato llamativo, el 1% del material audiovisual disponible en YouTube se parece o repite lo que hay en Netflix, y ese 1% de contenido premium (videoclips musicales, películas pay per view, series, etc) se lleva casi el 95% de la audiencia total. Solo el 0.1% de los videos alojados en YouTube supera el millón de vistas, mientras que más del 90% de los videos alojados en la plataforma no llega a las mil visualizaciones según la plataforma de análisis de flujos de datos en la web PEX (**Figura 3**).

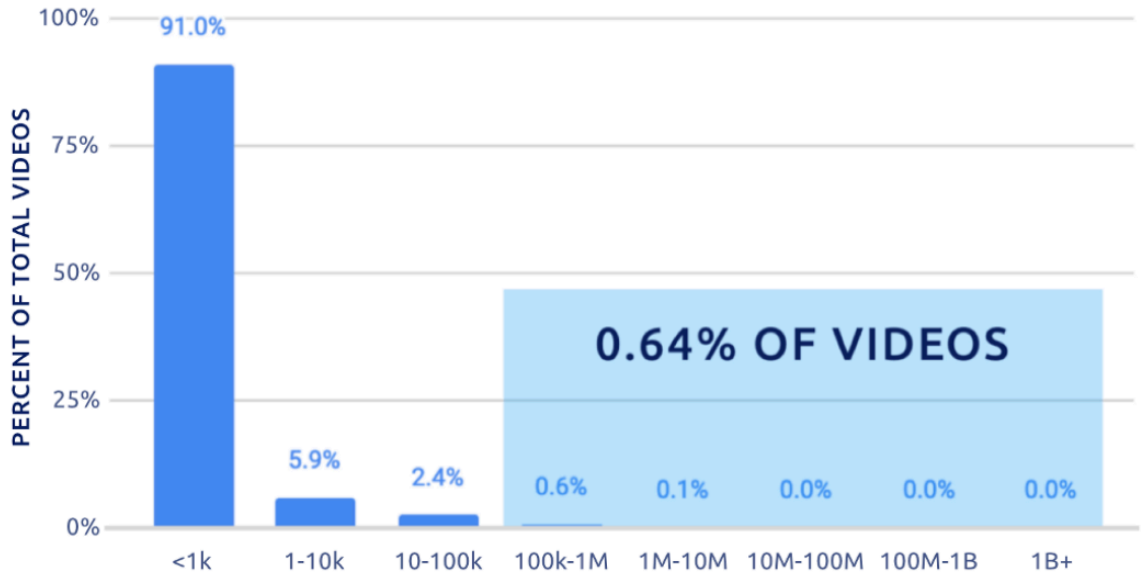

Figura 3: *Gráfico con el que la compañía de análisis de flujo de datos en la web PEX, exhibe la distribución de público por video en YouTube.*

Haciendo un poco más de matemáticas, sumando algunos otros datos mencionados anteriormente, es posible afirmar que a YouTube se suben más de 65 años de video por día, según cifras obtenidas en 2021 (esto significa 180 siglos de video por año), que muy posiblemente nunca, nadie verá (Zappin, A., Malika, H., Dampiera, D. A., Shakshukib, E.N., 2022).

Este colosal repositorio videográfico de la humanidad está disponible, es la videoteca más gigante que jamás existió, pero rara vez quien la visita se topa con esas piezas ocultas mientras plácidamente mira videos, saltando de uno a otro o dejando que la playlist monte su propia trayectoria para cada usuario, que por lo general merodea dentro del 1% de lo disponible, y el motivo es obra del montaje que realiza o sugiere el sistema de recomendaciones.

Qué es un sistema de recomendación, y por qué se parece al montaje

En el contexto de consumo de contenido audiovisual en plataformas como YouTube es prácticamente imposible no entrar en contacto con las sugerencias de los sistemas de recomendación, las miniaturas a la derecha de cada video que vemos en YouTube, el video que continúa al que estábamos viendo si dejamos que continúe automáticamente, y en algún punto la publicidad que interrumpe la visualización si es que no tenemos instalado un bloqueador de anuncios. Toda esa cadena de montaje está bajo el ámbito de acción del sistema de recomendaciones.

Citando a Gilbert Simondon en su *Modo de existencia de los objetos técnicos* accedemos a una lectura respecto a la profundidad filosófica de esta clase de esquemas humano y máquina, contrapuesta a la expresada por las aplicaciones en sus manuales de uso y publicidad, que abundan en los beneficios del ahorro de tiempo .

> El individuo se convierte solamente en el espectador de los resultados del funcionamiento de las máquinas, o en el responsable de la organización de los conjuntos técnicos que hacen funcionar las máquinas. Esta es la razón por la cual la noción de progreso se desdobla y se convierte en angustiante y agresiva, ambivalente; el progreso está a cierta distancia del hombre (Simondon, G. 2017, p 134).

Haciendo un ejercicio de *arqueología de medios* (Huhtamo, E., Parikka, J., 2011) pero yendo solo unos años atrás, es posible comprobar varios cambios en el aspecto del panel de uso, y en la sofisticación de las herramientas disponibles, en principio, en los elementos que se recomienda ver. Esto se debe a la sofisticación que ha cobrado el algoritmo responsable de seleccionar que vale la pena ver, una vez que se detectó lo determinante de su función (**figura 4**).

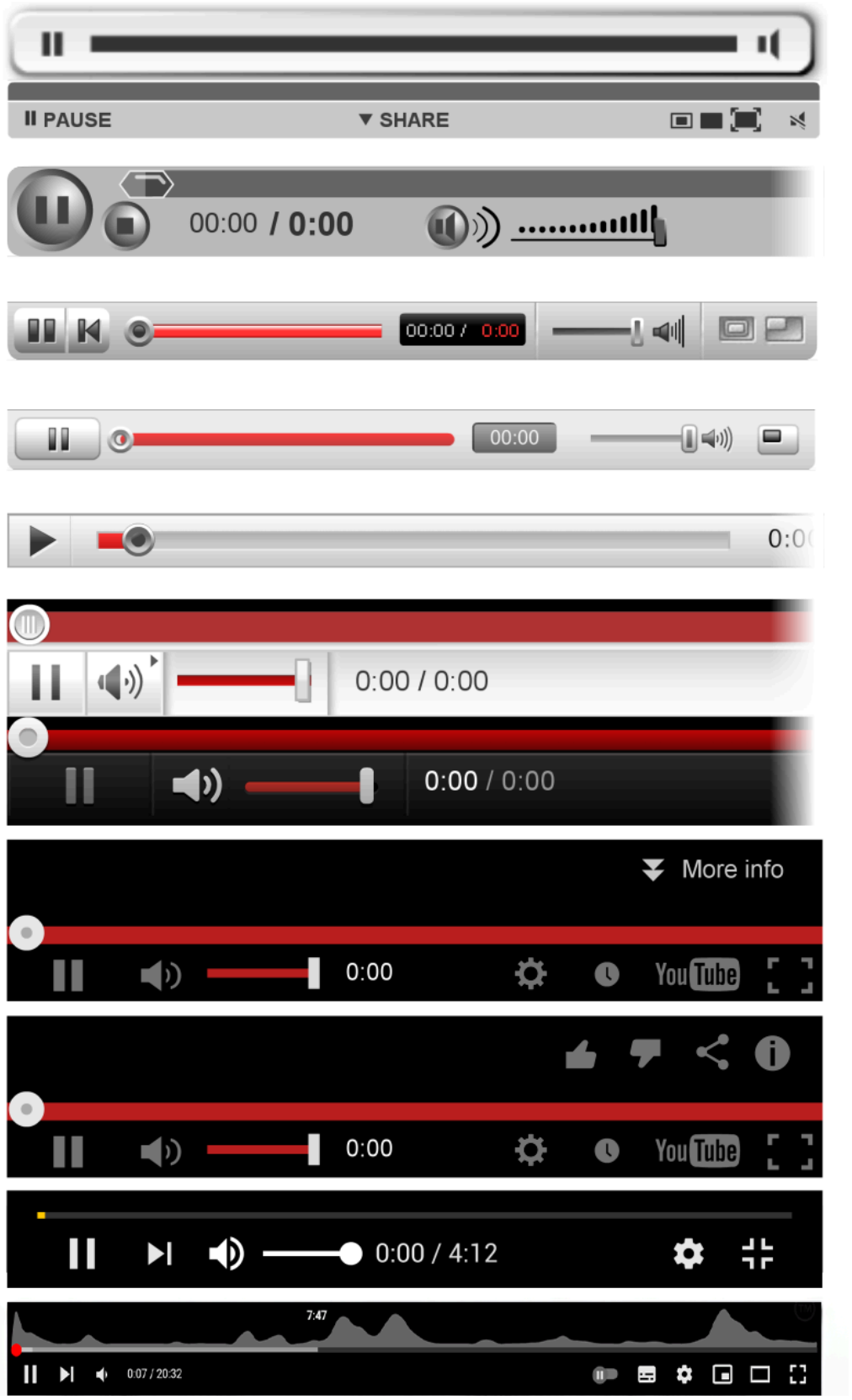

Figura 4: *Soficticación de la barra de control de YouTube de 2007 a 2023 (historia de Youtube wikipedia + capturas de autor desde la plataforma YouTube).*

Un sistema de recomendación es un sistema de montaje, porque en definitiva estructura secuencias temporales, y lo hace cada vez con más conocimiento, se trata de un acto deliberado que depende de una estructura absolutamente estudiada, es decir, hay un sentido detrás del orden de las recomendaciones. “La función esencial de un sistema de recomendación es predecir las preferencias personales de los usuarios” (Schrage, M. 2020, p 3).

El juicio de selección y recomendación puesto en práctica, depende de varias de las técnicas mencionadas en el capítulo 1, entrenamiento de redes neuronales, para conocer qué hay dentro de un video y en base a una cantidad metódicamente estudiada de valores contextuales saber qué corresponde hacer con el mismo. La columna vertebral del funcionamiento de lo descrito son los datos que entrenan sistemas para hacerlos autónomos, y que estos sistemas luego motorizan algoritmos, que muchas veces ayudan a encontrar algo con mayor facilidad, otras evitan contenido inapropiado. En cada acción dentro de la plataforma, está presente (Kane, F. 2018).

Reconocer la existencia del sistema de recomendación basado en estudio de datos y su sentido práctico parece apropiado para entender que otras cosas se podrían hacer, si la función del mismo fuese alterada, o modificado su contexto.Hacer un raconto de lo que se considera durante el proceso de programación de un sistema de recomendación, y cuales son las principales variables de corroboración para saber de su eficiencia, es fundamental en este punto.

A continuación algunas de esas variables explicadas en cursos genéricos para creación de sistemas de recomendación para todo tipo de plataformas, de compras, de citas, y desde luego de consumo audiovisual.

Diversidad

En el estudio de eficiencia de un sistema de recomendación, la diversidad es un valor fundamental, estrictamente representado por la variedad de características de los ítems sugeridos, en el caso de YouTube, estamos hablando de los videos en miniatura a la derecha del reproductor, cuyo orden de prioridad también depende del sistema de recomendaciones.

La diversidad está basada en la previa catalogación de esos ítems, y esta catalogación como sabemos es minuciosa y en la misma participan una enorme cantidad de factores, algunos de los cuales ya hemos mencionado, otros son el nombre del ítem, su contenido, su fecha de publicación, su duración, de grado de popularidad etc.

“El objetivo de un sistema recomendador es que cada consumidor reciba una lista acotada de ítems recomendados que le resulten más atractivos, extraída de una lista mucho más grande de elementos” (Kane, F. 2018).

Luego de ver un video en YouTube, aparecen disponibles en principio 10 recomendaciones entre las cuales seleccionar qué ver luego, la diversidad, en rigor, es el nivel o grado de variedad dentro de esos 10 elementos, cuanto menos se parecen entre sí, más alto es el nivel de diversidad del conjunto de ítems. Los sistemas de recomendación tratan de mantener el nivel de diversidad en el mínimo posible, de esos 10 ítems sugeridos para visualizar a continuación, probabilmente 8 comparten similares características al que acabamos de ver, es decir son poco diversos, mientras que los 2 restantes muy probablemente estén basados en similitudes de algo que vimos anteriormente, o algo que es popular entre la gente de nuestra edad o nuestra región geográfica.

Para los sistemas de recomendación la diversidad es un problema, porque se supone genera desconfianza en el usuario, que no se ve identificado con el contenido sugerido y deja de utilizarlo como herramienta. Es decir, los sistemas de recomendación recomiendan cosas que son muy parecidas entre sí, y esta necesidad tiene como soporte el extenso estudio de actitudes humanas en la red y la consecuente minimización de permanencia y consumo que busca la compañía propietaria del sistema, para cada uno de sus usuarios.

Para ejemplificar esto es valioso traer nuevamente el caso de Netflix y su concurso en 2009, para que programadores externos creen el mejor sistema de recomendaciones posible. El equipo ganador desarrolló un sistema que permitía acceder de manera eficiente al contenido de la plataforma, los testers y parámetros de su funcionamiento eras super positivos, excepto por uno, usando ese sistema la compañía ganaba menos dinero, porque la trayectoria de los usuarios agotaba más rápidamente el contenido del sitio, por lo tanto ese sistema de recomendación llamado “caos pragmático” de dejó de lado (Kane , F 2018).

Es entonces importante reconocer que valores como la diversidad deben ser percibidos como tales para quien consume, es decir la persona que interactúa con la plataforma debe experimentar variedad, y novedad, pero al mismo tiempo estos valores están regidos por la utilidad comercial de la plataforma, cosa que naturalmente manipula finalmente estos valores en función de sus propias necesidades.

Novedad

Otro de los valores fundamentales en la programación y entrenamiento de sistemas de recomendación es el de novedad. La novedad refiere a la diferencia entre el ítem que acabamos de ver y el que se sugiere que veamos luego, que se considera que el usuario no conoce, y nunca vió. Si el nivel de novedad es muy alto se trata de un valor negativo para los sistemas de referencia, porque significa tomar un riesgo, desinteresar a quien está usando la plataforma por presentarle algo que se parece poco a lo que acaba de ver.

La tendencia entonces es que la consecución de cosas que se parecen es la más sensata y nuevamente más provechosa y rentable para las compañías, y esto se verifica en click, compras y acciones positivas de usuarios en plataformas que utilizan sistemas de recomendación tan diversas como amazon, ebay, spotify, academia, tinder, y desde luego YouTube.

Los valores de novedad y diversidad son parámetros mediante los cuales se estudian y entrenan los sistemas de recomendación, siempre que un sistema es puesto en práctica se verifica antes que el valor de novedad y diversidad en las recomendaciones de los mismos sea bajo. Es llamativo sin embargo que la sensación al utilizar estas plataformas es muchas veces la contraria, se experimenta la sensación de estar aprendiendo y conociendo cosas nuevas gracias a las recomendaciones. La explicación de este fenómeno contradictorio en el que la máquina ofrece siempre de los mismo, y quien consume cree que hay gran diversidad y novedad en su consumo, tiene seguramente una explicación psicológica, pero otra que es más bien simple, y es que quien navega ignora el verdadero volumen de información que contiene cada repositorio cuando lo consulta o visita, porque en nuestra pantalla, vemos a los sumo 20 videos en miniatura, esto es inclusive desde la lógica de estudios de la experiencias de uso y de las interfases un problema resolver, el de cómo navegar el contenido de una base de datos

con más de *500 siglos de video*, sabiendo que llevaría la vida completa de 700 personas ver la totalidad de ese contenido.

Filtrado

Los sistemas de recomendaciones creados con fines comerciales, cuentan con una puerta de entrada que permita intervenir y direccionar el resultado de las búsquedas si fuese necesario, con el objetivo de hacerlas coincidir con un grupo de resultados más acotado que el disponible, para eso entre otras cosas, se crean se crean *filtros* (Kane , F 2018).

No es tan trascendente para este proyecto el caso de los filtros, porque implican principalmente el aspecto comercial de los sistemas de recomendación, para posicionar productos y ofrecerlos a la persona que más probablemente los necesite, o comience a creer que los necesita.

El filtrado requiere de la comprensión de los perfiles de quienes usan la plataforma, estos perfiles basados en comportamientos, acciones y datos identitarios cargados por cada usuario, dentro de las magnitudes inéditas de la Big Data.

Distintas estrategias son puestas en práctica por las plataformas, para filtrar y actualizar los sistemas de recomendación, en función de la interpretación de las conductas de cada usuario y los modelos socioeconómicos y geográficos a los que corresponden, en este sentido el sistema de recomendación comple una función semejante a la de los estudios de marketing generales, no obstante su modo de aplicación dentro de los sistemas de recomendación, es mediante lo lógica de montaje, construyendo secuencias.

Elementos a tener en cuenta en la interacción/ montaje

Como fue descrito con anterioridad, el cine interactivo ha estudiado durante no menos de 60 años las posibilidades de la interacción entre el espectador y las acciones que suceden en la pantalla, presentando distintos tipos de opciones entre las cuales quien observa debe decidir,

para de tal modo alterar los posibles derroteros del relato en cuestión y sus desenlaces (Koenitz H., Ferri G., Haahr M., Sezen D Sezen T.I., 2015).

Es común a estas experiencias la construcción a partir de diagramas de flujo para estructurar y visualizar las alternativas del relato, cada punto desde el cual se abren posibilidades y opciones, se conoce como punto de transición, y su aplicación representa la comprensión del ritmo del relato y la oportunidad de modificación del mismo.

Es coincidente con estos estudios, que interpretan el montaje como una herramienta común al creador e interactor, el estudio de los momentos propicios para generar un quiebre en el relato dónde el público tome en sus manos la decisión de cómo debe continuar la historia. En la detección de esos momentos se ha basado gran parte del estudio de comportamientos de los sistemas de recomendación de video, y no sólo eso, la duración de los videos nativos de Youtube y orientados a esa plataforma consideran estadísticamente cuando tiempo tarda el espectador promedio en saltear un video, adelantarlo, pasar de capítulo, etc.

El estudio general de comportamientos para saber en qué punto algo se torna aburrido para el público, es prácticamente una regla de oro para tener éxito con contenidos audiovisuales en la web, ese dato, es central en la información analítica que YouTube le provee a cada creador de contenido, y por tanto también es insumo de modelos para el desarrollo de formas de montaje automatizadas.

Otras prácticas de montaje audiovisual

Resumiendo lo expresado hasta ahora en este capítulo, se mencionó el montaje y las experiencias interactivas previas a la existencia del cine, y paralelas a la popularidad del mismo, fundamentalmente a lo largo del siglo 20, y luego se trazaron algunos puntos comunes entre las experiencias interactivas y el escenario de consumo audiovisual en YouTube, como caso de estudio específico de la presente investigación. Se detectaron prácticas experimentales del cine interactivo, que están vinculadas por similitudes de estrategias de moderación de material audiovisual en YouTube, mediante sus herramientas analíticas y sistemas de recomendación (Burgess, J. y Green, J. , 2018).

Lo que sigue a continuación tiene que ver con concepciones de montaje alternativas, llevadas a cabo en andariveles paralelos al cine popular, mediante estrategias que apelan fundamentalmente a la reflexión, alterando las reglas convencionales de construcción de espacio y tiempo, entre la obra y cada espectador.

Es fundamental mencionar estas prácticas alternativas (o experimentales), puesto que el cruce de las mismas con las herramientas potenciadas dentro del marco de los grandes repositorios de vídeo como YouTube (las redes neuronales convolucionales y recurrentes, los sistemas de recomendación y moderación mediante IA, etc,) abre un gran campo de acción y reflexión para experiencias artísticas y científicas futuras.

Las otras prácticas

Esta concepción de otras prácticas de montaje, fundamentalmente refiere a dos campos que en alguna medida tienen puntos de coincidencia, el del denominado Cine Estructural surgido en la década de 1960 y la práctica conocida como Found Footage, o metraje encontrado, asociada al reciclado de material audiovisual.

Tanto el Cine Estructural como el Found Footage, como se demostrará a continuación, tienen puntos de contacto con algunos principios seminales de transmisión de ideas mediante el uso de sonidos e imágenes y su ordenamiento y montaje. El primero de esos puntos a destacar es el que las asocia con el antes mencionado efecto Kuleshov como evidencia de la importancia del montaje al momento de interpretar la secuencia de dos planos yuxtapuestos como unidad de sentido (Monahan, M. 2015). Lo que pone de manifiesto este reconocido efecto es de gran importancia para lo que se estudia en el transcurso de este texto.

No se trata necesariamente grandes relatos estructurados de manera automática, sino más bien estructuras simples de consecución, que revelan caminos poco transitados dentro del mayúsculo volumen de contenido audiovisual disponible en YouTube, para que de las mismas emerjan también otras formas menos habituales de sentido.

También Sergei Eisenstein y su teoría del montaje, particularmente la idea del montaje de atracciones, es útil para dejar nuevamente en claro sobre qué clase de aspectos del montaje

pretende reflexionar este texto, actualizando problemáticas teóricas dentro de las magnitudes y con las herramientas inéditas que vengo mencionando antes.

Volviendo a la teoría de Eisenstein, “ Si dos partes de película, de cualquier tipo, son colocadas una detrás de otra, inevitablemente se combinan en un nuevo concepto, de nuevas cualidades, surgido de la yuxtaposición” (Eisenstein, S 2020, p 18). La importancia de esta definición no está sólo en función de la emoción o el poder estimulante del que habla Eisenstein en sus escritos, sino pensando en su poder como visualizador de continuidades y discontinuidades que nos permita comprender las variantes y matices casi infinitos que habilita el montaje. Queda en claro también, que esto antes descrito se enmarca dentro de un esfuerzo de prédica ideológica, que se fundamenta y transmite de manera intelectual, y que detrás del reconocimiento de este poder que tiene el montaje, se pretende encauzar un acto revolucionario. En lo que a este proyecto incumbe, lo importante es el poder de la combinación que reconoce Eisenstein, no para un uso concreto, sino para liberarlo en un terreno desconocido, que permita utilizarlo como máquina de exploración.

Por su parte el denominado Cine Estructural, se ha dedicado en parte a investigar el terreno de las continuidades de imágenes, no asociadas a un objetivo narrativo, o emocional en términos del relato clásico, sino buscando en lo profundo de la semiótica del dispositivo cinematográfico nuevas potencialidades que permitan comprender cómo se formulan y comunican las ideas audiovisuales.

Por esos motivos se hará referencia a algunas de las prácticas del cine estructural, asociándolas a posibles aplicaciones en la actualidad.

Cine estructural y montaje

Lo que interesa los fines de esta investigación respecto al cine estructural es su concepción fundante, que invierte la mirada habitual, no se interesa en lo relatos ni en el aspecto tradicional del cine, que suele promoverse como una suerte de espejo de la vida de las personas, el cine estructural ha revisado hacia adentro de los mecanismos técnicos, semánticos y estructurales que ponen en práctica esas clase de representaciones, en palabras del cineasta y pensador norteamericano Ken Jacobs, “ Observar las películas con una visión

al mismo tiempo de rayos X y de microscopio, descubriendo la parte secreta, y lo que generalmente es ignorado" (Pierson, James, Arthur, 2011).

De algún modo entender el origen de las capacidades de un sistema informático para construir montaje autónomamente, implica reconocer y actualizar el contexto tecnológico, su uso y su formato, descubriendo una parte secreta o menos conocida.

Otra particularidad del cine estructural es el acercamiento informal de sus principales exponentes, que provienen de disciplinas como la poesía, la pintura y la fotografía, y utilizan el medio cinematográfico, por lo general en formatos 8 y 16mm, como instrumento de pensamiento. Tal vez este origen diverso y cierta extranjeridad formal, es la que les permitió una mirada innovadora en sus comienzos, durante la década de 1960.

La inmediatez e informalidad en términos prácticos es evidente en muchas de las obras de este grupo de artistas, se trata muchas veces de ejercicios basados en palabras y secuencias de edición vertiginosas.

La curiosidad creativa que distingue a las obras asociadas al cine estructural, se puede decir, es siempre radical, quien nunca experimentó un cine fuera de los cánones formales institucionalizados, y se enfrenta a una obra de cine estructural, lo primero que piensa es que algo está fallando, prácticamente no hay nada en con común entre lo que el teórico estadounidense Noel Burch teórico definió como Modo de Representación Institucional[3] y el cine estructural, exceptuando tal vez la materialidad (Gidal, P., 1978).

A continuación serán descritas algunas obras de cine estructural, sus métodos de factura, y también referentes particulares sobre los que hace pie el presente trabajo, pero antes que eso se detalla una lista de paralelos de prácticas distintivas del cine estructural y las que esta investigación pretende actualizar.

[3] Son una serie de convenciones o normas estandarizadas que se adoptan en la década de 1910 codificando el lenguaje cinematográfico con el fin de que el mundo ficcional propuesto ofrezca coherencia interna, causalidad lineal, realismo psicológico y continuidad espacial y temporal.

Palabras/imágenes

El juego con palabras es recurrente en el cine estructural, el orden, la sonoridad de las mismas y su intercambio con las imágenes, es motivo destacado de muchas obras. Las palabras en forma de etiquetas que describen el contenido de videos en búsquedas en YouTube, permiten juegos de asociación similares formalmente, el cineasta estructural Hollis Frampton en uno de sus escritos habla del *¨Cine en la casa de las palabras¨*, haciendo referencia a que la estructura de yuxtaposición en el montaje cinematográfico está basada en la sintaxis con las que se estructuran las oraciones gramaticales (Jenkins, B. 2009).

Figura 1: *Título de imagen, tablas, cuadros, o gráficos. Autor y/o fuente de procedencia.*

Ilusiones de continuidad

En la factura estructural mediante técnicas de montaje, es frecuente encontrar secuencias de imágenes pasando a gran velocidad, obras de cineastas como Kurt Kren, Joyce Weiland, Paul Sharits, y el argentino Claudio Caldini, (que es pertinente sumar a la lista), entre otros, evidencian un profundo estudio de las bases perceptivas de la ilusión de movimiento, quebrando la lógica de secuencialidad cronológica que la historia del cine convencional atribuye a la obra del fotógrafo Eadweard Muybrdge.

En el caso de las películas estructurales, la continuidad se suele construir más que por secuencias de orden cronológico, por secuencias formales de consecución de un mismo elemento, desde distintos ángulos, o en distintos momentos del día, o de elementos que comparten características descriptivas. Esto último es otro ejemplo semejante a lo que ocurre cuando obtenemos muchos ejemplos de un mismo ítem, buscado a partir de una palabra clave en la web.

Como referencia directa de este tipo de secuencias es posible mencionar dos trabajos del cineasta austríaco Kurt Kren, que se enmarcan dentro de lo que él concibió como "*montaje seriado*" (Tscherkassky, P. 2012).

En el montaje seriado de Kren, la continuidad secuencial no es generada por la asociación cronológica natural entre fotogramas yuxtapuestos, recreando ilusión de movimiento y el paso del tiempo, sino por la continuidad de elementos en forma de serie, como cuando luego de una búsqueda se nos presentan elementos similares que responden a una misma categoría y descripción.

Las dos obras que mejor ilustran esta técnica de edición que Kren creaba *directo desde la cámara* son Árboles en otoño, de 1960, y Dos mil años de cine, de 1990. En la primera, de 1960, vemos una seguidilla incesante de árboles que sus ramas desprovistas de hojas, generando una epiléptica continuidad morfológica que se superpone en el tiempo (Tscherkassky, P. 2012).

El criterio de construcción del montaje, llevándolo a la contemporaneidad y a la búsqueda dentro de un repositorio como YouTube, podría ser el de limitar a un parámetro específico nuestra búsqueda, y ver sólo un elemento, por ejemplo solo fragmentos de videos donde aparecen tortas de cumpleaños, o cámaras de fotos.

Mediante lo que se ha expresado sobre las redes CNN esto sería posible y tal vez daría una pauta, o podría dejar visualizar, las actitudes miméticas que guardan millones de vídeos subidos a la red.

La otra obra que de Kren , Dos mil años de cine (**figura 6**), también ejemplo de montaje de series, justamente es una seguidilla de "personas apuntando con sus cámaras a un edificio popular en la ciudad de Nueva York, la catedral de Saint Stephan" (Tscherkassky, P. 2012).

Kren revela en este breve película la cantidad de personas que sacan la misma foto por día, repitiendo un exacto mismo gesto, en una conducta arquetípica, la de ser turista.

Figura 6: *Fotograma de la película Dos mil años de cine, Kurt Kren, 1990.*

Experiencias estroboscópicas, y largas escenas de ruido visual ilustran la concepción que los cineastas estructurales tienen sobre el material físico desde la que emerge la ficción. Películas de ¨flickeo¨, profunda alteración del fotograma, o extensas secuencias estroboscópicas, son otros de los elementos usuales en el cine estructural (Levi, P., 2012).

Dar con patrones visuales de vacío de información para revelar la cantidad de huecos y material ausente de figuración dentro del colosal repositorio de YouTube, es otro vector de creación que permite un paralelo entre las prácticas estructurales y la presente investigación.

Ritmo y matemática de edición

Otra característica reconocible de las películas estructurales, es el estudio rítmico y matemático del montaje, el desarrollo de técnicas seriales, y films que se despliegan en función de una fórmula matemática. Todo indica que se intenta *extraer otra clase de cosa, distinta a la emoción a partir del montaje, reconociendo su poder, pero buscando redireccionar el mismo hacia aspectos menos explorados y convencionales*, relativos al inconsciente de quien observa y a la fisiología de la percepción (Gidal, P., 1978).

En relación a esta característica está vinculada a la aplicación de parámetros de búsqueda que determinen un resultado. Todo es parametrizable en video digital, duración, color, brillo, e inclusive, gracias a las redes convolucionales anteriormente descritas, el reconocimiento uno por uno los items que hay en cada fotograma. Esto permitiría luego de pautar una serie de parámetros esperar que a partir de fragmentos de videos buscados en YouTube, la secuencia parametrizada se construya autónomamente, trayendo elementos novedosos aún para su creador.

Mediante los ejemplos de estrategias mencionadas, se intenta generar un paralelo de acciones posibles de llevar a cabo, proyectando qué tipos de operaciones se pueden realizar juntando fragmentos desde la inmensa y diversa cantidad de videos disponibles en YouTube. Considerando también, que lo que surgiría no tendría el aspecto de una narración tradicional, sino más bien de una exploración semejante a lo que se conoce como *minado de información*, en el terreno de la Big Data. En ese sentido es que las pautas de factura del cine estructural parecen adecuados como vectores de reflexión una vez transportados a otro contexto de uso, formato, (in)materialidad, y tecnología, es decir, a otro medio el de las plataformas de consumo audiovisual masivas, como YouTube.

Zorn´s lemma

Surgida de un proyecto previo llamado word pictures, Zorn's lemma, (**Figura 7**) de Hollis Frampton es una de las películas de cine estructural más representativas. Queda en evidencia en este film el cruce entre palabras e imágenes que tanto interesaba a Frampton, como poeta devenido cineasta. El montaje de Zorn´s lemma está basado en una proposición teórica que

habla de la lógica de relación entre conjuntos y sus subestructuras, las misma que se da entre palabras y oraciones (Jenkins, B. 2009).

Mediante este film, Frampton entra en diálogo con las teorías de montaje de Eisenstein, intentando reflexionar sobre el potencial del montaje.

> Eisenstein pretendía, más allá del montaje intelectual: la construcción de una máquina, muy parecida al cine, más eficiente que el lenguaje, que podría, entrando en competencia directa con el lenguaje, trascender su velocidad, abstracción, compacidad, democracia, ambigüedad, y poder, un proyecto, además, cuya última promesa era la constitución de una crítica externa del lenguaje mismo (Frampton, H. 1980, p 3).

Frampton integra un poema infantil que se utiliza para enseñar las letras del abecedario a los niños, pero en esta versión las palabras del mismo son recitadas en orden alfabético en lugar del original, que da sentido al texto.

Nuevamente, en este caso el montaje es parametrizado mediante reglas anteriores y lógicas previas a las imágenes, intentando liberar de la estructura otros sentidos e interpretaciones posibles, experimentando con la emoción y la percepción sin intervenir en el montaje para generar sentido de la manera narrativa institucionalizada.

Es en este sentido es valioso el ejemplo de Zorn´s lemma, porque en el campo de acción de un repositorio de video absolutamente parametrizado mediante el minucioso trabajo de las redes neuronales que habitan sistemas de recomendación, encontrando inspiración en la estrategia de este film, podríamos establecer cientos de estructuras que nos devuelvan ediciones del contenido innovadoras, semejantes a la que plantea este película.

Es posible comparar este caso con lo que sucede en la minería de datos, donde muchas veces se ponen en práctica estrategias de estudio de datos inusuales, para lograr encontrar vertientes y estructuras que describen información desconocida. La práctica que una estrategia así proyectada, podría ser minería de video, la misma será retomada en el capítulo 4.

Figura 7: *Fotogramas de la película Zorn´s Lemma, H. Frampton 1970)*

Metraje encontrado

El metraje encontrado, popularmente Found footage, es otra práctica cinematográfica y videográfica que consta fundamentalmente de una premisa, construir nuevas secuencias de imágenes y sonidos con fragmentos preexistentes. Entre los matices y las técnicas de distintos referentes del género es posible encontrar definiciones más detalladas de qué significa y qué técnicas supone el metraje encontrado, algunas de las mismas son propicias para entender porque esta práctica inspira también a este proyecto.

La idea de reciclar fragmentos que fueron ideados para otro fin e introducirlos en secuencias narrativas distantes en las que cobran otro sentido o utilidad, es según el investigador William Wees, una "práctica tan antigua como el cine, ya en películas de los primerísimos tiempos, en el siglo XIX, hay ejemplos de fragmentos reutilizados" (Wees, W. 1993, p 34).

Ahora bien, el found footage, como práctica contemporánea y en el sentido que pretendo recuperarla, no trata de tomar fragmentos cuando sean oportunos para el bien de un relato en el sentido clásico, el tipo de ejercicio que interesa a esta investigación es el que libera al poder de la yuxtaposición temporal, el sentido de partes de video de orígenes diversos y distantes.

La naturaleza y origen de los fragmentos audiovisuales previos, es descrita por el nombre que denomina la práctica como encontrados, lo que parece indicar que el mismo no tiene dueño, o que fue olvidado por desinterés, autores como Malcom Le Grise corrigen el nombre popular y hablan de metraje robado, dando a entender que encontrado, o no, alguien produjo ese material previamente, y muchas veces la autoría original queda oculta.

El concepto de reciclado que utiliza Weiss implica no desconocer que hay un autor, ni imaginar que la película cayó del cielo, parece ubicarse con lógica entre ambas definiciones, con cierta justeza (Wees, W. 1993).

No es el objetivo en este estudio hablar de categorías de reutilización audiovisual, sino más bien pensar cómo se actualizan las prácticas de montaje con material pre existente dentro de una plataforma con más de 500 siglos de video (por no decir cerca 500 millones de horas de video) disponibles. Es sabido que según estudios (**Figura 2**), casi el 90% de ese contenido es visto por menos de 100 personas, ni siquiera la gente que los creó lo ha vuelto a ver en muchos casos.

Figura 8: *Fotograma de la Introducción de Very Nice very Nice de 1961, Arthur Lipsett.*

Lo que abunda en esa marea audiovisual, contiene mucho contenido mimético y redundante, videos perdidos en su estado actual, que aún en su estatus de casi anonimato son cruciales como cápsula del tiempo de la humanidad

Los matices entre obras y artistas de este género, el del metraje reciclado, hay trabajos que se nutren de material significativo y popular, y otras cuya estrategia hurga en lo impopular y desechado, esa segunda vertiente es la que más interesa a esta investigación, porque se parece al contenido que menos visto en YouTube.

A continuación algunos casos puntuales de cine reciclado, útiles para repensar el montaje transportándolos a la escala, posibilidades tecnológicas del repositorio audiovisual de YouTube.

Very Nice, Very Nice

Como referencia directa a las películas hechas con lo que habitualmente es considerado es inútil o impopular vale la pena mencionar el caso de Very Nice, Very Nice (**figura 8**), del cineasta canadiense Arthur Lipsett.

Las características de Lipsett como realizador llamaron la atención desde sus primeras creaciones dentro del National Film Board de Canadá, sus primeras obras eran realizadas con recortes de fotos y textos, tomados de revistas y diarios, que cobraban vida dentro de sus collages animados. En Very Nice, Very Nice de 1961 la idea de recortes y fragmentos reutilizados fue llevada un poco más allá. Lipsett, que prácticamente no salía de su estudio en el NFB, solía por las madrugadas pasar por los estudios de sus colegas para husmear en los tachos de basura, y recoger fragmentos de películas que el resto había desechado. Esta actitud excéntrica no desentonaba con la extraña personalidad de Lipsett, lo que tal vez pocos imaginaban es que Lipsett estaba haciendo cine hecho de deshechos, es decir, montando todo ese material tirado a la basura, para hacer una película (Wees, W. 2007). El resultado del proceso fue Very Nice Very Nice, nominada al Oscar en 1962.

Lipsett unió todas esas partes inconexas y originalmente desechadas, encontrando denominadores comunes entre unas y otras, accesos a nuevos sentidos mediante la yuxtaposición, revelando mediante el montaje una visión crítica de la sociedad y del humor

de fines de la década de 1950, basándose en la discontinuidad de elementos, logró ilustrar un escenario contextual con una técnica innovadora y los recursos menos valorados (Does, A 2012).

Una práctica semejante es posible en la actualidad, ya que en YouTube, como se mencionó con anterioridad, hay millones de videos que apenas han sido vistos, equivalen a la basura en 16mm en los tachos de los colegas del Lipsett en el film board canadiense, estos videos de menos de 10 vistas, que casi nadie observó, pueden ser recuperados del anonimato, y montados, para formar parte de una visualización inédita de la sociedad que produce esa fenomenal cantidad de contenidos.

Variaciones sobre el montaje

Los afluentes históricos y estéticos mencionados en este capítulo intentan argumentar lo determinante de la continuidad como estructura en términos audiovisuales, no es ese el descubrimiento claro, el objetivo es señalar que estamos expuestos a estructuras de continuidad que de manera muy concreta y al mismo tiempo sigilosa, impactan en nuestra manera de asociar ideas, y que la naturaleza de ese impacto es semejante al tantas veces analizado poder del montaje dentro de los estudios sobre relatos cinematográficos.

El cuidado enhebrado de productos audiovisuales que consumimos on line, uno detrás de otro como se ha expuesto con anterioridad, no es casual, y esto queda en evidencia mediante la lógica de construcción de los sistemas de recomendación. El usufructo económico y la búsqueda de fidelidad de quien consume, son propósitos inamovibles en la programación de estos sistemas, que como hemos visto también cuentan con infinidad de sofisticadísimas herramientas que los hacen cada vez más eficientes.

Durante mediados del siglo XX, aún por fuera del poder magnético de la digitalidad, surgieron técnicas y géneros que combatieron la lógica de creación de sentido impuesta por las estructuras más populares dentro del ámbito audiovisual, las del cine comercial. Estas otras prácticas fueron en contra aún de los principios más elementales de construcción de sentido estandarizado, entre los cuales, el montaje cuenta con un rol principal.

En el caso de este proyecto, lo que se pretende es advertir y valorar la concepción y uso del montaje que estas otras prácticas alternativas aportaron, permitiendo imaginar también otras formas de montaje para las sofisticadísimas herramientas informáticas actuales, puestas en práctica por plataformas super populares, como YouTube. El objetivo es reconocer a través de ellas otras maneras de sacar provecho de la enorme cantidad de contenido audiovisual disponible, explotandolas como visualizadores sociales y juguetes estéticos, mediante el montaje automatizado.

Capítulo 3 Una imagen vale (sólo) un puñado de palabras

Mediante lo expresado en capítulos anteriores se presentó un contexto de convergencia, el de las tecnologías de base para la comprensión artificial de contenidos de videos, y sus orígenes y tareas distribuyendo contenido audiovisual, al que se añadieron luego concepciones sobre el montaje, intentando abarcar prácticas relacionadas con el trabajo con archivos y fórmulas estructurales de montaje. Advirtiendo ya entre ambas una suerte de punto en común, el que se expresa a diario en los sistemas de recomendación de videos.

En este capítulo se estudiarán herramientas disponibles en la actualidad que ejemplifican el montaje artificial, es decir la **c**onstrucción de secuencias de sentido yuxtaponiendo videos, mediante el uso de inteligencia artificial, para primero comprenderlos y luego ordenarlos.

Se suman a esta herramientas, otras semejantes que operan como puentes entre imágenes y palabras, para generar secuencias de video artificiales a partir de guiones literarios, o por el contrario, para extraer descripciones en *lenguaje natural*[4] de secuencias de video.

Las técnicas, Neural Edit, Neural Remake y ensamblajes automáticos entre palabras e imágenes

Una imagen vale un puñado de palabras

En el primer capítulo se mencionaron las características de las redes neuronales que interesaban a este proyecto, y a grandes rasgos, las habilidades de las mismas. Las redes convolucionales y su uso para reconocimiento de elementos dentro de imágenes fotos o dibujos, y las redes recurrentes, que son utilizadas en tareas que incluyen orden temporal y secuencias lógicas.

[4] Forma de lenguaje humano generada espontáneamente en un grupo de hablantes con propósito de comunicarse, en el caso de la IA, se utiliza para mencionar descripciones orales o escritas que sirven para dar instrucciones a máquinas de manera simple y casual, sin interponer un código específico o particular.

También sintéticamente se mencionó cómo es que esas redes pueden colaborar en la construcción automática de secuencias de montaje audiovisuales.

Volviendo atrás con la explicación, es importante recordar que una red neuronal convolucional, entrenada mediante un datasets como Cifar, es capaz de desglosar lo que reconoce en un fotograma en una serie de palabras. Esta traducción formal entre imágenes y palabras, que puede entregarnos con distinto grado de certeza unas 15 palabras por imagen, parece interponerse en un ámbito que hasta ahora dominaba la indeterminación, y que en el sentido común expresa la frase "una imagen vale más que mil palabras".

> Todos hemos oído decir que una imagen vale más que mil palabras. Pero si esta declaración es cierta ¿Por qué tiene que ser un dicho? Porque una imagen equivale a mil palabras sólo en circunstancias espaciales, y estas comúnmente incluyen un contexto de palabras dentro del cual se sitúa aquella (Ong,W 2021, p 41).

Este estudio del vínculo entre palabras e imágenes en el caso de cineastas como Eisentein contaba con fundamentos surgidos de un estudio detallado realizado sobre el funcionamiento enunciativo, expresivo y comunicacional de jeroglíficos, pictogramas e ideogramas, como antecedentes del montaje basado en imágenes. La síntesis argumental de estas capacidades ancestrales humanas para comunicarse mediante dibujos y sus relaciones condujo a Eisenstein a hablar de un "pensamiento sensible" (Eisenstein, S. 2020).

Para una red neuronal convolucional, la cantidad de cosas detectadas en una imagen, luego convertidas en palabras, es objetiva, y se puede manipular mediante filtrados el nivel de certeza y la cantidad de ítems detectados. El resultado de la detección realizado por una red neuronal convolucional, entre otras cosas, no depende de estados emocionales, a los que las personas son proclives, una fotografía puede emocionar hasta las lágrimas dependiendo a quién está allí fotografiado, y la asociación de palabras que puede extraer un ser humano de la experiencia de ese visionado, aunque genuina, sería de un alto grado de subjetividad.

Mediante entrenamiento, las redes sí son capaces de deducciones a posteriori. Por ejemplo, de elementos sustantivos detectados como nieve, ramas sin hojas y bufanda, es posible que deduzcan también que se trata de una imagen del invierno. Nuevamente, el objetivo aquí es permitir la proyección de lo posible con estos ejemplos, sin entrar en procesos de entrenamiento complejos mediante los cuales esto es posible.

El intercambio entre imágenes y palabras (etiquetas) que efectúan las redes neuronales convolucionales, parece ser bastante distinta a la que experimentan las personas ya que tienden a la objetividad. Esta característica distintiva, de la que podemos aparentemente aún diferenciamos, rige el orden de la gran mayoría de lo que se ve en la red, esas palabras detectadas, entre otras cosas, están detrás de la ubicación de cada video que sugiere ver YouTube, dando orden, tratando de lograr identificación en quien consume, apelando a un perfil que van quedando como rastro de cada acción en la red, para traccionar nuestras acciones y emociones. Todo eso se parece a una operación de montaje.

Neural Edit

Si es que los usos que se les da a las herramientas informáticas que se vienen describiendo a lo largo de este texto, tienen tanto en común con el montaje, en el sentido estricto en el que lo usan quienes trabajan en disciplinas audiovisuales, es decir, el proceso que consistente en unir trozos de película o video para crear distintas secuencias, generalmente siguiendo un guión o idea que termina en una producción final, lo que incluye una ardua tarea en la que se revisa el material audiovisual *crudo* y se seleccionan las partes que sirven a los fines narrativos originalmente planeados, entonces debería existir una herramienta basada en inteligencia artificial que haga exactamente eso. La hay, puesto que esta descripción genérica del montaje, que puede encontrarse en un manual de cine inicial, y se parece bastante a un algoritmo, ya que parece describir una serie de pasos, un proceso y un resultado.

“Un algoritmo es un conjunto de pasos que puede seguir con lápiz y papel, y puede estar seguro de que esta descripción aparentemente fácil está cerca de los utilizados por los matemáticos y la computadora científicos” (Lauridas, P. 2020, p 4).

La técnica que se describirá a continuación, informalmente denominada Neural Edit hace eso, cumple con un proceso de edición, ni necesariamente bien, ni mal, sólo lo hace. Se toman aquí como referencia softwares en estado *beta*, es decir en etapa de desarrollo, no consolidados como productos comerciales, la función de estos softwares es editar video mediante redes neuronales, nuevamente aclarando que se trata de una expresión algo vaga, porque no describe qué tipo de edición llevan a cabo. El término Neural Edit se utilizará de manera preliminar para describir los procesos y resultados que se pueden lograr mediante

estos tipos de aplicaciones. Es importante, tratándose de un algoritmo, comprender los procesos detrás del funcionamiento de un sistema de edición de video, con asistencia de redes neuronales.

Lo primero que hay que considerar es el material audiovisual, es decir los videos capturados y que copiosamente se acumulan en los dispositivos móviles de la contemporaneidad, obligando a menudo a expandir con medios de almacenamiento de mayor capacidad las computadoras, teléfonos móviles y cuentas en la nube para poder almacenarlos. Estos videos capturados, ya no forman solo parte de una ceremonia o eventos importantes de la vida, sino que registran acontecimientos de naturaleza cotidiana, prácticamente en continuado. Quedó en el pasado la tiranía de los casetes, las cintas de fílmico y fotografía, y las memorias minúsculas, junto con ellas desapareció la restricción física que limitaba el uso de cámaras de video en todo momento, para grabar videos casi sin limitación.

Desde hace más de una década, aparte de agolparse en teléfonos móviles, un porcentaje de ese registro cotidiano es eventualmente es subido a YouTube, de hecho, la consigna original de 2005 parecía una instrucción en ese sentido *Broadcast yourself*, transmite a ti mismo (Burgess, J. y Green, J., 2018).

Todo lo que hacen las personas a diario, por más intrascendente que sea tiene reservado su lugar, las plataformas no advierten sobre tomarse unos minutos y pensar si es importante, útil o valioso para alguien o algo, esto que está por subir y socializar.

Es sensato decir que existe un problema contemporáneo, ya conocido, el de la cantidad de información que producidas, que probablemente nunca se volverá a ver, y que tal vez desaparezca en la siguiente migración de tecnología, 5 o 6 años más adelante. En este contexto parece pertinente describir una herramienta que hace uso de la técnica neural edit, llamada FLO. Pero antes de entrar en detalles técnicos es importante ver cómo esta herramienta creada para teléfonos móviles, se describe a sí misma, en principio como publicidad muestra una captura de pantalla en la que Flo oyendo a su propietario toma nota, Hazme un video de mi gato (**Figura 9**).

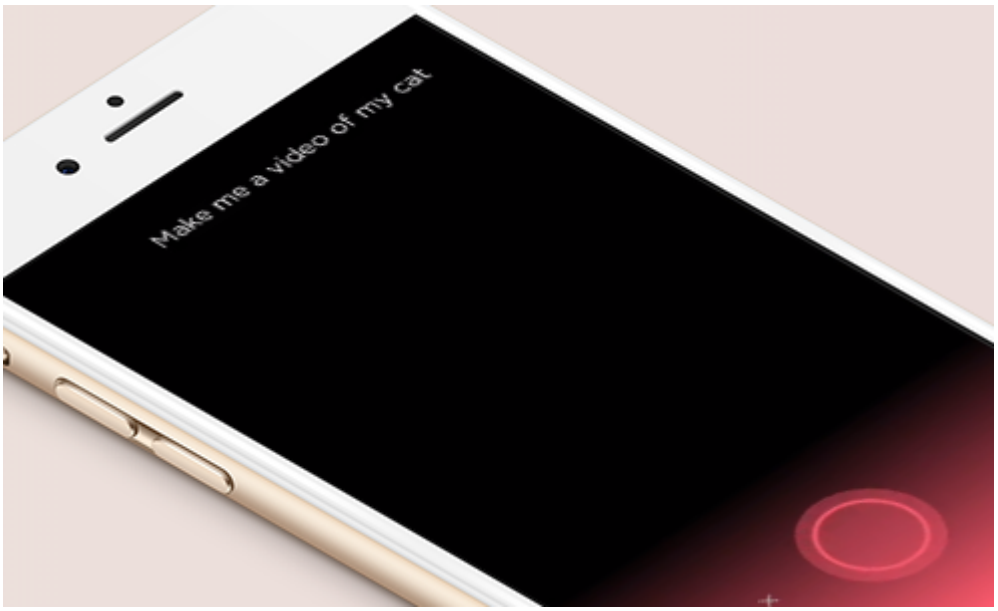

Figura 9: *Imagen publicitaria de la aplicación FLO.*

A todos les encanta grabar videos, pero no saben qué hacer con ellos. Editar videos es difícil, requiere mucho tiempo y es un trabajo que requiere habilidad. Para eso NexGear Technology ha creado Flo, una cámara inteligente y una aplicación de creación de películas que edita automáticamente secuencias de video sin procesar, y las convierte en películas cinematográficas cortas, utilizando aprendizaje profundo e inteligencia artificial. FLO combina el aprendizaje automático, la visión por computadora y el procesamiento del lenguaje natural para reunir los momentos interesantes de sus videos en una película cinematográfica corta. Flo reconoce objetos y entiende escenas en el video como un ser humano.

Sundar Pichai de Google I/O y, más recientemente, Tim Cook de WWDC destacaron la importancia de ejecutar redes neuronales localmente en los dispositivos.

Flo usa redes neuronales convolucionales (CNN) para comprender un video y luego memoria a corto plazo (LSTM) para describir en forma de texto legible por humanos. El modelo de IA ha sido entrenado en ~100.000 imágenes descritas por humanos. Actualmente, el sistema es bueno para describir las escenas correctamente, pero puede mejorar aún más si se entrena con más imágenes.

Flo Camera utiliza el aprendizaje profundo para comprender y describir una escena en tiempo real cuando el usuario está grabando un video. Flo va más allá de

capturar. Entiende lo que se está capturando y te describe la escena en tiempo real. Además, los videos capturados con la cámara de Flo no tardan en procesarse.

Flo Voice Assistant es una característica revolucionaria que le permite crear ediciones de video simplemente diciendo lo que quiere en su video. Simplemente solicite una historia por ubicación, período de tiempo, etiquetas o todo lo anterior y Flo le responderá como su propio asistente de creación de videos, creando la historia en video que elija, por ejemplo, "Hazme una historia en video de mi gato". o "Haz un video de mi viaje de fin de semana a la playa". Es como tener un amigo editor que puede hacer videos a pedido en segundos. Flo escanea los videos de su teléfono y crea una historia en video lista para compartir de su viaje a la playa del último fin de semana sincronizada correctamente con música y filtro cinematográfico. (Texto publicitario de la aplicación FLO, insidebigdata.com 10/20).

Desglosando el texto original, comparándolo con lo que se ha mencionado con anterioridad respecto de cómo funcionan las redes neuronales y cuán eficientes son, es posible valernos de una pauta para reconocer la veracidad de lo que propone esta aplicación. Al mismo tiempo es sabido que muchas herramientas que pone en práctica FLO, fiables o no, son las mismas que utiliza YouTube para recomendar qué ver luego, que videos rechazar por contenido inapropiado y bajar por infracciones de derechos de autoría.

La primera parte del texto describe el universo hiper abundante de la producción de videos hogareños, y plantea que la edición es una tarea compleja, que no obstante su complejidad puede ser resuelta gracias a aprendizaje profundo, es decir, redes neuronales, IA, etc. La primera tarea es discriminar que parte de los videos no sirven y pueden ser desechados y que partes son valiosas, tienen carácter ¨cinematográfico¨, o son ¨interesantes¨.

Al margen de la vaguedad de estas definiciones, lo que realiza FLO cuando proclama que *comprende las escenas*, es detectar errores comunes que pueden ser percibidos mediante parámetros visuales, es decir capturas fuera de foco o movidas, sub exposiciones o sobre exposiciones de luz, fragmentos de video abstractos en lo que no se observa ningún objeto reconocible, para luego desecharlos. Y por el contrario, detectar cosas básicas como una sonrisa, dos copas de champaña, la playa, una sombrilla o una torta de cumpleaños, conducen a la aplicación a valorar esa escena para el montaje definitivo.

La aplicación funda su eficiencia actual en el hecho de haber sido entrenada con 100000 imágenes, sin explicitar si son estáticas o en movimiento, aunque todo parece indicar que son imágenes estáticas de alguno de los datasets populares que fueron referenciados anteriormente.

La aplicación, gracias a sus consiguientes permisos, es todo el tiempo consciente, mediante el uso de redes CNN, como las detalladas en el primer capítulo, de lo que captura la cámara del móvil, y va generando un catálogo latente de ítems que definen lo que hemos grabado, para que en el momento de que se solicite, mediante comando de voz o escritura, este el material disponible y catalogado sin que sea necesaria una nueva revisión.

Saber que hace FLO con toda esa información que recopila, con quien la comparte y qué beneficio obtiene de eso es un tema importante, y que tal vez justifica la existencia de la aplicación, más que su supuesta utilidad, para la que pone en práctica sofisticadas tecnologías para una tarea elemental, las virtudes que se le atribuyen al proceso, parecen a priori exageradas.

La tarea de edición en forma de algoritmo semejante al utilizado por aplicaciones como FLO se puede resumir en los siguientes pasos:

Paso 1: Analizar mediante una red CNN todo el contenido capturado con un teléfono móvil, lo que resulta en un catálogo de ítems detectados en cada captura. Sumado esto a los datos cronológicos y de geolocalización con los que automáticamente ya cuenta el registro.

Paso 2: Cuando hay una palabra clave o dos en una oración mediante la cual el usuarios solicitan el montaje de un video, como por ejemplo, gato y casa, seleccionar una serie de videos que contengan esos ítems.

Si el usuario al ejecutar el pedido al programa para que edite el video utiliza palabras como felicidad o descanso en la descripción, estos valores serán utilizados también para seleccionar un tema musical con derechos de reproducción liberados que acompañe el montaje, extraído de la base de datos de la aplicación. Otras características descriptivas también pueden ser utilizadas para el ritmo y el tipo de cortes de la edición final del video, que por definición nunca dura más de 1 minuto y 30 segundos.

Paso 3: Una vez obtenidos los video candidatos para el corte final (fragmentos de video en el móvil, que contengan al gato y la casa preferentemente, o ambos por separado), cortar los momentos de mayor nitidez y claridad, sin fueras de foco o movimientos bruscos, sub exposiciones o sobre exposiciones. Luego ordenarlos de algún modo durante 1:30 aproximadamente. Aquí entran en funcionamiento las redes RNN, que aplican algún orden sensato a la secuencia, surgido de oraciones creadas a partir de las palabras extraídas (**Figura 10**).

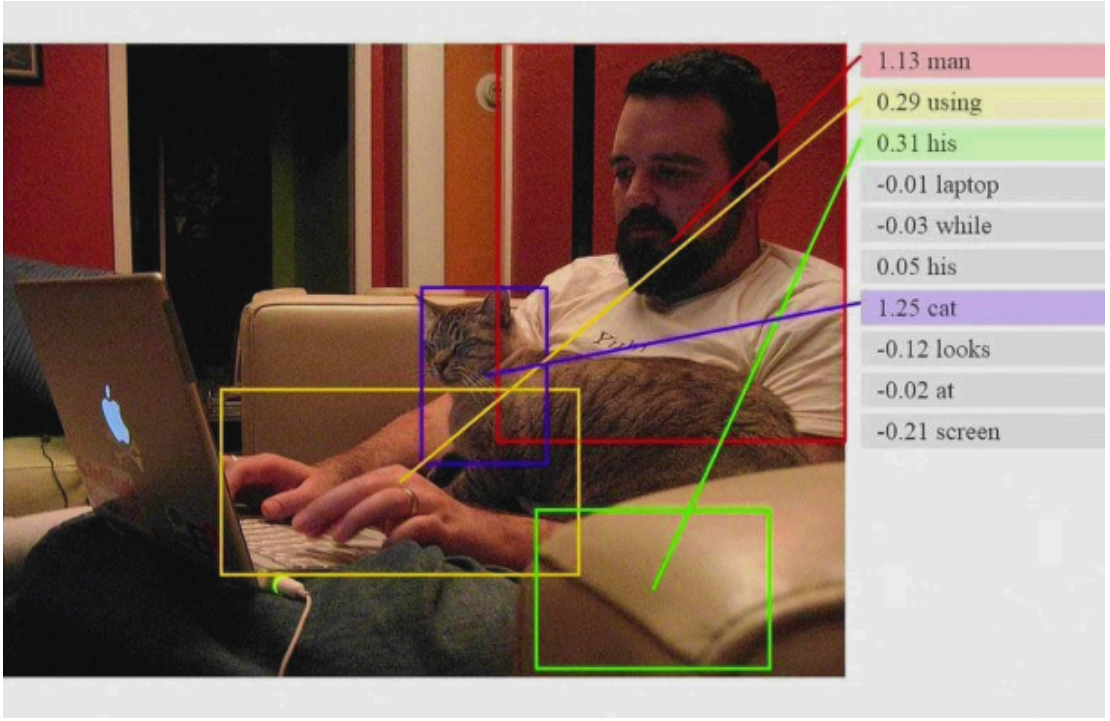

Figura 10: *Ejemplo de extracción de palabras con redes CNN y ordenamiento en forma de oración con red RNN.*

Al observar el funcionamiento de la aplicación muchas veces, se percibe que el valor principal que toma en cuenta es el ritmo de la música, sobre la que realiza todos los cortes en el video, cabe decir que esto no parece hacer un uso profundo de la redes RNN. No obstante se percibe secuencias que conservan algún tipo de lógica temporal, como por ejemplo que los videos capturados de noche tienden a estar al final del montaje, y los luminosos promediando la edición, lo que da cierta sensación de ciclo de día, y paso del tiempo. En este caso sí parece haber uso de habilidades de redes RNN, es decir que elaboran secuencias temporales de sentido.

En definitiva, ya conociendo con relativa profundidad que hace una aplicación de Neural Edit, es posible decir, al menos preliminarmente, que el rasgo técnico de la imágenes capturadas se prioriza al momento de elegir el material y que la expresividad comunicativa del montaje quede relegada más un bien a una suerte de *tema* o *template* genérico, plácido, melancólico o festivo, determinado por la música y las velocidad del los cortes.

La tarea de las redes RNN apenas se detecta, como se observa en ejemplos del capítulo I, sería el uso de esas redes el que aporte mayor sofisticación al montaje.

FLO claramente se soporta en el hecho de que sus usuarios no pretenden mucho como resultado, solo que corte lo que no sirve, que ponga música y que se ¨entienda¨ que se trata de unas vacaciones, o cumpleaños o una reunión entre amigos. La expresión ¨calidad cinematográfica¨, que expresa el texto de promoción de la app, es arbitraria y se refiere más bien a un compendio musicalizado, sin pretensiones singulares.

Gliacloud

Se describe aquí el caso de otra aplicación basada en IA para editar video, nuevamente partiendo de lo que la difusión del producto autoproclama

> *Glia Studio es una plataforma innovadora que permite crear videos en minutos con tecnología de* ***automatización de video****. Glia Studio brinda una experiencia de creación de videos más inteligente para editores, especialistas en marketing y blogueros (Texto publicitario de la aplicación Galicloud, gliacloud.com)*.

Aquí entra la descripción automatización de video (video automatization) que es aún más vaga que neural edit, no obstante eso, el perfil de lo que esta aplicación produce no menciona lo cinematográfico, sino que modestamente habla de una experiencia de ayuda para entre otras cosas blogueros o personas interesadas en publicación web, lo que reduce las expectativas de su potencial considerablemente.

> *Simplifique el proceso de producción de video con la tecnología de* ***texto a video,*** *produzca automáticamente videos personalizados de formato corto a escala. Utilice*

*recursos multimedia y **plantillas** de alta calidad para crear videos sorprendentes (Texto publicitario de la aplicación Galicloud, gliacloud.com).*

Una clave en esta segunda descripción es el uso de texto a vídeo, que ilustra el uso de redes CNN, es decir redes capaces de convertir imágenes, o fotogramas en palabras, es decir etiquetas (tags). Como se ha mencionado anteriormente, tanto en los procesos de entrenamiento, en la generación de gigantescos datasets y como el resultado de reconocimiento de elementos en imágenes, esta idea y vuelta entre palabras-imágenes-palabras parece estar siempre presente, siendo parte esencial de los mecanismos de este tipo. Para un sistema basado en redes neuronales convolucionales interpretar exitosamente una imagen es poder convertirla en palabras, palabras que por otro lado son esenciales para generar un orden lógico en secuencia, que luego volverá a convertirse en imágenes, ahora ¨montadas¨ en función de su orden y sentido literal, gracias a la tarea de redes neuronales recurrentes.

La herramienta también destaca el uso de templates[5], genéricos dentro de los que define su función, la de contribuir a la productividad, lo que permite concluir que la utilidad que se intenta extraer de la redes neuronales y del uso de Inteligencia Artificial en general está asociado a la eficiencia, a reducir costos y tiempos de producción, sin priorizar lo singular o lo sensible.

Es sabido que la lógica de acción que debe darse, para que un algoritmo que implementa redes neuronales pueda editar un video de manera coherente, es que el objetivo del mismo, es decir su coherencia, sea acotada al mínimo número de probabilidades. Es decir, esta clase de herramientas no exploran todo el poder ni la utilidad de la IA, sino que la ponen en práctica en tareas simples, basadas en perfiles y plantillas muy definidas. La función de la IA es entonces básicamente definir que tipo de plantilla corresponde el video que se quiere montar, sin explotar al parecer, y mucho menos liberar, el poder expresivo de la tecnología de la que se jacta.

[5] Base prediseñada de estructuración de los contenidos de una aplicación o una página web para que se ajusten a unos parámetros estandarizados.

Ejemplos contextuales

Abundan los ejemplos de aplicaciones para edición de video que utilizan redes CNN, sin necesariamente advertirlo como su principal característica, la mayoría de ellas mantiene la tendencia de las de lo comentado hasta ahora en el caso de FLO y GLIACLOUD, es decir son buenas para trabajo más bien genéricos basados en plantillas predeterminadas.

Es común que las personas ya conozcan aplicaciones semejantes gracias al servicio que proveen Google Photos y Facebook, a partir de eventualmente generar contenido automático al usuario para que revise los mejores momentos de los últimos meses, o ilustre la amistad que mantiene con alguien, generando video de menos de un minuto de duración que compilan contenidos asociados a un lugar, una época del año, o una persona.

Es importante destacar que la interacción que cada usuario tiene con esas fotos y videos en las redes y en su propio móvil, es también un factor determinante para el montaje final de la secuencia. Pero no obstante esos valores, la cantidad de veces que ese contenido es compartido con alguien, o es vuelto a ver, también se utiliza para hilar esos compilados de momentos.

Vale mencionar también que no solo herramientas para productividad y marketing genéricos utilizan redes en alguna escala, o para algún propósito específico. Otras herramientas, de uso profesional, como el software para edición de video Premier de Adobe, rápidamente va incorporando añadidos que utilizan redes neuronales convolucionales y de otros tipos, para diversas tareas, por ejemplo, de efectos visuales, generación de inter cuadros al momento de ralentizar videos y muy próximamente para ordenar el contenido del material crudo, previo a la edición manual llevada a cabo por el usuario.

Palabras a imágenes, imágenes a palabras

Existen y proliferan a principios de la década de 2020 proyectos que utilizan transferencias entre textos e imágenes, haciendo uso de las redes CNN. Extrayendo palabras de imágenes con distintos propósitos, pero también hay casos que hacen la operación a la inversa, o que juegan a manipular esta transposición, o transporte de ideas, del mundo de la palabra, al

mundo de la visual. Dentro del emergente universo de esos proyectos se mencionan dos, sólo con el objetivo de reforzar la idea de potenciales proyectos detrás de estas herramientas.

Wordseye

El primero de estos ejemplos es Wordseye, cuya traducción al español podría ser, ojos de palabras, que insta a sus usuarios a que *escriban* una imagen

Wordseye es un sitio web que permite que sus usuarios puedan crear escenas, en forma de renders 3D luego de describirlas en una breve oración. En este caso no se menciona Wordseye para describir sus cualidades técnicas, sino para habilitar la comprensión contextual de proyectos que exploran informalmente y con humor, la frontera entre la descripción escrita y la representación pictórica (**Figura 11**).

Figura 11: *Ejemplo de funcionamiento de Wordeye (wordeye.com).*

Text2image

Otro caso valioso para el análisis, en el que si intervienen procesos de entrenamiento de redes neuronales con fotografías, es el de la técnica de texto a imagen popular en la web. Por ejemplo el trabajo de Cristobal Valenzuela, el artista de origen chileno que desarrolla y explora expresivamente el uso de tecnologías ligadas a la IA (**Figuras 11, 12, 13**).

También en el caso de Text2image[6], se cuenta con un espacio para redactar un texto, que será luego de escrito, presentado a la derecha de la pantalla, por una imagen. A diferencia de Wordseye, lo que se conforma aquí luego de escribir la frase o la palabra, es una imagen que pretende ser fotorrealista, y está basada en representaciones generadas a partir de asociaciones entre palabras y fotografías, que involucran otros procesos y tipos de redes neuronales distintas a las que visitamos en este texto, pero que también forman parte de este emergente contexto de creación relacionados con la IA, que parece por momentos traer consigo modificaciones fundamentales en nuestra manera de comprender el vínculo entre expresiones escritas y representaciones visuales.

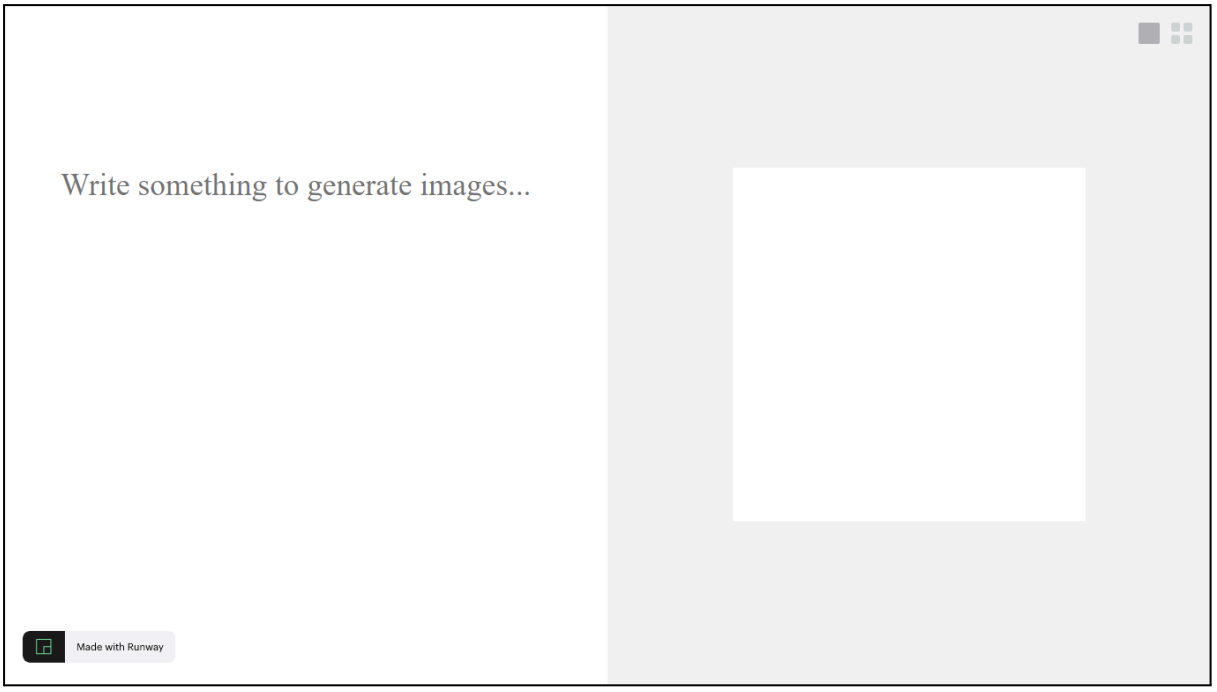

Figura 11

[6] Sistema que mediante el uso de Inteligencia Artificial permite generar imágenes a partir de una descripción textual en lenguaje natural.

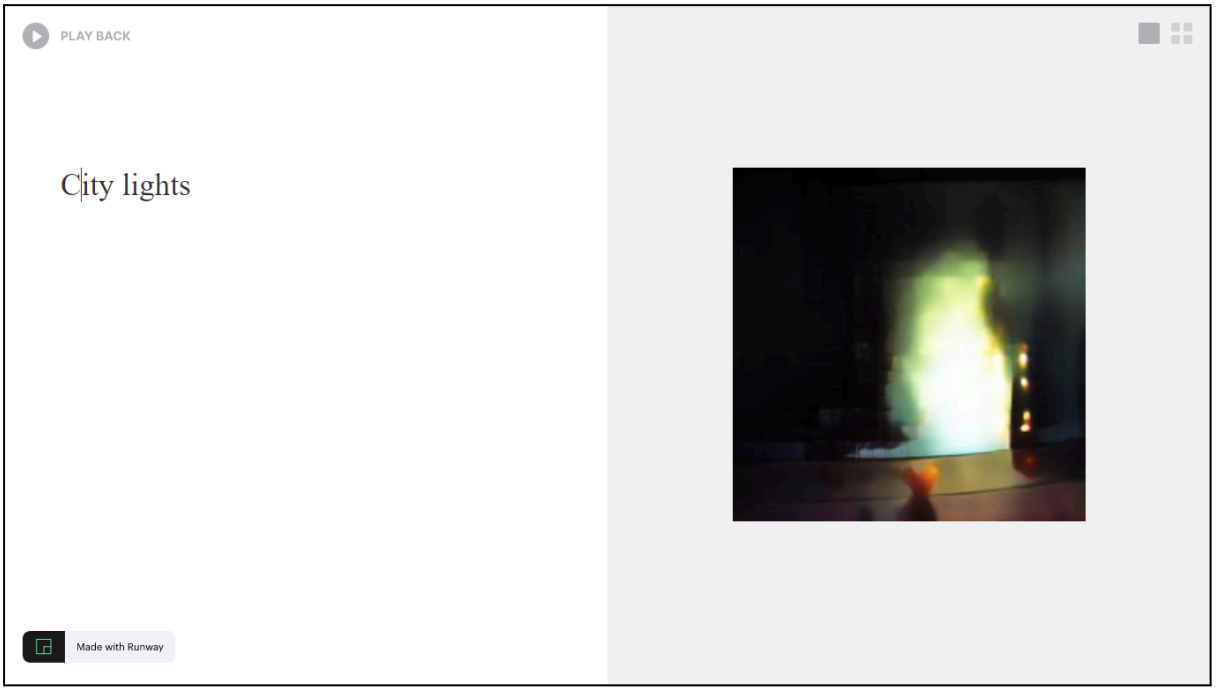

Figura 12

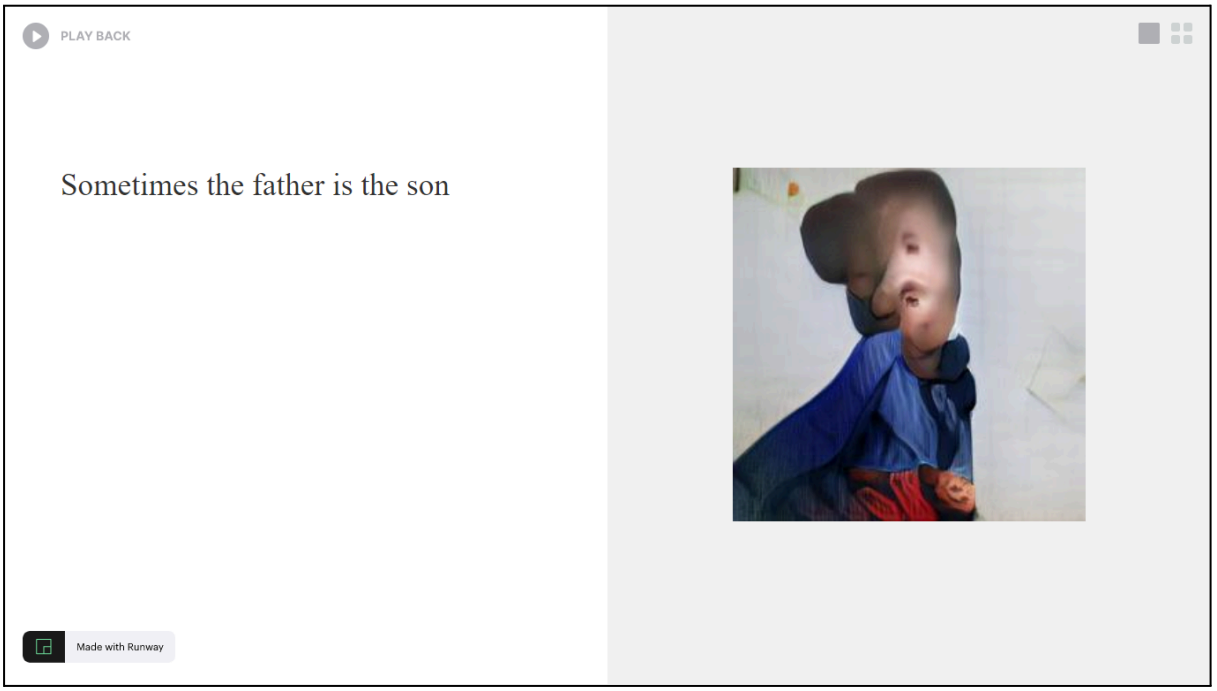

Figura 13

Hay un caso remoto pero valioso que es importante mencionar como antecedente de estas posibilidades tecnológicas, cuyo sentido y utilidad está aún en proceso. Es el del artista yugoslavo Monny de Boully que hace aproximadamente 100 años hablaba del *cine para la mente*, películas que no eran filmadas sino escritas, que requerían ser leídas para que cada quien las imagine con los ojos cerrados. La imagen no aparece en el sentido habitual, representada objetivamente, sino que se manifiesta en la mente de cada participante de manera indeterminada y subjetiva (Levi, P. 2012).

La naturaleza de uno de esos relatos surreales aquí a continuación.

> Hypnison se traga la luna, se mete las nubes en el bolsillo, levanta el brazo y transforma el paisaje arbitrario en otros decorados ilusorios: un interminable desierto púrpura que rodea una colina. Hay una tienda negra al pie de la colina; en el oriente, un sol púrpura y cuadrado está saliendo. Los ventiladores (en el cine) soplan un viento sahariano caliente. Agua y dátiles frente a los espectadores de la película. La banda toca melodías orientales. Hypnison encuentra el sol cuadrado bastante insípido. Lo cambia por otro: un elipsoide en llamas. Pero este sol tampoco le gusta, y enfadado, con la velocidad del rayo, lo patea como si fuera un balón de rugby. El sol vuela sobre la Vía Láctea, y en algún lugar, muy lejos en el Cosmos, se rompe en pedazos.Tambores africanos suenan con toda su fuerza. De repente, las cosas pierden el aspecto de la perspectiva. Las hipérbolas, las espirales, las elipses y las coordenadas bailan delante, detrás y en la pantalla: todo el mundo intuye que la explosión del Sol tiene algo que ver con el hiperespacio, la cuarta o la quinta dimensión (Levi, P., 2012, p. 46).

Este fragmento de ¨La Máquina de Vivir¨ escrito en 1923 por Monny De Boully, aparte de sumamente enigmático, permite ver como expresiones marginales a las prácticas cinematográficas convencionales, pueden servir como herramientas de exploración y comprensión de nuevas tecnologías.

Neural remake (reversión neuronal)

En este capítulo se visitaron diversos ejemplos de sistemas de de edición automatizada a partir de material audiovisual *crudo*, que utiliza redes convolucionales y recurrentes, para desestimar contenido no útil, y editar videos de manera automática.También se detallaron casos de sistemas que generan interacción entre descripciones en palabras de escenas o situaciones y la generación de las mismas de manera visual, reconociendo que el vínculo entre palabras e imágenes, o más bien la frontera que las separa, es traspasada por las redes neuronales que venimos estudiando, de manera informal, construyendo lo que parece ser un nuevo tipo de vínculo.

Aún dentro de ese contexto, pero tomando distancia de la utilidad productiva del Neural Edit, se continúa aquí con la descripción de otra técnica de edición automatizada, el Neural Remake, bautizado así por el artista alemán Mario Klingemann.

El Neural Remake es una técnica que explora el potencial expresivo de las capacidades de las redes CNN, convirtiendo un fragmento audiovisual, con sus cortes, escenas, planos, ritmos, etc, en un guión. Gracias a todo lo que se ha detallado quizás sea simple imaginar cómo funciona esta técnica, donde el uso de redes CNN y la capacidad de extraer elementos que se encuentran de una imagen y convertirlos en palabras es esencial, no obstante eso aquí un ejemplo.

Se toma un video de unos 3 minutos de duración, por ejemplo un videoclip, en el que se suceden planos y acciones a gran velocidad, supongamos que durante los 3 minutos hay 150 planos distintos, que ocurren en diversas locaciones. Ese material será la materia prima.

El paso siguiente es tomar palabras y descripciones de cada plano y escena del videoclip, para lo que naturalmente podemos utilizar una red CNN. Será necesario programar que la red extraiga unos 5 fotogramas por segundo del videoclip, y que de cada fotograma, se extraiga una serie de palabras que desglosen lo que hay en cada plano.

Al finalizar esta operación se obtendría una serie de palabras asociadas a cada momento del video, a cada segundo dividido en 5 partes, estas palabras sueltas, son las etiquetas que extrajo la red CNN. Luego de tener todas esas palabras asociadas a un momento del video

podemos utilizar una red RNN, que nos permita convertir esas palabras sueltas en oraciones con sentido, es decir semánticamente correctas.

En esta instancia del funcionamiento de nuestro algoritmo tenemos una serie de acciones detalladas en palabras y tipos de planos descritos visualmente y organizados en el tiempo. Lo que equivale a decir que tenemos un guión.

Recapitulando entonces, lo que habitualmente se conoce como desglose de un guión, es realizado automáticamente por una red neuronal convolucional a la que se le presenta un video, en este ejemplo eligiendo un videoclip de 3 minutos de duración, semejante a los que utiliza en creador de la técnica, Mario Kilngemann. Luego el texto desconectado que extrae la red CNN de cada momento del video cobra sentido gracias al orden que le asigna una red RNN.

En esta serie de pasos se invierte la naturaleza de la creación cinematográfica con más de 100 años de historia, es decir primero está la imagen y luego el guión.

Lo que continúa es poner en práctica ese guión de manera automatizada, buscando, gracias a redes CNN dentro de una base de datos de miles de videos que tengamos preparada, escenas que sean comparables en términos descriptivos con las de nuestro guión.

El siguiente paso de nuestra secuencia de órdenes sería entonces montar este nuevo video, reemplazando automáticamente cada plano en el video original por otros de equiparables características tomado del material de una base de datos, basándonos en la misma descripción segundo a segundo (**Figura 14**).

El proceso completo, el algoritmo, con alteraciones y ajustes corresponde a la técnica denominada Neural Remake.

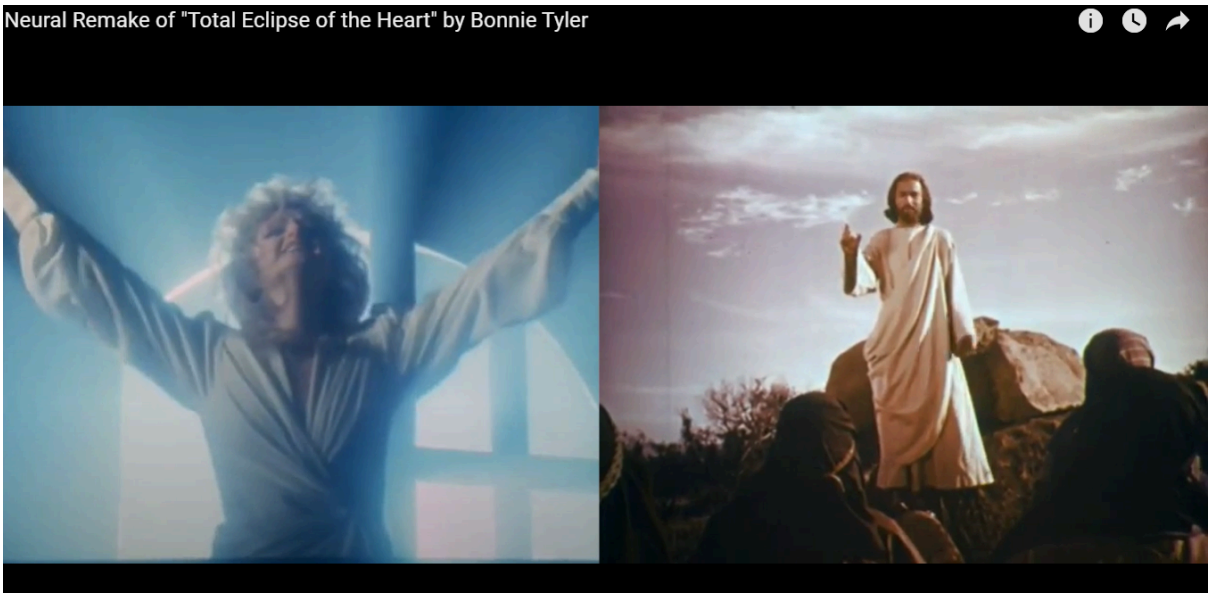

Figura 14: *Neural remake de Total eclipse of the heart del artista Mario Klingemann.*

El neural remake de Klingemann es un ejemplo valioso, por su originalidad y porque permite a las máquinas explorar un costado creativo asociado con el error y lo indeterminado, ya que esta técnica de montaje subordina a la máquina a un propósito artístico exploratorio.

> El verdadero perfeccionamiento de las máquinas, aquel del cual se puede decir que eleva el grado de tecnicidad, corresponde no a un acrecentamiento del automatismo, sino, por el contrario, al hecho de que el funcionamiento de una máquina preserve un cierto margen de indeterminación (Simondon, G., 2017, p. 33).

Área en proliferación

El variopinto de técnicas autorales e industriales, aplicaciones y experiencias artísticas vistas en esta capítulo, da cuenta de un espacio en vertiginosa expansión, cuyo motor fundamental es la comprensión artificial de videos y sus consecuentes y aún inexploradas utilidades, una de las cuales podría ser montar secuencias de sentido automáticamente.

Google, con su colaboratorio de código on line, el AI cloud y sus sistemas de visión artificial basados en modelos entrenados con colosales cantidades de datos extraídos de la vida de las

personas en la web, tiene quizás la mayor cantidad de herramientas disponibles para cualquier tarea que se proponga realizar de manera automatizada. Dentro de sus mismas plataformas de promociona sistemas de búsqueda de video mediante reconocimiento de elementos y contenido, e inclusive herramientas que permiten editar vídeos mediante técnicas descriptas como *highlight extraction*, extracción de momentos llamativos, para editar videos sólo utilizando las partes más atractivas, *most engaging moments*.

Amazon también cuenta con plataformas semejantes para trabajo colaborativo, para que usuarios de todo el mundo prueben herramientas como Amazon Rekognition, mientras aportan a la mejora de los algoritmos y encuentran nuevas utilidades y aplicaciones.

El detalle de estas plataformas es inabarcable para la escala de información que se propone estudiar esta investigación, pero de todos modos es importante reconocer estos espacios como el caldo de cultivo de aplicaciones de montaje de video automatizado en el futuro próximo.

Una imagen vale un puñado de palabras II

En este capítulo se enumeraron algunas de las técnicas específicas que se utilizan para montar videos con ayuda de inteligencia artificial en el presente, resultado del vínculo entre la Inteligencia Artificial, y el montaje, que permiten visualizar la zona aún experimental donde proliferan aplicaciones semejantes de uso expresivo, no comercial.

El estado actual de la cuestión, y la ebullición de proyectos que rondan las tecnologías centrales que se utilizan para comprender videos artificialmente, permite vislumbrar que surgirán de aquí aplicaciones de uso popular. Al mismo tiempo, la novedad que aportan estas herramientas, permite imaginar que igual de novedosas serán algunas de las aplicaciones que se les extraerán, tal vez modificando la concepción de algunas posibilidades, por ejemplo, las del montaje.

Capítulo 4 La región sin nosotros

Introducción al capítulo 4, La región sin nosotros

En este capítulo se discute el estado de la convergencia que se ha mencionado desde el comienzo de este texto, la del montaje y la informática, encarnada esta última en la Inteligencia artificial. Se realizan también consideraciones respecto a la pregunta inicial sobre los potenciales aportes de la IA al montaje, yuxtaposción y creación de sentido, editando videos. Se aportan a su vez ideas y aplicaciones surgidas de la información reunida para esta investigación.

La región sin nosotros

A lo largo de esta investigación se han descrito herramientas y tecnologías que han surgido dentro del contexto de producción de información que habitamos en la actualidad, contexto particularmente singular por el uso de herramientas utilizadas para moderar y distribuir gigantescas cantidades de información audiovisual. A lo largo de este texto se ha ejemplificado entorno a la sofisticación de los procesos de aprendizaje automatizado mediante los cuales han sido entrenadas redes neuronales,que motorizan la web y sus algoritmos de automatización, volviéndose omnipresentes y al mismo tiempo invisibles para gran parte del público consumidor.

Dentro de los repositorios de consumo audiovisual como YouTube hay algoritmos que utilizan Inteligencia Artificial, y que gracias a la misma cada usuario recibe a diario sugerencias para ver videos, o volver a ver lo que ya vieron hace un año. Al mismo tiempo, la naturaleza de ese algoritmo es ignorada por el público usuario, sólo es posible ver los resultados de su accionar. Apenas es posible modificar sus variables mediante la parametrización acotada y digitada de antemano que permiten las búsquedas. La interacción que se mantiene con el real volumen de los datos disponible es limitada.

De alguna manera, el público consumidor es parte de lo que se denomina *data compression*, esta compresión de datos significa que los intereses en las búsquedas de videos, por más

singulares que sean, tienden a ser generalizados dentro de lo que es más habitual, simplificando la tarea de distribución y moderación de la información para las compañías propietarias del contenido, en este caso audiovisual.

Al menos 200 mil años de video disponibles, un reservorio de la humanidad, del cual menos del 10% consigue ser visto por más de 1000 personas, el resto es prácticamente anónimo, pero representa una *imagen* de bastante detalle de la humanidad de principios del siglo XXI.

La lógica de moderación y recomendación que estructura todo ese contenido funciona como un sistema de montaje automatizado, crea secuencias y une un video con otro, hilvanando un maquínico orden de sentido. Ese sentido se sofistica aprendiendo a diario de lo que hace cada usuario, al mismo tiempo que introduce cambios e induce a ciertas conductas de consumo y de todo ese proceso de trastienda no es posible tener información, pues suele ser ese el secreto que hace rentable el funcionamiento de las plataforma.

Se han descrito también referencias de técnicas de montaje que podrían servirnos para atravesar ese enorme volumen de contenido audiovisual, extrayendo probablemente datos sensibles, aportes para construir un espejo contemporáneo de la humanidad.

Tecnologías de Video description, Video retrieval multi model understanding

El campo de análisis del contenido de vídeos e imágenes mediante aprendizaje automatizado, es muy activo en el presente, como se ha observado a lo largo de este texto, está vigente en varios espacios de investigación con diversas aproximaciones en desarrollo, complejas de enumerar y describir, en parte por el vertiginoso avance que las caracteriza.

Con el fin de hacer un corte transversal que de cuenta de que clase de cosas están sucediendo en torno a los métodos automatizados de descripción de video se hará aquí una breve descripción de las metas que se impone esta reciente área de estudio.

El propósito de un sistemas de descripción de video es el de extraer el contenido de un video, y convertir esos elementos detectados en una descripción mediante *lenguaje natural*, es decir oraciones que podríamos usar seres humanos, para describir lo que sucede en los planos, secuencias una a una, o en un grupo completo de las mismas. Muchas de las experiencias

utilizan videos de entre 2 y 3 minutos de duración en promedio para entrenamiento. Mientras más complejo y variado sea el contenido, más difícil se vuelve la tarea de extraer una oración, o *caption* que sintetice con agudeza lo que pasa en el video.

El concepto detrás del *multimodel understanding* implica que la segmentación de la información audiovisual es analizada de varios modos en simultáneo. Por un lado el audio, por otro los ítems que se detectan en cuadro y finalmente las acciones que realizan las personas que aparecen en escena. Son modelos independientes que luego confluyen para ser comprendidos globalmente con el fin de extraer una oración que los describa (**Figura 18**).

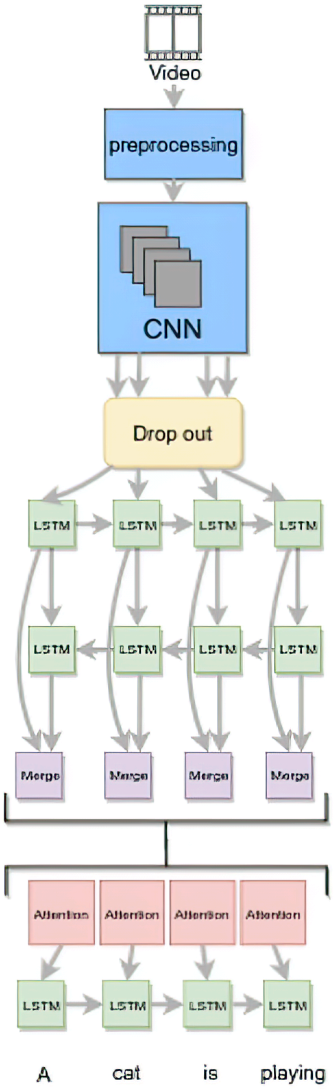

Figura 18: *Algoritmo de video descripción.*

Durante este texto se han mencionado redes neuronales y su uso para la detección de elementos y patrones en imágenes, se incorpora a esto el campo de la *video descripción*, que comple con tareas mas sofísticadas que el etiquetado de video, tareas que implican la comprensión artificial del contenido, involucrando un cruce de saberes y herramientas basado en lo que anteriormente se describió aquí, pero con un grado de complejidad veces superior.

La capacidad artificial de interpretar el contenido de un video representa un valioso nodo de potenciales usos, tan diversos como el cuidado a distancia vía streaming de ancianos en un geriátrico, o la ayuda en sitios web para que personas no videntes puedan saber qué contenidos están navegando. Los múltiples usos de esta tecnología, entre los que se encuentra la posibilidad del montaje automatizado, hacen que cientos de equipos humanos destinen energías a diario para hacer más eficientes sus algoritmos descripción de video.

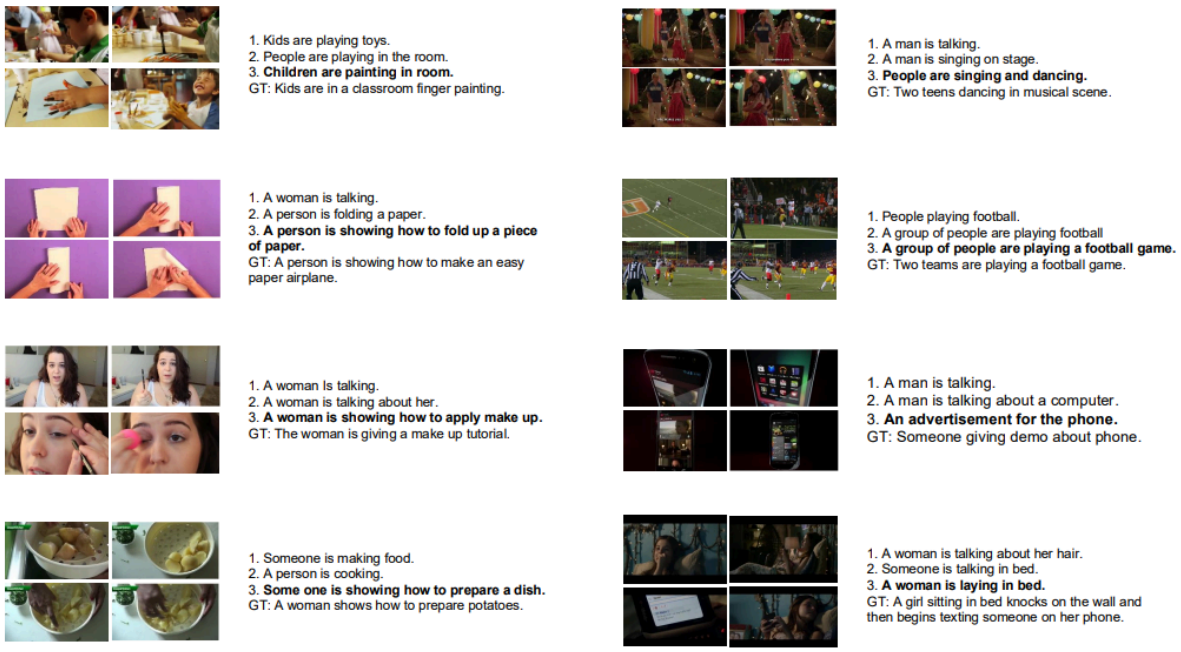

Figura 19: *Resultado de análisis de video y descripción automatizada del contenido mediante distintas estrategias.*

Uno de los desafíos más complejos es el de generar un dataset de video de tamaño suficiente, es decir de cantidades de videos y variedad, como para ser representativo del contenido general que puede tener un video, y por tanto útil para entrenamiento de un red neuronal. Exceptuando lo que sucede con YouTUbe y su manejo de datos, las investigaciones de

menores recursos *independiente*s buscan alternativas que no impliquen sólo analizar video, sino también datos contextuales aislados del contenido audiovisual, esto debido a la exigencia computacional que implica acopiar y estudiar millones de fotogramas de video.

Entre estas, llamadas *estrategias multimodales*, están las investigaciones que generan modelos a partir del estudio de descripciones de videos en *lenguaje natural* recogidas de todos los rincones de la web, estrategia denominada *from the wild* (desde lo salvaje). Mediante esta estrategia se compila un gran volumen de descripciones de video, generando un dataset de descripciones de video para entrenamiento. El modelo resultante de ese proceso se utiliza para colaborar sintáctica y semánticamente con la creación de descripciones automatizadas, reduciendo radicalmente el coste computacional del análisis de video, y comprendiendo que la estructura lingüística de cada descripción es fundamental en el proceso de creación de una descripción de video nueva (**Figura 19**).

Sub Plataformas que recirculan información audiovisual

En este mismo texto se ha hablado extensamente sobre YouTube, haciendo foco varias veces en volumen de su contenido y la manera cuantiosa en la crece a diario, esa clase de dimensiones de cantidad de información inéditas hasta hace relativamente pocos años, es parte de lo que popularmente se conoce como Big Data. Una de las características de la Big Data es las dificultades que implica el interactuar con ella En alguna medida ese conglomerado de imágenes y sonidos conforma una suerte de Aleph borgiano, contiene tanta información que para la comprensión de dimensiones del ser humano equivale al infinito, miles de siglos de videos.

El primer compromiso de una plataforma que contiene esa cantidad exorbitante de información, es desde luego como almacenarla, pero al mismo tiempo cómo lidiar con la misma y ofrecer su contenido a millones de personas. Esta dificultad significa para YouTube la tarea de generar distintos métodos de selección y compresión de información, muchos visibles por el simple hecho de que la plataforma ofrece un menú acotado de posibilidades de navegación, y de contenido relevante para el perfil de cada usuario.

En este contexto hay otras plataformas, independientes, que generan interfaces en la web, para exponer el contenido profundo del repositorio de Youtube, sin estar asociadas contractualmente con Google o YouTube.

Estas otras plataformas como *astronaut.io*, *petittube.com* o *underviewed.com*, han desarrollado maneras de presentar el vastísimo contenido sin basarse en estudio de perfil de cada usuario ni sofisticados sistemas de recomendación, ni en la popularidad de los videos.

La pretensión de estos sitios, que existen desde mediados de la década de 2010, es exhibir lo menos visto en YouTube, es decir el extremo opuesto de lo popular, que como se ha mencionado anteriormente, representa la enorme mayoría del contenido, siendo que casi el 90% de los vídeos disponibles es visto por menos de 100 personas, miles de millones de videos.

Debido al hermetismo con el que Youtube controla su contenido, y la imposibilidad de acceder a modificar el algoritmo discutiblemente monopólico con el que lo distribuye y modera sus videos, estas sub plataformas generan automatismos de búsqueda basados en el nombre de los videos subidos. Considerando que los videos que utilizan el nombre genérico que la cámara o la aplicación con las que fueron creados o grabados les asignó, por ejemplo MVI_0434.AVI, pueden ser un indicio de que el video no tuvo demasiada atención al realizarse, y que debido a ese nombre difícilmente escale en algún ranking de popularidad, sospechando que a nadie se le ocurriría buscar un video usando una serie alfanumérica aleatoria (**Figura 20**).

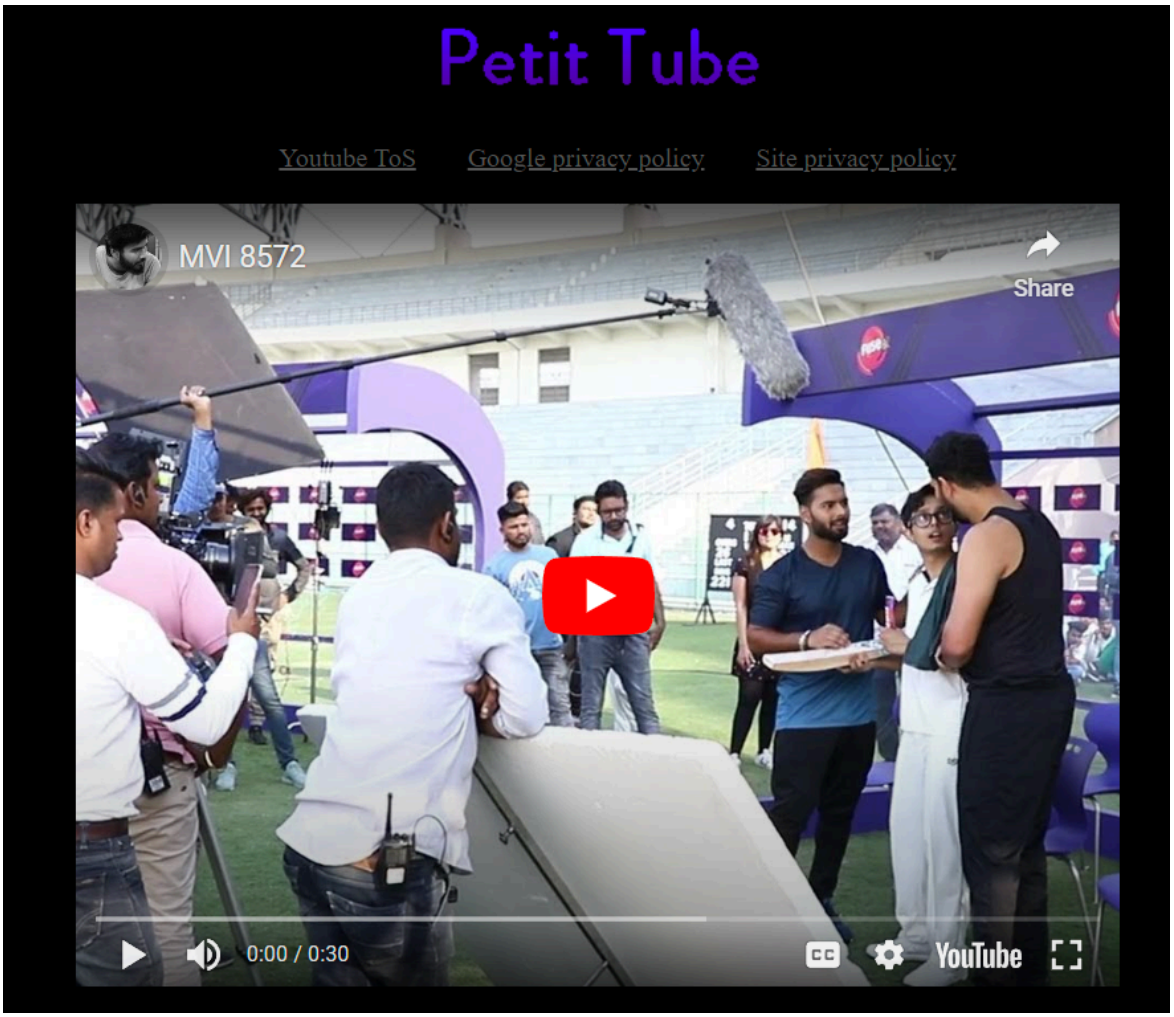

Figura 20: *Captura de pantalla de Petit Tube.*

Al margen del aspecto técnico y estadístico, la experiencia que proveen estas plataformas pone en evidencia el valor antropológico del contenido profundo de Youtube. Una hora de visionado de material en cualquiera de estas sub plataformas, nos ofrece un montaje aleatorio de escenas de origen remoto que revelan un mapeo de la diversidad humana en sus gestos cotidianos.

Eficiencia vs. creatividad

Como se ha exhibido en el presente texto, diversas estrategias en desarrollo suelen analizar modelos de descripciones en *lenguaje natural*, para mejorar las cualidades de la interpretación de un algoritmo de descripción de video. Es notorio que muchas veces el

resultado del análisis resulta más sorprendente que preciso. Cuanto más canónica sea la tarea a resolver para un sistema de montaje automatizado con I.A., más posibilidades hay de que el trabajo esté a la altura o supere el de un ser humano en términos de eficiencia. Por ejemplo, montar el material de un evento, una boda o una entrevista en el noticiero de la tarde, es una tarea que un sistema basado en IA, puede realizar con cierta efectividad en la actualidad, debido a que son las mencionadas, tareas que obedecen a un patrón muy bien establecido y reiterado. Se desprende de esto que la comprensión profunda no es necesaria, sino más bien el reconocer los puntos canónicamente más importantes.

Por lo tanto, expresar mediante el montaje de secuencias de vídeo, la sutileza de emociones complejas en la construcción de una relato, aparenta ser demasiado pedir para el nivel de comprensión, que tienen los algoritmos descritos, de cada fragmento de video a editar.

Para reemplazar personas en tareas que pueden ser entendidas como genéricas, los sistemas con I.A. de montaje pueden ser útiles, principalmente por la capacidad de hacerlo muy velozmente y a gran escala.

Pero las cualidades reflexivas humanas en obras de mayor sutileza no parecen aún haberse logrado, fundamentalmente porque los sistemas disponibles en la actualidad, no son capaces de identificar la enorme y subjetiva cantidad de variables que implica la construcción de sentido mediante la yuxtaposición de secuencias de video. Retomando aquí lo mencionado por Simondon en la cita del segundo capítulo, la condición de indeterminación en el resultado de la tarea de un sistema automatizado de montaje representaría un grado de desarrollo avanzado, ya que esa característica, la indeterminación es de algún modo análoga a la creatividad humana.

Volviendo a la segunda parte de la reflexión, la que contiene ideas habilitadas por el estudio de prácticas y autores cuyos métodos podemos denominar como alternativos, es donde se observa el potencial del cruce entre informática y montaje. Esa convergencia no solo está siendo, casi de manera invisible, la que modela la gran parte las rutinas de consumo humanas, sino desde la que posiblemente surgirán aún gran cantidad de aplicaciones y utilidades.

Es un patrón habitual en la historia de la tecnología, que el surgimiento de nuevos medios y posibilidades, suelen en primera instancia utilizarse para imitar las capacidades técnicas obtenidas por medios y tecnologías anteriores. Podemos entender este proceso, en estudios

genealógicos de medios, como un proceso de validación de utilidad, mediante el cual la nueva generación tecnológica se obstina en equiparar a la anterior, sin liberar su potencial aún a nuevas aplicaciones y concepciones.

Este patrón que incluye cierta validación de utilidad parece necesariamente marcar a las disciplinas que entran en contacto con la inteligencia artificial, ya que el parámetro de las capacidades humanas está siempre presente como estándar de comparación. La medida de evaluación intenta comparar lo que puede hacer un ser humano, y que tan bien podría imitarlo un sistema artificial (Alpaydin, E., 2019).

Conclusión

Una minería audiovisual

Así como existe la minería de datos, es posible imaginar una *minería audiovisual*. Al igual que en la minería o exploración de datos, el objetivo de la minería audiovisual sería el de extraer información de un conjunto de datos (en este caso contenido audiovisual) y transformarla en una estructura de sentido para su uso posterior.

El blanco de análisis, en términos de información, sería en una primera etapa los videos menos vistos de Youtube, los que jamás son recomendados por un algoritmo, e inclusive rara vez son vueltos a ver por quienes los subieron originalmente, ni por sus contactos o seguidores, en caso de que tengan alguno.

Los videos nunca vistos padecen su propia naturaleza, la de estar mal grabados en términos técnicos, el contenido de los mismos tal vez sea monótono y genérico, y ni siquiera sus títulos los hacen atractivos como para que alguien mediante una búsqueda azarosa, de con ellos.

Es importante reconocer que todo ese material olvidado, representa una biblioteca, o mas bien una videoteca, la mas grande que la humanidad haya generado en toda su historia, y está disponible y a pocos pasos de distancia, su contenido aún no cumpliendo con las normas que hacen que algo sea popular representa un *mapeo* de la sociedad de una precisión ignorada.

Lo documentado en este repositorio podría ser de enorme valor antropológico, rutinas y ceremonias que ilustran reveladores datos mediante videos subidos por millones de personas, de una sociedad que expone su vida en las redes.

Es prácticamente imposible dar azarosamente con lo impopular, debido a que el volumen descomunal de información que y la necesidad de una herramienta eficiente para el entretenimiento y la colocación de productos de consumo, han moldeado una interfaz que prioriza con sensatez contextual, ofrecer videos populares a ofrecer video ignotos. La función de los mencionados *sistemas de recomendación* no es pura malicia, estos sistemas intentan resolver un problema complejo, priorizando la eficiencia en el consumo.

Pero bien, si alguien quisiera ir por la vía contraria, navegar los millones de videos de la periferia, desde las opciones admitidas por la plataforma YouTube, es sumamente difícil, sino imposible. Esto queda en evidencia cuando se estudia el funcionamiento de los sistemas que utiliza la plataforma. Los mismos tienden a la generalización de lo que más se ve, y evitan la diversidad de contenidos y orígenes. Esa tendencia es irreversible para un solo usuario, aún explorando todos los parámetros de búsqueda durante meses, debido a la magnitud de la marea estadística que apunta a lo contrario.

Hacer *minería audiovisual*, sería tomar un volumen importante de contenido de YouTube muy poco visto, por ejemplo 1 millón de videos de menos de 100 vistas, extraídos de la manera más aleatoria posible. Luego, mediante una red CNN entrenada de manera independiente, desglosarlos en palabras, para una vez que se cuenta con todas esas palabras comenzar a hacer cruces exploratorios de montaje, utilizando como primera herramienta las prácticas del cine estructural, colocando por ejemplo una detrás de otra todas los fragmentos de video de los cuales se extrajo una misma palabra, a modo del *montaje seriado*, desarrollado por el artista Kurt Kren, reseñado en el segundo capítulo de este texto.

El término *minería*, para describir esta práctica proviene del hecho de que es prácticamente imposible, al menos para un usuario, visualizar todo ese contenido para tomar decisiones de

montaje a posteriori. Lo que se hace necesario es tener un estrategia de *minado*, o de exploración, la idea es que las que estrategias mencionadas anteriormente provenientes del cine estructural y del found footage serían provechosas si fueran traspuestas a este nuevo ámbito, puesto que evitarían los objetivos productivos asociados a los sistemas de recomendación que apuntan al rendimiento máximo en términos comerciales, y habilitarán forma de consumo aleatoria, que podría revelar magnitudes y diversidad de contenidos jamás explorados por cada usuario.

En la introducción de la presente investigación se planteó la convergencia de dos áreas de pensamiento y creación, originalmente distantes. La de la informática, representada por los desarrollos en el campo de la Inteligencia Artificial como punta de lanza, y por otro lado la del montaje, percibido como cualidad esencial en la creación de sentido en piezas de video. En función de esta convergencia, es que se formuló la pregunta respecto a las habilidades de sistemas basados en IA para montar secuencias de video, siendo capaces de comprender el contenido de las mismas y generando comunicación o expresión controlada, mediante el uso de edición.

A lo largo de este texto se describió la coyuntura que propició el acercamiento de las dos áreas mencionadas, la construcción en la web de repositorios de vídeo de cuya cantidad de contenido podemos identificar por su volumen como *Big Data*, y las estrategias de moderado y distribución automatizada, que necesariamente tuvieron que inventarse para que el funcionamiento e interacción con el material audiovisual de la plataforma sea posible.

Al día de hoy, en los comienzos de la tercera década del siglo XXI, es posible decir que estos avances en la catalogación de material de manera automática, están en camino a desarrollar la capacidad autónoma de comprender cabalmente el contenido de un video, y por tanto, a partir de esa comprensión, poder expresar y comunicar ideas a partir del montaje de distintas partes partes de videos. Esta posibilidad incluye cruces con la capacidad técnica de reconocer elementos puntuales en un fotograma y la de utilizar palabras que manifiesten ordenadamente, en *lenguaje natural*, ese nivel de comprensión.

Es posible resumir los avances recientes que han permitido que esa, aún relativa, *compresión artificial* sea posible en una breve línea de tiempo que describe las subcategorías de estudio surgidas y potencias en los últimos años.

Desde la capacidad de juicio autónomo, Inteligencia Artificial, y la visión por computadora como disciplinas muy abarcativas, a situaciones de uso puntuales, como el reconocimiento de elementos, dentro de lo que se podría incluir el reconocimiento óptico de caracteres como ejemplo. Dentro del campo del reconocimiento de ítems en información visual, el desarrollo de esta posibilidad en fotografías, y luego en fotogramas, donde aparecen sub disciplinas más específicas, como el *video tagging*, que abarca lo que se denomina ingesta, desglose, o etiquetado de video, pero realizado de manera autónoma. Como siguiente paso, en grado de especificidad se pueden mencionar técnicas de Video2text, es decir sistemas autónomos de transcripción del audio de un video a texto, para su lectura, traducción o comprensión. Finalmente, complementando sólo lo que interesa al esta investigación, como sub disciplina más reciente y sofisticada se mencionó la *video descripción*, es decir la posibilidad de que una máquina (un algoritmo) sea capaz de interpretar el contenido de un video y pueda convertir lo que *ve* en una descripción ajustada a la realidad y en *lenguaje natural*.

Por otra parte, desde el terreno expresivo artístico, se observaron algunos referentes periféricos o experimentales, que en la creación de sus piezas han indagado otras maneras de montar imágenes y sonidos para construir secuencias. La relectura y visionado de esos autores aportó una visión más amplia de lo que puede llegar significar el montaje como herramienta narrativa, al margen de su utilización clásica, y las normas canónicas cuya sintaxis el público ya tiene interiorizada. Ejemplo de estas normas de lenguaje estandarizado pueden encontrarse en la cinematografía mainstream y la gran mayoría del material producido para TV tradicional.

Haciendo esta salvedad entonces es que se enuncian dos reflexiones. Una con respecto a la potencialidad en el uso herramientas que utilizan IA en tareas que repiten patrones de montaje canónicos o clásicos, y otra distinta en las que experimentan con la información generando cruces de archivo y contemplando el resultado de la obra como insumo para la reflexión.

En primer término, con respecto a las posibilidades del uso de herramientas IA en el montaje de secuencias dentro de reglas de lenguaje preestablecidas, las de uso más habitual y canónico. Tareas como por ejemplo, construir una película de género policial completa a partir de fragmentos de material crudo. Por lo expuesto en este texto sabemos que la fiabilidad de los sistemas automatizados de detección de contenido, etiquetado e ingesta de

video, es aún relativa. El grado de comprensión artificial de un video es aún falible y las técnicas para mejorar esa capacidad enfrentan aún desafíos. Es evidente que para poder construir una secuencia de sentido, que se elabora desde la comprensión de cada imagen que la compone, es necesario reconocer primero de manera sensible y profunda que contiene cada una de esas partes. Si contamos solo con etiquetas y conceptos surgidos de las mismas es posible caer en varios posibles errores, por ejemplo, imaginar un video del que se obtienen estos cuatro elementos como etiquetas, "vehiculo traccionado a sangre", "zapato de tacos altos" y "escalera" y "medianoche", los mismos pueden configurar múltiples relatos, pero tal vez pase inadvertido que estamos hablando de La Cenicienta.

Como se analizó en el recorrido por las estrategias de montaje en obras y artistas del cine estructural y cine de Found Footage, algunas de las variables presentes en la factura y concepción de las obras son: las reglas matemáticas, la consecución de series de elementos y el trabajo con archivos, y estos tres elementos son naturales en el ecosistema de *entrenamiento* de las redes neuronales que hoy día se utilizan para todo tipo de tareas de análisis de imágenes, incluído el denominado Montaje Neuronal.

Este contexto común revela, y permite proyectar hacia que áreas apuntan las capacidades y potencialidades de estas redes neuronales, desde luego algunas de ellas ya en funcionamiento con reconocida eficacia, como hemos visto a lo largo de este texto, en funciones de moderación, distribución y recomendación de contenidos. Los datasets y procesos de entrenamiento que han permitido que estas redes sean hoy eficientes, están asociados entonces a esas funciones originales.

Cuando se extrapolan esas redes entrenadas con estos datasets para funciones sensibles como el montaje de secuencias con objetivos narrativos que no están basados en patrones genéricos o reconocimiento de elementos o acciones, se está forzando su utilidad hacia un área de otra lógica y proceso.

En resumidas cuentas, las tareas genéricas de montaje están al alcance de los algoritmos IA, porque son capaces de realizar el proceso mecánico de pegar una secuencia de video detrás de la otra, reconociendo algunos valores, y utilizándolos como fundamento para colocar orden y duración a la serie. Pero, por otro lado, la lógica reflexiva mediante la cual lleva a cabo esta tarea es más bien superficial, y que la cantidad de elementos que está evaluando para tomar una u otra decisión de corte es limitada.

De todos modos este no es un análisis sobre las facultades o posibilidades de que un sistema artificial pueda montar una película como lo haría un ser humano. Lo que se desea advertir es que al día de hoy, investigados los sistemas vigentes, hay evidencias para creer que el contexto genérico de entrenamiento de las redes neuronales que se utilizan para ese fin, no es el mejor que podría existir, y para que este sistema exista deberían de modificarse varias instancias previas en el entrenamientos de las redes, y planear un sistema *ad hoc* para esta función, es decir, un complejo sistema de entrenamiento basado entre otras cosas, como por ejemplo en la historia del cine.

Si lograse ese objetivo se crearía un sistema que sería capaz de editar una película como un ser humano, intentando emular mediante capacidades artificiales algo que ya existe, tan sólo para comprobar su utilidad.

Lo que hay disponible al día de hoy, muchas veces bajo la autoproclama de ser sistemas de montaje automatizados mediante I.A., con cualidades cinematográficas, son en realidad, como se ha dicho anteriormente, sistemas que fuerzan las capacidades de redes neuronales entrenadas con otros objetivos, para que puedan editar videos.

Todo este argumento, no significa que es absurdo pensar en que las redes neuronales artificiales no sirvan para editar videos bien, sino por el contrario, que pueden hacer un trabajo sobrehumano, siempre y cuando se abra la comprensión del montaje a las posibilidades que son naturales a estas redes neuronales, las que las vinculan con la detección de elementos y acciones, en volúmenes de información inabarcables en dimensiones humanas. Allí probablemente está el verdadero potencial, de alguna forma pre anunciado por algunas prácticas y vanguardias de las que he hecho mención, con la introducción de la lógica y matemática al montaje y el trabajo de catalogación de archivos.

Los algoritmos que incluyen redes neuronales, observan el contenido de fotografías y videos como si fueran jeroglíficos. Extraen conceptos en forma de palabras, y tratan de entender mediante una reflexión artificial, lo que representan y significan. Pasa algo similar cuando las personas analizamos nuestros sueños tratando de darles sentido, nos sentimos traduciendo algo como de otro tiempo, que intentamos trasladar a nuestro idioma y comprensión.

Es una misión enriquecedora apuntar a que las redes neuronales observen y narren en secuencias de videos, lo que perciben y comprenden en toda esa inédita cantidad de datos que

son capaces de analizar, articulando sentido, expresando esa forma de sentido, uniendo unas partes de archivos con otras, atravesando en la edición vidas de personas y lugares geográficos, generando patrones de montaje múltiples, liberando lo aprendido para ilustrarnos de nuevas formas de entender lo que producimos audiovisualmente como humanidad.

La tarea de montaje verdaderamente valiosa, y que da sentido a la convergencia del montaje y la de la informática, se parece a la anteriormente descrita exploración o *minería audiovisual*, una herramienta visual que nos permita extraer alguna forma de sentido del Big Data audiovisual existente en YouTube.

Exploración y minado, son términos ya en uso para hablar de trabajo intensivo sobre grandes volúmenes de datos *Big Data*, en este caso, el de la exploración audiovisual el montaje sería de asociaciones libres y automatizadas, fruto de la navegación de sistemas de montaje automatizado por los siglos de video disponibles en un repositorio como YouTube. Lo que la exploración devolvería son secuencias organizadas, una detrás de la otra, ilustrando patrones y continuidades detectadas por redes neuronales que en una fracción de tiempo infinitamente más veloz que un ser humano, son capaces de reconocer que hay en millones de videos, producidos por personas de todo el planeta que con distinta suerte y popularidad esperan ser vistos en la web.

Esta proyección es la de una función complementaria, en la que informática y montaje en conjunto revelarían información sobre culturas y sus coincidencias, y permitiría verificar la lógica y capacidad real mediante las cuales algoritmos similares comprenden perfiles y sugieren ver o comprar cosas. poniendo a las personas en contacto con la mecánica que opera y modera flujos de información que rodean sus vidas. Proyectar sistemas con esos objetivos, es en parte pensar funciones superadoras en la que ambas partes, informática y montaje, descubran su potencial como conjunto, sin reemplazar las personas y su sensibilidad, sino habilitando la complementación de saberes y habilidades.

Bibliografía

<u>Libro:</u>

Alpaydin, E. (2019). *Machine learning*. MIT press.

Buñuel, L. (1982). *Mi último suspiro (Memorias)*. España: Plaza & Janes.

Burch, N. (2017) Praxis del cine, Editorial Fundamentos, España

Burgess and Green (2018). *YouTube*. Cambridge: Polity Books.

Ceruzzi, P. (2012). *Computing*. MIT press.

Does, A. (2012). *Do Not Look Away, The life of Arthur Lipsett*. Canadá: Independiente.

Eisenstein, S. (2020). *La forma del cine.* Ciudad de México: Siglo XXI editores.

Eisenstein, S. (2020). *El sentido del cine.* Ciudad de México: Siglo XXI editores.

Gidal, P. (1978). *Structural Film Anthology*. Inglaterra: British Film Institute.

Gombrich, E.H. (2004). *El sentido del orden, estudio sobre la psicología de las artes decorativas*. Londres: Phaidon.

Hyde, R. (2004). *Myrioramas, Endless Landscapes The Story of a Craze*. Print Quarterly Publications.

Huhtamo, E. (2013). *Illusions of motion, Media archeology of the moving panorama and related spectacles*. Massachusetts: The MIT press.

Huhtamo, E., Parikka, J. (2011). *Media archeology, Approaches, applications, and implications.* Londres: University of California Press.

Jenkins, B. (2009). *On camera arts and consecutive matters. The writings of Hollis Frampton*. Londres: The MIT press.

Kane, F. (2018). *Building Recommender Systems with Machine Learning and AI*. Estados Unidos: Independiente.

Kelleher, J. (2019). *Deep learning*. MIT press.

Kelleher J. D, Tierney B. (2018). *Data Science*. MIT press.

Koenitz H., Ferri G., Haahr M., SezenD Sezen T. I. (2015). *Interactive Digital Narrative, History, Theory and Practice.* Estados Unidos: Routledge.

Koonce, B. (2021). *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization.* Apress Berkeley, CA.

Levi; Pavle. (2012). *Cinema by other means*. Estados Unidos: Oxford University Press.

Ong, W.J. (2021). *Oralidad y escritura: Tecnologías de la palabra*. Buenos Aires: Fondo de Cultura económica.

Pierson, M. , James, D., Arthur, P. (2011). *Optic Antics, the cinema of Ken Jacobs*. Estados Unidos: Oxford University Press.

Lauridas, P. (2020). *Algorithms*. MIT press.

Schrage, M. (2020). *Recommendation engines*. MIT press.

Simondon, G. (2007). *El modo de existencia de los objetos técnicos*. Buenos Aires: Prometeo.

Szeliski, R. (2011). *Computer vision, algorithms and applications*. Londres: Springer.

Tscherkassky, P. (2012). *Film Unframed: A History of Austrian Avant-Garde Cinema*. Austria: Austrian Film Museum Books.

Wees, W. (1993). *Recycled Images, The arte and Politics of Found Footage*. Estados Unidos: Anthology Film Archives.

Wees, W. (2007). *From film Compilation to Found Footage, Films of Arthur Lipsett.* Canadá: University of Toronto Press.

Zielinski. S. (1999). *Audiovisions.* Holanda: Amsterdam University Press.

Texto en compilación ajena:

Wukui, Y, Shan, G, Wenran, L, Xiangyang, ji, (2018) Stream Convolutional Networks for Video Action Recognition with Hybrid Motion Fieldm dept. Automation of Tsinghua University, Beijing, China

Frampton, H. (1980) Film in the house of the word

Laokulra, N., Okazaki, N., Nakayama, H (2018) Incorporating Semantic Attention in Video Description Generation, Artificial Intelligence Research Center, Universidad de Tokio, Japón

Jun Xu, Tao Mei, Ting Yao and Yong Rui (2019) Large Video Description Dataset for Bridging Video and Language, Microsoft Research, Beijing, China

Addair, T, (2017) Deep Learning YouTube Video Tags, Stanford University

Guadarrama, S., Krishnamoorthy, N, Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell,T, Saenko, K.(2014)YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-shot Recognition, The Computer Vision Foundation.

Knwal, Y., Tabassam, N. (2021)A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos, Department of Software Engineering, University of Engineering and Technology, Taxila, Pakistan

Zappin, A., Malika, H., Dampiera,D. A., Shakshukib, E.N. (2022) YouTube Monetization and Censorship by Proxy: A Machine Learning Prospective

Cool, K, Seitz, M., Mestrits, J., Bajaria, S, Yadati, U. (2017) YouTube, Google, and the Rise of Internet Video, Kellogg School of Management, NW University

Yousaf. K ,Nawaz,T. (2022) A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos, Department of Software Engineering, University of Engineering and Technology, Pakistan

Orozco,C., Xamena, E., Buemi M. E, Berlles, J.J. (2020) Reconocimiento de Acciones Humanas en Videos usando una Red Neuronal CNN LSTM Robusta) Ciencia y Tecnología, No 20, 2020, pp. 23-36

Monahan, M (2015) The Kuleschov effect, The poetry Ireland review, Irlanda

Sitios web

Van Buskirk, E. (2009) BellKor's Pragmatic Chaos Wins $1 Million Netflix Prize by Mere Minutes www.wired.com/2009/09/bellkors-pragmatic-chaos-wins-1-million-netflix-prize/

Turek,R. (2019) What content dominates on YouTube?

pex.com/blog/what-content-dominates-youtube/

Historia de YouTube https://en.wikipedia.org/wiki/History_of_YouTube

Gutierrez, D. (2017) Introducing Flo – Bringing Deep Learning to Video Editing. Inside Bigdata. Recuperado el 10/10/2020 de https://insidebigdata.com/2017/06/19/introducing-flo-bringing-deep-learning-video-editing/

Cristobal Valenzuela, sitio web del artista. cvalenzuelab.com

Gliacloud, sitio web gliacloud.com

Wordeye, sitio web wordeye.com

repositorio audiovisual web *astronaut.io*

repositorio audiovisual web *petittube.com*

repositorio audiovisual web *underviewed.com*

Cloudinary, sitio web de herramientas visuales IA cloudinary.com

Audiovisuales:

Frampton, H. (Director). (1970). Zorns Lemma [Película]. Estados Unidos

Lipsett, A. (Director). (1960) Very Nice Very Nice, Canadá, NFB

Lipsett, A. (Director). (1965) A trip down memory lane, Canadá, NFB

Lipsett, A. (Director). (1964) 21-87, Canadá, NFB

Kren, K, (Director) (1995) Thousandyearsofcinema, Austria

Kren, K, (Director) (1960) 3/60 Bäume im Herbst, Austria

Gaucher, E. (Director) (2007) A dot on the histomap, Canadá, NFB

Wieland, J. (Director) (1967) Sailboat, Estados Unidos

Klingemann, M. (Director) (2018) Neural remake of Total Eclipse of the Heart

Castle. W. (Director) (1965) Mr Sardonicus