Sistema Híbrido Inteligente para el Análisis de Contenido Multimedia

TESIS DOCTORAL

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

Autora

Lucía Martín Gómez

Dirigida por los Doctores

Juan Francisco De Paz Santana

Vivian Félix López Batista

Facultad de Ciencias
Departamento de Informática y Automática

JULIO 2020

Declaración de Autoría

El Dr. Juan Francisco De Paz Santana y la Dra. Vivian Félix López Batista, Profesores Titulares de Universidad del Departamento de Informática y Automática de la Universidad de Salamanca

HACEN CONSTAR

que la doctoranda Lucía Martín Gómez ha desarrollado este trabajo titulado *Sistema Híbrido Inteligente para el Análisis de Contenido Multimedia* bajo su supervisión, y por ello autorizan su presentación para la obtención del título de Doctora.

DE PAZ SANTANA JUAN FRANCISCO - 76125754D

Firmado digitalmente por DE PAZ SANTANA JUAN FRANCISCO - 76125754D
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-76125754D, givenName=JUAN FRANCISCO, sn=DE PAZ SANTANA, cn=DE PAZ SANTANA JUAN FRANCISCO - 76125754D
Fecha: 2020.07.27 09:23:44 +02'00'

Fdo. Juan Francisco De Paz Santana

LOPEZ BATISTA VIVIAN FELIX - 12420865Z

Firmado digitalmente por LOPEZ BATISTA VIVIAN FELIX - 12420865Z
Fecha: 2020.07.27 09:46:25 +02'00'

Fdo. Vivian Félix López Batista

En Salamanca, a 27 de julio de 2020

A ti, que aunque te has ido,

siempre estarás presente.

If a machine is expected to be infallible,
it cannot also be intelligent.

Alan Turing

Agradecimientos

Only a life lived for others
is a life worthwhile.

Albert Einstein

A lo largo del tiempo empleado para el desarrollo de este trabajo me he sentido acompañada y arropada por muchas personas. A todas ellas deseo expresarles mi gratitud.

En primer lugar, quiero agradecer a mis directores de tesis su apoyo, ayuda y atención a lo largo de este recorrido. Concretamente, a Juan Francisco De Paz Santana por poner a mi alcance soluciones a las dificultades que han ido surgiendo a lo largo de estos años. A Vivian Félix López Batista, por ser un impulso constante tanto en el desarrollo profesional como en el personal.

Debo agradecer a la Universidad de Salamanca y al Grupo de Investigación BISITE, por permitirme iniciar mi andadura en este arduo camino y por ayudarme a crecer como persona. Estoy muy agradecida también con el Grupo de Investigación ESALAB, por acogerme con los brazos abiertos y estar siempre dispuestos a ayudar. Y por supuesto, no puedo terminar este trabajo sin agradecer la admirable dedicación y labor de María N. Moreno y el trato que da a los estudiantes de este programa de doctorado.

A Lope y a Juan les agradezco la pasión que transmiten y contagian por la ciencia y la investigación; a Javi, su paciencia y su apoyo incondicional. Estoy enormemente agradecida con los tres por aportar siempre una bocanada de aire fresco y por ser un gran ejemplo de amistad.

Mi gratitud va también dirigida a la Universidad Pontificia de Salamanca por ofrecerme la oportunidad de satisfacer mi vocación profesional en el ámbito docente e investigador. Me gustaría dar las gracias a todos mis compañeros por su apoyo y por todo lo que he podido aprender de ellos en este tiempo. Quiero agradecer especialmente a Alfonso José López Rivero la confianza que ha depositado en mí desde el primer momento y la *presión* que ha ejercido sobre mí para que este trabajo llegara a buen puerto. Y, por supuesto, tengo que hacer una mención especial a Daniel, Fernando, Rubén, Luis y particularmente Rebeca porque, además de estar presentes en los buenos momentos, siempre han tenido palabras de aliento para ayudarme a superar los obstáculos y darme un baño de realidad.

Finalmente, me gustaría dar las gracias a mi familia por confiar en mí más que yo misma. A mis padres, porque, como ellos dicen, la mayor herencia que me dejan es el afán de superación y la educación. A mi hermana, por apoyar cada una de mis decisiones y ser mi compañera de batalla y un pilar fundamental. A ti, porque aunque ya no estés físicamente con nosotros has sido, eres y serás el mayor ejemplo de fuerza y superación para mí. Muchas gracias a todos por ser mi sustento y no dejarme caer.

Resumen

Los grandes avances en las áreas de las TIC, el IoT y la IA han propiciado una serie de sistemas cuyo uso se ha visto incrementado exponencialmente en los últimos años, fomentando la generación de ingentes cantidades de datos de naturaleza heterogénea. Las propuestas recogidas en la literatura para la explotación de estos datos están enfocadas a la resolución de problemas muy específicos, favoreciendo el desaprovechamiento de la información. Este trabajo plantea una arquitectura modular y flexible para implementar un sistema híbrido inteligente capaz de soportar diferentes procesos de análisis de contenido multimedia gracias a la adaptación del concepto de ETL y la aplicación de tuberías de datos. Con el objetivo de comprobar el potencial de la arquitectura propuesta, se diseñan dos *frameworks* para la automatización del proceso de composición musical descriptiva a partir de contenido audiovisual y se desarrollan dos casos de estudio bien diferenciados donde se aplican diversas técnicas de extracción de meta-información y algoritmos enmarcados en el área del aprendizaje automático. La discusión de los resultados obtenidos se realiza considerando el rendimiento de los algoritmos y la aceptación social de la música por medio de diferentes test de usuario. En conclusión, la propuesta favorece la validación de la hipótesis previamente establecida, evidenciando que los datos multimedia analizados mediante técnicas de IA permiten crear otro tipo de información útil para el usuario.

Abstract

Great advances in the fields of Information and Communication Technologies, IoT and AI have led to a series of systems whose use has increased exponentially over the past few years. This has encouraged the generation of huge amounts of data of a heterogeneous nature. The state of the art gathers many proposals for the exploitation of these data, but they all focus on the resolution of specific problems, favouring the waste of information. This work proposes a modular and flexible architecture to implement an intelligent hybrid system for the analysis of multimedia content. Thanks to the adaptation of the ETL concept and the application of data pipelines the system can support the concurrence of several analysis processes running in parallel. With the aim of verifying the potential of the proposed architecture, two frameworks are designed. Both are oriented to the automatic composition of descriptive music based on audiovisual content and they are put into operation in two well-differentiated case studies where diverse metadata extraction techniques and algorithms are applied within the context of machine learning. The performance of the algorithms and the social acceptance of the music are taken into account to validate the results obtained in this work. In closing, the proposal favours the validation of the previously established hypothesis, proving that the analysis of multimedia data through AI techniques allows the creation of other relevant information.

Índice general

Índice de figuras

Índice de tablas

Glosario

Aprendizaje Automático

Rama de la inteligencia artificial en la que los sistemas extraen conocimiento a partir de los datos y se adaptan a los cambios y problemas que suceden a su entorno de manera completamente autónoma. Conocido también como machine learning.

Armonía

Técnica que estudia la creación de un acorde y la concatenación de los mismos a partir de una serie de principios musicales relacionados con la percepción de varios sonidos de manera simultánea.

Clasificación

Técnica enmarcada en el aprendizaje automático supervisado que consiste en el análisis de datos y la evaluación de patrones con el objetivo de seleccionar y asignar la etiqueta categórica que resulta más apropiada para cada instancia del conjunto de etiquetas considerado en el estudio.

Creatividad Computacional

Área de investigación ubicada en la convergencia de la inteligencia artificial, el arte y la psicología cognitiva. El objetivo de esta rama de la ciencia es la obtención de obras artísticas y creativas por medio de la simulación de habilidades cognitivas humanas en las máquinas.

Ensemble

Conjunto de métodos o algoritmos de aprendizaje que se utilizan de manera combinada para solucionar un problema de tal forma que el rendimiento del conjunto sea mejor que el rendimiento de cada uno de los métodos que lo constituyen.

Fotograma

Cada una de las imágenes que se suceden en una película cinematográfica y que se pueden considerar de manera aislada. La traducción del término al inglés es ***frame****.*

Melodía

Sucesión lineal y ordenada de notas dotada de un ritmo definido y enmarcada en el contexto musical de una o varias escalas musicales que tiene un sentido propio y se percibe como una entidad musical.

META-INFORMACIÓN

Conjunto de metadatos o descriptores del contenido informativo de una entidad mayor. La meta-información de una imagen contemplará descriptores relacionados con las formas, los colores y la disposición de los elementos que aparecen en ella.

MULTIMEDIA

Combinación de diferentes tipos de información (texto, gráficos, animación, imágenes, vídeo y sonido entre otros) que da lugar a una entidad informativa multisensorial e interactiva.

MÚSICA DESCRIPTIVA

Estilo de música que tiene como objetivo evocar ideas, imágenes, estados de ánimo y otros elementos extra-musicales en la mente del oyente mediante la utilización de recursos musicales que se basan en la imitación de sonidos de la naturaleza y en la analogía de la información visible. Cuando el uso de elementos descriptivos es la base de la composición, la música se denomina programática.

SISTEMA HÍBRIDO INTELIGENTE

Sistemas software que aplican, en paralelo y de manera combinada, varios algoritmos enmarcados en el área de la inteligencia artificial. El objetivo del sistema se obtiene mediante la aplicación conjunta de todos y cada uno de los algoritmos que lo componen.

Test de Turing

Prueba de la capacidad de una máquina para exhibir un comportamiento inteligente similar al de un humano. El objetivo del test es que los usuarios no sean capaces de distinguir si una determinada tarea es realizada por una persona o por una máquina; de esta manera, según la propuesta de Alan Turing, se determina que una máquina es realmente inteligente.

Transfer Learning

Técnica aplicada en el ámbito del aprendizaje automático para la mejora del rendimiento de un algoritmo mediante la transferencia de conocimiento adquirido en la fase de entrenamiento de alguna tarea similar o relacionada.

Tuberías de Datos

*También conocidas como **data pipelines**, son una herramienta para la definición de un flujo flexible y automatizado en tareas de procesamiento de datos. La tubería recibe unos datos de entrada que transforma a modo de caja negra para proporcionar finalmente los datos de salida. Las tuberías permiten aplicar transformaciones en los datos de forma paralela y distribuida.*

Acrónimos

BoVW	***B**ag **o**f **V**isual **W**ords*
CENS	***C**hroma **E**nergy **N**ormalized **S**tatistics*
CNN	***C**onvolutional **N**eural **N**etwork*
DAG	***D**irected **A**cyclic **G**raph*
ETL	***E**xtract **T**ransform and **L**oad*
IA	***I**nteligencia **A**rtificial*
IoT	***I**nternet **o**f **T**hings*
LP	***L**abel **P**owerset*
LSTM	***L**ong **S**hort-**T**erm **M**emory*
MIDI	***M**usical **I**nstrument **D**igital **I**nterface*
MIR	***M**usic **I**nformation **R**etrieval*
ML-kNN	***M**ulti**L**abel **K**-**N**earest **N**eighbors*
NB	***N**aive **B**ayes*
RAKEL	***RA**ndom **K**-Lab**EL**sets*
RF	***R**andom **F**orest*
RMSE	***R**oot-**M**ean-**S**quare **E**rror*
RNN	***R**ecurrent **N**eural **N**etwork*
ROC	***R**eceiver **O**perating **C**haracteristic*
SIFT	***S**cale-**I**nvariant **F**eature **T**ransform*
SVM	***S**upport **V**ector **M**achine*

TIC ***T****ecnologías de la* ***I****nformación y la* ***C****omunicación*

TL ***T****ransfer* ***L****earning*

Parte I

Memoria de la tesis doctoral

Capítulo 1

Introducción

Resumen: *Este primer capítulo se desarrolla a modo de introducción del trabajo doctoral. En él se plantea la problemática que da pie al desarrollo de la investigación, se formula una hipótesis de partida, se definen los objetivos en el marco de los sistemas inteligentes para el análisis de contenido multimedia y se describe la metodología de investigación aplicada a lo largo de todo el proceso.*

El nacimiento de internet y la gran evolución en el área tecnológica a lo largo del último siglo han sido dos factores clave en la consecución de un mundo globalizado, donde los diferentes países, y como consecuencia sus ciudadanos, disfrutan de los beneficios de un alto y creciente nivel de interdependencia y facilidad de comunicación, independientemente de la distancia que los separe [9]. La inversión en investigación científica para el desarrollo de productos y servicios tecnológicos innovadores hace posible la obtención de sistemas de comunicación avanzada enmarcados en la rama de las Tecnologías de la Información y la Comunicación (TIC). Los objetivos de estos sistemas se enfocan, entre otras cosas, en garantizar la seguridad personal

o empresarial, facilitar y optimizar la ejecución de actividades laborales o proporcionar un recurso recreativo para la diversión y entretenimiento [108].

El acceso a estos sistemas, en un primer momento, se realizaba exclusivamente mediante un ordenador. Sin embargo, actualmente existen innumerables dispositivos dotados de conexión a internet y con capacidad para transmitir y recibir información, lo que hace aún más sencilla la comunicación interpersonal. Los avances tecnológicos en materia de inteligencia artificial (IA) y robótica han dado lugar a dispositivos como los actuales teléfonos móviles, los relojes inteligentes, televisores inteligentes o altavoces inteligentes —en inglés, *smartphones*, *smartwatches*, *smart TVs* y *smart speakers*, respectivamente—. En todos ellos, el adjetivo "inteligente" hace referencia, precisamente, a su capacidad para recoger, procesar y transmitir información [144]. En esta misma línea, el concepto de *Internet of Things* (IoT) es el fruto de la convergencia de diferentes áreas de conocimiento como las telecomunicaciones, la informática, la electrónica y las ciencias sociales, que tiene como fin la creación de una red de objetos cotidianos interconectados y dotados de una inteligencia ubicua para la recopilación y el tratamiento de datos [11, 138].

Un claro ejemplo de este tipo de sistemas cuya base es la combinación entre las TIC, el IoT y la IA son los asistentes virtuales inteligentes, que ofrecen un servicio interactivo capaz de procesar el lenguaje natural del usuario y mantener una conversación con él a fin de satisfacer algunas de sus necesidades [30]. Estos sistemas están incorporados en muchos tipos de dispositivos físicos entre los que destacan los ordenadores, los teléfonos móviles y los altavoces inteligentes. Esto permite realizar tareas muy diversas como controlar los dispositivos de automatización del hogar, buscar y reproducir contenido multimedia o realizar compras en internet mediante el control por voz [62]. Algunos de los asistentes más populares actualmente son Siri de Apple,

Alexa de Amazon, Cortana de Microsoft o el Asistente de Google.

Las plataformas de *streaming* como Netflix[1], HBO[2], Amazon Prime Video[3] o Twitch[4] son otro medio de difusión de información que ha atraído a muchos usuarios e investigadores debido a su proliferación en internet a lo largo de los últimos años [119]. Atendiendo a los intereses de los usuarios, estas plataformas permiten compartir y/o consumir contenido audiovisual generalmente con fines de entretenimiento. De hecho, [118] publicaba que el uso de plataformas de *streaming* se ha incrementado hasta en un 108 % en España y un 82 % en Francia durante la pandemia ocasionada por el COVID-19 (*coronavirus disease 2019*) a causa de la transformación de hábitos desencadenada por las medidas de distanciamiento social y la cuarentena. Por otra parte, los investigadores ponen el foco en el análisis de los patrones de interacción entre los usuarios de estas plataformas para mejorar dichos servicios. En esta línea, algunos tipos de sistemas de recomendación integrados en estas plataformas (como los basados en contexto o los basados en filtrado colaborativo [106]) analizan la información generada por un determinado usuario mediante sus interacciones para hacer recomendaciones a otros usuarios.

Por otra parte, [123] afirma que el medio de comunicación por excelencia en la actualidad son las redes sociales por la facilidad que ofrecen para compartir opiniones, ideas, percepciones y experiencias en diferentes formatos (texto, imágenes, audio y vídeo) garantizando además que esta información pueda ser accesible desde cualquier lugar del mundo. Si bien el número de usuarios de este tipo de sistemas ha sufrido un gran incremento en los últimos años, la edad de sus consumidores ha generado un crecimiento asimétrico.

[1] https://www.netflix.com/es/
[2] https://es.hboespana.com/
[3] https://www.primevideo.com/
[4] https://www.twitch.tv/

Así, el número de usuarios de entre 18 y 29 años es de un 12 % de la población en 2005, y en 2015 asciende al 90 %, mientras que los adultos entre 30 y 49 años pasan de un 8 % a un 77 % en el mismo período de tiempo, y para las personas entre 50 y 64 años los porcentajes parten de un 5 % en 2005 y ascienden a un 51 % una década después [105]. En España, el número de usuarios de redes sociales entre 16 y 65 años asciende de un 51 % de la población en el año 2009 a un 85 % en 2019 [63]. La Figura 1.1 muestra el porcentaje de personas con una edad comprendida entre los 16 y los 65 años que se pudieron contabilizar como usuarios de las redes sociales más utilizadas en España a lo largo del año 2019. Es interesante ver cómo algunas aplicaciones de mensajería instantánea como WhatsApp han evolucionado tanto que recogen funcionalidades propias de las redes sociales.

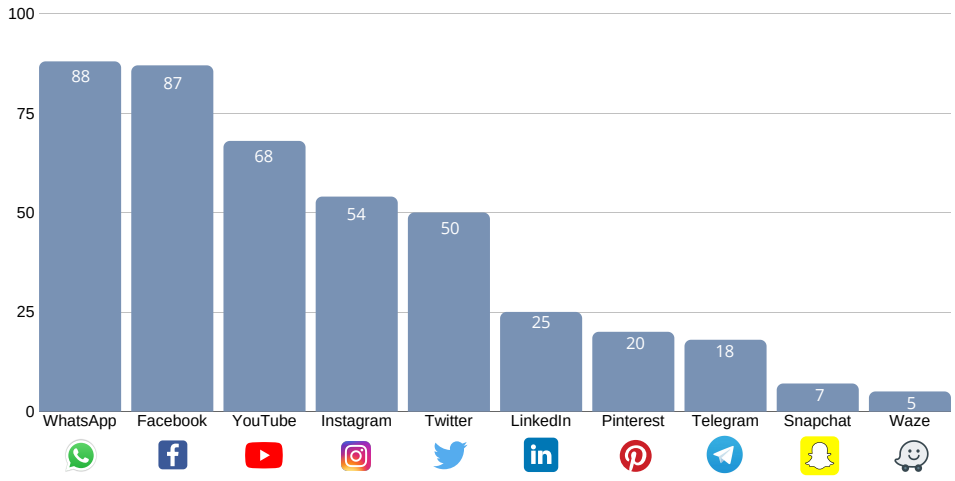

Figura 1.1: Porcentaje de personas entre 16 y 65 años que eran usuarios de las redes sociales más utilizadas en España durante el año 2019 *(Fuente: IAB Spain y Elogia [63])*

El uso cotidiano de todos estos tipos de sistemas basados en las TIC hace que generemos, inconscientemente, una inconmensurable cantidad de datos que además tienen formatos totalmente dispares (imágenes, vídeos, audio,

texto y atributos numéricos entre otros). Este gran volumen de datos supone, por una parte, una gran oportunidad para extraer información valiosa mediante técnicas de la IA. Sin embargo, a su vez, la ingente cantidad de datos y su condición heterogénea desencadena una serie de problemas y retos en términos de eficiencia y capacidad de procesamiento, que conllevan el desaprovechamiento de gran parte de la información que se genera en internet y que puede dar lugar a la obtención de conocimiento útil.

Para hacer frente al desafío que supone el análisis de los datos generados en sistemas de comunicación avanzada existen numerosas propuestas con propósitos muy divergentes; por ejemplo, [142] analiza publicaciones realizadas en Instagram para detectar posibles casos de *bullying*, [76] plantea un método para la recomendación de canales personalizados en plataformas de *stream* desarrollando un caso de estudio concreto en Twitch, y [90] diseña un sistema para la detección e identificación de actividades físicas como abdominales o sentadillas mediante los acelerómetros de un *smartwatch*. Sin embargo, la mayoría de estos trabajos se centran en la solución de un problema concreto con unos datos de entrada predefinidos, homogéneos e invariables. Por este motivo, los sistemas diseñados en estas propuestas desarrollan soluciones *encorsetadas* a una problemática concreta, contribuyendo así a la dilapidación de la información.

Este trabajo doctoral propone un sistema flexible para el análisis de datos heterogéneos generados en sistemas de comunicación avanzada como los anteriormente descritos. En este sentido se hace indispensable el uso de técnicas de la IA que permitan analizar los datos generados y expedidos en las interacciones de los usuarios con otros sistemas. El objetivo de estos análisis va encaminado a la obtención de patrones y a la generación de información nueva de valor para los usuarios, independientemente del formato de los datos de partida. La adaptabilidad del sistema a distintas problemáticas y la

concurrencia de varios procesos completos de análisis para cubrir diferentes necesidades conlleva la utilización de técnicas muy dispares para el análisis de datos, lo que dota al sistema de un carácter híbrido.

1.1. Hipótesis y objetivos

Atendiendo a la problemática que plantea la ingente cantidad de datos de naturaleza heterogénea generados en sistemas de comunicación y tras el estudio de diferentes técnicas de la IA y su aplicación en el área del análisis de información multimedia, se formula la siguiente hipótesis:

> *La gran cantidad de información expedida en las numerosas interacciones entre consumidores de sistemas multiusuario, a pesar de la naturaleza heterogénea de los datos, puede ser utilizada para la generación de otro tipo de información útil para el usuario mediante el diseño y desarrollo de sistemas basados en inteligencia artificial.*

Por ello se plantea que es posible diseñar un sistema que sea capaz de analizar información multimedia proveniente de sistemas de usuario como las redes sociales con el propósito de generar otro tipo de información útil y de interés para el usuario.

Para poder validar esta hipótesis es necesario establecer un enfoque que permita abordar y afrontar la problemática que supone. Por ello, el objetivo principal de este trabajo es investigar en el diseño de un conjunto de servicios novedosos relacionados con el tratamiento y análisis de contenido multimedia y dirigidos a los usuarios de sistemas existentes, como las redes sociales. Este objetivo principal se divide, a su vez, en los siguientes objetivos específicos:

- Tratar con fuentes de datos externas y heterogéneas, que contengan información multimedia generada en un proceso de interacción entre los diferentes usuarios de un sistema.
- Estudiar las diversas técnicas de extracción de características en función de la naturaleza de los datos y de su adecuación tanto a la problemática general como a los casos de estudio particulares que se plantean en el trabajo.
- Extraer meta-información del contenido multimedia de tal forma que se obtengan las características necesarias para poder realizar diferentes análisis de minería de datos de manera transparente.
- Investigar en el diseño y desarrollo de servicios novedosos enmarcados en el área de la IA que permitan tratar datos complejos en beneficio de los usuarios.
- Realizar un proceso de selección, adaptación y optimización de los algoritmos de aprendizaje automático más idóneos con base en la problemática a resolver.
- Diseñar metodologías para el análisis de datos y la generación de información de interés que se adapten a la naturaleza de los datos y de los casos de estudio.

La consecución de los objetivos generales y específicos planteados permitirá diseñar, llevar a cabo y evaluar los resultados obtenidos en una serie de casos de estudio que a su vez servirán para contrastar la hipótesis de este trabajo. El análisis multimedia en los diferentes enfoques planteados en esta propuesta dará lugar a un proceso de transformación de la información. De esta manera, a partir de los datos expedidos en los sistemas de comunicación avanzada se podrá extraer conocimiento suficiente para generar nueva información de valor para el usuario.

1.2. Metodología de investigación

El procedimiento aplicado a lo largo de todo el proceso de investigación para el desarrollo del presente trabajo se ha fundamentado en la metodología *Action Research* [89]. Este proceso metodológico, basado en un enfoque práctico que pretende integrar la experimentación científica con la acción social, comienza con la detección de un problema, su identificación y contextualización en un área de estudio concreta y el planteamiento de una hipótesis a modo de meta u objetivo para dar solución al mismo. Posteriormente tiene lugar un proceso de recopilación y análisis de la información que desemboca en el diseño de una propuesta enfocada en solucionar el problema inicialmente detectado. La fase final consiste en la obtención y discusión de resultados y en la extracción de conclusiones.

Aunque las fases de la metodología se realizan de manera secuencial, las conclusiones obtenidas en la investigación desencadenan una acción que en ocasiones conlleva la modificación de la hipótesis. Por lo tanto, estas fases se realizan varias veces a lo largo de la investigación, dando lugar a un proceso cíclico de exploración, actuación y valoración de los resultados. La condición iterativa e incremental de la metodología *Action Research* facilita la comprensión del problema, la exploración práctica del mismo y la puesta en marcha de planes de mejora, además de que permite la optimización de los resultados de la investigación. Como consecuencia, la metodología persigue la obtención de un conocimiento y experiencia incremental en la materia analizada y la solución, de manera óptima, del problema planteado.

En este caso, dado que la propuesta describe un sistema híbrido que encapsula diferentes técnicas y pretende solucionar diversos planteamientos del problema detectado, la aplicación de la metodología elegida para este proce-

so de investigación favorece la retroalimentación del sistema para optimizar los resultados.

A fin de fomentar el acceso abierto y la educación científica, en el período de realización de este trabajo se ha asistido y participado en diversas conferencias, cursos, seminarios y congresos internacionales y se han realizado publicaciones de los avances realizados en la investigación. Adicionalmente, la cooperación con otras universidades también ha favorecido la transferencia del conocimiento y la investigación colaborativa, aportando otros enfoques de investigadores del área y enriqueciendo el trabajo realizado.

1.3. Organización de la memoria

Atendiendo a la hipótesis anteriormente formulada y con el propósito de satisfacer los objetivos establecidos como punto de partida del trabajo, la memoria se divide en dos partes. En la Parte I del documento se puede encontrar una descripción detallada del trabajo. Esta memoria está dividida en cinco capítulos, siendo este el primero de ellos.

El Capítulo 2 recoge los principales conceptos y técnicas que circunscriben este trabajo. En primer lugar, se realiza una presentación del concepto de contenido multimedia y se introduce el concepto de creatividad computacional, destacando la rama de la composición musical automática. Posteriormente se realiza un estudio de los algoritmos de aprendizaje automático existentes en la literatura con el objetivo de optimizar la selección en cada proceso de análisis del sistema y finalmente se estudian las diferentes técnicas existentes para la extracción de meta-información a partir de imágenes y sonido. Tras un proceso de revisión de las propuestas más influyentes en la materia, se analizan las necesidades de la sociedad pendientes de satisfacer

y se orienta la propuesta del trabajo en esa dirección.

Una vez analizadas las principales investigaciones en relación con la hipótesis de partida, y tras haber detectado las carencias y necesidades de los sistemas existentes, en el Capítulo 3 se desarrolla la propuesta de este trabajo. Concretamente, se plantea la arquitectura de un sistema heterogéneo capaz de englobar diferentes metodologías y técnicas para el análisis de contenido multimedia. Asimismo se describe detalladamente cada uno de los enfoques que se han considerado, constituyéndose cada uno de ellos como un marco de trabajo de esta propuesta.

En el Capítulo 4 se describen dos casos de estudio que se han llevado a cabo en entornos reales con el objetivo de validar el sistema previamente propuesto. El primer caso de estudio, recogido en la Sección 4.1, presenta un estudio para la composición automática de melodías descriptivas a partir de vídeos. La Sección 4.2 expone detalladamente el segundo caso de estudio, donde se aplican algunas técnicas de la IA para generar composiciones armónicas de manera dinámica mientras una persona realiza una ilustración digital.

El Capítulo 5 completa la primera parte de esta tesis doctoral. En él se resumen las principales contribuciones del trabajo, se presentan las conclusiones inferidas a lo largo del proceso de investigación y se discute el valor de la propuesta en relación con el cumplimiento de los objetivos inicialmente planteados. Adicionalmente, este capítulo esboza algunas líneas de trabajo futuro que podrán ser llevadas a cabo partiendo de esta investigación.

La Parte II contiene información adicional y complementaria para facilitar una mejor comprensión de investigación. Concretamente, con el objetivo de favorecer la aprehensión del trabajo a aquellas personas que no tengan una base de conocimiento musical previa, se desarrolla el Apéndice A. En

él se describen una serie de conceptos musicales básicos que se aplican en el trabajo y cuyo entendimiento es necesario para la correcta interpretación de esta tesis. Por otra parte, el Apéndice B reúne la información relativa a las encuestas que se han realizado a los usuarios en los casos de estudio. Cerrando este trabajo, el Apéndice C recoge varios ejemplos de composiciones musicales melódicas y armónicas obtenidas por los dos enfoques desarrollados en la propuesta. En cada ejemplo se puede consultar la imagen de partida y una transcripción de la música obtenida por el sistema.

En el próximo capítulo...

Una vez presentada la problemática con la que se va a lidiar en el presente trabajo de investigación, y formulados tanto la hipótesis de partida como los objetivos, en el próximo capítulo se recogen las contribuciones principales del estado del arte alineadas con las preocupaciones que nos ocupan. Concretamente, se introducen diversos conceptos relacionados con el contenido multimedia que son indispensables para la comprensión del trabajo y se presentan diferentes técnicas de extracción de meta-información a partir de imágenes y sonido. Adicionalmente se realiza una revisión de los diferentes algoritmos de aprendizaje automático presentes en la literatura y se analizan diversas técnicas para la composición automática de música enmarcadas en el área de la creatividad computacional.

Capítulo 2

Antecedentes

Resumen: Antes de profundizar en la propuesta que desarrolla este trabajo como respuesta a la problemática previamente planteada, el presente capítulo relaciona y describe los distintos formatos del contenido multimedia y las técnicas existentes para la extracción de meta-información que contienen, especialmente en los casos de imágenes y sonido. Adicionalmente se hace un estudio de los diferentes algoritmos de aprendizaje automático presentes en la literatura y se introduce el concepto de creatividad computacional, poniendo el énfasis en la composición musical automática. Finalmente, se revisan algunos de los trabajos más relevantes en la materia con el objetivo de orientar de manera óptima la propuesta de este trabajo. Este análisis es un paso necesario para diseñar el proceso de generación de información de relevancia para el usuario, tal y como definía la hipótesis.

Los avances del sector tecnológico, concretamente en las áreas de las TIC, el IOT y la IA, han dado lugar a numerosas herramientas que facilitan la comunicación entre las personas, independientemente de la distancia que las

separe [9]. Estos sistemas generan diariamente ingentes cantidades de datos de naturaleza muy heterogénea, lo que supone, por una parte, una fuente de datos muy útil para la extracción de conocimiento mediante técnicas de análisis basadas en IA, y por otra, un reto tecnológico por la dificultad de procesamiento y explotación que conllevan.

Los grandes volúmenes de datos generados en este tipo de sistemas junto con las gran diversidad de formatos que presentan y la pluralidad de fuentes de procedencia suscitan la necesidad de diseñar propuestas tecnológicas innovadoras que permitan avanzar en la materia y explotar los datos de manera eficiente. En este sentido, existen muchas propuestas en la literatura que abordan el problema del análisis de contenido multimedia desde diferentes perspectivas, y el objetivo de este capítulo es ofrecer una síntesis de la revisión de todas ellas.

Este capítulo comienza con una introducción al concepto de contenido multimedia en la Sección 2.1. A continuación, la Sección 2.2 presenta algunas técnicas utilizadas para el procesamiento y la transformación de la información multimedia basadas en creatividad y la Sección 2.3 se centra en algunos de los algoritmos de IA que facilitan el análisis de datos multimedia. La Sección 2.4 pormenoriza algunas de las técnicas existentes para la extracción de meta-información. Finalmente, en la Sección 2.5 se realiza un análisis de las propuestas más relevantes en la materia a fin de acotar, modelar y construir una solución al problema planteado en este trabajo.

2.1. Nuevas tecnologías de la comunicación y contenido multimedia

Los grandes avances en el área de la informática han facilitado el nacimiento y la evolución de las llamadas nuevas tecnologías de la información y de los medios audiovisuales. Por su parte, las nuevas tecnologías de la información se fundamentan en el avance tecnológico y están relacionadas con los conocimientos, procedimientos o instrumentos utilizados para la generación, tratamiento y difusión de la información verbal o icónica, independientemente de la naturaleza de su soporte. Por otro lado, los medios audiovisuales toman como referencia la utilización de herramientas como los proyectores y los magnetófonos, y facilitan la representación audiovisual y verboicónica de la información. Como resultado de la combinación entre la digitalización de la información y los avances tecnológicos que facilitan el desarrollo de los medios audiovisuales se obtienen una serie de progresos científicos que dan lugar a las nuevas tecnologías de la comunicación [84]. Estos resultados tecnológicos han tenido una acogida muy satisfactoria y han supuesto un gran impulso para muchas actividades necesarias en la sociedad, entre las que destacan el turismo, la educación o el desarrollo empresarial [23, 13, 10].

Al hablar de las nuevas tecnologías de la comunicación se hace imposible no pensar en el concepto multimedia. Etimológicamente, el término es redundante puesto que *media* significa, por sí mismo, "varios medios". La definición, según el Diccionario de la lengua española, es la siguiente: *que utiliza conjunta y simultáneamente diversos medios, como imágenes, sonidos y texto, en la transmisión de una información.* El concepto, sin embargo, se ha venido aplicando con significados y matices diversos, pudiendo encontrar en la literatura las siguientes acepciones:

- *La multimedia es el uso de texto, gráficos, animación, imágenes, vídeo y sonido para presentar información. Dado que estos medios pueden ahora integrarse utilizando una computadora, ha habido una explosión virtual de aplicaciones instructivas multimedia basadas en la computadora* [93].

- *Multimedia no es un producto, ni siquiera una tecnología. Se trata de una plataforma que combina elementos hardware y software para crear un entorno informativo multisensorial* [84].

- *Multimedia engloba una clase de sistemas de comunicación interactiva controlada por ordenador que crea, almacena, transmite y recupera redes de información textual, gráfica y auditiva* [100].

- *Los sistemas multimedia, en el sentido que hoy se da al término, son básicamente sistemas interactivos con múltiples códigos. Un aspecto clave en ellos es la integración de diferentes tipos de información soportada por diferentes códigos* [13].

A pesar de ser enfoques diferentes para definir un mismo concepto, todos ellos comparten que las tecnologías multimedia suponen la combinación de varios medios bien diferenciados. Estos medios pueden ser un conjunto de dispositivos interconectados entre sí o un conjunto de módulos que forman un único dispositivo. Sin embargo, en cualquier caso cada componente del todo está especializado en el procesamiento de un tipo de documento (textos, imágenes, vídeos, gráficos, animaciones, sonido...) La integración de todos estos tipos de datos da lugar a un documento audiovisual al que se puede denominar contenido multimedia [8]. Algunos autores afirman que el término más apropiado para referirse a este tipo de datos sería documento multilenguaje, ya que involucra los lenguajes verbal, visual, sonoro y audiovisual, o documentos multisensoriales porque su procesamiento requiere de la

utilización de varios sentidos [84]. De hecho, a estos materiales generalmente se les añade una característica de interactividad, que es la posibilidad de relación y respuesta mutua entre el usuario y el medio [14, 57]. Actualmente, la gran mayoría de dispositivos entre los que se encuentran los ordenadores, las tabletas inteligentes o incluso los teléfonos inteligentes aglutinan los medios suficientes para procesar individualmente todo tipo de documentos multimedia.

En 1995, Vaughan ya clasificaba los ámbitos de la aplicación multimedia en las siguientes categorías [132]:

- **Negocios**: necesaria para la creación y tratamiento de bases de datos, comunicaciones en red, presentaciones, tareas de marketing y publicidad y la gestión de las empresas en general.
- **Educación**: útil para el proceso de enseñanza adaptando los métodos tradicionales a la evolución tecnológica.
- **Hogar**: presente en dispositivos IoT relacionados con la domótica y las actividades de ocio.
- **Lugares públicos**: ventajoso para la adaptación tecnológica de hoteles, hospitales, estaciones de tren, centros comerciales, museos y tiendas.
- **Realidad Virtual**: lentes, cascos e interfaces especiales utilizadas para simular experiencias similares a la vida real.

En las últimas décadas, la utilización de sistemas multimedia se ha incrementado significativamente. [91] afirma que los niños y los jóvenes se mueven en un universo de dinamismo e inmediatez de continua estimulación y donde todo es simultáneo. Como consecuencia, surgen nuevos modos de percepción, relación de los jóvenes con la cultura popular, sociabilidad, y nuevas

dinámicas familiares que definen a una sociedad con sobreabundancia de información y datos: una sociedad multimedia [43].

De acuerdo con lo comentado previamente, el contenido multimedia incluye información de diversos tipos (textos, gráficos, sonidos, animaciones, videos, etc.) y los integra de manera coherente, incitándonos a utilizar varios de nuestros sentidos para su correcta y completa comprensión. Un ejemplo claro de este tipo de tecnología serían las aplicaciones de mensajería instantánea, que nacieron con el objetivo de transmitir mensajes textuales y actualmente permiten compartir información textual, visual, auditiva y audiovisual. En la misma línea podríamos destacar el avance de las redes sociales, que permiten compartir información de naturaleza heterogénea sin que apenas nos paremos a pensar sobre ello.

Como veremos a continuación, existen numerosas propuestas que permiten analizar el contenido multimedia con diferentes enfoques, pero con un objetivo común: explotar la información que compartimos por internet para generar conocimiento útil. Sin embargo, a los diferentes tipos de información que conforman el contenido multimedia y a las ingentes cantidades de esta información que se comparten diariamente por internet debemos añadir la gran variedad de formatos en los que se puede presentar un determinado contenido y las numerosas fuentes de datos de donde se puede extraer la información. Todos estos aspectos son problemas que dificultan la tarea de análisis del contenido multimedia.

2.2. Creatividad computacional

La meta de esta tesis doctoral es probar que los datos generados en sistemas de comunicación avanzada pueden ser explotados para generar informa-

ción nueva y de valor para el usuario a pesar de su naturaleza heterogénea. El proceso de análisis que se lleva a cabo para la transformación de los datos se va a realizar mediante la aplicación de IA; sin embargo, dentro de este marco, el proceso puede tomar diferentes direcciones en función del carácter de los resultados que se quieran obtener. Una de las posibles orientaciones del proceso de análisis está relacionada con la generación de resultados dotados de un componente artístico.

Muchas de las actividades llevadas a cabo por los seres humanos de manera cotidiana exigen cierta inteligencia: la comprensión del lenguaje, la extracción de patrones de comportamiento de la sociedad o incluso la conducción de un automóvil, entre otras muchas. A lo largo de las últimas décadas se ha avanzado en el diseño de sistemas informáticos que pueden desarrollar algunas de estas tareas, dando lugar a los llamados sistemas inteligentes [96]. Cuando la IA se aplica para que, además de simular habilidades cognitivas, las máquinas puedan desarrollar procesos creativos, los sistemas obtenidos se enmarcan en el área de la creatividad computacional [131].

Las creaciones artísticas comienzan con un motivo que impulsa al autor a llevar a cabo una idea; este factor puede ser una iniciativa personal, una fuente de inspiración o una combinación de ambas [125]. La inspiración es un estímulo externo capaz de evocar un estado de motivación que fomenta la creatividad [126]. Pero, ¿puede una máquina reaccionar a un estímulo externo para obtener motivación artística? Existe cierta controversia con la teoría de que una máquina pueda concebir ideas, y con ello, ser creativa [136].

Algunos autores como David Cope afirman que la dificultad de asimilar la inventiva de las máquinas reside en una definición tan estricta de lo que es la creatividad que ni siquiera un ser humano cumpliría los requisitos para poder considerarse creativo. Tras este análisis, el músico y científico

define la creatividad como el resultado de la combinación entre diferentes aspectos como la extracción de patrones, alusiones, inferencia y jerarquía de elementos de una rama artística. Como resultado, obtiene un modelo de asociación inductiva que asegura que soluciona el problema de la creatividad computacional en la creación automática de contenido artístico [35]. Por este motivo, los trabajos enmarcados en este área del conocimiento no tienen como único objetivo el estudio de la capacidad de las máquinas para generar contenido creativo, sino que también abarcan la evaluación de la obra de arte generada.

La creatividad computacional se aplica en muchas áreas relacionadas con creaciones artísticas, y como consecuencia existen diversas propuestas en la literatura que afrontan problemáticas muy diferentes. Para dar solución a un problema de creatividad visual, [2] diseña una propuesta híbrida que combina dos algoritmos de inteligencia de enjambre para esbozar dibujos a partir de una imagen de entrada. Por otra parte, en el contexto del entretenimiento digital, [33] y [34] afrontan la automatización del diseño de un videojuego en su totalidad, tanto en lo relativo a su historia como a los elementos que aparecen en él. En el ámbito de la lingüística, [37] aborda el problema de la composición de poemas en bengalí, un lenguaje rico en morfosintáctica y parcialmente fonético, buscando obtener calidad poética, una gramática correcta y un significado coherente en las creaciones.

Otra rama artística donde se aplica la creatividad computacional es la música. Existen numerosos trabajos que combinan la IA y la música con objeto de generar contenido artístico; [135] estudia el fenómeno de la expresividad en la interpretación, [83] propone un método para la construcción de modelos computacionales de interpretación expresiva para agrupaciones de músicos, y en [124] se diseña un sistema basado en aprendizaje automático para seleccionar entorno de interpretación óptimo para una composición mu-

sical. La Sección 2.2.1 se centra en analizar las técnicas de la IA con mayor relevancia en la composición musical mediante máquinas.

2.2.1. Composición musical automática

La convergencia entre la ciencia de la computación (concretamente en lo tocante a la rama de la IA), la psicología y la música ha dado lugar a numerosas propuestas que tratan de abordar el complejo problema de la automatización de la composición musical. A continuación se exponen algunas de las técnicas más explotadas y con mejores resultados en este dominio [79].

La formalización de la totalidad de los enunciados admitidos en un idioma mediante las gramáticas generativas de Noam Chomsky supuso un antes y un después en la teoría lingüística y la ciencia cognitiva [29]. Algunos autores, inspirados en esta teoría, comenzaron a trabajar en la generación de una gramática musical basándose en las similitudes entre la lingüística y la teoría musical [60, 74]. Ciertos trabajos más recientes enfocan estas técnicas a la generación automática de progresiones con base armónica [109] y al análisis armónico automático, que puede considerarse un paso previo a la composición automática [38]. Las gramáticas generativas, a día de hoy, no se aplican únicamente para la composición automática de música clásica, sino que se extienden a otros géneros musicales; concretamente [28] y [27] plantean respectivamente dos sistemas para la generación de secuencias de acordes y la improvisación de jazz.

Otra técnica ampliamente utilizada en este contexto son los modelos de Markov. Estos modelos, basados en cálculos estadísticos, modelan secuencias melódicas o armónicas definiendo una serie de estados (por ejemplo, el conjunto de las notas musicales) y calculando la probabilidad de transición entre los diferentes estados a partir de un conjunto de datos de partida [113]. De

esta manera se obtienen prediciones de tono y duración de los sonidos, dando lugar a una melodía o a una composición polifónica. Una de las propuestas más significativas en esta rama de la composición musical automática es *The Continuator*, un sistema interactivo basado en modelos de Markov para generar material musical con un estilo definido tratando diversos elementos musicales como el ritmo y la armonía [101].

Los algoritmos bioinspirados conforman una rama de la IA que emula el modo de procesar información y de resolver problemas de un ser vivo [17]. Dentro de esta rama se pueden distinguir diferentes técnicas como las redes neuronales (de las que se hablará más adelante), los sistemas inmunológicos artificiales o la inteligencia de enjambre. Por otra parte, en la década de 1970, Holland planteó la posibilidad de automatizar algunos procesos de adaptación y mecanismos naturales de supervivencia propios de los seres vivos para la resolución de problemas de optimización. Los procedimientos resultantes son conocidos como algoritmos genéticos y simulan la evolución por selección natural con el objetivo de identificar a los individuos "mejor adaptados" a la hora de realizar una determinada tarea [59]. Todos estos tipos de algoritmos bioinspirados también son técnicas recurrentes a la hora de diseñar e implementar sistemas compositores de música. Ejemplo de ello es [68], que propone un sistema híbrido que combina la optimización de enjambre de partículas o *particle swarm optimization* (PSO) con un algoritmo genético para la composición musical interactiva. Otros trabajos proponen una aplicación del algoritmo de optimización de colonia de hormigas –*ant colony optimization* en inglés– (ACO) para la composición musical; concretamente [50] aplica la técnica descrita para la generación de melodías y su posterior armonización con estilo barroco. Desde otra perspectiva pero en el mismo marco de los algoritmos bioinspirados, [94] se basa en un sistema inmunológico artificial para generar progresiones de acordes.

Por su parte, el aprendizaje profundo o *deep learning* ha supuesto un gran avance para el estudio de datos complejos. La arquitectura basada en capas de estos modelos computacionales da lugar a un proceso iterativo de análisis que facilita la representación de los datos en múltiples niveles de abstracción [73]. Dadas sus características, este tipo de métodos resuelve con éxito problemas en el campo del reconocimiento de voz, la visión artificial y la composición musical automática, entre otros. En materia de composición musical, concretamente, se han realizado numerosas propuestas aplicando diferentes arquitecturas de redes neuronales. [16] demuestra la viabilidad del uso de redes de creencia profunda (*deep belief networks*) para la composición automática de música; concretamente, la propuesta se basa en creación automatizada de improvisaciones jazzísticas. Desde otro ángulo, algunos autores aplican redes neuronales recurrentes *recurrent neural networks* (RNN) para la generación musical; sin embargo, la música compuesta con este tipo de sistemas a menudo carece de coherencia global. Una de las soluciones más comunes para este problema conlleva la utilización de bloques o celdas de memoria *Long-Short Term Memory* (LSTM) [53]. [1] y [45] plantean diferentes enfoques para la composición musical automática aplicando LSTM, dando lugar a creaciones musicales con una estructura bien formada y un estilo definido. Asimismo, las redes neuronales paralelas o *parallel neural networks* han cobrado mucho interés en la literatura por la alta eficiencia que supone la computación en paralelo [121]. En este sentido, [67] presenta una arquitectura específica de redes paralelas que permite desarrollar tareas de predicción y composición de música polifónica basadas en modelos probabilísticos. Dadas todas estas aplicaciones, otros autores optan por la combinación de varios de estos métodos dando lugar a sistemas híbridos como [51], que combina las redes de creencia profunda y las RNNs para la generación de música polifónica.

Todas las técnicas utilizadas para la composición musical automática se

incluyen bajo el paraguas de la IA; sin embargo, se enmarcan en ramas bien diferenciadas. Analizando la diversidad de algoritmos que se han abierto paso en este área de la creatividad computacional, es interesante comprobar que algunos de los algoritmos más presentes en la literatura para resolver tareas de minería de datos como la clasificación o la regresión no se consideran una opción para la solución de este tipo de problemas.

2.3. Algoritmos de aprendizaje automático

El análisis provechoso de la gran cantidad de datos generada en los sistemas de comunicación avanzada supone, simultáneamente, un problema y una necesidad. La minería de datos es un campo multidisciplinar que resulta de la convergencia entre las matemáticas, la estadística y la ciencia de la computación. Su objetivo es solucionar la dificultad del análisis de la información, y para ello proporciona herramientas útiles para el diseño de sistemas que extraen conocimiento de los datos [56]. Cuando los sistemas pueden aprender y adaptarse a los cambios que suceden en su entorno de manera autónoma evitando así que su diseñador deba prever y proporcionar soluciones para todos los problemas posibles, hablamos de aprendizaje automático o *machine learning* [6]. Así, un algoritmo enmarcado en esta rama de la IA consiste en la automatización de la identificación de patrones o tendencias en los datos.

Dentro del concepto de aprendizaje automático se pueden realizar numerosas categorizaciones de los algoritmos con diferentes criterios. En cualquier caso, la selección del tipo de algoritmo óptimo en cada caso debe ir supeditada al objetivo que se persiga en cada problema y a la naturaleza y la estructura de los datos. La literatura avala la adecuación de los algoritmos para la resolución de diferentes problemas: clasificación, predicción, agrupamiento, regresión... Una vez identificado el tipo de algoritmo más adecuado

en cada caso, muchos autores recomiendan probar y comparar el rendimiento de varios algoritmos sobre el mismo y sobre diferentes conjunto de datos [40] a fin de garantizar una selección óptima para la resolución del problema.

Una de las divisiones que se establece habitualmente para los algoritmos de aprendizaje automático teniendo en cuenta la naturaleza de los datos de entrenamiento permite diferenciar entre aprendizaje supervisado y no supervisado [112]. En la primera categoría, los algoritmos trabajan con datos etiquetados o valores numéricos; esto quiere decir que, de antemano, conocemos la etiqueta o clase o el valor de predicción que le corresponde a cada instancia del conjunto de entrenamiento. El objetivo, en este caso, es aprender de los datos utilizados en la fase de entrenamiento los patrones que presentan los datos para las etiquetas que se consideran y buscar dichos patrones en cada instancia de los datos de prueba para asignarle la etiqueta o el valor más adecuado. Estos algoritmos se utilizan comúnmente en problemas de regresión y clasificación. Por otra parte, el aprendizaje no supervisado consiste en el análisis de datos no etiquetados con el objetivo de realizar agrupaciones de las diferentes instancias basadas en sus similitudes. A partir de esta categorización de los algoritmos existen otras muchas técnicas de aprendizaje como el semi-supervisado y el aprendizaje por refuerzo.

El concepto de aprendizaje automático comprende una inmensa cantidad de técnicas y algoritmos que no es factible abarcar en su totalidad en este trabajo. Por ello, la revisión de la literatura, en este apartado, se enfoca de manera general sobre las técnicas más apropiadas para la resolución de los problemas que se van a abordar y de manera específica sobre los algoritmos que se van a aplicar en este trabajo. El objetivo del estudio es seleccionar los algoritmos más adecuados para modelar un criterio que permita establecer una relación entre la información visual y la información auditiva. Para ello, las secciones 2.3.1 y 2.3.2 abordan el problema de clasificación de una

instancia de datos para una única etiqueta y para varias, respectivamente.

2.3.1. Modelado en problemas de clasificación multiclase

La clasificación, como ya se ha explicado, se engloba dentro del aprendizaje supervisado y consiste en la evaluación de los patrones existentes en los datos con el objetivo de asignar una categoría a cada instancia [56]. Para ello, los algoritmos se dividen en dos fases: la primera, orientada al análisis de los datos de entrenamiento y donde se construye el modelo, y la segunda, donde se realiza la clasificación aplicando el modelo matemático generado en la fase anterior.

Cuando el problema trata de identificar si una instancia pertenece a una determinada categoría o no, estamos hablando de clasificación binaria. En este caso, para cada instancia de los datos, el algoritmo aplica el modelo determinando si los atributos o características son propios de la categoría evaluada (1) o no (0). Por extensión, la clasificación binaria permite distinguir dos categorías o clases diferentes; de esta forma, un problema de este tipo podría consistir en la identificación de tareas matutinas o vespertinas (si una tarea es clasificada como *no-matutina*, entonces será vespertina). De la misma forma, se podría considerar un problema de clasificación binaria la identificación de dos razas diferentes y excluyentes de perro. Sin embargo, en este caso el problema podría residir en la evaluación de una instancia que no se corresponde con ninguna de las dos razas consideradas en el estudio. En este caso sería más apropiado considerar tantas categorías (también excluyentes) como razas de perro se quieran considerar, y el estudio sería, por tanto, un problema de clasificación multiclase.

Las siguientes secciones detallan los aspectos más relevantes y la lógica de tres algoritmos válidos para la clasificación multiclase. La selección de *Ran-*

dom Forest, *Support Vector Machines* y *Naive Bayes* de entre los numerosos algoritmos que podrían haber sido útiles en este trabajo se fundamenta, por una parte, en su adecuación a la cantidad, naturaleza y dominio de los datos y al objetivo perseguido en el presente trabajo, y por otra, en su excelente reputación en problemas de clasificación con características similares en la literatura más actual.

2.3.1.1. *Random Forest*

En 2001, Breiman presentó la idea de bosque aleatorio o *Random Forest* (RF) [22], demostrando su buen rendimiento en comparación con otros clasificadores como las máquinas de vectores de soporte y las redes neuronales. RF es un grupo de árboles de decisión robustos al ruido y útiles para la resolución de problemas de clasificación y de regresión que se construyen a partir de la selección aleatoria de muestras de los datos de entrenamiento. Los árboles de decisión conforman un enfoque para el aprendizaje supervisado que genera una jerarquía de construcciones lógicas basada en la estructura de un árbol [5]. Atendiendo a los valores de los atributos utilizados para el entrenamiento, se construye, comenzando por la raíz, el árbol de decisión. Cada nodo interno representa una determinada característica o premisa, y cada nodo final —llamado hoja— representa una clase en un problema de clasificación. Así, cada rama, que es un conjunto de premisas que desembocan en una etiqueta, constituye una regla. Los árboles de decisión conforman una técnica muy potente que proporciona reglas de clasificación fácilmente interpretables por los humanos, aunque en algunos casos el coste computacional es relativamente alto. La utilización de este tipo de algoritmos se puede observar en diferentes áreas como la clasificación y extracción de textos, la comparación estadística de datos o la selección de genes.

De manera más específica, un RF es un *ensemble* compuesto por L clasificadores con estructura de árbol $h(X, \theta_n), N = \{1, 2, ..., L\}$, donde X representa el conjunto de datos de entrenamiento y θ_n son vectores aleatorios independientes con una distribución uniforme. Así, cada árbol de decisión se construye mediante una selección independiente, uniforme y aleatoria de los datos disponibles. La predicción se realiza mediante la agregación de las predicciones del conjunto: por voto mayoritario en los problemas de clasificación, y mediante el cálculo del promedio para los problemas de regresión.

Algunas de las ventajas del algoritmo incluyen su baja sensibilidad a la existencia de *outliers* en los datos, su capacidad para evitar y solucionar el problema de *overfitting* que además hacen que la poda no sea necesaria y la obtención automática de indicadores de la relevancia de cada uno de los atributos en la predicción [5]. Numerosos estudios demuestran, adicionalmente, la efectividad del algoritmo en conjuntos de datos con numerosas instancias, datos desbalanceados y con valores perdidos. El muestreo aleatorio y las estrategias de *ensemble* favorecen la consecución de un buen rendimiento con datos continuos, categóricos y binarios y también en problemas de clasificación multiclase.

2.3.1.2. *Support Vector Machines*

Las máquinas de vectores de soporte o *Support Vector Machines* (SVM) en inglés, son un conjunto de algoritmos de aprendizaje automático originalmente desarrollados por Vladimir Vapnik y su equipo de trabajo que han sufrido una gran evolución en la literatura [49]. Estos algoritmos tienen buen rendimiento con datos escasos o imprecisos, y son muy robustos en problemas binarios y con variables numéricas, por lo que se utilizan para la resolución de problemas de clasificación o regresión en áreas muy diversas [19].

En problemas de clasificación, SVM sitúa cada instancia de los datos con etiquetas binarias como un punto en un espacio con tantas dimensiones como atributos se consideren y busca un hiperplano que se encuentre a la máxima distancia entre las instancias etiquetadas para ambas clases. De esta manera, las instancias de cada clase quedan separadas por el hiperplano. Los puntos más cercanos al hiperplano de separación conforman un vector denominado vector de soporte.

La división más sencilla viene dada por una línea recta, un plano recto o un hiperplano N-dimensional. Sin embargo, SVM debe lidiar en numerosas ocasiones con varias dimensiones, curvas para la separación de las clases, datos en los que no es posible la separación total de las instancias por clases mediante hiperplanos y problemas multiclase. En estos casos donde no es posible una separación lineal, las SVM trabajan con funciones de *kernel*, que consisten en realizar automáticamente un mapeo no lineal en un espacio de dimensión superior. En estos casos, el hiperplano encontrado por el SVM en el espacio de características se corresponde con un límite de decisión no lineal en el espacio de entrada.

2.3.1.3. *Naive Bayes*

Los modelos Naive Bayes (NB) son una técnica de aprendizaje automático basada en el teorema de Bayes muy utilizada en problemas de clasificación [107]. El pilar de la lógica del algoritmo se fundamenta en la idea de la independencia de los atributos; es decir, que la presencia de un atributo en el conjunto de datos no conlleva la presencia de otro atributo.

NB se aplica teniendo en cuenta que cada instancia x se describe mediante un conjunto de valores para los atributos y donde la función objetivo $f(x)$ puede tomar cualquier valor del conjunto finito V [69]. Tras la fase de

entrenamiento para la consecución de la función objetivo se considera una nueva instancia, descrita por la tupla de valores de atributos $a_1, a_2, ..., a_n$. NB debe clasificar la instancia mediante la función objetivo. El enfoque bayesiano para clasificar la nueva instancia consiste en asignar el valor objetivo más probable v_{MAP} dados los valores de los atributos que describen dicha instancia, tal y como se puede observar en la Ecuación 2.1.

$$v_{MAP} = \underset{v_j \in V}{arg\,max}\, P(v_j|a_1, a_2, ..., a_n) \tag{2.1}$$

Aplicando el teorema de Bayes se obtiene la Ecuación 2.2.

$$v_{MAP} = \underset{v_j \in V}{arg\,max}\, \frac{P(a_1, a_2, ..., a_n)P(v_j)}{P(a_1, a_2, ..., a_n)} = \underset{v_j \in V}{arg\,max}\, P(a_1, a_2, ..., a_n)P(v_j) \tag{2.2}$$

Considerando que NB simplifica el problema de clasificación presuponiendo que los valores de los atributos son independientes dado el valor de la etiqueta, por lo que se llega a la Ecuación 2.3 donde v_{NB} denota el valor obtenido por el clasificador NB.

$$v_{NB} = \underset{v_j \in V}{arg\,max} \left(P(v_j) \prod_i P(a_i|v_j) \right) \tag{2.3}$$

NB supone una técnica sencilla pero a la vez muy potente para la resolución de problemas de clasificación binarios y multiclase. A menudo las estimaciones de probabilidad obtenidas son inexactas, sin embargo su rendimiento en problemas de clasificación está a la altura e incluso supera a otros algoritmos más sofisticados cuando es apropiada una presunción de independencia y cuando los conjuntos de datos tienen un escaso número de

instancias. El mejor rendimiento de NB se obtiene en dos casos extremos: cuando los atributos son completamente independientes y cuando los atributos son funcionalmente dependientes. En los casos intermedios, NB los resultados obtenidos por NB son sustancialmente más pobres.

2.3.2. Modelado en problemas de clasificación multi-etiqueta

A diferencia de los problemas de clasificación binarios o multiclase en los que cada instancia de los datos se corresponde con una única etiqueta del conjunto predefinido para la clasificación, los algoritmos de clasificación multi-etiqueta consisten en el análisis de patrones de los datos de manera que pueda inferirse más de una etiqueta simultánea para una instancia concreta [6]. Muestra de ello podría ser un problema de análisis de sentimiento de la música basado en clasificación en función del sentimiento que la composición despierta en el oyente. En este caso, las diferentes etiquetas serían los sentimientos que se consideraran en el estudio y el algoritmo podría determinar que una pieza musical puede producir simultáneamente alegría y calma o tristeza, molestia y excitación, por ejemplo.

De manera general, existen dos enfoques para el desarrollo de este tipo de algoritmos [6]. El primero de ellos, consiste en dividir o transformar el problema en varios problemas simples, aplicando k veces el algoritmo de clasificación binaria elegido sobre los datos y considerando iterativamente cada una de las k etiquetas predefinidas en el problema. De esta manera, una clasificación con valor 0 significaría que la instancia no se clasifica para la etiqueta evaluada, y una clasificación con valor 1 conllevaría que el algoritmo sí clasifica la instancia para la etiqueta evaluada. En el segundo enfoque se adaptan los algoritmos utilizados comúnmente en problemas de clasificación binaria y multiclase para poder lidiar con la simultaneidad de etiquetas.

De los numerosos algoritmos existentes en la literatura para la solución de problemas multi-etiqueta, las siguientes secciones se centran en detallar la lógica de los dos algoritmos más relevantes o apropiados para el objetivo de la tesis; a saber, *Random k-Labelsets* y *Multilabel k-Nearest Neighbors*.

2.3.2.1. *Random k-Labelsets*

Uno de los métodos de transformación de problemas multi-etiqueta en problemas de clasificación de etiqueta única más utilizados es *Label Powerset* (LP) [128]. El procedimiento que aplica para la transformación consiste en considerar cada conjunto de etiquetas único presente en el *dataset* multi-etiqueta de entrenamiento como una clase nueva. Como resultado, la tarea de clasificación etiqueta cada nueva instancia del *dataset* de prueba con la clase más probable, que es un conjunto de etiquetas.

LP consigue mejores resultados que otras aproximaciones computacionalmente más sencillas como *Binary Relevance*, que aprende un modelo binario para cada etiqueta de manera independiente al resto [139]. Adicionalmente, otra de las grandes ventajas de este algoritmo es que, al considerar cada conjunto de etiquetas como una clase nueva para el clasificador simple, se tiene en cuenta la correlación entre etiquetas. Sin embargo, presenta varias desventajas: en primer lugar, tiene un alto coste computacional; en segundo lugar, con frecuencia las clases compuestas por varias etiquetas y utilizadas para el clasificador simple suelen tener poca representación en los datos, y esto dificulta considerablemente la tarea de aprendizaje; por último y más importante, LP predice sólo conjuntos de etiquetas presentes en el *dataset* de entrenamiento.

Para afrontar estos problemas, *Random k-Labelsets* (RAKEL) construye un *ensemble* de LP donde el conjunto de etiquetas se divide aleatoriamente

en diferentes *labelsets* más pequeños (el tamaño viene determinado por el parámetro k), que son los valores de clase para el clasificador de una sola etiqueta de LP [129]. Una vez obtenidos los *labelsets* a partir del conjunto de entrenamiento, para la clasificación multi-etiqueta, RAKEL aplica el método LP y como consecuencia se reduce el coste computacional y el sesgo de la distribución de valores. Además, RAKEL mantiene la ventaja de considerar las correlaciones entre las clases.

Para la construcción de los *labelsets* con un tamaño reducido, los autores propusieron dos estrategias; en la primera de ellas los *labelsets* serán disjuntos (dos *labelsets* no compartirán la misma etiqueta) y en la segunda podrá existir solapamiento (dos o más *labelsets* pueden considerar una misma etiqueta). Para los conjuntos disjuntos, el conjunto C de todas las etiquetas de clase se divide, aleatoriamente, en m conjuntos disjuntos de etiquetas con tamaño k. En el segundo enfoque, donde se consideran las etiquetas superpuestas, los m conjuntos de etiquetas de tamaño k se muestrean aleatoriamente a partir de los conjuntos de etiquetas de tamaño k contenidos en C. El valor de k debe ser pequeño para evitar las debilidades de LP, por lo que los autores recomiendan utilizar un mayor número de modelos de clasificación para lograr un alto nivel de rendimiento predictivo [65]. Ambos enfoques de RAKEL obtienen mejores resultados que LP, especialmente en conjuntos de datos con muchas clases. Sin embargo, el enfoque con *labelsets* superpuestos obtiene un mejor rendimiento predictivo y permite corregir problemas de correlación mediante la agregación de muchas predicciones para cada etiqueta y la decisión de la etiqueta final por voto.

2.3.2.2. *Multilabel k-Nearest Neighbors*

Existen numerosos métodos para la solución de problemas de clasificación multi-etiqueta basados en el algoritmo *k-Nearest Neighbors* (kNN) [129]. Este algoritmo pertenece al llamado *lazy learning*, que es un aprendizaje automático donde el peso y el mayor coste computacional recaen en la fase de prueba (en este caso, clasificación) en lugar de hacerlo en la fase de entrenamiento. El primer paso de todas estas aproximaciones consiste en recuperar las k instancias más cercanas; sin embargo, la diferencia reside en el segundo paso: la obtención de los conjuntos de etiquetas.

El algoritmo *Multilabel k-Nearest Neighbors* (ML-kNN) es la primera aproximación para la clasificación multi-etiqueta basado en *lazy learning*, y como su propio nombre indica, deriva del popular kNN [140]. Como consecuencia, no se trata, como en el caso de LP, de un algoritmo que transforme el problema multi-etiqueta en varios problemas de etiqueta simple, sino que la adaptación del algoritmo aborda el problema multi-etiqueta en su totalidad. Los resultados experimentales demuestran que ML-kNN tiene mejor rendimiento que otros algoritmos relevantes en la clasificación multi-etiqueta como el BoosTexter, el AdaBoost y el Rank-SVM [140].

El flujo de ML-kNN se divide en dos fases bien diferenciadas. En la primera se identifican los k vecinos más cercanos de la instancia atendiendo a una medida de distancia concreta; en la segunda se obtienen las etiquetas de la instancia mediante un cálculo estadístico que atiende a las etiquetas de los k vecinos hallados en la primera fase del algoritmo.

Más concretamente, cada instancia x tiene conjunto de etiquetas asociadas $Y \subseteq \mathscr{Y}$. Así, el vector $\vec{y}_x$ representa las etiquetas de x de manera que cada componente $l : \vec{y}_x(l)(l \in \mathscr{Y})$ representa una de las etiquetas conside-

radas en el problema y toma valor 1 si $l \in Y$ y 0 en caso contrario. $N(x)$ representa el conjunto de los k vecinos más cercanos a x en el conjunto de entrenamiento. De esta manera, teniendo en cuenta los conjuntos de etiquetas de $N(x)$ se obtiene un vector de recuento de miembros definido como se expresa en la Ecuación 2.4, donde $\vec{C}_x(l)$ representa el número de vecinos de x pertenecientes a la clase l-ésima.

$$\vec{C}_x(l) = \sum_{a \in N(x)} \vec{y}_a(l), \quad l \in \mathscr{Y} \tag{2.4}$$

Para cada instancia de prueba t, el ML-kNN primero identifica sus vecinos más cercanos $N(t)$ en el conjunto de entrenamiento. H_1^l y H_0^l representan respectivamente los hechos de que t esté o no etiquetado con la etiqueta l, y que E_j^l $(j \in 0, 1, ..., k)$ representa el hecho de que, entre los k vecinos más cercanos de t hay exactamente j instancias que tienen la etiqueta l. Teniendo en cuenta el vector de recuento de miembros $\vec{C}_t$, el vector de etiquetas $\vec{y}_t$ se determina usando el principio del MAP definido en la Ecuación 2.5. Aplicando el teorema de Bayes, se obtiene la Ecuación 2.6.

$$\vec{y}_t(l) = \underset{b \in 0,1}{arg\ max}\, P(H_b^l | E^l_{\vec{C}_t(l)}), \quad l \in \mathscr{Y} \tag{2.5}$$

$$\vec{y}_t(l) = \underset{b \in 0,1}{arg\ max} \frac{P(H_b^l) P(E^l_{\vec{C}_t(l)} | H_b^l)}{P(E^l_{\vec{C}_t(l)})} = P(H_b^l) P(E^l_{\vec{C}_t(l)} | H_b^l) \tag{2.6}$$

Como consecuencia, el cálculo de las clases para la instancia t está basado en un análisis probabilístico que tiene en cuenta las etiquetas de los k vecinos más cercanos.

2.4. Extracción de meta-información del contenido multimedia

La minería de datos consiste en la automatización de la detección de patrones y tendencias con el objetivo de explicar y formalizar el comportamiento de los datos [56]. El conocimiento generado por los algoritmos de IA facilita y optimiza la toma de decisiones en las áreas en las que se aplica. Si bien el avance de la tecnología ha fomentado un acceso directo y sencillo a la información que circula por internet, la gran cantidad de fuentes de datos y la heterogeneidad de los mismos dificultan la tarea de recopilación de información representativa de un problema. Por ello, una de las fases más complejas y a la vez más importantes en un proceso de análisis es, precisamente, la obtención de un conjunto de datos apropiado para extraer el conocimiento necesario y cumplir los objetivos del estudio.

La selección de un algoritmo de aprendizaje automático y el diseño del proceso de análisis en general está condicionado por el objetivo que se persigue en el estudio y por la naturaleza de los datos que se quieren analizar [6]. En el contexto del análisis de contenido multimedia se pueden diferenciar dos enfoques diferentes para la creación de un conjunto de datos que represente correctamente el problema. Por una parte, se puede distinguir un enfoque macroscópico donde el *input* es el conjunto de datos formado por los archivos multimedia, como sucede en el caso de las redes neuronales convolucionales (CNN); desde otro punto de vista, existe otro tipo de análisis con un enfoque microscópico donde el análisis de datos se realiza sobre meta-información extraída a partir de los archivos multimedia que son objeto de análisis. Sin embargo, teniendo en cuenta que en el primer enfoque los algoritmos que procesan los ficheros realizan una extracción de las características que definen al contenido multimedia a modo de caja negra como paso previo al análi-

sis, se podría decir que en ambos casos, el análisis de contenido multimedia se reduce a la extracción de meta-información y su posterior procesamiento para la generación de conocimiento útil.

Las técnicas de extracción son específicas para cada formato de fichero multimedia. Esta sección realiza una revisión de las técnicas de extracción de meta-información a partir de contenido multimedia más relevantes en la literatura. Debido a la gran cantidad de formatos existentes y teniendo en cuenta el problema que se plantea en este trabajo, el contenido se centra en la extracción de meta-datos a partir de imágenes y sonido. El objetivo del estudio de estas técnicas va enfocado a la generación de diversos conjuntos de datos que permitan resolver el problema concreto que nos planteamos para dar respuesta a la hipótesis de esta tesis doctoral.

2.4.1. Descriptores de imagen

Esta sección realiza un estudio de las técnicas más destacadas para la obtención de meta-información a partir de una imagen. En otro orden de ideas y analizando el marco teórico que establece Vaughan en [132], se puede considerar que un vídeo es una secuencia de imágenes denominadas fotogramas que se suceden con una frecuencia (fotogramas por segundo o FPS) lo suficientemente alta como para que el ojo humano pueda percibir el movimiento. Teniendo esto en cuenta, la extracción de meta-información de un vídeo se puede reducir a la extracción de descriptores de los fotogramas o imágenes que lo componen.

En lo relativo a la obtención de descriptores gráficos de las imágenes, la literatura recoge una serie de técnicas que se pueden agrupar en función del tipo de información que se extrae. Así, la Sección 2.4.1.1 detalla las técnicas más relevantes para la extracción de información cromática, la Sección 2.4.1.2

recoge algunas de las técnicas más utilizadas para la detección de formas y disposición de elementos en la imagen y en la Sección 2.4.1.3 se describen otras técnicas que obtienen descriptores de la imagen de alto nivel.

2.4.1.1. Información cromática

El color es una de las características visuales con mayor carga expresiva, y como consecuencia se ha convertido en un foco de estudio en el área de recuperación de imágenes basada en contenido, sistemas multimedia y bibliotecas digitales a lo largo de las últimas décadas [81]. Sin embargo, existen una serie de factores psicológicos, circunstanciales, ambientales y fisiológicos que atribuyen a la percepción del color un cariz subjetivo. Esto desencadena una dificultad en la encapsulación y representación del color en el contenido multimedia.

La digitalización del color conlleva un proceso de formalización para su representación numérica mediante expresiones matemáticas para que, de esta manera, se pueda visualizar en dispositivos tecnológicos de manera muy precisa. La formalización cromática se puede realizar de numerosas formas distintas teniendo en cuenta la información que se quiere representar. Así, los modelos de color son los diferentes sistemas de intepretación cromática [64]. El modelo RGB toma su nombre de las siglas de los tres colores primarios en la luz: rojo, verde y azul. La representación del color, en este caso se representa como una mezcla regulable de estos tres colores. El modelo CMY es similar al RGB, pero considera los colores primarios en pigmento: cian, magenta y amarillo. Por otra parte, la familia de modelos HSI definen el color en función de su tono (*hue*), su saturación (*saturation*) y su intensidad (*intensity*), y están basados en el sistema de visión humana. En cuanto a la familia de modelos YUV, el color se representa en función de la luminancia

y la cromaticidad. La selección del modelo de color para una determinada aplicación es una tarea compleja que depende de las propiedades del modelo y de la naturaleza del problema.

En muchos casos, otro de los problemas de la digitalización del color es la limitación de la gama de colores que pueden mostrar los dispositivos [99]. Como solución se aplican técnicas para la cuantificación del color que permiten reducir el rango de valores cromático de una imagen dando lugar a una compresión con pérdida [26]. Este concepto es uno de los pilares de la extracción de información cromática, aplicándose, por ejemplo, para la obtención de paletas de colores como se verá a continuación.

El grupo *Moving Picture Experts Group* (MPEG) formado por la Organización Internacional de la Estandarización (ISO) y la Comisión Electrotécnica Internacional (IEC) tiene como objetivo el desarrollo de una representación estándar para la información audiovisual que englobe la descripción de sus metadatos. El estándar *Multimedia Content Description Interface* más conocido como MPEG-7, define numerosos procedimientos para la extracción de meta-información entre los que se incluyen algunos para la extracción del color [82]. A continuación se describen los descriptores de color más relevantes según el MPEG-7.

Histograma de color. Los histogramas de color obtienen información sobre la distribución del color en la imagen. Los descriptores de color originados por el análisis de histogramas han jugado un papel central en el desarrollo de los descriptores visuales en MPEG-7, por lo que son una de las técnicas más frecuentes para la extracción de información cromática [81]. Los histogramas son muy adecuados para la representación global del color en una imagen, sin embargo, su alta dimensionalidad y su dependencia con el modelo del color, con la cuantificación en el espacio de color y con la cuantificación en

bins de los valores son cuestiones a tener en cuenta de cara a su aplicación [55]. La Figura 2.1 representa el histograma extraído para el modelo de color RGB teniendo en cuenta 255 bins para cada uno de los colores rojo, verde y azul.

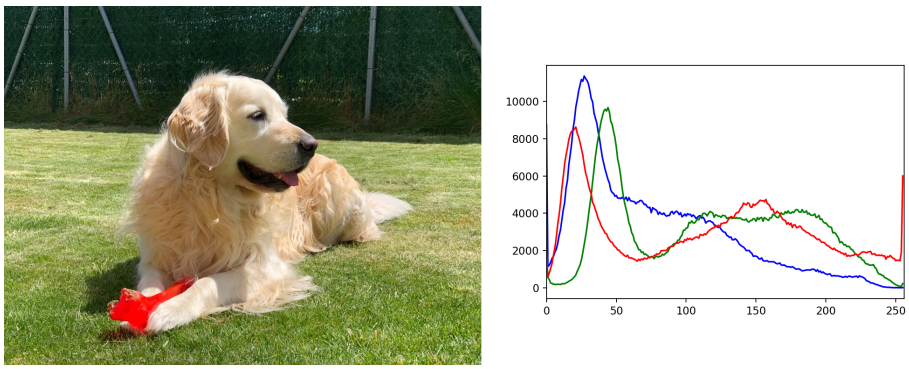

Figura 2.1: Obtención del histograma de color en el modelo RGB de una imagen

Dominant color descriptor. El descriptor de color dominante obtiene la distribución de los colores más destacados de la imagen [81]. La limitación de este descriptor está relacionada con la selección del modelo de color y la cuantificación del espacio de color. El resultado que se obtiene es una reducción de la información cromática de la imagen a un pequeño número de colores representativos [114]. Aunque existen técnicas basadas en histogramas de color para la obtención de paletas cromáticas que describen una imagen [39], la mayor parte de las propuestas se basan en el color dominante. Uno de los ejemplos más ilustrativos podría ser [41], que facilita la obtención dinámica de una paleta de colores y del color dominante de una imagen mediante una técnica de cuantificación. En la Figura 2.2 se muestra un ejemplo de una imagen en la que se extrae el color dominante y una paleta de colores dominantes a partir de una imagen mediante la herramienta Color Thief.

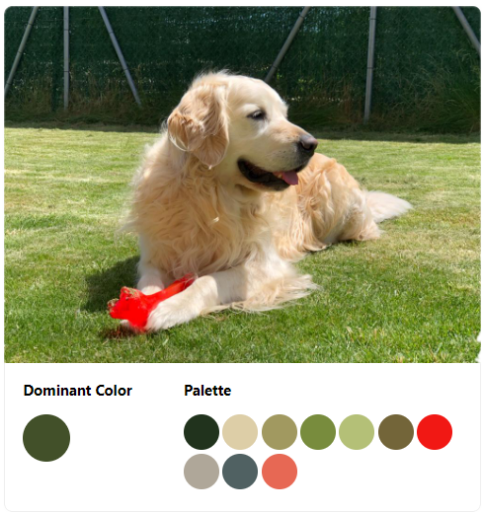

Figura 2.2: Obtención del color dominante y de una paleta de colores dominantes mediante la herramienta Color Thief

Layout color descriptor. Este descriptor captura la disposición espacial de los colores dominantes en una cuadrícula superpuesta en la región de interés o en la imagen completa [81]. Se trata de un descriptor muy compacto y efectivo en aplicaciones para la comparación de imágenes fijas y segmentos de vídeo cuya obtención es equivalente a la división de una imagen en una cuadrícula y la posterior obtención de los colores dominantes de cada una de las regiones obtenidas. La Figura 2.3 representa la información extraída con este descriptor del color teniendo en cuenta una cuadrícula de 3x4.

Figura 2.3: Representación del *layout color descriptor* para una imagen con una cuadrícula de 3 filas por 4 columnas

2.4.1.2. Descriptores de forma y disposición de los elementos

La extracción de características gráficas a partir de las imágenes es una tarea muy importante en los campos de visión artificial y robótica, y por consecuencia los descriptores de características son la base de cualquier sistema de reconocimiento de imágenes [70]. El diseño de nuevos descriptores y la optimización de las propuestas existentes en la literatura conforman una fuente de inagotable de estudios de investigación. Las imágenes se representan por medio de su descriptor, y en problemas de reconocimiento de imágenes la comparación no se realiza con los píxeles de la imagen sino con estos atributos que la describen.

La extracción de los descriptores de imagen se puede realizar mediante diferentes tipos de algoritmos cuyo objetivo es doble: la precisión y la eficiencia. La mayoría de los descriptores se basan en características locales de diferentes subzonas de la imagen descritas por los términos *keypoint* (puntos de interés de la imagen), *feature descriptor* (representación de los descriptores de una subzona concreta de la imagen) e *image descriptor* (representación completa de la imagen) [70]. A continuación se presentan algunos de los algoritmos más destacados para la extracción de este tipo de meta-información de la imagen.

***Scale-Invariant Feature Transform* (SIFT)**. SIFT es un algoritmo para la extracción de características de las imágenes [80]. En una primera etapa, el algoritmo obtiene los *keypoints* de la imagen y les asigna una orientación. Posteriormente, para cada punto de interés se obtiene un descriptor que incluye información relacionada con el espacio que le rodea. SIFT se considera como una de las opciones con mayor calidad debido a la distintividad e invariabililidad que consigue en una gran variedad de transformaciones comunes de imágenes tales como rotaciones y escalado [75]. Además de ser uno de los

algoritmos más utilizados, se ha utilizado como base de muchas propuestas posteriores.

***Speeded-Up Robust Features* (SURF)**. Se trata de un detector y descriptor invariante a escala y rotación especialmente destacable en cuanto a robustez, repetibilidad y distintividad [15]. El motivo de la reducción del coste computacional de SURF reside en el uso de imágenes integrales, la utilización de una medida basada en la matriz de Hesse y la simplificación de los métodos existentes.

***KAZE* y *Accelerated KAZE* (A-KAZE)**. Los algoritmos KAZE [3] y A-KAZE [4] fueron propuestos por Alcantarilla en 2012 y 2013 respectivamente. Ambos explotan el espacio a escala no lineal mediante el filtrado de difusión no lineal, sin embargo, la principal diferencia reside en el método para la construcción de dichos espacios. A-KAZE aplica un marco computacional llamado *Fast Explicit Diffusion* (FED) incrustado en un enfoque piramidal que acelera significativamente los cálculos. La detección en ambos casos se realiza mediante la matriz de Hesse, y los descriptores obtenidos son invariables a escalas y rotación.

***Binary Robust Independent Elementary Features* (BRIEF)**. BRIEF es un descriptor binario basado en tests de intensidad entre píxeles de diferentes regiones de la imagen [25]. Si bien el algoritmo tiene como principal objetivo la reducción del tiempo de computación, la precisión no queda relegada a un segundo plano.

***Oriented FAST and Rotated BRIEF* (ORB)**. En 2011, se introducía el algoritmo ORB [110] como una combinación de *Features from Accelerated Segment Test* y BRIEF [25]. La detección de *keypoints* se realiza, en este caso, mediante el algoritmo FAST. Sin embargo, los resultados se optimizan mediante la aplicación de la detección de esquinas de Harris, obteniendo

los N *keypoints* de mayor calidad y descartando el resto. A continuación, el algoritmo aplica una pirámide a escala de la imagen para la generación de atributos FAST para cada nivel establecido. Para la obtención de los descriptores de imagen, ORB aplica una modificación de BRIEF basada en la dirección de los *keypoints* que soluciona la inestabilidad a la rotación.

***Binary Robust Invariant Scalable Keypoints* (BRISK)**. BRISK se presenta como un método novedoso para la detección, descripción y comparación de *keypoints* de imágenes [75]. El algoritmo obtiene unos descriptores considerablemente invariantes a rotación y escala a un bajo coste computacional. BRISK obtiene los *keypoints* aplicando el algoritmo FAST sobre un espacio de escalas generado a partir de la imagen. Posteriormente identifica la dirección característica de cada *keypoint* para la obtención de los descriptores invariables a la rotación.

La selección del mejor algoritmo para la extracción de descriptores de la imagen no es trivial; es más, es una decisión crítica en la mayoría de sistemas de visión artificial. [122] realiza un análisis comparativo muy detallado de los algoritmos SIFT, SURF, KAZE, AKAZE, ORB y BRISK desde dos puntos de vista: por una parte, se estudia la eficiencia de cada uno de los algoritmos, y por otra, se valora qué algoritmo es más invariable a escala, rotación y puntos de vista. Los resultados de este estudio ponen de manifiesto las fortalezas y debilidades de cada uno de los algoritmos. Los algoritmos más precisos son SIFT y BRISK y el más eficiente es ORB. Como consecuencia, la selección del algoritmo óptimo para la extracción de descriptores de imagen debe realizarse en función de los objetivos de cada problema.

2.4.1.3. Descriptores de alto nivel

El *deep learning* es una rama de la IA enfocada en el aprendizaje automático. Los algoritmos enmarcados en este área constan de una arquitectura basada en capas que a su vez están compuestas por unidades de procesamiento denominadas neuronas artificiales, especializadas en la detección de características de los objetos percibidos [73]. La concatenación de diferentes capas de procesamiento dota a estos algoritmos de un gran poder para extraer múltiples niveles de abstracción de los datos, por lo que su utilización en problemas de visión artificial está muy extendida.

En otro orden de ideas, el *transfer learning* (TL) es una técnica emergente en el ámbito del aprendizaje automático que favorece la mejora del aprendizaje mediante la transferencia de conocimientos entre tareas relacionadas [102]. Si bien el diseño de la mayor parte de los algoritmos de aprendizaje automático está enfocado en la solución de una tarea particular, hay tareas que comparten parcialmente objetivos o fases del análisis. En estos casos, el aprendizaje por transferencia presenta numerosas ventajas, como la simplificación en la resolución de ciertos problemas, la reducción del tiempo de computación o el aumento del rendimiento de los algoritmos.

En el caso de las redes neuronales, la obtención de altas precisiones en problemas de clasificación de imágenes conlleva el procesamiento de grandes conjuntos de datos etiquetados y, en consecuencia, un largo tiempo de entrenamiento [52]. Sin embargo, la adquisición de estos conjuntos de datos en muchos casos es un problema determinante. La utilización de modelos preentrenados con conjuntos de datos de imágenes a gran escala como ImageNet [111] tales como VGG-16, Inception o GoogLeNet facilita en gran medida la tarea de clasificación de imágenes. El TL supone mejoras en el problema de clasificación a pesar de las posibles diferencias entre los dominios y las

etiquetas de los datos de entrenamiento y clasificación. En este caso, la fortaleza del TL reside en la capacidad del modelo pre-entrenado para extraer características diferenciadoras de las imágenes. Dado que las últimas capas de la red neuronal son las encargadas del proceso de clasificación y las capas anteriores realizan la extracción de los descriptores de las imágenes, la eliminación de las capas de clasificación conlleva la obtención de un vector de características de alto nivel para cada imagen procesada.

De esta manera, la aplicación del TL a problemas de *deep learning* y la obtención de los vectores de características por medio de las últimas capas de la red neuronal se puede considerar una técnica útil para la extracción de meta-información de un conjunto de datos formado por imágenes. Sin embargo, es importante comprender que los descriptores que se obtienen representan una abstracción de los datos, por lo que podemos considerar que los atributos no son fácilmente comprensibles por un humano, sino que representan información de alto nivel.

2.4.2. Características del sonido

El análisis automatizado de música se enfoca, en numerosas ocasiones, como un análisis de contenido de audio, y este problema requiere de técnicas y herramientas fiables y versátiles para su correcta resolución [130]. La extracción de meta-información del sonido es un área de investigación en continuo crecimiento donde confluyen diferentes disciplinas como la musicología, la ciencia de la computación, la teoría musical, la física y las telecomunicaciones [88]. Este área, conocida como *Music Information Retrieval* (MIR), engloba técnicas para la extracción de contenido de audio de siete facetas musicales diferentes: tono, tiempo, armonía, timbre, editorial, textual y bibliográfico [44]. Para el desarrollo de la presente tesis doctoral no

se consideran aspectos relacionados con el ritmo o el tempo, ni tampoco con el timbre, la letra o la información bibliográfica de una composición, por lo que la revisión de descriptores se centrará en las facetas de tono y armonía.

Uno de los problemas iniciales para la extracción de características del sonido relacionadas con tono y armonía es la representación digital de las notas musicales. En algunos casos, la representación de estas notas se basa en el cifrado americano, donde a cada nota se le asigna una letra. Sin embargo, en este caso las diferentes escalas, y por tanto las diferentes alturas para una misma nota, quedan reducidas a una misma letra. Para solucionar este y otros problemas relacionados con la representación digital de la música nacen algunos estándares como el MIDI (*Musical Instrument Digital Interface*) [47]. Se trata de un estándar basado en eventos que permite representar características tonales, temporales, agógicas y dinámicas de las notas musicales de manera numérica.

La representación del sonido en cualquier composición musical viene dada por una onda compleja, es decir, por la composición de varias ondas simples. Fourier propuso un método para la descomposición de series temporales complejas en un conjunto de ondas simples sinusoidales [20]. El paso del dominio del tiempo al dominio de la frecuencia facilita la obtención de las notas musicales contenidas en una onda ya que la onda simple con menor frecuencia se denomina frecuencia fundamental, y se corresponde con un tono o nota musical. La introducción del algoritmo de la transformada rápida de Fourier (FFT) ha ampliado considerablemente el ámbito de aplicación de la transformada de Fourier al análisis de datos y a la representación digital del sonido en general. Por este motivo, el análisis de Fourier ha sido ampliamente utilizado en tareas de MIR tal y como se verá a continuación.

Tomando el trabajo de Fourier como base, existen diferentes propuestas

que obtienen la intensidad de cada una de las notas musicales de la escala cromática para abordar problemas relacionados con la armonía. Uno de los descriptores más relevantes en la literatura son los vectores de chroma, que son vectores con 12 elementos donde cada uno de ellos representa el valor normalizado para la intensidad de cada nota en un fragmento musical [46]. En esta misma dirección, los descriptores CENS (*Chroma Energy Normalized Statistics*) se obtienen de la normalización de los vectores de croma y la aplicación de una cuantificación dado un determinado valor umbral, expresando así una distribución relativa de la energía que permite una extracción más robusta y eficiente [92].

Desde otra perspectiva, otro de los descriptores más comunes del sonido son los *Mel-Frequency Cepstral Coefficients* (MFCCs) [104]. Su obtención comienza con el muestreo de la onda de sonido para dividir la señal en varios *frames*. El muestreo se obtiene generalmente aplicando una función de ventana a intervalos fijos —típicamente una ventana de Hamming—. Posteriormente, a cada *frame* se le aplica la FFT y se conservan los valores del logaritmo del espectro de amplitud. Tras suavizar el espectro y enfatizar las frecuencias perceptualmente significativas de acuerdo con la escala de Mel. Dicha escala se basa en que el sistema auditivo humano no percibe el tono de manera lineal, y propone un mapeo entre la frecuencia real del sonido y el tono percibido para solucionar el problema. Finalmente se aplica una transformación mediante la transformada discreta del coseno, y como consecuencia se obtiene un vector de características cepstrales para cada *frame* de la señal muestreada. Estos descriptores del sonido son las características dominantes en tareas de reconocimiento de voz; sin embargo, la información que proporcionan los MFCCs no es óptima para el análisis y el modelado musical [78].

Partiendo de la base de los MFCCs, [143] introduce los LFCCs (*Linear*

Frequency Cepstral Coefficients). Estos descriptores también parten de un muestreo de la señal en diferentes ventanas o *frames*. Para cada una de dichas ventanas se aplica la FFT con el objetivo de traducir la señal del dominio del tiempo al dominio de la frecuencia. En este caso, los filtros aplicados a la señal posteriormente son lineales. Finalmente, se aplica la transformada discreta del coseno igual que en el caso anterior. Estos descriptores se aplican especialmente en tareas de reconocimiento de voz.

Considerando algunas de las técnicas y descriptores anteriormente citados, algunos autores han perseverado en el desarrollo de herramientas útiles en el área de MIR. [18] presenta Essentia 2.0, una biblioteca de código abierto desarrollada en C++ para el análisis de audio y la recuperación de información musical basada en audio. Essentia permite obtener descriptores en el dominio del tiempo y de la frecuencia así como descriptores tonales tonales y rítmicos y otros atributos de alto nivel. En [130] se presenta MARSYSAS (*MusicAl Research SYstem for Analysis and Synthesis*), un *framework* implementado en C++ para la experimentación, evaluación e integración de diversas técnicas interactivas para el análisis de audio. En la misma dirección, LibROSA es una biblioteca desarrollada en Python para el análisis de música y audio que proporciona la funcionalidad necesaria para los sistemas de MIR [88].

2.5. Conclusión de la revisión de antecedentes

Los grandes progresos tecnológicos aplicados a los sistemas de comunicación avanzada han favorecido la combinación de información heterogénea; así, la mayor parte de los dispositivos tecnológicos pueden procesar individualmente textos, imágenes, vídeos, gráficos, animaciones, sonido y otros tipos de datos. El auge de estos sistemas conlleva la generación y expedición

diaria de grandes cantidades de contenido multimedia. Existen numerosas propuestas enfocadas al análisis de estos tipos de datos, pero el procesamiento de cada una de ellas se centra en un tipo de dato y un proceso de análisis muy concretos, favoreciendo así al desaprovechamiento de gran parte del contenido multimedia despachado en internet.

En consideración a la composición musical automática, son muchas las técnicas recogidas en la literatura con este fin. Sin embargo, resulta interesante analizar el motivo por el que algunos algoritmos de aprendizaje automático como los árboles de decisión o las máquinas de vectores de soporte no se contemplan como opción en este área. Este tipo de algoritmos se utilizan comúnmente para solucionar problemas de clasificación, por lo que relacionan de alguna manera atributos descriptivos con el atributo a predecir o clasificar, conocido como clase. La dificultad de utilizar estos algoritmos para la composición automática viene desencadenada por la identificación de los atributos descriptivos y, como consecuencia, por la definición de un criterio para realizar la clasificación. En este sentido, una posible solución podría ser el apoyo en la música programática y la música descriptiva. Ambos estilos de música se basan en la imitación de sonidos de la naturaleza y en la analogía, mediante información audible, de la información visible [95]. La relación que la música descriptiva establece entre lo visual y lo sonoro podría utilizarse como criterio para la composición musical automática mediante algoritmos como los previamente mencionados.

Por otra parte, el análisis de contenido multimedia conlleva, en muchos casos, la extracción de meta-información. La selección de la información a extraer a partir de los datos de partida y la técnica a utilizar para la extracción se realizará con base en algunos factores como la naturaleza de cada problema, el objetivo del estudio, el tipo de datos y la técnica o algoritmo que se vaya a utilizar para el análisis. No existe una técnica óptima de ex-

tracción para cada tipo de dato, sino que la selección debe ir supeditada a las características de cada problema.

En conclusión, las propuestas realizadas para el análisis de contenido multimedia solucionan problemas individuales y muy específicos, desperdiciando así una gran cantidad de datos que podrían resultar útiles para la generación de nuevo conocimiento. En este sentido, se hace ineludible la necesidad de diseñar una arquitectura flexible que soporte la coexistencia de varios procesos de análisis que generen datos útiles para los usuarios a partir de la información que comparten en diversos sistemas de comunicación avanzada como las redes sociales. Dicha aglutinación de propuestas conlleva la aplicación de diversas técnicas, lo que da lugar a un único sistema híbrido que favorezca la explotación de múltiples datos mediante diferentes procesos de análisis de manera centralizada y eficiente.

En el próximo capítulo...

Tras analizar diversas propuestas de la literatura actual en relación con la problemática que nos ocupa y detectar ciertas carencias y necesidades, en el próximo capítulo se describe detalladamente la propuesta diseñada para dar respuesta a la hipótesis establecida y llevar a buen término los objetivos fijados en la fase inicial de la investigación.

Capítulo 3

Propuesta

Resumen: *Tras analizar los diversos enfoques existentes en la literatura actual que tratan de dar respuesta a la problemática planteada y detectar diversos problemas y necesidades que plantean, este capítulo detalla la propuesta del presente trabajo. El planteamiento ha sido diseñado para contrastar la hipótesis, en la que se sugería que la información generada en la interacción entre consumidores de sistemas multiusuario, a pesar de su naturaleza heterogénea, puede dar lugar a otro tipo de información de valor mediante su análisis con técnicas de la inteligencia artificial. De manera global, se describe un sistema híbrido inteligente para el análisis de contenido multimedia; de manera particular, los dos enfoques que se han desarrollado y que dan respuesta a diferentes carencias detectadas en el campo de estudio.*

La utilización de sistemas de comunicación avanzada ha crecido exponencialmente en los últimos años. El mayor ejemplo de ello son las redes sociales, que han pasado de ser utilizadas por un 7 % de la población en el año 2005 a contar con un número de usuarios correspondiente al 63 % de la población

en 2015 [105]. En este tipo de sistemas se genera una inmensa cantidad de información a lo largo de un día. Este trabajo parte de la hipótesis de que estos datos, a pesar de su naturaleza heterogénea, pueden ser analizados y utilizados para la generación de otro tipo de información de valor mediante técnicas de la IA. Tras un proceso de estudio de la problemática y un análisis de la literatura existente, este capítulo describe un sistema cuyo objetivo es obtener información de fuentes de datos heterogéneas, extraer metadatos que los definan y aplicar técnicas de la IA que faciliten la generación de información nueva y útil para el usuario.

La estructura de este capítulo incluye dos secciones: por una parte, la Sección 3.1 define la arquitectura del sistema, haciendo distinción entre los diferentes módulos que lo componen y especificando la funcionalidad de cada uno de ellos; por otra parte, en la Sección 3.2 se presentan dos marcos de trabajo que, haciendo uso de la arquitectura propuesta, ponen de manifiesto distintas aplicaciones del sistema.

3.1. Arquitectura del sistema híbrido inteligente

Tras analizar la problemática y las necesidades que debe cubrir el sistema, se llega a la conclusión de que se puede definir un proceso genérico de análisis y transformación de datos que se adapte a los diversos escenarios y dé solución a los diferentes problemas a los que se enfrenta el sistema. Este proceso consta de tres fases bien diferenciadas: la extracción de meta-información a partir del contenido publicado por los usuarios, el análisis de la misma mediante técnicas de la IA y la preparación y presentación de los resultados obtenidos al usuario.

Salvando las diferencias, este marco de transformación de la información

se asemeja mucho al conocido proceso de Extracción-Transformación-Carga o *Extract-Transform-Load* (ETL), que permite la obtención de información de diversas fuentes de naturaleza heterogénea, su posterior procesamiento y análisis y el almacenamiento y explotación de la nueva información obtenida [7]. Este concepto ha tomado mucha fuerza en las herramientas relacionadas con el *Business Intelligence*, donde la transformación de información en conocimiento es un factor clave para optimizar la toma de decisiones en una empresa.

Aunque la definición no encaja exactamente con el proceso diseñado en este trabajo, el concepto de ETL va a ser la base de la arquitectura del sistema. Así, se podrán distinguir tres módulos o capas relacionadas con la extracción de la información, su posterior análisis para la generación de nuevo conocimiento y su presentación al usuario.

La Figura 3.1 muestra la aplicación del concepto de ETL como base de la arquitectura de este trabajo. La información generada por los usuarios en sistemas interactivos como las redes sociales es el *input* del sistema propuesto. A partir de estos datos se inicia un proceso de extracción de meta-información, análisis mediante diversas técnicas de la IA y preparación de la información para que sea fácilmente accesible y consumible por el usuario.

Figura 3.1: Esquema general de la ETL propuesta para el sistema

El usuario inicia un proceso de creación, compartiendo algún tipo de

contenido en sistemas externos, como las redes sociales. La primera tarea del sistema es la obtención de los datos multimedia y la posterior extracción de meta-información de los mismos, que serán necesarios para realizar el análisis. Tras la fase de extracción, tiene lugar la fase de transformación, donde los datos se adaptan y se preparan para un análisis basado en técnicas y algoritmos de aprendizaje automático. Finalmente, en la fase de carga de la ETL se almacena y se proporciona al usuario la nueva información generada por el sistema a partir de los contenidos multimedia iniciales. De esta manera, el usuario comienza el proceso de generación de contenidos que el sistema analiza y transforma para acabar proporcionándole nuevos contenidos, dando lugar a un ciclo de aprovechamiento, reciclaje y creación de información.

En la Figura 3.2 se puede ver el flujo de transformación de información que genera el sistema propuesto. En ella se pueden distinguir las tres fases generales de la ETL donde se produce el análisis de la información, y los módulos específicos del sistema que se enmarcan en cada una de ellas.

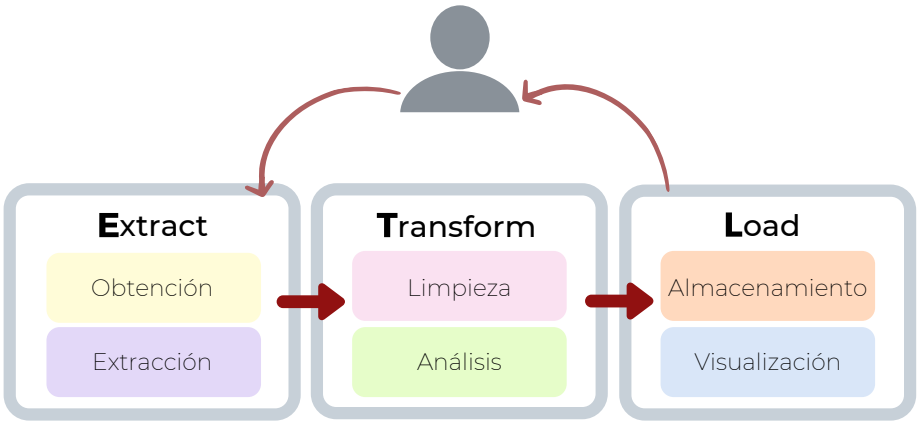

Figura 3.2: Esquema de la ETL con los módulos propuestos para el sistema

Las ETLs se utilizan, generalmente, para cargas de trabajo por lotes,

especialmente a gran escala; esto quiere decir que el proceso se aplica para un conjunto de datos concreto en un momento determinado. Sin embargo, la generación masiva de datos en internet supone una fuente de información muy interesante para la toma de decisiones de cualquier sistema de *Business Intelligence* o de aprendizaje automático. La necesidad de consumir estos datos en tiempo de ejecución o de manera eficiente ha provocado un aumento en el número de problemas y su impacto, obligando a adaptar el diseño de estos procesos. Como consecuencia, las ETLs se encuentran en un proceso de evolución notable y constante desde su modelo clásico a las diferentes variantes que satisfacen las necesidades actuales [137].

La eficiencia en el flujo de datos de estos sistemas es un factor crítico, puesto que las operaciones tienen cierto carácter secuencial: hasta que los datos no se hayan extraído y preparado, el análisis no puede comenzar. Esto puede desencadenar una serie de dificultades y problemas que penalicen la eficiencia del sistema. ¿Qué pasa si un proceso de extracción lleva demasiado tiempo y causa latencia? ¿Y si existieran dos tareas que se pudieran realizar en paralelo para optimizar el proceso? Estos problemas podrían ser comunes en nuestra propuesta. Por ejemplo, si se quiere realizar el procesamiento de una imagen habría que realizar varios procesos de extracción para obtener datos relativos al color, a las formas, a los elementos... Posteriormente habría que preparar y adaptar estos datos y finalmente realizar el análisis. Las tareas de extracción son independientes, por lo que podrían realizarse en paralelo reduciendo así el tiempo de espera para poder llevar a cabo el aprendizaje por parte del sistema. Por ello, es necesario adaptar y reforzar el concepto de ETL para que el sistema satisfaga los objetivos del trabajo de una manera óptima.

Para el procesamiento de flujos de datos, en este trabajo se van a utilizar las tuberías de datos o *data pipelines*. Se trata de una herramienta que

engloba la idea de ETL y la hace más flexible, permitiendo además que los datos se procesen como flujo en tiempo de ejecución, y no en lotes [71]. Esto es especialmente interesante en el contexto del presente trabajo, puesto que permite realizar análisis del contenido multimedia que se genera en los sistemas de usuario en flujo continuo, como si se tratara de un sensor que va captando información de su entorno. Las tuberías de datos permiten aplicar transformaciones de datos de forma paralela y distribuida, resolviendo así el problema que planteábamos con anterioridad para el procesamiento de una imagen.

El diseño modular favorece, además, la condición heterogenea del sistema propuesto. Por una parte, permite paralelizar y distribuir diferentes procesos de análisis de contenido multimedia para que se puedan desarrollar de manera simultánea. Así, el sistema puede abarcar todos procesos de análisis que se diseñen siguiendo la arquitectura propuesta, independientemente de la naturaleza de los datos de entrada (imágenes, audio, texto...) Por otra parte, el desarrollo modular de las tareas del sistema favorece la reutilización de las mismas para más de un proceso de análisis. Es decir, si dos o más estudios del sistema tienen una tarea básica en común, esta tarea, que será única, se aplica indistintamente en cada uno de ellos.

El término *tubería* ilustra perfectamente el concepto que pretende representar, ya que cada *data pipeline* recibe unos datos de entrada, los transforma (o no) y proporciona unos datos de salida, siendo el proceso que se produce en su interior perfectamente hermético. El proceso completo está formado por un conjunto o red de tuberías de manera que la distribución de tareas facilita y agiliza el flujo de transformación de la información, reduciendo la carga o presión que se ejerce sobre las máquinas que lo ejecutan.

La relación que se establece entre las diferentes tuberías de la red viene

determinada por las dependencias entre las tareas que realizan. Es decir, si una tubería A necesita que su *input* sea lo que otra tubería B obtiene como *output* se establece una relación de dependencia $A \longrightarrow B$. Esta dependencia conlleva que la tarea definida para la tubería B debe ejecutarse antes que la tarea definida para la tubería A.

El resultado de establecer dependencias entre todas las tareas o tuberías que componen el sistema da lugar a un grafo dirigido donde las tareas no pueden utilizar como *input* su propio *output*. Es decir, el grafo que obtenemos de la representación de dependencias es un grafo dirigido acíclico o *Directed Acyclic Graph* (DAG) [71]. Cada nodo de este grafo representa una tarea o tubería, y los arcos indican las relaciones de dependencia entre ellas, y con ello, el flujo de transformación de los datos.

A modo de resumen, la arquitectura del sistema está basada en el concepto de proceso ETL donde se diferencian tres fases: extracción, transformación y carga. En cada una de estas fases se diferencian dos módulos que englobarán una serie de tareas básicas. Sin embargo, debido a las necesidades del problema, el procesamiento de los datos se va a realizar mediante la aplicación de tuberías de datos. Como resultado se obtiene una arquitectura completamente flexible, modular y escalable. Así, en la sencillez del diseño arquitectónico del sistema reside su fortaleza.

El sistema está planteado para implementar soluciones a diversos problemas de manera simultánea. Para ello, cada proceso de análisis debe ser previamente diseñado como un DAG de tareas básicas que se enmarcan en cada una de las fases de la ETL. A continuación se describe la funcionalidad contenida en cada fase y en cada módulo del sistema. Los módulos relacionados con extracción se presentan en las Secciones 3.1.1 y 3.1.2. En ellos se detalla, respectivamente, el proceso de obtención de datos y de extracción de

meta-información. En las Secciones 3.1.3 y 3.1.4 se desglosan los módulos de limpieza de datos y la aplicación de algoritmos de IA asociadas a la fase de transformación de la ETL. La fase de carga de datos está compuesta por el módulo de almacenamiento, definido en la Sección 3.1.5 y el de visualización o reproducción de los resultados obtenidos, descrito en la Sección 3.1.6.

3.1.1. Obtención de información

El primer paso para el análisis de contenido multimedia es la obtención de información. Este módulo se enmarca dentro de la fase de extracción y consiste en obtener el contenido multimedia creado y compartido por un usuario en un sistema externo (como puede ser una red social) para utilizarlo como punto de partida y como objeto de análisis en la presente propuesta.

La evolución de la tecnología a lo largo de los últimos años ha fomentado el desarrollo de sistemas que permiten compartir información en muchos formatos distintos: texto, imagen, audio, vídeo... Dado que el objeto del presente trabajo es desarrollar un sistema para el análisis de contenido multimedia, es importante que las tareas que engloba este módulo estén enfocadas en la obtención de la información necesaria para cada caso, atendiendo a sus características y a las restricciones de las fuentes de datos.

Para cada proceso de análisis será necesario establecer, al menos, una tarea de obtención de información en función de las necesidades del problema y del enfoque de la solución. Algunos métodos que podrían aplicarse en las tareas de este módulo serían la obtención de los datos mediante la API del sistema original, la publicación de un formulario para la carga de datos manual por parte del usuario, la utilización de *crawlers* o el desarrollo de una extensión de navegador.

3.1.2. Extracción de meta-información

Existen muchas propuestas para el análisis de contenido multimedia. En el caso del análisis de imágenes, por ejemplo, algunos autores trabajan en el diseño de arquitecturas especiales de redes neuronales para poder llevar a cabo una tarea de clasificación [31, 120]. En estas propuestas la propia red neuronal, por su arquitectura, realiza una extracción de características y una tarea de clasificación. Sin embargo, por la condición heterogénea del sistema, uno de los objetivos del trabajo es extraer características descriptivas de los diferentes tipos de datos. Adicionalmente, se debe considerar que los datos extraídos se adapten a las necesidades de cada problema.

En función de la naturaleza de los datos de entrada al sistema se deben diseñar las tareas necesarias para la extracción de meta-información. Esta tarea dará lugar a un conjunto de datos sobre el que se realizará el análisis. Siguiendo con el ejemplo anterior del análisis de una imagen, si se quiere realizar un estudio que establezca una relación entre la información cromática con el estado de ánimo de su autor, la tarea de extracción estará enfocada únicamente en los datos del color. Sin embargo, si el objetivo es establecer una relación entre los componentes gráficos de la imagen y el estado de ánimo de su autor será necesario realizar varias tareas de extracción que permitan obtener información sobre el color, las formas, la disposición de los elementos en la imagen, el contraste de la misma...

Como consecuencia, las tareas incluídas en este módulo supondrán la obtención de los datos específicos, descriptivos e intrínsecos del contenido multimedia sobre los que se realizará el análisis. Por ello, las tareas de extracción estarán diseñadas en función del tipo de contenido que se quiera analizar (imagen, vídeo, audio...) y de las características descriptivas que se necesiten en el estudio.

3.1.3. Limpieza de datos

Una vez extraídos los datos del contenido multimedia que el sistema va a analizar es necesario limpiarlos, adaptarlos y prepararlos para optimizar los resultados del estudio. Independientemente de las características del algoritmo de aprendizaje automático y de su adecuación al problema, su eficiencia está determinada e incluso condicionada por la calidad de los datos. Por este motivo, las tareas de este módulo tienen una relevancia especial en el proceso de análisis [141].

Es muy frecuente que existan ciertos errores o que no haya armonía entre los datos obtenidos de diversas fuentes. Esto es lo que se denomina *suciedad* en los datos [97]. Por ejemplo, en un proceso de análisis de textos es necesario que la codificación y el formato sean homogéneos y que no existan errores ortográficos. Hay otros casos en los que los registros no están completos porque hay alguna característica de la que no se ha podido obtener información, y otros en los que la información está duplicada. En todos estos casos es necesario aplicar un criterio y tomar decisiones para obtener, calcular o estimar los datos que faltan y eliminar aquellos que no aportan nada al estudio. Esta fase es especialmente importante en sistemas como el que se plantea en este trabajo, que obtienen los datos de diferentes fuentes dando lugar a una información completamente heterogénea.

En otros casos, el exceso de datos sólo aporta ruido y distorsión al estudio. Para evitar esto, se aplican análisis estadísticos para estudiar la aleatoriedad de los datos, la dependencia entre los mismos y el peso y relevancia que pueden tener para el algoritmo. Tras estas consideraciones, frecuentemente se aplican transformaciones en los datos que permiten generar nuevos campos en función de los que ya se tienen.

Este proceso de limpieza y preparación de la información será específico para cada uno de los conjuntos de datos que el sistema analice y para los algoritmos de aprendizaje automático que se apliquen en cada estudio.

3.1.4. Análisis de datos

El objetivo de las tareas de este módulo es implementar los algoritmos de aprendizaje automático en el sistema. Tomando como entrada el conjunto de datos preparado por las tareas del módulo anterior, estos algoritmos buscan patrones y generalizan comportamientos a fin de extraer conocimiento e inferir una expresión formal, denominada modelo, que los defina.

La selección de algoritmos de IA se realizará en función del problema que se quiera resolver y de la naturaleza de los datos de partida. Si el problema consiste en la obtención de una etiqueta en función de una serie de características será necesario estudiar algoritmos de clasificación. Sin embargo, si el objetivo es dividir el conjunto en una serie de subconjuntos o *clusters* con características similares, será interesante valorar la aplicación de los diferentes tipos de algoritmos de *clustering*. Si los datos responden a un patrón a lo largo de un período temporal y se quiere predecir un dato en un momento del futuro será interesante estudiar la aplicación de algoritmos de series temporales. En ocasiones, un proceso de análisis concreto puede conllevar la integración de varios de estos algoritmos con el objetivo de analizar diferentes conjuntos de datos o para complementarse y optimizar los resultados. Esta combinación de técnicas de IA da lugar a sistemas híbridos inteligentes.

En cualquier caso, las tareas de este módulo deben suministrar tanto la funcionalidad para la extracción de conocimiento o creación del modelo a partir de los datos como la aplicación del modelo sobre un conjunto nuevo de datos.

3.1.5. Almacenamiento de la información generada

En este módulo del ETL se enmarcan las tareas de almacenamiento del sistema. Como resultado de las tareas definidas para los módulos previos se genera información: el contenido multimedia de partida, uno o varios conjuntos de datos procedentes de las diversas tareas de extracción, el conjunto de datos global, el resultado de las múltiples tareas de limpieza de datos [141], los modelos obtenidos por los algoritmos de aprendizaje automático [97], las métricas que evalúan la eficiencia de dichos algoritmos [66]... Y finalmente, como producto de cada proceso de análisis, se obtendrá un resultado con información valiosa para el usuario. Toda esta información puede ser útil para una posible retroalimentación del sistema y para el posterior tratamiento de la información con objeto de futuros estudios y mejoras del sistema.

Debido a esta necesidad y considerando la condición heterogénea del sistema y de los datos, se ha considerado la utilización de una base de datos no relacional para este trabajo.

3.1.6. Visualización o reproducción de la información generada

Tal y como se establece en la hipótesis y los objetivos del trabajo, cada proceso de análisis finaliza con la generación de algún tipo de información basada en el contenido multimedia que crea y comparte un determinado usuario. Las tareas de este módulo están enfocadas a cerrar el ciclo presentado en la Figura 3.2 facilitando esta nueva información al usuario.

Considerando que esta nueva información puede tener diferentes formatos para cada proceso de análisis, es necesario que las tareas de visualización de cada estudio estén diseñadas para proporcionarle el resultado al usuario de la

manera más adecuada. Así, en el caso de que el sistema genere información numérica, se podrá considerar la representación de la misma en gráficos; si el sistema genera una imagen o un vídeo, se deberán desarrollar las tareas necesarias para que el usuario pueda visualizarlo y reproducirlo de la mejor manera posible.

3.2. Formalización de marcos de trabajo para el análisis de contenido multimedia

En esta sección se proponen dos marcos de trabajo para la aplicación de la arquitectura diseñada en la sección anterior. El primero de ellos, detallado en la Sección 3.2.1, presenta un flujo de trabajo para la composición automática de melodías descriptivas de un vídeo. El segundo se presenta en la Sección 3.2.2, y su objetivo es la creación automática de armonías descriptivas a modo de banda sonora durante un proceso de ilustración con medios digitales.

Ambos marcos de trabajo tienen características comunes que, desde un punto de vista superficial, hacen que parezcan aplicaciones similares del sistema: la información de partida es, en ambos casos, contenido gráfico (en un caso un vídeo, y en otro una ilustración), los dos *frameworks* abordan el problema de la composición musical descriptiva mediante la utilización de técnicas de la IA y en los dos casos se aprende y se aplica el criterio de una película musical preexistente para relacionar la información gráfica con la información auditiva. Sin embargo, se trata de dos enfoques completamente diferentes: se tratan problemas distintos, los datos de partida no son los mismos, el planteamiento y las técnicas utilizadas dan lugar a procesos inconexos entre sí y los resultados musicales que se obtienen en ambos casos son totalmente dispares.

Los dos *frameworks* que se plantean en este trabajo como ejemplo de aplicación de la arquitectura propuesta son totalmente independientes. Dada la condición modular y el carácter genérico y flexible de la arquitectura, se podrían implementar otros *frameworks* que dieran respuesta a la hipótesis de una manera sencilla. El proceso sería definir un nuevo caso enmarcado en la problemática que se trata en este trabajo, definir las técnicas y el flujo de trabajo para solucionarlo y añadir la funcionalidad necesaria a los módulos correspondientes del sistema. Así, por ejemplo, se podría llevar a cabo la creación de contenido gráfico a partir de información auditiva (proceso inverso al que planteamos en este trabajo), la generación de algún tipo de información de valor a partir del análisis de los comentarios de una publicación en una red social aplicando técnicas de la minería de textos o el análisis de influencia de un determinado usuario basado en las interacciones de una red social mediante técnicas de análisis de grafos.

La coexistencia en un mismo sistema de diferentes marcos de trabajo como los dos que se describen a continuación permite la integración de diversas técnicas de IA con el objetivo común de analizar contenido multimedia proveniente de diferentes fuentes. La combinación de estos algoritmos dota al sistema de un carácter híbrido que permite ampliar el rango de análisis de este tipo de datos y, con ello, incrementar la riqueza y la calidad de los resultados obtenidos.

La descripción de los *frameworks* se realiza con base en la arquitectura del sistema, detallando una serie de tareas que, a alto nivel, describen el flujo de trabajo de ambas propuestas.

3.2.1. *Framework* para la creación de melodías descriptivas basadas en vídeos

El primer proceso de análisis que se propone en este trabajo está relacionado con la composición automática de música [77]. La descripción del marco de trabajo se realiza con base en la arquitectura del sistema, por lo que es importante que el flujo de trabajo se pueda describir como una serie de tareas sencillas que tienen dependencias entre sí.

En este *framework* se propone la utilización de un vídeo inicial para establecer un criterio para relacionar elementos gráficos y auditivos. Para ello, será necesario analizar por separado la imagen y el sonido. Esta información se analizará mediante algoritmos enmarcados en el área de la IA con el objetivo de extraer y formalizar un patrón de relación entre las características de la imagen y los sonidos más importantes. Una vez extraído este conocimiento de los datos de partida se podrá llevar a cabo un proceso de composición musical descriptiva. Dado un contenido audiovisual que un usuario haya creado y compartido en otro sistema, el análisis comienza por la supresión de la información auditiva original, puesto que el objetivo de este estudio es componer música que describa la imagen. A partir de las características gráficas de este segundo vídeo y aplicando el patrón extraído por el algoritmo en la etapa anterior se obtiene una composición musical descriptiva de manera automática. Finalmente, la unión entre la imagen original del vídeo proporcionado por el usuario y la composición musical creada por el sistema da lugar a un nuevo vídeo.

Para poder comprender la aplicación de la arquitectura diseñada en este *framework* es necesario tener en cuenta que el flujo de trabajo está formado por dos etapas bien diferenciadas. En la primera de ellas se describe el proceso de aprendizaje, estrechamente ligado a la etapa de entrenamiento

del algoritmo enmarcado en el área de la IA. En la segunda etapa tiene lugar el proceso de creación, donde las tareas se encargan de la aplicación del modelo anteriormente obtenido para la obtención de la composición musical descriptiva.

La arquitectura de ambas etapas se presenta mediante el conjunto de todas las tareas que conforman el proceso, divididas según su funcionalidad en las etapas de extracción, transformación y carga. El color de cada tarea hace referencia al módulo que la engloba; así, las tareas de obtención se representan en amarillo, las de extracción de meta-información en morado, las de limpieza y preparación de datos en rosa, las de análisis en verde, las de almacenamiento en naranja y las de visualización o reproducción en azul.

La representación de las tuberías de datos se realiza, en ambos casos, mediante un DAG que indica las dependencias entre las diferentes tareas. Cada dependencia $A \longrightarrow B$ indica que la tarea A hará uso de la información obtenida por la tarea B. Por ello, es importante observar que el flujo de ejecución de las tareas se realizará en sentido contrario a la dependencia entre las mismas, llevándose a cabo B antes que A.

En la Figura 3.3 se pueden observar las tuberías que constituyen la etapa de aprendizaje del sistema. La primera tarea consiste en la obtención de un vídeo del que se inferirá el criterio para relacionar las características de la imagen con el sonido. Para ello, en primer lugar, es necesario extraer tanto información gráfica (datos del color, forma y disposición de los elementos) como auditiva (la nota principal de cada fragmento). Tras pasar por un proceso de limpieza y preparación, estos datos serán utilizados como entrada a un algoritmo de aprendizaje automático que formalice el patrón de relación de las características de la imagen con el sonido principal. En esta etapa, se almacenará el conjunto de datos extraídos y el modelo, que será necesario

para la etapa de creación.

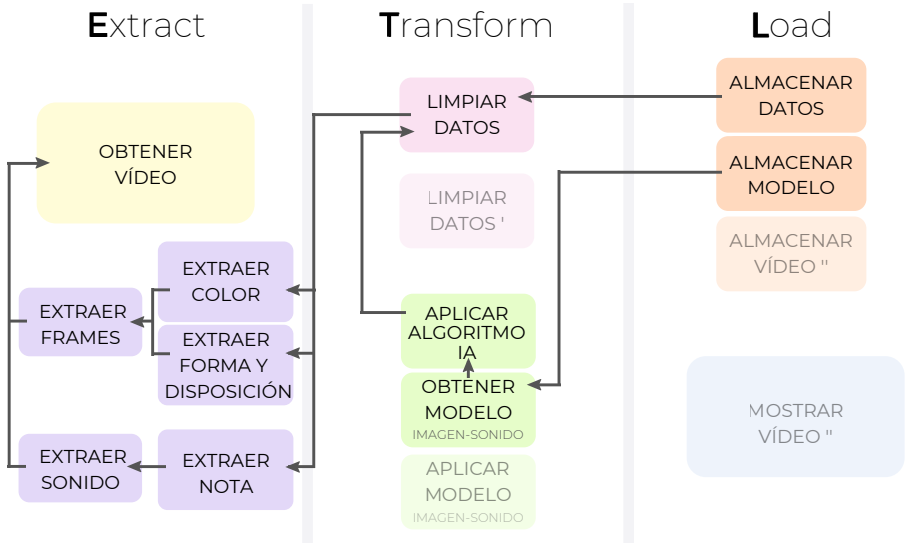

Figura 3.3: Grafo de tareas para la etapa de aprendizaje del primer *framework* propuesto para el análisis de contenido multimedia

Para la extracción del color se puede utilizar cualquiera de las técnicas descritas en la Sección 2.4.1.1, y en la Sección 2.4.1.2 se puede encontrar una recopilación de las técnicas más utilizadas para la extracción de descriptores de forma y disposición de los elementos en una imagen. Las técnicas empleadas para la obtención de información musical se describen en la Sección 2.4.2, y algunos de los algoritmos de clasificación enmarcados en el área de la IA que se podrían aplicar en este problema se detallan en la Sección 2.3. En la implementación de este *framework* desarrollada en la Sección 4.1 se llevan a cabo dos métodos para la extracción de información gráfica. En el primero de ellos se ha utilizado un histograma del color en el espacio RGB [127] y el algoritmo SIFT [80] para la extracción de color y forma y disposición respectivamente; en el segundo método se aplica la técnica de TL sobre una CNN previamente entrenada [103]. La extracción de la información musical se ha realizado a manos de un experto en este caso, y los algoritmos aplicados para obtener el patrón de relación entre atributos visuales y auditivos han sido

NB [107], SVM [19] y RF [22].

La Figura 3.4 esquematiza la etapa de creación del *framework*. El punto de partida es la obtención de un nuevo vídeo, creado y compartido por un usuario en otro sistema. La fase de extracción en este caso considera únicamente los atributos gráficos: color y forma de los elementos. Estos datos son procesados para darles el mismo formato que en la etapa anterior y utilizados como entrada al modelo previamente obtenido. Como resultado para cada fotograma, con base en sus características y al criterio establecido por el vídeo anterior para relacionar atributos visuales e información auditiva, se predice una nota. La concatenación de estas predicciones para el conjunto de todos los *frames* del vídeo da lugar a una composición musical descriptiva, que se utiliza como audio para el vídeo del usuario. En esta etapa se almacena el conjunto de datos utilizado para la predicción y el vídeo del usuario con la composición musical creada por el sistema.

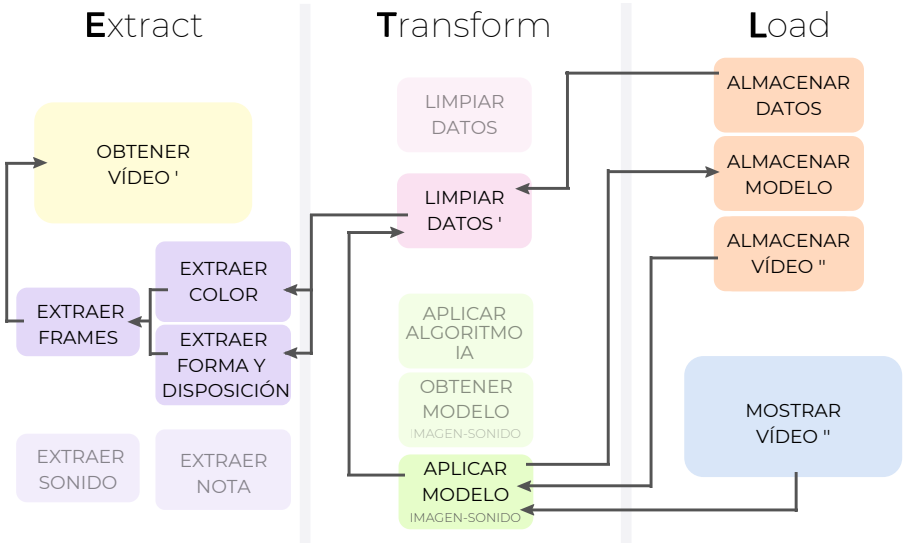

Figura 3.4: Grafo de tareas para la etapa de creación del primer *framework* propuesto para el análisis de contenido multimedia

Este marco de trabajo se aplica en un caso de estudio detallado en la

Sección 4.1. En él, la aplicación de tuberías de datos se lleva a cabo con Luigi[1], un módulo de Python utilizado por Spotify, Skyscanner y Red Hat entre otros.

3.2.2. *Framework* para la composición dinámica de armonías durante el proceso de ilustración con medios digitales

En este caso se propone un *framework* para la composición automática a partir de un proceso de ilustración. El objetivo de este estudio es diseñar un sistema compositor cuya música describa una creación artística [95] y que a su vez sea capaz de inspirar al artista en su proceso de creación [98]. Para ello, el análisis de la ilustración y el proceso de composición debe ser dinámico.

Para poder realizar una traducción de elementos gráficos a elementos auditivos es necesario establecer un criterio. Para ello, de nuevo se van a utilizar los descriptores de imagen y la información armónica de un material audiovisual de partida. Un algoritmo basado en IA analizará esta información a fin de obtener un modelo de relación entre las características de ambos tipos extraídas. Por otra parte, un segundo algoritmo de aprendizaje automático se va a utilizar para el proceso de composición musical. Una vez extraído este conocimiento, para aplicarlo es necesario realizar un análisis constante de la ilustración del usuario: periódicamente, se extraerán los descriptores de imagen y se aplicará el patrón imagen-sonido previamente obtenido para predecir el sonido más adecuado con base en el criterio de traducción aplicado. Esta información auditiva será la entrada del proceso de composición, haciendo que la música resultante esté determinada por la armonía obtenida y, por tanto, por las características del dibujo.

[1] https://github.com/spotify/luigi

Igual que en el caso anterior, la arquitectura de este marco de trabajo se define mediante la aplicación de una serie de tareas divididas en dos etapas: aprendizaje y creación. Cada tarea estará enmarcada en una de las tres fases de la ETL. Adicionalmente, con base en su funcionalidad, cada tarea estará incluida en uno de los módulos definidos en la arquitectura del sistema. Esta distinción se podrá realizar mediante el código de colores aplicado en la Figura 3.2. Las diferentes tuberías que dan lugar al flujo de trabajo se representan como un DAG para indicar la dependencia entre tareas.

En la Figura 3.5 se puede observar el proceso de aprendizaje de este marco de trabajo. Igual que en el caso anterior, el criterio para relacionar atributos gráficos de la imagen con información auditiva se establece mediante el análisis de un vídeo inicial. Para ello, a través de una serie de descriptores del color, forma y disposición de los elementos y la información armónica más relevante de cada fragmento del audio se aplica un primer algoritmo de aprendizaje automático que permita formalizar los patrones de relación entre dichas características. De esta manera, el algoritmo extrae patrones en la relación entre los atributos de la imagen y el contenido musical del vídeo de partida. Por otra parte, un segundo proceso de aprendizaje se lleva a cabo para el desarrollo de la composición automática de música. Con este objetivo, un conjunto de piezas musicales codificadas en un formato digital se utilizan como entrada para un segundo algoritmo de IA. Como resultado se obtiene un segundo modelo que permite completar el proceso creativo. Este modelo analiza una serie de composiciones musicales codificadas en formato digital para extraer los patrones de la relación entre las diferentes notas de la escala musical y, posteriormente, aplica dichos patrones en el proceso de creación para componer música. En esta etapa, tanto el conjunto de datos extraídos del vídeo como los dos modelos se almacenan en el sistema.

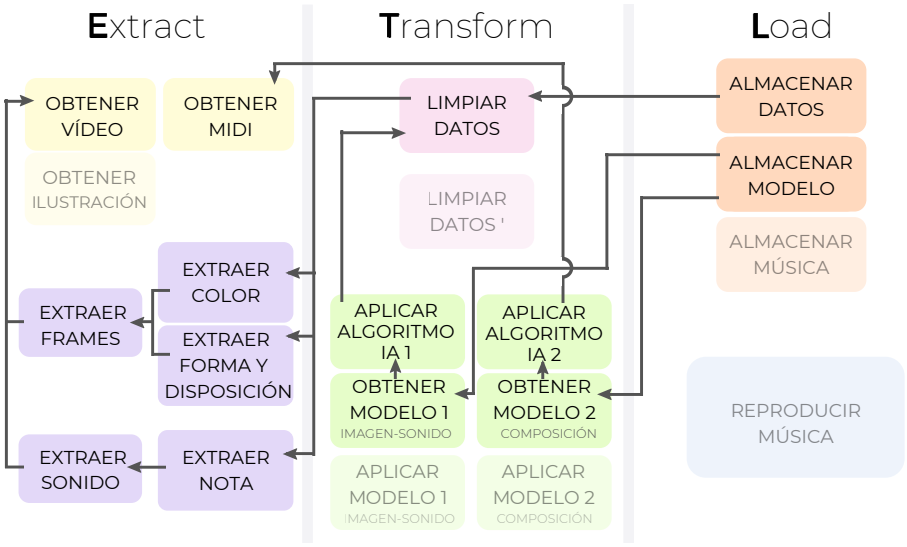

Figura 3.5: Grafo de tareas para la etapa de aprendizaje del segundo *framework* propuesto para el análisis de contenido multimedia

Para la extracción del color, forma y disposición de los elementos —es decir, la información gráfica del vídeo— las técnicas utilizadas se pueden encontrar en las Secciones 2.4.1.1 y 2.4.1.2. Las técnicas para la extracción de información auditiva se pueden consultar en la Sección 2.4.2. En el desarrollo de este *framework* detallado en la Sección 4.2 del trabajo, las técnicas utilizadas para la extracción de información de la imagen son la cuantificación del color mediante un algoritmo de *clustering* [26] y el algoritmo SIFT para la obtención de forma y disposición de los elementos [80]. La extracción de información musical se realiza, en este caso, mediante el algoritmo CENS [92]. Para la formalización de un patrón que relacione las características de la imagen con la información auditiva es necesaria la obtención de un primer modelo. En este caso se han empleado los algoritmos de clasificación multietiqueta RAKEL [129] y ML-kNN [140]. Como segundo algoritmo de IA se aplica una RNN basada en LSTM [116] con el objetivo de llevar a cabo el proceso de composición musical.

La Figura 3.6 ilustra la etapa creativa de la propuesta, en la que la ilustración del usuario se utiliza como punto de partida. La información gráfica se extrae y se procesa para poder ser utilizada como entrada al modelo de traducción imagen-sonido. El resultado será una nube de sonidos que representa el dibujo inicial y que servirá de entrada al modelo de composición musical. Este segundo modelo aplica, a partir de la nube de sonidos predicha, los patrones de relación entre notas inferidos del conjunto de datos de entrenamiento. Como consecuencia, el acorde obtenido por el primer modelo con el objetivo de describir la ilustración determina la información armónica de la música compuesta finalmente por el sistema. En esta etapa, la creación musical del sistema será almacenada y reproducida para que el usuario pueda utilizarla como fuente de inspiración en su proceso ilustrativo.

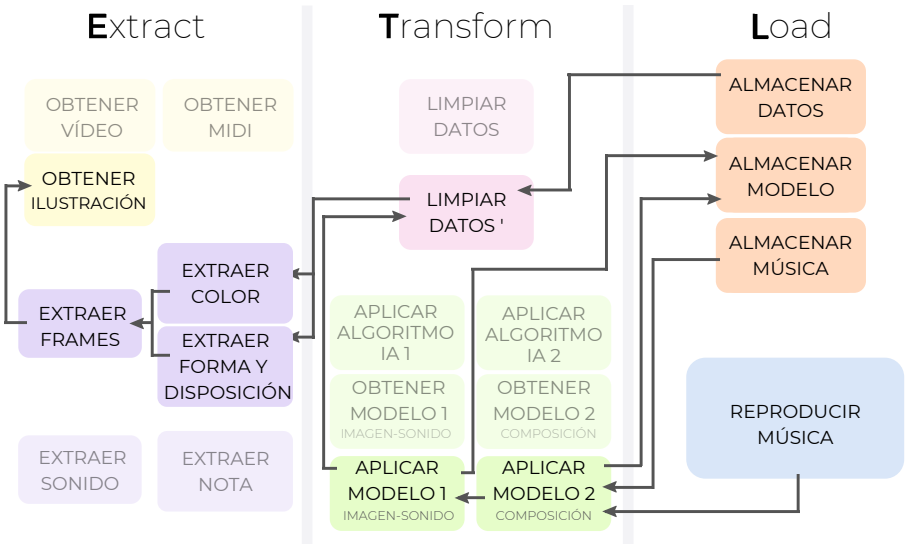

Figura 3.6: Grafo de tareas para la etapa de creación del segundo *framework* propuesto para el análisis de contenido multimedia

La Sección 4.2 presenta todos los detalles de un caso de estudio que aplica este *framework* para la composición musical descriptiva mediante IA. Al igual que en el caso anterior, el desarrollo de las tuberías de datos se lleva

a cabo mediante el módulo de Python Luigi[2].

En el próximo capítulo...

Una vez formalizada, descrita y desarrollada la propuesta, para poder probar su validez se va a poner en explotación en un entorno real. El siguiente capítulo presenta los detalles de dos casos de estudio que permiten analizar el funcionamiento de la propuesta en circunstancias reales y particulares. Dada la condición heterogénea del sistema propuesto, cada aproximación irá orientada a la obtención de resultados y observaciones y a su posterior análisis. La finalidad de todo este proceso es contrastar la validación de la hipótesis inicialmente formulada y comprobar si los resultados del proceso de investigación están alineados con la consecución de los objetivos planteados en el primer capítulo de la memoria.

[2] https://github.com/spotify/luigi

Capítulo 4

Casos de estudio

Resumen: *La fase de diseño, formulación y desarrollo de la propuesta es sucedida por una etapa de explotación del mismo en un entorno real. Dada la heterogeneidad del sistema y la independencia de cada uno de los enfoques desarrollados, este capítulo describe dos casos de estudio bien diferenciados que contribuyen a la validación de la hipótesis. El primer caso de estudio describe un proceso de composición de melodías descriptivas a partir de un vídeo; el segundo, detalla un método para la generación automática y dinámica de armonías mientras un artista realiza una ilustración con medios digitales. En ambos se utiliza la relación entre elementos visuales y musicales que se establece en la película Fantasia de Disney.*

Actualmente, la utilización de sistemas basados en TIC con objetivos de comunicación avanzada como las redes sociales genera un gran tráfico de información de naturaleza muy diversa: imágenes, vídeos, mensajes de texto, mensajes de voz... El presente trabajo establece su hipótesis entorno a este hecho, proponiendo que estos datos sean analizados mediante técnicas de la

IA para crear otro tipo de información útil para el usuario. Para solventar este problema, se propone una arquitectura modular que permita combinar diferentes técnicas para el análisis y creación de nuevo contenido multimedia. Adicionalmente, de manera teórica, se proponen dos aplicaciones de dicha arquitectura que están relacionados con la composición musical automática. Para poder comprobar la eficacia del problema, y con ello si el trabajo realizado valida la hipótesis previamente establecida, es necesario poner el sistema propuesto en explotación en un contexto real.

Este capítulo pormenoriza dos casos de estudio en los que se aplican los dos marcos de trabajo propuestos en la Sección 3.2 del capítulo anterior. Aunque ambos tienen como resultado una composición musical descriptiva y los dos aplican el mismo criterio para la traducción de elementos gráficos a sonido, tanto el punto de partida, como el proceso de composición y los resultados musicales obtenidos son totalmente diferentes. En el primero de ellos, partiendo de un vídeo, se desecha su información auditiva y se hace únicamente uso del contenido visual. A partir de esta información, analizando diversos descriptores como el color, la forma y la disposición de los elementos se componen melodías capaces de describir los cambios visuales que suceden a lo largo de todos los fotogramas. En el segundo, se desarrolla una herramienta para la creación y edición de imágenes por medio de una tableta gráfica. Las características visuales de la ilustración generada se utilizan como fuente de inspiración para el sistema, que compone, de manera dinámica y automática, composiciones descriptivas a modo de banda sonora del proceso ilustrativo.

El contenido de este capítulo está estructurado en dos secciones. En primer lugar, la Sección 4.1 detalla el primer caso de estudio llevado a cabo en este trabajo. En él, el sistema compone melodías descriptivas a partir de un vídeo. Por otra parte, en la Sección 4.2 se describen todos los aspectos relativos a la ejecución del segundo caso de estudio, en el que el sistema to-

ma como fuente de inspiración una ilustración realizada por medios digitales para desarrollar un proceso de composición musical armónica.

4.1. Composición de melodías que describen vídeos aplicando el estilo de la película *Fantasia* de Disney

El objetivo de este estudio es probar la validez del *framework* descrito en la Sección 3.2.1. Para ello se define un escenario real y concreto, se seleccionan los vídeos que se van a utilizar como entrada del sistema y se aplican las técnicas y algoritmos necesarios para la obtención de melodías descriptivas.

Para poder establecer una relación entre los elementos gráficos y los elementos auditivos en este estudio se utiliza la película *Fantasia* de Disney [134]. Esta obra clásica de animación no destaca por los diálogos de sus personajes, que además son casi inexistentes, sino que se distingue por ilustrar diversos temas de la música clásica que se caracterizan por ser piezas descriptivas. Para ello, un equipo de animadores profesionales crearon un modelo que relacionaba los diferentes colores con las emociones que provoca la música y analizaron cómo diseñar a los personajes y sus movimientos para que existiera una coherencia entre la animación y los detalles musicales [32].

La Figura 4.1 presenta un fotograma de la película en el que se puede ver al personaje de Mickey Mouse con un gorro mágico que le otorga ciertos poderes. Esta imagen pertenece al fragmento de la película que ilustra la pieza musical de *El aprendiz de brujo*, compuesta por Paul Dukas en 1897 y perteneciente al estilo de la música programática.

Figura 4.1: Fotograma de la película *Fantasia* de Disney que ilustra la pieza *El aprendiz de brujo*

El flujo de trabajo de este caso de estudio se divide en dos etapas: la de aprendizaje y la de creación. Concretamente, en la etapa de aprendizaje, se ha seleccionado el fragmento de la película donde se ilustra la pieza de *El cascanueces* de Piotr Ilich Tchaikovsky porque está compuesta por varias danzas que evocan las cuatro estaciones y por lo tanto tiene una gran diversidad de formas y colores que resulta muy interesante para el entrenamiento del sistema.

La Figura 4.2 ilustra el flujo de trabajo establecido para esta primera etapa del sistema. Como se puede observar, el punto de partida es el fragmento de vídeo seleccionado, que se divide en fotogramas con el objetivo de analizar los descriptores de la imagen. No todos los fotogramas que conforman el vídeo inicial se consideran para el proceso de análisis; se realiza un proceso de limpieza previo al estudio para desestimar aquellos fotogramas que son especialmente similares, reduciendo así la complejidad en la extracción de información musical y la probabilidad de un desajuste del algoritmo por sobre-entrenamiento (*overfitting*) [42].

Para cada uno de los fotogramas relevantes finalmente considerados para el presente estudio se lleva a cabo la extracción de dos tipos de datos: por un lado, se extrae información gráfica de los fotogramas considerados, y en segundo lugar se obtiene el sonido principal que se escucha durante la reproducción de cada uno de los fotogramas. Toda esta información servirá como entrada a los algoritmos de aprendizaje del sistema, y como resultado final se obtendrá un modelo matemático que establece una relación entre los descriptores de la imagen (atributos) y el sonido principal (clase).

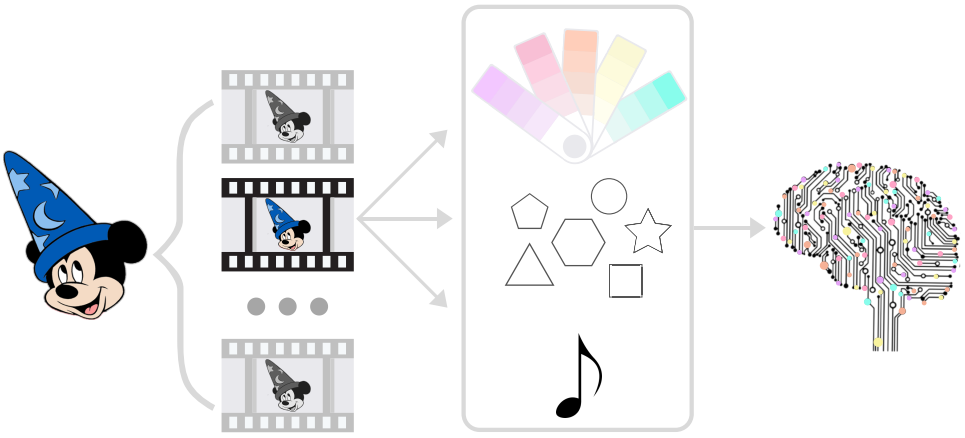

Figura 4.2: Vista global de la etapa de aprendizaje del primer caso de estudio

La Figura 4.3 representa la dependencia entre tareas que se van a llevar a cabo para el desarrollo de esta etapa del caso de estudio. En este DAG se concretan las técnicas seleccionadas para llevar a cabo el marco de trabajo planteado en la Figura 4.2 de la Sección 3.2.1. Para este estudio se plantean dos métodos de extracción de descriptores de la imagen. Por una parte, el primer método (M1) hace uso de un histograma del color en el espacio RGB [127] junto con el algoritmo SIFT [80] y su optimización mediante la técnica *Bag of Visual Words* (BoVW) [115]; el segundo método (M2) hace uso de una CNN [120] previamente entrenada, aprovechando los beneficios que esto supone mediante la aplicación de TL [103]. Para la generación de un modelo que relacione descriptores gráficos con atributos musicales en el vídeo de partida, se aplican los algoritmos NB [107], SVM [19] y RF [22].

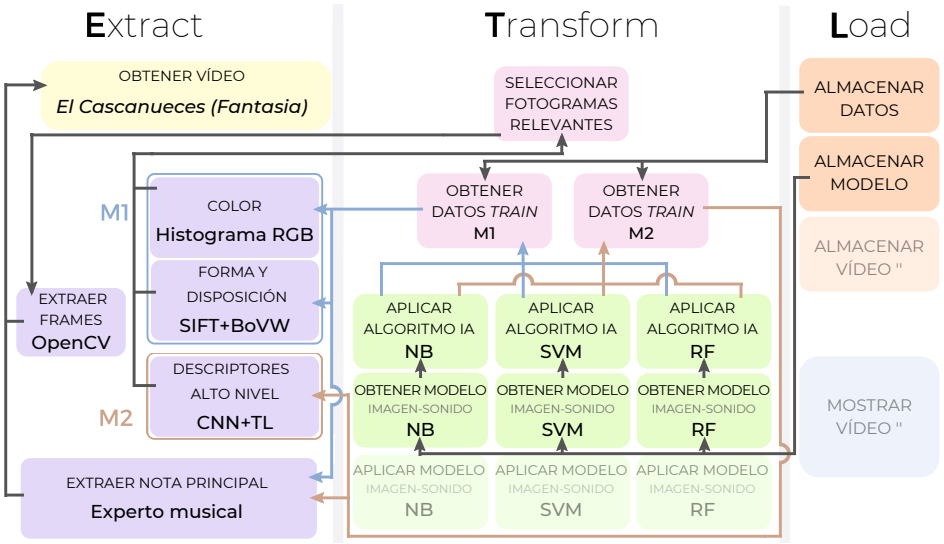

Figura 4.3: Grafo de tareas para la etapa de aprendizaje del primer caso de estudio

En la etapa de creación de este caso de estudio se utiliza, como vídeo de partida, un fragmento perteneciente a la película *Fantasia 2000* de Disney [133]. Esta producción fue creada para celebrar el 60 aniversario de la película *Fantasia*, y sigue la línea y el estilo narrativo de su predecesora. Por su

variedad en colores y formas, el fragmento correspondiente a la pieza musical *Pájaro de fuego*, compuesta por Igor Stravinksi en 1919, se utiliza para la composición musical en esta etapa. La ilustración de este fragmento incluye diversas formas y colores que cuentan la historia de un hada que despierta a un espíritu de lava gigante y violento con forma de fénix.

La Figura 4.4 ilustra el proceso de creación musical del sistema. En esta etapa, el sonido original del fragmento seleccionado se desecha. El vídeo se divide en un conjunto de fotogramas con el mismo criterio aplicado en la etapa anterior, y de cada uno de ellos se extraen una serie de descriptores de imagen que sirven como entrada al modelo generado en la fase de aprendizaje. Este modelo, aplicando el criterio del primer fragmento de vídeo y teniendo en cuenta las características de la imagen, predice el sonido más adecuado para la información gráfica que recibe. De esta manera se obtiene una consecución de sonidos que representan a cada uno de los *frames* del vídeo, dando lugar a una melodía.

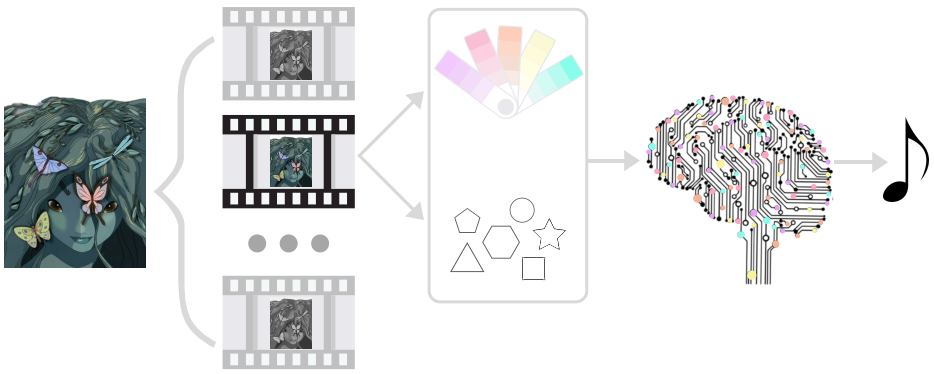

Figura 4.4: Vista global de etapa de creación del primer caso de estudio

La arquitectura de esta etapa del caso de estudio se detalla en la Figura

4.5. En este caso, se parte de un vídeo diferente del de la etapa anterior, pero las técnicas aplicadas para la extracción de descriptores gráficos son las mismas. En este caso no se extrae información musical del vídeo puesto que es, precisamente, lo que se va a generar al finalizar el proceso de análisis.

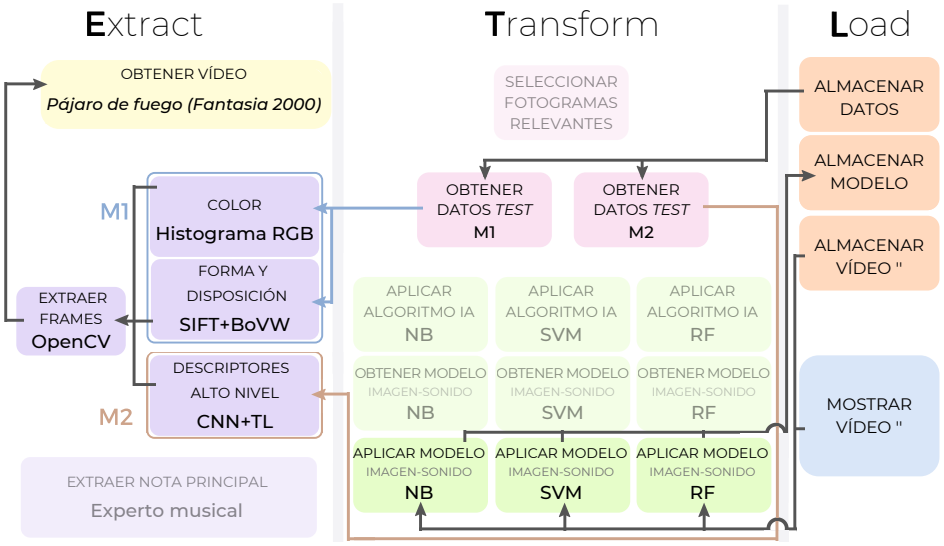

Figura 4.5: Grafo de tareas para la etapa de creación del primer caso de estudio

Los detalles técnicos del caso de estudio se estructuran de la siguiente manera: en la Sección 4.1.1 se presenta el proceso diseñado para la extracción de meta-información gráfica y auditiva, en la Sección 4.1.2 se describe la fase de aplicación de la IA en el estudio, la Sección 4.1.3 presenta los resultados desde el punto de vista analítico y desde el punto de vista musical y la Sección 4.1.4 desarrolla las conclusiones obtenidas tras la realización del caso de estudio.

4.1.1. Extracción y preparación de la meta-información gráfica y musical de partida

La película seleccionada para el análisis está rodada a 24 fotogramas por segundo y en este estudio se considera el primer fotograma de cada 8, lo que da lugar a la descomposición del fragmento del vídeo en imágenes, obteniendo 3 representaciones gráficas por cada segundo de película. De esta cantidad de fotogramas se descartan aquellos que son demasiado similares y no presentan cambios en el color, forma o disposición de los elementos, dando lugar a un conjunto de 483 imágenes.

Para la obtención de fotogramas a partir del vídeo se utiliza la librería OpenCV [21]. El Algoritmo 1 detalla el pseudocódigo del proceso llevado a cabo para la obtención de los fotogramas relevantes que finalmente serán analizados en fases posteriores del estudio. El proceso para obtener la representación de las imágenes en vectores con el formato deseado se puede observar en el Algoritmo 2. En este punto, se aplica una medida de distancia —en este caso, el coseno— entre los vectores de características del último fotograma relevante almacenado y el fotograma que se está evaluando, permitiendo identificar como imágenes similares aquellas cuya distancia supere un umbral previamente establecido. En este proceso, si dos o más fotogramas consecutivos son muy similares, se conserva solo el primero de ellos.

Tal y como indica el Algoritmo 1, la detección de imágenes similares se basa en el cálculo de un vector de características que resume la imagen. Este proceso se refleja en el Algoritmo 2, y la aplicación de estas técnicas se describe detalladamente en la Sección 4.1.1.1. Las características que definen la imagen son los descriptores de forma y distribución de los elementos obtenidas mediante el algoritmo SIFT [80] y los valores del histograma del color en el espacio RGB [127].

Algoritmo 1 Obtención de fotogramas relevantes de un vídeo

Input $fragmento$: vídeo de partida para la obtención de fotogramas
Output $fotogramas[\,]$: vector con el conjunto de fotogramas relevantes

$fotogramas\_raw[\,] \longleftarrow ObtenerFotogramas(fragmento)$
$j \longleftarrow 1$

for $i : 1$ **to lenght**$(fotogramas\_raw[\,])$ **do**
  if $i \,\%\, 8 = 0$ **then**
    $descriptores1[\,] \longleftarrow ObtenerDescriptores(fotogramas\_raw[i])$
    $descriptores2[\,] \longleftarrow ObtenerDescriptores(fotogramas[j-1])$
    if $\mid coseno(descriptores1, descriptores2) \mid > umbral$ **then**
      $fotogramas[j] \longleftarrow fotogramas\_raw[i]$
      $j \longleftarrow j + 1$
    end if
  end if
end for

return $fotogramas[\,]$

Algoritmo 2 Obtención de un vector de descriptores de una imagen

Input $imagen$: imagen para obtener los descriptores
Output $descriptores[\,]$: vector con el conjunto de descriptores de la imagen

$descriptores[\,] \longleftarrow ObtenerSIFT(imagen)$
$descriptores[\,]$ **append** $ObtenerHistogramaR(imagen)$
$descriptores[\,]$ **append** $ObtenerHistogramaG(imagen)$
$descriptores[\,]$ **append** $ObtenerHistogramaB(imagen)$

return $descriptores[\,]$

En este estudio, el proceso de extracción de meta-información se aplica para cada uno de los fotogramas o *frames* relevantes del vídeo. La obtención de los descriptores de imagen se ha llevado a cabo mediante la aplicación de dos técnicas diferentes. Como resultado, se obtienen dos métodos para la composición automática de melodías descriptivas cuya única diferencia reside en la extracción de meta-información de los fotogramas. A continuación se detalla el proceso de extracción de descriptores gráficos y musicales para cada uno de los métodos propuestos.

4.1.1.1. Primer método de extracción de descriptores gráficos

En este trabajo, para poder establecer una relación entre imagen y sonido es necesario un proceso previo de extracción de patrones de las características básicas. Los descriptores de forma y disposición de los elementos en la imagen se extraen mediante la combinación del algoritmo SIFT [80] y la técnica BoVW [115]. En primer lugar, para cada fotograma se extraen una serie de descriptores o características que pueden entenderse como puntos clave de la imagen y la información de los píxeles que los rodean. Para ello se utiliza el algoritmo SIFT puesto que se prioriza la precisión sobre la rapidez y la eficiencia [70]. Estos vectores de características se agrupan mediante técnicas de *clustering* para obtener lo que se denominan palabras visuales. La técnica de BoVW nos permite representar cada fotograma a partir de un vector que contiene un recuento ponderado de cada palabra visual. Como consecuencia, los primeros 400 atributos que se extraen de cada fotograma se corresponden con un vector de *visual words* (palabras visuales) obtenido de la aplicación de BoVW a los descriptores SIFT.

Para la extracción de color se han utilizado histogramas de color [127]. El espacio de color seleccionado para la extracción de información mediante

histogramas es el RGB por presentar ciertas similitudes con la percepción del ojo humano y por ser uno de los más utilizados en conjunción con los histogramas de color [24]. De esta manera, para cada uno de los canales (R, G y B) se ha obtenido un vector con 256 *bins* donde cada uno de ellos representa un valor de intensidad en el rango [0-255]. El valor numérico para cada uno de estos *bins* hace referencia al número de píxeles de la imagen que se corresponden con cada intensidad del canal. Así, para cada canal (rojo, verde y azul) se obtienen 256 valores, dando lugar a un total de 798 atributos relativos al color por cada fotograma.

La Figura 4.6 muestra un fotograma de la película *Fantasia* (**a**), los descriptores SIFT que se obtienen mediante la extracción propuesta en este método (**b**) y el histograma de color en el espacio RGB (**c**).

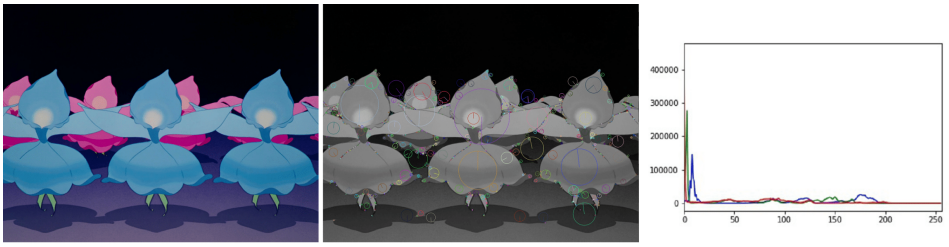

Figura 4.6: Descriptores de la imagen sobre un *frame* del fragmento *El Cascanueces* de la película *Fantasia* en el primer caso de estudio. La Figura a) contiene el fotograma original. La Figura b) muestra los descriptores SIFT extraídos. La Figura c) presenta el histograma de color

Cada descriptor SIFT representa un borde o una forma concreta que posiciona el elemento en el espacio, y el conjunto de todos ellos componen la información de la imagen completa. El histograma de colores muestra que los tres canales R (rojo), G (verde) y B (azul) tienen un alto número de píxeles en los diez primeros *bins*, lo que se corresponde con los colores más oscuros presentes en el fondo (negro) y el suelo (azulado). Sin embargo, en el

resto de *bins* se representa que hay menos píxeles que obtienen valores altos para los tres canales. Destacan las ligeras variaciones o repuntes del canal verde (*bins* 130-150) y del canal azul (*bins* 175-200) que se corresponden, respectivamente, con los colores más brillantes de los tallos y los cuerpos de las flores.

La extracción de meta-información auditiva del vídeo se lleva a cabo con el objetivo de obtener el sonido de mayor relevancia armónica que se escucha durante la reproducción de cada fotograma. En ocasiones, este sonido coincide con la nota fundamental del acorde que se escucha, pero en otros casos se considera la nota que destaca por encima de las demás. Dada la complejidad del problema y para evitar que la obtención automática de sonidos penalice la calidad predictiva del sistema, este proceso lo realiza un músico profesional. Para este proceso de recolección de información se consideran las 12 notas de la escala cromática, restringiendo el registro a la octava correspondiente al do central del piano para simplificar la tarea. De esta forma, no se trabaja con diferentes escalas que permitan tener un gran registro de graves y agudos, sino que se focaliza en la distinción de las 12 notas musicales establecidas. Para el desarrollo de este trabajo se utiliza la codificación MIDI, que define cada sonido con una serie de características musicales (nota musical, tempo, volumen, duración...) a las que les atribuye valores numéricos para que sean comprensibles por una máquina [47]. La Tabla 4.1 muestra la codificación de las 12 notas musicales utilizadas en este trabajo en cifrado latino, americano y MIDI.

C. Latino	Do	Do#	Re	Re#	Mi	Fa	Fa#	Sol	Sol#	La	La#	Si
C. Americano	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
MIDI	60	61	62	63	64	65	66	67	68	69	70	71

Tabla 4.1: Notación numérica de las notas musicales

Tras obtener la meta-información auditiva del vídeo, se ha realizado un breve análisis para conocer la frecuencia de cada una de las notas musicales consideradas en el estudio. La Tabla 4.2 muestra el número total y el porcentaje de fotogramas contabilizados para cada una de las notas musicales.

	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
Frecuencia	6	24	93	12	39	21	30	66	0	78	42	72
Frecuencia (%)	1.24	5	19.25	2.48	8.07	4.35	6.21	13.66	0	16.15	8.7	14.9

Tabla 4.2: Frecuencia de cada una de las notas en el conjunto de datos

De los valores de la Tabla 4.2 se pueden extraer dos conclusiones. La primera de ellas es que no existe representación de la nota `G#`. Esto significa que para ninguno de los fotogramas del fragmento de vídeo utilizado en el entrenamiento del sistema la nota principal es el `Sol#`. Esto conlleva que en la fase de creación el sistema no asignará esta nota a ningún fotograma. En segundo lugar, el número de ocurrencias de cada nota en el conjunto de datos no es el mismo. En minería de datos esto se conoce como desbalanceo de los datos y da pie a que pueda existir una predisposición del sistema a predecir unas notas concretas (aquellas que tienen más presencia en el conjunto de datos de entrenamiento). La justificación de ambos fenómenos se puede realizar desde un punto de vista musical: en una obra musical se establece una armonía principal, y aunque existan modulaciones hay notas que son más lejanas al marco sonoro establecido. El posible problema podría solucionarse mediante la ampliación del conjunto de datos del entrenamiento de manera que todas las notas tuvieran una representación similar.

La fase de extracción con este primer método obtiene, para cada fotograma, 1198 atributos que describen características gráficas y un atributo que representa la información auditiva. El conjunto de datos generado por este método se puede encontrar en [85].

4.1.1.2. Segundo método de extracción de descriptores gráficos

En este caso, la información gráfica que describe cada uno de los fotogramas del vídeo se extrae mediante una CNN ya entrenada. Concretamente se ha utilizado el modelo Inception-v3, diseñado para tareas de visión artificial [120]. Este modelo alcanzó una exactitud superior al 78.1 % en la tarea de reconocimiento de imágenes sobre el conjunto de datos de ImageNet [111]. De esta manera, en lugar de etiquetar cada uno de los fotogramas con su sonido correspondiente y utilizar la CNN para realizar una tarea de clasificación, aprovechamos las ventajas que ofrece un modelo de clasificación ya entrenado y lo utilizamos para la extracción de características de cada uno de los fotogramas del vídeo. Este proceso, conocido como TL [103] permite obtener el vector de características de alto nivel obtenido por la última capa de *pooling* de la CNN. Como consecuencia se obtienen 2048 descriptores de alto nivel de cada uno de los fotogramas del vídeo.

En la etapa de aprendizaje, la extracción del sonido principal en este método se realiza de la misma forma que en el método anterior. Por cada uno de los fotogramas un músico profesional obtiene la nota más relevante de la armonía que se escucha en el instante de su reproducción. En la etapa de creación, como en el caso anterior, no se lleva a cabo un proceso de extracción de meta-información auditiva puesto que el resultado musical creado por el sistema sustituirá al audio original.

Como resultado final del segundo método de extracción de características de este caso de estudio se obtienen 2048 atributos que describen la imagen y un sonido para cada uno de los fotogramas del vídeo. El conjunto de datos generado se puede consultar en [86].

4.1.2. Aplicación de algoritmos para la generación de melodías

La gran mayoría de procesos artísticos requieren de una fuente de inspiración, y concretamente para los compositores ese estímulo es un factor clave [126]. En los trabajos relacionados con la creatividad computacional, la generación de contenido creativo se lleva a cabo haciendo que las máquinas imiten el comportamiento inventivo de un humano [131]. En este trabajo, se diseña un sistema para la composición descriptiva de música a partir de las características gráficas de un vídeo. Concretamente, el objetivo del trabajo es aplicar el criterio de los animadores de la película *Fantasia* de Disney y realizar el proceso inverso: formalizar el criterio que empleaban para diseñar los personajes y escoger los colores en función de las emociones generadas por la música que pretendían ilustrar y aplicarlo dado un nuevo conjunto de imágenes para generar una secuencia de sonidos. De esta manera, la fuente de inspiración del sistema es el modelo creado por los animadores profesionales para relacionar información gráfica e información auditiva, y la formalización de dicho criterio se realiza mediante algoritmos enmarcados en el área de la IA.

La tarea de aprendizaje supone la aplicación de una serie de algoritmos de clasificación y consta de dos fases bien diferenciadas. La primera de ellas, la fase de entrenamiento, se enmarca en la etapa de aprendizaje de la propuesta (Figura 4.2). Los datos de entrada a los algoritmos cuentan con la información de todos los fotogramas del vídeo correspondientes a la obra *El cascanueces* de la película seleccionada en este caso de estudio. Esta información está compuesta por los atributos extraídos de la imagen (descriptores de forma, disposición y color) y con la nota musical que representa cada uno de los fotogramas, que será la llamada etiqueta o clase del problema de clasifi-

cación, puesto que será el atributo a predecir en la segunda fase del sistema. Con esta información, los algoritmos crean un modelo en el que formalizan el criterio de los animadores profesionales: relacionan los diferentes descriptores de imagen con un sonido concreto. La segunda fase es la de prueba o test, y se enmarca en la etapa de creación del sistema (Figura 4.4). En ella, se considera únicamente la información gráfica de un nuevo vídeo: el fragmento de *El pájaro de fuego* de la película *Fantasia 2000*. Se realiza la extracción de los descriptores de imagen de la misma manera que se realizaba en el vídeo inicial. Esta información es utilizada como entrada para el modelo que se ha creado en la fase anterior, y en este momento es donde se aplica el criterio de los animadores de Disney en el proceso inverso. A partir de una serie de características representativas de las imágenes el modelo será capaz de predecir el sonido más adecuado.

Tras un análisis teórico de las técnicas más utilizadas en problemas de aprendizaje supervisado y considerando la naturaleza de los datos de este caso de estudio se seleccionaron tres algoritmos con el objetivo de optimizar los resultados [72]. En primer lugar se ha aplicado el clasificador probabilístico NB [107]; en segundo lugar, SVM con el objetivo de realizar una separación lineal de los datos mediante hiperplanos [19]; finalmente, también se utiliza una combinación de árboles de decisión, concretamente con el algoritmo RF [22].

En este caso de estudio se proponen dos métodos bien diferenciados para la extracción de descriptores de imagen, y por lo tanto se generan dos conjuntos de datos diferentes. Por este motivo, cada uno de los algoritmos seleccionados se aplica sobre los dos conjuntos de datos obtenidos. Como consecuencia se realizan seis procesos de aprendizaje, cada uno de ellos con su fase de entrenamiento y su fase de prueba.

4.1.3. Resultados y discusión del rendimiento de los algoritmos y análisis de la calidad musical

Dado que el estudio consiste en la automatización de un proceso creativo, para poder evaluar la calidad de la propuesta es necesario analizar dos tipos de resultados. Por una parte, desde un punto de vista técnico es necesario estudiar la eficacia de la máquina. Para ello se considerarán diferentes métricas que permitan valorar el rendimiento de los algoritmos de aprendizaje y la efectividad de la IA en esta solución. Por otra parte, al tratarse de un proceso creativo, la aceptación y satisfacción de los usuarios con los resultados tienen una gran relevancia. Con este objetivo se ha diseñado una encuesta en la que aparecen diferentes fragmentos de un vídeo con melodías compuestas por el propio sistema, y los usuarios deben valorar la adecuación de la música con los elementos visuales. Toda la información relativa a este estudio se puede encontrar detallada en el Apéndice B.

El primer enfoque que se va a realizar para analizar la calidad de los resultados tiene como objetivo medir el rendimiento de los algoritmos utilizados en la etapa de aprendizaje. Para poder cuantificar lo apropiados que son los algoritmos empleados se van a utilizar una serie de métricas de calidad que ofrecen información complementaria sobre la calidad del proceso de clasificación [54]. Es importante recordar que con el primer método de extracción de descriptores gráficos (M1) se obtienen 1198 atributos, y con el segundo método (M2), el número de atributos asciende a 2048. En ambos casos se considera, adicionalmente, un atributo relacionado con el sonido que será el objeto de predicción mediante el algoritmo de IA. Este atributo —o clase— tomará el valor de una de las doce notas musicales de la escala cromática (`Do`, `Do♯`, `Re`, `Re♯`, `Mi`, `Fa`, `Fa♯`, `Sol`, `Sol♯`, `La`, `La♯` o `Si`). El rendimiento de los algoritmos se fundamenta en la calidad de la predicción de la nota que mejor

describe a cada fotograma en función de la información gráfica que contiene y del modelo creado por los algoritmos de aprendizaje automático.

La selección de las métricas adecuadas es un factor crítico para discriminar y obtener el algoritmo óptimo para el proceso de clasificación. Tras un análisis de las diferentes métricas de evaluación existentes y la valoración de su aplicación en este estudio, se han seleccionado algunas que ofrecen información complementaria del éxito y error del entrenamiento de los algoritmos aplicados en este estudio. Toda la información relativa a las técnicas seleccionadas se puede consultar en [61]. La precisión y la sensibilidad (*recall*) son medidas de la exactitud en la tarea de clasificación: examinan cuántas de las instancias devueltas son correctas y cuántos positivos devolvió el modelo respectivamente. La métrica *F-score* combina la precisión y la sensibilidad determinando así un único valor ponderado. El índice Kappa es una evaluación de consistencia que permite ajustar el efecto del azar. La raíz del error cuadrático medio o *root-mean-square error* (RMSE) es una métrica que da información sobre la concentración de datos en torno a su mejor ajuste, y la característica operativa del receptor o *receiver operating characteristic* (ROC) representa los intercambios entre los verdaderos positivos y los falsos positivos. De esta manera, todas estas métricas son analizadas en el proceso de discusión del caso de estudio.

La Tabla 4.3 recoge los valores de las métricas correspondientes a la aplicación de cada uno de los algoritmos (NB, SVM y RF) a los conjuntos de datos extraídos mediante por cada uno de los métodos propuestos en las Secciones 4.1.1.1 (M1) y 4.1.1.2 (M2). El análisis de estos resultados desencadena una discusión sobre la adecuación de ambos métodos de extracción de características en el trabajo y facilita la selección del algoritmo con mejor rendimiento sobre los datos obtenidos en este caso de estudio.

		Precision	Recall	F-score	Kappa	RMSE	ROC
NB	**M1**	0.474	0.429	0.421	0.389	0.324	0.720
	M2	0.552	0.536	0.533	0.464	0.289	0.868
SVM	**M1**	0.795	0.793	0.791	0.767	0.254	0.950
	M2	0.810	0.807	0.807	0.779	0.266	0.948
RF	**M1**	0.843	0.832	0.832	0.807	0.185	0.983
	M2	0.683	0.615	0.598	0.546	0.246	0.930

Tabla 4.3: Rendimiento de los algoritmos NB, SVM y RF para ambos métodos de extracción de descriptores (M1 y M2) en el primer caso de estudio

De acuerdo con los resultados, el rendimiento de los algoritmos es considerablemente divergente para los datos de los dos métodos de extracción de características visuales. Con respecto a la precisión, el *recall* y la métrica *F-score*, NB es el peor clasificador en ambos casos. El índice Kappa muestra que el mejor valor del segundo método de extracción lo logra SVM (0.779); sin embargo, el clasificador RF obtuvo un valor aún menor con los datos extraídos con el algoritmo SIFT (0.807). El mejor valor para RMSE fue obtenido por RF con los datos del primer método de extracción (0.185), demostrando que los sonidos mal clasificados están más cerca del correcto que en los otros casos. RF casi alcanzó el valor máximo para la métrica ROC con los datos del primer método (0.983). Por lo tanto, en términos generales SVM es el mejor clasificador para la meta-información extraída con las CNNs. Sin embargo, el rendimiento del clasificador RF con los datos extraídos por el algoritmo SIFT fue aún mejor.

El segundo factor importante para analizar la validez de la propuesta es la valoración de la calidad descriptiva de la música por parte de los usuarios.

Para poder medir la aceptación social de la relación entre imagen y sonido se ha realizado una encuesta. El vídeo utilizado en la etapa de creación de este caso de estudio se ha dividido en cinco fragmentos. A continuación, tomando esos cinco fragmentos de vídeo, se ha realizado el proceso de composición musical para cada uno de los dos métodos de extracción de descriptores de imagen. Para cada uno de los métodos propuestos se ha utilizado el algoritmo que mejores resultados obtenía; el primer método de extracción se ha combinado con el RF, y el segundo método con SVM. Como consecuencia se han obtenido 10 vídeos con melodías descriptivas compuestas por el sistema, donde los usuarios debían evaluar la calidad descriptiva musical. Para ello, atendiendo a la adecuación entre el color y las formas y el sonido a lo largo de toda la creación debían valorar cada fragmento del 1 al 10. Todos los detalles sobre el diseño y los resultados de la encuesta se pueden consultar en la Sección B.2 del Apéndice B.

Tras la realización de esta encuesta por parte de 47 usuarios, se ha realizado una recopilación y análisis de los resultados. La Figura 4.7 presenta, de forma gráfica, un resumen estadístico de las valoraciones de los usuarios. Concretamente, el gráfico utilizado para la visualización de las valoraciones son los *boxplots* por la gran cantidad de información sobre la distribución de los valores de la variable que ofrece. En este estudio, tal y como se ha comentado anteriormente, para cada uno de los métodos de extracción de decriptores gráficos (M1 y M2) se selecciona el modelo que mejores resultados ofrece. Así, en color naranja se pueden observar los resultados obtenidos para cada uno de los fragmentos compuestos con el primer método de extracción de características (SIFT y BoVW) y el algoritmo RF, y en morado se muestran las valoraciones para los fragmentos compuestos con el segundo método de extracción de características (CNN aplicando TL) y el algoritmo SVM.

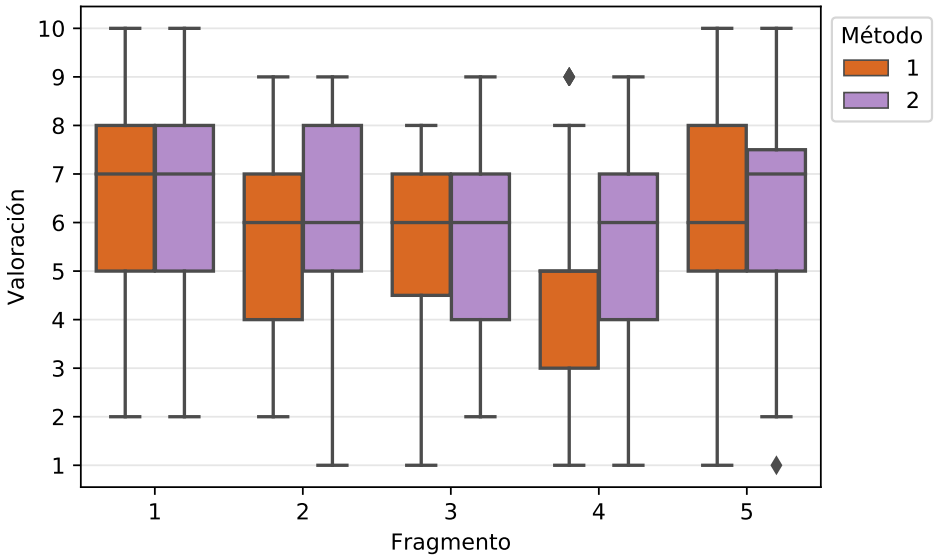

Figura 4.7: *Boxplot* con los resultados de la encuesta para la valoración de la calidad descriptiva de las composiciones musicales en el primer caso de estudio

Tras el análisis de los resultados se obtienen varias conclusiones. En primer lugar, el rango de valoración para cada uno de los fragmentos es muy amplio, llegando a variar entre el valor mínimo (1) y el valor máximo (10) para una misma composición. Esto pone de manifiesto que, como en toda creación musical, existe un componente de percepción que desemboca en un problema de subjetividad. En segundo lugar y a pesar de lo anterior, las valoraciones de todos los fragmentos son similares y demuestran que existe una aceptación social general para el criterio de traducción imagen-sonido establecido en el trabajo. Por último, analizando las valoraciones de ambos métodos se puede determinar que la calidad descriptiva se considera bastante buena en general. Sin embargo, el segundo método de extracción de características (CNN) tiene una mejor aceptación entre los usuarios para los fragmentos presentados en la encuesta.

4.1.4. Conclusiones

La propuesta desarrolla un sistema que compone, de manera automática, melodías descriptivas a partir de un vídeo. Para ello se plantean dos métodos diferenciados por la técnica de extracción de meta-información de la imagen. Por una parte, la aplicación de histogramas de color y del algoritmo SIFT junto con la técnica BoVW permiten la extracción de descriptores de color, forma y disposición de elementos. Por otra parte, la complejidad asociada a tareas de visión artificial se ven reducidas mediante el uso de CNN y TL, obteniendo como resultado una gran cantidad de descriptores de imagen de alto nivel.

Para poder establecer la relación entre imagen y sonido es necesario extraer información auditiva del vídeo. En este trabajo, un músico profesional analizó el vídeo y relacionó cada fotograma con el sonido más destacado en su instante de reproducción. Aunque existen técnicas que permiten realizar una extracción automatizada de una nota o un acorde a partir de una nube de sonidos, se ha considerado que para este trabajo la obtención *manual* de la nota principal era un método más adecuado, a pesar de su complejidad. A la vista de los resultados, se puede considerar un esfuerzo bien invertido.

Tras la composición del conjunto de datos para la fase de entrenamiento se realizó un breve análisis de los datos donde se detectó que una de las doce notas musicales no tenía representación en los datos de entrenamiento y que existía un problema de desbalanceo de los datos. Sin embargo, los resultados obtenidos para las métricas seleccionadas con los diferentes algoritmos aplicados demuestra que los clasificadores se comportan adecuadamente. No obstante, una ampliación del conjunto de datos de entrenamiento podría solucionar los problemas y, probablemente, dar lugar a un mejor comportamiento de los algoritmos de clasificación.

Atendiendo al rendimiento de los algoritmos de clasificación para ambos métodos propuestos, el NB es el que obtiene peores resultados para ambos métodos, mientras que el mejor resultado se obtiene con el algoritmo RF para los datos del primer método de extracción. Considerando los resultados de la encuesta de usuario para medir la calidad descriptiva de la música, las melodías compuestas con los datos extraídos por la CNN (segundo método de extracción de características de la imagen) están ligeramente mejor valoradas. Esto tiene dos lecturas: por una parte, significa que los dos métodos de extracción de descriptores gráficos solucionan el problema planteado en el caso de estudio; por otra parte, demuestra que los resultados obtenidos de trabajos de creatividad computacional no pueden analizarse sólo de manera estadística para medir el rendimiento de la máquina, sino que la percepción, aceptación y gusto del usuario por el material artístico que se crea es un elemento muy importante a tener en cuenta.

Los dos vídeos utilizados en este caso de estudio (el fragmento de *El cascanueces* de la película *Fantasia* en la etapa de aprendizaje y el fragmento de *El pájaro de fuego* de la película *Fantasia 2000*) pertenecen al género de animación y son obra de Disney. En ambos casos existe una gran diversidad de color, forma y estilo musical, lo que demuestra la robustez del modelo de clasificación.

Aunque el movimiento de los personajes no es analizado de manera específica en este trabajo, al considerar el vídeo como una concatenación de fotogramas y analizar las diferentes formas y disposición de elementos en cada uno de ellos se realiza un análisis del cambio de posición, y por lo tanto del movimiento. De la misma forma, aunque no se realice una composición melódica compacta a partir del vídeo completo, la traducción de cada *frame* en un sonido da lugar a una concatenación de sonidos que describen la secuencia de fotogramas. Además, cuando dos o más *frames* consecutivos

dan lugar a la misma nota musical, el resultado musical es un único sonido con una duración proporcional al número de fotogramas que dan lugar a la predicción. Así la melodía está dotada de un ritmo básico que también tiene un componente descriptivo.

4.2. Composición musical armónica a partir de ilustraciones realizadas con tableta gráfica empleando el estilo de la película *Fantasia* de Disney

Este caso de estudio implementa el marco de trabajo propuesto en la Sección 3.2.2, cuyo objetivo es la composición musical descriptiva y automática basada en un proceso ilustrativo. La finalidad del trabajo es que cada usuario pueda hacer uso de un sistema que muestree y analice las ilustraciones que él mismo realiza en una tableta gráfica y componga música de manera dinámica a modo de banda sonora, dando lugar a un proceso de creación cooperativa. Para poder llevar a cabo este caso de estudio se diseña y delimita un escenario de aplicación donde se seleccionan los datos a utilizar y las técnicas y algoritmos más apropiados para optimizar los resultados musicales deseados.

En este caso de estudio se llevan a cabo dos tareas de aprendizaje automático en paralelo. La primera tarea enmarcada en el área de la IA tiene como objetivo establecer un vínculo y extraer un patrón de relación entre elementos gráficos y elementos auditivos de un vídeo. De nuevo, en este caso de estudio se va a extraer información gráfica y auditiva de la película *Fantasia* de Disney [134] con el objetivo de formalizar el criterio de los animadores profesionales [32] y aplicarlo en la fase de creación de manera inversa. El fragmento seleccionado para la extracción de la meta-información necesaria en esta tarea es también el correspondiente a la obra *El cascanueces* de

Tchaikovski, por la variedad de colores, formas y tonalidades de la obra.

La segunda tarea de aprendizaje automatizado está relacionada con la composición musical. En este estudio la música creada no es una simple concatenación de notas o acordes, sino que además contiene información relacionada con la interpretación musical como la dinámica o la agógica y cumple las directrices de un estilo musical concreto. Como resultado, el sistema imitará el estilo de música y el estilo de interpretación de los datos que se utilicen en la fase de entrenamiento. Por ello, el conjunto de datos utilizado en el aprendizaje del sistema debe estar codificado en un formato que comprenda información sobre las frecuencias de las notas musicales e información de interpretación, como el MIDI [47]. El conjunto de datos seleccionado para el entrenamiento de esta parte del sistema consiste en un total de 59 archivos en formato MIDI recopilados de la interpretación de una serie de piezas de un pianista sobre un piano digital. Su intérprete es un pianista estadounidense llamado Kenzie Smith que se dedica a publicar vídeos, partituras y grabaciones en formato digital en canales como Spotify, iTunes, Youtube o Facebook con el objetivo de motivar a sus seguidores a que aprendan a tocar el piano. De todas las grabaciones en formato MIDI que tiene publicadas en su blog, se han seleccionado las bandas sonoras de series de animación anime [117]. De esta manera, las composiciones musicales del sistema estarán enmarcadas en este estilo de música y su reproducción contendrá elementos musicales que recordarán al estilo de interpretación del pianista.

En el flujo de trabajo de este caso de estudio se distinguen, de nuevo, dos etapas. La primera de ellas abarca la extracción de meta-información relevante y la fase de entrenamiento de los algoritmos de aprendizaje. La segunda etapa describe el funcionamiento del sistema una vez entrenados los algoritmos. En ella se utiliza una herramienta desarrollada para la creación y

la edición de ilustraciones mediante tabletas gráficas, y se realiza la extracción de los descriptores de imagen de la ilustración del usuario, aplicando posteriormente los modelos construidos en la etapa anterior para obtener la composición musical. Este proceso de creación se realiza de manera periódica, por lo que se crea un flujo bidireccional de inspiración: la composición musical se va adaptando a los nuevos elementos que el usuario dibuja y la música influye en las decisiones de trazo y color del artista.

La Figura 4.8 muestra las diferentes tareas que se realizan en la primera etapa del sistema, en la que se pueden diferenciar dos tareas de aprendizaje.

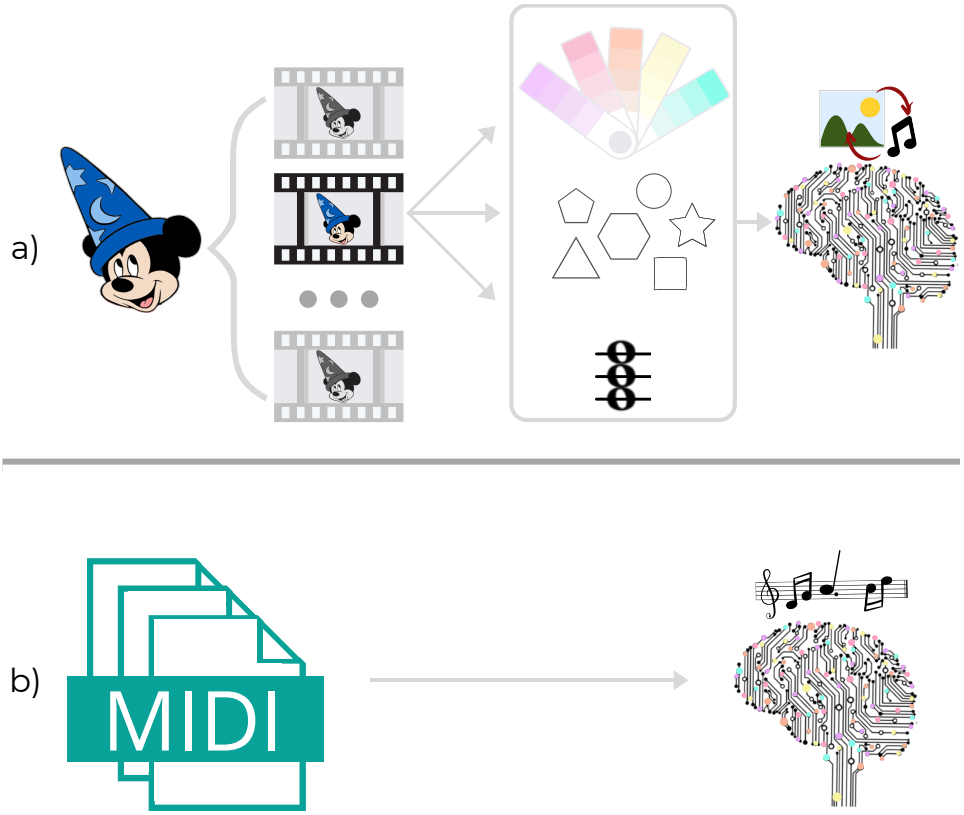

Figura 4.8: Vista global de etapa de aprendizaje del segundo caso de estudio

En primer lugar, la parte **a)** de la figura refleja el proceso diseñado para

establecer un criterio de relación o traducción entre la información gráfica y la información auditiva del vídeo inicial; en segundo lugar, en la parte **b)** se describe un proceso de aprendizaje para la composición musical polifónica en la que el algoritmo de IA toma como datos de entrada un conjunto de piezas musicales en formato digital para extraer patrones como la duración de las notas, las secuencias de notas más repetidas o la relación entre notas musicales.

Para la primera tarea de aprendizaje automático, como se ha comentado anteriormente, el fragmento de *El cascanueces* de la película seleccionada es el punto de partida de este trabajo. El vídeo se divide en fotogramas, seleccionando el primero de cada ocho y descartando el resto. Como resultado, por cada segundo se consideran 3 fotogramas, lo que da lugar a un total de 2575 *frames* para la etapa de entrenamiento. De cada una de estas imágenes se extraen tres tipos de características: descriptores del color, de la forma y distribución de los elementos, y el acorde de tres notas que tienen mayor relevancia armónica en el momento que el fotograma se está reproduciendo. Esta información es la entrada para el primer algoritmo de IA que desarrolla una tarea de clasificación, dando lugar a un modelo cuyo objetivo es relacionar los descriptores de la imagen (atributos) con el acorde de los tres sonidos (clase). En esta primera etapa tiene lugar otra tarea de aprendizaje adicional que se desarrolla de manera paralela a la anterior. En este caso, una serie de ficheros codificados en formato MIDI sirven de entrada a un segundo algoritmo de IA que busca patrones de repetición y relación entre las notas musicales. Como consecuencia se obtiene un segundo modelo que se utilizará para la composición automática de música.

El DAG de tareas de esta primera etapa del estudio se refleja en la Figura 4.9. Para establecer una relación entre atributos gráficos y musicales a partir del vídeo inicial, es necesaria la extracción de esta información. Pa-

ra la extracción de color en este caso de estudio se aplica una técnica de cuantificación mediante k-Means [26], y los descriptores de forma y disposición de elementos se obtienen mediante el algoritmo SIFT [80]. En este caso, la extracción de información musical conlleva la obtención de un acorde en lugar de una nota aislada, y se realiza de manera automática mediante la combinación de las técnicas *Harmonic Percussive Source Separation* (HPSS) [48] y CENS [92]. Puesto que este proceso de análisis conlleva la predicción de tres notas musicales (un acorde tríada) se trata de un problema de clasificación multi-etiqueta, que se va a resolver con los algoritmos RAKEL [129] y ML-kNN [140]. La segunda tarea de aprendizaje de esta etapa aborda el problema de la composición musical automática, y se lleva a cabo mediante una RNN basada en LSTM [116]. El conjunto de datos utilizado para su aprendizaje es una serie de bandas sonoras de películas de estilo anime interpretadas en un teclado con salida digital [117], dando lugar a un conjunto de datos expresados en MIDI.

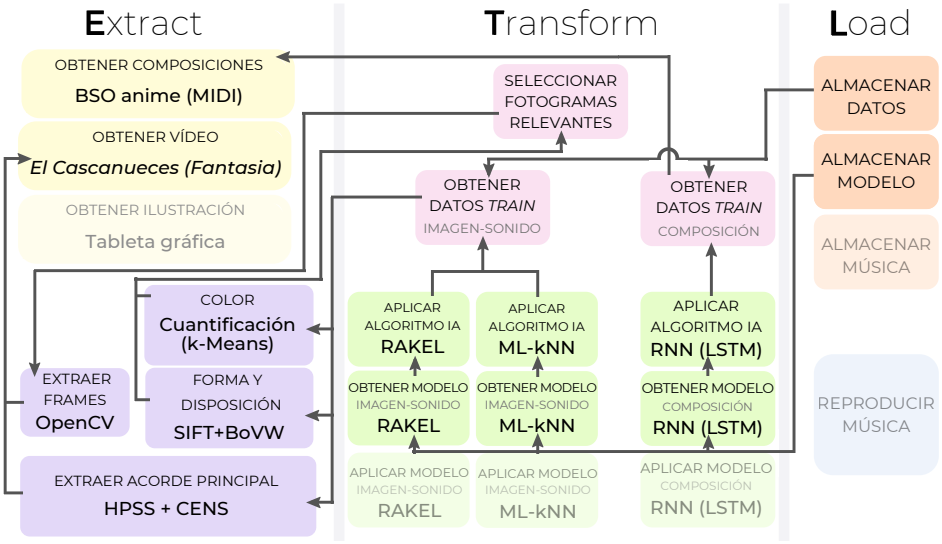

Figura 4.9: Grafo de tareas para la etapa de aprendizaje del segundo caso de estudio

La Figura 4.10 muestra el flujo de trabajo de la etapa de creación del

sistema. Para abordar el problema, este trabajo involucra tanto al usuario como a la máquina en el proceso creativo. Por ello, se ha desarrollado una herramienta para la creación y edición de imágenes mediante una tableta gráfica. La ilustración del usuario es el punto de partida de esta segunda etapa. Tras la extracción de descriptores de la imagen (color, forma, disposición de los elementos...) se aplica el primer modelo creado en la etapa anterior para la predicción del acorde de tres notas más adecuado. Este acorde sirve como entrada para el segundo modelo de la fase anterior. Así, aunque la música tiene un estilo y unos matices de interpretación predefinidos por el conjunto de datos utilizado en el entrenamiento, se establece un marco armónico para la composición musical en función de las características de la ilustración. Este proceso se repite cada 10 segundos, recogiendo información de la ilustración periódicamente para la composición musical dinámica a modo de banda sonora.

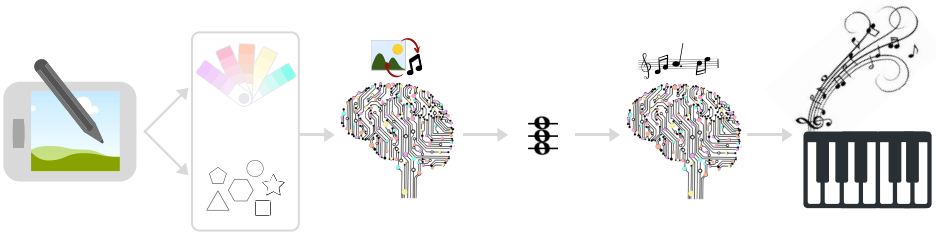

Figura 4.10: Vista global de etapa de creación del segundo caso de estudio

Este flujo de trabajo se representa en la Figura 4.11 a modo de DAG. En esta etapa del caso de estudio, cabe resaltar que el punto de partida no es un vídeo sino una ilustración que el usuario realiza en una tableta gráfica, por lo que se hace necesario implementar una herramienta que permita que el usuario gestione el proceso ilustrativo y que recupere periódicamente la ilustración en formato digital. En este punto se extraen los descriptores gráficos de la ilustración con las mismas técnicas que en la etapa anterior y se aplican los modelos obtenidos con los algoritmos RAKEL y ML-kNN para

predecir el acorde que mejor describe la imagen. Esta información musical se toma como entrada para el segundo modelo del sistema, permitiendo que la composición musical esté condicionada por la armonía previamente inferida. Finalmente, tras la obtención de la creación artística, esta se reproduce con la intención de que pueda ser escuchada por el usuario.

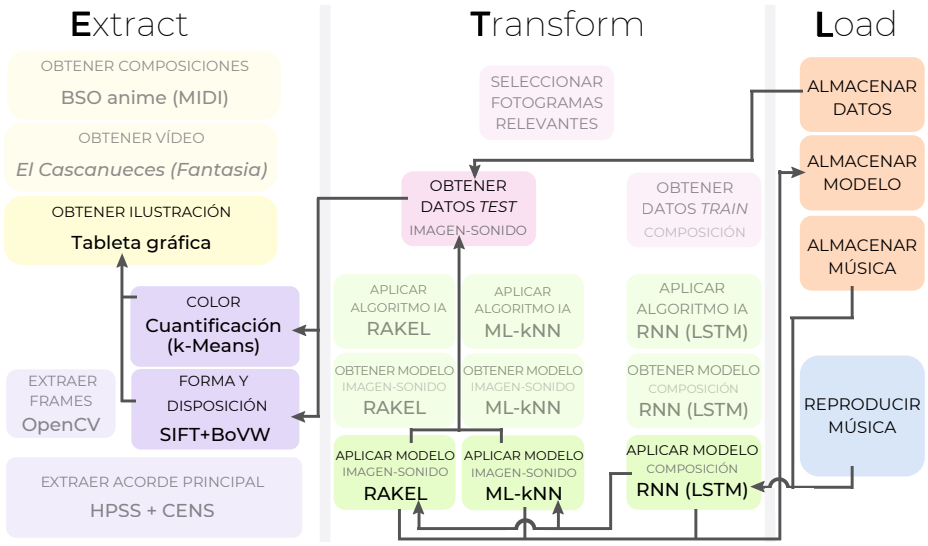

Figura 4.11: Grafo de tareas para la etapa de creación del segundo caso de estudio

La Sección 4.2.1 detalla el proceso de extracción de meta-información visual y auditiva que se lleva a cabo en el caso de estudio. En la Sección 4.2.2 se describe la selección y aplicación de algoritmos de aprendizaje automático que darán pie a la composición musical. La Sección 4.2.3 presenta y analiza los resultados obtenidos en el caso de estudio desde una perspectiva más técnica y desde un enfoque musical, y la Sección 4.2.4 desarrolla las conclusiones extraídas del estudio.

4.2.1. Obtención y limpieza de los descriptores de imagen y meta-información musical

En este caso de estudio es necesaria la obtención de características de dos tipos: descriptores de imagen y meta-información musical. Esta división va a ser la que marque la estructura de esta subsección con el objetivo de facilitar la comprensión del proceso de extracción necesario en cada una de las etapas del sistema.

La extracción de descriptores de imagen se lleva a cabo tanto en la etapa de aprendizaje como en la etapa de creación. En la primera de ellas, el fragmento de la película correspondiente a la obra *El cascanueces* [134], se divide en fotogramas haciendo uso de la librería de visión artificial OpenCV [21], considerando en este caso el primero de cada ocho. Como resultado, en este trabajo se utilizan 2574 *frames*, para cada uno de los cuales se debe extraer información relativa al color, forma y disposición de los elementos con el objetivo de establecer una relación con la información musical. El Algoritmo 3 muestra el proceso para la selección de fotogramas en este segundo caso de estudio.

Algoritmo 3 Obtención de fotogramas representativos de un vídeo

Input $fragmento$: vídeo de partida para la obtención de fotogramas
Output $fotogramas[\,]$: conjunto de fotogramas representativos

$fotogramas\_raw[\,] \longleftarrow ObtenerFotogramas(fragmento)$

for $i : 1$ **to lenght**$(fotogramas\_raw[\,])$ **do**
  if $i \,\%\, 8 = 0$ **then**
    append $fotogramas\_raw[i]$ **to** $fotogramas[]$
  end if
end for

return $fotogramas[\,]$

En la segunda etapa del sistema, la ilustración del usuario es la fuente de extracción de características gráficas. Para ello, se desarrolla una herramienta básica de diseño para realizar y editar ilustraciones con la tableta gráfica y recuperarlas en formato digital. El proceso de extracción de meta-datos a partir de la ilustración se realiza periódicamente, cada 10 segundos, a fin de realizar todo el proceso de creación y que la música esté continuamente adaptándose a los trazos y los colores que el usuario emplea en su obra.

La información gráfica relativa a la forma y la distribución de los elementos en la imagen se ha obtenido mediante la combinación del algoritmo SIFT [80] y la técnica BoVW [115]. Para cada imagen se extrae información sobre sus puntos clave y el área de píxeles que los rodean mediante el algoritmo SIFT. La aplicación de este algoritmo en el trabajo presente se debe, por una parte, a su robustez y la invariabilidad de las características extraídas a las traslaciones, rotaciones y transformaciones, y por otra parte a que, a pesar de ser más lento que otros algoritmos, los resultados obtenidos son más precisos [70]. Posteriormente, las características se agrupan para la obtención de 50 palabras visuales o *visual words*, y mediante la técnica de BoVW, cada imagen se representa como un vector que contiene un recuento ponderado de cada palabra visual.

La extracción de color en este caso se lleva a cabo mediante un algoritmo de cuantificación basado en *clustering*. Concretamente, el algoritmo K-Means se aplica a los píxeles de cada imagen para reducir la información del color y extraer los colores más relevantes [26]. Para este trabajo, la reducción del número de colores de una imagen no supone una gran pérdida de información, sino que la simplificación de la gran cantidad de información relativa al color que contienen los píxeles de una imagen facilita la tarea de relacionar atributos gráficos y auditivos. En este caso, se ha decidido que la información cromática de cada imagen quede sintetizada con sus tres colores

más destacados.

La Figura 4.12 ilustra un ejemplo de extracción de descriptores gráficos sobre un fotograma del fragmento de la película *Fantasia* seleccionado para la etapa de aprendizaje. En primer lugar se muestra el fotograma original (**a**), en segundo lugar se presenta la información relativa a forma y disposición de elementos obtenida mediante el algoritmo SIFT (**b**) y por último, el fotograma se representa con los tres colores proporcionados por la cuantificación (**c**). Estos colores se corresponden con las siguientes codificaciones en RGB: (2,29,71), (126,140,129) y (36,89,91).

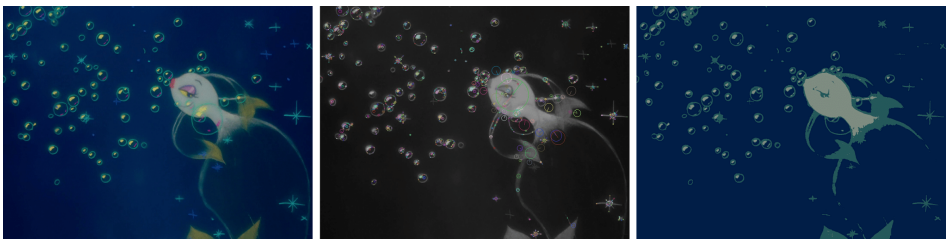

Figura 4.12: Descriptores de la imagen sobre un *frame* del fragmento *El Cascanueces* de la película *Fantasia* en el segundo caso de estudio. La Figura a) contiene el fotograma original. La Figura b) muestra los descriptores SIFT extraídos. La Figura c) presenta el fotograma original con la cuantificación de color

El propósito de la extracción de características auditivas en este trabajo es obtener las tres notas que tienen una mayor influencia armónica en la nube de sonidos que se escucha durante la reproducción de cada fotograma. El método seleccionado para la extracción automática de esta información hace uso de un cromagrama, que es una representación del espectrograma del sonido que identifica doce perfiles correspondientes a las doce notas musicales de la escala cromática. Concretamente, se utiliza el método CENS por su facilidad para obtener características del audio relacionadas con la progresión armónica de la señal auditiva [92]. Estas características presentan un alto

nivel de robustez a las variaciones como la dinámica, el timbre, la articulación y las desviaciones del tempo local. En el estudio se utiliza la implementación de la bilbioteca LibROSA [88].

Para mejorar la calidad de las características del croma, el audio se descompone en primer lugar mediante la separación de fuentes armónica y rítmica conocida como *Harmonic Percussive Source Separation* (HPSS) [48], dando lugar a dos componentes musicales bien diferenciados. Por una parte, se obtienen los componentes rítmicos, de poco interés en este trabajo; por otra parte, los armónicos, que son los únicos considerados para la obtención del cromagrama.

En la Figura 4.13 se puede visualizar el vector de croma extraído para el fragmento correspondiente a un fotograma de la *Danza Rusa* (perteneciente a la obra musical *El Cascanueces*) de la película analizada. En él se puede visualizar la energía de cada una de las notas musicales a lo largo de 1s del fragmento musical de acuerdo a una escala de color.

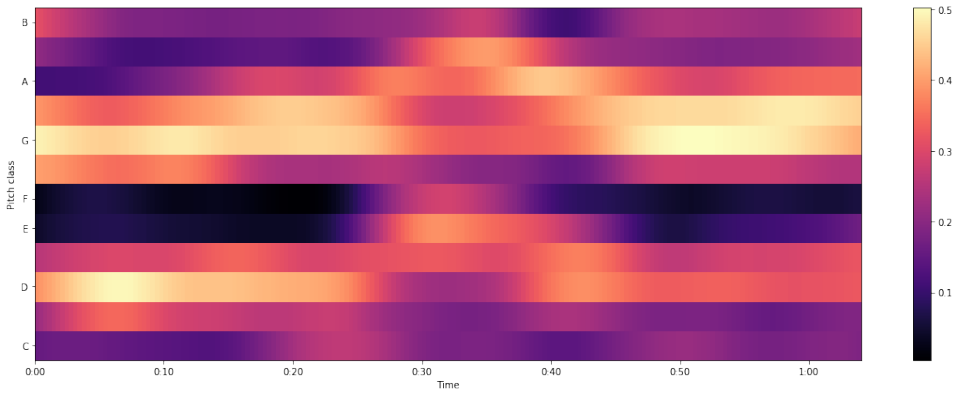

Figura 4.13: Vector de características de croma obtenido por el método CENS en la *Danza Rusa* de la obra musical *El Cascanueces*

De esta tarea de extracción de meta-información auditiva se obtiene, para cada fotograma del vídeo, un vector que contiene la intensidad en un rango

entre [0,1] para cada una de las notas musicales de la escala cromática. La señal auditiva del vídeo se muestrea a $22050Hz$, y el número de muestras (o columnas del espectrograma) se ha fijado en $2^6 * 115$, por lo que existe una separación entre cada fragmento auditivo analizado de $334ms$, que coincide con el muestreo de 3 fotogramas por segundo del vídeo. Para obtener un acorde con las tres notas más relevantes del fragmento auditivo se seleccionan los tres perfiles del vector que presentan una mayor intensidad en cada vector de croma.

Como resultado final de este proceso, cada imagen (un fotograma del vídeo en la etapa de aprendizaje y una captura de la ilustración del usuario en la etapa de creación) se expresa como un vector de características relacionadas con la forma, la distribución de los elementos y el color. Concretamente, el vector está compuesto por 50 *visual words* y 3 colores RGB. En las imágenes utilizadas en la etapa de aprendizaje a esta información se le añaden las tres notas musicales que representan cada uno de los fotogramas. El conjunto de datos generado para el entrenamiento del modelo que relaciona características visuales e información musical se puede encontrar en [87].

4.2.2. Algoritmos de aprendizaje automático para la composición descriptiva

Para poder llevar a cabo la composición musical descriptiva, este caso de estudio involucra dos procesos paralelos de aprendizaje automático. El primero de ellos tiene como objetivo la creación de un vínculo entre la información gráfica de la ilustración y la información auditiva que se va a componer. En el segundo, se establece como meta el proceso de composición musical como tal, atendiendo no sólo a aspectos armónicos y melódicos, sino

también a ciertos detalles de interpretación de la obra.

La asociación entre elementos gráficos y elementos musicales se establece atendiendo al criterio establecido por una serie de animadores profesionales de Disney. Igual que en el caso de estudio anterior, un fragmento de vídeo de la película de Disney será el punto de partida de este trabajo. Aprovechando el concepto que aplica Disney para la creación de esta película donde los elementos gráficos se diseñan y se crean para describir los detalles musicales de una serie de obras de música clásica, este análisis permitirá realizar la traducción en sentido inverso. Para ello, los descriptores de imagen que se extraen de cada uno de los fotogramas del fragmento seleccionado junto con la información musical extraída de la señal auditiva se utilizan como entrada para el algoritmo de aprendizaje automático en su fase de entrenamiento. Puesto que el objetivo del sistema es predecir el acorde que mejor describe una determinada imagen, los descriptores gráficos son los atributos en este problema, y las tres notas musicales actúan como la clase. Como resultado se obtiene un modelo donde el criterio diseñado por los animadores para relacionar imagen y sonido queda formalizado.

En este caso la predicción no es un dato simple, sino que las características de cada imagen se utilizarán para predecir el acorde o conjunto de las tres notas más apropiadas. Esto es un problema de clasificación multi-etiqueta, donde cada instancia (cada imagen) se clasifica con tres valores diferentes y no excluyentes [128]. Tras un análisis de los datos y de los algoritmos más indicados para este tipo de problemas, se decide aplicar el RAKEL [129] y el ML-kNN [140].

RAKEL es un método para la clasificación multi-etiqueta que se basa en la combinación de múltiples algoritmos de aprendizaje simples para optimizar el rendimiento predictivo [129]. Cada uno de estos clasificadores de una

etiqueta se entrenan con pequeños subconjuntos aleatorios de etiquetas. Debido a su adecuación a los datos de este trabajo y al problema, se emplea RF como clasificador de una sola etiqueta [72]. En este experimento, el tamaño de los subconjuntos (k) es 6, y se realizan 10 iteraciones (m). Además, el número de árboles generados para RF es de 100 en este experimento.

Por su parte, ML-kNN es un método de clasificación multi-etiqueta basado en el algoritmo k-Nearest Neighbors (kNN) [140]. Este método busca los k vecinos más cercanos de una instancia y en función de sus conjuntos de etiquetas se aplica el principio *Maximum A Posteriori* (MAP) para determinar el conjunto de etiquetas de la nueva instancia. Concretamente, en el trabajo se aplica el algoritmo para un valor de $k = 3$ y considerando la distancia de coseno entre las diferentes instancias.

Finalmente, en la etapa de creación del sistema hace uso del modelo obtenido para predecir el conjunto de tres notas musicales que mejor describe la ilustración del usuario. Esta vez, el modelo necesita como entrada únicamente los descriptores de imagen, por lo que cada vez que se quiere componer música es necesario que se realice el proceso de extracción de descriptores gráficos de la ilustración. Los dos algoritmos seleccionados para la creación del modelo se aplican sobre estos datos, y como consecuencia, se completan dos procesos de aprendizaje, cada uno con su fase de entrenamiento y su fase de prueba.

Ádemás de formalizar el patrón de relación entre imagen y sonido, este trabajo incluye un segundo proceso de aprendizaje automático que consiste en la creación de música polifónica. El objetivo de este sistema no es únicamente crear armonías o secuencias de notas, sino que cuenten con detalles de interpretación que hagan que su reproducción las aleje de parecer una creación sintética. Para ello se utiliza un conjunto de archivos en formato MIDI

que se corresponden con la interpretación de varias bandas sonoras de series de estilo anime. Cada uno de estos archivos es un flujo de eventos musicales que capturan el inicio y el final de una nota, su velocidad y otros detalles relativos a la interpretación del músico, como las dinámicas y las agógicas. Para contar con una mayor cantidad de información en la fase de entrenamiento se ha aplicado un proceso de aumento de datos, creando ejemplos adicionales mediante técnicas como la transposición musical y el cambio de ritmo en las obras originales.

Para realizar este proceso de composición musical se utiliza una RNN basada en LSTM. Específicamente, en este trabajo se utiliza el modelo Magenta de Performance-RNN de Google [116]. Aunque la arquitectura de este modelo puede ser modificada, está formada por 3 capas con 512 bloques LSTM por defecto. La RNN modela motivos musicales y características interpretativas como el fraseo y la dinámica. Así, los tiempos y velocidades de las notas se basan en la actuación humana, lo que hace que la máquina parezca una persona que compone e interpreta la música. Para evitar problemas de sobreajuste, se emplea una arquitectura de codificador-decodificador [12]. Además, una vez que la red ya está entrenada, se puede modificar un parámetro llamado *temperatura* para regular la aleatoriedad de la música generada.

El modelo generado por la red neuronal no necesita ninguna fuente de inspiración para empezar a componer música. Sin embargo, en este trabajo, la composición musical está determinada por el acorde obtenido a partir de la ilustración del usuario. Así, cuando el usuario comienza a dibujar, cada 10 segundos se extraen las características visuales y se obtiene un acorde. Este acorde se utiliza como entrada para el modelo de composición musical determinando el marco tonal de la composición de la red. Como resultado se obtiene una composición inspirada en el color, la forma y la distribución de

los elementos del dibujo del usuario.

4.2.3. Presentación y análisis de los resultados desde un punto de vista técnico y artístico

El objetivo de este caso de estudio es diseñar un sistema creativo para la composición automática de música descriptiva a partir de las ilustraciones de un usuario mediante una tableta gráfica. La Figura 4.14 muestra un ejemplo del proceso creativo del sistema; en primer lugar se puede observar la ilustración realizada por uno de los usuarios que probaron el sistema, y a su derecha se puede visualizar la partitura de la composición musical realizada por el sistema.

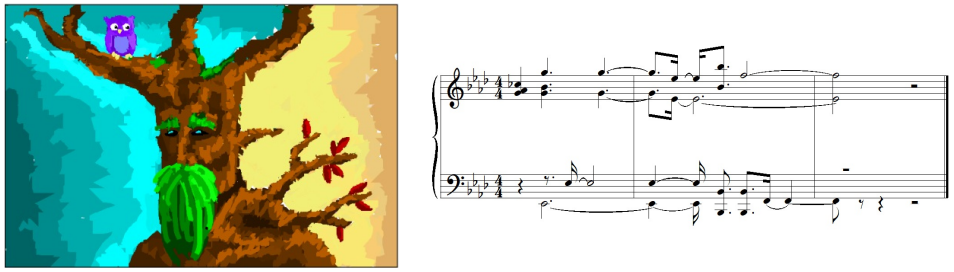

Figura 4.14: Ilustración realizada por el usuario con una tableta digital y grafía de la composición descriptiva del segundo caso de estudio

Para poder evaluar la calidad de los resultados obtenidos y el grado de consecución de los objetivos de este trabajo es necesario realizar dos tipos de análisis. Por una parte, se analizará la eficacia de la máquina en el proceso de aprendizaje automático mediante la obtención de una serie de métricas que reflejen datos estadísticos del rendimiento de los diferentes algoritmos. Por otra parte, dado que el resultado del sistema es una creación artística, se considera necesario realizar un análisis de la aceptación del usuario. Para ello se emplean dos tipos de encuestas que permiten discutir la capacidad

creativa de la máquina y la validez del vínculo establecido para relacionar las ilustraciones con la música compuesta.

En primer lugar se evalúa el rendimiento de los algoritmos utilizados para formalizar el vínculo entre los descriptores de la imagen y el sonido. Para realizar esta tarea se utilizan dos clasificadores multi-etiqueta: RAKEL [129] y ML-kNN [140]. En ambos casos se utiliza una técnica de validación cruzada de 10 iteraciones, con una división de los datos correspondiente al 80 % para el conjunto de entrenamiento y el 20 % restante para el conjunto de prueba.

Para poder determinar la eficacia de los procesos de aprendizaje utilizados en esta propuesta se calculan algunas métricas que permiten describir la calidad de los algoritmos en el proceso de clasificación. Las métricas utilizadas en este estudio son diferentes a las utilizadas en el caso de estudio anterior puesto que los problemas de clasificación multi-etiqueta requieren de un análisis diferente al que se aplica en la clasificación con una única etiqueta. La selección de las métricas más adecuadas, en este caso, se ha realizado atendiendo a las recomendaciones de [128].

La métrica *F-score* representa la media armónica de la precisión y la sensibilidad, y *exact match* muestra el porcentaje de instancias en las que todas las etiquetas han sido clasificadas correctamente. Por otra parte, se emplean tres métricas relacionadas con los errores de clasificación, y cuyo valor debe acercarse a 0. *Hamming loss* es una métrica que captura la fracción de etiquetas mal clasificadas, *one-error* evalúa la frecuencia de la etiqueta mejor clasificada que no estaba en el conjunto de etiquetas correctas de la instancia y *ranking loss* mide la fracción media de pares de etiquetas que están ordenadas incorrectamente para la instancia.

La Tabla 4.4 muestra los valores de las métricas anteriormente seleccionadas para cada uno de los dos algoritmos (RAKEL y ML-kNN) empleados

en la tarea de aprendizaje automático para relacionar descriptores gráficos e información musical.

	F-score	**Exact match**	**Hamming loss**	**One-error**	**Ranking loss**
RAKEL	0.796	0.571	0.092	0.146	0.093
ML-kNN	0.610	0.285	0.150	0.204	0.131

Tabla 4.4: Rendimiento de los algoritmos RAKEL y ML-KNN en el segundo caso de estudio

Los resultados obtenidos por ambos algoritmos son considerablemente diferentes. Las métricas *F-score* y *exact match* muestran que la exactitud del algoritmo RAKEL, con valores de 79,6 % y 57,1 % respectivamente, es considerablemente mejor que la del algoritmo ML-kNN, con valores 61 % y 28,5 %. Adicionalmente, las métricas de error revelan menos fallos en el algoritmo RAKEL. Concretamente, la métrica *hamming loss* muestra que únicamente el 0,92 % de etiquetas se clasifican erróneamente con RAKEL, mientras que el valor de esta métrica para ML-kNN alcanza el 15 %. Además, la diferencia en los valores de *one-error* y *ranking loss* en ambos casos es del 8 %.

Adicionalmente se ha analizado la exactitud de cada uno de los algoritmos utilizados para la clasificación multi-etiqueta con cada una de las posibles clases consideradas en este problema. La Tabla 4.5 muestra el porcentaje de instancias que se predijeron correctamente en la etapa de prueba para cada una de las notas musicales.

	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
RAKEL	0.94	0.91	0.87	0.88	0.92	0.96	0.90	0.91	0.92	0.89	0.88	0.91
ML-kNN	0.90	0.85	0.78	0.81	0.87	0.94	0.85	0.83	0.88	0.83	0.82	0.86

Tabla 4.5: Precisión de los algoritmos RAKEL y ML-KNN en la predicción de cada nota musical

En el caso del algoritmo RAKEL, la exactitud es superior al 86 % en todas las etiquetas, llegando a alcanzar el 96,3 % para la nota musical `F`, y el 93,9 % para la nota `C`. ML-kNN alcanza valores ligeramente menores en todos los casos, comprendidos entre el 78,3 % y el 93,5 %. Como conclusión, el algoritmo RAKEL muestra un mejor rendimiento y por ello es el elegido para obtener el modelo que relaciona características visuales y sonoras de la película *Fantasia* de Disney.

Para evaluar la calidad de la música compuesta por el sistema y su capacidad creativa se realizó un test de Turing [36]. Todos los detalles de esta prueba están recogidos en la Sección B.3 del trabajo. El objetivo de esta prueba es determinar si la máquina es lo suficientemente creativa como para hacer que un ser humano no pueda distinguir sus piezas musicales de las de un compositor profesional. De manera adicional, se solicita a los usuarios que evalúen las composiciones musicales para obtener información relacionada con la aceptación social de las composiciones.

En esta prueba se presentan 20 fragmentos musicales de 10 a 20 segundos de duración, de los cuales la mitad están creados por un compositor profesional y la otra mitad, por una máquina. La encuesta ha sido realizada por 46 usuarios que, tras escuchar cada composición musical, han indicado cuáles estaban compuestas por una persona y cuáles por la máquina y han calificado entre 1 (desagradable) y 10 (muy agradable) la calidad musical según su percepción y su gusto personal.

La Figura 4.15 muestra la tasa de éxito para cada uno de los fragmentos de la encuesta; es decir, el porcentaje de personas que acertaron (en verde) y que erraron (en rojo) en la diferenciación de la naturaleza del compositor de cada fragmento. La figura de la izquierda (**a**) muestra los porcentajes de acierto y fallo para cada uno de los diez fragmentos musicales creados por un

compositor profesional, mientras que la figura de la derecha (**b**) muestra los porcentajes de acierto y fallo para los fragmentos compuestos por el sistema. En ambas figuras el color verde representa acierto y el rojo, fallo.

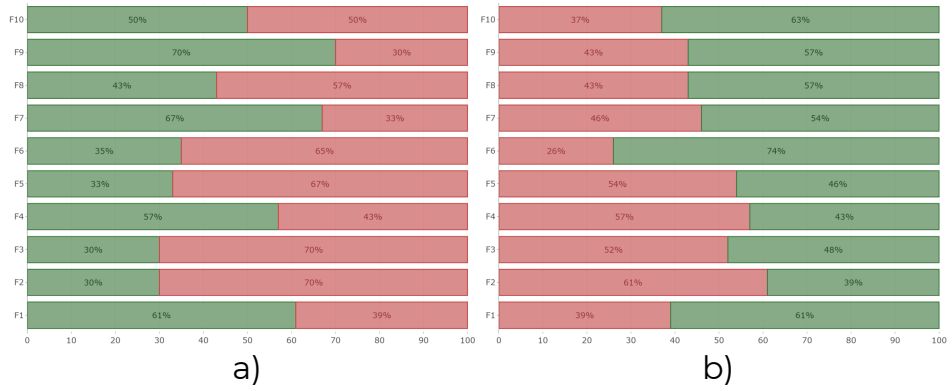

Figura 4.15: Resultados de las tasas de acierto para las creaciones musicales del compositor profesional (a) y de la máquina (b)

De los resultados obtenidos en el test de Turing se pueden extraer dos conclusiones. A la vista de los resultados de las composiciones musicales creadas por el compositor profesional, en muchos casos las personas no pudieron distinguir el tipo de compositor, ya que el porcentaje de éxito de cinco fragmentos musicales es inferior al 50 %. Además, la tasa de éxito de cuatro de los fragmentos compuestos por la máquina es inferior al 50 %, y otros cinco tienen una tasa de aciertos comprendida entre el 50 % y el 60 %. Esto significa que un porcentaje cercano al 50 % de las personas que hicieron la prueba pensaron que estos fragmentos musicales estaban compuestos por un ser humano y no por una máquina, y por lo tanto la diferenciación entre máquina y compositor podría tener cierto componente de aleatoriedad.

Las Figuras 4.16 y 4.17 muestran las valoraciones de los usuarios para cada uno de los diez fragmentos creados por el compositor y por la máquina respectivamente.

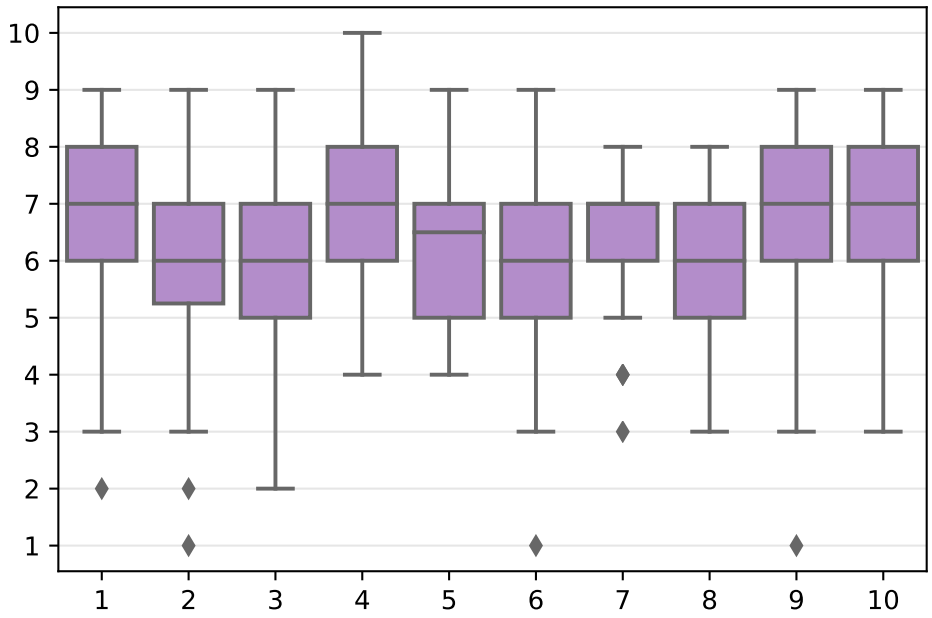

Figura 4.16: Resultados del Test de Turing para las creaciones musicales de un compositor profesional en el segundo caso de estudio

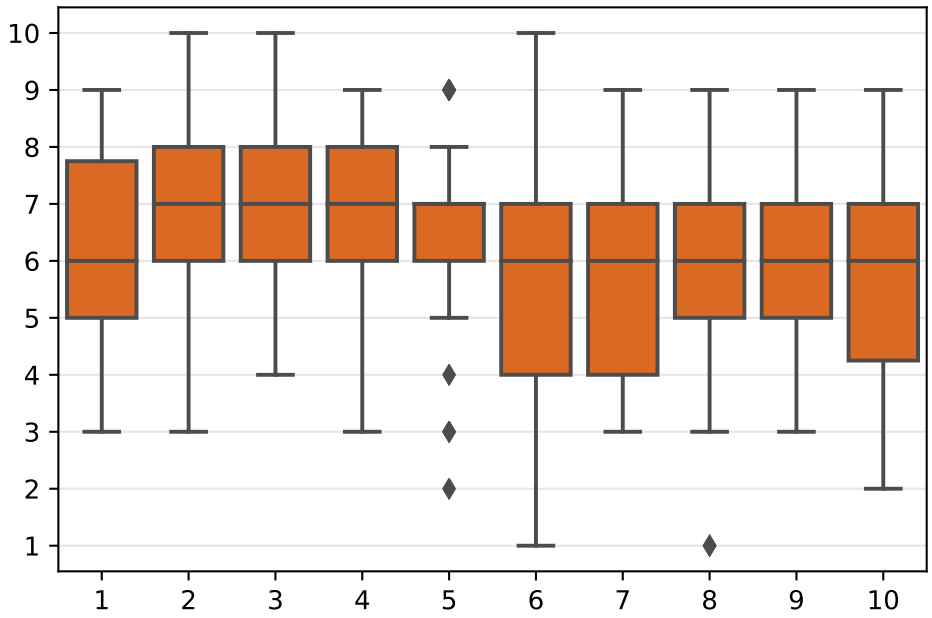

Figura 4.17: Resultados del Test de Turing para las creaciones musicales de la máquina en el segundo caso de estudio

La mediana de las valoraciones es muy similar para todos los fragmentos en ambos casos, tomando en todos los casos valores 6 y 7. Sin embargo, la dispersión estadística (H-Spread) muestra que, en términos generales, la diferencia entre la valoración máxima y mínima de las composiciones musicales de la máquina es mayor que la de las composiciones humanas. Esto significa que, aunque el valor mediano es menor en algunos casos, algunas personas dieron a la máquina una mejor valoración que al compositor. Por otra parte, esto pone de manifiesto que existe un componente subjetivo que toma mucha fuerza en este tipo de trabajos artísticos, ya sea creados por personas o por máquinas.

Con el objetivo de comprobar si existe una dependencia entre el compositor de cada obra (músico profesional o máquina) y las valoraciones que los usuarios dan a cada uno de los fragmentos, se ha aplicado un test U de Mann Whitney [58]. La hipótesis nula establece que las valoraciones de la música son iguales, independientemente de la naturaleza de su compositor. Partiendo de los datos disponibles en la Sección B.3.2 del Apéndice B se aplica un contraste de comparación de la tendencia central con un nivel de significación α =0.01, considerando que si las valoraciones son iguales para los dos tipos de compositores, las variables no serán dependientes. Tras la realización del test, se obtiene un p-valor de 0.023 que nos permite aceptar la hipótesis nula, concluyendo que no se detectan diferencias entre las valoraciones de las piezas del compositor y la máquina. Este análisis respalda el hecho de que la máquina desarrolla ciertas habilidades inteligentes y creativas para la composición de música.

Finalmente se diseñó una encuesta para evaluar el vínculo establecido entre la imagen y el sonido (todos los detalles del diseño y los resultados de este experimento se pueden encontrar en la Sección B.4). De entre los usuarios que probaron la herramienta desarrollada para crear y editar ilus-

traciones mediante el uso de una tableta gráfica y pudieron experimentar con la composición descriptiva del sistema, se seleccionó a 10 personas para calificar del 1 (totalmente en desacuerdo) al 10 (totalmente de acuerdo) una serie de preguntas con el fin de medir la calidad de la interconexión entre las características visuales (color, forma y disposición) de los elementos que dibujaban y la música que componía el sistema. Las preguntas empleadas en este experimento fueron las siguientes:

- **Q1**: ¿Cree que existe una adecuación entre su dibujo y los sonidos generados?
- **Q2**: ¿Cree que el color influye en los sonidos de la composición?
- **Q3**: ¿Cree que la forma influye en los sonidos de la composición?
- **Q4**: ¿Le ha influido la música a la hora de dibujar?

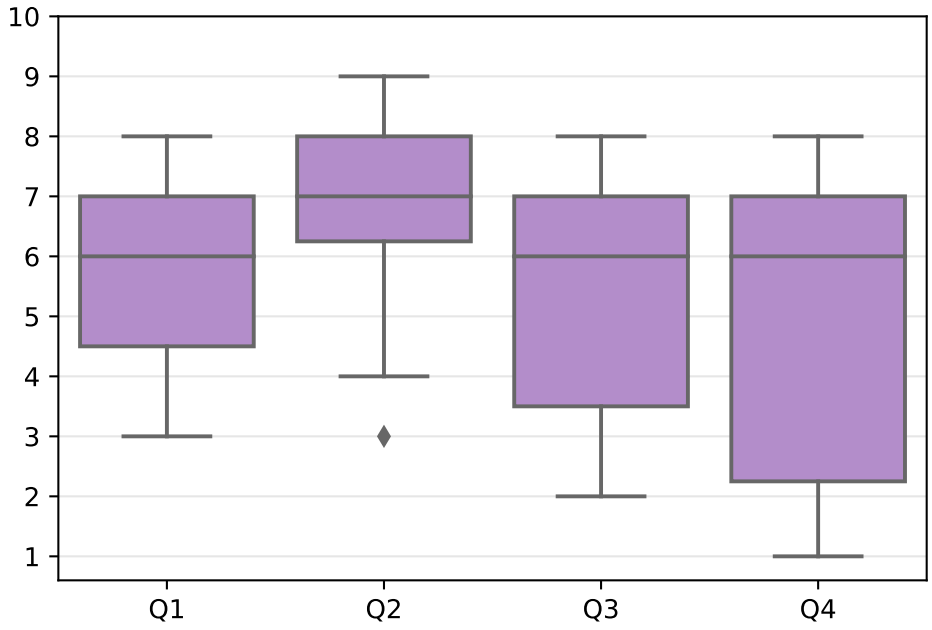

Figura 4.18: Resultados de la encuesta para evaluar la relación entre imagen y sonido en el segundo caso de estudio

En términos generales, los usuarios que han realizado la encuesta consideran que sí existe una buena relación entre los elementos gráficos de sus ilustraciones y la música compuesta por el sistema. Los resultados de las preguntas Q2 y Q3 muestran que, según la percepción de los usuarios, el factor más influyente del vínculo establecido es el color, por encima de la forma y la disposición de los elementos. Finalmente, los resultados para la pregunta Q4 ponen de manifiesto que la propuesta crea un flujo bidireccional entre el usuario y la máquina durante el proceso creativo. Esto significa que la composición musical está determinada por la características visuales de los dibujos del ususario, al mismo tiempo que el usuario se ve influido por la música a la hora de crear la ilustración. Algunos usuarios, además, trataron de encontrar una nube de sonido específica con sus ilustraciones.

4.2.4. Conclusiones

Esta propuesta desarrolla un sistema para la composición automática de música descriptiva. En este caso, la fuente de inspiración o el elemento a describir es una ilustración que el usuario realiza por medio de un dispositivo digital como una tableta gráfica. Para ello se llevan a cabo dos procesos bien diferenciados de aprendizaje automático. El primero de ellos formaliza el criterio utilizado por Disney en la película *Fantasia* para relacionar características gráficas con información musical, y el segundo permite la composición musical polifónica con detalles de interpretación para que la reproducción de la música compuesta por el sistema se aleje de ser un producto sintético. Estos dos procesos utilizan técnicas y algoritmos bien diferenciados en sus diferentes fases, y la combinación de todos ellos obtiene como resultado una composición musical a modo de banda sonora para el proceso ilustrativo de un usuario.

Para la formalización del criterio de traducción imagen-sonido se ha hecho uso de la película *Fantasia* de Disney, por ser una de las películas que mejor explotan el concepto de música descriptiva. El algoritmo SIFT se ha utilizado para la extracción de información clave de las imágenes junto con la técnica BoVW, que permite simplificar la complejidad de los datos y obtener meta-información muy valiosa para el proceso de aprendizaje automático de la propuesta. Por otra parte, la cuantificación del color mediante *k-Means* permite simplificar la información cromática de una imagen sin perder datos relevantes. Por último, la extracción de información musical se ha realizado mediante un proceso automatizado en esta propuesta, simplificando enormemente el trabajo y obteniendo datos clave para la resolución del problema.

La necesidad de predecir un acorde (conjunto de tres notas), hace necesaria la aplicación de algoritmos de clasificación multi-etiqueta. Concretamente, los algoritmos RAKEL y ML-kNN se aplican en este problema, obteniendo como resultado una mayor adecuación del primero de ellos. Por otra parte, la utilización de RNN con bloques LSTM da lugar a una serie de composiciones musicales polifónicas dotadas de información de agógica y dinámica muy útiles para su reproducción.

Las dos encuestas de usuario que se han realizado para validar este experimento han arrojado resultados satisfactorios. En primer lugar, el test de Turing respalda el hecho de que la máquina desarrolla una actividad de manera similar a como lo haría una persona, simulando cierta habilidad inteligente y creativa. A la vista de los resultados, se puede concluir que muchas personas son incapaces de distinguir si la música está compuesta por una máquina o por un ser humano. Además, las valoraciones de las composiciones musicales son muy similares y no existe una gran diferencia entre las composiciones del profesional y las del sistema. Esto significa que, a pesar del componente subjetivo presente en la valoración de cualquier creación ar-

tística, los resultados del sistema tienen una buena calidad musical. Por otra parte, la encuesta realizada tras hacer uso del sistema pone de manifiesto que los usuarios perciben que los colores de sus ilustraciones tienen mayor influencia que las formas y la distribución de elementos en el proceso de composición musical. El sistema crea un flujo bidireccional y una realimentación por la cual la ilustración del usuario da pie a la composición musical, y la música inspira al usuario para seguir dibujando.

El sistema desarrollado compone música de estilo anime, puesto que el conjunto de datos con el que se entrena la RNN determina el estilo musical que se obtiene como resultado en las composiciones. Para mejorar la experiencia del usuario, se pueden entrenar varios modelos con diferentes conjuntos de datos, cada uno de ellos enmarcado en un estilo musical concreto. De esta manera todos los modelos estarán disponibles en el sistema y el usuario podrá decidir y seleccionar el estilo musical que quiere que el sistema componga mientras él dibuja.

En el próximo capítulo...

Tras haber analizado y discutido los resultados obtenidos en los casos de estudio anteriormente descritos, el capítulo siguiente cierra el trabajo doctoral presentando las conclusiones que se han podido inducir a lo largo de todo el proceso de investigación y resumiendo las principales contribuciones del mismo. Además, esta información desemboca en la discusión de la validez de la hipótesis y el análisis de consecución de los objetivos definidos al inicio del trabajo. Finalmente, dado que se trata del último capítulo, también se esbozan algunas líneas de trabajo futuro que permitirán seguir enriqueciendo la investigación.

Capítulo 5

Conclusiones

Resumen: *Este capítulo expone las conclusiones derivadas del trabajo de investigación y analiza la validez de la hipótesis inicialmente formulada. Tras la aplicación del sistema en ambos casos de estudio y el análisis de los resultados obtenidos, se discute la validez de la propuesta y se resumen las principales contribuciones de la misma. Finalmente, se plantean algunas líneas de trabajo futuras que se pueden estudiar para contribuir al enriquecimiento de esta investigación.*

De acuerdo con la hipótesis establecida en el Capítulo 1, el presente trabajo propone un sistema heterogéneo que combina y engloba diferentes técnicas, entre las que destacan los algoritmos de aprendizaje automático, para el análisis del contenido multimedia expedido en otros sistemas. El fruto de este análisis es, en todo caso, la creación de otro tipo de información útil y de valor para el usuario.

El objetivo principal de este trabajo fijaba el foco de estudio en la investigación de técnicas y herramientas relacionadas con el tratamiento y el análisis de contenido multimedia, a fin de diseñar un sistema capaz de dar

respuesta a la hipótesis de partida y de afrontar la problemática planteada. El desarrollo de la presente tesis doctoral comienza con el planteamiento de una arquitectura flexible, escalable y modular enfocada al diseño de procesos de aprendizaje automático a partir de información multimedia y orientada a facilitar la combinación de diversas técnicas para el análisis. Posteriormente, se han diseñado dos marcos de trabajo concretos e independientes, cuyo objetivo es el análisis de contenido multimedia. Para ello se ha establecido el propósito de cada uno de ellos y se llevado a cabo la identificación de tareas básicas y su consecutiva adaptación a la arquitectura propuesta. Por último, se han implementado ambos *frameworks* y se han puesto en explotación en un entorno real con el propósito de confirmar que el sistema da respuesta a la hipótesis establecida al inicio del trabajo.

La arquitectura planteada en el Capítulo 3 surge de la abstracción de un proceso de análisis y la identificación de tareas básicas y el flujo de trabajo que las conecta. Como resultado se diseña una metodología genérica para el análisis de datos, y se propone una arquitectura flexible, modular y escalable que permite la convivencia de varios procesos de análisis de contenido multimedia independientes en un único sistema. Por estos motivos el sistema se puede calificar como híbrido y heterogéneo. La arquitectura engloba la adaptación de un proceso de análisis como las ETLs, utilizadas principalmente en el ámbito del *Business Intelligence*, y su aplicación mediante *datapipelines* en el ámbito de la IA. La fortaleza del sistema propuesto reside, por una parte, en la sencillez de su diseño, y por otra, en los beneficios que presenta en estos tipos de análisis de datos.

Algunos de los objetivos específicos de la presente investigación se centraban en la obtención y el tratamiento de datos de distinta naturaleza. Los dos marcos de trabajo diseñados en la Sección 3.2 tienen su punto de partida en la obtención del contenido multimedia. En el primer caso, el contenido

multimedia es un vídeo; en el segundo, una ilustración realizada por medios digitales. Adicionalmente, ambos *frameworks* trabajan con contenido musical tanto en la etapa de aprendizaje como en la de creación. La extracción de meta-información se realiza, en el Capítulo 4, atendiendo a la naturaleza de los datos iniciales y a las características que son realmente útiles y necesarias para cada estudio. En el caso del vídeo, como primer paso se obtiene la secuencia de fotogramas que lo componen. A partir de este momento, se lleva a cabo un proceso de extracción de meta-información similar en cada uno de los fotogramas y en la ilustración. Concretamente, se extrae información relativa al color, a la forma y a la disposición de los elementos de cada una de las imágenes. Las técnicas específicas utilizadas para la extracción del color son los histogramas de color en el espacio RGB y la cuantificación del mismo mediante una agrupación con el algoritmo *k-Means*. Para la extracción de las formas y la disposición de los elementos se utiliza el algoritmo SIFT por su excelente efectividad. La extracción de meta-información musical se lleva a cabo mediante el método CENS.

Para poder llevar a cabo el análisis de los datos, las tareas asociadas a la aplicación de los algoritmos de aprendizaje automático cobran un especial interés en este trabajo. Para el diseño de los dos casos de estudio desarrollados en este trabajo se ha realizado un examen detallado de los algoritmos más adecuados en función de diversos factores como los datos de partida, la problemática a resolver y los objetivos establecidos. Como resultado, en el primer caso de estudio se seleccionan los algoritmos NB, RF y SVM; en el segundo caso de estudio, al tratarse de un problema de clasificación multi-etiqueta se seleccionan los algoritmos RAKEL y ML-kNN. En ambos casos, los algoritmos son adaptados con el objetivo de optimizar los resultados y aplicados a los datos previamente extraídos para la obtención de resultados. La selección final del mejor algoritmo en cada caso se realiza mediante el estudio y discusión de diversas métricas que reflejan la eficiencia, y la valo-

ración de la aceptación de los resultados obtenidos por parte de los usuarios.

Los dos procesos de análisis diseñados como marcos de trabajo en la Sección 3.2 y aplicados en como casos de estudio en el Capítulo 4 obtienen como resultado una composición musical. En ambos estudios se lleva a cabo la simulación o imitación de un proceso artístico mediante la aplicación de algoritmos de IA en una máquina —concretamente, en este caso, la composición musical automática—, por lo que se enmarcan en el área de la creatividad computacional. Las obras creativas producidas por una máquina, al igual que las elaboradas por artistas profesionales, tienen como objeto la emoción y suscitación de sentimientos en las personas. En relación con este aspecto, el trabajo pone de manifiesto la necesidad de analizar y discutir los resultados del sistema desde dos perspectivas bien diferenciadas; el análisis de la eficiencia de los algoritmos de aprendizaje automático aplicados en cada caso no es suficiente para validar la capacidad creativa de la máquina. La aceptación social de las creaciones artísticas se convierte en un elemento esencial a considerar a la hora de aprobar la calidad de los resultados obtenidos. Sin embargo, es importante no perder de vista que la valoración de cualquier obra artística tiene una naturaleza subjetiva ligada a factores culturales, sentimentales, educativos e incluso perceptivos.

En lo relativo al proceso de composición musical automática efectuada en este trabajo, cabe destacar el componente descriptivo que acerca las composiciones al concepto de música programática. Para poder llevar a efecto dicho proceso creativo se propone establecer una relación entre los componentes gráficos y los auditivos de un vídeo. La calidad descriptiva de la música compuesta por el sistema depende, en parte, de los patrones de conexión entre los descriptores de la imagen y el sonido. Concretamente, en este trabajo, parte del éxito de la propuesta se debe al vínculo establecido mediante el análisis de la película *Fantasia* de Disney. El hecho de que un grupo de animado-

res profesionales dedicaran tiempo y esfuerzo a diseñar a los personajes y a aplicar los colores en función de los sentimientos que evocaba la música que pretendían ilustrar, ha favorecido la extracción de patrones y la traducción inversa en este trabajo.

Para concluir, en virtud de las grandes cantidades de datos expedidos diariamente en sistemas como las redes sociales y dada su naturaleza heterogénea, el trabajo plantea una arquitectura capaz de englobar, de manera eficiente, diferentes procesos de análisis de contenido multimedia basados en IA. Adicionalmente, los dos marcos de trabajo diseñados y llevados a la práctica en esta investigación, ponen de manifiesto que los análisis realizados mediante el sistema propuesto pueden generar información de valor para los usuarios. Si bien el diseño de nuevos *frameworks* para el análisis de contenido multimedia requiere del conocimiento técnico para el desarrollo y aplicación de algoritmos de aprendizaje automático, la creación de ciertos resultados por parte del sistema podrían hacer necesario el estudio de otras áreas o de la colaboración interdisciplinar de varias personas. La informática debe ser entendida como una herramienta para lograr la satisfacción de necesidades y la solución de problemas.

5.1. Contribuciones de la investigación

El proceso de investigación llevado a cabo en este trabajo para dar respuesta a la hipótesis establecida y satisfacer los objetivos planteados en la introducción obtiene como resultado una serie de aportaciones al estado del arte de la materia tratada. Concretamente, estas aportaciones se enmarcan en el ámbito del análisis de datos mediante técnicas de IA y en la composición automática de música.

La revisión del estado del arte ha sido una pieza clave en el desarrollo de este trabajo. En lo tocante al diseño de la arquitectura del sistema, ha sido necesario el estudio detallado de las diferentes fases de un proceso de análisis de datos, así como la aplicabilidad y adecuación de los diferentes algoritmos de aprendizaje automático. Para el diseño de los marcos de trabajo propuestos y la implementación de los casos de estudio ha sido necesario un análisis exhaustivo de las técnicas de extracción de datos descriptores de imagen y sonido así como un acercamiento a la rama de creatividad computacional y su aplicación para la composición musical automática. El fruto de este trabajo integra todos estos conceptos y teorías, dando lugar a una propuesta multidisciplinar para el análisis de contenido multimedia.

Considerando la arquitectura propuesta para el diseño de un sistema de análisis de contenido multimedia, es imprescindible incidir en los beneficios técnicos que ofrece y que son esenciales para validar la hipótesis de este trabajo. El diseño modular facilita las tareas de desarrollo y mantenimiento y el diseño e integración de nuevos procesos de análisis en el sistema, así como la reutilización de tareas. La flexibilidad de la arquitectura permite que el sistema incluya estudios para el análisis de contenido multimedia completamente heterogéneos. Por otra parte, la escalabilidad del sistema posibilita la concentración de diversos procesos de análisis salvaguardando su independencia y fomentando su existencia simultánea. En este sentido, la condición híbrida del sistema es sustancial.

Los dos marcos de trabajo diseñados comparten el objetivo de componer música mediante algoritmos de IA. La mayor aportación de este trabajo, en materia de creatividad computacional, es el diseño de una metodología para la composición de música descriptiva. En este trabajo se utiliza un fragmento de la película *Fantasia* de Disney para extraer patrones de relación entre descriptores de imagen y sonido que serán posteriormente aplicados en un

proceso de traducción gráfico-musical.

Finalmente, la participación activa en cursos de formación, debates científicos, y la asistencia a conferencias relacionadas con la materia que ocupa este trabajo ha fomentado la realimentación y el intercambio de conocimiento para establecer una base sólida sobre la que construir la propuesta. Del mismo modo, la asistencia a congresos internacionales y las diferentes vías de difusión del trabajo aplicadas a lo largo de la investigación han favorecido la consideración de diferentes teorías y puntos de vista y han supuesto un refuerzo para el trabajo realizado, contribuyendo a la optimización de los resultados obtenidos al término de este estudio.

5.2. Líneas de trabajo para la prosecución de la investigación

Para cerrar esta tesis, en esta sección se proponen diversas líneas de ampliación de la investigación a fin de coadyuvar al perfeccionamiento y enriquecimiento del trabajo realizado.

La primera propuesta de ampliación y continuación de este trabajo está relacionada con la arquitectura del sistema. Si bien el diseño planteado en este trabajo satisface las necesidades y soluciona los problemas identificados al inicio de esta investigación, sería interesante estudiar si la aplicación del concepto de agentes inteligentes contribuye a la optimización de la arquitectura planteada. Dada la condición modular del sistema, la aplicación de un sistema multi-agente o incluso de una organización virtual de agentes podrían suponer una mejora en la propuesta.

Las fortalezas técnicas que supone la arquitectura propuesta favorecen

la coexistencia de procesos de análisis heterogéneos en el sistema. Por este motivo, otra de las líneas futuras de investigación que proponemos es el diseño de nuevos marcos de trabajo para el análisis de contenido multimedia. Esto engloba, por una parte, la opción de seleccionar como punto de partida conjuntos de datos en otros formatos, como puede ser la información textual contenida en los comentarios de una publicación de una red social con su correspondiente procesamiento de lenguaje natural y análisis de sentimiento. Por otra parte, el diseño de nuevos procesos de análisis involucra la necesidad de definir un objetivo relacionado con la transformación de los datos y de la creación de información de valor para el usuario. En este sentido, de la misma forma que los dos *frameworks* diseñados en este trabajo conjugan diversas competencias de la informática con el conocimiento de la teoría musical, hay que considerar que los estudios propuestos podrían implicar el estudio de nuevas disciplinas o áreas del conocimiento o la colaboración multidisciplinar de expertos en dos o más áreas del conocimiento.

La traducción de descriptores de imagen a información musical para dar lugar a una composición musical descriptiva se realiza, en este caso, mediante el análisis de un fragmento de la película *Fantasia* de Disney. Como consecuencia, el vínculo entre contenido gráfico y contenido auditivo está supeditado al criterio establecido por los animadores de Disney. ¿Qué sucedería si en lugar de utilizar un fragmento de esta película para construir dicha conexión se analizara otro tipo de vídeo? Por este motivo, se plantea, a fin de contribuir al enriquecimiento de esta propuesta, la opción de utilizar otro tipo de vídeos para extraer patrones de relación imagen-sonido, como podría ser un fragmento de una película musical de un género diferente a la animación.

Parte II

Apéndices

Apéndice A

Conceptos de la teoría musical

Without music, life would be a mistake.

Friedrich Wilhelm Nietzsche

Resumen: *Este apéndice recoge la descripción de una serie de conceptos musicales cuya comprensión es necesaria para el correcto y completo entendimiento de la presente tesis doctoral.*

El elemento básico de la música es el sonido. Se trata de un fenómeno vibratorio que se transmite en forma de ondas a través de un medio elástico. Estas vibraciones se transmiten haciendo oscilar la presión del aire hasta llegar a nuestro cerebro. Hay dos características de estas ondas que son especialmente relevantes desde el punto de vista musical: la frecuencia y la amplitud.

En primer lugar, la frecuencia de una onda simple expresa el número de repeticiones de un evento periódico que se producen por unidad de tiempo. En el caso del sonido, la frecuencia representa el número de vibraciones del

medio que transmite el sonido en un segundo. Esta característica da lugar al tono del sonido de manera que las frecuencias altas dan lugar a sonidos agudos y las frecuencias bajas, a sonidos graves. La inmensa mayoría de los sonidos que se producen en la naturaleza dan lugar a ondas complejas. Sin embargo, estas se pueden descomponer fácilmente en ondas simples con ayuda de la Transformada de Fourier [20]. La frecuencia más baja de las ondas simples obtenidas de esta transformación se denomina frecuencia fundamental, y se corresponde con el tono o nota musical que nuestro oído percibe al escuchar un sonido.

Por otra parte, la amplitud de la onda es una medida de la variación máxima de una propiedad que varía periódicamente en el tiempo. En el caso de las ondas sonoras, la amplitud representa la intensidad o volumen del sonido. Así, las ondas con mayor amplitud darán lugar a sonidos más intensos y las ondas con menor amplitud representan sonidos con un volumen bajo.

Tras un acercamiento a la explicación física del elemento básico de la música, se van a introducir algunos conceptos básicos de la teoría musical. Todos los conceptos que se presentan en este trabajo se enmarcan en el contexto de la música occidental. La Sección A.1 presenta información sobre las notas musicales. La Sección A.2 detalla algunos aspectos relevantes en la concatenación de notas musicales y la Sección A.3 describe el marco armónico en el que se desarrollan las obras musicales. En las Secciones A.4 y A.5 se recoge información sobre la interpretación musical y la notación o grafía de la misma respectivamente.

A.1. Distancia entre notas musicales

Las notas musicales se corresponden con las diferentes frecuencias del sonido, y son `Do`, `Re`, `Mi`, `Fa`, `Sol`, `La` y `Si` según el sistema de notación latino o `C`, `R`, `E`, `F`, `G`, `A` y `B` según el sistema de notación anglosajón.

La distancia, en frecuencia, entre dos notas musicales se denomina intervalo. En la música occidental, el intervalo más pequeño entre dos notas se denomina semitono, y dos semitonos equivalen a un tono. Para poder obtener estos intervalos es necesario *alterar* las notas musicales mediante la utilización de los signos ♯ y ♭. El signo ♯ permite alterar una nota musical para obtener el sonido correspondiente a un semitono más agudo, y el signo ♭ rebaja en un semitono la nota. Adicionalmente, existe otro signo denominado becuadro (♮) que permite deshacer los efectos de los dos anteriores.

Como consecuencia, se obtienen 12 sonidos diferentes que se corresponden con las 12 notas musicales. La Figura A.1 muestra la posición de cada una de las notas sobre las teclas del piano. Las teclas blancas representan las notas naturales, y las teclas negras representan las notas alteradas (con ♯ o ♭).

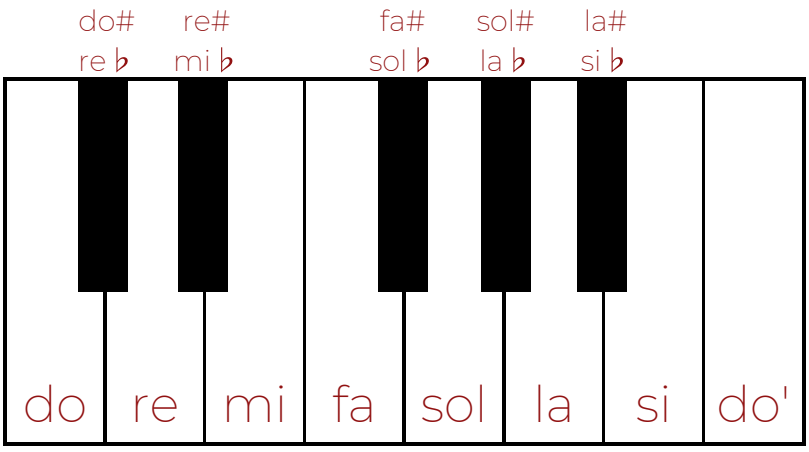

Figura A.1: Notas musicales sobre las teclas de un piano

La Figura A.1 muestra que distancia entre dos teclas consecutivas de un piano es de un semitono. La distancia entre dos teclas blancas siempre es de un tono (dos semitonos) salvo en los intervalos `Mi-Fa` y `Si-Do`, ya que, al no haber teclas negras entre ellas, la distancia es de un semitono. Esto define los intervalos que existen entre dos notas musicales consecutivas.

Hay otro aspecto que llama la atención en la imagen, y es que las teclas negras, correspondientes a las notas alteradas, reciben dos nombres diferentes. Este fenómeno se denomina enarmonía, y conlleva que un mismo sonido pueda tener más de un nombre. Ejemplo de ello es la nota `La♯` o `Si♭`, cuyo sonido se corresponde con el de la última tecla negra que aparece en la imagen.

Para medir la distancia entre dos notas, consecutivas o no, los intervalos pueden medirse en tonos y semitonos o haciendo alusión al grado, que es el número de notas que separan dos sonidos concretos. En este recuento, la nota de inicio y la nota de fin también deben tenerse en cuenta. Así, con el primer sistema para medir intervalos, la distancia entre las notas `Do` y `Mi♭` es de un tono y un semitono. Con la segunda notación, el intervalo sería una tercera.

Sin embargo, para poder diferenciar el intervalo `Do-Mi♭` de `Do-Mi` (que también es una tercera) es necesario especificar además de qué tipo de intervalo se trata: mayor, menor, justo, aumentado o disminuido. Este adjetivo se añade en función de los semitonos que separen a ambas notas. Así, la Tabla A.1 muestra los diferentes tipos de intervalo para cada grado en función de los semitonos que separan a las dos notas. Como se puede observar en la tabla, el intervalo `Do-Mi♭` es una tercera menor (3m), mientras que el intervalo `Do-Mi` es una tercera mayor (3M).

Grado	Tipo	Tonos
2	Menor (m)	0.5
	Mayor (M)	1
3	Menor (m)	1.5
	Mayor (M)	2
4	Disminuida (d)	2
	Justa (J)	2.5
	Aumentada (A)	3
5	Disminuida (d)	3
	Justa (J)	3.5
	Aumentada (A)	4
6	Menor (m)	4
	Mayor (M)	4.5
7	Menor (m)	5
	Mayor (M)	5.5

Tabla A.1: Relación entre el número de tonos y semitonos y el tipo de intervalo para cada grado

El intervalo entre dos notas con el mismo nombre es una octava. La diferencia entre ambas notas es la frecuencia, siendo la frecuencia de la más grave la mitad de la más aguda. Así, el intervalo `Do-Do'` correspondiente a la primera y la última tecla blanca del piano de la Figura A.1 sería una octava donde la nota situada más a la izquierda es más grave y le corresponde una frecuencia equivalente a la mitad de la frecuencia de la nota `Do'`.

Así, los intervalos pueden ser simples o compuestos dependiendo de si la distancia que separa a las dos notas es menor o mayor a una octava. Para la obtención del tipo de estos intervalos se analiza su intervalo simple correspondiente; es decir, el intervalo entre una nota Do y una nota Mi♭ de la octava siguiente sería una décima menor porque el número de notas que existen entre ellas son 10 y la tercera correspondiente al intervalo simple (Do-Mi♭) es un intervalo menor.

A.2. Sucesión de notas musicales

La ordenación de las notas con una distancia concreta y predefinida da lugar a una escala. Las escalas reciben el nombre de su nota inicial, y el orden de las notas, atendiendo a la frecuencia, puede ser ascendente o descendente. Existe un tipo de escala, denominada cromática, que está formada por un conjunto de 12 notas a distancia de semitono. Es decir, esta escala estaría formada por las notas Do, Do♯, Re, Re♯, Mi, Fa, Fa♯, Sol, Sol♯, La, La♯ y Si.

Existe también un tipo de escala denominada diatónica, formado por un conjunto de 7 notas a distancia de intervalos de segunda. Estas escalas se pueden clasificar como mayores o menores en función del orden en el que estén dispuestos los intervalos (segundas mayores o menores). Por lo general, las escalas mayores tienen una sonoridad que se relaciona con un sentimiento de alegría, mientras que las escalas menores despiertan sentimientos de tristeza.

Las escalas diatónicas mayores y menores son la base del sistema tonal, que se considera el eje de la música culta occidental de los siglos XVII-XX. En este sistema musical, cada posición en una escala diatónica se denomina grado, y se representa con números romanos. Cada grado de una escala tiene asignada una función musical que hace referencia a la tensión que provoca al

escucharla. Se pueden diferenciar cuatro tipos principales de funciones en la escala: la tónica (T), que es el sonido principal y más estable en la escala; la mediante o modal (M), que como su propio nombre indica es la que establece el modo de la escala (mayor o menor); la dominante, que presenta un alto grado de inestabilidad musical; la sensible, que es el sonido que presenta más inestabilidad y atracción hacia la tónica. Los grados II, IV y VI toman su nombre de estas cuatro funciones añadiendo además el prefijo sub (Sb-) o súper (S-) según se encuentren por debajo o por encima de dicha función. La Tabla A.2 muestra las funciones de los grados para la escala de Do Mayor.

Do Mayor

Nota	Do	Re	Mi	Fa	Sol	La	Si
Grado	I	II	III	IV	V	VI	VII
Función	T	S-T	M	Sb-D	D	S-D	S

Tabla A.2: Grados y funciones de la escala de Do Mayor

Otra de las características del sonido es su duración. En una composición musical, las notas tienen una duración concreta y finita cuya medida básica es un pulso. La sucesión de las duraciones de las notas que conforman una composición musical da lugar al ritmo. Adicionalmente, el silencio es una parte indispensable de la música, y aunque no tiene una frecuencia ni una amplitud específica asociada, sí tiene una duración concreta. La Figura A.2 muestra los pulsos de duración de las figuras musicales básicas para sonidos (arriba) y silencios (abajo).

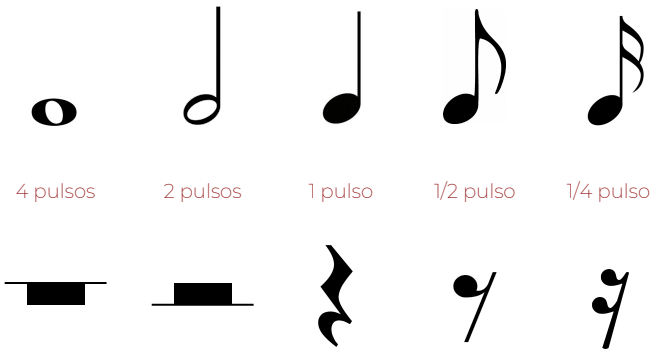

Figura A.2: Pulsos de duración de las figuras musicales y silencios básicos

La concatenación de notas sin un orden concreto y con un ritmo definido se percibe como una entidad musical y se denomina melodía. Las melodías tienen un sentido propio y se enmarcan en el contexto musical de una o varias escalas particulares. Su característica más relevante es que es una secuencia lineal de sonidos, es decir, que no suena más de una nota de manera simultánea.

A.3. Contexto armónico de la música

Cuando dos sonidos se reproducen de forma consecutiva se obtiene un intervalo melódico. En cambio, si las dos notas suenan simultáneamente, entonces el intervalo es armónico. Por otra parte, un acorde es un conjunto de tres o más sonidos que suenan en paralelo. El acorde básico recibe el nombre de tríada, y está compuesto por tres notas separadas entre sí por intervalos de tercera.

En relación con estos conceptos es necesario hablar de un componente musical esencial denominado armonía. La armonía es la técnica que estudia la creación de un acorde y la concatenación de los mismos con base en una serie

de principios musicales. A diferencia de la melodía, que es una concatenación lineal o secuencial de sonidos, la armonía entiende la música de manera vertical, haciendo que dos o más notas puedan sonar simultáneamente.

El contexto musical de una composición viene determinado por su tonalidad, que es la comprensión de la escala desde un punto de vista armónico. Atendiendo a este aspecto, sobre cada grado de la escala se puede crear un acorde tríada que tendrá una sonoridad característica. Por este motivo, a cada grado de la escala se le asigna una función tonal relacionada con la atracción o fuerza de su sonoridad, que puede ser tónica (T), dominante (D) o subdominante (S). La tónica está relacionada con la estabilidad de la tonalidad, la dominante crea una sensación de inestabilidad y tensión, y la subdominante crea sensación de movimiento, a pesar de tener una sonoridad estable. La Tabla A.3 muestra las funciones tonales de la tonalidad de Do Mayor.

Do Mayor							
Nota	Do	Re	Mi	Fa	Sol	La	Si
Grado	I	II	III	IV	V	VI	VII
Función	T	S	T	S	D	T	D

Tabla A.3: Funciones tonales de los grados en la tonalidad de Do Mayor

A pesar de que, generalmente, el contexto musical de las composiciones está enmarcado en una tonalidad concreta, existe la posibilidad de que se produzca un cambio de tonalidad. Este cambio se denomina modulación y suele tener un carácter temporal. Los cambios de contexto musical (tonalidad) de una pieza habitualmente son breves y provisionales; las composiciones acostumbran a terminar en la misma tonalidad en la que empiezan.

Al reproducirse dos o más notas de manera paralela, la persona que escucha puede experimentar sensaciones agradables o sensaciones de tensión. Así, surgen los conceptos de consonancia y disonancia, que hacen referencia a la estabilidad o la tensión respectivamente que experimenta una persona al percibir dos o más sonidos de manera simultánea. Muchas personas califican la consonancia como una nube de sonidos que resulta agradable y placentera, y la disonancia como una amalgama de sonidos que generan una sensación desagradable de tensión.

A.4. Interpretación musical

Uno de los componentes más importantes en la música es la interpretación, que consiste en la ejecución de una obra musical aplicando los conocimientos para la lectura de partituras, el dominio del instrumento y la expresión. En este último aspecto es donde recae toda la carga emotiva y sentimental de la música. La interpretación es el factor de la música que la aleja de la ciencia, puesto que la música no es simplemente la aplicación de una serie de técnicas y teorías, sino que tiene un componente emocional. Esto hace que dos personas tocando una única partitura realicen interpretaciones completamente distintas. Aunque gran parte de esta expresión depende del intérprete, hay ciertos aspectos como la dinámica y la agógica que facilitan la consecución de una buena interpretación.

La dinámica musical hace referencia a la graduación de intensidad de los sonidos. Para poder expresar diferentes niveles de intensidad se utilizan los denominados matices. Aunque esta cualidad del sonido varía en función del instrumento, del estilo e incluso del contexto musical, se establecen una serie de niveles para facilitar que el intérprete pueda reproducir la obra tal y como el compositor la pensó. De mayor a menor nivel de intensidad, los

símbolos y términos en italiano utilizados en las partituras son ***f*** (*forte*), ***mf*** (*mezzo-forte*), ***mp*** (*mezzo-piano*) y ***p*** (*piano*). Estos símbolos indican que el pasaje debe tocarse fuerte, moderadamente fuerte, moderadamente suave o suave.

Para indicar que la intensidad del sonido debe crecer a lo largo de un pasaje se puede utilizar el ángulo < o la palabra *crescendo* (***cresc.***). De igual manera, la reducción de intensidad puede indicarse mediante el signo > o las palabras *decrescendo* (***decresc.***) y *diminuendo* (***dim.***).

El *tempo* de una obra es la velocidad con la que debe ejecutarse una pieza musical. Aunque el *tempo* permite establecer una velocidad para la sucesión de pulsos de una obra, existen recursos complementarios que conceden ciertas licencias y permiten su variación o modificación en momentos puntuales. Estos recursos, que en conjunto se engloban dentro de la agógica musical, facilitan la transmisión de sentimientos por parte del intérprete. Así, los términos *ritardando* (***rit.***) y *accelerando* (***accel.***) permiten reducir o aumentar la velocidad gradualmente en un pasaje concreto. El uso de estas licencias por parte del propio compositor le dan al intérprete un margen de interpretación, pero a la vez le guían para que la obra sea más expresiva y se interprete tal y como él la pensó al escribirla.

A.5. Notación musical

Para poder escribir composiciones musicales se hace uso de los pentagramas, que son un conjunto de cinco líneas paralelas y equidistantes en las que se pueden colocar las figuras musicales. Para establecer el registro musical se utilizan las claves. Con estos signos se indica la referencia de una nota concreta, a partir de la cual se pueden representar el resto de sonidos en diferentes

alturas ocupando líneas y espacios. La clave de sol, por ejemplo, establece que la nota situada en la segunda línea del pentagrama comenzando desde la parte inferior se corresponde con la nota `Sol`. Las notas se disponen en el pentagrama de manera que las que están en las líneas y espacios inferiores representan sonidos más graves que las que están en las líneas y espacios superiores. Cuando el pentagrama no permite representar un sonido agudo o grave se pueden utilizar las denominadas líneas adicionales, que permiten ampliar el registro de sonidos que se representan con una clave.

Para poder representar correctamente el componente rítmico de la composición es necesario utilizar una serie de elementos adicionales en la partitura. En primer lugar, se establece el compás, que es la entidad métrica que determina la disposición de los pulsos fuertes y débiles de la música. Los compases se representan y diferencian mediante una línea vertical en la partitura. También se establece el *tempo*, que permite identificar la rapidez de los pulsos, y con ello la rapidez de interpretación de la obra musical. Para representar la duración de cada sonido se utilizan las figuras musicales de la Figura A.2.

La tonalidad de la composición se representa al inicio de la partitura, tras la clave. En ella se especifican las alteraciones (♯ o ♭) necesarias para la tonalidad principal. El conjunto de alteraciones de una tonalidad se denomina armadura. De manera adicional, si en algún momento es necesrio alterar una nota concreta, se pueden utilizar estas alteraciones junto con el becuadro (♮) de manera individual delante de la nota que se quiera alterar.

En la Figura A.3 se puede observar una breve partitura donde se representa la escala de La Mayor. En ella se pueden distinguir elementos como el pentagrama, la armadura (conjunto de los tres signos ♯), el compás y la línea divisoria, la línea adicional necesaria para representar la nota `La` más aguda

y el *tempo*, *Adagio*.

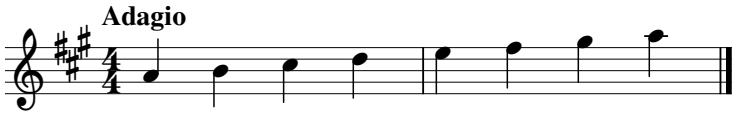

Figura A.3: Ejemplo de partitura musical simple

Para poder representar la simultaneidad de varios sonidos, las figuras se disponen verticalmente en el pentagrama. Adicionalmente, en ocasiones se realizan composiciones orquestales o escritas para ser interpretadas por instrumentos que pueden reproducir más de un sonido a la vez como el piano, el acordeón, el arpa o el violín. En estos casos se puede hacer necesaria la utilización de dos o más pentagramas, donde cada uno tendrá una clave en función del registro de la música que se desee escribir.

Apéndice B

Encuestas realizadas durante la investigación

Music is an experience, not a science.

Ennio Morricone

Resumen: *En este apéndice se expone toda la información de las encuestas realizadas a usuarios durante la ejecución de los casos de estudio del trabajo. Estos test de usuario se han utilizado como herramientas para la obtención de resultados en los diferentes estudios realizados para validar el sistema. Concretamente, en esta investigación se han realizado tres pruebas con usuarios, dos encuestas para medir la satisfacción del usuario con los resultados de la propuesta y un test de Turing.*

B.1. Introducción

La hipótesis de este trabajo plantea que los datos que se intercambian en las interacciones entre usuarios de sistemas como las redes sociales pueden ser analizados y utilizados para generar otro tipo de información de valor. Tras un profundo análisis de la literatura existente, se propone una arquitectura que da respuesta al problema planteado y se realizan dos casos de estudio que permiten utilizar contenido multimedia como vídeos o ilustraciones para componer música descriptiva de manera automática. La composición de música es un proceso creativo que obtiene como resultado una invención artística. Este proceso, localizado en la intersección de la IA y la música, se enmarca dentro de un área denominada creatividad computacional.

Para poder realizar una evaluación completa de la calidad del contenido artístico creado mediante el sistema propuesto es necesario analizar varios factores. En primer lugar, desde un punto de vista técnico y basado en la estadística, se deben considerar los resultados obtenidos por los algoritmos para comprobar la adecuación y la efectividad del método propuesto. Por otra parte, la aceptación y satisfacción de los usuarios con el sistema y especialmente con los resultados obtenidos son elementos a tener en cuenta, ya que tienen una relevancia especial en este tipo de trabajos. Finalmente, es interesante estudiar si el sistema manifiesta un comportamiento inteligente y desempeña su tarea de una forma similar a como lo haría un ser humano.

Como consecuencia, a lo largo del proceso de investigación se han llevado a cabo tres encuestas con usuarios: la primera tiene como objetivo la medición de la calidad de la relación establecida para la traducción entre elementos visuales (color y forma) y elementos musicales (sonido); la segunda incluye un test de Turing [36] para determinar si el sistema ha alcanzado el

objetivo de simular inteligencia durante el proceso de composición musical; la tercera evalúa la calidad descriptiva de la música compuesta en relación con las ilustraciones que el sistema utiliza como punto de partida. Todos los formularios se han diseñado en la plataforma Jotform[1] y se han difundido por medio del correo electrónico a listas de usuarios que encajaban con *target* de este trabajo.

Todos los formularios realizados en esta investigación tienen un carácter anónimo y una estructura común; existe una diferenciación de tres partes principales, accesibles en diferentes páginas:

- **Presentación:** Se pone en contexto al usuario, explicando brevemente el proceso de composición automática con el que se han generado las composiciones que se van a evaluar y el objetivo que se persigue con la realización del cuestionario. Adicionalmente, se presenta una explicación sencilla de la información que se va a encontrar en las páginas posteriores y se dan unas instrucciones básicas para que el usuario no tenga dudas a la hora de realizar la encuesta. Finalmente, se incluye un texto a modo de declaración de consentimiento para garantizar que el sujeto participa de forma voluntaria y que permite que los datos se utilicen de manera anónima en la investigación.

- **Evaluación:** Se presenta la información propia de cada formulario y el método de puntuación o valoración más adecuado en cada caso.

- **Agradecimiento:** Se realizan algunas preguntas adicionales para tener información del tipo de usuario que contesta a las preguntas y se da la opción de añadir, de manera voluntaria, un comentario para enriquecer los resultados de la encuesta o proponer mejoras de cara a futuros trabajos. Finalmente se agradece al usuario el tiempo que ha

[1] https://eu.jotform.com/

dedicado a realizar la encuesta y se deja un correo de contacto para cualquier duda o problema que haya podido surgir durante el proceso de evaluación.

En la Sección B.2 se describe el test de escucha realizado con el objetivo de evaluar la calidad de la relación entre imagen y sonido en el primer caso de estudio aplicado en este trabajo. La Sección B.3 recoge toda la información relativa al test de Turing realizado para evaluar la habilidad compositiva del *framework* para la composición armónica a partir de ilustraciones digitales. Por último, la Sección B.4 describe todos los detalles del test de usuario realizado para medir la calidad descriptiva de las armonías compuestas en el segundo caso de estudio.

B.2. Encuesta para evaluar la relación entre imagen y sonido

En esta sección se describe la encuesta llevada a cabo en el caso de estudio detallado en la Sección 4.1 de este trabajo. El objetivo de este experimento es realizar una composición musical descriptiva a partir de un vídeo. La película *Fantasia* de Disney [134] se utiliza como criterio para relacionar y traducir los elementos visuales y el sonido. La extracción de descriptores de imagen se realiza de dos maneras en este estudio: en primer lugar, mediante el algoritmo SIFT [80]; en segundo, mediante la aplicación del TL a las CNNs [103]. Una vez obtenidas las características visuales y auditivas del vídeo inicial se obtienen dos modelos (uno para cada método de extracción de los metadatos de la imagen). Finalizada la etapa de entrenamiento del experimento, la etapa de test consiste en la extracción de características visuales de un nuevo vídeo y la aplicación de los modelos previamente generados para obtener una

secuencia de sonidos.

El vídeo sobre el que se realiza la evaluación de los modelos en este estudio es la pieza de *The Firebird* de la película *Fantasia 2000* de Disney [133]. El vídeo se divide en cinco fragmentos, y para cada uno de ellos se elimina el audio original y se obtienen melodías con los dos modelos generados en la etapa de entrenamiento. En la encuesta que se describe en esta sección se muestran los cinco fragmentos de vídeo con las melodías compuestas por cada uno de los modelos (en total se componen 10 melodías, 2 para cada fragmento). El usuario debe evaluar de 1 (muy mala) a 10 (muy buena) la calidad descriptiva de la composición musical. De esta manera, el test tiene un doble objetivo: por una parte, evaluar la aceptación de la relación entre imagen y sonido; por otra, la calidad descriptiva de las composiciones musicales para cada uno de los dos métodos de extracción de características aplicados.

En la Sección B.2.1 se detalla el diseño de la encuesta, y la Sección B.2.2 desglosa las respuestas de los usuarios y ofrece un resumen estadístico de las mismas.

B.2.1. Diseño de la encuesta

El formulario está dividido en cuatro páginas. En la primera de ellas se realiza la presentación de la encuesta, en las dos siguientes se realiza la evaluación, y en la página final se agradece al usuario el tiempo dedicado a esta tarea. Para la evaluación, en este caso, se cuenta con 10 fragmentos de vídeo acompañados de sus respectivas melodías descriptivas compuestas por el sistema. Estos fragmentos aparecen divididos en dos páginas para reducir la carga cognitiva del usuario (5 en una página y otros 5 en la siguiente). Cada composición debe ser evaluada en una escala del 1 al 10 para indicar

la correspondencia entre los elementos visuales del vídeo (color, forma, disposición de los elementos, etc.) y el sonido. A continuación se presenta la información disponible en la encuesta:

Evaluación de la calidad descriptiva de la música

El propósito de este formulario es evaluar la calidad descriptiva de la música. No hay respuestas correctas o incorrectas ni se pretende evaluar la calidad de la música como tal, sino cómo los sonidos describen a la imagen que acompañan. Para el proceso de composición se ha considerado un fragmento de la película Fantasía de Disney. Se han utilizado dos técnicas de la IA para la extracción de un patrón de relación entre las características de la imagen y el sonido en este vídeo preliminar. Posteriormente, los patrones obtenidos por cada una de estas dos técnicas se han aplicado sobre las imágenes de un nuevo fragmento de vídeo de Fantasía 2000 de Disney donde el audio original no se ha tenido en cuenta. De esta manera, se han obtenido dos composiciones musicales que describen el vídeo. El objetivo del formulario es valorar la calidad descriptiva de cada una de estas dos composiciones.

En este experimento se presentan algunos fragmentos de los resultados obtenidos por dos sistemas inteligentes que componen música descriptiva a partir de un vídeo de forma automática. Por favor, reproduzca los fragmentos de vídeo y evalúe de 1 a 10 la calidad descriptiva de la música; es decir, cómo los sonidos describen a la imagen que acompañan y cómo cambian con los cambios de color, forma, disposición de elementos, etc.

Este experimento no llevará más de 10 minutos. Por favor, lea con atención las siguientes instrucciones antes de comenzar:

1. *Reproduzca los vídeos, escuche las composiciones y evalúe la calidad descriptiva de las mismas.*

2. *Por favor, realice el test en un ambiente tranquilo y use cascos para escuchar mejor los sonidos.*

3. *Puede reproducir los vídeos tantas veces como sea necesario.*

¡Gracias por su tiempo!

Puede decidir no participar en este estudio y, si comienza la participación, puede decidir detenerse y retirarse en cualquier momento. Puede salir de la encuesta cerrando la ventana del navegador donde está visualizando la encuesta. Su decisión será respetada y no dará lugar a la pérdida de los beneficios a los que de otra manera tiene derecho. No se recopilará información de identificación en esta encuesta. Al hacer clic en SIGUIENTE, usted otorga permiso para su participación. Al completar este cuestionario, usted acepta que los datos se utilizarán de forma anónima para los fines de este experimento.

Por favor, reproduzca los fragmentos del vídeo y evalúe la calidad descriptiva de la música en cada caso. La calidad descriptiva hace referencia a cómo los sonidos describen las características de la imagen a la que acompañan y cómo cambian con el color, la forma, la disposición de los elementos, etc.

Preguntas finales

- *¿Cómo escuchó el audio?*
  - *Cascos*
  - *Auriculares*
  - *Altavoces*
- *¿Cuál es su formación musical?*
  - *Músico profesional o musicólogo*
  - *Amateur (sé tocar un instrumento, soy miembro de un coro...)*
  - *No tengo formación musical*
- *Inserte un comentario (OPCIONAL)*

¡Muchas gracias por su tiempo!

Si quiere saber más acerca de este proyecto: luciamg@usal.es

La encuesta se ha realizado con la aplicación Jotforms y está disponible en la web[2]. La Figura B.1 muestra un ejemplo de evaluación de uno de los fragmentos; en ella se puede apreciar una captura del vídeo para el que la máquina ha realizado una composición y un ítem con una escala del 1 al 10 de opción única.

[2] https://form.jotformeu.com/80381752910354

Figura B.1: Imagen de ejemplo de la encuesta para evaluar la relación entre imagen y sonido

B.2.2. Respuestas de los usuarios a la encuesta

La relación entre imagen y sonido para los fragmentos compuestos por las dos variantes de la propuesta se ha evaluado por un total de 47 usuarios. En la Tabla B.1 se muestran, por filas, las evaluaciones de cada usuario, y por columnas, la información de los fragmentos de cada variante (extracción de metadatos de la imagen mediante los métodos M1 o M2 definidos en el Capítulo 4). Las dos últimas columnas de esta tabla indican el modo de reproducción de los vídeos (1-cascos, 2-auriculares, 3-altavoces) y la formación musical de cada uno de los usuarios que realizan la encuesta (1-Músico profesional, 2-Amateur, 3-Sin formación). La Tabla B.2 resume toda esta información con una serie de métricas estadísticas como son los valores má-

ximo, mínimo, medio, central (mediana) y el rango intercuartílico de cada uno de los fragmentos. Además, en esta tabla se muestran los valores medio, central y el rango intercuartílico para las evaluaciones de cada uno de los modelos utilizados para la composición.

	M1					M2					AUDIO	FORMACIÓN
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**		
U1	7	9	8	5	7	6	8	6	6	7	3	1
U2	6	7	5	4	4	5	6	5	6	7	3	3
U3	8	6	7	4	8	7	9	9	6	9	1	3
U4	6	6	6	8	8	7	7	7	8	8	1	3
U5	10	8	8	8	10	10	9	8	9	9	1	3
U6	2	4	6	9	5	5	7	6	9	7	2	2
U7	3	2	4	1	4	6	4	3	2	3	2	2
U8	7	6	4	3	5	8	6	4	5	6	3	3
U9	4	3	3	1	4	5	4	3	3	3	2	3
U10	6	2	2	2	2	2	2	2	2	2	3	2
U11	7	7	8	5	6	7	6	5	5	8	1	3
U12	3	2	1	1	4	5	5	3	3	6	2	2
U13	5	4	5	5	10	4	6	7	5	10	1	3
U14	5	2	5	1	4	4	3	2	4	4	1	2
U15	5	7	3	2	7	5	7	3	2	6	2	2
U16	7	6	8	3	5	8	8	9	6	8	3	3
U17	7	8	5	5	6	5	5	7	5	7	2	3
U18	8	4	6	2	4	8	7	7	4	4	3	2
U19	3	4	6	3	6	5	5	5	5	5	3	2
U20	3	4	4	3	5	3	4	4	4	5	2	1
U21	5	2	2	4	6	5	3	3	5	6	3	1
U22	8	3	5	4	7	8	8	7	9	7	1	3
U23	9	7	7	5	3	9	8	5	3	4	3	3

	M1					M2					AUDIO	FORMACIÓN
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**		
U24	2	3	2	2	5	4	5	5	4	6	1	3
U25	7	3	5	3	4	4	5	8	7	6	2	2
U26	7	6	7	4	5	7	7	6	7	7	2	3
U27	7	5	5	7	5	5	5	7	6	8	2	3
U28	8	7	7	8	7	8	7	7	8	7	3	2
U29	6	2	1	1	1	2	1	3	1	1	1	3
U30	8	7	7	5	8	8	7	7	7	8	2	3
U31	7	9	8	6	8	7	7	7	8	8	1	2
U32	8	7	4	5	4	7	8	4	5	4	2	1
U33	7	6	5	5	8	9	6	7	4	6	1	3
U34	7	6	4	2	5	8	6	4	4	4	2	2
U35	10	6	8	7	8	9	9	7	6	7	2	3
U36	9	8	8	9	9	9	9	7	8	7	1	3
U37	8	5	7	5	6	7	7	8	5	6	2	3
U38	9	6	8	8	10	10	7	8	8	8	2	3
U39	8	8	8	9	7	7	8	9	6	4	3	3
U40	4	4	6	4	6	7	5	8	4	4	2	2
U41	6	5	6	5	7	7	6	6	7	6	1	3
U42	4	7	8	4	8	7	8	9	7	8	3	3
U43	6	4	5	5	7	7	4	6	6	5	3	3
U44	4	4	5	5	7	4	3	3	3	7	1	3
U45	8	4	6	2	8	8	9	6	6	7	1	2
U46	9	7	8	7	9	8	8	7	9	9	1	3
U47	5	6	7	5	6	5	4	8	8	7	2	2

Tabla B.1: Respuestas del test de escucha aplicado al primer caso de estudio

	M1					M2				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Máximo	10	9	8	9	10	10	9	9	9	10
Mínimo	2	2	1	1	1	2	1	2	1	1
Media	6.3	5.3	5.6	4.5	6.1	6.4	6.1	5.9	5.3	6.2
Mediana	7	6	6	5	6	7	6	6	6	7
RI	3	3	2.5	2	3	3	3	3	3	2.5
Media global			5.6					6		
Mediana global			6					6		
RI global			3					3		

Tabla B.2: Resumen estadístico de las respuestas del test de escucha aplicado al primer caso de estudio

De manera adicional, algunos de los usuarios que realizaron la encuesta decidieron, de manera voluntaria, dejar un comentario al respecto. Estos comentarios se reproducen a continuación de manera textual, indicando la formación musical de la persona que los hizo y el modo de reproducción de audio que utilizó:

- *Se repite un montón la nota Re. Todas las canciones empiezan por esa nota, indiferentemente de los colores o dinámica del vídeo. Las dos primeras muy bien conseguidas, pero las que reflejan mas dinamismo, la música apenas acompaña a ese dinamismo. Pero muy chulo aún así* **(Músico profesional o musicólogo / Auriculares)**
- *Me ha costado elegir una puntuación, en el caso de tenerlo que repetir no tengo claro si escogería la misma opción* **(Sin formación musical / Auriculares)**
- *Durante la realización de la prueba puede interpretar la melodía acorde*

a la representación de la imagen, sin embargo en ocasiones algunas notas comparten demasiado tiempo distintos momentos de la imagen lo que me lleva a pensar si existe la posibilidad de aumentar más el número de notas musicales **(Sin formación musical / Cascos)**

- *Muy interesante el estudio. El último vídeo de la primera página me parece que es espectacular cómo se adapta la música a los cambios de color. Gran trabajo y encantada de colaborar!* **(Sin formación musical / Auriculares)**
- *Me ha parecido muy interesante. Enhorabuena por el trabajo y mucha suerte* **(Sin formación musical / Auriculares)**
- *Hay fragamentos del audio, especialmente en las 5 primeras muestras, donde mantiene demasiado la nota y en la escena sí hay cambios* **(Sin formación musical / Cascos)**

B.3. Test de Turing para evaluar la habilidad compositiva del sistema

En segundo caso de estudio que se lleva a cabo en este trabajo y que se detalla en la Sección 4.2, se lleva a cabo una creación musical descriptiva basada en la información visual de un dibujo. Para poder medir la capacidad inteligente de un sistema, Alan Turing propuso un tipo de test en el que los usuarios debían diferenciar si una tarea es realizada por una persona o por una máquina. Por ello, el test que se describe en esta sección tiene como objetivo evaluar la habilidad de la máquina para exhibir un comportamiento inteligente similar al de un humano. Concretamente, el comportamiento que se evalúa es la composición automática, teniendo en cuenta que en esta composición se incluyen aspectos musicales como la armonía, la duración de

las notas, la dinámica y la agógica. Para ello, se presentan 20 fragmentos musicales de entre 10 y 20 segundos de duración. La mitad de ellos estarán creados por un compositor humano (son fragmentos utilizados para el entrenamiento de la red neuronal) y la otra mitad los habrá compuesto la máquina. El orden será aleatorio para dificultar un poco más la tarea de distinción del usuario. El usuario debe valorar si cada uno de ellos está creado por un compositor profesional o por el sistema, y evaluar del 1 al 10 la calidad musical (armonía e interpretación) de las piezas.

En la Sección B.3.2 se muestra la información del formulario diseñado y la Sección B.3.2 recoge la información de las respuestas de los usuarios.

B.3.1. Diseño del test

El formulario se divide en cuatro páginas. La primera de ellas incluye una presentación del trabajo y le ofrece al usuario unas pautas para realizar la encuesta correctamente. Las dos siguientes páginas contienen 20 fragmentos musicales, 10 de los cuales están compuestos por la máquina, y el resto por compositores profesionales. Estos fragmentos están divididos en dos páginas (10 en cada una) para reducir la carga cognitiva del usuario, y están ordenados de manera aleatoria para dificultar la tarea de distinción del compositor real. El usuario debe discernir el tipo de compositor de cada pieza musical y evaluar, del 1 al 10 la calidad de las mismas. Finalmente, en la última página, se agradece el usuario el tiempo dedicado en realizar la encuesta.

Test para evaluar la calidad musical

En este test se presentan una serie de fragmentos de varias composiciones musicales. Algunas de ellas han sido creadas por un compositor profesional y otras se han obtenido a partir

de un sistema inteligente. El objetivo de esta prueba es reproducir un test de Turing. Este tipo de test fue propuesto por Alan Turing para evaluar la habilidad de una máquina para exhibir un comportamiento inteligente similar al de un humano. Por favor, escuche con atención cada uno de los fragmentos musicales e intente descubrir por quién ha sido creada (compositor profesional o máquina). Además, valore en una escala del 1 al 10 la calidad musical de las piezas.

Este experimento no llevará más de 15 minutos. Por favor, lea con atención las siguientes instrucciones antes de comenzar:
1. *Reproduzca los vídeos, escuche las composiciones y conteste a las preguntas.*
2. *Por favor, realice el test en un ambiente tranquilo y use cascos para escuchar mejor los sonidos.*
3. *Puede reproducir los audios tantas veces como sea necesario.*

¡Gracias por su tiempo!

Puede decidir no participar en este estudio y, si comienza la participación, puede decidir detenerse y retirarse en cualquier momento. Puede salir de la encuesta cerrando la ventana del navegador donde está visualizando la encuesta. Su decisión será respetada y no dará lugar a la pérdida de los beneficios a los que de otra manera tiene derecho. No se recopilará información de identificación en esta encuesta. Al hacer clic en SIGUIENTE, usted otorga permiso para su participación. Al completar este cuestionario, usted acepta que los datos se utilizarán de forma anónima para los fines de este experimento.

Evaluación

20 fragmentos musicales acompañados con un ítem de respuesta única para evaluar si el compositor es la máquina o un ser humano, y un segundo ítem de evaluación con una escala del 1 al 10 para valorar la calidad musical.

Inserte un comentario (OPCIONAL)

¡Muchas gracias por su tiempo!

Si quiere saber más acerca de este proyecto: luciamg@usal.es

El formulario, realizado con la aplicación Jotforms, está accesible en la web[3]. La Figura B.2 ilustra la evaluación correspondiente al séptimo fragmento de la encuesta. En ella se puede reproducir una pieza musical compuesta por la máquina disponible en la plataforma Soundcloud[4]. El usuario debe adivinar si el compositor original de dicha pieza es la máquina o un ser humano, y evaluar en una escala del 1 (mala) al 10 (buena) la calidad musical de la misma.

[3] https://form.jotform.com/200825869922363
[4] https://soundcloud.com/

Figura B.2: Imagen de ejemplo del test de Turing para evaluar la habilidad compositiva del sistema

B.3.2. Respuestas de los usuarios al test

Para realizar este test, es necesario considerar creaciones musicales de un compositor profesional y otras de la máquina. Por este motivo, el test no se ha podido realizar durante la utilización del sistema, ni tras la utilización de la misma (las composiciones en ambos casos son exclusivamente de la máquina). En este caso, 46 personas han participado en este estudio.

Aunque los fragmentos en la encuesta no aparecían en orden en función de quién fuera su compositor, a continuación se presentan los resultados de la encuesta haciendo diferenciación con este criterio. Así, la Tabla B.3 presenta las respuestas de los usuarios para los fragmentos creados por el compositor profesional, y la Tabla B.4, las respuestas para los fragmentos compuestos

por la máquina. En ambas tablas se representan, por filas, las respuestas de cada uno de los usuarios. Las columnas hacen referencia a cada uno de los diez fragmentos del compositor en cuestión, diferenciando en cada caso entre (♬) el criterio del usuario para determinar si la música ha sido creada por un compositor (C) o por una máquina (M) y (★) la puntuación del usuario para evaluar la calidad musical.

	1		**2**		**3**		**4**		**5**		**6**		**7**		**8**		**9**		**10**	
	♬	★	♬	★	♬	★	♬	★	♬	★	♬	★	♬	★	♬	★	♬	★	♬	★
U1	M	7	C	9	C	4	M	7	C	6	M	7	M	6	C	7	C	5	C	6
U2	M	6	C	6	M	5	M	6	M	5	C	5	C	6	M	6	C	5	C	6
U3	M	2	M	3	C	5	C	7	M	4	M	5	C	6	C	6	M	4	M	5
U4	C	9	C	6	C	6	M	7	M	6	C	8	C	7	M	6	C	9	C	8
U5	C	7	M	7	M	7	C	7	C	7	C	7	C	7	C	7	C	7	C	7
U6	C	8	M	7	M	4	C	8	M	7	M	6	C	7	C	7	M	8	C	9
U7	M	5	M	6	M	5	C	5	M	5	M	5	C	6	C	5	M	5	C	6
U8	C	8	M	5	C	6	C	7	M	4	C	5	C	6	C	8	C	6	C	7
U9	C	8	M	6	M	5	C	6	C	7	M	5	C	7	M	5	C	7	M	7
U10	C	9	M	7	M	4	C	9	M	7	M	6	C	8	M	7	C	9	C	8
U11	M	8	C	6	C	9	M	8	M	5	M	5	M	6	C	7	M	6	M	5
U12	C	5	M	3	M	6	C	5	M	4	M	5	C	5	M	5	C	6	C	4
U13	C	8	M	8	M	8	M	7	M	7	M	6	C	7	M	6	C	7	M	6
U14	C	6	M	4	M	5	M	7	C	7	M	6	M	7	M	4	C	6	M	6
U15	C	8	C	7	M	5	M	8	C	6	C	6	C	6	M	5	C	7	C	8
U16	C	8	C	6	M	4	M	5	C	5	M	5	M	3	C	3	M	9	C	9
U17	C	8	M	6	C	7	M	8	M	7	M	7	C	8	C	8	M	9	C	9
U18	C	7	C	7	M	3	C	8	M	4	C	4	C	4	C	6	C	3	C	8
U19	C	7	M	6	M	6	C	7	M	7	M	5	C	8	M	6	C	8	C	6
U20	C	8	M	6	M	7	C	9	M	6	C	7	M	7	M	7	C	6	M	7
U21	M	8	C	7	M	7	C	7	M	7	M	7	C	6	C	6	C	7	C	8

	1		2		3		4		5		6		7		8		9		10	
	♫	★	♫	★	♫	★	♫	★	♫	★	♫	★	♫	★	♫	★	♫	★	♫	★
U22	M	6	C	9	C	8	C	8	M	7	C	6	M	5	M	6	M	5	C	6
U23	M	5	M	6	C	7	C	6	C	8	M	6	C	7	C	6	M	6	M	6
U24	M	7	M	5	M	5	M	7	C	7	C	7	C	7	C	8	C	7	M	8
U25	C	8	M	8	M	7	C	10	M	9	M	6	M	8	M	7	M	8	M	8
U26	C	7	C	6	M	5	M	7	C	6	M	6	C	7	C	7	C	7	C	6
U27	C	8	M	7	M	7	C	9	C	8	C	9	C	8	M	7	C	8	C	9
U28	M	8	M	4	C	7	C	8	M	7	M	7	C	7	M	5	C	7	M	6
U29	C	6	M	5	M	6	M	6	M	5	M	4	M	5	C	6	M	4	M	4
U30	M	6	M	6	C	8	M	6	M	6	M	6	C	6	M	7	M	6	M	6
U31	M	7	M	8	C	9	M	7	M	8	M	6	C	7	M	7	M	7	M	7
U32	M	7	M	7	C	6	C	8	M	7	C	7	C	6	M	4	C	6	M	7
U33	C	9	C	8	M	6	M	7	M	7	C	8	C	8	C	8	C	7	C	8
U34	C	7	C	8	M	5	C	6	M	5	C	7	C	8	C	6	C	8	C	6
U35	C	7	M	2	M	2	M	4	C	5	C	3	M	4	M	5	C	6	C	7
U36	M	6	M	5	M	5	M	6	C	6	M	7	M	8	C	7	C	8	M	8
U37	M	8	M	8	M	5	C	8	M	6	M	6	C	8	M	7	C	8	M	6
U38	C	7	M	6	M	3	C	6	M	5	M	6	C	6	M	5	C	7	M	5
U39	C	8	M	6	M	7	M	7	M	6	M	5	M	6	M	6	M	6	M	7
U40	M	4	M	1	M	4	C	6	M	4	M	1	C	7	M	5	C	1	M	3
U41	C	9	M	7	M	7	C	9	M	6	M	6	C	7	M	7	C	8	M	7
U42	C	7	C	7	C	8	C	7	M	8	C	7	C	7	M	6	C	6	C	6
U43	M	3	M	5	M	8	M	4	C	8	C	7	M	8	C	8	C	5	M	7
U44	M	7	M	5	C	5	C	6	C	9	M	4	M	5	M	5	C	6	M	7
U45	C	8	C	9	M	7	C	9	M	8	M	8	M	7	C	8	C	9	C	9
U46	C	9	M	6	M	7	M	7	C	8	M	5	M	6	M	5	M	5	M	7

Tabla B.3: Respuestas del test de turing aplicado al segundo caso de estudio para las composiciones del compositor

	1		2		3		4		5		6		7		8		9		10	
	♫	★	♫	★	♫	★	♫	★	♫	★	♫	★	♫	★	♫	★	♫	★	♫	★
U1	C	8	C	6	M	7	C	7	C	6	M	7	C	8	C	7	C	5	M	6
U2	M	5	M	5	C	7	C	7	M	7	M	3	M	3	M	1	M	4	M	3
U3	M	4	C	5	C	7	C	6	C	5	M	5	C	6	C	6	M	5	M	5
U4	M	9	M	6	C	6	C	8	M	7	M	7	C	8	C	8	M	7	C	7
U5	C	7	C	7	M	7	M	7	M	7	M	7	M	7	M	7	M	7	M	7
U6	M	8	C	6	C	8	M	6	M	6	M	5	M	7	M	4	C	6	M	5
U7	M	5	M	6	C	5	C	5	C	6	C	5	C	7	M	6	C	5	C	6
U8	M	4	M	7	M	4	M	3	C	6	M	2	M	4	M	1	C	5	M	2
U9	M	6	M	6	C	7	M	6	M	7	M	3	C	5	M	5	C	7	M	5
U10	M	7	C	8	M	7	C	8	M	7	M	5	C	7	M	5	M	7	M	6
U11	C	8	M	7	M	6	M	5	C	8	C	7	C	7	C	7	C	7	C	8
U12	M	4	C	7	M	8	C	4	C	7	M	2	M	5	C	6	M	6	M	5
U13	M	6	C	8	C	9	C	7	M	7	M	6	M	6	M	7	C	6	M	7
U14	C	6	M	6	C	8	M	6	C	7	C	6	C	5	C	4	M	5	C	5
U15	C	8	C	6	M	7	M	6	C	8	M	3	M	4	M	4	M	5	M	5
U16	M	3	M	7	C	8	M	4	M	3	M	6	C	4	C	5	M	4	M	4
U17	M	9	M	9	C	8	M	8	C	9	C	10	C	9	C	9	M	8	M	8
U18	M	3	C	6	C	8	C	5	M	2	M	2	C	5	C	6	M	3	C	8
U19	M	6	C	8	C	8	M	6	C	7	C	6	M	6	C	6	M	5	M	5
U20	M	7	M	7	M	6	C	9	C	9	C	7	M	7	C	7	M	6	C	6
U21	M	7	C	9	C	8	C	7	M	6	M	4	M	7	M	6	M	8	M	4
U22	M	3	C	10	C	8	M	5	C	5	M	1	M	5	M	4	C	9	M	4
U23	C	7	M	6	C	7	C	6	C	9	C	7	C	7	C	7	M	6	C	6
U24	C	7	C	8	C	8	M	6	M	5	M	3	M	3	M	5	C	7	M	3
U25	C	8	C	9	C	10	M	9	M	7	M	7	M	6	M	7	C	8	C	7
U26	C	8	C	7	C	7	M	6	M	6	M	6	C	6	M	5	M	5	M	5
U27	M	8	C	9	M	8	M	8	M	7	M	7	M	8	M	6	C	8	M	7

	1		2		3		4		5		6		7		8		9		10	
	♬	★	♬	★	♬	★	♬	★	♬	★	♬	★	♬	★	♬	★	♬	★	♬	★
U28	C	7	C	6	M	6	C	8	C	7	C	5	C	6	C	5	M	6	C	7
U29	C	6	C	7	M	6	M	5	C	6	M	5	M	4	C	6	M	5	M	5
U30	C	7	C	8	M	5	M	5	C	8	C	7	M	5	C	7	M	5	C	7
U31	C	8	C	7	M	8	C	7	C	8	C	8	C	8	C	8	C	7	C	8
U32	C	6	M	5	M	5	C	8	C	8	M	4	C	5	M	7	M	4	M	4
U33	M	7	C	8	C	8	C	8	C	8	M	6	M	5	M	6	C	8	M	6
U34	M	5	C	8	M	7	C	7	C	7	M	4	M	4	M	4	M	6	M	6
U35	M	5	M	3	C	5	M	5	M	4	M	3	M	3	M	3	M	3	M	2
U36	M	6	C	8	C	6	C	6	C	7	M	7	C	9	M	7	C	8	C	9
U37	C	6	M	9	M	8	C	7	M	6	M	7	C	8	M	8	C	7	M	8
U38	M	3	M	4	M	5	C	7	M	3	M	1	M	3	M	5	C	5	M	4
U39	C	8	M	7	M	7	C	8	C	7	M	5	C	7	M	6	C	7	C	7
U40	C	4	M	4	M	5	C	6	M	6	C	7	M	3	C	8	M	3	C	4
U41	M	6	C	9	C	9	M	6	C	8	M	6	C	7	C	7	C	8	C	7
U42	M	5	C	8	C	7	C	8	M	6	M	6	M	4	M	5	C	5	M	4
U43	M	6	C	7	C	7	C	8	M	6	M	3	M	3	M	3	C	7	M	4
U44	C	8	C	7	M	5	M	9	C	5	C	6	M	5	C	5	M	6	C	5
U45	M	6	M	8	M	8	C	9	M	9	M	5	M	3	C	7	M	8	M	8
U46	M	7	C	8	M	6	C	8	C	7	M	6	C	6	M	5	M	7	C	8

Tabla B.4: Respuestas del test de turing aplicado al segundo caso de estudio para las composiciones de la máquina

A continuación, las Tablas B.5 y B.6 resumen los resultados de las tablas anteriores haciendo uso de estadísticas. Por una parte, se hace una diferencia entre la selección del compositor de cada fragmento por parte de los usuarios (♬) mostrando en cada caso el porcentaje de votos para cada uno de los dos posibles compositores. Por otra parte, se recogen los valores máximo, mínimo,

medio, central (mediana) y el rango intercuartílico de las votaciones de los usuarios para evaluar la calidad musical (★).

		1	2	3	4	5	6	7	8	9	10
♫	**C (%)**	60.87	30.43	30.43	56.52	32.61	34.78	67.39	43.48	69.57	50
	M (%)	39.13	69.57	69.57	43.48	67.39	65.22	32.61	56.52	30.43	50
★	**Máximo**	9	9	9	10	9	9	8	8	9	9
	Mínimo	2	1	2	4	4	1	3	3	1	3
	Media	7	6.13	5.91	7	6.5	6.91	6.54	6.20	6.52	6.76
	Mediana	7	6	6	7	6.5	6	7	6	7	7
	RI	2	1.75	2	2	2	2	1	2	2	2

Tabla B.5: Resumen estadístico de las respuestas del test de Turing aplicado al segundo caso de estudio para las composiciones del compositor

		1	2	3	4	5	6	7	8	9	10
♫	**C (%)**	39.13	60.87	52.17	56.52	54.35	26.09	45.65	43.48	43.48	36.96
	M (%)	60.87	39.13	47.83	43.48	45.65	73.91	54.35	56.52	56.52	63.04
★	**Máximo**	9	10	10	9	9	10	9	9	9	9
	Mínimo	3	3	4	3	2	1	3	1	3	2
	Media	6.24	7	6.96	6.63	5.57	65.23	5.65	5.72	6.07	5.70
	Mediana	6	7	7	7	7	6	6	6	6	6
	RI	2.75	2	2	2	1	3	3	2	2	2.75

Tabla B.6: Resumen estadístico de las respuestas del test de Turing aplicado al segundo caso de estudio para las composiciones de la máquina

Adicionalmente, el diseño del formulario permitía a los usuarios añadir comentarios si lo consideraban oportuno. La información recogida en este apartado se presenta, sin alteraciones de los textos originales, a continuación:

- *Me ha resultado difícil, me intriga mucho conocer las respuestas correctas*

- *Todas se parecen demasiado y las pistas son tan breves que no da tiempo a observar un desarrollo mayor y con ello dar una mejor respuesta. Reconozco que en algunas he respondido sin mucho criterio. Ánimo y gracias*
- *Me encantaría saber si he acertado alguna...*
- *No soy entendido en música, he asignado los valores según me sonaban bien o mal, si me sonaba algo raro en el cambio de ritmo he dicho máquina y si me sonaba todo "bien" compositor, probablemente por esta razón haya fallado*
- *Hay melodías preciosas y el sonido me ha sorprendido. Resulta muy complicado saber si las ha creado un compositor o una máquina*
- *Me gusta mucho la temática ;)*
- *Los dos primeros fragmentos se parecían bastante. Me dio la sensación de que uno era el original compuesto por un humano y el segundo una versión ligeramente alterada (para mal) por el ordenador*
- *Buen trabajo*

B.4. Encuesta para medir la calidad descriptiva de la música compuesta

En el caso de estudio presentado en la Sección 4.2, el sistema compone música de manera dinámica y automática durante un proceso de ilustración con medios digitales. Con el objetivo de complementar los resultados obtenidos por el test de Turing que se describen en la sección anterior, en este experimento se realiza una encuesta para medir la aceptación social de

la conexión entre los dibujos realizados por el usuario y las composiciones musicales creadas por el sistema.

La encuesta consta de cuatro preguntas que deben ser contestadas del 1 al 10, y se realiza únicamente a usuarios que ya han hecho uso del sistema. Esta encuesta permite evaluar qué características del dibujo considera el usuario que influyen más en el resultado musical y en qué grado lo hacen. Además, al tratarse de una composición que se realiza de manera dinámica mientras el usuario pinta, se estudia la opción de que la música también tenga una influencia en los trazos y colores que el usuario decide utilizar.

En la Sección B.4.1 se reproduce la información incluida en el formulario y la Sección B.4.2 expone las respuestas de los usuarios que han realizado la encuesta.

B.4.1. Diseño de la encuesta

El formulario está dividido en tres páginas. En la primera de ellas se realiza la presentación y contextualización del experimento, en la siguiente se realiza la evaluación y en la página final se agradece el tiempo dedicado a contestar las preguntas formuladas. La evaluación de esta encuesta consiste en contestar cuatro preguntas en una escala del 1 (nada) al 10 (mucho) para indicar en qué grado se cumple la cuestión que se plantea. La información del formulario se presenta a continuación:

> ***Test para evaluar la relación entre la ilustración y la música***
>
> *Este test se realiza tras utilizar el sistema de composición automática que genera música descriptiva a partir de una ilustración creada por medios digitales. El objetivo es evaluar la*

calidad de la relación entre los elementos de la ilustración (color, forma...) y la música generada.

Por favor, lea cada una de las cuatro preguntas que se le plantean y valore, en una escala del 1 al 10, la adecuación que considera que se produce en cada caso.

Este experimento no llevará más de 5 minutos. Por favor, lea con atención las preguntas y tómese su tiempo para contestar.

¡Gracias por su tiempo!

Puede decidir no participar en este estudio y, si comienza la participación, puede decidir detenerse y retirarse en cualquier momento. Puede salir de la encuesta cerrando la ventana del navegador donde está visualizando la encuesta. Su decisión será respetada y no dará lugar a la pérdida de los beneficios a los que de otra manera tiene derecho. No se recopilará información de identificación en esta encuesta. Al hacer clic en SIGUIENTE, usted otorga permiso para su participación. Al completar este cuestionario, usted acepta que los datos se utilizarán de forma anónima para los fines de este experimento.

PREGUNTA 1 *- ¿Cree que existe una adecuación entre su dibujo y los sonidos generados?*

PREGUNTA 2 *- ¿Cree que el color influye en los sonidos de la composición?*

PREGUNTA 3 *- ¿Cree que la forma influye en los sonidos*

de la composición?

PREGUNTA 4 *- ¿Le ha influido la música a la hora de dibujar?*

Inserte un comentario (OPCIONAL)

¡Muchas gracias por su tiempo!

Si quiere saber más acerca de este proyecto: luciamg@usal.es

La encuesta se ha realizado con la aplicación Jotforms y está accesible en la web[5]. La Figura B.3 muestra el diseño del formulario (concretamente la tercera pregunta) y un ítem para su respuesta con una puntuación del 1 al 10.

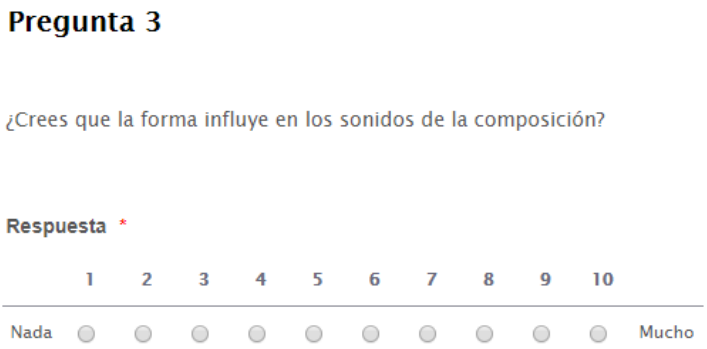

Figura B.3: Imagen de ejemplo de la encuesta para medir la calidad descriptiva de la música

[5] https://form.jotform.com/200824962130347

B.4.2. Respuestas de los usuarios a la encuesta

El sistema descrito en el segundo caso de estudio de este trabajo y para el cual se diseña esta encuesta estuvo desplegado temporalmente en un servidor para que los usuarios pudieran acceder a él y probarlo en un ambiente tranquilo. Tras haber hecho uso de la herramienta, se seleccionó a diez usuarios para contestar a las cuatro preguntas que se plantean en este estudio.

Las respuestas de cada uno de ellos se pueden observar en la Tabla B.7, donde cada fila representa las respuestas de un usuario, y cada columna hace referencia a una pregunta. Por otra parte, la Tabla B.8 refleja un resumen estadístico de estas preguntas, considerando los valores máximo, mínimo, medio, central (mediana) y el rango intercuartílico de cada una de las preguntas.

	Q1	**Q2**	**Q3**	**Q4**
U1	7	9	5	7
U2	4	9	3	7
U3	6	3	8	2
U4	3	4	2	1
U5	6	7	7	3
U6	4	8	6	8
U7	6	7	7	5
U8	8	7	3	2
U9	7	8	6	8
U10	8	6	7	7

Tabla B.7: Respuestas del test de escucha aplicado al segundo caso de estudio

	Q1	**Q2**	**Q3**	**Q4**
Máximo	8	9	8	8

	Q1	**Q2**	**Q3**	**Q4**
Mínimo	3	3	2	1
Media	5.9	6.8	5.4	5
Mediana	6	7	6	6
RI	2.5	1.75	3.5	4.75

Tabla B.8: Resumen estadístico de las respuestas del test de escucha aplicado al segundo caso de estudio

Apéndice C

Fragmentos musicales compuestos por el sistema

In scientific thinking are always present elements of poetry. Sciences and music requires a thought homogeneous.

Albert Einstein

Resumen: *En este apéndice se reproducen algunos resultados obtenidos con la metodología propuesta. Esta información comprende tanto el punto de partida (el vídeo o imagen que el sistema utiliza para realizar la composición musical) como el resultado musical descriptivo que se obtiene. Concretamente, la Sección C.1 ilustra algunos ejemplos de composición melódica obtenidos mediante el primer enfoque de la propuesta. La Sección C.2 recoge algunos fragmentos musicales armónicos generados durante el proceso de creación de ilustraciones digitales.*

C.1. Ejemplos de fragmentos musicales melódicos

A continuación se presentan algunos ejemplos de composiciones descriptivas realizadas con el caso de estudio desarrollado en la Sección 4.1. En este trabajo la música describe los elementos gráficos de un vídeo, y para ello cada fotograma se traduce en una nota. La concatenación de todas las notas obtenidas por el sistema dan lugar a la melodía final.

Los vídeos seleccionados para la obtención de melodías descriptivas pertenecen a la película *Fantasia 2000* [133]. Concretamente, los tres documentos audiovisuales utilizados como punto de partida para la composición se corresponden con algunos fragmentos de las animaciones creadas por Disney para la obra musical *Pájaro de fuego*. Para poder representar cada uno de ellos gráficamente, se muestra una recopilación de algunos de sus fotogramas más icónicos.

En la Figura C.1 se muestra el fragmento de vídeo representado por una selección de *frames* y la melodía obtenida en el primer ejemplo de composición realizada con el primer caso de estudio del trabajo.

Figura C.1: Primer ejemplo de ilustración y composición musical obtenida con el primer caso de estudio

La Figura C.2 representa una selección de fotogramas del vídeo utilizado para el segundo ejemplo de composición con el primer caso de estudio junto con la partitura de la melodía compuesta por el sistema.

Figura C.2: Segundo ejemplo de ilustración y composición musical obtenida con el primer caso de estudio

En la Figura C.3 se puede observar la selección de *frames* para el vídeo utilizado en el tercer ejemplo de composición de melodías con el sistema propuesto y la partitura de la melodía descriptiva creada.

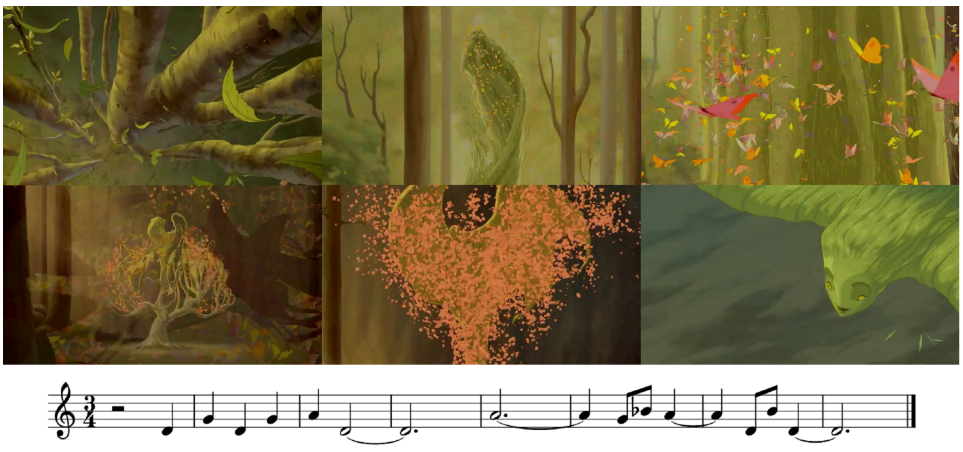

Figura C.3: Tercer ejemplo de ilustración y composición musical obtenida con el primer caso de estudio

C.2. Ejemplos de fragmentos musicales armónicos

En esta sección se muestran tres ejemplos de composición automática obtenidos en el segundo caso de estudio del trabajo (Sección 4.2). En este caso, la fuente de inspiración para el sistema es una ilustración realizada por el usuario mediante una tableta gráfica, y las composiciones musicales son armónicas.

La Figura C.4 constituye el primer ejemplo de composición musical armónica realizado por el sistema propuesto. En ella se puede observar la ilustración realizada por un usuario en la parte superior, y la partitura de la música que el sistema compuso para describirla en la parte inferior.

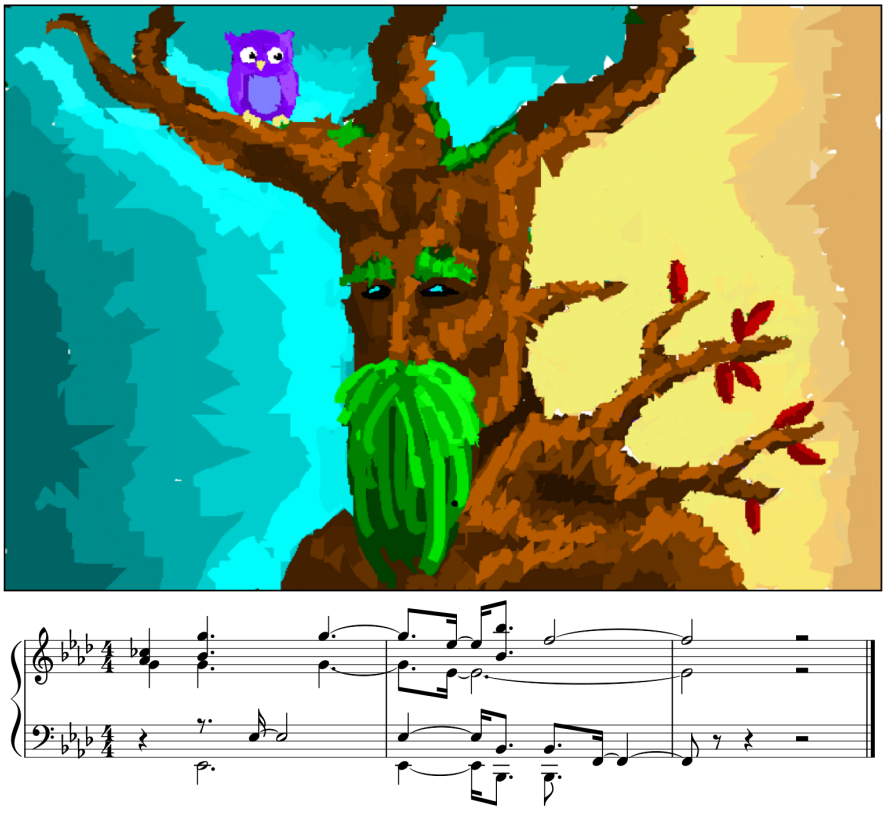

Figura C.4: Primer ejemplo de ilustración y composición musical obtenida con el segundo caso de estudio

El segundo ejemplo de composición armónica del sistema puede encontrarse en la Figura C.5, distinguiéndose la ilustración realizada con la tableta gráfica en la parte superior y la partitura de la composición en la parte inferior.

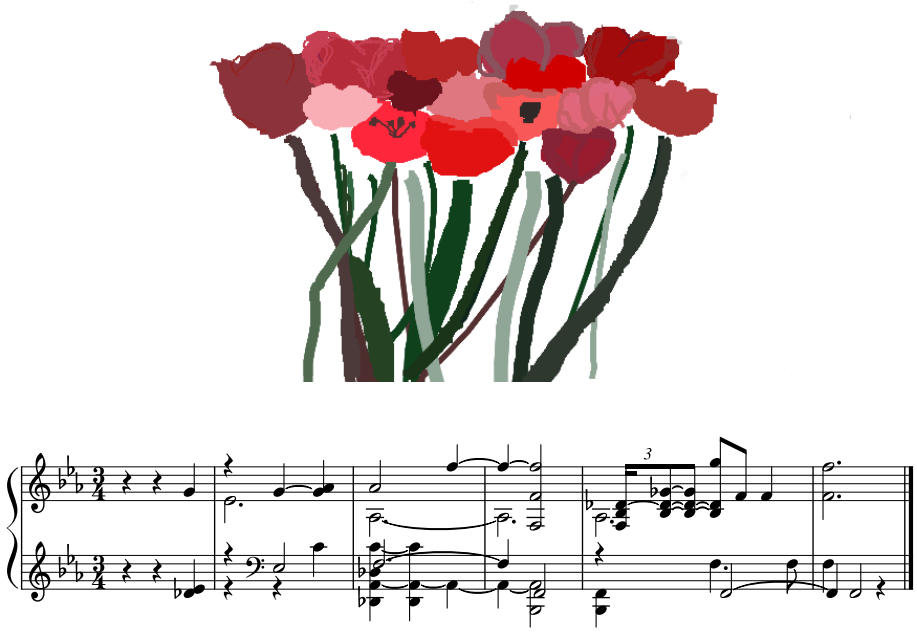

Figura C.5: Segundo ejemplo de ilustración y composición musical obtenida con el segundo caso de estudio

Con la misma estructura que en las dos figuras anteriores, la ilustración realizada por el usuario y la partitura de la composición del sistema que constituyen el tercer ejemplo de creación armónica se pueden visualizar en la Figura C.6.

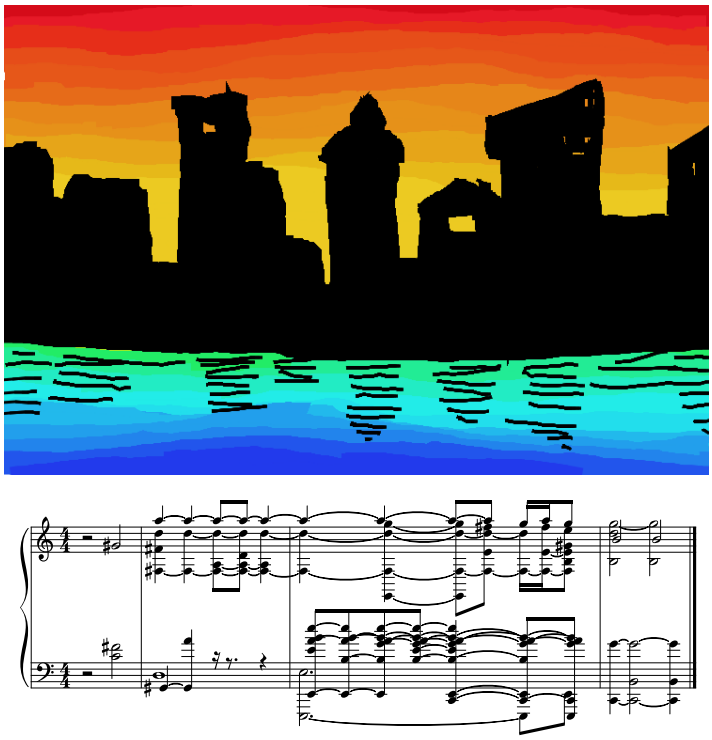

Figura C.6: Tercer ejemplo de ilustración y composición musical obtenida con el segundo caso de estudio

Bibliografía

[1] N. Agarwala, Y. Inoue, and A. Sly, *Music composition using recurrent neural networks*, CS 224n: Natural Language Processing with Deep Learning, Spring, (2017).

[2] M. M. al Rifaie, J. M. Bishop, and S. Caines, *Creativity and autonomy in swarm intelligence systems*, Cognitive computation, 4 (2012), pp. 320–331.

[3] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, *KAZE features*, in European Conference on Computer Vision, Springer, 2012, pp. 214–227.

[4] P. F. Alcantarilla and T. Solutions, *Fast explicit diffusion for accelerated features in nonlinear scale spaces*, IEEE Trans. Patt. Anal. Mach. Intell, 34 (2011), pp. 1281–1298.

[5] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, *Random forests and decision trees*, International Journal of Computer Science Issues (IJCSI), 9 (2012), p. 272.

[6] E. Alpaydin, *Introduction to machine learning*, MIT press, 2020.

[7] N. Anand, *ETL and its impact on business intelligence*, International Journal of Scientific and Research Publications, 4 (2014), p. 1.

[8] R. APARICI MARINO, *La revolución de los medios audiovisuales*, Educación y nuevas tecnologías, (1996).

[9] D. Archibugi and S. Iammarino, *The globalization of technological innovation: definition and evidence*, Review of International Political Economy, 9 (2002), pp. 98–122.

[10] M. Attaran, *Information technology and business-process redesign*, Business process management journal, (2003).

[11] L. Atzori, A. Iera, and G. Morabito, *The internet of things: A survey*, Computer networks, 54 (2010), pp. 2787–2805.

[12] D. Bahdanau, K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate*, arXiv preprint arXiv:1409.0473, (2014).

[13] A. R. Bartolomé Pina, *Multimedia interactivo y sus posibilidades en educación superior*, Pixel-Bit. Revista de Medios y Educación, 1, 5-14., (1994).

[14] G. B. Bauzá, *El guión multimedia*, Univ. Autònoma de Barcelona, 1997.

[15] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, *Speeded-up robust features (SURF)*, Computer vision and image understanding, 110 (2008), pp. 346–359.

[16] G. Bickerman, S. Bosley, P. Swire, and R. M. Keller, *Learning to create jazz melodies using deep belief nets.*, in ICCC, 2010, pp. 228–237.

[17] S. Binitha, S. S. Sathya, et al., *A survey of bio inspired optimization algorithms*, International journal of soft computing and engineering, 2 (2012), pp. 137–151.

[18] D. Bogdanov, N. Wack, E. Gómez Gutiérrez, S. Gulati, H. Boyer, O. Mayor, G. Roma Trepat, J. Salamon, J. R. Zapata González, X. Serra, et al., *Essentia: An audio analysis library for music information retrieval*, in 14th Conference of the International Society for Music Information Retrieval (ISMIR), International Society for Music Information Retrieval (ISMIR), 2013, pp. 493–498.

[19] L. Bottou and C.-J. Lin, *Support vector machine solvers*, Large scale kernel machines, 3 (2007), pp. 301–320.

[20] R. N. Bracewell and R. N. Bracewell, *The Fourier transform and its applications*, vol. 31999, McGraw-Hill New York, 1986.

[21] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*, O'Reilly Media, Inc., 2008.

[22] L. Breiman, *Random forests*, Machine learning, 45 (2001), pp. 5–32.

[23] D. Buhalis and P. O'Connor, *Information communication technology revolutionizing tourism*, Tourism recreation research, 30 (2005), pp. 7–16.

[24] L. Busin, N. Vandenbroucke, and L. Macaire, *Color spaces and image segmentation*, Advances in imaging and electron physics, 151 (2008), p. 1.

[25] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, *Brief: Binary robust independent elementary features*, in European conference on computer vision, Springer, 2010, pp. 778–792.

[26] M. E. Celebi, *Improving the performance of k-means for color quantization*, Image and Vision Computing, 29 (2011), pp. 260–271.

[27] M. Chemillier, *Improvising jazz chord sequences by means of formal grammars*, in Journées dínformatique musicale, 2001, pp. 121–126.

[28] ——, *Toward a formal study of jazz chord sequences generated by steedman's grammar*, Soft Computing, 8 (2004), pp. 617–622.

[29] N. Chomsky, *Studies on semantics in generative grammar*, vol. 107, Walter de Gruyter, 2013.

[30] H. Chung, M. Iorga, J. Voas, and S. Lee, *Alexa, can i trust you?*, Computer, 50 (2017), pp. 100–104.

[31] D. Ciregan, U. Meier, and J. Schmidhuber, *Multi-column deep neural networks for image classification*, in 2012 IEEE conference on computer vision and pattern recognition, IEEE, 2012, pp. 3642–3649.

[32] M. Clague, *Playing in'toon: Walt disney's"fantasia"(1940) and the imagineering of classical music*, American Music, 22 (2004), pp. 91–109.

[33] M. Cook, S. Colton, and J. Gow, *The angelina videogame design system— Part I*, IEEE Transactions on Computational Intelligence and AI in Games, 9 (2016), pp. 192–203.

[34] ——, *The angelina videogame design system— Part II*, IEEE Transactions on Computational Intelligence and AI in Games, 9 (2016), pp. 254–266.

[35] D. Cope, *Computer models of musical creativity*, MIT Press Cambridge, 2005.

[36] B. J. Copeland, *The turing test*, Minds and Machines, 10 (2000), pp. 519–539.

[37] A. Das and B. Gambäck, *Poetic machine: Computational creativity for automatic poetry generation in bengali.*, in ICCC, 2014, pp. 230–238.

[38] W. B. De Haas, M. Rohrmeier, R. C. Veltkamp, and F. Wiering, *Modeling harmonic similarity using a generative grammar of tonal harmony*, in Proceedings of the Tenth International Conference on Music Information Retrieval (ISMIR), 2009.

[39] J. Delon, A. Desolneux, J. L. Lisani, and A. B. Petro, *Automatic color palette*, in IEEE international conference on image processing 2005, vol. 2, IEEE, 2005, pp. II–706.

[40] J. Demšar, *Statistical comparisons of classifiers over multiple data sets*, Journal of Machine learning research, 7 (2006), pp. 1–30.

[41] L. Dhakar, *Color thief*. https://lokeshdhakar.com/color-thief/, 2019.

[42] T. Dietterich, *Overfitting and undercomputing in machine learning*, ACM computing surveys (CSUR), 27 (1995), pp. 326–327.

[43] A. Dorr, B. E. Rabin, and S. Irlen, *Parenting in a multimedia society*, Handbook of parenting, 5 (2002), pp. 349–373.

[44] J. S. Downie, *Music information retrieval*, Annual review of information science and technology, 37 (2003), pp. 295–340.

[45] D. Eck and J. Schmidhuber, *Finding temporal structure in music: Blues improvisation with LSTM recurrent networks*, in Proceedings of the 12th IEEE workshop on neural networks for signal processing, IEEE, 2002, pp. 747–756.

[46] D. P. Ellis, *Classifying music audio with timbral and chroma features*, in Proceedings of the 8th International Conference on Music Information Retrieval, 2007, pp. 23–27.

[47] R. England, *Standard MIDI file production as the focus of a broad computer science course*, Journal of Computing Sciences in Colleges, 32 (2017), pp. 4–10.

[48] D. Fitzgerald, *Harmonic/percussive separation using median filtering*, in Proc. of DAFX, vol. 10, 2010.

[49] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, *Support vector machine classification and validation of cancer tissue samples using microarray expression data*, Bioinformatics, 16 (2000), pp. 906–914.

[50] M. Geis and M. Middendorf, *Creating melodies and baroque harmonies with ant colony optimization*, International Journal of Intelligent Computing and Cybernetics, (2008).

[51] K. Goel, R. Vohra, and J. Sahoo, *Polyphonic music generation by modeling temporal dependencies using a rnn-dbn*, in International Conference on Artificial Neural Networks, Springer, 2014, pp. 217–224.

[52] K. Gopalakrishnan, S. K. Khaitan, A. Choudhary, and A. Agrawal, *Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection*, Construction and Building Materials, 157 (2017), pp. 322–330.

[53] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, *LSTM: A search space odyssey*, IEEE transactions on neural networks and learning systems, 28 (2016), pp. 2222–2232.

[54] F. Guillet and H. J. Hamilton, *Quality measures in data mining*, vol. 43, Springer, 2007.

[55] J. Han and K.-K. Ma, *Fuzzy color histogram and its use in color image retrieval*, IEEE Transactions on image Processing, 11 (2002), pp. 944–952.

[56] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*, Elsevier, 2011.

[57] M. C. Herreros, *La radio en la convergencia multimedia*, Editorial Gedisa, 2018.

[58] J. J. Higgins, *An introduction to modern nonparametric statistics*, Brooks/Cole Pacific Grove, CA, 2004.

[59] J. H. Holland et al., *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*, MIT press, 1992.

[60] S. Holtzman, *Using generative grammars for music composition*, Computer Music Journal, 5 (1981), pp. 51–64.

[61] M. Hossin and M. Sulaiman, *A review on evaluation metrics for data classification evaluations*, International Journal of Data Mining & Knowledge Management Process, 5 (2015), p. 1.

[62] M. B. Hoy, *Alexa, siri, cortana, and more: an introduction to voice assistants*, Medical reference services quarterly, 37 (2018), pp. 81–88.

[63] IAB Spain and Elogia, *Estudio anual de redes sociales 2019*. https://iabspain.es/wp-content/uploads/2019/06/estudio-anual-redes-sociales-iab-spain-2019_vreducida.pdf, 2019.

[64] N. A. Ibraheem, M. M. Hasan, R. Z. Khan, and P. K. Mishra, *Understanding color models: a review*, ARPN Journal of science and technology, 2 (2012), pp. 265–275.

[65] M. Ivasic-Kos and M. Pobar, *Multi-label classification of movie posters into genres with rakel ensemble method*, in International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer, 2017, pp. 370–383.

[66] N. Japkowicz and M. Shah, *Evaluating learning algorithms: a classification perspective*, Cambridge University Press, 2011.

[67] D. D. Johnson, *Generating polyphonic music using tied parallel networks*, in International conference on evolutionary and biologically inspired music and art, Springer, 2017, pp. 128–143.

[68] M. A. Kaliakatsos-Papakostas, A. Floros, and M. N. Vrahatis, *Interactive music composition driven by feature evolution*, SpringerPlus, 5 (2016), p. 826.

[69] M. Kantarcioglu, J. Vaidya, and C. Clifton, *Privacy preserving naive bayes classifier for horizontally partitioned data*, in IEEE ICDM workshop on privacy preserving data mining, 2003, pp. 3–9.

[70] N. Khan, B. McCane, and S. Mills, *Better than SIFT?*, Machine Vision and Applications, 26 (2015), pp. 819–836.

[71] K. Koitzsch, *Data pipelines and how to construct them*, in Pro Hadoop Data Analytics, Springer, 2017, pp. 77–90.

[72] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, *Supervised machine learning: A review of classification techniques*, 2007.

[73] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, nature, 521 (2015), pp. 436–444.

[74] F. Lerdahl and R. S. Jackendoff, *A generative theory of tonal music*, MIT press, 1996.

[75] S. Leutenegger, M. Chli, and R. Y. Siegwart, *BRISK: Binary robust invariant scalable keypoints*, in 2011 International conference on computer vision, Ieee, 2011, pp. 2548–2555.

[76] C.-Y. Lin and H.-S. Chen, *Personalized channel recommendation on live streaming platforms*, Multimedia Tools and Applications, 78 (2019), pp. 1999–2015.

[77] C.-H. Liu and C.-K. Ting, *Evolutionary composition using music theory and charts*, in 2013 IEEE Symposium on Computational Intelligence for Creativity and Affective Computing (CICAC), IEEE, 2013, pp. 63–70.

[78] B. Logan et al., *Mel frequency cepstral coefficients for music modeling.*, in Ismir, vol. 270, 2000, pp. 1–11.

[79] O. Lopez-Rincon, O. Starostenko, and G. Ayala-San Martín, *Algoritmic music composition based on artificial intelligence: A survey*, in 2018 International Conference on Electronics, Communications and Computers (CONIELECOMP), IEEE, 2018, pp. 187–193.

[80] D. G. Lowe, *Object recognition from local scale-invariant features*, in Proceedings of the seventh IEEE international conference on computer vision, vol. 2, Ieee, 1999, pp. 1150–1157.

[81] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, *Color and texture descriptors*, IEEE Transactions on circuits and systems for video technology, 11 (2001), pp. 703–715.

[82] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: multimedia content description interface*, John Wiley & Sons, 2002.

[83] M. Marchini, R. Ramirez, P. Papiotis, and E. Maestre, *The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets*, Journal of New Music Research, 43 (2014), pp. 303–317.

[84] A. G. Martín et al., *Educación multimedia y nuevas tecnologías*, vol. 9, Ediciones de la Torre, 2010.

[85] L. Martín-Gómez and J. Pérez-Marcos, *Automatic composition of descriptive music: A case study of the relationship between image and sound*. https://figshare.com/articles/Automatic_composition_of_descriptive_music_A_case_study_of_the_relationship_between_image_and_sound/6682998, 2018.

[86] ——, *Image and sound data from film Fantasia produced by Walt Disney*. https://figshare.com/articles/FantasiaDisney_ImageSound/5999207, 2018.

[87] ——, *Musical Chords and Image Descriptors from Film Fantasia (Disney)*. https://figshare.com/articles/Image_and_Sound_Data_from_Film_Fantasia_Disney_/12110712, 2020.

[88] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, *LibROSA: Audio and music signal analysis in python*, in Proceedings of the 14th python in science conference, vol. 8, 2015, pp. 18–25.

[89] J. McNiff, *Action research: Principles and practice*, Routledge, 2013.

[90] V. Mendiola, A. Doss, W. Adams, J. Ramos, M. Bruns, J. Cherian, P. Kohli, D. Goldberg, and T. Hammond, *Automatic exercise recognition with machine learning*, in International Workshop on Health Intelligence, Springer, 2019, pp. 33–44.

[91] R. Morduchowicz, *La generación multimedia: significados, consumos y prácticas culturales de los jóvenes*, Paidós Buenos Aires, 2008.

[92] M. Muller, F. Kurth, and M. Clausen, *Chroma-based statistical audio features for audio matching*, in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005., IEEE, 2005, pp. 275–278.

[93] L. J. Najjar, *Multimedia information and learning*, in Journal of educational multimedia and hypermedia, Citeseer, 1996.

[94] M. Navarro, M. Caetano, G. Bernardes, L. N. de Castro, and J. M. Corchado, *Automatic generation of chord progressions with an artificial immune system*, in International Conference on Evolutionary and Biologically Inspired Music and Art, Springer, 2015, pp. 175–186.

[95] F. Niecks, *Programme Music in the last four Centuries; a Contribution to the History of musical Expression*, Ardent Media, 2013.

[96] N. J. Nilsson, *Principles of artificial intelligence*, Morgan Kaufmann, 2014.

[97] D. E. O'Leary, *Artificial intelligence and big data*, IEEE intelligent systems, 28 (2013), pp. 96–99.

[98] V. C. Oleynick, T. M. Thrash, M. C. LeFew, E. G. Moldovan, and P. D. Kieffaber, *The scientific study of inspiration in*

the creative process: challenges and opportunities, Frontiers in human neuroscience, 8 (2014), p. 436.

[99] M. T. Orchard, C. A. Bouman, et al., *Color quantization of images*, IEEE transactions on signal processing, 39 (1991), pp. 2677–2690.

[100] J. d. Pablos Pons and J. Jiménez Segura, *Nuevas tecnologías, comunicación audiovisual y educación*, Cedecs, 1998.

[101] F. Pachet, *The continuator: Musical interaction with style*, Journal of New Music Research, 32 (2003), pp. 333–341.

[102] S. J. Pan and Q. Yang, *A survey on transfer learning*, IEEE Transactions on knowledge and data engineering, 22 (2009), pp. 1345–1359.

[103] ——, *A survey on transfer learning*, IEEE Transactions on knowledge and data engineering, 22 (2010), pp. 1345–1359.

[104] J. Pérez-Marcos, D. M. Jiménez-Bravo, J. F. De Paz, G. V. González, V. F. López, and A. B. Gil, *Multi-agent system application for music features extraction, meta-classification and context analysis*, Knowledge and Information Systems, 62 (2020), pp. 401–422.

[105] A. Perrin, *Social media usage*, Pew research center, (2015), pp. 52–68.

[106] F. Ricci, L. Rokach, and B. Shapira, *Recommender systems: introduction and challenges*, in Recommender systems handbook, Springer, 2015, pp. 1–34.

[107] I. Rish et al., *An empirical study of the naive bayes classifier*, in IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, 2001, pp. 41–46.

[108] G. Riva, F. Vatalaro, and F. Davide, *Ambient intelligence: the evolution of technology, communication and cognition towards the future of human-computer interaction*, vol. 6, IOS press, 2005.

[109] M. Rohrmeier, *A generative grammar approach to diatonic harmonic structure*, in Proceedings of the 4th sound and music computing conference, 2007, pp. 97–100.

[110] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, *ORB: An efficient alternative to sift or surf*, in 2011 International conference on computer vision, Ieee, 2011, pp. 2564–2571.

[111] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., *Imagenet large scale visual recognition challenge*, International journal of computer vision, 115 (2015), pp. 211–252.

[112] R. Sathya and A. Abraham, *Comparison of supervised and unsupervised learning algorithms for pattern classification*, International Journal of Advanced Research in Artificial Intelligence, 2 (2013), pp. 34–38.

[113] W. Schulze and B. Van Der Merwe, *Music generation with markov models*, IEEE MultiMedia, (2010), pp. 78–85.

[114] H. Shao, Y. Wu, W. Cui, and J. Zhang, *Image retrieval based on mpeg-7 dominant color descriptor*, in 2008 The 9th International Conference for Young Computer Scientists, IEEE, 2008, pp. 753–757.

[115] R. Shekhar and C. Jawahar, *Word image retrieval using bag of visual words*, in 2012 10th IAPR International Workshop on Document Analysis Systems, IEEE, 2012, pp. 297–301.

[116] I. Simon and O. Sageev, *Performance RNN: Generating music with expressive timing and dynamics*. https://magenta.tensorflow.org/performance-rnn, 2017. Accessed: 2018-02-19.

[117] K. Smith, *Kenzie smith piano - anime covers for piano*. https://kenziesmithpiano.com/anime-midi/, 2018. Accessed: 2018-01-27.

[118] Statista, *Variación porcentual del consumo de plataformas de streaming durante la cuarentena por el coronavirus en países seleccionados a fecha de marzo de 2020*. https://es.statista.com/estadisticas/1108893/covid-19-aumento-del-uso-de-plataformas-de-streaming-por-pais/, 2020.

[119] D. Stohr, T. Li, S. Wilk, S. Santini, and W. Effelsberg, *An analysis of the younow live streaming platform*, in 2015 IEEE 40th Local Computer Networks Conference Workshops (LCN Workshops), IEEE, 2015, pp. 673–679.

[120] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, *Rethinking the inception architecture for computer vision*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[121] Y. Takefuji, *Neural network parallel computing*, vol. 164, Springer Science & Business Media, 2012.

[122] S. A. K. Tareen and Z. Saleem, *A comparative analysis of sift, surf, kaze, akaze, orb, and brisk*, in 2018 International conference on computing, mathematics and engineering technologies (iCoMET), IEEE, 2018, pp. 1–10.

[123] G. Thevenot, *Blogging as a social media*, Tourism and hospitality research, 7 (2007), pp. 287–289.

[124] M. Thorogood and P. Pasquier, *Impress: A machine learning approach to soundscape affect classification for a music performance environment.*, in NIME, 2013, pp. 256–260.

[125] T. M. Thrash and A. J. Elliot, *Inspiration as a psychological construct.*, Journal of personality and social psychology, 84 (2003), p. 871.

[126] T. M. Thrash, E. G. Moldovan, V. C. Oleynick, and L. A. Maruskin, *The psychology of inspiration*, Social and Personality Psychology Compass, 8 (2014), pp. 495–510.

[127] L. V. Tran and R. Lenz, *Compact colour descriptors for colour-based image retrieval*, Signal Processing, 85 (2005), pp. 233–246.

[128] G. Tsoumakas and I. Katakis, *Multi-label classification: An overview*, International Journal of Data Warehousing and Mining (IJDWM), 3 (2007), pp. 1–13.

[129] G. Tsoumakas, I. Katakis, and I. Vlahavas, *Random k-labelsets for multilabel classification*, IEEE Transactions on Knowledge and Data Engineering, 23 (2010), pp. 1079–1089.

[130] G. Tzanetakis and P. Cook, *Marsyas: A framework for audio analysis*, Organised sound, 4 (2000), pp. 169–175.

[131] L. R. Varshney, F. Pinel, K. R. Varshney, A. Schörgendorfer, and Y.-M. Chee, *Cognition as a part of computational creativity*, in Cognitive Informatics & Cognitive Computing (ICCI* CC), 2013 12th IEEE International Conference on, IEEE, 2013, pp. 36–43.

[132] T. Vaughan, *Todo el poder de multimedia*, McGraw-Hill México, 1995.

[133] Walt Disney Pictures & Walt Disney Feature Animation, *Fantasia 2000*, 1999.

[134] Walt Disney Productions, *Fantasia*, 1940.

[135] G. Widmer, *Using ai and machine learning to study expressive music performance: Project survey and first report*, AI Communications, 14 (2001), pp. 149–162.

[136] G. A. Wiggins, *Searching for computational creativity*, New Generation Computing, 24 (2006), pp. 209–222.

[137] K. Wilkinson, A. Simitsis, M. Castellanos, and U. Dayal, *Leveraging business process models for ETL design*, in International Conference on Conceptual Modeling, Springer, 2010, pp. 15–30.

[138] F. Xia, L. T. Yang, L. Wang, and A. Vinel, *Internet of things*, International journal of communication systems, 25 (2012), p. 1101.

[139] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, and X. Geng, *Binary relevance for multi-label learning: an overview*, Frontiers of Computer Science, 12 (2018), pp. 191–202.

[140] M.-L. Zhang and Z.-H. Zhou, *ML-KNN: A lazy learning approach to multi-label learning*, Pattern recognition, 40 (2007), pp. 2038–2048.

[141] S. Zhang, C. Zhang, and Q. Yang, *Data preparation for data mining*, Applied artificial intelligence, 17 (2003), pp. 375–381.

[142] H. Zhong, H. Li, A. C. Squicciarini, S. M. Rajtmajer, C. Griffin, D. J. Miller, and C. Caragea, *Content-driven detection of cyberbullying on the instagram social network.*, in IJCAI, 2016, pp. 3952–3958.

[143] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, *Linear versus mel frequency cepstral coefficients for speaker recognition*, in 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, IEEE, 2011, pp. 559–564.

[144] J. L. Zimmermann, M. Clegg, E. de Bellis, and R. Hofstetter, *Smart products report 2020*, Research Platform Alexandria, (2020).

I never am really satisfied that I understand anything;
because, understand it well as I may, my comprehension
can only be an infinitesimal fraction of all I want to understand
about the many connections and relations which occur to me,
how the matter in question was first thought of or arrived at...

Ada Lovelace