UNIVERSIDAD CARLOS III DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

PROYECTO FIN DE CARRERA

INGENIERÍA TÉCNICA DE INFORMÁTICA DE GESTIÓN

Sistema de clasificación y exposición de características faciales SICECAF

AUTOR: ALBERTO GARCÍA GARCÍA-CASTRO

TUTOR: LUIS PUENTE RODRÍGUEZ

Septiembre de 2012

Proyecto Fin de Carrera

Sistema de clasificación y exposición de características faciales SICECAF

AUTOR:

Alberto García García-Castro

TUTOR:

Luis Puente Rodríguez

La defensa del presente Proyecto Fin de Carrera se realizó el día 28 de Septiembre de 2012 y fue evaluada por el siguiente tribunal:

PRESIDENTE: Israel González Carrasco

VOCAL: Diana M. Vásquez Bravo

SECRETARIA: Karina Tatiana Robles Palacios

A mi familia, a mis tutores/as, a mis amigos, a los alumnos de la facultad de Comunicación Audiovisual de la UC3M y a todos los voluntarios que han hecho posible este proyecto.

Alberto.

PRÓLOGO

En la actualidad las tecnologías relacionadas con el reconocimiento automático del habla se han desarrollado de manera exponencial. Gracias a la investigación en este campo se ha mejorado la interacción persona-máquina, obteniendo nuevos tipos de aplicaciones relacionadas con la comunicación.

Aunque las capacidades de los reconocedores del habla han aumentado en los últimos años siguen teniendo carencias importantes. Entre las más habituales destacan el ruido en el canal de transmisión y las ambigüedades del lenguaje, lo que provoca una falta de acierto considerable.

Para solucionar estos problemas se necesita aumentar las prestaciones de los sistemas anteriormente descritos, tanto las capacidades de los dispositivos de sonido, como los algoritmos de reconocimiento, teniendo en cuenta las señales visuales presentes en el habla.

En esta memoria se expone un sistema de reconocimiento facial que aumente las prestaciones de los reconocedores actuales. Se crea un sistema que combina diferentes métodos de visualización y discriminación de zonas faciales.

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

Las tecnologías que soportan los sistemas reconocedores del habla se han convertido en uno de los principales puntos de atención por parte de muchas empresas y entidades [Google 2012] [Apple 2012]. Como consecuencia, la investigación ha crecido notablemente en los últimos años en un intento de ofrecer implementaciones de sistemas más precisos y seguros.

Para conseguir unos mejores resultados, diferentes organizaciones y la comunidad científica están desarrollando varias líneas de investigación [Gales et al, 2008] [Norris et al, 2004] para mejorar y aumentar las capacidades de estos sistemas.

A partir de varios artículos de investigación [Duchnowski et al, 1994] y [Eveno et al, 2004] se plantea el desarrollo de un sistema reconocedor del habla que permitiese la innovación y la obtención de resultados fiables, mediante la adicción de nuevas fuentes de información.

Este PFC se integra dentro de las tecnologías de reconocimiento del habla y su objetivo es la ampliación de sus capacidades para obtener unas tasas de error inferiores a las que se presentan en la actualidad mediante la adición de características visuales. Para lograrlo es necesario extraer información del movimiento de los labios del individuo excluyendo el resto zonas faciales.

En concreto, el proyecto *"Sistema de Clasificación y Exposición de Características Faciales"* persigue establecer y evaluar algoritmos para reconocer las diferentes partes de la cara, posicionando a continuación la zona labial del hablante.

1.1. OBJETIVOS

En esta sección se enumeran las principales metas para establecer las bases de este proyecto: creación de una base de datos audiovisual, posicionamiento de la región labial en la imagen y reconocimiento de las partes que forman el rostro.

Mediante la grabación de esta nueva colección de vídeos se crea una herramienta de ayuda para los investigadores en este tipo de tecnología. En la actualidad la mayoría de bases de datos tienen como idioma característico el anglosajón, por lo que se ofrece un nuevo método de estudio para las personas de habla hispana.

En cuanto al reconocimiento automático de la región labial, se desarrolla un nuevo software capaz de localizar de manera automática la posición exacta de la boca del interlocutor. Ofrece además información acerca de las longitudes tanto verticales como horizontales de los extremos de la boca.

Por último, se crea una aplicación que clasifica y reconoce las zonas que forman el rostro humano. Para ello se identifican unívocamente cada una de ellas, obviando el resto de áreas corporales que no resultan de interés.

1.2. METODOLOGÍA

Esta investigación se lleva a cabo con el fin último de mejorar la tecnología relacionada con los dispositivos de reconocimiento del habla. El propósito principal consiste en añadir información visual de lectura labial al sistema reconocedor para mejorar sus resultados. Finalmente, este proyecto desarrolla un sistema que clasifica y expone las distintas partes del rostro. Para poder llevar a cabo el objetivo primario, se efectuarán nuevos trabajos en el futuro.

El sistema de clasificación y exposición de características faciales elaborado por el equipo de investigación de este proyecto está compuesto de:

- **Base de datos AV-UC3M:** se crea una colección de vídeos con un gran número de sujetos diferentes que sirven como entrada de datos a los sistemas de reconocimiento facial creados en fases posteriores. Para ello se escogen a una serie de personas que reúnan las características físicas necesarias para esta investigación. Se trata de la fuente de información principal de la que se van a extraer todas las características apropiadas tanto para desarrollar las aplicaciones, como para ejecutar sus pruebas de fiabilidad.

- **Sistema SIM-RC:** esta aplicación se desarrolla por este equipo investigador para realizar una clasificación y una exposición de las características faciales pertenecientes a un interlocutor de la base de datos. En concreto, el sistema selecciona automáticamente la parte bucal del rostro discriminando el resto de elementos de la cara mediante una serie de filtros por color. Además muestra al investigador las distancias verticales y horizontales de los extremos de los labios del hablante. Mediante esta funcionalidad se extraerán, en posteriores trabajos, las características labiales necesarias para proseguir con la adición de información al sistema reconocedor del habla.

- **Sistema MAT-RP:** este sistema se lleva a cabo como una nueva vía de investigación para la detección y discriminación de las partes del rostro. Se basa en la clasificación de diferentes zonas mediante la utilización de patrones de características faciales. Para poder desarrollar esta funcionalidad, se utilizan sistemas de clasificación como son las redes neuronales y las máquinas de vectores soporte. Finalmente este sistema diferencia automáticamente las zonas de la cara del hablante.

1.3. ESTRUCTURA DEL DOCUMENTO

El resto del documento se estructura de la siguiente forma: el capítulo 2 recoge el Estado del Arte de los reconocedores del habla. El capítulo 3 expone la base de datos AV-UC3M, que aporta toda la información necesaria para el desarrollo del proyecto. El documento continúa con el capítulo 4, en el que se presenta el sistema de reconocimiento SIM-RC. El capítulo 5 recoge el sistema MAT-RP encargado de mejorar las prestaciones de los reconocedores habituales. A continuación, el capítulo 6 muestra la gestión del proyecto y finalmente, el capítulo 7 ofrece las conclusiones y las líneas futuras de trabajo.

2. ESTADO DEL ARTE

A lo largo de este capítulo se presenta una visión general de los reconocedores del habla. En primer lugar se describen los más populares actualmente, a continuación se enumeran las principales bases de datos audiovisuales y finalmente se describen las diferentes etapas que forman el nuevo método de reconocimiento del habla mediante la adición de nuevos tipos de características.

2.1. INTRODUCCIÓN

En la siguiente sección se realiza un breve resumen de las principales características y atributos de la lengua. En concreto, se le da una mayor relevancia a las partes que forman el lenguaje, ya que es el motivo principal por el que se va a desarrollar esta investigación.

Los principales canales de comunicación se dividen según [Chen 2001] en dos: habla y señales visuales. Estos dos canales a menudo se complementan para hacer la comunicación más efectiva. Según [Chen 2001] otra característica importante del habla es su bimodalidad, compuesta tanto de audio, como de los siguientes aspectos visuales: la boca, la cavidad nasal, los dientes, la lengua, etc. Mediante la combinación de estos dos canales aparece el efecto McGurk [McGurk et al, 1976] que establece ciertas peculiaridades dentro de este campo. En él se describe cómo el discurso que percibe una persona no sólo depende de las señales acústicas, sino también de las señales visuales. En contraste con esta idea, en [Easton et al, 1982] se asegura que varios estudios psicológicos certifican que existe el inverso del efecto McGurk, es decir, que los resultados de la percepción visual del habla pueden ser afectados por el audio del discurso. Gracias a estas dos investigaciones se establece que ambos aspectos de la comunicación son lo suficientemente importantes como para tenerlos en cuenta.

Además de los estudios anteriormente citados, destaca el denominado efecto Lombard [Junqua et al, 1993] en el que se explica cómo la voz del hablante sufre modificaciones debido al ruido de fondo o a la degradación del audio, por las características adversas del canal de transmisión [Potamianos et al, 2001].

2.2 RECONOCIMIENTO DEL HABLA

Los sistemas de reconocimiento del habla se definen como sistemas informáticos con los que los seres humanos interactúan y en los que el lenguaje natural es parte importante de la comunicación [Fraser 1997]. Más concretamente, su cometido es convertir la expresión de entrada del usuario, formada por una señal continua en el tiempo, en una secuencia de unidades discretas, como son los fonemas y las palabras [McTear 2002].

Actualmente las aplicaciones de este tipo de tecnologías son muy numerosas, ya que cualquier tarea en la que se interactúe con un ordenador puede llevar implícito el reconocimiento de voz. Las más habituales son:

- **Dictado automático**: en algunos casos, como en el dictado de recetas médicas y diagnósticos o en el dictado de textos legales, se usan corpus especiales para mejorar los resultados del sistema.

- **Control por comandos**: sistemas de reconocimiento de habla diseñados para dar órdenes a un computador. Estos sistemas reconocen un vocabulario muy reducido, lo que incrementa su tasa de acierto.

- **Telefonía**: actualmente los teléfonos móviles permiten a los usuarios ejecutar comandos mediante el habla, en lugar de pulsar tonos.

- **Sistemas portátiles**: los sistemas portátiles de pequeño tamaño tienen unas restricciones muy concretas de tamaño y forma, de modo que el habla es una solución natural para introducir datos en estos dispositivos.

- **Sistemas diseñados para personas con discapacidad**: los sistemas de reconocimiento de voz pueden ser útiles para personas con discapacidades que no pueden teclear con fluidez, así como para personas con problemas auditivos, que consiguen usarlos para obtener texto escrito a partir del habla.

Aunque esta tecnología posea un potencial de desarrollo muy grande, el principal obstáculo aparece debido al alto grado de variabilidad de la señal de voz. Esta variación se produce en los siguientes casos [McTear, 2002]:

- **Variabilidad lingüística:** aparición de la variabilidad fonética, en el que los sonidos que se producen al hablar están determinados tanto por los fonemas anteriores como por los posteriores.

- **Variabilidad del hablante:** cada persona posee un tono de voz que lo diferencia del resto de sujetos. Además, en ciertas ocasiones ocurre que un mismo individuo puede cambiar su tono. Más concretamente, se pueden diferenciar características:

    - **Entre interlocutores**: diferencias como la edad, el género y el lugar de nacimiento.

    - **En el propio sujeto**: las mismas palabras dichas en diferentes ocasiones por el mismo hablante tienden a diferir en sus propiedades acústicas. También hay que tener en cuenta los factores físicos como el cansancio, la congestión de las vías respiratorias o los cambios en el estado de ánimo.

- **Variabilidad del canal:** introducción de sonidos de fondo en la conversación, lo que dificultará una comunicación efectiva entre los interlocutores.

A continuación se describen brevemente dos de los reconocedores de habla de código abierto más utilizados en la actualidad.

2.2.1 Sphinx

El reconocedor de habla Sphinx es un reconocedor de voz desarrollado por la Universidad de Carnegie Mellon en Estados Unidos. Se trata de un sistema creado en base a los fundamentos de los Modelos Ocultos de Markov (HMM). El lenguaje de programación que se ha utilizado para su desarrollo ha sido Java, favoreciendo de este modo la modularidad y la multiplataforma.

El sistema integra varios componentes dedicados a tareas específicas que realizan el reconocimiento, permitiendo conectarlos en tiempo de ejecución. La arquitectura del sistema se muestra a continuación:

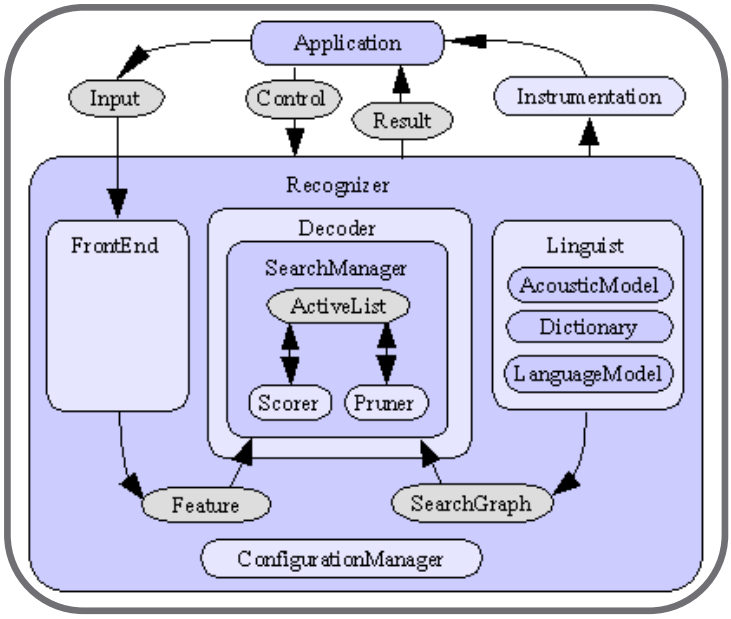


Figura 2.1: sistema de decodificación de Sphinx-4

Los módulos descritos en la figura anterior se unen para llevar a cabo la tarea del reconocimiento. Los más importantes son:

- **FrontEnd**: obtiene una o más señales de entrada y las transforma en parámetros que dan lugar a una secuencia de *features* (características).
- **Lingüista**: traduce todo tipo de lengua estándar recibida desde el FrontEnd, a un lenguaje reconocible por el módulo "Decoder" descrito a continuación.
- **Decoder**: el decodificador utiliza las características del FrontEnd y del módulo Lingüista para realizar la decodificación, generando los resultados.

Las principales ventajas que ofrece esta tecnología son:

- Funciona en una gran variedad de plataformas.
- El rico conjunto de APIs de plataforma reduce el tiempo de codificación.
- Soporta multithreading, lo que aumenta la velocidad de decodificación del lenguaje.
- La recolección automática de basura ayuda a los desarrolladores a concentrarse en el desarrollo de algoritmos en lugar de en las pérdidas en la memoria.

2.2.2 HTK

El *Hidden Markov Model Toolkit* (HTK) es una herramienta desarrollada para la construcción y manipulación de modelos ocultos de Markov. Se diseñó mediante una técnica de modelización de datos secuenciales aplicada al campo del reconocimiento automático del habla [Rabiner, 1989]. Actualmente es una herramienta casi imprescindible, que ha encontrado aplicación en disciplinas diversas: análisis de imagen [Aas et al, 1999], psicología [Visser et al, 2002] o bioinformática [Durbin et al, 1998].

HTK se emplea principalmente para la investigación de reconocimiento de voz, aunque se ha utilizado en numerosas otras aplicaciones, incluyendo la investigación de síntesis de voz, reconocimiento de caracteres y en la secuenciación del ADN.

Se compone de un conjunto de módulos y herramientas disponibles en código C. El software se puede utilizar para construir sistemas complejos HMM (*Hidden Markov Model)*. El desarrollo completo de una aplicación de HTK se realiza en cuatro etapas: preparación de las muestras, entrenamiento, pruebas y análisis (resultados).

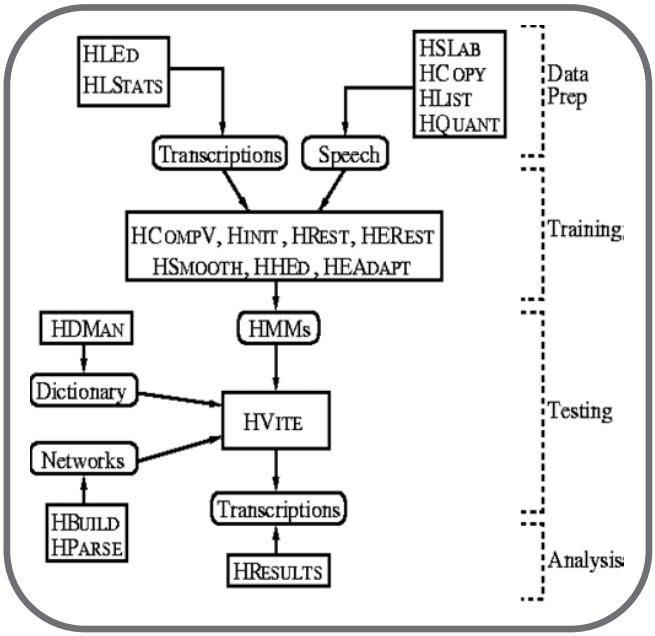


Figura 2.2: esquema de desarrollo de HTK

Este reconocedor es ideal para experimentar con modelos de lenguajes y modelos de Markov, dando la posibilidad de probar diferentes metodologías de entrenamiento. Otra ventaja es su gran capacidad como sistema etiquetador, definido como el proceso de asignar a cada una de las palabras de un texto su categoría gramatical. Con él se puede etiquetar cualquier base de datos lingüística.

2.3 RECONOCIMIENTO DEL HABLA MULTIMODAL

Para mejorar el reconocimiento automático del habla se han creado diferentes técnicas, que logran unos resultados exitosos en ambientes ruidosos, como se puede apreciar en: [Wang et al, 2007], [Citengul et al, 2005], [Jiang et al, 2001].

Las afirmaciones expuestas en [Potamianos et al, 2004] describen como el reconocimiento automático audiovisual del habla es la mejor forma de añadir información a los reconocedores de habla tradicionales, ya que no interfieren en el sistema ni el sonido ambiente ni el ruido.

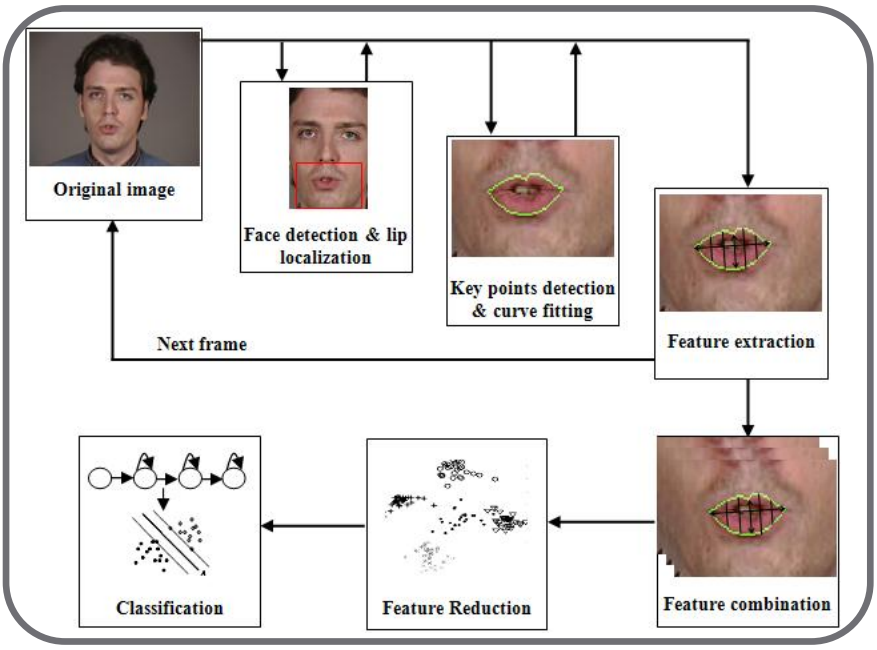


Figura 2.3: reconocimiento multimodal

A pesar del entusiasmo inicial generado a partir del desarrollo del primer sistema audiovisual [Petajan 1984], aparecieron problemas difíciles de abordar, como fueron el diseño visual de la interfaz o la fusión de las señales visual y auditiva. Para intentar minimizar todas estas dificultades se llevaron a cabo importantes trabajos de investigación en la comunidad científica [Potamianos et al, 2004]. A continuación se desarrolló un reconocedor únicamente visual con un vocabulario de 100 palabras que incluía dígitos y letras, descrito en [Rabiner et al, 1993].

Tres años más tarde, se estableció un grupo encabezado por Christian Benoit en el Instituto de comunicación oral PIC, en Grenoble, Francia. Su trabajo continuó con un nuevo reconocedor audiovisual realizado por [Heckmann et al, 2001], en el que se realiza una combinación de redes neuronales (NN) junto con modelos ocultos de Markov (HMM).

Actualmente, otros investigadores continúan trabajando en el campo de los reconocedores multimodales mejorando las tasas de error y aumentando la robustez de estos sistemas [Hazen 2006], [Nilsson et al, 2007].

A pesar del progreso realizado en la extracción de características visuales, la incertidumbre sigue apareciendo al tratar de precisar qué tipo de características son las más importantes en los entornos

visuales. Además tanto la cámara utilizada como el entorno que rodea a los hablantes siguen condicionando la tecnología de habla audiovisual.

Aun así, vale la pena mencionar dos argumentos en favor de este nuevo tipo de tecnología:

- Está demostrado que la información más significativa entre la comunicación de los seres humanos es la visual. La información aportada por todos los elementos de la boca (dientes, lengua) es mayor que la aportada únicamente por los labios [Summerfield et al, 1989].
- El rendimiento necesario para la extracción de características es computacionalmente razonable. Es cierto que en la etapa de detección bucal el rendimiento del sistema es bajo. Aunque en la etapa de creación de las matrices de características la velocidad de extracción es muy elevada [Prensa et al, 1995]. Estos hechos permiten la aplicación en tiempo real de los sistemas automáticos de la lectura labial.

El citado sistema automático audiovisual de reconocimiento del habla se divide en las siguientes fases [Potamianos et al, 2003]:

- Seguimiento de la cara y extracción de la región de interés de la boca (ROI) [Hermansky et al, 1994].
- Extracción de las características visuales [You et al, 1999].
- Integración de audio y de vídeo [Potamianos et al, 2004].

2.4 BASES DE DATOS AUDIOVISUALES

A diferencia de los sistemas de reconocimiento de voz estándar, donde las bases de datos son abundantes, sólo un pequeño número están disponibles para el reconocimiento audiovisual de la voz. Esto se debe a una serie de factores como la complejidad de adquisición y procesamiento de datos o a que la investigación se lleva a cabo principalmente por investigadores individuales o pequeños grupos [Ahmad et al, 2008].

Las bases de datos que están disponibles a menudo tienen una mala calidad de vídeo, con un limitado número de hablantes y, por tanto, son poco adecuadas para los experimentos de reconocimiento de voz [Ahmad et al, 2008]. En la actualidad hay varias bases de datos apropiadas para realizar investigaciones en lo referente a sistemas de reconocimiento audiovisual:

- **VidTIMIT**: se compone de 43 sujetos (24 varones y 19 mujeres). Cada persona recita ocho oraciones diferentes delante de una cámara centrada en la parte frontal del interlocutor. Las sentencias en la base de datos son ejemplos de discurso continuo y contienen un total de 216 frases con un vocabulario de 925 palabras.

  El audio se grabó con una tasa de 32 kHz, y el vídeo con una tasa de 25 frames por segundo [Sanderson et al, 2002] [Hazen et al, 2004].

Figura 2.4: base de datos VidTIMIT

- **FERET** [Philips et al, 2000]: fue desarrollada para el reconocimiento automático facial. Contiene 14.051 imágenes en 2D de rostros, incluyendo diferentes posturas, expresiones faciales y variaciones en la iluminación. Hay varios individuos que buscan la diferenciación mediante gafas de sol, diferentes cortes de pelo, etc. Es una de las bases de datos más populares.

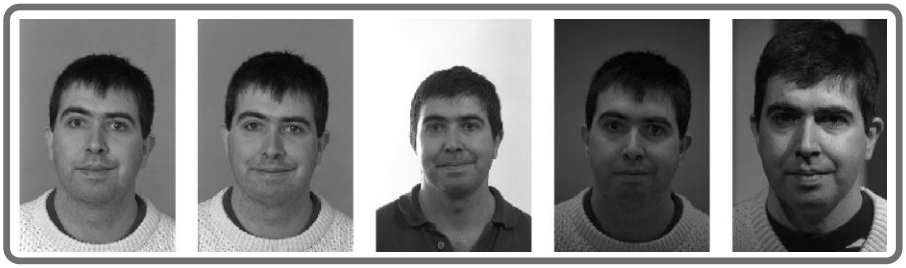


Figura 2.5: base de datos FERET

- **Yale DB** [Belhumeur et al, 1997]: creada por el Centro de Visión Computacional y Control de la Universidad de Yale. Contiene imágenes de rostros en escala de grises en 2D de 15 individuos. Hay 11 imágenes diferentes por cada uno de ellos: imágenes normales, con gafas y sin ellas, con variaciones de iluminación (varía según la posición de la fuente de luz: en el centro, a la izquierda y a la derecha) e imágenes de rostros con diferentes expresiones faciales, tales como felicidad, tristeza, sorpresa, etc.

Figura 2.6: base de datos de Yale DB

- **AR** [Martínez et al, 1998]: fue creado por el Centro de Visión por Computador de la Universidad Autónoma de Barcelona (España). Contiene imágenes del rostro en color 2D correspondientes a 126 individuos (70 hombres y 56 mujeres).

  Presenta cuatro expresiones faciales diferentes, distintas condiciones de iluminación, gafas de sol, etc. Las imágenes fueron capturadas bajo un estricto control de las condiciones. No hubo restricciones sobre la presencia de gafas, maquillaje, estilo de la audición, etc. Cada persona ha colaborado en dos sesiones, separadas por dos semanas.

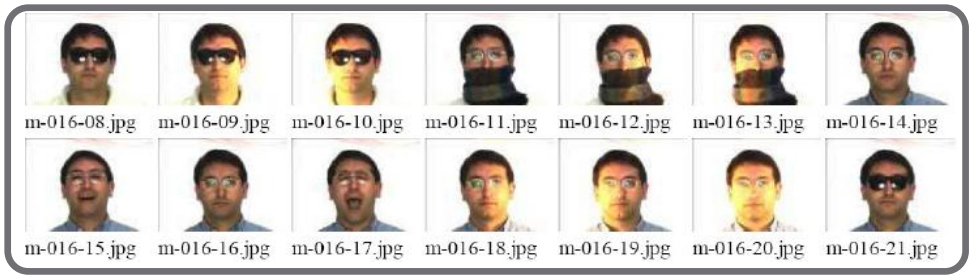


Figura 2.7: base de datos AR FaceDatabase

- **MIT Database** [Turk et al, 1991]: fue construida por el Laboratorio del MIT (Massachusetts Institute of Technology). Contiene imágenes de la cara en 2D que corresponden a 16 individuos del sexo masculino. Incluye imágenes con diferentes tipos de orientaciones faciales (izquierda y derecha), 3 variaciones de iluminación y 3 variantes de escala (variando la zoom de la cámara). Las imágenes poseen diferentes resoluciones.

Figura 2.8: base de datos del MIT

- **ORL** [Samaria et al, 1994]: creada por la Fundación AT&T (Cambridge). Contiene imágenes en 2D de 40 personas, en el que hay 10 imágenes diferentes por persona.

  Las imágenes fueron capturados en momentos diferentes, variando la iluminación, las expresiones faciales (ojos abiertos/ojos cerrados, risa/no risa) y otros detalles (gafas/sin gafas). El fondo es oscuro y homogéneo. Cada sujeto fue capturado mirando hacia arriba, a la derecha y en posición frontal.

Figura 2.9: base de datos ORL

- **PF01**: contiene imágenes de rostros de personas de Asia, la mayoría de ellos procedentes de Corea. Incluye variaciones con respecto a la iluminación, la pose y la expresión facial. Esta base de datos ofrece suficientes variaciones entre las imágenes de cada individuo. Es apropiada para la evaluación de sistemas de reconocimiento automático de caras sobre los individuos con rasgos asiáticos.

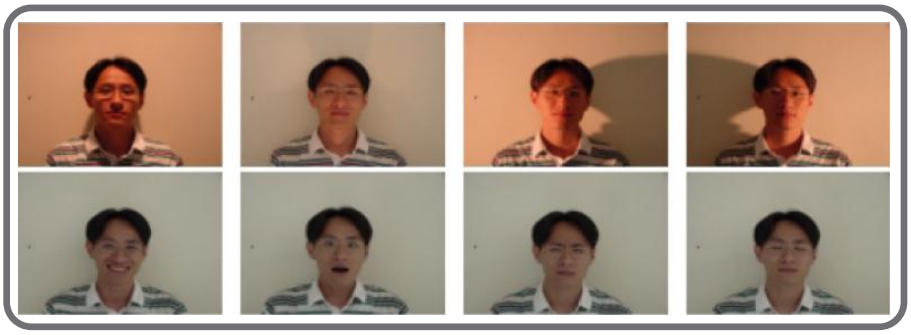


Figura 2.10: base de datos PF01

- **XM2VTS** [Messer et al, 1999] [Matas et al, 2000]: es una gran base de datos orientada hacia pruebas de verificación multimodal. Fue creada por el Centro para la Visión, el Habla y el Procesamiento de la señal de la Universidad de Surrey (Reino Unido). Contiene 4 sesiones de captura de 295 personas, en intervalos de 1 mes.

  En cada sesión se capturó el habla de un individuo mientras iba girando su cabeza. Como resultado se extrajeron imágenes en color de alta calidad, archivos de sonido y secuencias de vídeo digitalizadas. Más tarde se obtuvo un modelo 3D de cada individuo en el tercer período de sesiones. Se trata de una base de datos comercial.

Figura 2.11: base de datos XM2VTS

- **Base de datos 3D de la Universidad de York**: las bases de datos faciales en 3D mencionadas con anterioridad figuran como un conjunto reducido de imágenes faciales por sujeto. Esta base de datos tiene imágenes correspondientes a 97 individuos. Contiene 10 capturas por individuo con diferentes posturas. Sin embargo, en sólo 2 realiza diferentes expresiones faciales (sonrisa y ceño fruncido).

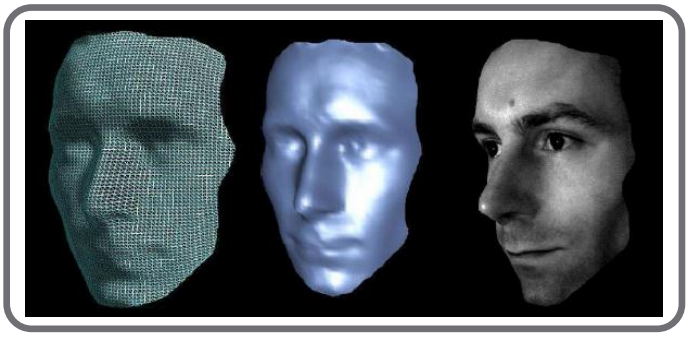


Figura 2.12: 3D FaceDatabase

- **CUAVE** [Patterson et al, 2002] está dividida en dos partes bien diferenciadas. En primer lugar se presenta a 36 sujetos (17 mujeres y 19 hombres) y a continuación 20 parejas de sujetos que realizarán un discurso similar. Está comprimido en formato MPEG-2, con un audio de 44kHz y con una velocidad de transmisión de datos de 5000 kbps.

Figura 2.13: base de datos CUAVE

En el siguiente cuadro resumen se pueden comparar las diferentes características de la mayoría de bases de datos expuestas anteriormente. Como se puede observar, existen datos que no se han podido obtener de las bases de datos debido a la falta de información proporcionada por parte de sus creadores.

	Nombre	Hablantes	Hombres	Mujeres	Palabras	Color	Formato	Audio	AudioRate	FrameRate	Resolución	2D	3D	Población
1	VidTIMIT	43	24	19	925	No	Mpeg-1	WAV	32 kHz	25 fps	384 x 512	Sí	Sí	Caucásica
2	FERET	1148	-	-	-	No	-	-	-	-	256 x 384	Sí	No	-
3	Yale DB	15	-	-	-	No	No	No	-	-	640 x 480	Sí	No	-
4	AR	126	70	56	-	Si	-	-	-	-	768 x 576	Sí	No	-
5	ORL	40	-	-	-	No	-	-	-	-	92 x 112	Sí	No	-
6	PF01	103	53	50	-	Sí	-	-	-	-	1280 x 960	Sí	No	Asiática (Corea)
7	XM2VTS	295	-	-	-	Sí	-	-	23kHz	-	576 x 720	Sí	Sí	-
8	CUAVE	36	19	17	60	Sí	Mpeg-2	WAV	44kHz	29.97 fps	720 x 480	Sí	No	Caucásica, africana

Tabla 2.1: características Bases de Datos Audiovisuales

Como se puede observar, cada una de ellas posee características que las diferencian de las demás. Por ejemplo, FERET tiene un número de participantes que supera con creces la media o FT01 está formada únicamente por personas asiáticas. En cuanto al color de imagen, casi la mitad de ellas carece de esta propiedad.

2.5 LOCALIZACIÓN FACIAL

En esta sección se describe una visión general de los distintos métodos que se han utilizado hasta ahora para la localización del rostro dentro de una imagen. En concreto se exponen casos desde la última década del siglo pasado hasta la actualidad.

En un primer momento, los investigadores [Senior 1999] y [Neti et al, 2000] desarrollan un nuevo algoritmo para la detección y localización de las características faciales. Se basa en la segmentación de los tonos de piel utilizando como base de información vídeos grabados en color. Mediante esta técnica se reduce rápidamente el tiempo de búsqueda de candidatos. Una vez que el sistema encuentra la cara dentro de la imagen, se utiliza un conjunto de detectores de rasgos faciales para estimar la ubicación de 26 tipos de características dentro del rostro.

Unos años más tarde, el equipo de [Viola et al, 2004] desarrolló un método extremadamente robusto de detección facial en tiempo real. En primer lugar, el sistema crea una nueva representación de la imagen eliminando las zonas no relevantes, para poder llevar a cabo una exploración de las zonas faciales mucho más rápida. A continuación se clasifican las diferentes partes de la cara mediante la combinación de varios sistemas clasificadores de características en cascada. Gracias a esta metodología se consiguió aumentar significativamente la velocidad de los nuevos detectores.

Figura 2.14: detección facial

Dos años después [Arsic et al, 2006] desarrolló un nuevo método de búsqueda, realizando primero una detección facial y buscando a continuación las zonas de interés. Una vez localizadas estas áreas, se enmarcan dentro de un cuadrado superpuesto encima de la imagen original. A continuación se utiliza una plantilla para identificar la región nasal en los sucesivos fotogramas, estabilizando de esta manera el proceso de detección.

Al encontrar la posición de la nariz, se utilizan propiedades antropométricas humanas como la distancia entre ojos y la zona labial para encontrar de este modo el centro de la boca desde el primer fotograma. Como consecuencia, la región de interés es localizada y extraída como se muestra en la siguiente imagen:

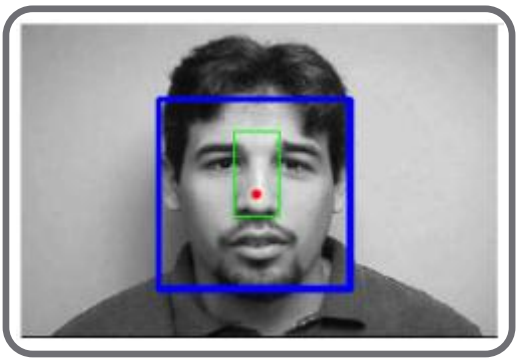


Figura 2.15: ejemplo de detección facial

Por último, el equipo de investigación de [Nilsson et al, 2007] lleva a cabo una transformación de la iluminación en las imágenes que se introducen en el sistema para eliminar el ruido (variación aleatoria del brillo o el color) presente en las mismas. A continuación se utiliza la función de transformación SMQT expuesta en [Nilsson et al, 2005] para extraer las características buscadas en la imagen. Por último, se lleva a cabo la identificación de las regiones de interés, mediante la modificación de un clasificador llamado SNoW, descrito en [Froba et al, 2004].

Figura 2.16: ejemplo de detección de rostros

2.6 LOCALIZACIÓN BUCAL

Gracias a los métodos expuestos con anterioridad se llega a la identificación final de la región de la boca de manera automática. A continuación se utilizan diferentes técnicas para obtener estimaciones del contorno labial. Los métodos que se utilizan más a menudo son los siguientes: formas activas y modelos de apariencia [Cootes et al, 1998], serpientes [Kass et al, 1988] y plantillas [Silsbee 1994]. A continuación se explicarán de manera breve todas estas técnicas.

La forma activa y los modelos de apariencia son algoritmos que dan como resultado diseños con forma labial mediante elementos estadísticos. Estos modelos se utilizan para el seguimiento de los movimientos labiales mediante la actualización constante de los parámetros que indican el lugar que ocupa la boca en la imagen [Cootes et al, 1998]. El algoritmo que se utiliza para ajustar el modelo a los datos de la imagen aparece en [Matthews et al, 1998].

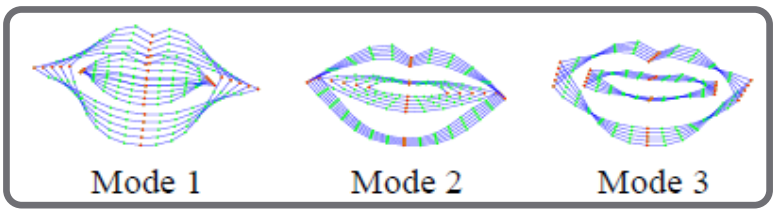


Figura 2.17: técnica de forma activa

En cuanto al método de serpientes, está basado en la creación de una curva representada por un conjunto de puntos de control que delimitan el contorno labial. Estos puntos son actualizados periódicamente para ajustarse lo máximo a la forma exacta de la boca [Kass et al, 1988].

Otro ejemplo de empleo de la técnica de la serpiente se expone en [Eveno et al, 2004]. En este artículo se describe la creación de un nuevo método casi automático que sólo requiere de la selección manual de un único punto situado encima de la boca. A continuación se emplea la nueva técnica llamada "*jumping snake*", en la que se van posicionando diferentes puntos de manera automática aproximándose a la zona bucal.

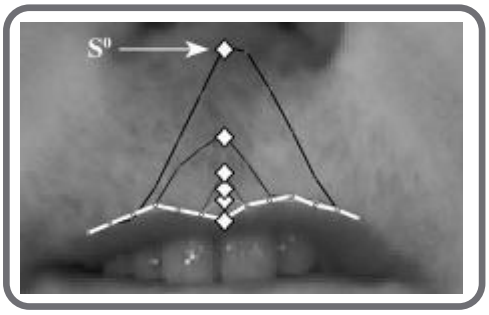


Figura 2.18: jumping snake

La última técnica descrita en este apartado está compuesta por las plantillas labiales. Esta metodología utilizada por [Chandramohan et al, 1996] crea plantillas compuestas por curvas parametrizadas que se ajustan a la forma de la boca del individuo.

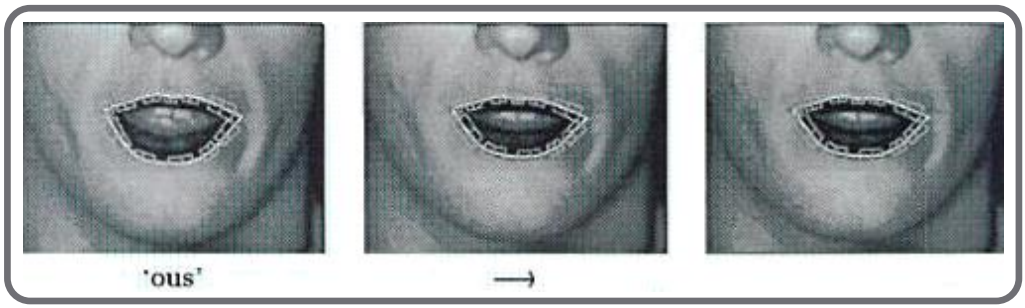


Figura 2.19: ejemplo de la técnica de plantillas

2.7 EXTRACCIÓN DE CARACTERÍSTICAS BUCALES

En los últimos 20 años se han propuesto en la literatura varios métodos de extracción de características visuales para la lectura automática labial. En general, se pueden agrupar en tres categorías: apariencia, contorno y perfil.

2.7.1 APARIENCIA

En este enfoque de extracción de características visuales, la parte de la imagen que contiene la región de la boca considerada óptima para leer los labios se denomina región de interés (ROI) [Potamianos et al, 2004]. Dicha región es un rectángulo que delimita la boca, y posiblemente incluye grandes partes inferiores de la cara, como la mandíbula y las mejillas [Potamianos et al, 2001] o toda la cara [Matthews et al, 2001].

A menudo, la región de interés se puede resaltar mediante un rectángulo de tres dimensiones, en un esfuerzo por capturar la información en el habla [Potamianos et al, 1998]. Por otra parte, la ROI puede corresponder a una serie de imágenes de perfil [Dupont et al, 2000] o ser simplemente un disco alrededor del centro de la boca [Duchnowski et al, 1994]. Por la concatenación de los píxeles en escala de grises [Dupont et al, 2000] o los valores de color [Chiou et al, 1997] se obtiene un vector de características.

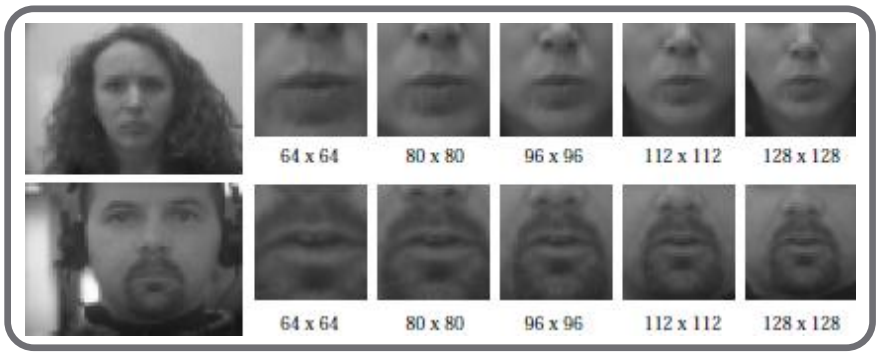


Figura 2.20: enfoque de apariencia

2.7.2 CONTORNO

La forma basada en la extracción de características de contorno asume que la mayor parte de la información relativa a la lectura labial se encuentra dentro de los contornos labiales del hablante [Matthews et al, 2001] o más generalmente en los contornos faciales (la mandíbula y la forma de las mejillas).

Hay dos tipos de extracción de características dentro de esta categoría: la de tipo geométrico y el modelo de forma basado en características. En ambos casos, el algoritmo recoge tanto el interior como el exterior del contorno labial.

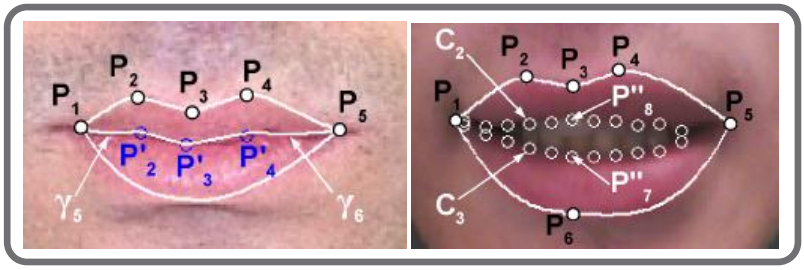


Figura 2.21: enfoque de forma

2.7.3 PERFIL

Otra forma de extraer las características labiales del hablante se estudiaron en [Potamianos et al, 2006]. Los autores compararon los resultados extraídos de las imágenes frontales de la cara con los de las imágenes de perfil.

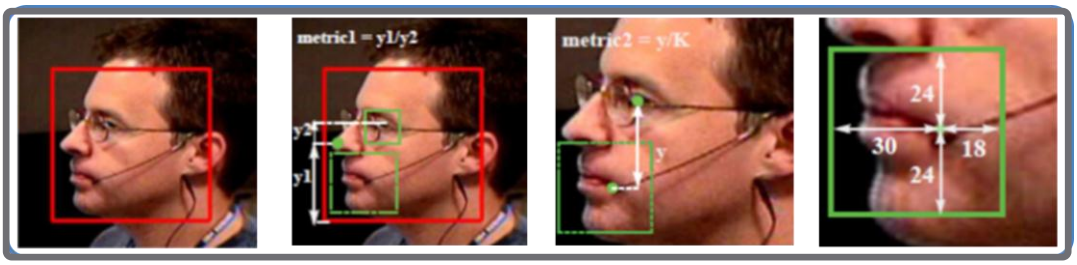


Figura 2.22: enfoque de perfil

Se concluye que la mejor manera de extraer las características es combinar tanto la imagen frontal como la lateral. Gracias a ello se mejoró la tasa de error aproximadamente en un 20% con respecto al reconocimiento único de la voz.

2.8 FUSIÓN AUDIO-VISUAL

Una vez decididas las características a extraer, es necesario encontrar un método fiable de combinación de las mismas. Con el fin de hacer frente a este problema se han explorado diferentes enfoques: redes neuronales [Lippman 1990] [Rothkrantz et al, 2000], máquinas de vectores soporte [Ganapathiraju 2002] y Modelos Ocultos de Markov. Al final, la mayoría de autores se ha decantado por utilizar este último método para realizar las aproximaciones de probabilidades [Chitu et al, 2007].

Sin embargo, al realizar la combinación de varios tipos de datos surgen los problemas esbozados a continuación:

- Las señales pueden tener diferentes rangos: la duración del sonido de un fonema puede ser menor que la duración correspondiente al labio en movimiento.
- Puede haber un desfase de tiempo entre las señales. La lectura del movimiento labial se suele realizar antes que la señal de audio. Ambas suelen sincronizarse de manera asíncrona.

- Las señales pueden ser grabadas a diferentes ritmos. En particular, las tasas de vídeo suelen ser más lentas que las tasas de audio (generalmente la velocidad de fotogramas de vídeo es de 25-30fps mientras que la señal de audio es de 100fps).

En la siguiente imagen se aprecia el proceso que se realiza desde la obtención de características por separado hasta su integración en un único vector de características:

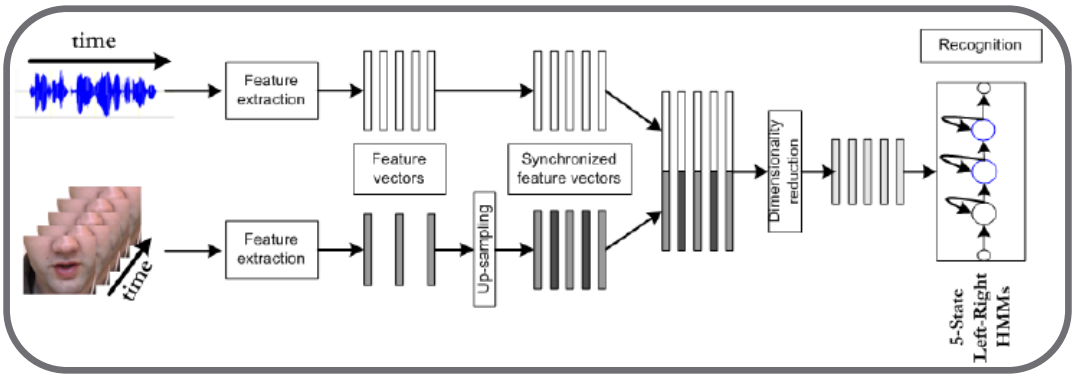


Figura 2.23: integración audiovisual

En esta etapa la mayoría de los investigadores realizan el mismo procedimiento. Unifican todos los datos obtenidos tanto de la parte visual como del sonido para introducirlos a continuación en un clasificador que fusione los datos.

En [Iwano et al, 2007] se realiza la fusión de las características de audio y vídeo. Cada trama de voz se convierte en 38 parámetros acústicos. En cuanto a las señales visuales, están definidas por los parámetros RGB (Red Green Blue) de cada píxel.

Tanto el audio como el vídeo se graban de forma simultánea. A continuación se realiza una normalización de ambas características y se lleva a cabo una fusión de los vectores resultantes introduciéndolos en un clasificador HMM *(Hidden Markov Model)*.

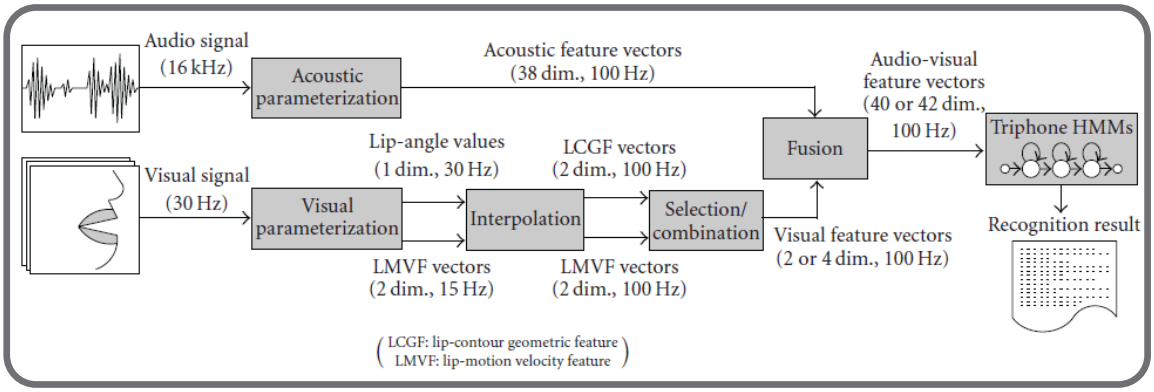


Figura 2.24: sistema bimodal del habla

3 BASE DE DATOS AV-UC3M

Este capítulo presenta una visión general de las partes de las que se compone la base de datos AV-UC3M. En primer lugar se describen los objetivos que debe cumplir, a continuación se muestra el proceso de elaboración y, por último, se detallan sus características.

3.1 INTRODUCCIÓN

En la actualidad, el interés de los investigadores por el procesamiento del lenguaje audiovisual ha aumentado de manera considerable debido a la mejora de la potencia de cálculo en los computadores actuales [Li et al, 2009]. Al añadir información visual a la hora de realizar el reconocimiento de voz se compensa la pérdida de información que se produce debido a ruidos externos. Por todo ello, el reconocimiento audiovisual de voz puede superar a los reconocedores de habla tradicionales en ambientes ruidosos o con más de un interlocutor.

Debido a la novedad de este tipo de investigaciones, así como a la dificultad de grabar muestras representativas con una gran cantidad de sujetos, se ha limitado la creación de este tipo de bases de datos. Por ello, la mayoría de los investigadores han tenido que crear sus propias muestras.

Las carencias expuestas con anterioridad provocan la creación de esta base de datos que permite el acceso a muestras audiovisuales de alta calidad, con diferentes condiciones de iluminación y un número elevado de participantes.

3.2 OBJETIVOS

En esta sección se exponen los diferentes objetivos por los que es necesaria la creación de esta base de datos audiovisual. A partir de ellos se desarrollan todas las funcionalidades para poder ejecutar el reconocimiento facial en posteriores etapas.

El primer objetivo consiste en realizar una colección de vídeos que cuente con un número elevado de personas. La base de datos ofrece una diversidad enorme de sujetos para llevar a cabo trabajos de investigación de manera robusta.

A continuación, esta colección de vídeos debe contar con una alta calidad tanto de imagen como de sonido. El vídeo será en color, contando con un equipo de grabación de altas prestaciones y con unas condiciones estudiadas de iluminación. En cuanto al sonido, se elimina cualquier ruido de fondo que no pertenezca al discurso del hablante.

El siguiente objetivo establece la elección de un corpus que ofrezca diferentes tipos de discurso. Si la alocución es siempre lineal, las pruebas que a los que se someten a los sistemas reconocedores se reducen. Los sujetos de esta base de datos deben recitar números, frases cortas y ejemplos del lenguaje natural.

Por último, se debe fijar como idioma predeterminado el castellano. Actualmente en el mercado la mayor parte de bases de datos están realizadas por personas angloparlantes. Por este motivo, se ofrece una nueva funcionalidad a los investigadores que deseen trabajar en lengua hispana.

3.3 FUNDAMENTACIÓN

En esta sección se describen las etapas por las que este equipo de investigación ha pasado hasta lograr elaborar esta base de datos. En un primer momento se elige el formato televisivo hasta que finalmente se decide llevar a cabo esta colección de vídeos.

La implantación de la televisión digital terrestre (TDT) en España se mejora notablemente tanto la resolución de la imagen como la modulación del sonido. Por estas razones, se decide comenzar a utilizar este formato como el predeterminado para el proyecto.

En un principio se realizaron las pruebas necesarias en diferentes tipos de formato audiovisual: películas, series y programas de televisión. Finalmente se llega a la conclusión de que el formato que más se adecúa a los objetivos del proyecto son los programas de informativos.

Figura 3.1: imágenes informativos

Una vez tomada la decisión de utilizar este formato se producen una serie de dificultades que ponen en riesgo la consecución de los objetivos propuestos inicialmente. En concreto, aparecen en el momento de reunir una batería de vídeos suficiente para poder continuar con la investigación. Los problemas se pueden resumir en:

- **Cambios de plano:** se realizan cambios de plano frecuentemente, lo que dificulta la obtención de imágenes adecuadas que se ajusten a los objetivos.

Figura 3.2: tipos de plano

- **Sombras en la imagen:** en muchas ocasiones el interlocutor aparece de perfil o con sombras en la cara debido a la iluminación.

Figura 3.3: imagen con sombra

- **Discurso variable:** para realizar una comparación en los resultados de clasificación es necesario que el discurso posea las mismas características para todos los integrantes. En este formato resulta casi imposible conseguir discursos similares.

Tras la aparición de estas dificultades en la edición de los telediarios como fuente de información se decide realizar una base de datos audiovisual propia, a fin de que posteriormente se cumplan todos los requisitos establecidos.

3.4 CARACTERÍSTICAS

El proceso de grabación se realiza en la Facultad de Audiovisuales de la Universidad Carlos III de Madrid, en su Campus de Getafe. Esta ubicación es seleccionada debido a la existencia de varios platós de televisión, beneficiando de esta manera el cumplimiento de los objetivos de esta investigación. Este espacio cuenta con un sistema profesional de iluminación y con áreas específicas que para favorecer el manejo de cámaras.

Los criterios establecidos para llevar a cabo la base de datos AV-UC3M se basan en tres aspectos diferentes:

- **Sujetos:** se tienen en cuenta aspectos como el número de la muestra, su género, complementos de moda, maquillajes y la edad.
- **Características técnicas:** los elementos más importantes se resumen en iluminación, equipo de grabación y posibilidades de edición de imagen.
- **Corpus:** el discurso que reciten los hablantes posee varios apartados, en idioma castellano y con una duración apropiada.

Para proporcionar un mayor conocimiento de estas tres características, se describen de manera exhaustiva en las siguientes secciones.

3.4.1 SUJETOS

Una buena selección de los sujetos que forman la colección de vídeos resulta fundamental para obtener unos resultados óptimos. Este aspecto es la parte más importante de la base de datos pues de dichos sujetos se van extraer todas las peculiaridades necesarias para las siguientes fases, como pueden ser las partes del rostro o la posición de la boca.

La base de datos está compuesta por un gran número de individuos seleccionados en base a las siguientes características: acentos en el habla, color de piel, color de ojos, y complementos de moda como pendientes, anillos o maquillajes.

En total, la muestra consta de 108 sujetos, en concreto 69 mujeres y 39 hombres. Todos ellos tienen la nacionalidad española y son castellanoparlantes. Los colores de piel varían desde el blanco caucásico hasta el negro africano.

En cuanto al color de ojos, la mayor parte de los sujetos los tiene de color marrón oscuro, aunque también existe cierta representación de tonos azules y verdes. La forma de los ojos es europea, ya que no hay ningún interlocutor asiático.

La presencia de complementos en los sujetos dificulta la discriminación de zonas faciales, ya que habitualmente las características que se tienen en cuenta a la hora de reconocerlas son el color y la posición. Por ejemplo, en caso de que el color se modifique por medio de maquillajes como puede ser un pintalabios morado, el sistema de selección de partes de la cara no asociará esa zona con los labios, descartando esos píxeles, perjudicando de esta manera la clasificación. Aun así no se lleva a cabo ninguna restricción para que la muestra sea representativa.

Las edades de los participantes están comprendidas entre los 13 y los 58 años. La mayoría de los integrantes tienen entre veinte y treinta años, por lo que predomina la gente joven.

3.4.2 CARACTERÍSTICAS TÉCNICAS

La posición de la cámara siempre ha sido fija, realizando únicamente primeros planos a una distancia constante.

Existen dos tipos de iluminación: una con tonos cálidos en la cual participan 12 participantes y otra con un matiz más claro, con una muestra de 96 sujetos.

El fondo de la imagen es un *Chroma key* de color verde. Este fondo se ha seleccionado por dos motivos: el primero es que puede ayudar a encontrar el contorno facial debido al contraste de color y el segundo es la posibilidad de añadir diferentes tipos de fondos dependiendo del interés del investigador.

Figura 3.4: modificación del chroma

La base de datos fue grabada con una cámara Canon con una resolución de 576x720 píxeles, con una tasa de 25 fps, utilizando el formato AVI. En cuanto al sonido, fue grabado directamente con el micrófono de la cámara con un rate de 48 kHz y una resolución de 16 bits.

3.4.3 CORPUS

El corpus de la base de datos se define como el texto recitado por los integrantes que la forman, y que sirve de base para realizar la investigación. El discurso está formado por 439 palabras, dividido en cuatro partes:

- **Serie de nombres y apellidos:** consta de 21 palabras y trata de unir elementos del habla con pausas entre palabra y palabra.
- **Serie de números del 0 al 10:** está constituido por 11 palabras y da lugar a una serie de palabras con pausas entre las mismas.
- **Primer párrafo del libro "El Quijote":** formado por 33 palabras conocidas por la mayoría de los las personas que forman la base de datos.
- **Cuento:** compuesto de 374 palabras que forma una muestra representativa de habla habitual.

En el Anexo B se adjunta el corpus comentado en este apartado.

3.5 ESTADÍSTICAS

En esta sección se ofrecen una serie de estadísticas extraídas de la base de datos. Con ello se pretende ofrecer una visión más específica de las características que poseen los participantes que la forman. Son las siguientes:

- **Género:** en esta figura se muestra que predomina el género femenino en un 30% con respecto al masculino.

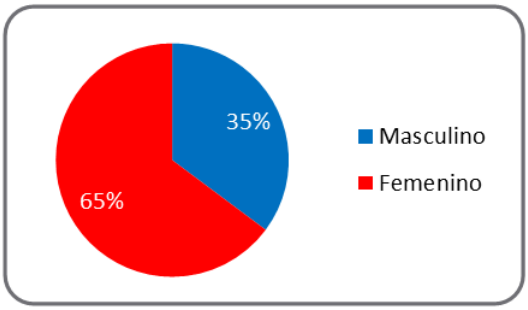


Figura 3.5: géneros

- **Color de pelo:** el color predominante es el castaño muy por delante del resto.

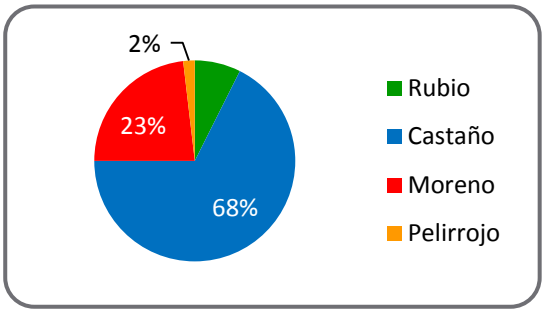


Figura 3.6: color de pelo

- **Color de ojos:** la información que nos ofrece esta gráfica nos indica el predominio del color de ojos marrón con respecto a los azules y verdes.

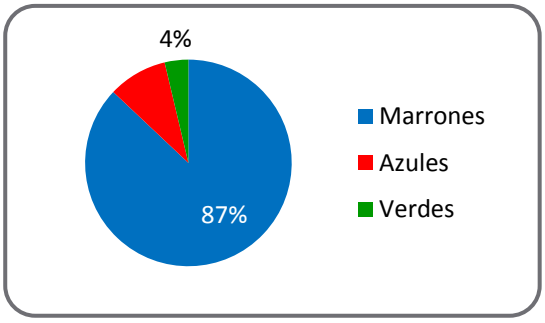


Figura 3.7: color de ojos

- **Barba:** la diferencia entre las personas que sí tienen barba a las que no, supera el 80%. Esto se debe al menor número de hombres. En caso contrario, la cifra seguramente sea mayor.

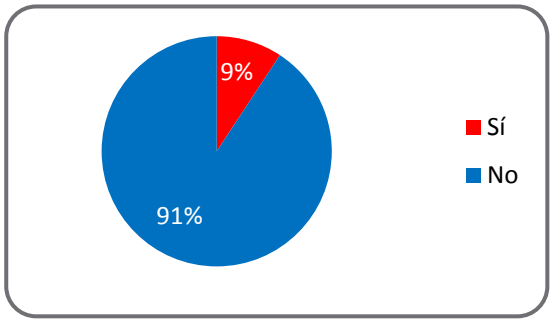


Figura 3.8: barba

- **Maquillajes:** al contrario que ocurre anteriormente, el porcentaje de personas maquilladas es mayor. Esto se debe a que las mujeres predominan sobre los hombres en la base de datos y son las que suelen maquillarse más habitualmente.

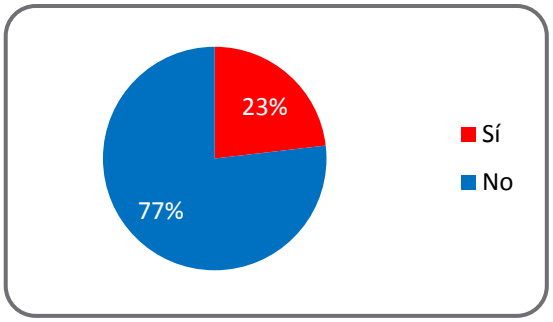


Figura 3.9: maquillajes

- **Gafas:** existe una diferencia del 54% entre los sujetos que llevan gafas con respecto a los que no las llevan.

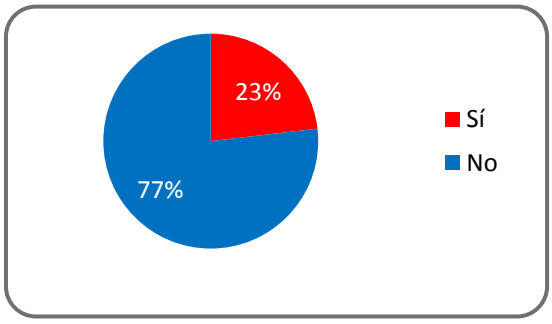


Figura 3.10: gafas

- **Pendientes:** en este apartado hay una mayor paridad, ya que en los últimos años tanto los hombres como las mujeres visten con pendientes.

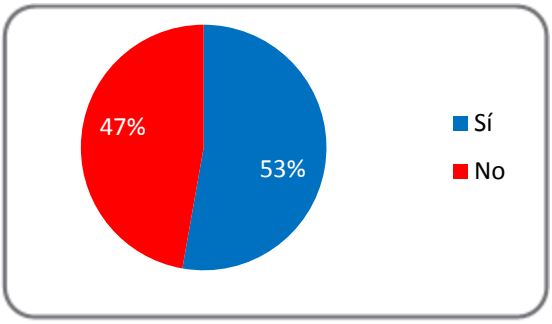


Figura 3.11: pendientes

- **Color de piel:** esta figura nos muestra la predominancia de las personas con un tono de piel blanco sobre las que tienen un tono de piel oscura.

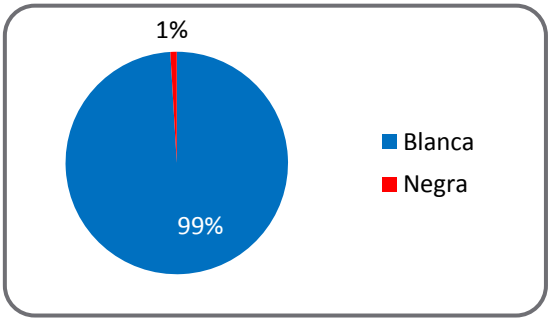


Figura 3.12: color de piel

En el Anexo C se muestra una tabla resumen con todos estos detalles especificados para cada vídeo de la colección.

3.6 CONCLUSIONES

En esta sección se exponen las conclusiones a las que el equipo de investigación ha llegado una vez terminada la base de datos AV-UC3M. A continuación se evalúa si los resultados obtenidos cumplen con los objetivos fijados con anterioridad.

El número de sujetos que participan en la colección de vídeos asciende a 108. Esta cifra cumple con creces con las metas establecidas al principio de esta fase del proyecto. En cuanto al género de los mismos, a pesar de que no se ha podido establecer una paridad completa entre hombres y mujeres, las diferencias no son excesivamente relevantes. Además, mediante la autorización tanto de complementos como de maquillajes en los participantes se consigue una fiel representación del mundo real.

En cuanto a la utilidad de esta colección de vídeos en otras áreas de investigación, sus posibilidades son muy variadas. Por ejemplo, se puede utilizar en campos tan diversos como la biometría para el reconocimiento facial de personas, hasta la robótica con el reconocimiento de expresiones faciales.

Otra de las conclusiones a las que se ha llegado tiene que ver con los aspectos técnicos de la base de datos. Se ha conseguido una iluminación profesional gracias a la utilización de los equipos de la

Universidad Carlos III de Madrid. En cuanto al sonido, se graba en salas insonorizadas que evitan la adición de ruidos de fondo al discurso propio del interlocutor.

Con respecto al corpus, también se logran cumplir con los objetivos planteados. Posee diferentes partes: series de números, nombres, fragmentos de libros y lenguaje real. Además se utiliza el castellano como idioma principal.

En conclusión, por todos los motivos expuestos con anterioridad se puede afirmar que esta base de datos desempeña las metas que se establecieron al principio del proyecto de manera adecuada y, en algunos casos como en el número de participantes, las supera.

4 SISTEMA SIM-RC

En este capítulo se realiza una descripción general del sistema basado en reconocimiento de color. En primer lugar, se describen los objetivos que se tienen que cumplir, a continuación se detallan sus características y por último las pruebas a las que se le ha sometido.

4.1 INTRODUCCIÓN

En la actualidad, la comunicación entre los seres humanos y los dispositivos electrónicos está mejorando cada día gracias a la investigación en el campo de los reconocedores del habla, donde se están añadiendo nuevas funcionalidades para mejorar su rendimiento [Ahmad et al, 2008].

Las características gráficas extraídas de la imagen del interlocutor están basadas tanto en la luz, como en los diferentes colores que forman el rostro. Gracias a estos parámetros, el sistema SIM-RC es capaz de distinguir entre las partes más importantes de la cara.

La clasificación de los elementos que forman el rostro ofrece la oportunidad de extraer información importante de cada uno de ellos. Por ejemplo, la posición de los ojos o el movimiento de los labios. Con esta última característica se realizará una aproximación al lenguaje natural, aportando nueva información al reconocedor.

4.2 OBJETIVOS

En esta sección se exponen los objetivos establecidos para poder llevar a cabo la aplicación SIM-RC. Son los siguientes:

- **Visualización:** el sistema debe reproducir los vídeos de entrada para su visualización por parte del investigador. Mediante esta exposición se evalúa si el reconocimiento facial llevado a cabo por la aplicación se realiza correctamente.
- **Separación de frames:** la aplicación debe separar todos los frames pertenecientes al vídeo de entrada, procesar la información y volver a unirlos para su reproducción.
- **Características faciales:** la aplicación debe ser capaz de mostrar las distancias tanto horizontales como verticales que existen entre los extremos de la boca. Además, debe obtener información de todos los píxeles de la boca que se consideren relevantes para la investigación.
- **Exposición de características:** SIM-RC debe informar al usuario de la distancia, tanto vertical como horizontal, existente entre los labios superior e inferior, mediante gráficos superpuestos en la imagen.

4.3 IMPLEMENTACIÓN

El lenguaje de programación utilizado para realizar la aplicación ha sido Simulink. Es un entorno de programación visual, que funciona sobre Matlab aunque posee un nivel de abstracción más elevado.

Simulink es una herramienta de simulación de modelos o sistemas, con cierto grado de abstracción de los fenómenos físicos involucrados en los mismos. Se hace hincapié en el análisis de sucesos, a través de la concepción de sistemas.

Se emplea arduamente en Ingeniería Electrónica en temas relacionados con el procesamiento digital de señales (DSP), involucrando temas específicos de ingeniería biomédica, telecomunicaciones, entre otros. También es muy utilizado en Ingeniería de Control y Robótica. En este caso se ha utilizado el módulo de procesamiento de imágenes y vídeos, ya que permite adquirir varias imágenes dentro de un mismo vídeo o modificar vídeos en tiempo real.

4.4 CARACTERÍSTICAS

En esta sección se exponen las fases que se efectúan a lo largo de todo el proceso de ejecución de la aplicación. Al mismo tiempo se describen de forma detallada sus características técnicas. Las etapas principales son:

- **Entrada:** en ella se reproduce el vídeo de entrada a la aplicación. Además se realiza la separación de cada uno de los frames del vídeo.
- **Discriminación zonal:** en esta fase la aplicación discrimina las partes de la cara para extraer la longitud, tanto vertical como horizontal, que existe entre los extremos de la boca.
- **Visualización de resultados:** en esta etapa se unifican de nuevo todos los frames del vídeo y se muestran las distancias que existen entre los extremos bucales.

En el siguiente esquema se muestra la relación existente entre las distintas fases que forman la aplicación:

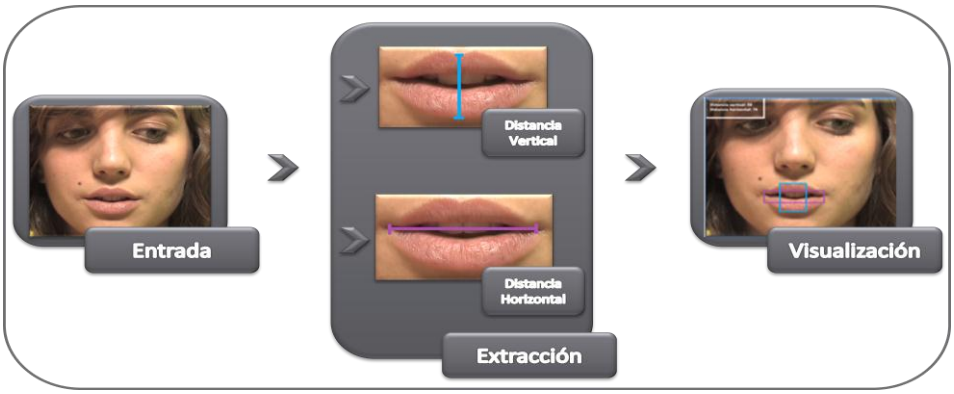


Figura 4.1: fases de la aplicación

A continuación se describen exhaustivamente todas las etapas que forman la aplicación SIM-RC. El objetivo es ofrecer una visión específica de todas las fases que la componen.

4.4.1 ENTRADA

En esta primera etapa se introducen los datos de entrada a la aplicación. En este caso se trata de una grabación de vídeo en la que un sujeto habla mirando a una cámara. Se extrae de la base de datos creada para este proyecto.

Se trata de un vídeo en color en formato AVI (Audio Video Interleave), con un formato de color RGB. Se elige este formato de vídeo debido a que el proyecto se realizará sobre plataforma Windows y es el estándar utilizado por este sistema operativo.

En cuanto a los datos de entrada, únicamente se van a extraer los datos de vídeo. Los datos de audio en el caso de la lectura de labios no son relevantes. Dichos datos se utilizan más adelante para incorporarlos al reconocedor del habla.

Figura 4.2: imágenes de muestra

4.4.2 EXTRACCIÓN

Esta etapa recibe los datos de vídeo de la etapa anterior. Éstos son procesados para extraer las distancias tanto verticales como horizontales de separación entre los labios.

El primer paso es la separación del vídeo en los tres colores que lo forman: rojo, verde y azul. El resultado es el que se ve en la figura 4.3, dónde se aprecia que son imágenes en blanco y negro. Esto se debe a la extracción de la intensidad del color de cada píxel de la imagen, no al color de la misma. Cuanto mayor es el nivel de color, más claro es el tono de la imagen final y al contrario.

Figura 4.3: extracción de color

A continuación se llevan a cabo dos fases dentro de este apartado: extracción de distancia vertical y extracción de distancia horizontal.

4.4.2.1 EXTRACCIÓN DISTANCIA VERTICAL

En la fase de extracción de la distancia vertical labial se efectúa una transformación de los espacios de color. Un espacio de color es un modelo matemático abstracto que describe la forma en la que los colores pueden representarse como tuplas de números, normalmente como tres o cuatro valores, como puede ser el RGB. Cada color de píxel de la imagen posee estas tres características en mayor o menor grado. En concreto el rango varía desde el 0 para la ausencia de color hasta el 255 que representa el máximo.

En este caso, se transforma el espacio de color RGB original del vídeo, al espacio de color HSV (Matiz, Saturación, Valor), ya que según [Vezhnevets et al, 2003] la transformación de RGB a HSV puede formar una muy buena opción para los métodos de detección de piel, ya que a la imagen resultante no le afecta ni una alta intensidad en las luces blancas, ni la orientación de la que proviene la luz.

Figura 4.4: transformación HSV

A continuación se binarizan (transformación de píxeles a blanco y negro) las imágenes resultantes mediante un filtrado de intensidad. Este filtro colorea en blanco los píxeles cuya intensidad sea mayor a un umbral establecido y en caso contrario, los píxeles se colorean en negro.

La binarización se realiza en los siguientes espacios de color pertenecientes al HSV:

- **Matiz (H):** se extrae la información del lugar que ocupa la boca, tanto el contorno labial como el interior de la misma, eliminando casi en su totalidad el resto de partes de la cara.
- **Valor (V):** se elimina casi por completo los píxeles correspondientes al pelo. Únicamente extrae información de los píxeles con color de piel.

En la siguiente figura se puede apreciar el resultado de la binarización de estos espacios de color:

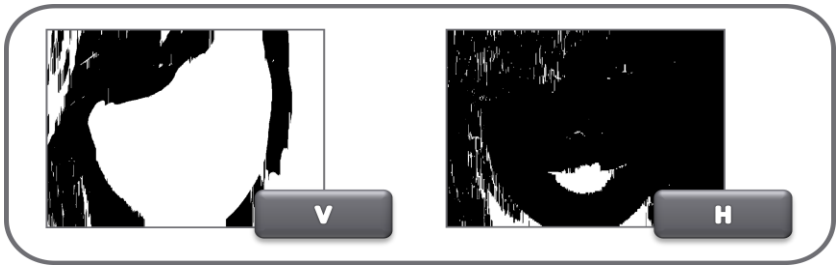


Figura 4.5: transformación HSV

A continuación se realiza una comparación de píxeles de todas las imágenes. El objetivo es extraer en un único frame, los píxeles que se han considerado relevantes en todos los espacios de color anteriores.

El método de extracción es el siguiente: si un píxel se considera importante en todos los frames extraídos con anterioridad, es relevante en la imagen final. Esto conlleva que dicho píxel se pinte de blanco. En caso contrario no se tendrá en cuenta, por lo que el píxel final estará pintado en negro. En la figura 4.9 se muestra este método de extracción.

En la siguiente figura se muestra la imagen resultante. Como se puede apreciar, la boca queda en blanco y el resto de la imagen queda coloreada en negro casi en su totalidad. Mediante este método se permite establecer la región de interés bucal buscada.

Figura 4.6: imagen binarizada

A pesar de que la mayoría de partes coloreadas de blanco corresponden a la boca, existen otras que pueden llevar a confusión. Por eso se ha implementado un analizador de píxeles que se encarga de discriminar los grupos de píxeles más relevantes.

El funcionamiento se basa en fijar una cantidad mínima de píxeles que cumplan con un umbral establecido por el investigador. Si el grupo es menor al creado con anterioridad se descarta. El resultado final consiste en la extracción del conjunto de píxeles de mayor tamaño, en este caso, los correspondientes a la boca.

El resultado de todo este proceso se divide en dos: extracción de la región de interés y longitud vertical de la boca. Gracias a estos datos en la fase de visualización se podrán mostrar gráficamente tanto la zona como la longitud de la misma en el vídeo original.

4.4.2.2 EXTRACCIÓN DISTANCIA HORIZONTAL

En esta etapa, además de la transformación de RGB en HSV, se ha realizado una nueva transformación al espacio de color YCbCr, ya que según [Vezhnevets et al, 2003] éste es otra de las opciones más populares en la detección de la piel. Representa el color de la imagen con dos componentes completamente independientes: luminancia (Y) y crominancia (Cb y Cr). Gracias a ellos se elimina la redundancia de las imágenes en RGB.

Figura 4.7: espacios de color YCbCr

El siguiente paso es la binarización de los frames resultantes de los cambios en el espacio de color. Para ello se utiliza el mismo método que en el apartado anterior. Se realiza un filtrado por nivel de intensidad del color en las imágenes.

Los espacios de color utilizados en esta etapa son los siguientes:

- **Luminancia Y:** se eliminan todos los píxeles correspondientes a la piel del interlocutor. Sólo escoge los que tienen tonos más oscuros.
- **Matiz HSV:** extrae la totalidad de la boca incluyendo labios, dientes, y lengua.
- **Valor HSV:** elimina todos los píxeles que pertenecen al pelo.
- **Región de interés (ROI):** con este filtro se eliminan todos los píxeles que no se encuentran en la parte inferior de la imagen. Se da por hecho que la boca siempre estará en la parte inferior de la misma.

En la siguiente figura se puede ver el resultado de la binarización de los diferentes espacios de color:

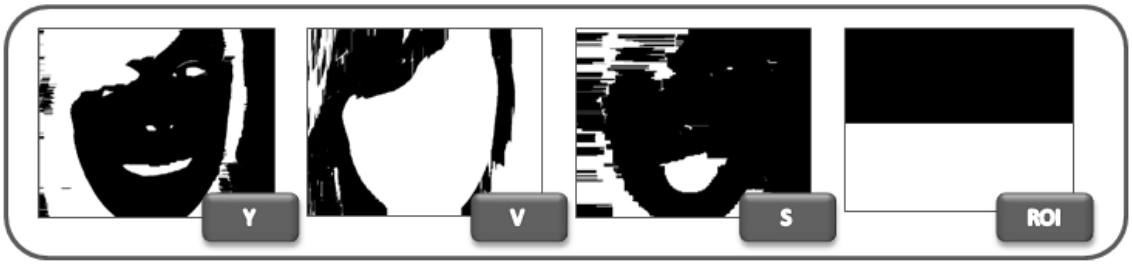


Figura 4.8: binarización

A continuación se realiza una comparación de píxeles en todas las imágenes como en el apartado anterior. En la siguiente figura se muestra un esquema de su funcionamiento:

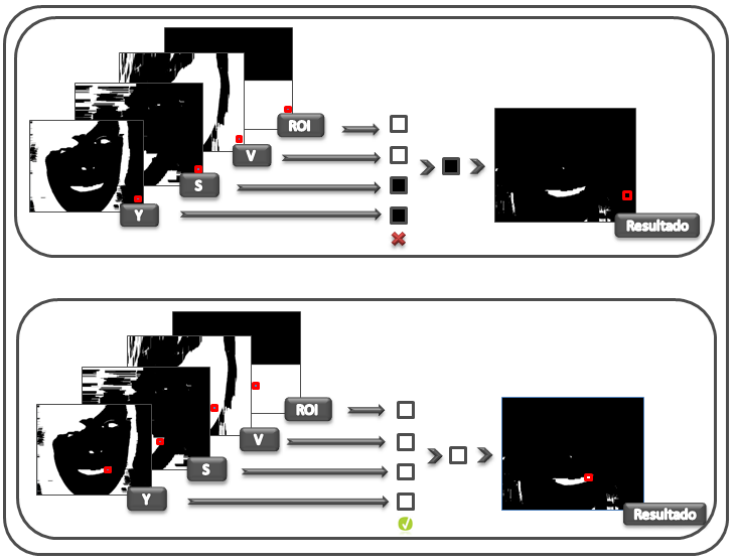


Figura 4.9: combinación frames

A continuación se procede exactamente igual que en el apartado anterior: se realiza un análisis de los grupos de píxeles que forman la imagen para extraer el grupo más grande correspondiente a la boca.

Por último se obtienen dos parámetros: extracción de la región de interés y longitud horizontal de la boca. Gracias a estos datos, en la siguiente fase se muestra el lugar que se corresponde con la boca y su longitud.

4.4.3 VISUALIZACIÓN

El objetivo de esta fase es mostrar el resultado de la extracción de características para realizar las comprobaciones necesarias.

Esta fase se desarrolla para poder llevar a cabo la fase de pruebas. Se divide a su vez en cuatro apartados diferentes: parte interior de la boca, parte exterior, distancias y combinación final.

4.4.3.1 INTERIOR

Los datos de entrada que se necesitan para visualizar la parte interior de la boca son:

- **Coordenadas**: se extraen en la etapa anterior y coinciden con la posición del interior de la boca.
- **Frame original**: se extrae del vídeo de la fase de entrada de datos.

El siguiente paso es la superposición de las coordenadas en la imagen original. Para ello se extraen las posiciones que ocupan los píxeles relevantes y se recuadran con un rectángulo que los delimite.

El rectángulo se mostrará frame a frame delimitando un contorno aproximado de la parte interna de la boca. En la siguiente figura se puede observar el resultado final:

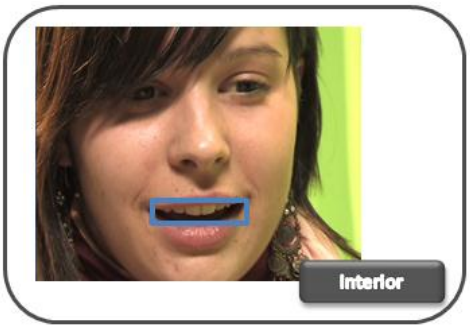


Figura 4.10: visualización interior

4.4.3.2 EXTERIOR

Esta etapa se realiza exactamente igual que la anterior. Gracias a la combinación de los datos de coordenadas de la parte exterior de la boca con el frame de vídeo, se visualiza una aproximación de la misma. En la siguiente figura se muestra un ejemplo de visualización:

Figura 4.11: visualización exterior

4.4.3.3 DISTANCIAS

En este apartado se calculan las longitudes que separan los labios. Más adelante se utiliza esta información para realizar el reconocimiento labial en combinación con la información extraída del reconocedor de voz.

La distancia que separa ambos labios se extrae a partir de las coordenadas de las que se compone cada frame de la imagen. Se tiene en cuenta la fila y la columna en la que está situado el primer píxel y el último de los que forman los labios.

Con esta información se realiza una comparación entre las posiciones de los píxeles. A continuación se muestra cómo se calcula:

$$fila_{px\,min} - fila_{px\,max} = dist_{vert}$$

$$col_{px\,min} - col_{px\,max} = dist_{horiz}$$

Figura 4.12: distancias

Con estas ecuaciones se extraen las distancias a nivel de píxel, existentes tanto vertical como horizontalmente. A continuación se muestran en todos los frames que aparecen en el vídeo. Aparece recuadrado en la parte superior izquierda de la imagen, de esta manera:

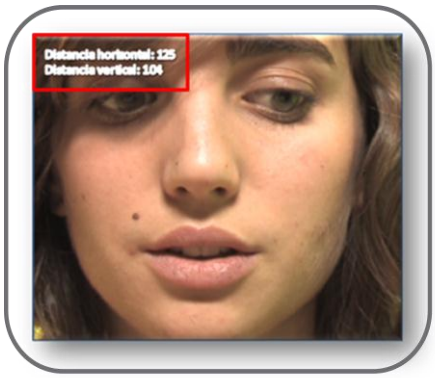


Figura 4.13: visualización distancias

4.4.3.4 COMBINACIÓN FINAL

En esta última fase se realiza una superposición de todas las visualizaciones descritas anteriormente. Para conseguirlo, se combinan los frames que resaltan las posiciones tanto exteriores como interiores de los labios. A continuación se vuelve a combinar la imagen resultante con el frame que muestra las distancias.

El resultado final se puede apreciar en la siguiente figura:

Figura 4.14: visualización final

4.5 PRUEBAS

En este capítulo se especifican las fases de las que se compone el desarrollo de las pruebas del SIM-RC. En primer lugar se describen las baterías de pruebas a las que se somete al sistema y, por último, se realiza una evaluación de los resultados obtenidos en las mismas.

4.5.1 INTRODUCCIÓN

En este apartado se comprueba el correcto funcionamiento de la aplicación mostrando los resultados obtenidos. Se exponen las tasas de error y de acierto con las que se lleva a cabo la evaluación.

Para realizar las pruebas necesarias se utiliza la base de datos AV-UC3M, elaborada a lo largo del proyecto. Contiene la suficiente variedad de participantes como para poder realizar los test que se consideran necesarios.

Además se exponen los métodos de evaluación utilizados y los factores que se han tenido en cuenta a la hora de evaluar el sistema.

4.5.2 BATERÍA DE PRUEBAS

La batería de pruebas utilizada se compone de un número limitado de frames pertenecientes a vídeos de entrada de la base de datos AV-UC3M. Estos frames se han seleccionado en función de sus características, ya que reúnen la mayoría de tipos de fotogramas de los que está compuesta la base de datos.

Se ha seleccionado a cinco sujetos al azar como muestra representativa de todo el conjunto de personas que forman la base de datos. Tiene participantes de diferentes edades, con género tanto masculino como femenino y con varios tipos de complementos.

En cuanto a las muestras, se realizan pruebas con cien frames de cada sujeto. En concreto, se seleccionan cincuenta para la extracción horizontal de características y otras cincuenta para la extracción vertical.

4.5.3 EVALUACIÓN

El objetivo de la evaluación es comprobar si los elementos que mide la aplicación han sido los adecuados. Gracias a ellos se realiza una evaluación fiable de los resultados obtenidos.

Se ha tenido en cuenta tanto la distancia vertical como horizontal de la boca. En este caso se comprueba que el recuadro que utiliza la aplicación para la medición se encuentra en el lugar adecuado, calculando la distancia de manera satisfactoria.

En la siguiente figura se muestra un cuadro resumen con los resultados extraídos de la comprobación de la aplicación:

Vídeo	Distancia vertical				Distancia horizontal			
	Aciertos	Fallos	%Acierto	%Error	Aciertos	Fallos	% Acierto	%Error
Vídeo 1	49	1	98%	2%	41	9	82%	8%
Vídeo 2	47	3	94%	6%	33	17	66%	34%
Vídeo 3	28	22	56%	44%	30	20	60%	40%
Vídeo 4	25	25	50%	50%	32	23	64%	36%
Vídeo 5	31	19	62%	38%	23	32	36%	64%
Totales	36	14	72%	28%	31,8	20,2	62%	36%

Tabla 4.1: tasas de acierto y error

Las conclusiones que se pueden obtener son las siguientes:

- **Tasas de acierto:** para la distancia bucal vertical se han obtenido unas tasas de acierto de un 72%. En cuanto a la distancia horizontal, se obtiene un porcentaje del 62%. Aunque la primera tiene una tasa mayor, ambas poseen unos porcentajes aceptables.

- **Tasas de error:** tanto la distancia vertical como la horizontal poseen unas tasas de error no muy elevadas. En función de estos datos podemos extraer unas conclusiones positivas con respecto a las pruebas realizadas.

Según la información extraída de las pruebas a las que se ha sometido a la aplicación SIM-RC, se puede llegar a la conclusión de que realiza un reconocimiento de características lo suficientemente bueno como para poder seguir investigando en este campo. A pesar de que aún existe un margen de mejora, se puede afirmar que la aplicación cumple con éxito los objetivos perseguidos antes de su realización.

5 SISTEMA MAT-RP

En este capítulo se realiza una descripción general del sistema MAT-RP. En primer lugar se describen los objetivos fijados a priori, a continuación se detallan sus principales características y por último se exponen los resultados obtenidos de la fase de pruebas.

5.1 INTRODUCCIÓN

Desde que Sutherland [Sutherland 1963] creó el primer programa de dibujo por computadora en 1963, la comunicación persona-máquina ha crecido exponencialmente. Mediante este software, se amplía el intercambio de información con las computadoras gracias a elementos como el ratón del computador, las pantallas con mapas de bits, sistema de ventanas, punteros para clicar o como en este caso, sistemas de reconocimiento del habla.

La mejora de estos dispositivos de reconocimiento mediante la adición de características visuales permite aumentar sus capacidades de forma notable. Para poder agregar este tipo de información es necesario extraer diferentes características faciales de forma automática.

La extracción de información visual se realiza de múltiples formas: reconocimiento por color, por posición, tonos de piel, iluminación, etc. En este caso, se utiliza el método de extracción mediante sistemas clasificadores de características.

5.2 OBJETIVOS

El principal objetivo propuesto por el equipo de investigación de este proyecto es la discriminación de zonas faciales, de manera que se consigan unos resultados óptimos. En caso de conseguirlo, se llevará a cabo el reconocimiento automático del habla con unas mayores garantías de éxito.

La siguiente meta consiste en la extracción de características mediante sistemas clasificadores de características. Con su utilización se permite la obtención de unos resultados fiables en un menor tiempo de procesamiento.

Por último, a la hora de realizar la clasificación automática el sistema debe ser capaz de distinguir todas las zonas que forman el rostro humano. Por ello, se crea una herramienta que muestra el modo de diferenciar las regiones automáticamente y así poder evaluarlo de forma adecuada.

5.3 CARACTERÍSTICAS

En esta sección se detallan las diferentes funcionalidades que ofrece este sistema. En una primera fase se muestran los diversos métodos, tanto de extracción de frames como de características. A continuación se enumeran las fases en las que se divide la aplicación MAT-RP y, por último, se detallan los resultados obtenidos con sus conclusiones correspondientes.

5.3.1 EXTRACCIÓN DE FRAMES

Los frames utilizados para realizar la visualización de zonas faciales se obtienen de la base de datos AV-UC3M creada durante la realización de este proyecto. De cada vídeo se extraen los fotogramas suficientes para efectuar las pruebas correspondientes.

Para llevar a cabo la extracción se estudian varios formatos de imagen. En un primer momento surge la duda entre dos tipos: jpeg y bmp. El primero tiene la ventaja de que sus archivos ocupan menos espacio en disco, aunque su calidad es más baja debido a la compresión de los datos que forman el frame. Finalmente, se escoge el formato bmp, ya que aunque los archivos son más grandes, tienen una mayor calidad.

5.3.2 DISCRIMINACIÓN DE ZONAS FACIALES

La primera fase de esta aplicación consiste en la introducción de datos de características al sistema. El objetivo es obtener el mayor número de ellas, con la mayor precisión posible.

Las partes de la cara relevantes para el proyecto, y de las que se va a extraer información, son las siguientes:

- **Ojo izquierdo:** se necesita la posición del ojo para que, en caso de que la imagen no tenga la inclinación adecuada, pueda corregirse utilizando dicha posición como referencia.
- **Ojo derecho:** cumple la misma función que el otro ojo. Se realiza una separación de los mismos, para facilitar su clasificación.
- **Lengua:** es una parte fundamental en el desarrollo del lenguaje. Es necesaria para pronunciar ciertos tipos de fonemas y, por tanto, relevante en el reconocimiento del habla.
- **Dientes:** su aparición en la imagen puede ofrecer indicios sobre el tipo de fonema que está pronunciando el interlocutor.
- **Labios:** es la parte más importante y sobre la que se basa el proyecto. Con la información que nos proporcionan se puede realizar el reconocimiento.
- **Otros:** estas características se utilizan para contrastar con el resto y así poder distinguir las partes relevantes de las que no lo son.

Las características que se deben extraer se deciden en función de la posición en la que se encuentran y del color de cada una de las zonas:

- **Posición:** se recopila la posición exacta de cada píxel que forma una región de interés. En concreto serán la fila y la columna dentro de cada frame.
- **Color:** se almacenan los tres colores RGB que forman el píxel.

A continuación se detallan los métodos utilizados para extraer la información.

5.3.3 EXTRACCIÓN MANUAL

En primer lugar se realiza una extracción manual de las características faciales. Para este fin se desarrolla un software capaz de realizar un listado con los siguientes elementos de cada píxel: fila y columna que ocupan dentro de la imagen y los colores que forman el formato RGB: rojo, verde y azul. En la siguiente imagen se muestra un ejemplo con los vectores de características resultantes:

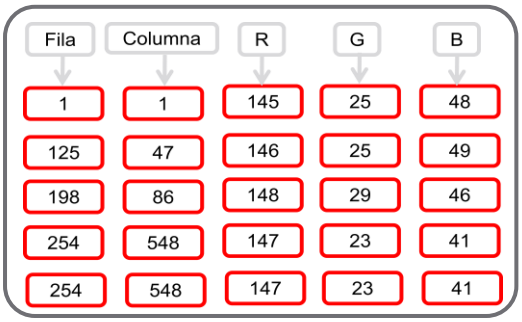


Figura 5.1: características

El software está desarrollado en lenguaje Matlab. Está compuesto de un editor de imágenes que extrae información de los píxeles de un frame desde el vídeo de entrada. Se marcan las posiciones deseadas por el usuario gracias a un puntero situado sobre la imagen. Además se permite realizar este procedimiento sobre imágenes que pertenecen a diferentes sujetos.

El usuario se posiciona en la parte de la imagen sobre la que desea extraer información y hace clic sobre la misma con el ratón. Automáticamente se almacenan los datos referentes a la posición y al color del píxel seleccionado, en referencia a su imagen.

Figura 5.2: ejemplo extracción

Esta operación se realiza tantas veces como el usuario estime oportuno dentro de una misma zona. Para seguir extrayendo características de otra región, se repite el mismo procedimiento tantas veces como zonas haya. El resultado final son seis listados con la información referente a cada zona: ojo derecho, ojo izquierdo, labios, dientes, lengua y un conjunto de píxeles de otras zonas: piel, pelo, fondo, etc.

Esta opción de extracción de características se utiliza durante las fases preliminares del proyecto. Más adelante se decide integrar una fase de extracción automática dentro del sistema mediante la edición de imágenes. Sus características se muestran en el siguiente capítulo.

5.4 APLICACIÓN MAT-RP

En este capítulo se describen las características principales de la aplicación MAR-RP. En primer lugar, se detallan los elementos que forman la entrada de datos a la aplicación. A continuación, se normalizan estos datos y, por último, se introducen en los clasificadores obteniendo los resultados correspondientes.

5.4.1 VECTOR DE CARACTERÍSTICAS

Estas características se van a utilizar como modelo de datos de entrada a los clasificadores. Para poder crear el vector es necesario tener dos tipos de datos de cada región: datos genuinos y datos falsos.

5.4.2 DATOS GENUINOS

Para obtener datos genuinos se lleva a cabo una extracción automática de características. El objetivo es obtener una mayor información sin necesidad de recurrir a la forma manual. De esta manera se consiguen dos objetivos: más características faciales y un menor tiempo de procesamiento.

El primer paso es el retoque de los frames utilizando un editor de imágenes. A cada zona de la cara se le asigna un color determinado que no se encuentre en ningún otro lugar de la imagen. Los colores que se han seleccionado, siguiendo el código de color RGB, son los siguientes:

Zona	Rojo	Verde	Azul
Ojo derecho	**255**	**0**	**0**
Ojo izquierdo	**218**	**0**	**118**
Labios	**12**	**240**	**221**
Dientes	**15**	**9**	**221**
Lengua	**251**	**255**	**0**

Tabla 5.1: códigos de color

A continuación se realiza este procedimiento en una serie de imágenes que sirven como modelo de características. En la siguiente figura se puede observar el resultado obtenido al editar los frames:

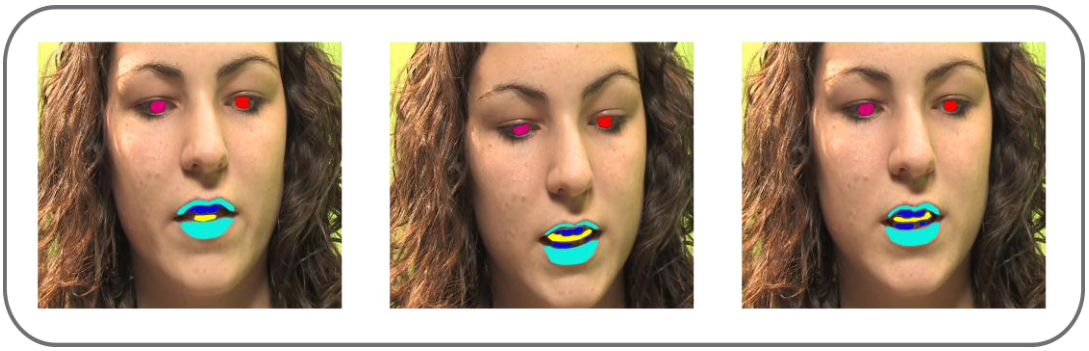


Figura 5.3: edición de imagen

El siguiente paso consiste en la comparación automática de los píxeles en busca de aquellos colores que coincidan con los que se han retocado. Se realiza un recorrido por todos los píxeles de la imagen buscando unas coordenadas RGB exactamente iguales a las expuestas en la tabla anterior.

El procedimiento consiste en acceder a todas los píxeles que forman la imagen. Cada uno de ellos están formados por tres componentes de color: rojo, verde y azul, cada uno con un valor asignado. Para ilustrar cómo están almacenados se ha creado un esquema representado en la siguiente figura:

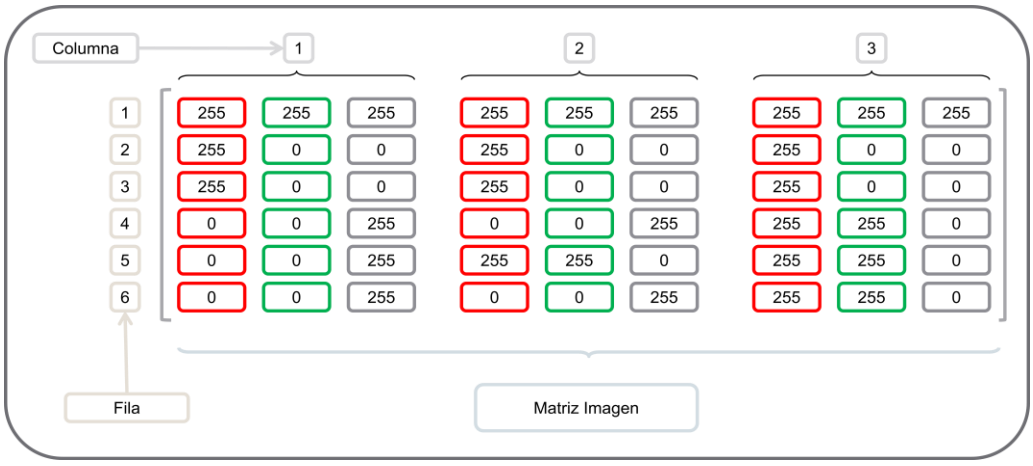


Figura 5.4: formato de imagen

El siguiente paso consiste en comparar uno a uno todos los píxeles de la imagen retocada hasta encontrar una coincidencia en los valores de color, con la imagen primaria. Cuando esto ocurre, se almacenan sus coordenadas espaciales (posición vertical y horizontal de la imagen) y la región de la cara a la que pertenecen los mismos. La siguiente figura explica el funcionamiento de lo expuesto en este párrafo:

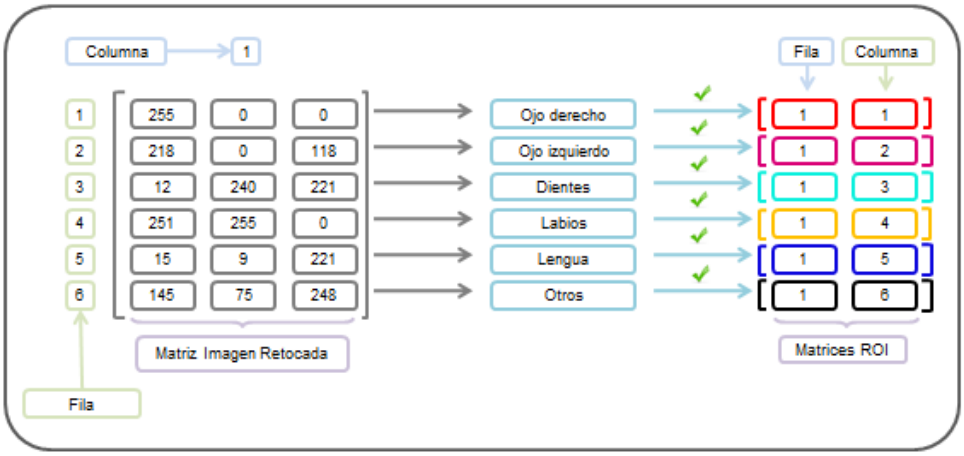


Figura 5.5: comparación de píxeles

La comparación se extiende a todos los píxeles de la imagen. Se obtienen seis listados: cinco con coordenadas de zonas de interés y otro con las coordenadas que no se corresponden a ninguna de ellas. Teniendo en cuenta que los frames tienen unas dimensiones de 720x576, el número de píxeles asciende a 414.720.

A continuación se vuelve a hacer un recorrido por todos los píxeles de la imagen original accediendo a las posiciones que se han registrado con anterioridad. De esta forma se obtienen las características RGB originales de cada píxel del frame, así como las características necesarias de color y posición de todas las regiones de interés. En la siguiente figura se muestra el mecanismo de extracción:

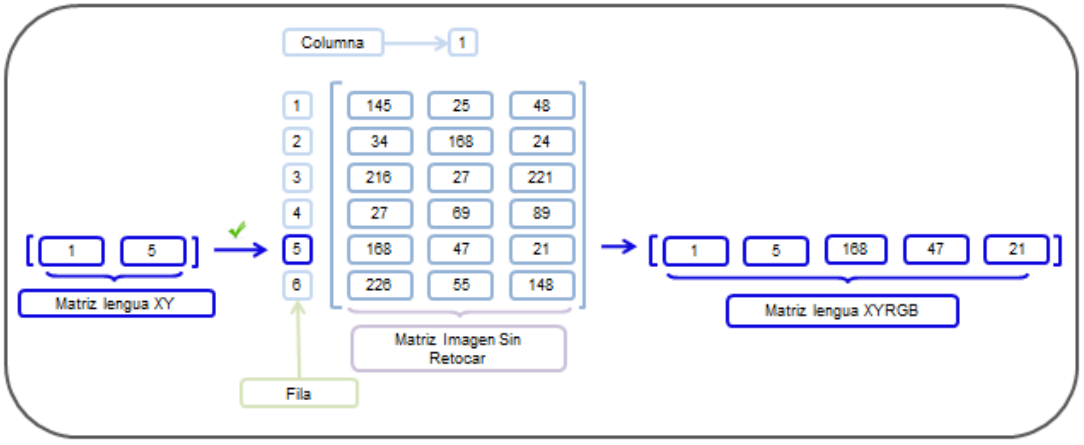


Figura 5.6: extracción RGBXY

5.4.2.1 DATOS FALSOS

El modelo de datos falsos se crea para realizar el entrenamiento de los clasificadores utilizados para el reconocimiento de zonas. Su función es la de “enseñarle” el tipo de dato que no pertenece a una zona en particular.

Para poder crear este modelo se necesitan características pertenecientes a las otras partes de la cara. Se seleccionan de manera aleatoria tantos datos como el usuario estime oportuno.

En la siguiente figura se puede ver un ejemplo del tipo de datos que forman este modelo para la región del ojo derecho:

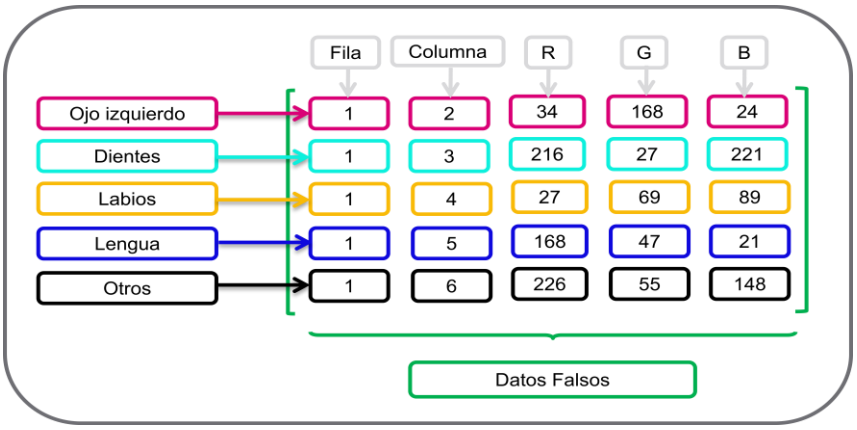


Figura 5.7: datos falsos ojo derecho

5.4.2.2 FASE FINAL

El último paso consiste en la unión de los dos modelos anteriores en uno solo, añadiendo una nueva característica que distinga los datos verdaderos de los falsos. Este dato permite al clasificador diferenciar qué tipo de píxel se corresponde con cada región de interés.

El modelo final está compuesto por una matriz con las siguientes características: coordenadas de color RGB, coordenadas de posición y dato de confianza. Esta característica informa al clasificador sobre la correspondencia de un píxel a una zona facial en concreto.

Si los datos no pertenecen a una región de la cara el valor de la confianza es menos uno y en caso contrario es de uno. Por ejemplo, en la figura siguiente se puede apreciar como en un modelo que recoge los datos de los labios, las dos primeras filas corresponden a píxeles de esa zona y las dos últimas son de otras áreas como los dientes y la lengua.

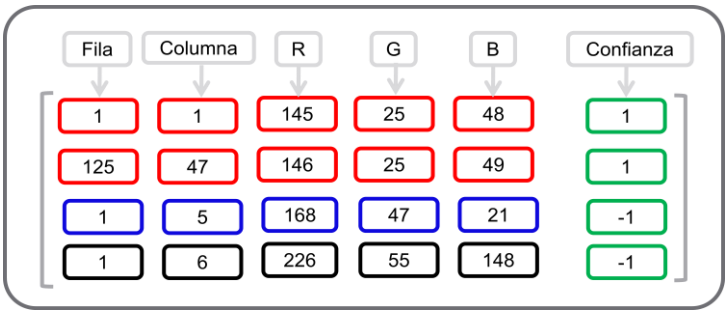


Figura 5.8: vector de características

5.4.3 NORMALIZACIÓN

En el momento de utilizar los datos de entrada extraídos de los frames, se decide homogeneizarlos, ya que se obtuvieron de diferentes ámbitos (color y posición) pudiendo dar lugar a errores. Con esta normalización se consigue transformar los resultados (*scores)* de los sistemas individuales a un dominio común, antes de combinarse entre sí [Jain et. al. 2005].

Los tres tipos de normalización utilizados en este proyecto son la normalización *min-max,* la normalización *z-score* y la normalización *median.*

5.4.3.1 NORMALIZACIÓN *MIN-MAX*

La normalización *min-max* es la técnica más simple: los valores mínimos y máximos de los *scores* se desplazan a los valores 0 y 1, respectivamente., y todos los *scores* se transforman en el rango [0,1], de manera que la distribución original se mantiene (excepto para el factor de escala).

La ecuación que se utiliza se desarrolla a partir de una serie de scores $\{s_k\}, k = 1,2, \dots n$, los scores normalizados se obtendrán de la siguiente manera:

$$s_k' = \frac{s_k - min}{max - min}$$

Figura 5.9: normalización min-max

Este tipo de normalización es muy sensible a los valores atípicos en los datos utilizados para la estimación. Conserva la distribución original de *scores* excepto por un factor de escala y transforma todos los scores en un rango común [0,1].

5.4.3.2 NORMALIZACIÓN *Z-SCORE*

La normalización *z-score* es la técnica más utilizada. Consiste en transformar la distribución original de los *scores* en una distribución de media cero y variancia unitaria.

Se puede esperar un buen rendimiento de este método si se conoce a priori la media y la variación de *scores.* La normalización viene definida por la siguiente ecuación:

$$s_k' = \frac{s_k - \mu}{\sigma}$$

Figura 5.10: normalización z-score

Dónde μ es la media aritmética y σ es la desviación estándar del conjunto de datos. Sin embargo, tanto la media como la desviación son muy sensibles a los valores anómalos, haciendo perder robustez al sistema.

5.4.3.3 NORMALIZACIÓN *MEDIAN*

La normalización *median* es más robusta que las dos normalizaciones precedentes, ya que tanto la *mediana* como la *desviación absoluta de la mediana (MAD)* que la forman son muy poco sensibles a valores atípicos. La ecuación de la que se obtienen los valores, es la siguiente:

$$s_k' = \frac{s_k - mediana}{MAD}$$

Figura 5.11: normalización median

En el que MAD se define como: $MAD = mediana\ (|\ s_k - mediana|)$. Sin embargo, la media y la mediana tienen una menor eficiencia que la media y la desviación estándar utilizadas en el método anterior.

5.4.4 UMBRALES Y TIPOS DE ERROR

Al normalizar los datos de entrada al clasificador, se consigue homogeneizar las muestras. El siguiente paso es obtener el EER (Equal Error Rate), en el que se establece un umbral medio entre las tasas de falsa aceptación (False Acceptance Rate FAR) y la de falso rechazo (False Reject Rate FRR).

La tasa de falsa aceptación (FAR) es una estadística utilizada para medir el rendimiento del sistema durante la verificación. Mide el porcentaje de veces en el que un píxel es vinculado a una región a la cual no pertenece.

En cuanto a la tasa de falso rechazo (FRR) es otra estadística que mide el porcentaje de veces en el que un píxel no es vinculado a la región a la que pertenece.

En la siguiente figura se muestra una gráfica sobre los parámetros que forman el EER:

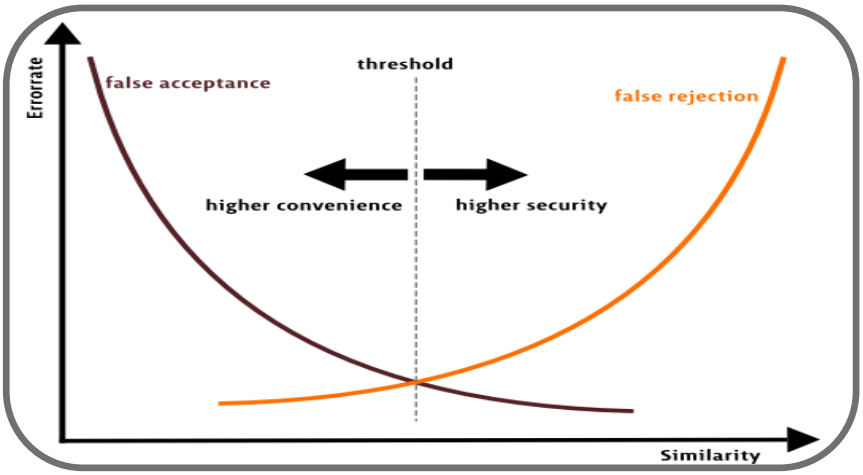


Figura 5.12: gráfica EER

Como se puede observar, si el umbral estuviera situado más a la izquierda, se obtendría un sistema altamente permisivo con los errores. En caso contrario, el sistema tendría una mayor tasa de falso rechazo aumentando la seguridad.

El objetivo deseado es hallar un umbral que consiga estar en un término medio entre seguridad y permisividad en los errores.

5.4.5 CLASIFICADORES

La aplicación da la posibilidad al usuario de elegir el tipo de clasificador a utilizar en la experimentación. Se puede elegir entre dos tipos: redes neuronales y máquinas de vectores de soporte.

5.4.5.1 TIPOS

Para llevar a cabo la clasificación de características visuales se han utilizado dos herramientas creadas para desarrollar proyectos de investigación. Las redes neuronales se han obtenido del kit de herramientas del programa Matlab. En cuanto a las máquinas de vectores soporte se ha utilizado un programa de libre distribución llamado “SVM Light Machine”.

5.4.5.1.1 MATLAB NEURAL NETWORK TOOLKIT

MATLAB es un lenguaje de alto desempeño diseñado para realizar cálculos técnicos. Integra el cálculo, la visualización y la programación en un ambiente fácil de utilizar, donde los problemas y las soluciones se expresan en una notación matemática. Permite resolver muchos problemas computacionales, específicamente aquellos que involucren vectores y matrices, en un tiempo mucho menor al requerido para escribir un programa en un lenguaje escalar no interactivo tal como C o Fortran. MATLAB se utiliza ampliamente en:

- Cálculos numéricos
- Desarrollo de algoritmos
- Modelado, simulación y prueba de prototipos
- Análisis de datos, exploración y visualización
- Realización de gráficos de datos con fines científicos o de ingeniería
- Desarrollo de aplicaciones que requieran de una interfaz gráfica de usuario (GUI, Graphical User Interface).

En el ámbito académico y de investigación, es la herramienta estándar para los cursos introductorios y avanzados de matemáticas, ingeniería e investigación. En la industria se utiliza para el análisis, investigación y desarrollo de nuevos productos tecnológicos. Su ventaja principal es el uso de familias de comandos de áreas específicas llamadas toolboxes.

Lo más importante para los usuarios de la herramienta es que los toolboxes le permiten aplicar la teoría relacionada a su ámbito de trabajo. Son grupos de comandos que resuelven problemas de áreas específicas de la ciencia y la ingeniería. Por ejemplo, existen toolboxes para las áreas de Procesamiento Digital de Señales, Sistemas de Control, Redes Neuronales, Lógica Difusa, etc.

Para llevar a cabo este proyecto se utiliza la funcionalidad de Neural Network Toolbox™. Contiene las herramientas necesarias para diseñar, implementar, visualizar y simular redes neuronales. Se utiliza para aplicaciones donde los análisis habituales harían difíciles o imposibles los reconocimientos de patrones y los sistemas no lineales de identificación y control. Esta herramienta soporta la mayoría de tipos de redes: radiales, dinámicas, mapas auto-organizativos, etc.

5.4.5.1.2 SVM LIGHT

Para la realización de esta fase se ha utilizado la herramienta SVM Light Machine [Joachims et. al. 1999]. Se trata de un software desarrollado por el departamento de informática de la universidad de Dormund.

Este desarrollo es una implementación de la máquina de soporte vectorial ideada por [Vapnik, 1995] para investigar sobre el problema de reconocimiento de patrones. Los algoritmos de optimización utilizados se describen en [Joachims, 2002], [Joachims, 1999]. Posee varios requisitos de escalabilidad de memoria y puede manejar los problemas formados por miles de vectores de soporte de manera eficiente.

El software proporciona métodos para evaluar el rendimiento de manera eficiente. Incluye varios métodos de estimación para las tasas de error. Entre ellas están las estimaciones de XiAlpha [Joachims, 2002, Joachims, 2000] que se pueden calcular sin costo computacional.

El código ha sido utilizado en una amplia gama de problemas, incluyendo la clasificación de texto [Joachims, 1999], [Joachims, 1998], tareas de reconocimiento de imagen, bioinformática y aplicaciones médicas. Esta aplicación proporciona una representación muy compacta y eficiente.

5.4.5.2 REDES NEURONALES

El sistema ofrece la posibilidad de realizar la clasificación de los datos de entrada mediante redes neuronales. Este tipo de redes fue propuesto por dos neurocientíficos en los años 50, llamados Warren McCulloch y Walter Pitts. En él se modelizaba una estructura y un funcionamiento simplificado de las neuronas del cerebro, considerándolas como dispositivos de varias entradas, una única salida y dos estados posibles: activo o inactivo [McCulloch 1943].

Su aprendizaje adaptativo, auto-organización, tolerancia a fallos, operación en tiempo real y fácil inserción, han hecho que su utilización se haya extendido en áreas como la biológica, financiera, industrial, medio ambiental, militar, salud, etc. [Hilera 1995]. Están funcionando en aplicaciones que incluyen identificación de procesos [González 98], detección de fallos en sistemas de control [Aldrich 1995], modelación de dinámicas no lineales [Meert 1998] [Wang 1998], control de sistemas no lineales [Bloch 1997] y optimización de procesos [Altissimi 1998].

5.4.5.2.1 NEURONAS

La primera funcionalidad que ofrece el sistema a la hora de llevar a cabo la clasificación mediante redes neuronales es la de definir el número de neuronas que tendrá cada capa en la red. Se considera una neurona como un elemento formal, módulo o unidad básica de la red que recibe información de otros módulos o del entorno; la integra, la computa y emite una única salida que se va a transmitir idéntica a múltiples neuronas posteriores [Wasserman 89].

El siguiente elemento importante dentro de las redes neuronales artificiales es el peso o fuerza sináptica. Cada neurona recibe multitud de entradas de forma simultánea, cada una con un peso que pondera la importancia de la misma con respecto a las demás. Cada peso se calcula siguiendo esta fórmula:

$$w_{ij} = Peso\ de\ la\ conexión\ entre\ la\ neurona\ j\ (que\ emite)\ y\ peso\ entre\ la\ neurona\ i\ (que\ recibe)$$

Figura 5.13: fuerza sináptica

Después de recibir todas las entradas con sus correspondientes pesos, la neurona emite una salida a partir de tres funciones que se aplican de forma consecutiva:

- **Función de propagación**: se trata del sumatorio de cada entrada multiplicada por su peso. Si el resultado es positivo se considera una conexión excitatoria, en caso contrario se denomina inhibitoria.

$$NET_i = \sum_{j=0}^{n-1} [W_{ij} * O_j]$$

Figura 5.14: función de propagación

- **Función de activación**: recibe el resultado de la función de propagación y calcula el estado de activación de la neurona correspondiente. Existen dos modelos de esta función: acotados y no acotados.

$$A_i(t) = FA(A_i(t-1), NET_i(t))$$

Figura 5.15: función de activación

- **Función de transferencia**: tiene como entrada la salida de la función de activación. Se encarga de acotar la salida de la neurona. Existen varios tipos de funciones: lineal, escalón, sigmoidea, gaussiana, etc.

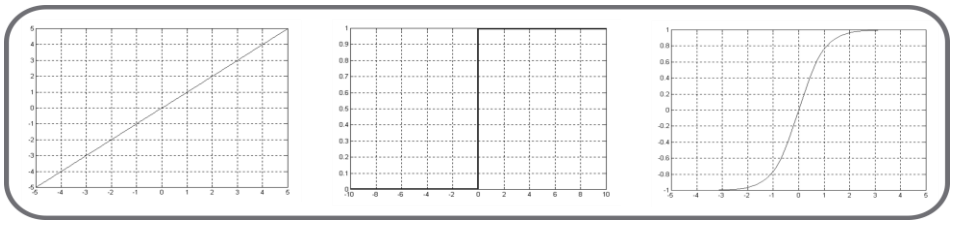


Figura 5.16: funciones de transferencia

En la siguiente figura se pueden observar de manera gráfica los elementos que componen una neurona:

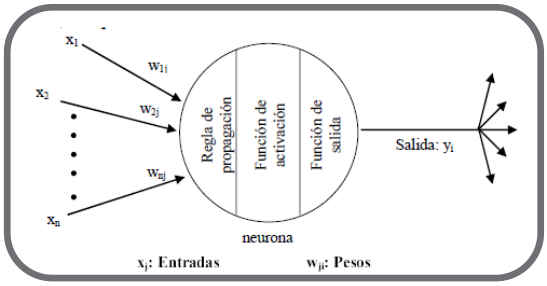


Figura 5.17: neurona

5.4.5.2.2 ARQUITECTURAS RNA

El usuario elige el número de capas con el que debe contar la red. Cada una de ellas agrupa a un conjunto de neuronas que reciben informaciones de las neuronas de la capa anterior y emiten salidas hacia las neuronas de la capa siguiente. Entre las neuronas de una misma capa no hay conexiones. Existen tres tipos de capas:

- **De entrada:** compuesta por neuronas que reciben datos desde fuera de la red neuronal.
- **Oculta:** es aquella capa que no se conecta con ningún elemento del entorno, puede a su vez conectarse con otras capas ocultas de similares características.
- **De salida:** establece los datos de salida de la red neuronal. Habitualmente está conectada con una capa oculta y con el entorno.

En cuanto a las redes, podemos diferenciar entre dos tipos:

- **Redes monocapa:** formadas por una sola capa de neuronas.

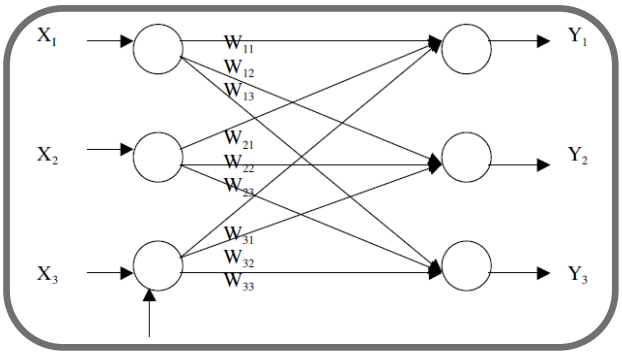


Figura 5.18: red monocapa

- **Redes multicapa:** formadas por más de una capa de neuronas.

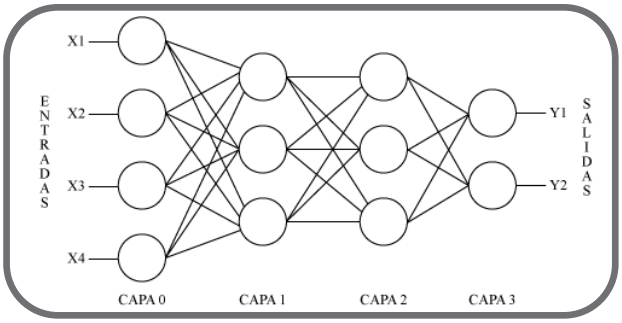


Figura 5.19: red multicapa

5.4.5.3 MÁQUINAS DE VECTORES SOPORTE SVM

Como segunda opción la aplicación ofrece al investigador otro tipo de clasificador denominado máquina de soporte vectorial (SVM). Esta técnica fue desarrollada por [Vapnik 1998] y se centra en lo que se conoce como Teoría del Aprendizaje Estadístico [González 2003]. Consiste en buscar una función apropiada que permita llevar a cabo una buena generalización, mediante una tarea de aprendizaje dada y un conjunto finito de datos.

La SVM mapea los datos de entrada a un espacio de características de una dimensión mayor y encuentra un hyperplano que los separe y maximice el margen m entre las clases [Betancourt 2005]. En la siguiente figura se ilustra el funcionamiento:

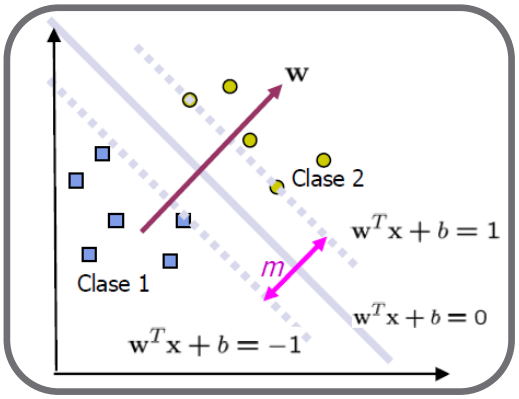


Figura 5.20: funcionamiento SVM

5.4.5.3.1 CLASIFICACIÓN

En este capítulo se realiza un resumen de los métodos de clasificación de las máquinas de vectores soporte. El objetivo es dar a conocer las opciones que presenta este clasificador.

En primer lugar está el caso de los datos linealmente separables. Imaginemos un conjunto de coordenadas *S* en un espacio bidimensional:

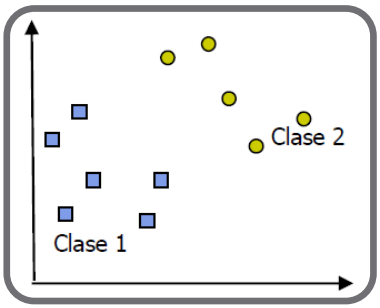


Figura 5.21: caso linealmente separable

Como se puede observar en la figura anterior, existen dos clases de datos de entrenamiento $x_i \in R^n$. Para poder realizar una discriminación apropiada de los mismos, el clasificador busca un hyperplano en una dimensión mayor. Sea $z = \varphi(x)$ la notación del correspondiente vector en el espacio de características con un mapeo de $\varphi\ de\ R^n$ a un espacio de características Z. El hyperplano está definido con la siguiente ecuación:

$$w * z + b = 0$$

Figura 5.22: hyperplano

Definido por el par (w, b), tal que podamos separar el punto x_i de acuerdo a la siguiente función:

$$f(x_i) = sign(w * z_i + b) = f(x) = \begin{cases} 1, & y_i = 1 \\ -1, & y_i = -1 \end{cases}$$

Figura 5.23: hyperplano

Sean válidos para todos los elementos del conjunto *S*. Para el caso linealmente separable de *S* podemos encontrar un único hyperplano óptimo, para el cual, el margen entre las proyecciones de los puntos de entrenamiento de dos diferentes clases, es maximizado.

El siguiente caso que se puede presentar es que los datos no sean linealmente separables, lo que complica algo más la clasificación.

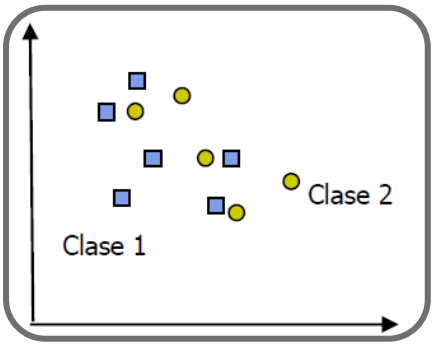


Figura 5.24: caso linealmente no separable

Para poder establecer una diferenciación en este tipo de casos, es necesaria una función llamada *Kernel* que calcule el producto punto de los puntos de entrada en el espacio de características Z, esto es [Betancourt 2005]:

$$z_i * z_j = \varphi(x_i) * \varphi(x_j) = K(x_i, x_j)$$

Figura 5.25: kernel

La aplicación ofrece la posibilidad al investigador de seleccionar el tipo de kernel que más convenga a sus necesidades, puede elegir entre los siguientes:

- **Polinomial:** el kernel está definido en función del grado *p* mediante la siguiente función genérica:

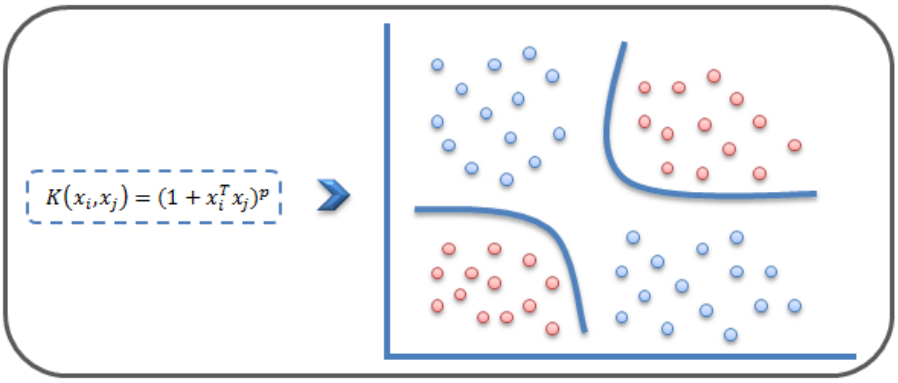


Figura 5.26: kernel polinomial

- **Lineal:** el kernel estará determinado por la siguiente función:

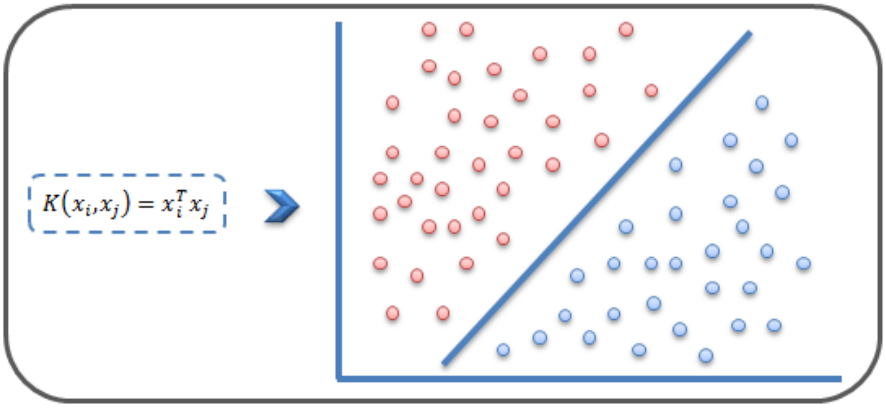


Figura 5.27: kernel lineal

- **Gaussiano:** en este caso el la función kernel está definida de la siguiente manera:

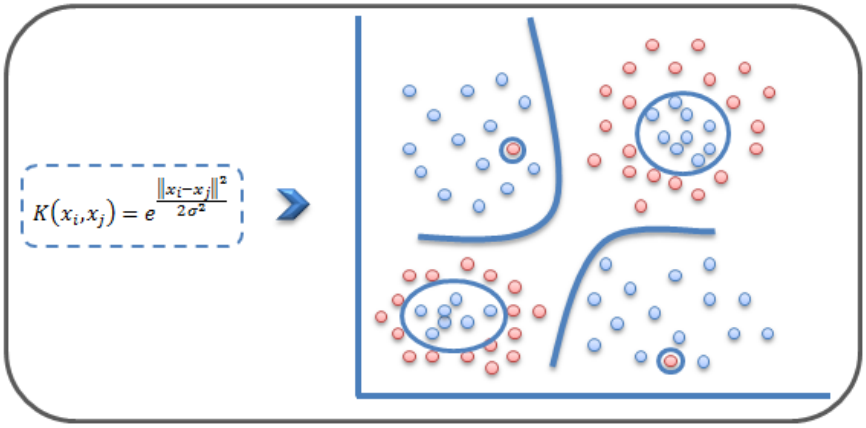


Figura 5.28: kernel gaussiano

5.4.6 FASE DE ENTRENAMIENTO

Una vez diseñada la arquitectura de la red se tiene que proceder a realizar el entrenamiento. En él la red "*aprende"* el comportamiento que debe tener, es decir, "*aprende"* a dar la respuesta adecuada de acuerdo a los patrones de entrada que se le presentan.

Durante la fase de aprendizaje se entrena a la red para que vaya modificando sus pesos sinápticos, adaptándolos paulatinamente para que la respuesta de la red sea la correcta. Para la realización de este proyecto se ha realizado un aprendizaje supervisado.

Mediante esta técnica se presentan a la red diferentes patrones de estímulo de entrada de forma repetitiva. Las características de ensayo están formadas por parejas: patrón de estímulos y respuesta. En este caso el patrón de estímulos está formado por las características RGB y por las coordenadas espaciales, y la respuesta sería el valor uno en caso correcto y el menos uno en el incorrecto.

Al realizar el entrenamiento la respuesta que da la red a cada patrón se compara con la respuesta correcta ante dicho patrón. A continuación se vuelven a reajustar los pesos sinápticos en función de lo correcta que sea la comparación. El reajuste de los pesos está orientado para que la respuesta de la red se acerque cada vez más a la respuesta real.

Para proporcionar una mayor flexibilidad al investigador, la aplicación da la posibilidad de escoger el número de datos de entrada al clasificador.

5.4.7 FASE DE VALIDACIÓN O TEST

Tras la fase de entrenamiento se realiza la fase de ejecución. En ella se le pide a la red neuronal que responda a diferentes datos de entrada de los proporcionados en la fase anterior. Mediante los ejemplos aprendidos en la fase de ensayo la red debe ser capaz de generar una respuesta correcta.

Para operar con una red entrenada se procede igual que en la fase precedente. La única diferencia son los datos de entrada a la red. En este caso se obtienen nuevos datos mediante la extracción de características de un nuevo frame.

5.5 PRUEBAS MAT-PR

En el siguiente capítulo se exponen las pruebas que se han realizado sobre la aplicación MAT-PR. Se realiza una introducción sobre las mismas y, a continuación, se enumeran las baterías de pruebas.

5.5.1 INTRODUCCIÓN

En este apartado se comprueba el correcto funcionamiento de la aplicación y se muestran los resultados que devuelve para así poder ofrecer un tanto por ciento de acierto, sobre el que se pueda calificar a la aplicación.

Para poder probar el sistema se requiere un conjunto de datos sobre los que comprobar si el reconocimiento es correcto. Se opta por elegir la base de datos AV-UC3M, ya que se trata de una batería de pruebas hecha especialmente para este proyecto.

Se seleccionan los frames considerados más representativos del mundo real. Se escogen vídeos de prueba tanto de hombres como de mujeres, teniendo en cuenta diferentes tipos de iluminación, complementos, maquillajes, etc.

5.5.2 BATERÍA DE PRUEBAS

Para realizar todas las pruebas a la aplicación, se escogen una serie de frames pertenecientes a vídeos de la base de datos AV-UC3M. Los fotogramas son seleccionados intentando reunir el mayor número de características diferentes, haciendo de ellos una muestra representativa de todos los vídeos.

En este caso se opta por un único frame perteneciente a un interlocutor de género femenino, debido a que reúne las características necesarias para realizar un buen reconocimiento. El fotograma contiene todas las regiones de interés que se necesitan para extraer las conclusiones oportunas de cada clasificador.

Tanto para las redes neuronales como para las máquinas de vectores soporte, se realizan cincuenta pruebas con diferentes tipos de datos. Además se varía tanto el número de neuronas y capas en las NN, como los núcleos de las SVM.

5.5.3 EVALUACIÓN

En este apartado se detallan todas las pruebas realizadas al sistema para comprobar si se han cumplido los objetivos fijados con anterioridad. Una vez realizadas se extraen las conclusiones pertinentes. Las pruebas a las que se somete al sistema están divididas en dos apartados: test de redes neuronales y pruebas a la máquina de vectores de soporte.

5.5.3.1 TEST REDES NEURONALES

La batería de pruebas elegida se extrae de la base de datos creada para este PFC. De todos los vídeos que la forman se selecciona un único frame, ya que de esta manera se pueden comprobar los resultados con una misma imagen de test. Si se hubieran seleccionado varios tipos de frames, no se podrían contrastar los resultados.

Las partes del rostro importantes para la evaluación del sistema están resaltadas mediante colores no presentes en el resto de la imagen. De esta forma se pueden extraer las conclusiones apropiadas a partir de una simple observación del frame. Estos colores están compuestos por los siguientes colores RGB:

Zona	Rojo	Verde	Azul
Ojos	**251**	**0**	**222**
Labios	**2**	**240**	**255**
Dientes	**255**	**255**	**0**
Lengua	**227**	**143**	**236**

Tabla 5.2: colores test

Un factor muy importante a la hora de realizar esta fase de evaluación son los datos que se introducen en la red neuronal. Para que el sistema cuente con varias opciones se varía tanto el número de datos

correctos pertenecientes a cada una de las zonas de interés (ojos, labios, dientes y lengua), como los datos incorrectos pertenecientes al resto de la cara (pelo, cejas, piel, ropa, etc.). Por lo tanto, se introducen desde 15.000 datos por región, hasta llegar a los 30.000.

El siguiente elemento a tener en cuenta es el número de capas y neuronas que posee la red. Ya que el sistema también permite seleccionarlas en función de los intereses del investigador, se realizan varias pruebas con distintos tipos de redes.

La siguiente tabla resumen presenta la batería de pruebas seleccionada para realizar el test. En la columna de la izquierda se muestra el número de capas de la red y las neuronas que forman la misma, mediante la siguiente nomenclatura:

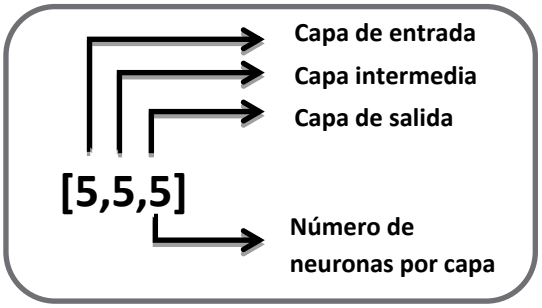
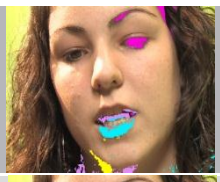
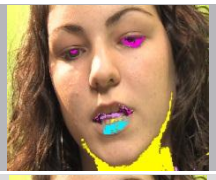
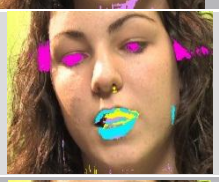
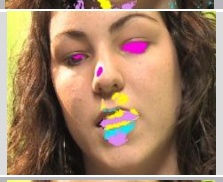
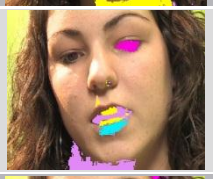
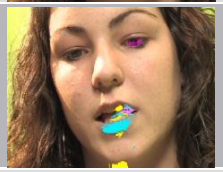
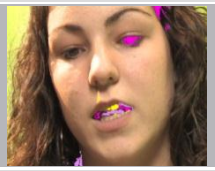
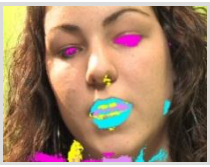
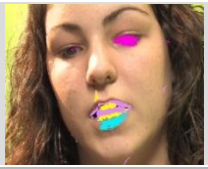
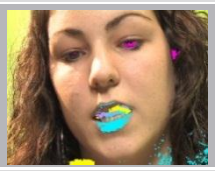
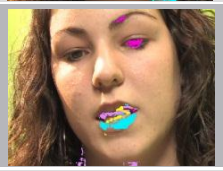
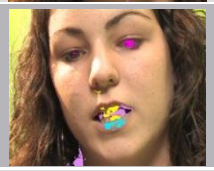
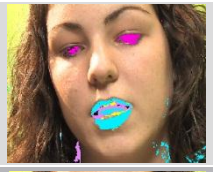
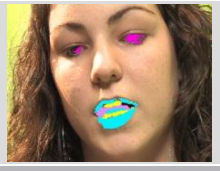
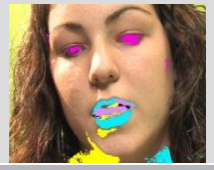
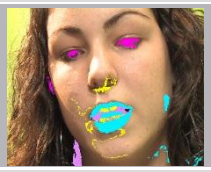
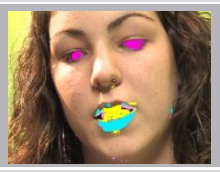
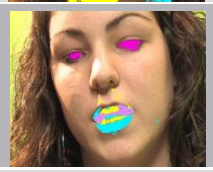
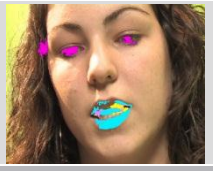
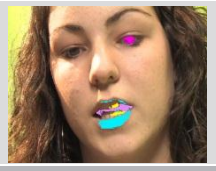
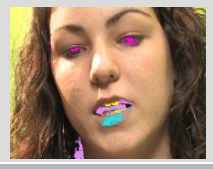
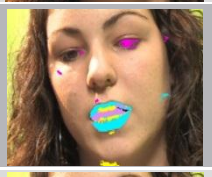
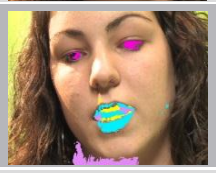
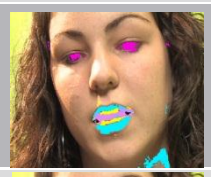
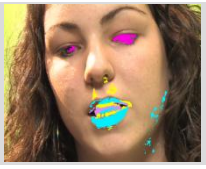
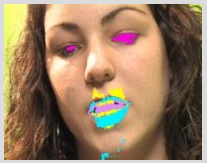
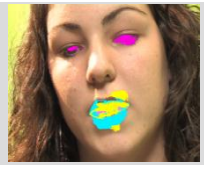


Figura 5.29: nomenclatura Red Neuronal

Redes Neuronales	Datos válidos: 15000 Datos falsos: 15000	Datos válidos: 15000 Datos falsos: 30000	Datos válidos: 20000 Datos falsos: 30000
[1,3,1]			
[1,3,3]			
[1,4,1]			
[1,4,4]			
[1,5,1]			
[1,5,5]			
[5,5,5]			
[1,6,1]			
[1,6,6]			
[6,6,6]			

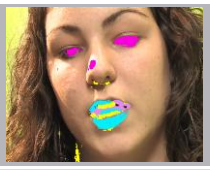
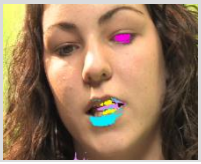
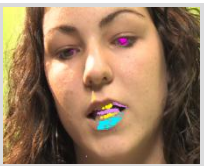
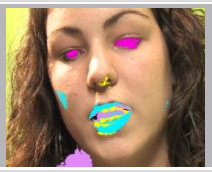
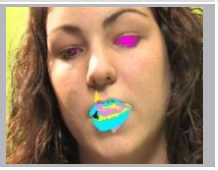
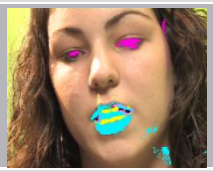
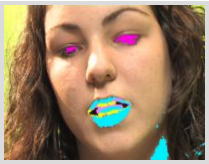
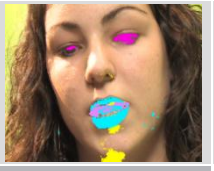
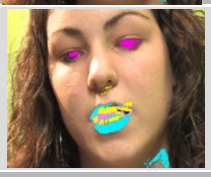
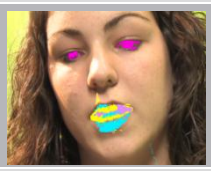
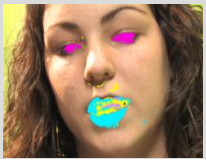
Redes Neuronales	Datos válidos: 15000 Datos falsos: 15000	Datos válidos: 15000 Datos falsos: 30000	Datos válidos: 20000 Datos falsos: 30000
[6,6,6,6]	--	--	
[1,7,1]			
[1,7,7]			
[7,7,7]			
[8,8,8,8]	--	--	
[9,9,9,9]	--	--	

Tabla 5.3: resumen resultados NN

En base a la anterior tabla se pueden extraer las siguientes conclusiones:

- **Número de datos válidos:** las diferencias que existen entre 15.000 y 20.000 datos no son representativas, ya que aunque hay veces que parece que mejoran las tasas de acierto en ciertas regiones, existen otras en las que las tasas de falsos positivos aumentan.
- **Número de datos falsos:** disminuye la tasa de falsos positivos, manteniendo de forma similar la tasa de aciertos entre los 15.000 y los 30.000 datos falsos. Es una buena herramienta para mejorar el porcentaje de acierto del sistema.
- **Número de neuronas por capa:** a medida que el número de neuronas crece en cada capa, el reconocimiento tiende a ser mejor. Por ejemplo, las diferencias existentes entre asignar tres neuronas por capa a asignar seis, son bastante notables.
- **Número de capas:** el aumentar el número de capas en la red neuronal no asegura un porcentaje mayor de éxito. En general, los resultados son similares y sin grandes cambios. Incluso hay ocasiones en las que empeora el reconocimiento.

En una segunda fase se realizan pruebas sometiendo a las características a un proceso de normalización. Como se puede observar en la tabla siguiente se crea una nueva región de interés dividiendo la zona ocular en dos partes: derecha e izquierda.

Mediante esta técnica se intentan mejorar las tasas de reconocimiento facial presentadas con anterioridad. Los test se realizaron en varios tipos de redes neuronales, dando como resultado los siguientes frames:

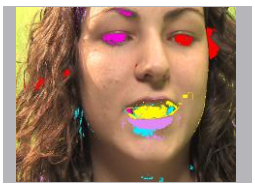
	Min-Max	Median	Z-Score
Red Neuronal			

Tabla 5.4: resultados normalización

Las conclusiones que se pueden extraer de estos test fueron poco satisfactorias, ya que no se mejoraron las tasas de acierto en comparación con el método utilizado anteriormente. La normalización min-max ha sido la que ha dado unos resultados más satisfactorios, aunque con casi ningún aporte nuevo.

5.5.3.2 TEST MÁQUINAS DE VECTORES SOPORTE

Para llevar a cabo estos test se vuelve a utilizar la base de datos AV-UC3M. En este caso, se realizan menos pruebas que en el caso de las redes neuronales. La mayoría de resultados han sido similares y sólo se exponen los frames más representativos.

Los fotogramas utilizados están sometidos a pruebas con los distintos kernels de SVM. En primer lugar se muestra el lineal, a continuación el radial y, por último, el polimonial. El resultado es exactamente igual que en el apartado anterior, se resaltan mediante color las zonas de interés (labios, ojos, dientes y lengua).

A continuación se muestra la tabla resumen con los resultados obtenidos:

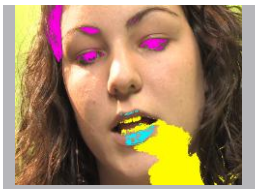
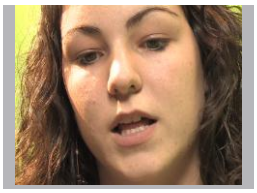
	Lineal	Polinomial	Radial	Sigmoide
SVM				--

Tabla 5.5: resultados SVM

Los resultados de las máquinas de vectores soporte no son los esperados. En la mitad de los casos realizan reconocimientos acertados, aunque con unas tasas de falsos positivos muy elevadas. En otros casos, el sistema ni siquiera logra realizar una clasificación.

En una siguiente fase se realiza una normalización de las características mediante tres métodos: min-max, median y z-score. Los resultados de las pruebas son los siguientes:

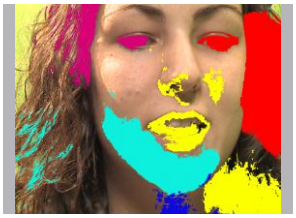
	Median	Min-Max	Z-Score
SVM			

Tabla 5.6: resultados SVM normalizado

Las conclusiones que se pueden extraer de estos test son concluyentes: ninguna de las pruebas a las que se somete a las máquinas de vectores soporte mediante la normalización de características proporciona ninguna mejora con respecto al anterior método.

5.5.4 CONCLUSIONES GLOBALES

En esta sección se extraen las conclusiones a las que se ha llegado a partir de los resultados obtenidos en la fase de test. Están divididas en dos apartados: redes neuronales y máquinas de vectores soporte.

En primer lugar se comparan las tasas de reconocimiento entre los dos clasificadores. El número de características acertadas por parte de las redes neuronales supera de manera elevada a las de las máquinas de vectores soporte. En cuanto a los falsos positivos, las NN poseen unos mejores resultados.

En segundo lugar, cabe destacar que el tiempo de procesamiento empleado para realizar las pruebas de uno y otro clasificador ha sido elevado, ya que se trata de un proceso muy costoso computacionalmente hablando. Aun así, las redes neuronales han realizado los test en un menor tiempo.

Por último, por todos los motivos expuestos con anterioridad se puede concluir que las redes neuronales son la mejor opción para realizar el reconocimiento de caracteres faciales dentro del campo de los clasificadores.

6 GESTIÓN DEL PROYECTO

En este apartado se detalla la planificación que ha tenido el proyecto en su realización. Además, se especifica el presupuesto necesario para llevarlo a cabo. A continuación se describen los recursos necesarios para su desarrollo, tanto económicos como humanos y, por último, se enumeran las etapas del mismo.

6.1 INTRODUCCIÓN

A la hora de emprender un nuevo proyecto resulta determinante establecer una planificación para llevar a cabo un plan de trabajo eficiente. Es una tarea imprescindible para cumplir con los objetivos propuestos en el tiempo que se considere adecuado. El objetivo es realizar una estimación aproximada del tiempo, del riesgo y los costes que conlleva el proyecto.

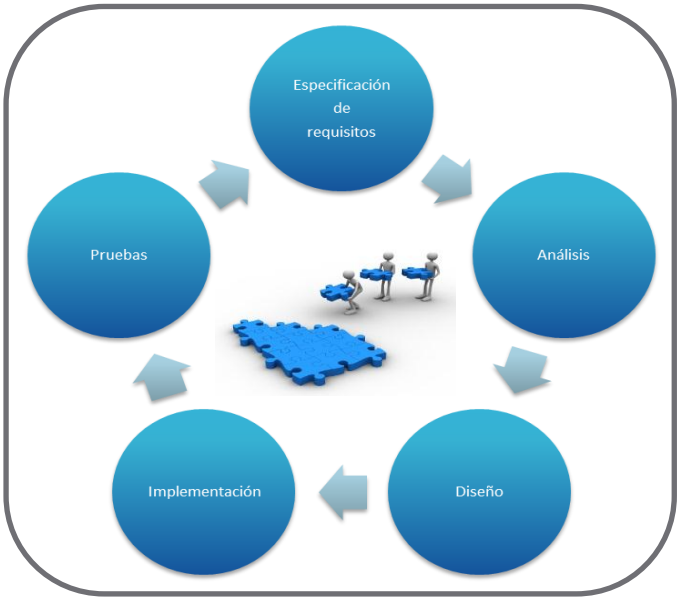


Figura 6.1: metodología software

El proceso software es la descripción de las etapas que se siguen durante la ejecución de un proyecto. Su correcta definición permite asegurar una adecuada asignación de recursos y conocer el estado del proyecto en cada momento. Establece el marco de trabajo, tanto técnico como de gestión, en la aplicación de los métodos, las herramientas y las personas a las tareas de desarrollo de software [Cuevas, 2003].

Una vez definido, es necesario establecer una metodología de desarrollo. Se trata de una serie de procedimientos, técnicas, herramientas y soporte documental que ayudan a los desarrolladores a realizar un nuevo software. En la figura anterior se muestra un ejemplo de metodología. Estas etapas se utilizan para estructurar, planear y controlar el proceso de desarrollo software.

6.2 POSIBLES ALTERNATIVAS

En esta sección se explican brevemente las diferentes posibilidades que se barajan en durante la realización del proyecto en distintos campos tecnológicos.

6.2.1 LENGUAJES DE PROGRAMACIÓN

Las posibles alternativas que se estudiaron a la hora de ejecutar la implementación fueron las siguientes:

- **JAVA:** Java es un lenguaje de programación orientado a objetos desarrollado por Sun Microsystems. El lenguaje en sí mismo toma mucha de su sintaxis de C y C++, pero tiene un modelo de objetos más simple y elimina herramientas de bajo nivel, que suelen inducir a muchos errores, como la manipulación directa de punteros o memoria.

- **Visual C++:** Esta especialmente diseñado para el desarrollo y depuración de código escrito para las API's de Microsoft Windows, DirectX y la tecnología Microsoft .NET Framework. Visual C++ hace uso extensivo del framework Microsoft Foundation Classes, el cual es un conjunto de clases C++ para el desarrollo de aplicaciones en Windows. Está basado en C++, y es compatible en la mayor parte de su código con este lenguaje, a la vez que su sintaxis es exactamente igual.

- **C#:** Lenguaje de programación orientado a objetos desarrollado y estandarizado por Microsoft como parte de su plataforma .NET, que después fue aprobado como un estándar por la ECMA e ISO. Su sintaxis básica deriva de C/C++ y utiliza el modelo de objetos de la plataforma.NET el cual es similar al de Java aunque incluye mejoras derivadas de otros lenguajes.

- **Simulink:** Simulink es una herramienta de simulación de modelos o sistemas, con cierto grado de abstracción de los fenómenos físicos involucrados en los mismos. Se hace hincapié en el análisis de sucesos, a través de la concepción de sistemas.

- **Matlab:** es un lenguaje de alto desempeño diseñado para realizar cálculos técnicos. Integra el cálculo, la visualización y la programación en un ambiente fácil de utilizar, donde los problemas y las soluciones se expresan en una notación matemática. Permite resolver muchos problemas computacionales, específicamente aquellos que involucren vectores y matrices, en un tiempo mucho menor al requerido para escribir un programa en un lenguaje escalar no interactivo tal como C o Fortran.

6.2.2 ENTORNOS DE PROGRAMACIÓN

Los entornos de desarrollo estudiados para llevar a cabo el proyecto son los siguientes:

- **Microsoft Visual Studio:** Microsoft Visual Studio es un entorno de desarrollo integrado (IDE, por sus siglas en inglés) para sistemas operativos Windows. Soporta varios lenguajes de programación tales como Visual C++, Visual C#, Visual J#, ASP.NET y Visual Basic .NET, aunque actualmente se han desarrollado las extensiones necesarias para muchos otros. Visual Studio permite a los desarrolladores crear aplicaciones, sitios y aplicaciones web, así como servicios web en cualquier entorno que soporte la plataforma .NET (a partir de la versión net 2002). Así se pueden crear aplicaciones que se intercomuniquen entre estaciones de trabajo, páginas web y dispositivos móviles.

- **Eclipse:** Eclipse es un entorno de desarrollo integrado de código abierto multiplataforma para desarrollar lo que el proyecto llama "Aplicaciones de Cliente Enriquecido", opuesto a las aplicaciones "Cliente-liviano" basadas en navegadores. Esta plataforma, típicamente ha sido usada para desarrollar entornos de desarrollo integrados (del inglés IDE), como el IDE de Java llamado Java Development Toolkit (JDT) y el compilador (ECJ) que se entrega como parte de Eclipse (y que son usados también para desarrollar el mismo Eclipse).

- **Matlab:** MATLAB ofrece un entorno de desarrollo integrado (IDE) muy versátil con un lenguaje de programación propio (lenguaje M). MATLAB está disponible para sistemas operativos Windows, Unix y Apple Mac OS X.

6.2.3 FORMATOS DE VÍDEO

En esta sección se exponen los formatos de vídeo estudiados para la ejecución de la base de datos AV-UC3M, son los siguientes:

- **AVI, AVI2**: Audio Video Interleave que significa algo así como intercalado de audio y vídeo, es un formato que contiene audio y vídeo, y funciona guardando una capa de vídeo seguida por otra de audio, fue diseñado por Microsoft y mejorado por Matrox llamándose AVI2.

- **Microsoft Windows Media Vídeo**: Desarrollado también por Microsoft para su reproductor Windows Media Player, los archivos de este formato son el .wmv que es el archivo que contiene vídeo, .wma que contiene audio, y .asf.

- **QuickTime**: Es un formato desarrollado por Apple para su sistema Mac, pueden visualizarse en el reproductor QuickTime y en Windows disponemos del QuickTime for Windows, la extensión de este formato es .mov, su reproductor nos permitirá realizar vídeos del mismo formato y algunas opciones básicas para editarlos.

6.2.4 FORMATOS DE IMAGEN

Los formatos más populares que se analizaron para la ejecución del proyecto son [Ordoñez 2005]:

- **BMP:** El formato bmp (Bit Map) es el formato de las imágenes de mapa de bits de Windows. Su uso fue muy extendido, pero los archivos son muy grandes dado la escasa compresión que alcanzan.

- **TIF:** El formato TIF (Tag Image File Format) se utiliza para imágenes de mapa de bits y es admitido prácticamente por todas las aplicaciones de autoedición y tratamiento de imágenes. Este formato fue desarrollado por Aldus Corporation. Lo reconocen casi todos los programas. Además, es compatible con PC y Mac. Su uso es de los más extendidos en la industria gráfica por la calidad de imagen y de impresión que presenta.

- **GIF:** El formato GIF corresponde a las siglas de Graphics Interchange Format propiedad de eCompuServe. El formato GIF es preferible para las imágenes de tonos no continuos o cuando hay grandes áreas de un mismo color ya que utiliza una paleta de color indexado que puede tener un máximo de 256 colores. Una de sus mayores ventajas es que podemos elegir uno o varios colores de la paleta para que sean transparentes y podamos ver los elementos que se encuentren por debajo de estos. También es uno de los pocos formatos de imagen con el que podemos mostrar animaciones porque hace que distintos frames se ejecuten secuencialmente. Además, es un formato de compresión diseñado para disminuir el tiempo de transferencia de datos por las líneas telefónicas.

- **JPG o JPEG:** Este formato toma su nombre de Joint Photographic Experts Group, asociación que lo desarrollo. Se utiliza usualmente para almacenar fotografías y otras imágenes de tono continuo. Gracias a que utiliza un sistema de compresión que de forma eficiente reduce el tamaño de los archivos. En contraste con GIF, JPEG guarda toda la información referente al color con millones de colores (RGB) sin obtener archivos excesivamente grandes. Además, los navegadores actuales reconocen y muestran con fidelidad este formato.

6.2.5 FORMATOS DE AUDIO

Los dos formatos de audio estudiados se exponen a continuación:

- **PCM:** los formatos PCM contienen toda la información que salió del convertidor analógico a digital, sin ninguna omisión y por eso, tienen la mejor calidad. Dentro de esta categoría se encuentran los formatos WAV, AIFF, SU, AU y RAW (crudo). La diferencia principal que tienen estos formatos es el encabezado, alrededor de 1000 bytes al comienzo del archivo.

- **mp3:** es un formato de audio digital comprimido con pérdida. Descarta información que no es perceptible por el oído humano para lograr que el mismo fragmento de audio pueda ocupar en la memoria hasta una décima parte.

6.2.6 ESPACIOS DE COLOR

En esta sección se describen los diferentes espacios de color:

- **RGB:** es conocido como un espacio de color aditivo (colores primarios) porque cuando la luz de dos diferentes frecuencias viaja junta, desde el punto de vista del observador, estos colores son sumados para crear nuevos tipos de colores. Los colores: rojo, verde y azul fueron escogidos porque cada uno corresponde aproximadamente con uno de los tres tipos de conos sensitivos al color en el ojo humano (65% sensibles al rojo, 33% sensibles al verde y 2% sensibles al azul). Con la combinación apropiada de rojo, verde y azul se pueden reproducir muchos de los colores que pueden percibir los humanos.

- **HSV:** Es un espacio cilíndrico, pero normalmente asociado a un cono o cono hexagonal, debido a que es un subconjunto visible del espacio original con valores válidos de RGB.

  - Matiz: (**H**ue): Se refiere a la frecuencia dominante del color dentro del espectro visible. Es la percepción de un tipo de color, normalmente la que uno distingue en un arcoíris, es decir, es la sensación humana de acuerdo a la cual un área parece similar a otra o cuando existe un tipo de longitud de onda dominante. Incrementa su valor mientras nos movemos de forma anti horaria en el cono, con el rojo en el ángulo 0.

  - Saturación: (**S**aturation): Se refiere a la cantidad del color o a la "pureza" de éste. Va de un color "claro" a un color más vivo (azul cielo/azul oscuro). También se puede considerar como la mezcla de un color con blanco o gris.

  - Valor: (**V**alue): Es la intensidad de luz de un color. Dicho de otra manera, es la cantidad de blanco o de negro que posee un color.

- **YCbCr** es una familia de espacios de colores usada en sistemas de transmisión de vídeo y fotografía digital. No es un espacio de color absoluto, sino que es una *forma de codificar* información RGB. El color que se muestra depende del primario RGB usado para mostrar la señal. Los componentes que forman la base colorimétrica son:

    - **Y:** es el componente de luminancia. En sistemas de televisión analógica en blanco y negro es la luz captada por el sensor, que es reflejada de los objetos de una determinada escena. En los sistemas en color, los sensores captan las componentes RGB y se calcula la luminancia como la suma ponderada de las tres componentes de color, Y(R, G, B)=0'299*R+0'587*G+0'114*B. Es la que contiene la información más relevante de una imagen porque el Sistema visual es mucho más sensible a esta información que a los colores.

    - **Cb:** Componente de color azul. (B-Y).

    - **Cr:** Componente de color rojo. (R-Y).

El sistema elegido para transmitir la señal es la combinación de la luminancia (Y), y dos señales diferencia de color R-Y, B-Y. Se utilizan estas dos señales diferencia porque así se consigue una mayor protección frente a las interferencias y el ruido.

6.2.7 CLASIFICADORES

Las posibles opciones sopesadas en cuanto a los sistemas clasificadores se ofrecen a continuación:

- **Redes neuronales:** las redes de neuronas artificiales (denominadas habitualmente como RNA o en inglés como: "ANN") son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. En inteligencia artificial es frecuente referirse a ellas como redes de neuronas o redes neuronales.

- **Modelos ocultos de Markov:** un modelo oculto de Markov o HMM (por sus siglas del inglés, *Hidden Markov Model*) es un modelo estadístico en el que se asume que el sistema a modelar es un proceso de Markov de parámetros desconocidos. El objetivo es determinar los parámetros desconocidos (u *ocultos*, de ahí el nombre) de dicha cadena a partir de los parámetros observables. Los parámetros extraídos se pueden emplear para llevar a cabo sucesivos análisis, por ejemplo en aplicaciones de reconocimiento de patrones. Un HMM se puede considerar como la red bayesiana dinámica más simple.

- **Máquinas de vectores soporte:** Las máquinas de vectores soporte o máquinas de vectores de soporte (Support Vector Machines, SVMs) son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik y su equipo en los laboratorios AT&T. Estos métodos están propiamente relacionados con problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases por un espacio lo más amplio posible. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de su proximidad pueden ser clasificadas a una u otra clase.

- **Modelo de Mezclas Gaussianas:** un modelo de mezcla gaussiana (GMM) es una función de densidad de probabilidad paramétrica representada como una suma ponderada de Gauss. El GMM se utiliza comúnmente como modelo paramétrico en la distribución de probabilidad para clasificar las características de un sistema biométrico, como por ejemplo, un sistema de reconocimiento del habla.

6.3 ALTERNATIVA ELEGIDA

En esta sección se enumeran todas las decisiones tecnológicas que se han tomado de entre todas las opciones descritas en el capítulo anterior:

- **Lenguajes de programación:** para efectuar la implementación de las aplicaciones pertenecientes a este proyecto se escogen dos lenguajes de programación diferentes: Simulink y Matlab. El lenguaje Simulink se selecciona por sus herramientas de edición de imagen y vídeo, su alto nivel de extracción y la facilidad de programación. En cuanto a Matlab, posee las mejores funcionalidades para trabajar con matrices de datos y redes neuronales.

- **Entorno de programación:** se selecciona el entorno que trae por defecto Matlab, ya que es el lenguaje que se va a utilizar para el desarrollo.

- **Formato de audio:** para realizar la grabación de la base de datos AV-UC3M se escoge el formato de audio WAV, ya que posee mejor calidad de sonido que el mp3.

- **Espacios de color:** en el desarrollo de la aplicación SIM-RC se llevan a cabo pruebas con todos los espacios de color estudiados, por lo que se escogieron RGB, HSV y YCbCr.

- **Clasificadores:** finalmente se eligieron dos tipos: redes neuronales y máquinas de vectores soporte. Los Modelos Ocultos de Markov se descartan para buscar otras alternativas, ya que la mayoría de autores lo utilizan actualmente. El GMM se decide no utilizar después de los irregulares resultados obtenidos con los otros clasificadores.

6.4 ESTIMACIÓN DE RECURSOS TEMPORALES

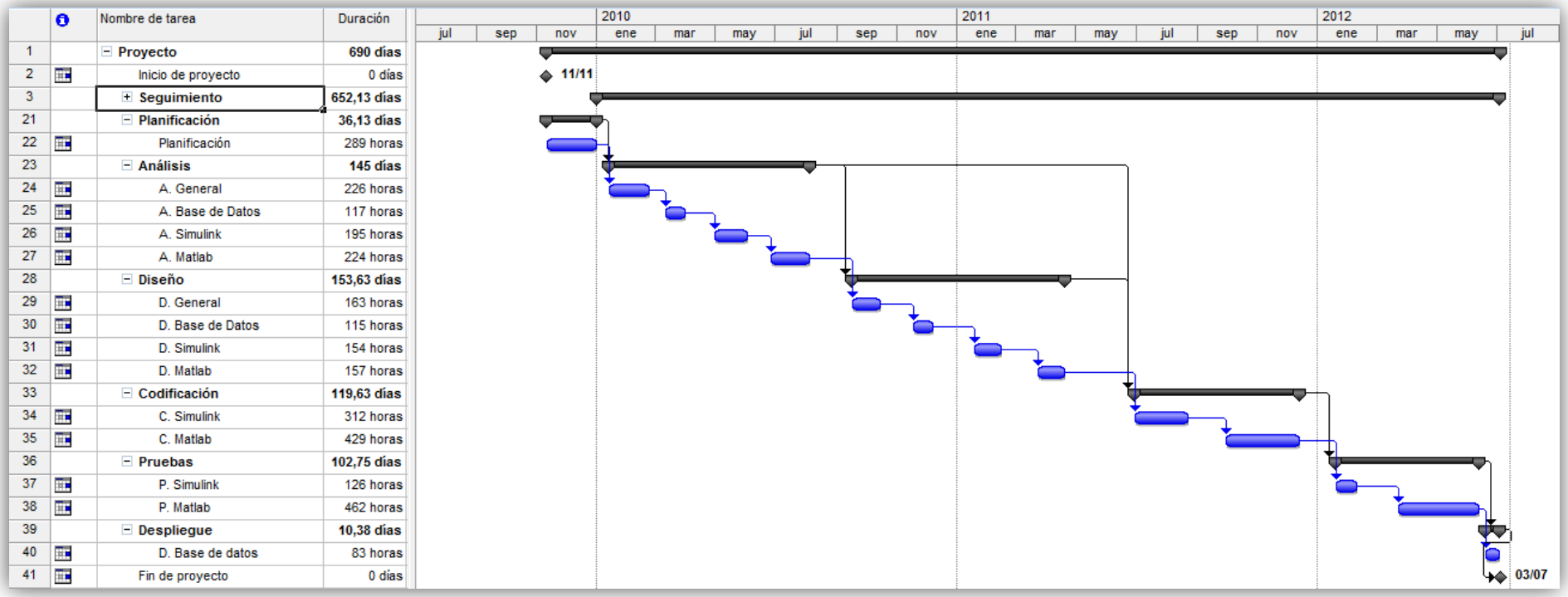
El proyecto se inicia el 11 de noviembre del año 2009. Se desarrolla durante los tres años siguientes, dándolo por finalizado el día 28 de Septiembre de 2012.

Se ha empleado un total de 3.086 horas, cuyo desglose se muestra a continuación:

TAREA		HORAS
1	Planificación	289
2	Análisis	762
3	Diseño	589
4	Codificación	741
5	Pruebas	588
6	Despliegue	83
7	Seguimiento	34
TOTALES:		**3.086**

Tabla 6.1: recursos temporales por fases del proyecto

El diagrama de Gantt de la tabla anterior muestra la sucesión de cada una de las tareas que componen el desarrollo del proyecto.



	Nombre de tarea	Duración
1	**Proyecto**	**690 días**
2	Inicio de proyecto	0 días
3	**Seguimiento**	**652,13 días**
21	**Planificación**	**36,13 días**
22	Planificación	289 horas
23	**Análisis**	**145 días**
24	A. General	226 horas
25	A. Base de Datos	117 horas
26	A. Simulink	195 horas
27	A. Matlab	224 horas
28	**Diseño**	**153,63 días**
29	D. General	163 horas
30	D. Base de Datos	115 horas
31	D. Simulink	154 horas
32	D. Matlab	157 horas
33	**Codificación**	**119,63 días**
34	C. Simulink	312 horas
35	C. Matlab	429 horas
36	**Pruebas**	**102,75 días**
37	P. Simulink	126 horas
38	P. Matlab	462 horas
39	**Despliegue**	**10,38 días**
40	D. Base de datos	83 horas
41	Fin de proyecto	0 días

Tabla 6.2: diagrama de Gantt

6.5 ESTIMACIÓN DE RECURSOS ECONÓMICOS

Se detallan a continuación los recursos tanto materiales como humanos necesarios para la ejecución del proyecto.

6.5.1 RECURSOS MATERIALES

La estimación del coste y de los periodos de amortización de los recursos materiales se han obtenido en base a lo descrito en [RD1777, 2004]. En este documento se establece un plazo de amortización de 8 años para los elementos de tratamiento de información y de 6 años para los programas informáticos. En base a esta información, el coste asociado a cada elemento viene determinado por la siguiente expresión:

$$Coste = Precio \cdot Coef = Precio \cdot \left(Unidades \cdot \frac{d}{a}\right)$$

Figura 6.2: expresión para el cálculo de las amortizaciones

En la expresión anterior, a y d corresponden al periodo de amortización del elemento y a la duración del proyecto en años, respectivamente. La tabla muestra el desglose de los costes asociados a los recursos materiales.

CONCEPTO	PRECIO	COEFICIENTE	COSTE
PC Portátil Lenovo G550	375 €	9,4 %	35,25 €
Impresora HP D2460	41,16 €	9,4 %	3,86 €
Conexión ADSL a 10 Mbps durante 28 meses a 18,80 €/mes	526.4 €	100,0 %	526,40 €
Microsoft Office 2010	219,27 €	12,5 %	27,41 €
Microsoft Project 2010	149,00 €	12,5 %	18,63 €
Microsoft Visio 2010	221,00 €	12,5 %	27,63 €
MATLAB & Simulink Student Version Release 2010a - Windows	29 €	12,5 %	3,56 €
Adobe Potoshop CS 4	849€	12,5 %	106,12 €
TOTAL			**748,86 €**

Tabla 6.3: recursos materiales

6.5.2 RECURSOS HUMANOS

A partir de los recursos temporales presentados en la sección 6.3 y de los perfiles profesionales descritos a continuación, se determina el coste asociado a los recursos humanos. En la siguiente tabla se muestran los perfiles de los profesionales requeridos para la ejecución de las diferentes tareas y los costes de cada uno de ellos.

DESCRIPCIÓN	€/HORA	HORAS	COSTE
Director de Proyecto	113,00 €	98	11.074 €
Jefe de Proyecto	75,00 €	184	13.800 €
Analista de Servicios	63,00 €	374	23.562 €
Desarrollador de Aplicaciones	50,00 €	522	26.100 €
Probador de Software	25,00 €	358	8.950 €
TOTAL			**83.486 €**

Tabla 6.4: recursos humanos

A continuación se detallan los perfiles profesionales que han desarrollado este proyecto, detallando sus estudios, habilidades y costes. Son los siguientes:

- **Director de Proyecto**: debe coordinar y supervisar el avance del proyecto en todos los aspectos: Organización, Planificación, Ejecución, Administración y Control. Los honorarios de este profesional se sitúan en 113 €/hora.

- **Jefe de Proyecto**: Profesional titulado en Ingeniería Informática con amplia formación en gestión de proyectos y recursos humanos. Los honorarios de este profesional se encuentran en 75 €/hora.

- **Analista de Servicios**: Profesional titulado en Ingeniería de Informática con conocimientos de arquitectura de redes, protocolos de comunicaciones y seguridad en entornos web. Así mismo, ha de poseer conocimientos en biometría. Los honorarios de este profesional se encuentran en 63 €/hora.

- **Desarrollador de Aplicaciones**: Profesional titulado en Ingeniería Informática con conocimientos en Matlab y Simulink. Sus honorarios se sitúan en 50 €/hora.

- **Probador de Software**: Profesional titulado en Ciclo Formativo de Grado Superior en Informática. Su función es la de validar la aplicación a nivel de usuario y reportar los problemas. Sus honorarios se encuentran en 25 €/hora.

6.5.3 COSTES TOTALES

El coste total asociado al desarrollo del proyecto asciende a 119.276,44 € (ciento diez y nueve mil doscientos setenta y seis euros con cuarenta y cuatro céntimos). La siguiente tabla recoge el desglose de los costes.

CONCEPTO	VALOR
Recursos Materiales	748,86 €
Recursos Humanos	83.486,00 €
Gastos Generales (20% Recursos Humanos)	16.846,87 €
Subtotal	**101.081,73 €**
I.V.A. (18% Base Imponible)	18.194,71 €
TOTAL	**119.276,44 €**

Tabla 6.5: costes del proyecto

6.6 PLAN DE PROYECTO

Aquí se detallan las diferentes etapas del plan del proyecto. La primera es la enumeración de sus fases. A continuación, se realiza un pequeño resumen sobre el seguimiento del proyecto llevado a cabo por los responsables del mismo. Por último, se listan las herramientas que han hecho posible la consecución de los objetivos.

6.6.1 FASES DEL PROYECTO

En el siguiente apartado se detallan las etapas en las que se ha dividido el proyecto. Especificando en cada una de ellas las tareas realizadas durante la misma.

6.6.1.1 PLANIFICACIÓN

La etapa de planificación pretende abordar los costes temporales y económicos de la ejecución del proyecto. Esta tarea se corresponde con la elaboración de la planificación contenida en el actual capítulo.

6.6.1.2 RECOLECCIÓN DE INFORMACIÓN

En esta fase del se establecen las bases sobre las que se asienta este proyecto. Para lograrlo se efectúa un proceso de aprendizaje continuo mediante la lectura de decenas de artículos científicos extraídos de diversas fuentes de información: libros de texto e internet.

6.6.1.3 GRABACIÓN DE LA BASE DE DATOS

El primer paso realizado para llevar a cabo la base de datos es el análisis en profundidad de los diferentes aspectos que la forman. En concreto, se tienen en cuenta los siguientes aspectos: lugar de grabación, número de individuos, iluminación, corpus e idioma.

A continuación se estudian distintas opciones para completar el corpus que reciten los sujetos que forman la base de datos. Para ellos se tuvieron en cuenta todo tipo de discursos: palabras sueltas, frases pertenecientes a la literatura clásica, números, nombres, mítines políticos, ensayos y artículos periodísticos.

Finalmente se ejecuta la grabación en un plazo de 3 días consecutivos, mediante la participación de una serie de voluntarios.

6.6.1.4 DESARROLLO APLICACIÓN SIM-RC

En esta fase se desarrolla la aplicación SIM-RC de discriminación y exposición de características faciales mediante el color. El primer paso consiste en el estudio y análisis de los diferentes lenguajes de programación que más se ajusten a las necesidades del sistema. Además se examinan otros aspectos como la entrada de datos al sistema, el procesamiento de la señal de vídeo, los espacios de color y el tipo de exposición al investigador.

El siguiente paso consiste en la implementación de la aplicación por medio de la plataforma Matlab, en concreto se codifica mediante la herramienta Simulink. Se decide utilizar debido a la funcionalidad que lleva integrada para la edición de vídeo e imagen. La decisión de seleccionar este módulo se toma debido al gran abanico de posibilidades que ofrece para conseguir los objetivos de este proyecto. El entorno de desarrollo de Matlab se ha elegido para realizar la codificación, ya que Simulink es un módulo de la herramienta Matlab.

Finalmente se somete a la aplicación a una batería de pruebas especialmente seleccionada dentro de la base de datos AV-UC3M. Mediante estos test se obtienen una serie de resultados que dan lugar a nuevas conclusiones con las que evaluar el sistema.

Estas dos últimas etapas de implementación y pruebas se repiten de forma cíclica hasta conseguir unos resultados satisfactorios.

6.6.1.5 DESARROLLO APLICACIÓN MAT-RP

El desarrollo de la aplicación MAT-RP comienza con el análisis de los aspectos fundamentales con los que tiene que contar para lograr los objetivos. Se estudian características como: el lenguaje de programación, la entrada de datos, los tipos de clasificador y su configuración por parte del usuario, la normalización de características y la visualización de las imágenes de salida.

En cuanto a la codificación del sistema se toma la decisión de utilizar el lenguaje de programación Matlab. Es la herramienta que más se ajusta a los requerimientos técnicos que necesita la aplicación. En cuanto al entorno de desarrollo se utiliza el predeterminado del entorno de Matlab.

Las pruebas ejecutadas en la aplicación se seleccionan dentro de la base de datos AV-UC3M. Se escogen una serie de frames procedentes de los vídeos de test para editarlos posteriormente para que los clasificadores discriminen las zonas faciales.

Las fases de implementación y pruebas se repiten de manera continua hasta que el equipo de investigación crea que se han cumplido los objetivos planteados al realizar el análisis previamente.

6.6.1.6 DESPLIEGUE

La fase de despliegue se corresponde con la puesta en marcha de la base de datos y de las aplicaciones una vez que las pruebas se han realizado de forma satisfactoria.

6.6.1.7 MEMORIA DEL PROYECTO

En esta fase se lleva a cabo la elaboración del documento de Memoria del proyecto. En él se refleja toda la información obtenida y los productos generados a lo largo del desarrollo del proyecto.

6.6.2 SEGUIMIENTO DEL PROYECTO

A lo largo del proyecto se realizan reuniones periódicas entre los miembros del equipo para comprobar que los objetivos se cumplían. A su vez se proponían nuevas tareas que mejoraran las funcionalidades propuestas al comienzo del mismo.

6.7 HERRAMIENTAS

A continuación se realiza una breve descripción de las herramientas utilizadas en el desarrollo del proyecto, tanto en el aspecto técnico como en la redacción de la memoria del mismo.

6.7.1 MATLAB

MATLAB es un lenguaje de computación técnica de alto nivel y un entorno interactivo para desarrollo de algoritmos, visualización de datos, análisis de datos y cálculo numérico. Con MATLAB, podrá resolver problemas de cálculo técnico más rápidamente que con lenguajes de programación tradicionales, tales como C, C++ y FORTRAN. Además, MATLAB contiene una serie de funciones para documentar y compartir su trabajo. Puede integrar su código de MATLAB con otros lenguajes y aplicaciones, y distribuir los algoritmos y aplicaciones que desarrollo usando MATLAB.

6.7.2 PHOTOSHOP

Photoshop CS4 Extended ofrece todas las funciones de edición y composición más avanzadas, y herramientas para trabajar con contenido en 3D y basado en el movimiento.

Se ha convertido, casi desde sus comienzos, en el estándar *de facto* en retoque fotográfico, pero también se usa extensivamente en multitud de disciplinas del campo del diseño y fotografía, como diseño web, composición de imágenes bitmap, estilismo digital, fotocomposición, edición y grafismos de vídeo y básicamente en cualquier actividad que requiera el tratamiento de imágenes digitales.

Photoshop ha dejado de ser una herramienta únicamente usada por diseñadores / maquetadores, ahora Photoshop es una herramienta muy usada también por fotógrafos profesionales de todo el mundo, que lo usan para realizar el proceso de "positivado y ampliación" digital, no teniendo que pasar ya por un laboratorio más que para la impresión del material.

Esta herramienta ha sido utilizada en este proyecto para realizar la modificación de las imágenes de entrada en la aplicación de discriminación de zonas faciales.

6.7.3 MICROSOFT OFFICE 2010

Microsoft Office 2010 es la versión más reciente del paquete ofimático de Microsoft. Cuenta con herramientas para editar textos, realizar hojas de cálculo, presentaciones de diapositivas y otras muchas aplicaciones. La elaboración de la presente memoria se ha realizado mediante MS Word, la planificación del proyecto se ha realizado con MS Project y los costes del proyecto se han calculado mediante MS Excel.

6.7.4 MOZILLA FIREFOX 3.6

Mozilla Firefox es un navegador de Internet libre y de código abierto desarrollado por la Corporación Mozilla, la Fundación Mozilla y un gran número de voluntarios externos. Firefox es un navegador multiplataforma y está disponible en versiones para Microsoft Windows, Mac OS X, GNU/Linux y algunos sistemas basados en Unix. Firefox se ha empleado para evaluar el funcionamiento de la aplicación web de la plataforma.

6.7.5 FFMPEG

Ffmpeg es un programa sin interfaz gráfica que permite convertir o transformar entre formatos multimedia, tanto de video como de audio. Aunque existen otros programas, algunos sin necesidad de usar comandos, es una de las opciones con más posibilidades y es muy rápida.

Se ha utilizado para extraer frames de todos los vídeos de la base de datos audio-visual. Una vez seleccionados se modifican con el programa de edición de imagen Photoshop y se introducen a modo de prueba en la aplicación de extracción de información.

6.7.6 VIRTUALDUB

VirtualDub es una herramienta de código abierto para capturar vídeo y procesarlo. Dispone de funciones muy avanzadas, es capaz de usar plugins para añadir diferentes técnicas de procesado de vídeo, y puede trabajar con cualquier fichero AVI, independientemente del códec que use, mientras esté instalado.

6.7.7 MEDIAINFO

MediaInfo es un analizador de archivos multimedia que proporciona información técnica sobre un archivo de vídeo o audio, tal como los codec utilizados, características de las pistas de vídeo y audio

(bitrate, framerate, frecuencia...), etc. Nos servirá para saber los datos técnicos de un archivo y así por ejemplo conocer qué codecs necesitamos para poder reproducirlo.

7 CONCLUSIONES Y TRABAJO FUTURO

Este último capítulo presenta los resultados obtenidos y los compara con los objetivos planteados inicialmente. Así mismo, ofrece las líneas futuras de trabajo sobre la plataforma desarrollada.

7.1 LOGROS

En esta sección se enumeran los logros obtenidos al finalizar el proyecto. A grandes rasgos, las metas conseguidas en este trabajo son las siguientes:

- Se crea una base de datos audiovisual con un gran número de participantes y una buena calidad tanto de vídeo como de sonido.
- Se crea un nuevo método a seguir dentro de la investigación de los reconocedores del habla. Al combinar varios tipos de información se logra aumentar de manera significativa la calidad de los dispositivos.
- Se desarrolla una aplicación basada en el color y en la iluminación de la imagen, que obtiene las características suficientes para seguir investigando en el campo del movimiento labial.
- Se desarrolla una aplicación basada en sistemas clasificadores de color y posición. Mediante este método se obtiene suficiente información para extraer características gráficas en futuros trabajos.
- Se investigan nuevos métodos de extracción de características mediante clasificadores. Mediante la comparación de sus resultados se extraen conclusiones que pueden resultar útiles en el desarrollo de nuevos proyectos.

7.2 CONCLUSIONES TÉCNICAS

Las conclusiones planteadas a continuación surgen de la confrontación de los objetivos propuestos al principio del proyecto y los resultados obtenidos a raíz de la realización del mismo. Esta técnica se ha efectuado para conseguir una evaluación fiable del sistema.

En las fases preliminares de este proyecto se estableció como meta la combinación de información visual y sonora para mejorar los sistemas de reconocimiento de voz. A medida que se avanzó en la realización del mismo el equipo de investigación optó por terminarlo de manera anticipada debido a la dificultad de ejecución del objetivo. A pesar de no conseguir alcanzar la meta principal, se han logrado ejecutar varias fases que posibilitarán llevarla a cabo en investigaciones futuras.

La base de datos AV-UC3M resulta una herramienta fundamental en el sistema, ya que es la única fuente de información sobre la que se realiza toda la investigación. Esta colección posee una serie de elementos como el número de individuos o las condiciones de iluminación de buena calidad, que hacen de ella una base de datos excelente para este tipo de investigaciones.

La aplicación SIM-RC es capaz de realizar un visionado de características faciales con una gran rapidez, ofreciendo una exposición clara de las distancias de los extremos labiales de los sujetos estudiados. Se

trata de una primera aproximación para la extracción de características labiales. Mediante esta funcionalidad se logrará en futuros trabajos añadir esta información a los reconocedores del habla.

Por lo que respecta al sistema MAT-RP, consigue clasificar y mostrar todas las partes de la cara de los hablantes. Se trata de una herramienta de investigación muy potente, ya que permite la selección de diferentes tipos de clasificadores y a su vez modificar sus propiedades.

Una vez finalizado el sistema SICECAF y después de la obtención de resultados a partir de las pruebas efectuadas se puede afirmar que en su mayor parte los objetivos planteados se han cumplido de manera satisfactoria.

7.3 CONCLUSIONES PERSONALES

La realización de este proyecto de investigación ha supuesto una fuente de aprendizaje continua que me ha dado la posibilidad de formarme en campos tan diversos como:

- **Reconocedores del habla:** al comenzar este trabajo poseía unas nociones muy básicas de su funcionamiento. Finalmente he logrado conocer plenamente su labor gracias al estudio de sus características.
- **Edición de imagen:** para utilizar las imágenes como fuente de información he necesitado aprender nuevos programas de edición como el Photoshop.
- **Formatos de imagen:** he aprendido acerca de la variedad de formatos de almacenamiento de una imagen digital y las partes en las que se divide el archivo interiormente.
- **Edición de vídeo:** para que la base de datos AV-UC3M fuera completamente operativa he necesitado editar sus muestras, con el correspondiente trabajo de autoformación por mi parte.
- **Espacios de color:** he realizado muchas transformaciones de espacio de color a lo largo de todo el proyecto, por este motivo he estudiado a fondo multitud de tipos y sus características.
- **Propiedades del sonido:** para comprender mejor el funcionamiento de los reconocedores ha sido necesario estudiar apropiadamente como se forma el sonido.
- **Clasificación de patrones:** para desarrollar la aplicación MAT-RP he necesitado formarme profundamente sobre clasificadores como pueden ser las Redes Neuronales y las Máquinas de vectores soporte. Además de estas he investigado sobre otras opciones que al final se han terminado descartando.
- **Lenguajes de programación:** en cuanto a la programación, he aprendido nuevos lenguajes como son Simulink y Matlab. Mediante estas herramientas se han desarrollado las aplicaciones que componen el proyecto.
- **Documentación:** he necesitado mejorar mi forma de escribir para ajustarla a los estándares del lenguaje científico. Esta técnica me ha sido muy útil a la hora de llevar a cabo nuevos proyectos fuera del mundo académico.

En conclusión, creo que este proyecto me ha dado la oportunidad de adentrarme en el terreno de la investigación, aportándome nuevas y útiles experiencias para el mundo laboral. En cuanto a la duración

del mismo, ha habido momentos muy duros que he conseguido superar y de los que he obtenido finalmente una buena recompensa.

7.4 TRABAJOS FUTUROS

En esta sección se enumeran los trabajos futuros que se pueden realizar sobre el reconocimiento multimodal del habla:

- Mejora de las tasas de reconocimiento de los reconocedores del habla tradicionales mediante nuevas técnicas algorítmicas.
- Creación de nuevos métodos de identificación facial.
- Mejora de la extracción de características faciales en la imagen.
- Extracción de características labiales en tiempo real.
- Creación de una base de datos que guarde información sobre la correspondencia entre la distancia interlabial y su fonema.
- Sincronización de información entre los elementos audiovisuales.
- Método de toma de decisiones entre los dos tipos de información: visual y auditiva.

8 REFERENCIAS

- **[McCulloch et al, 1943]** McCulloch W.S.; Pitts W. A logical Calculus of Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 1943, No. 5, 115-133
- **[Sutherland 1963]** Sutherland, Ivan Edward (January de 1963). "Sketchpad: A man-machine graphical communication system (courtesy Computer Laboratory, University of Cambridge UCAM-CL-TR-574 September 2003)". Massachusetts Institute of Technology.
- **[McGurk et al, 1976]** H. McGurk and J. MacDonald, *"Hearing lips and seeing voices," Nature*, pp. 746-748, Dec. 1976.
- **[Easton et al, 1982]** R.D. Easton and M. Basala, *"Perceptual dominance during lipreading," Perception and Psychophysics*, vol. 32, pp. 562-570, 1982.
- **[Petajan 1984]** Petajan, E.D. (1984) Automatic lipreading to enhance speech recognition, *Proceedings of the IEEE Communication Society Global Telecommunications Conference, November 26-29*, PDF and eBook Search Engine, Atlanta, Georgia.
- **[Boehm 1986]** Boehm B, A Spiral Model of Software Development and Enhancement, ACM SIGSOFT Software Engineering Notes, ACM, 11(4):14-24, Agosto 1986.
- **[Kass et al, 1988]** Kass, M., Witkin, A., and Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision, 1*(4):321–331.
- **[Kohonen 1988]** Kohonen, T. Self-Organization and Associative Memory, Springer-Verlag, New York, second edition, 1988.
- **[Rabiner 1989]** Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2): 257-286.
- **[Summerfield et al, 1989]** Summerfield, Q., MacLeod, A., McGrath, M., and Brooke, M. (1989). Lips, teeth, and the benefits of lipreading. In Young, A.W. and Ellis, H.D. (Eds.), *Handbook of Research on Face Processing*. Amsterdam, the Netherlands: Elsevier Science Publishers, pp. 223–233.
- **[Lippman 1990]** R. P. Lippmann, "Review of neural networks for speech recognition", Neural Comput., vol. 1, no. 1, pp. 1–38, 1990.
- **[Turk et al, 1991]** Turk, M. and Pentland, A. 1991. Eigenfaces for recognition. *Journal of Cognitive Neuroscience.*
- **[Junqua et al, 1993]** J.C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," J. Accout. Soc. Amer., vol. 43, no. 1, pp. 637-642, 1993.
- **[Rabiner et al, 1993]** Rabiner, L. and Bing-Hwang, J. 1993. *Fundamentals of Speech Recognition*. Prentice-Hall.
- **[Duchnowski et al, 1994]** Duchnowski, P., Meier, U., and Waibel, A. (1994). See me, hear me: Integrating automatic speech recognition and lipreading. *Proc. International Conference on Spoken Language Processing*, Yokohama, Japan, pp. 547–550.

- **[Hermansky et al, 1994]** H. Hermansky, N. Morgan, "RASTA processing of speech", *IEEE Transactions on Speech and Audio Processing,* Vol. 2, No. 4, pp. 578-589 (1994).
- **[Samaria et al, 1994]** Samaria F, Harter AC. Parameterisation of a stochastic model for human face identification. Proceedings of the Second IEEE Workshop on Applications of Computer Vision.1994.
- **[Duchnowski et al, 1994]** P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: Integrating automatic speech recognition and lip-reading," in *Proc. Int. Conf. Spoken Language Processing*, 1994, pp. 547–550.
- **[Silsbee 1994]** P. L. Silsbee, "Sensory integration in audiovisual automatic speech recognition," in *Conf. Rec. 28th Asilomar Conf. Signals, Systems, and Computers*, 1994, pp. 561–565.
- **[Aldrich 1995]** Aldrich C.; Deventer J.S. Comparison of different artificial neural nets for the detection and location of gross errors in process systems, Ind.Eng.Chem.Res., 1995, Vol. 34, 216-224
- **[Hilera 1995]** Hilera J.R.; Martínez V.J. Redes neuronales artificiales. Fundamentos, modelos y aplicaciones. Addison-Wesley Iberoamericana S.A., España, 1995
- **[Vapnik 1995]** V.Ñ. Vapnik. The nature of statistical learning theory. New York: Springer-Verlag, 1995.
- **[Adjoudani et al, 1996]** Adjoudani, A. and Benoît, C. (1996). On the integration of auditory and visual parameters in an HMM-based ASR. In Stork, D.G. and Hennecke, M.E. (Eds.), *Speechreading by Humans and Machines*. Berlin, Germany: Springer, pp. 461–471.
- **[Chandramohan et al, 1996]** Chandramohan, D. and Silsbee, P.L. (1996). A multiple deformable template approach for visual speech recognition. *Proc. International Conference on Spoken Language Processing*, Philadelphia, PA, pp. 50–53.
- **[Adjoudani et al, 1997]** Adjoudani, A., Guiard-Marigny, T., Le Goff, B., Reveret, L., and Benoıt, C. (1997). A multimedia platform for audiovisual speech processing. *Proc. European Conference on Speech Communication and Technology*, Rhodes, Greece, pp. 1671–1674.
- **[Belhumeur et al, 1997]** BELHUMEUR, P.N., HESPANHA, J. P. & KRIEGMAN, D. J. (1997) 'Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection', *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19: 711–20.
- **[Bloch 1997]** Bloch G. Neural intelligent control for a steel plant. IEEE Trans. Neural Networks, 1997, Vol. 8(4), 910-917
- **[Chiou et al, 1997]** Chiou, G. and Hwang, J.-N. (1997). Lipreading from color video. *IEEE Transactions on Image Processing, 6*(8):1192–1195.
- **[Fraser 1997]** FRASER N. 1997. Assessment of interactive systems. In *Handbook of Standards and Resources for Spoken Language Systems*, D. Gibbon, R. Moore, and R. Winski, Eds. Mouton de Gruyter, New York, NY, 564–614.
- **[Altissimi 1998]** Altissimi R. Optimal operation of a separation plant using artificial neural networks. Comp.Chem.Eng., 1998, Vol. 22, 939-942

- **[Cootes et al, 1998]** Cootes, T.F., Edwards, G.J., and Taylor, C.J. (1998). Active appearance models. *Proc. European Conference on Computer Vision*, Freiburg, Germany, pp. 484–498.
- **[Durbin *et al,* 1998]** Durbin, R., Eddy, S., Krogh, A. y Mitchison, G. 1998. *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*, Cambridge University Press: Cambridge.
- **[Joachims, 1998]** T. Joachims, *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. Proceedings of the European Conference on Machine Learning, Springer, 1998.
- **[Martínez et al, 1998]** Martínez, A. & Benavente, R. (1998). The AR face database. Computer Vision Center, Technical Report.
- **[Potamianos et al, 1998]** Potamianos, G. and Graf, H.P. (1998). Discriminative training of HMM stream exponents for audio-visual speech recognition. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Seattle, WA, pp. 3733–3736.
- **[Pressman 1998]** R. S. Pressman. *Ingeniería del software. Un enfoque práctico*. 4ª Edición. McGrawHill (1998).
- **[Wang 1998]** Wang H.; Oh Y.; Yoon E.S. Strategies for modeling and control of nonlinear chemical process using neural networks, Comp.Chem.Eng., 1998, Vol. 22, 823-826.
- **[Aas *et al,* 1999]** Aas, K., Eikvil, L. & Huseby, R.B. 1999. Applications of hidden Markov chains in image analysis. *Pattern Recognition*, 32, pp.703-713.
- **[Joachims et al, 1999]** T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods – Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
- **[Messer et al, 1999]** K. Messer, J. Matas, J. Kittler, J. Luettin and G. Maître. "Xm2vtsdb: The extended m2vts database", Second International Conference on Audio and Video-based Biometric Person Authentication, March 1999.
- **[Morik et al, 1999]** K. Morik, P. Brockhausen, and T. Joachims, *Combining statistical learning with a knowledge-based approach - A case study in intensive care monitoring*. Proc. 16th Int'l Conf. on Machine Learning (ICML-99), 1999.
- **[Senior 1999]** Senior, A.W. (1999). Face and feature finding for a face recognition system. *Proc. International Conference on Audio and Video-based Biometric Person Authentication*, Washington, DC, pp. 154–159.
- **[You et al, 1999]** K. H. You, H. C. Wang, "Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences", *Speech Communication*, Vol. 28, pp. 13-24 (1999)
- **[Dupont et al, 2000]** Dupont, S. and Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia, 2*(3):141–151.
- **[Joachims, 2000]** T. Joachims, *Estimating the Generalization Performance of a SVM Efficiently*. Proceedings of the International Conference on Machine Learning, Morgan Kaufman, 2000.

- **[Matas et al, 2000]** J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F.Smeraldi, J. Bigun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued and E. Mayoraz, "Comparison of face verification results on the XM2VTS database", Proceedings of the 15th International Conference on Pattern Recognition, Barcelona (Spain), vol. 4, September, pp. 858-863, 2000.

- **[Neti et al, 2000]** Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., and Zhou, J. (2000). *Audio-Visual Speech Recognition*. Final Workshop 2000 Report. Baltimore, MD: Center for Language and Speech Processing, the Johns Hopkins University.

- **[Philips 2000]** Phillips, P.J., Moon, H., Rizvi, S.A., and Rauss, P. "The FERET Evaluation Methodology for Face Recognition Algorithms". IEEE PAMI, Vol. 22, p. 1090-1104, 10, 2000.

- **[Rothkrantz et al, 2000]** L. J. Rothkrantz and D. Nollen, "Automatic speech recognition using recurrent neural networks"

- **[Chen 2001]** T. Chen, *"Audiovisual speech processing,"* IEEE Signal Processing Mag., vol. 18, pp. 9–21, Jan. 2001.

- **[Heckmann et al, 2001]** Heckmann, M., Berthommier, F., and Kroschel, K. (2001). A hybrid ANN/HMM audio-visual speech recognition system. *Proc. International Conference on Auditory-Visual Speech Processing*, Aalborg, Denmark, pp. 190–195.

- **[Matthews et al, 2001]** Matthews, I., Potamianos, G., Neti, C., and Luettin, J. (2001). A comparison of model and transform-based visual features for audio-visual LVCSR. *Proc. International Conference on Multimedia and Expo*, Tokyo, Japan.

- **[Potamianos et al, 2001]** Potamianos, G. and Neti, C. (2001). "Automatic speechreading of impaired speech." Proc. International Conference on Auditory-Visual Speech Processing, Aalborg, Denmark, pp. 177–182.

- **[Jiang et al, 2001]** Jiang, H., Soong, F., and Lee, C. (2001). Hierarchical stochastic feature matching for robust speech recognition. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, pp. 217–220.

- **[Elizondo 2002]** Esqueda Elizondo, J.J. (2002)."Matlab e Interfaces Graficas". Documento de internet.

- **[Ganapathiraju 2002]** A. Ganapathiraju, Support vector machines for speech recognition. PhD thesis, 2002. Major Professor-Joseph Picone.

- **[Joachims, 2002]** Thorsten Joachims, *Learning to Classify Text Using Support Vector Machines*. Dissertation, Kluwer, 2002.

- **[McTear 2002]** M. F. McTear, "Spoken dialogue technology: Enabling the conversational user interface," *ACM Comput. Surv.*, vol. 34, pp. 90–169, 2002.

- **[Patterson et al, 2002]** Patterson, E.K., Gurbuz, S., Tufekci, Z., and Gowdy, J.N. (2002). CUAVE: A new audio-visual database for multimodal human-computer interface research. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Orlando, FL, pp. 2017–2020.

- **[Visser *et al,* 2002]** Visser, I., Raijmakers M.E.J. y Molenaar P.C.M. 2000. Confidence intervals for hidden Markov model parameters, *British Journal of Mathematical and Statistical Psychology* 53: 317–327.

- **[Cuevas, 2003]** A. Cuevas: "Gestión del proceso software". Centro de Estudios Ramón Areces, S.A. 2003

- **[Potamianos 2003]** G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, BRecent advances in the automatic recognition of audiovisual speech, Proc. IEEE, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.

- **[Vezhnevets et al, 2003]** V. Vezhnevets, V. Sazonov, A. Andreeva, A survey on pixel-based skin color detection techniques, GRAPHICON03, 2003, pp. 85–92.

- **[Eveno et al, 2004]** N. Eveno, A. Caplier, and P.-Y. Coulon, "Accurate and quasi-automatic Lip tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 706–715, May 2004.

- **[Froba et al, 2004]** B. Froba and A. Ernst, "Face detection with the modified census transform," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, May 2004, pp. 91–96.

- **[Norris et al, 2004]** Norris, D., McQueen, J. M., & Smits, R. (2004). *Shortlist II: A Bayesian model of continuous speech recognition.*

- **[Potamianos 2004]** G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-Visual Speech Processing* (G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, eds.), MIT Press, 2004.

- **[RD1777, 2004]** Ministerio de Economía y Hacienda: "Real Decreto 1777/2004, de 30 de Julio, por el que se aprueba el Reglamento del impuesto de Sociedades. Anexo: Tablas de Coeficientes de Amortización". 2004.

- **[Viola et al, 2004]** Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *Int. J. Comput. Vis.*, *57*(2), 137–154.

- **[Walker et al, 2004]** Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, Joe Woelfel, "Sphinx-4: A Flexible Open Source Framework for Speech Recognition" SMLI TR2004-0811 c2004 SUN MICROSYSTEMS INC.

- **[Citengul et al, 2005]** H. Cetingul, Y. Yemez, E. Erzin, and A. Tekalp, "Robust Lip-motion Features for Speaker Identification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 509–512, 2005.

- **[Nilsson et al, 2005]** M. Nilsson, M. Dahl, and I. Claesson, "The successive mean quantization transform," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2005, vol. 4, pp. 429–432.

- **[Ordoñez 2005]** Revista Digital Universitaria. 10 de mayo 2005, Volumen 5 Número 7, ISSN: 1067-6079. *FORMATOS DE IMAGEN DIGITAL*. Cristian Andrés Ordoñez Santiago. Disponible en: http://www.revista.unam.mx/vol.6/num5/art50/may_art50.pdf

- **[Arsic et al, 2006]** Ivana Arsic, Roger Vilagut, Jean-philippe Thiran, "Automatic Extraction of Geometric Lip Features with Application to Multi-Modal Speaker Identification," icme, pp.161-164, 2006 IEEE International Conference on Multimedia and Expo, 2006

- **[Hazen 2006]** T. Hazen, "Visual model structures and synchrony constraints for audiovisual speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 1082–1089, May 2006.

- **[Potamianos et al, 2006]** G. Potamianos P. Lucey, "Lipreading using profile versus frontalviews," IEEE Multimedia Signal Processing Workshop, pp. 24–28, October 2006.

- **[Carrillo 2007]** Roberto Carrillo Aguilar, "Diseño y manipulación de modelos ocultos de Markov utilizando herramientas HTK. Una tutoría", Ingeniare. Revista chilena de ingeniería, vol. 15 Nº 1, 2007, pp. 18-26

- **[Chitu et al, 2007]** A.G. Chitu, L.J.M. Rothkrantz, J.C. Wojdeł, P. Wiggers, "Comparison Between Different Feature Extraction Techniques for Audio-Visual Speech Recognition", JMUI, pp. 7-20, Springer, 2007.

- **[Iwano et al, 2007]** Audio-visual speech recognition using lip information extracted from side-face images. In EURASIP JASMP, 2007(1):4-4, 2007.

- **[Nilsson et al, 2007]** M. Nilsson, J. Nordberg I. Claesson, "Face Detection Using Local SMQT Features and Split-up Snow Classifier", *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, Vol. 2, pp. II-589-II-592 (2007)

- **[Wang et al, 2007]** Wang, S.L., Lau, W.H., Liew, A.W.C., and Leung, S.H. (2007). Robust lip region segmentation for lip images with complex background. *Pattern Recognition*, 40(12), 3481-3491.

- **[Ahmad et al, 2008]** Ahmad, N., Datta, S., Mulvaney, D., Farooq, O.: A comparison of visual features for audiovisual automatic speech recognition. In: Acoustics 2008, Paris, pp. 6445–6448 (2008). DOI 10.1121/1.2936016.

- **[Gales et al, 2008]** M. Gales and S. Young. The Application of Hidden Markov Models in Speech Recognition. Now Publishers Inc, 2008.

- **[Li et al, 2009]** M. Li and Y. M. Cheung, "Automatic lip localization under face illumination with shadow consideration," *Signal Processing*,vol. 89, no. 12, pp. 2425–2434, 2009.

- **[Apple 2012]** Página web del reconocedor de habla "Siri" de la empresa Apple. Disponible en: http://www.apple.com/iphone/features/siri.html

- **[Google 2012]** Página web del reconocedor del habla de Google. Disponible en: http://support.google.com/chrome/bin/answer.py?hl=es&answer=140789

ANEXO A. GLOSARIO

A lo largo de las siguientes páginas se presentan las definiciones de los conceptos utilizados a lo largo de la memoria.

CLASIFICACIÓN

La clasificación trata de asignar las diferentes partes del vector de características a grupos o clases, basándose en las características extraídas. En esta etapa se usa lo que se conoce como aprendizaje automático, cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender.

CORPUS

Conjunto de datos, textos u otros materiales sobre determinada materia que pueden servir de base para una investigación o trabajo.

FRAME

Se denomina frame en inglés, a un fotograma o cuadro, una imagen particular dentro de una sucesión de imágenes que componen una animación. La continua sucesión de estos fotogramas producen a la vista la sensación de movimiento, fenómeno dado por las pequeñas diferencias que hay entre cada uno de ellos.

PIXEL

Un píxel o pixel, plural píxeles (acrónimo del inglés picture element, "elemento de imagen") es la menor unidad homogénea en color que forma parte de una imagen digital, ya sea esta una fotografía, un fotograma de vídeo o un gráfico.

RECONOCIMIENTO AUTOMÁTICO DEL HABLA

El Reconocimiento Automático del Habla (RAH) o Reconocimiento Automático de Voz es una parte de la Inteligencia Artificial que tiene como objetivo permitir la comunicación hablada entre seres humanos y computadoras electrónicas.

ANEXO B. CORPUS

Alberto García, Luis Puente, Cristina García, Laura Rodelgo, Santiago Ortega, Daniel Sierra, Leticia Robles, José Antonio Cuenca, Alberto Utrero, Javier Usero.

0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor.

Un científico que vivía preocupado por los problemas del mundo, estaba decidido a encontrar las respuestas necesarias para solucionarlos. Por eso, pasaba día tras día en el estudio de su casa en busca de respuestas para sus dudas.

Una tarde, su hijo de cinco años entró en el estudio con la intención de ayudarle a trabajar. El científico, nervioso por la interrupción, le pidió al niño que fuese a jugar a otro sitio. Pero después de comprobar que no le hacía ni caso, pensó en algo que pudiese distraer su atención.

¡Perfecto! Encontró una revista y vio que en una de sus páginas había un mapa del mundo... ¡Justo lo que necesitaba! Arrancó la hoja, recortó el mapa en muchos trozos y, junto con un rollo de celo, se lo dio a su hijo diciendo:

-“Mira hijo, como te gustan tanto los puzles, te voy a dar el mundo en trocitos para que lo arregles sin ayuda de nadie”.

Así, el padre quedó satisfecho y el niño también. El padre porque pensó que el niño tardaría más de una hora en hacerlo. El niño porque creyó que estaba ayudando a su padre. Pero después de unos minutos el niño exclamó:

-“¡Papá, ya!”.

El padre, en un primer momento, no dio crédito a las palabras del niño. Era imposible que, a su edad, hubiera conseguido recomponer un mapa que nunca había visto antes. Desconfiado, el científico levantó la vista del libro que leía, convencido de que vería el resultado desastroso propio de un niño de cinco años. Pero, para su sorpresa, comprobó que el mapa estaba perfectamente reconstruido: cada trocito había sido colocado y pegado en el lugar correspondiente.

Sin salir de su asombro y mirando fijamente el mapa, le dijo al niño:

- “Hijo, si tu no sabías cómo era el mundo, ¿Cómo has podido hacerlo?”

- “¡Muy fácil papá!” – contestó el niño -, cuando arrancaste la hoja de la revista vi que por el otro lado había un hombre. Di la vuelta a los trocitos que me diste y me puse a hacer el puzzle del hombre, que sabía cómo era. Cuando conseguí arreglar el hombre di la vuelta a la hoja y vi que había arreglado el mundo...”

¡Cambia tu corazón y el mundo cambiará!

ANEXO C: TABLA RESUMEN AV-UC3M

Vídeo	Género	Color de pelo	Color de ojos	Barba	Maquillaje	Gafas	Pendientes	Color piel
1	Femenino	Rubio	Marrones	No	Sí	No	Sí	Blanco
2	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco
3	Masculino	Moreno	Marrones	No	No	Sí	No	Blanco
4	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco
5	Masculino	Castaño	Azules	No	No	No	No	Blanco
6	Femenino	Castaño	Azules	No	Sí	No	Sí	Blanco
7	Masculino	Moreno	Marrones	No	No	No	No	Blanco
8	Femenino	Pelirrojo	Marrones	No	Sí	Sí	Sí	Blanco
9	Masculino	Castaño	Marrones	No	No	No	No	Blanco
10	Femenino	Castaño	Marrones	No	No	No	No	Blanco
11	Masculino	Castaño	Marrones	Sí	No	No	Sí	Blanco
12	Femenino	Castaño	Marrones	No	No	Sí	Sí	Blanco
13	Masculino	Moreno	Marrones	No	No	Sí	No	Blanco
14	Femenino	Moreno	Marrones	No	No	Sí	No	Blanco
15	Masculino	Castaño	Marrones	No	No	No	No	Blanco
16	Femenino	Castaño	Marrones	No	No	Sí	Sí	Blanco
17	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco
18	Femenino	Moreno	Marrones	No	No	No	Sí	Blanco
19	Masculino	Moreno	Marrones	Sí	No	Sí	No	Blanco
20	Masculino	Moreno	Marrones	No	No	No	No	Blanco
21	Femenino	Castaño	Marrones	No	No	No	No	Blanco
22	Masculino	Castaño	Marrones	No	No	No	Sí	Blanco
23	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco
24	Femenino	Castaño	Marrones	No	No	No	No	Blanco
25	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco
26	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco
27	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco
28	Femenino	Castaño	Marrones	No	No	No	No	Blanco
29	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco
30	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco
31	Masculino	Castaño	Marrones	Sí	No	No	No	Blanco
32	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco
33	Masculino	Castaño	Marrones	No	No	No	No	Blanco
34	Femenino	Castaño	Marrones	No	Sí	Sí	Sí	Blanco
35	Masculino	Castaño	Marrones	No	No	Sí	No	Blanco
36	Femenino	Castaño	Marrones	No	No	No	No	Blanco
37	Femenino	Castaño	Marrones	No	No	No	No	Blanco
38	Femenino	Moreno	Marrones	No	No	No	No	Blanco
39	Femenino	Rubio	Azules	No	No	No	Sí	Blanco

40	Femenino	Castaño	Marrones	No	No	No	No	Blanco
41	Femenino	Castaño	Marrones	No	No	No	No	Blanco
42	Femenino	Moreno	Azules	No	No	Sí	No	Blanco
43	Femenino	Castaño	Marrones	No	No	Sí	No	Blanco
44	Femenino	Castaño	Marrones	No	No	Sí	Sí	Blanco
45	Masculino	Castaño	Verdes	Sí	No	Sí	No	Blanco
46	Masculino	Castaño	Verdes	No	No	No	No	Blanco
47	Femenino	Moreno	Marrones	No	Sí	No	No	Blanco
48	Femenino	Castaño	Marrones	No	Sí	No	Sí	Blanco
49	Femenino	Castaño	Marrones	No	No	No	Sí	Negro
50	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco
51	Femenino	Castaño	Marrones	No	No	Sí	Sí	Blanco
52	Femenino	Moreno	Marrones	No	No	No	Sí	Blanco
53	Masculino	Castaño	Marrones	No	No	No	No	Blanco
54	Femenino	Moreno	Marrones	No	No	No	Sí	Blanco
55	Masculino	Moreno	Azules	No	No	No	No	Blanco
56	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco
57	Femenino	Castaño	Marrones	No	No	Sí	Sí	Blanco
58	Femenino	Castaño	Marrones	No	Sí	Sí	Sí	Blanco
59	Femenino	Castaño	Marrones	No	Sí	No	Sí	Blanco
60	Masculino	Castaño	Marrones	Sí	No	No	No	Blanco
61	Masculino	Rubio	Azules	No	No	No	No	Blanco
62	Masculino	Moreno	Marrones	No	No	Sí	No	Blanco
63	Femenino	Pelirrojo	Marrones	No	Sí	Sí	Sí	Blanco
64	Femenino	Castaño	Marrones	No	Sí	No	Sí	Blanco
65	Femenino	Rubio	Marrones	No	Sí	Sí	Sí	Blanco
66	Femenino	Castaño	Marrones	No	Sí	No	Sí	Blanco
67	Masculino	Castaño	Verdes	No	No	No	No	Blanco
68	Femenino	Castaño	Marrones	No	Sí	No	Sí	Blanco
69	Masculino	Moreno	Azules	No	No	Sí	No	Blanco
70	Masculino	Moreno	Marrones	Sí	No	No	No	Blanco
71	Femenino	Moreno	Marrones	No	No	Sí	Sí	Blanco
72	Femenino	Castaño	Marrones	No	Sí	Sí	No	Blanco
73	Femenino	Castaño	Marrones	No	No	No	No	Blanco
74	Masculino	Castaño	Marrones	No	No	No	No	Blanco
75	Masculino	Castaño	Marrones	No	No	No	No	Blanco
76	Masculino	Castaño	Marrones	No	No	No	No	Blanco
77	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco
78	Femenino	Moreno	Marrones	No	No	Sí	Sí	Blanco
79	Femenino	Castaño	Marrones	No	Sí	No	Sí	Blanco
80	Masculino	Moreno	Marrones	Sí	No	No	No	Blanco
81	Masculino	Moreno	Marrones	No	No	Sí	No	Blanco
82	Femenino	Castaño	Azules	No	Sí	No	No	Blanco
83	Masculino	Castaño	Marrones	Sí	No	No	No	Blanco

84	Masculino	Castaño	Marrones	No	No	No	Sí	Blanco
85	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco
86	Femenino	Castaño	Marrones	No	Sí	No	Sí	Blanco
87	Masculino	Castaño	Marrones	No	No	No	No	Blanco
88	Femenino	Castaño	Marrones	No	Sí	No	No	Blanco
89	Femenino	Castaño	Marrones	No	Sí	No	Sí	Blanco
90	Masculino	Castaño	Marrones	No	No	No	No	Blanco
91	Femenino	Rubio	Marrones	No	Sí	No	Sí	Blanco
92	Femenino	Castaño	Marrones	No	No	No	No	Blanco
93	Femenino	Castaño	Marrones	No	No	No	No	Blanco
94	Femenino	Castaño	Marrones	No	Sí	No	Sí	Blanco
95	Masculino	Moreno	Marrones	No	No	No	No	Blanco
96	Masculino	Moreno	Marrones	No	No	Sí	No	Blanco
97	Masculino	Castaño	Verdes	No	No	No	No	Blanco
98	Masculino	Moreno	Marrones	Sí	No	No	No	Blanco
99	Femenino	Rubio	Marrones	No	Sí	No	No	Blanco
100	Femenino	Rubio	Marrones	No	Sí	No	No	Blanco
101	Femenino	Castaño	Marrones	No	Sí	No	Sí	Blanco
102	Femenino	Moreno	Marrones	No	No	No	Sí	Blanco
103	Femenino	Rubio	Azules	No	No	No	No	Blanco
104	Femenino	Moreno	Marrones	No	Sí	No	Sí	Blanco
105	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco
106	Masculino	Castaño	Marrones	No	No	No	No	Blanco
107	Masculino	Castaño	Azules	Sí	No	No	No	Blanco
108	Femenino	Castaño	Marrones	No	No	No	Sí	Blanco