Universitat Ramon Llull

TESI DOCTORAL

Títol Síntesis Audiovisual Realista Personalizable

Realitzada per Javier Melenchón Maldonado

en el Centre Enginyeria i Arquitectura La Salle

i en el Departament Comunicacions i Teoria del Senyal

Dirigida per Elisa Martínez Marroquín

C. Claravall, 1-3
08022 Barcelona
Tel. 936 022 200
Fax 936 022 249
E-mail: urlsc@sec.url.es
www.url.es

Resumen

Se presenta un esquema único para la síntesis y análisis audiovisual personalizable realista de secuencias audiovisuales de caras parlantes y secuencias visuales de lengua de signos. En el primer caso, con animación totalmente sincronizada a través de una fuente de texto o voz; en el segundo, utilizando la técnica de deletreo de palabras mediante la mano. Sus posibilidades de personalización facilitan la creación de secuencias audiovisuales por parte de usuarios no expertos. El espectro de aplicaciones posibles de este esquema de síntesis comprende desde la creación de personajes virtuales realistas para interacción natural o vídeo juegos hasta vídeo conferencia de muy bajo ancho de banda y telefonía visual para las personas con problemas de oído, pasando por ofrecer ayuda en la pronunciación y la comunicación a este mismo colectivo.

La información del aspecto del objeto de interés se extrae de una secuencia de vídeo de no más de dos minutos de duración y se almacena en un modelo visual mediante un algoritmo causal de seguimiento y aprendizaje simultáneos. A partir de este modelo se pueden sintetizar toda la variedad de aspectos que presentaba el objeto en la secuencia inicial en cualquier orden y de forma realista gracias a un nuevo algoritmo de interpolación no lineal de alta dimensionalidad y otro de selección de unidades visuales. Esta variedad de aspectos se puede exagerar utilizando el propio algoritmo de selección, enfatizando visualmente la secuencia resultante. Si, además, existe una relación entre la información de audio y vídeo, ésta se puede almacenar en el modelo (ya sea mediante texto o sonido, directamente) para poder sincronizar el proceso de síntesis con información auditiva. Se han realizado diferentes pruebas basadas en la percepción humana para evaluar el nivel de realismo conseguido por el modelo en el proceso de síntesis y se ha analizado el impacto del énfasis visual sobre éste.

El uso de aparatos de adquisición domésticos, la carencia de elementos intrusivos y la inexistencia de restricciones de iluminación, añadidos a la total automatización del proceso de síntesis y la reducida interacción durante el análisis (del orden de unos pocos minutos) persiguen la máxima facilidad de uso posible al obtener secuencias audiovisuales personalizadas.

El sistema permite procesar secuencias largas con un consumo de recursos muy reducido, sobretodo, en lo que al almacenamiento se refiere, gracias al desarrollo de un nuevo procedimiento de cálculo incremental para la descomposición en valores singulares con actualización de la información media. Este procedimiento se complementa con otros

tres: el decremental, el de partición y el de composición.

En el caso del procesamiento de las caras parlantes, se ha establecido una nueva vía para obtener objetivamente grupos de alófonos visualmente similares (o visemas) a partir de habla natural. Además, se definen y estudian los efectos de incertidumbre audiovisual asociada a la predicción de información entre los modos auditivo y visual.

Este trabajo viene acompañado de diferentes aplicaciones que implementan algunas de sus funcionalidades para acercarlo más a la sociedad. Además, ha participado en numerosos proyectos públicos y privados, eventos sociales y científicos y ha aparecido en diferentes medios de comunicación a lo largo de su desarrollo.

Agradecimientos

Informales

Me gustaría ofrecer mi mayor agradecimiento a Laura, por estar siempre ahí, animándome en los momentos bajos, ayudándome a no desfallecer y dándome puntos de vista valiosísimos que, de no haberlos tenido, este trabajo no existiría. Espero poder devolverle algún día toda la ayuda que me ha brindado.

A mis padres y a mi hermana, por ofrecerme su infinito apoyo, ayuda y comprensión en la larga carrera que me propuse seguir desde que acabé la ingeniería superior.

También es mi deseo agradecer el empuje, decisión y motivación que me ha aportado Antonio, ya desde muy niño, a seguir con mi educación y a no tirar la toalla cuando se tienen momentos de baja moral.

A mi directora de tesis, Elisa, por su inestimable implicación, guía y apoyo que me ha prestado en todo momento.

A Borja, Sanchis, Lourdes, Alex, Marc e Ignacio, por su implicación, ayuda e interés mostrados en este trabajo de investigación.

A mis amigos Álvaro, Oriol, Pere, David, Luís y sus parejas y a toda la gente de Ripollet, por sus comentarios y críticas, y por ofrecerme esa distracción que tanto necesita un doctorando en determinados momentos.

A la obra fantástica de J.R.R. Tolkien, por sus valores sobre esfuerzo, amistad y coraje: «Sólo tú puedes decidir qué hacer con el tiempo que se te ha dado» (Tolkien, 1954).

A todo el *Departament de Comunicacions i Teoria del Senyal* de *Enginyeria La Salle*, por dotarme de un entorno de tanta calidad humana, amistosa y profesional.

A todas aquellas personas que han colaborado en las evaluaciones de este trabajo, por ayudarme a superar las barreras que las máquinas aún no son capaces de romper.

Y, finalmente pero no por eso menos importante, a La Salle, por ofrecerme su confianza y la oportunidad de contribuir a la sociedad con esta tesis doctoral.

A todos ellos, gracias.

Formales

En primer lugar, agradecer el reconocimiento que realiza la Fundación EPSON Ibérica al proyecto de tesis asociado a este trabajo, al otorgarle el primer premio Rosina Ribalta al mejor proyecto de tesis doctoral de España y Portugal en su VIII edición en junio del año 2006.

Este trabajo de investigación ha tenido el soporte de diferentes proyectos dotados con financiación pública y privada. Los agradecimientos en este aspecto van dirigidos a los siguientes proyectos:

- Detección de estados incipientes de somnolencia en conductores de vehículos a motor, DPI2002-02279, 2000 a 2005.
- Detector Autónomo de Estados Incipientes de Somnolencia, MCYT. PROFIT. FIT-1101100, 2002
- Personalització LV, PGR-PR2002-03, 2002.
- Desarrollo de un locutor virtual, FIT-150500-2002-410, 2002.
- Locutor Virtual Personalizable, PGR-PR2002-03, 2002 a 2003.
- Smart and Adaptive System-Human Interaction through Multimodal Interfaces (SASHIMI), PGR-PR200302, 2003 a 2004.
- Locutor virtual per interacció natural, FIT-340100-2004-20, 2004.
- Processament Multimodal per interacció natural, PGR-PR200402, 2004 a 2005.
- Videófono, FIT-350300-2004-44, 2004.
- IntegraTV-4all, FIT-350301-2004-2, 2005.
- Síntesis AudioVisual Expresiva (SAVE), TEC2006-08043/TCM, 2006 a 2009.

Índice general

Índice de figuras

Índice de cuadros

Acrónimos

AAM (modelos de apariencia activa)
Active Appearance Models

ANOVA (análisis de varianza)
ANalisis Of VAriance

BAP (parámetros de animación corporal)
Body Animation Parameters

CIV (correspondencia identificador-visema)

CVCV (consonate-vocal-consonante-vocal)

DMOS (puntuación diferencial media de opinión)
Differential Mean Opinion Score

EVD (descomposición en autovalores)
Eigenvalue Decomposition

HMM (modelos ocultos de Markov)
Hidden Markov Models

IRLS (minimización cuadrática ponderada iterativamente)
Iteratively Reweighted Least Squares

KNN (vecino k más cercano)
K-Nearest Neighbour

LAR (ratios del logaritmo de las áreas)
Log Area Ratios

LPC (coeficientes de predicción lineal)
Linear Prediction Coefficients

LSA (lengua de signos americana)

LSAl (lengua de signos alemana)

LSB (lengua de signos británica)

LSE (lengua de signos española)

LSEsl (lengua de signos eslovena)

LSF (frecuencias espectrales lineales)
Linear Spectrum Frequencies

LSFr (lengua de signos francesa)

LSG (lengua de signos griega)

LSH (lengua de signos holandesa)

LSI (lengua de signos irlandesa)

LSJ (lengua de signos japonesa)

LSP (lengua de signos polonesa)

MAP (máximo a posteriori)
Maximum a Posteriori

MFCC (coeficientes cesptrum de escala frecuencial Mel)
Mel Frequency Cepstrum Coefficients

ML (máxima verosimilitud)
Maximum Likelihood

MMM (modelos morfables multidimensionales)
Multidimensional Morphable Models

MMSE (mínimo error cuadrático medio)
Minimum Mean Square Error

MOS (puntuación media de opinión)
Mean Opinion Score

MSE (error cuadrático medio)
Mean Square Error

NMF (factorización de matrices no negativas)
Non-negative Matrix Factorization

PC (ordenador personal)
Personal Computer

PCA (análisis de componentes principales)
Principal Component Analysis

PSF (función de dispersión puntual)
Point Spread Function

PSFAM (modelos de apariencia facial personalizables)
Person Specific Facial Appearance Models

PSNR (relación entre el pico y ruido de la señal)
Peak to Signal Noise Ratio

RC (coeficientes de reflexión)
Reflection Coefficients

RNA (redes neuronales artificiales)

RVI (relación voz-identificador)

RVV (relación voz-visema)

SGPF (síntesis guiada por fonética)

SVD (descomposición en valores singulares)
Singular Value Decomposition

TEAM (traducción de inglés a lengua de signos americana por máquina)
Translation from English to ASL by Machine

TTS (texto a habla)
Text To Speech

Definiciones

actualización columna Tipo de actualización de la SVD incremental, en la que se añade una nueva columna a la matriz de datos supuesta en la iteración anterior.

actualización en bloque Tipo de actualización de la SVD incremental, en la que se añade un conjunto de nuevas columnas a la matriz de datos de la iteración anterior.

alófono Realización sonora de un *fonema*.

análisis Distinción y separación de las partes de un todo hasta llegar a conocer sus principios o elementos.

apariencia clave Imagen asociada a una *unidad visual real* distinguible entre el resto que se encuentra al inicio o fin de las interpolaciones.

apariencia visual Información visual relativa a la textura del objeto de interés en un instante temporal, que se puede obtener a partir de la reconstrucción de una *unidad visual real*.

aprendizaje Conjunto de procesos orientados a construir un *modelo visual* a partir de objetos sin movimiento rígido.

audiovisual Que se refiere conjuntamente al oído y a la vista, o los emplea a la vez.

blanquear Proceso de eliminación de la correlación interdimensional en un conjunto de datos.

caras parlantes O cabezas parlantes, son apariencias de caras con una actitud comunicativa activa. Estas apariencias pueden tener aspecto real o no, se pueden comunicar mediante gestos y movimientos y suelen tener un canal de audio asociado.

carga computacional Cantidad de operaciones a realizar en un proceso.

causal Que no depende de instantes futuros.

cepstrum Codificación de la señal de voz basada en transformaciones homomórficas.

coarticulación visual Apariencias del objeto asociadas a transiciones entre *unidades visuales reales*.

condiciones de iluminación Propiedades específicas relativas a la distribución de luz en un determinado espacio.

conjunto de entrenamiento Conjunto de datos ejemplo de referencia.

conjunto visual Subconjunto de *unidades visuales reales* asociadas al mismo *alófono*.

conversión fonética Proceso que pasa de una especificación fonética de un mensaje a una especificación visual del mismo.

corpus Conjunto lo más extenso y ordenado posible de datos que pueden servir de base a una investigación.

coste computacional Suponiendo la cantidad de operaciones como una función de la dimensión de los datos, el coste computacional se puede traducir como la función dominante cuando la dimensión de los datos tiende a infinito.

cuefrencia Término que se forma al reordenar letras de la palabra "*frecuencia*" y asociado al espacio que queda despues de realizar una antitransformada de fourier a una transformación homomórfica de la transformada de Fourier de una señal de voz.

descomposición QR Factorización de matrices basada en el método de ortogonalización Gram-Schmidt. Se basa en descomponer una matriz en el producto de otras dos, una ortonormal (Q) y otra triangular superior (R).

dinámica visual Información visual referente a los cambios en la *apariencia visual* de un objeto a lo largo del tiempo.

distancia geodésica Longitud del camino más corto entre dos puntos sin salir de un espacio concreto.

eje de visión Línea imaginaria perpendicular a la lente de un sistema de captura de imágenes y que pasa por el centro de la misma.

enfatización visual Exageración de movimiento o cambio de apariencia que experimenta un objeto, que puede ir asociado a una actividad comunicativa.

facilidad de uso Grado de simplicidad en la utilización de algo.

fiabilidad Cualidad del método que ofrece seguridad o buenos resultados.

flexibilidad Cualidad del método susceptible de cambios o variaciones según las circunstancias o necesidades.

flujo óptico El flujo óptico es un campo vectorial que asigna a cada pixel de una imagen las dos componentes de desplazamiento, utilizando la información de intensidad de una secuencia de imágenes. Este campo vectorial corresponde al movimiento aparente en el plano de la imagen.

fondo de la imagen Región de la imagen que no pertenece ningún objeto de interés de la misma.

fonema Unidad mínima con significado usada para distinguir diferentes palabras.

foto realismo Cualidad de imágenes sintéticas que poseen una *apariencia visual* indistinguible respecto a su versión natural.

grafo de coarticulación Grafo que cuantifica las similitudes y diferencias entre todas las *unidades visuales reales*.

imagen alineada Imagen cuyos objetos de interés se encuentran en la misma posición, orientación y escala de los de una de referencia, cuyos únicos cambios respecto a esta última se deben a variaciones de apariencia.

imagen mascara Imagen con valor de 1 en los píxeles pertenecientes a la región especificada por la misma y valor de 0 para el resto.

imagen vectorizada Imagen con las columnas dispuestas una debajo de otra, o las filas una al lado de otra.

información a posteriori Información disponible después de realizar un experimento o una observación.

información a priori Información disponible desde un inicio, típicamente usada para resolver un problema.

información media Media aritmética de un conjunto de datos.

información visual esencial Datos resultantes de aplicar el proceso de *análisis* sobre un *conjunto de entrenamiento*.

modelo acústico Parte del *modelo visual* que contiene la información relativa a la sincronía entre la *apariencia visual* de un objeto y la información auditiva existente.

modelo articulatorio tridimensional Tipo de representación de la información de los objetos de interés de una imagen que recoge información de posición tridimensional sobre puntos concretos, con mayor o menor densidad.

modelo basado en imágenes Tipo de representación de la información de los objetos de interés de una imagen que recoge información de intensidad sobre todos los píxeles de éstos.

modelo de movimiento composicional Representación incremental del movimiento como la concatenación de dos movimientos del mismo tipo.

modelo facial Particularización del *modelo visual* aplicado a una cara humana.

modelo visual Contenedor de la *información visual esencial* que se extrae a partir del *corpus audiovisual*.

orofacial Relativo a la parte bucal de la cara humana.

outlier Muestra que estadísticamente no posee el mismo significado que la mayoría dentro de un conjunto de muestras.

personalizable Que se le puede dotar de características a elegir.

pirámide resolutiva Conjunto de imágenes obtenidas a partir de diferentes diezmos de la imagen original, ordenadas de menor a mayor resolución.

plantilla Imagen que contiene el objeto a localizar en un proceso de *seguimiento*.

problema de optimización Planteamiento que busca encontrar los extremos de una función bajo unas ciertas restricciones.

realismo Cualidad simultánea de *foto realismo* y *vídeo realismo*.

reconstrucción Expresión en base canónica de la proyección de una imagen.

región visual Conjunto conectado de píxeles que cumple una misma propiedad.

síntesis Composición de un todo por la reunión de sus partes.

síntesis guiada por fonética O *SGPF*, creación de secuencias de imágenes a partir de la especificación de una transcripción fonética.

síntesis guiada por voz Creación de secuencias de imágenes a partir de la especificación de una forma de onda correspondiente a un texto oral. También llamada *SGPV*.

señales audiovisuales Magnitud física que se percibe por el oído y la vista y que evoluciona a través del tiempo, formada, típicamente, por imágenes y sonido.

seguimiento En general, acción de ir en busca de algo. En particular y en *visión por ordenador*, proceso de encontrar la localización de un objeto en la escena en cada cuadro de una secuencia de imágenes.

sistema sobredeterminado Sistema de ecuaciones con más condiciones que incógnitas.

sonidos homófenos Representaciones visuales indistinguibles de los fonemas.

subespacio de apariencia Subconjunto de imágenes capaz de representar toda una gama de apariencias con características comunes.

subespacio visual Componente del *modelo visual* que representa la apariencia estática del mismo.

suposición de iluminación constante Propone que la iluminación de un píxel se mantiene igual al cambiar éste de posición.

suposición de subespacio constante Propone que una imagen se puede reconstruir a partir de su proyección sobre un subespacio determinado, el cual no cambia.

SVD compuesta Método de cálculo de la *SVD* a partir de la combinación de dos *SVD*s existentes concatenadas en vertical.

SVD decremental Método de cálculo de la *SVD* a partir de una existente y la supresión de un conjunto de columnas de la matriz original.

SVD incremental Método de cálculo de la *SVD* a partir de una existente y un conjunto de nuevos datos u otra *SVD*.

SVD partida Método de cálculo de la *SVD* a partir de la separación de una o más filas de una *SVD* existente.

tiempo real El que transcurre de forma natural. Un sistema se dice que funciona a tiempo real si su tiempo de respuesta es lo suficientemente corto como para interaccionar con él del mismo modo que se haría con otras personas.

tracto vocal Conjunto de órganos humanos destinados a la producción del habla.

tramas de voz Conjunto de muestras ordenadas correspondientes a una señal de voz.

trayectoria geodésica Camino entre dos puntos dentro de un espacio dado.

trivisemas *Visemas* correspondientes a una serie de tres *alófonos* consecutivos.

unidad visual Señal visual correspondiente a una realización de un *fonema*.

unidad visual real Proyección de una imagen real en el *subespacio de apariencia*. Punto perteneciente a la *dinámica visual*.

unidad visual virtual Proyección de una imagen que no tiene porqué haber sido observada ni pertenecer estrictamente al muestreo existente en la *dinámica visual*.

vídeo realismo Cualidad de las secuencias de imágenes sintéticas que poseen una *dinámica facial* visualmente no distinguible respecto a su versión natural.

vector de apariencia Codificación de una *unidad visual virtual*.

vector de apariencia real Codificación de una *unidad visual real*.

ventaneo Proceso de partición una señal de voz en trozos y multiplicación de éstos por otra señal que se llama ventana.

visema Representación visual de un fonema, visualmente distinguible de otras.

visemas clave *Visemas* asociados a conjuntos de *alófonos* de apariencia visual similar a partir de los cuales se generan transiciones entre ellos. Son las *apariencias clave* de las regiones *orofaciales* del *modelo visual*.

visión por ordenador Subcampo de la inteligencia artificial que persigue el objetivo de programar un ordenador para que pueda "*comprender*" una escena o las características de una imagen.

webcam Instrumento doméstico de adquisición de imágenes, vídeo y sonido que se puede conectar a un PC.

Capítulo 1

Introducción

La presente tesis se enmarca dentro del programa de doctorado en *Tecnologías de la Información y las Comunicaciones y su Gestión* y se ha realizado en el *Grupo de Investigación en Procesamiento Multimodal* de *Enginyeria i Arquitectura La Salle*, perteneciente a la *Universitat Ramon Llull*, bajo la dirección de la doctora *Elisa Martínez Marroquín*.

1.1. Contexto

En los últimos años se han experimentado extraordinarios avances en la tecnología asociada a las comunicaciones y la informática, los cuales han desembocado en un aumento progresivo de la información digital multimedia (perceptible a través de diferentes canales como el visual, auditivo, táctil, ...) disponible (Stephanidis, 1999, André, 2000, Robbe-Reiter et al., 2000). A su vez, este aumento en la variedad de la oferta de información ha provocado un aumento en la demanda de la misma por parte de todos los sectores de la sociedad en general. Debido a la naturaleza digital de la información digital, se necesitan transductores (altavoces, monitores, ...) que permitan convertirla a estímulos que puedan captar los sentidos humanos. La interacción con estos transductores se realiza a través de un ordenador: en el proceso de búsqueda de la información digital deseada siempre hay que enfrentarse ante la interfaz de una unidad de proceso. De hecho, el éxito en la obtención de la información digital buscada depende del nivel de conocimiento de dicha interfaz, lo cual puede devenir en una barrera al acceder a ella, en especial para el colectivo discapacitado.

1.2. Motivación

Esta tesis busca facilitar el acceso a la información digital mediante la creación de interfaces más sencillas e intuitivas de utilizar. Una posible solución es emular el lenguaje natural propio de la comunicación interpersonal, el cual es multimodal y posee una gran capacidad expresiva (André, 2003). Por ejemplo, mirar las noticias por televisión es una

actividad más cómoda que leer un periódico, del mismo modo que hablar con una persona cara a cara (aunque sólo se la represente aproximadamente) mejora la calidad de comunicación y evita los malentendidos que puedan surgir a través del teléfono, la correspondencia o internet (un ejemplo de este último caso se puede consultar en el trabajo de Chandrasiri et al. (2004)). Existen diversas razones que pueden justificar estos hechos, entre las cuales se encuentran las siguientes:

1. Las personas humanas están acostumbradas a ver imágenes desde los pocos meses después de nacer, sin embargo, no aprenden a leer hasta al cabo de unos años. Se podría decir, pues, que las personas tienen mucha más experiencia viendo que leyendo. No es de extrañar, entonces, que sea más fácil ver y escuchar las noticias en boca de alguien, que tener que leerlas de un papel escrito.

2. El cerebro humano posee una región especialmente dedicada a reconocer imágenes faciales (Zhao et al., 2003). Este hecho explica que las personas humanas tengan más facilidad para reconocer caras que otro tipo de objetos. La interacción social que efectúa el ser humano es a través de las personas, las cuales se comunican con muchos más elementos que la voz, como son los gestos, expresiones y movimientos de la cara (Forner, 1999). Es lógico que exista una parte del cerebro especialmente dedicada a este tipo de objetos, los cuales se va a encontrar tan a menudo a lo largo de su existencia.

3. Desde que nace, el ser humano experimenta el habla humana como una actividad bimodal, es decir, que posee dos canales de comunicación asociados (Massaro, 2001, Chen, 2001): el visual y el acústico. Es de sentido común pensar también que la información de ambos canales se complemente mutuamente y vaya unida de alguna manera. De hecho, hace más de cincuenta años que se conoce que la componente visual del habla puede compensar las pérdidas de calidad del canal acústico Sumby y Pollack (1954). Este comportamiento es particularmente muy útil en entornos ruidosos y lo que es más importante, aunque ambos canales sean ambiguos, su aparición conjunta puede ayudar a obtener una interpretación con sentido (Massaro et al., 1999) del mensaje transmitido.

4. Los individuos con discapacidad auditiva adquieren una mayor soltura de comunicación en lengua de signos que leyendo (Huenerfauth, 2006), debido a su inherente dificultad para asociar grafías a sonidos y construir representaciones mentales auditivas de las palabras. Este hecho permite identificar la extendida solución de la subtitulación como no óptima para este colectivo. Ésta debería ser reemplazada por una subtitulación en el lenguaje de signos adecuado a cada región.

1.3. El problema de investigación

Si se pudiera conseguir una interfaz que emulara totalmente la comunicación con otra persona, el ser humano no necesitaría ningún entrenamiento especial para acceder a la preciada información digital. Una forma de aproximarse puede consistir, por ejemplo, en

dotar a los ordenadores de una cara humana capaz de hablar, gesticular y moverse como su contrapartida real, o generar algún tipo de lengua de signos, en el caso del colectivo con dificultades para percibir sonidos.

En el caso de utilizar caras virtuales, éstas reciben el nombre de caras parlantes o cabezas parlantes (Bailly, 2001, Ostermann y Weissenfeld, 2004) y se empezaron a desarrollar a principios de la década de los setenta. Las caras parlantes iniciales eran muñecos animados de pocos polígonos y movimientos limitados y han ido evolucionando hasta llegar a caras de apariencia real (o casi), con gran cantidad de movimientos, expresiones, gestos y posibilidades de personalización, aunque aún no se ha encontrado una solución que ofrezca todas estas características simultáneamente.

La generación de lengua de signos ha seguido un camino paralelo parecido, aunque las primeras pruebas datan de la década de los años ochenta y consistían en animaciones casi abstractas de una mano (Schantz y Poizner, 1982). No obstante, mediante técnicas de captura de movimiento (utilización de *Data Gloves*) y la creciente potencia de los renderizadores en tres dimensiones, se han conseguido modelos tridimensionales mucho más reales en movimiento y apariencia. Sin embargo, el progreso en esta línea revela un avance más lento que en el de las caras parlantes, al conseguir unos niveles de foto realismo más bajos.

1.4. Objetivos

En este contexto, el presente trabajo de investigación pretende definir **un esquema único** para la **síntesis** y **análisis audiovisual personalizable realista** de secuencias audiovisuales de **caras parlantes** y secuencias visuales de **lengua de signos**. En el primer caso, con animación totalmente sincronizada a través de una fuente de texto o voz; en el segundo, utilizando la técnica de deletreo de palabras mediante la mano.

La posibilidad de **personalización** viene asociada a la idea de facilitar al máximo la creación de secuencias audiovisuales, con lo que la utilización de transductores domésticos tanto de entrada como de salida se convierte en una prioridad en esta propuesta, así como la **facilidad de uso** de los métodos propuestos, automatizándolos siempre que es posible.

El abanico de aplicaciones posibles de este esquema de síntesis audiovisual es muy amplio, abarcando desde personajes virtuales realistas para interacción natural y sistemas de diálogo hasta vídeo conferencia de muy bajo ancho de banda y telefonía visual para las personas con problemas de oído, pasando por ofrecer ayuda en la pronunciación a este mismo colectivo.

1.5. Estado de la cuestión

A continuación se muestra un resumen del trabajo desarrollado en los ámbitos de generación de caras parlantes (apartado 1.5.1) y lengua de signos (apartado 1.5.2), a partir

de la cual se identifican las carencias actuales y se definen las aportaciones de esta tesis (ver apartado 1.6) para intentar completar algunas de ellas. Debido a que el cálculo incremental de la descomposición en valores singulares (SVD) ha sido objeto de investigación en esta tesis debido a su estrecha relación con el método de síntesis desarrollado, se adjunta también una revisión del estado de la cuestión sobre el trabajo realizado acerca de esta temática en el apartado 1.5.3.

1.5.1. Caras parlantes

Desde los trabajos iniciales sobre los temas relacionados con la animación facial (Parke, 1972) y la influencia de la información visual sobre la comprensión del habla (McGurk y McDonald, 1976), se ha llevado a cabo una actividad investigadora creciente en el campo de la síntesis de habla audiovisual, considerando diferentes aproximaciones para traducir información fonética o auditiva a secuencias de imágenes de caras que hablan. Todas ellas se pueden agrupar mediante dos aspectos fundamentales:

1. El modelo utilizado para representar las diferentes apariencias faciales, estrechamente relacionado con la manera de animar las caras (ver apartado 1.5.1.1).
2. El modo de controlar la animación, ligado al tipo de información de entrada y su agrupación (ver apartado 1.5.1.2).

Adicionalmente, las diferentes propuestas poseen ciertas características relativas a la facilidad de personalización (apartado 1.5.1.3) y al nivel de realismo (apartado 1.5.1.4) que ofrecen. Además, este último puede ser a nivel estático o de imagen, llamado foto realismo, o a nivel de secuencia, identificado por el término vídeo realismo. En el cuadro 1.1 se puede observar una clasificación resumida de los principales trabajos realizados hasta la actualidad sobre síntesis audiovisual de cabezas parlantes en base al tipo de modelo usado, el modo de controlar la animación, la facilidad de personalización y su foto realismo asociado. No ha sido posible evaluar el vídeo realismo de muchos de ellos, con lo que no se adjunta en la clasificación esta característica.

1.5.1.1. Modelos faciales

Los modelos faciales citados en la bibliografía se pueden clasificar en dos grandes grupos: los basados en imágenes bidimensionales y los articulatorios tridimensionales. Los primeros, Beier y Neely (1992), Ezzat y Poggio (1997), Bregler et al. (1997), Cosatto y Graf (1998), Brooke y Scott (1998), Ezzat et al. (2002), Cosatto (2002) y Cosker et al. (2004), usan un conjunto finito de imágenes de caras para generar la apariencia final, mientras que los segundos, Parke (1972), Platt y Badler (1981), Waters (1987), Lee et al. (1995), Ostermann et al. (1998), Brand (1999), Liu et al. (2001), Revéret et al. (2000), Chen (2001), Pelachaud et al. (2001), Morishima (2001), Hong et al. (2002), Beskow (2003), Blanz et al. (2003), Kuratate (2004), Gutiérrez-Osuna et al. (2005), Cao et al. (2005), Choi y Hwang (2005), Vlasic et al. (2005), Deng y Neumann (2006), Sifakis et al. (2006) y Pei y Zha (2007),

se basan en deformar mallas poligonales tridimensionales para conseguir la animación final. Los modelos basados en imágenes consiguen una apariencia más detallada y precisa que los articulatorios, aunque éstos poseen mayor libertad de movimiento. Con el objetivo de reducir las limitaciones en cada uno de los dos tipos de modelos, se han intentado aplicar técnicas propias del otro. Este es el caso de los modelos articulatorios que proponen el mapeo de texturas reales sobre su malla poligonal (Lee et al., 1995, Kuratate et al., 1997, Ostermann et al., 1998, Brand, 1999, Revéret et al., 2000, Liu et al., 2001, Morishima, 2001, Hong et al., 2002, Blanz et al., 2003, Cao et al., 2005, Choi y Hwang, 2005, Pei y Zha, 2007) para aumentar el detalle y realismo de su apariencia, aunque presentan problemas en la representación de elementos complejos como, por ejemplo, el pelo (Pei y Zha, 2007). Por otro lado, algunos modelos basados en imágenes han intentado incorporar información tridimensional en su proceso de síntesis, como los de Brooke y Scott (1998), Cosatto y Graf (2000) y Theobald et al. (2004), permitiendo una más variedad de movimientos, pero aún lejos del ofrecido por los modelos articulatorios. Recientemente, trabajos como los de Fagel (2004), están mostrando las compatibilidades entre los dos tipos, al crear modelos híbridos con una distribución equilibrada de las características de ambos.

Durante el desarrollo de esta tesis se han propuesto dos modelos faciales basados en imágenes bidimensionales (Melenchón et al., 2002a, 2003a), cuya relación con la misma se puede consultar en el apartado 1.6.1.

Seguidamente se detalla cada uno de los dos tipos de modelos faciales. Se refiere al lector a los trabajos de Ostermann y Weissenfeld (2004) y Radovan y Pretorius (2006) para otros puntos de vista en la clasificación de los principales trabajos de síntesis facial.

Basados en modelos articulatorios 3D La actividad investigadora relacionada con los modelos articulatorios empezó con el trabajo pionero de Parke (1972), un modelo 3D no foto realista que se ha convertido en una referencia clave para todos los investigadores que trabajan en este campo. El modelo de Parke requería grandes cantidades de supervisión experta y no podía ser sintetizado en tiempo real. Desde entonces, han aparecido gran cantidad de propuestas, las cuales se pueden clasificar en dos grupos principales: modelado de la estructura muscular (Platt y Badler, 1981, Waters, 1987, Lee et al., 1995, Morishima, 2001, Pelachaud et al., 2001, Gutiérrez-Osuna et al., 2005, Sifakis et al., 2006) y descripción paramétrica de mallas tridimensionales (Parke, 1972, Beskow, 1995, Ostermann et al., 1998, Revéret et al., 2000, Liu et al., 2001, Chen, 2001, Hong et al., 2002, Blanz et al., 2003, Kuratate, 2004, Cao et al., 2005, Choi y Hwang, 2005, Vlasic et al., 2005, Deng y Neumann, 2006, Pei y Zha, 2007). Estos últimos carecen de la precisión de movimientos de los primeros, aunque se controlan más fácilmente; no obstante, esta diferenciación cada vez es más pequeña gracias a avances recientes en ambas, como el control muscular simplificado a través de fisemas (o configuraciones musculares asociadas a fonemas), presente en la propuesta de Sifakis et al. (2006) y el aumento de los niveles de precisión en trabajos como el de Kuratate (2004) y los que versan sobre el desarrollo de elementos multilineales como el de Vlasic et al. (2005). De todas maneras, el resultado que ofrecen estas técnicas aún posee una apariencia más sintética comparado con el ofrecido por los modelos basados en imágenes bidimensionales (apartado 1.5.1.1), debido a la alta complejidad de generar algunos elementos, como el pelo (Pei y Zha, 2007).

Basados en imágenes bidimensionales Uno de los primeros ejemplos de este tipo de modelos lo constituye el modelo de *morphing* 2D de Beier y Neely (1992), que necesita información de características faciales de alto nivel (como contornos de ojos, por ejemplo) para una generación correcta de transiciones sintéticas. Posibles soluciones a esas dificultades aparecen paralelamente en dos trabajos: el de Ezzat y Poggio (1997), con un método de *morphing* automático, y el de Bregler et al. (1997), mediante la técnica de *Video Re-Write*; no obstante, ambos presentan nuevas dificultades: el *morphing* automático de Ezzat y Poggio (1997) es muy sensible a objetos de nueva aparición (como los dientes al hablar) y no modela la cabeza completamente (sino elementos sueltos, como la boca); por otro lado, la propuesta de Bregler et al. (1997) padece de un coste de almacenamiento en memoria muy alto. Cosatto y Graf (1998) propusieron en 1998 una representación modular de los diferentes elementos faciales para reducir los requisitos de memoria, aunque requiere supervisión manual (al igual que los anteriores) y genera movimientos espasmódicos especialmente cuando la base de datos no es suficientemente grande. Con el objetivo de reducir estos efectos, Ezzat et al. (2002) presentó el uso de los modelos morfables multidimensionales (MMM), necesitando únicamente la especificación de cuatro máscaras en la imagen inicial, lo cual simplificaba, además, el proceso de entrenamiento en gran medida; no obstante, el coste computacional no era despreciable, limitándose la duración de las secuencias de entrenamiento. Para contrarrestar esta desventaja, Chang y Ezzat (2005) redujo el tiempo necesario de entrenamiento del trabajo de Ezzat et al. (2002) a través de técnicas de clonación, aunque a costa de aumentar los requisitos de interacción manual. De todos modos, si la minimización de la intervención manual no es prioritaria, Theobald et al. (2004) propusieron el uso de los modelos de apariencia activa (AAM), de Cootes et al. (1998), para incrementar el control de los movimientos y permitir todo tipo de rotaciones tridimensionales.

Asociada a esta tesis, la propuesta de Melenchón et al. (2002a) presentó el uso de análisis de componentes principales (PCA) para parametrizar las apariencias orofaciales y facilitar el proceso de interpolación de apariencias labiales existentes con el objetivo de reducir los efectos espasmódicos de Cosatto y Graf (1998), aunque la interacción necesaria por parte del usuario para crear una instancia del modelo facial era demasiado grande, al igual que en los casos anteriormente propuestos. Además, de forma paralela a Ezzat et al. (2002), Melenchón et al. (2003a) presentaron una propuesta muy parecida pero basada en los modelos de apariencia facial personalizables (PSFAM) desarrollados por Torre y Black (2002), aunque no era aplicable a secuencias largas debido a sus requisitos de memoria. La diferencia más importante entre estos dos trabajos radica en que mientras que Ezzat et al. (2002) se basa en recrear los flujos ópticos observados, Melenchón et al. (2003a) recrea las imágenes directamente. Adicionalmente, se presentó una nueva propuesta, inicialmente en (Melenchón et al., 2004) y más adelante, de forma extendida, en (Melenchón et al., 2005), que habilitó el entrenamiento con secuencias largas al reducir el coste computacional y los resuisitos de memoria mediante el uso de un nuevo método de cálculo de la SVD incremental con actualización de la media.

1.5.1.2. Control de la síntesis

En un sistema de síntesis de caras parlantes, la entrada consiste en información semántica o auditiva y la salida, en información visual. La información semántica viene representada como una serie de elementos mínimos de significado llamados fonemas, los cuales poseen realizaciones sonoras concretas, o alófonos, en función de los fonemas vecinos.

La información visual relacionada con los fonemas se encuentra en la región de la boca. Esta región presenta una variabilidad de aspecto que se puede agrupar mediante lo que se conoce como visemas (Fisher, 1968) o sonidos homófenos (diferente de homófonos), que son representaciones visuales indistinguibles asociadas a los fonemas. A lo largo de las décadas de los 60, 70 y 80 se realizaron estudios para intentar encontrar un conjunto universal de visemas a partir de la agrupación de los diferentes alófonos (Owens y Blazek, 1985), llegando a un consenso que se ha aplicado en nuevos estándares como el MPEG-4 (Tekalp y Ostermann, 2000).

Por otro lado, la relación entre alófonos y visemas se trata de forma diferente dependiendo del formato de la información de entrada, que puede ser, básicamente, de dos tipos: transcripciones fonéticas y habla humana. En el primer caso, en el que se realiza síntesis guiada por fonética, se suelen utilizar tablas que mapean diferentes alófonos a un conjunto concreto de apariencias orofaciales que los articulan (asociadas a los visemas); en el segundo caso se realiza síntesis guiada por voz, donde las soluciones se basan en la búsqueda de una función que relaciona la información de voz con la facial. El primer tipo se viene utilizando desde el primer modelo propuesto por Parke (1972), mientras que los primeros trabajos acerca del segundo aparecen dos décadas más tarde (Yehia et al., 1998). Esta tardía aparición puede deberse a que los sistemas de síntesis guiada por voz se pueden transformar en sistemas de síntesis guiada por fonética mediante reconocedores de voz (Huang et al., 2001) simplificados, ya que no necesitan distinguir entre todos los alófonos, sino solamente entre visemas. Además, estos procesos de reconocimiento normalmente contemplan unos niveles de precisión sintáctica y semántica mayores que los utilizados en los procesos de síntesis guiada por voz, añadiendo dificultades adicionales innecesarias en la síntesis de información visual. Este razonamiento podría explicar, al menos en parte, porqué ninguno de los principales esquemas de síntesis de cabezas parlantes ha utilizado reconocedores de voz.

Agrupación de alófonos en visemas Una de las primeras reflexiones acerca de la importancia del modo visual del habla la dio Cotton (1935): «*There is an important element of visual hearing in all normal individuals*». La importancia del modo visual fue en aumento hasta que Fisher (1968) definió el concepto de visema, juntamente con un estudio preliminar que sentó las bases de posteriores trabajos para su búsqueda. Binnie et al. (1974) encontraron que el punto de articulación era la característica más discriminativa entre visemas al analizar la pronunciación de diversas consonantes precedidas y seguidas de la vocal [a]. El efecto McGurk, que consiste en la percepción de un tercer mensaje diferente cuando el audio y el vídeo expresan sonidos e imágenes distintas y sincronizadas (por ejemplo, los sonidos [ba], juntamente con la articulación [ga], son percibidos como el

mensaje [da]), fue hallado poco después por McGurk y McDonald (1976). Seguidamente, Walden et al. (1977) encontró que la discriminación de visemas podía variar sustancialmente dependiendo de la persona y de si ésta había sido enseñada o no a distinguirlos. Más adelante, Benguerel y Pichora-Fuller (1982) estudió el agrupamiento visémico en función de los efectos de transición entre sonidos, o efectos de coarticulación, producidos al incluir vocales cerradas y abiertas en el estudio. Todos estos trabajos se basaban en estudios subjetivos (Owens y Blazek, 1985) mediante el uso de personas evaluadoras en las diferentes pruebas que llevaron a cabo. El trabajo de Finn (1986) propuso medidas objetivas, asociadas a diferentes características bucales, para evitar el uso de personas en las evaluaciones. Goldschen (1993) siguió esta línea y encontró una agrupación de visemas similar a las obtenidas previamente. No obstante, las medidas utilizadas fueron de alto nivel (por ejemplo, la obertura de boca y la distancia entre las esquinas de la misma) y no se utilizaba habla natural, sino palabras aisladas o grupos fonéticos sin sentido. A partir de ese punto no se han encontrado avances significativos en esta temática y las nuevas propuestas se basan en los resultados obtenidos por estos trabajos, como, por ejemplo, la agrupación de visemas del MPEG-4 (Tekalp y Ostermann, 2000), basada en el trabajo de Binnie et al. (1974).

Síntesis guiada por fonética Las técnicas de síntesis guiada por fonética se basan en la existencia de una correspondencia entre alófonos y visemas. En la bibliografía se usan dos tipos de correspondencias: un primer tipo que relaciona un solo alófono con el visema asociado (Ezzat y Poggio, 1997, Revéret et al., 2000, Ezzat et al., 2002, Blanz et al., 2003, Theobald et al., 2004, Fagel, 2004, Vlasic et al., 2005, Deng y Neumann, 2006, Sifakis et al., 2006, Pei y Zha, 2007) y otro en el que se relacionan grupos de tres alófonos consecutivos con sus visemas objetivos (obteniendo lo que se conoce como trivisemas) (Bregler et al., 1997, Cosatto y Graf, 1998, Brooke y Scott, 1998, Cosatto, 2002). Los primeros poseen unos requisitos de memoria menores mientras que los segundos poseen un mayor realismo en las transiciones entre visemas (o efectos de coarticulación visual, modelados por Cohen y Massaro (1993)), ya que las tienen implícitamente presentes en los trivisemas. No obstante, los requisitos de memoria de estos últimos se pueden reducir al mejorar la generación de transiciones entre visemas mediante los MMM de Ezzat et al. (2002); al obtener transiciones suficientemente realistas, no se hace necesario reproducir trivisemas enteros.

Los dos modelos asociados a esta tesis (Melenchón et al., 2002a, 2003a) se basan en el primer tipo de correspondencia. No obstante, el segundo modelo ofrece la interpolación no lineal de alta dimensionalidad, la cual generaliza la creación de transiciones y aumenta su realismo de manera similar a como ocurre con el trabajo de Ezzat et al. (2002).

Síntesis guiada por voz La síntesis guiada por voz busca una representación funcional que relaciona la información acústica y visual a lo largo del tiempo, que, en el caso de utilizar un modelo articulatorio (ver apartado 1.5.1.1) es aproximadamente lineal en un 65 % según Yehia et al. (1998). De aquí aparece la necesidad del uso de técnicas no lineales de predicción, para obtener el restante 35 % de relación, al menos en este tipo de modelos. La investigación en este campo ha estado mayoritariamente orientada al uso de modelos ocultos de Markov (HMM) (Agelfors et al., 1999, Brand, 1999, Chen, 2001, Choi y Hwang, 2005) y redes neuronales artificiales (RNA) (Massaro et al., 1999, Morishima, 2001, Hong

et al., 2002). Los HMM son muy sensibles al ruido mientras que las RNA tienen una alta carga computacional en su fase de entrenamiento así como dificultades inherentes para encontrar la mejor arquitectura en cada caso. Para una comparación exhaustiva entre este tipo de técnicas, se refiere al lector al trabajo de Beskow (2003). También han aparecido otras alternativas para evitar los inconvenientes de los HMM y las RNA. Ejemplos de éstas son los mapeos directos entre vectores concatenados de características visuales y acústicas que se pueden encontrar en el trabajo de Cosker et al. (2004) y en el de Gutiérrez-Osuna et al. (2005). El primero utiliza técnicas de predicción lineal local juntamente con la construcción de un modelo combinado de forma, apariencia y voz, obteniendo también resultados correctos para su conjunto de pruebas, aunque el sistema es muy sensible a las condiciones de entorno iniciales. En el segundo caso se utiliza la técnica del vecino k más cercano (KNN) y un modelo 3D articulatorio, proclamando resultados correctos con un bajo coste computacional. Todos estos trabajos utilizan modelos computacionalmente poco costosos y, aunque limitan el rango de sonidos a sintetizar, son capaces de encontrar relaciones correctas entre audio e imagen. No obstante, no existe aún ninguna solución capaz de definir completamente la relación entre los modos auditivo y visual del habla.

Melenchón et al. (2003b) presentó, en el marco de esta tesis doctoral, el uso de la estimación bayesiana combinada con un modelo basado en imágenes, obteniendo resultados limitados con un coste computacional muy reducido, similar al de los trabajos de Cosker et al. (2004) y Gutiérrez-Osuna et al. (2005).

1.5.1.3. Facilidad de personalización

La habilidad de un método para personalizar la apariencia visual de la síntesis también es un aspecto que se ha tenido en cuenta en las aproximaciones presentadas a lo largo de los años. En general, la construcción del modelo facial es tedioso debido a la necesidad de marcar imágenes (Lee et al., 1995, Bregler et al., 1997, Brooke y Scott, 1998, Brand, 1999, Zhang et al., 2004, Cosker et al., 2004, Fagel, 2004, Theobald et al., 2004), supervisar el contenido (Lee et al., 1995, Ezzat y Poggio, 1997, Bregler et al., 1997, Brooke y Scott, 1998, Cosatto y Graf, 1998, Revéret et al., 2000, Morishima, 2001, Chen, 2001, Hong et al., 2002, Zhang et al., 2004, Kuratate, 2004, Cao et al., 2005, Chang y Ezzat, 2005) o, incluso, tener que intervenir manualmente en el proceso de creación del modelo (Parke, 1972, Platt y Badler, 1981, Waters, 1987, Ostermann et al., 1998, Blanz y Vetter, 1999, Pelachaud et al., 2001, Blanz et al., 2003, Cao et al., 2005, Vlasic et al., 2005, Pei y Zha, 2007). Se han propuesto diferentes soluciones para reducir la cantidad de esfuerzo manual en la personalización de la apariencia de las cabezas parlantes:

- Utilizar algún tipo de algoritmo que minimice el número de imágenes y puntos por imagen a marcar (Liu et al., 2001). Aunque se reduce la intervención manual, no se elimina totalmente en estos casos.
- Uso de marcas o aparatos intrusivos (Revéret et al., 2000, Kuratate, 2004, Cao et al., 2005, Deng y Neumann, 2006, Sifakis et al., 2006). Este tipo de soluciones simplifica los procesos de marcado pero impone un tipo de restricciones que puede limitar la libertad de movimientos y gestos del sujeto a entrenar.

- Utilizar algoritmos de adaptación de modelos genéricos como los de Blanz et al. (2003), Cosker et al. (2004), Theobald et al. (2004) y Choi y Hwang (2005). En estos casos no hay que crear el modelo facial cada vez, sino adaptar automáticamente los datos a partir de uno genérico. Por tanto, las ventajas de estos sistemas se aprovechan a partir del segundo modelo creado.
- Utilizar algoritmos de seguimiento automáticos como los de Ezzat et al. (2002). Tienen la ventaja de simplificar enormemente el proceso dada una inicialización, aunque necesitan de ésta y tienen un coste computacional y de memoria elevados. No obstante, han aparecido, recientemente, métodos que reducen estos costes (Lim et al., 2005), basados en el cómputo incremental de la SVD.

A pesar de estas soluciones, no existe aún ningún método que pueda crear de forma completamente automática un modelo facial para generar caras parlantes.

Se debe aclarar que uno de los primeros trabajos asociados a esta tesis (Melenchón et al., 2002a) necesitaba de la marcación de imágenes, obteniendo una pobre facilidad de uso. En la siguiente propuesta, Melenchón et al. (2003a) ofrecía un algoritmo de seguimiento automático que eliminaba la necesidad de marcación anterior. Además, éste fue optimizado mediante un esquema novedoso de cálculo incremental de la SVD con actualización de la información media presentado inicialmente por Melenchón et al. (2004). Adicionalmente, este trabajó se expandió a otros casos de uso, además del incremental, en (Melenchón y Martínez, 2007).

1.5.1.4. Realismo

Los niveles de realismo perseguidos han variado a lo largo de las diferentes propuestas. Existen trabajos que únicamente se han centrado en el vídeo realismo, como los de Parke (1972), Platt y Badler (1981), Waters (1987), Chen (2001), Pelachaud et al. (2001), Beskow (2003), Gutiérrez-Osuna et al. (2005) y Deng y Neumann (2006) mientras que en otros el objetivo principal es el foto realismo, como para Ostermann et al. (1998) y Liu et al. (2001); no obstante, la mayoría de propuestas intentan conseguir ambas características (Lee et al., 1995, Ezzat y Poggio, 1996, Bregler et al., 1997, Cosatto y Graf, 1998, Brooke y Scott, 1998, Brand, 1999, Revéret et al., 2000, Morishima, 2001, Ezzat et al., 2002, Melenchón et al., 2002a, Hong et al., 2002, Beskow, 2003, Blanz et al., 2003, Theobald et al., 2004, Cosker et al., 2004, Fagel, 2004, Kuratate, 2004, Gutiérrez-Osuna et al., 2005, Choi y Hwang, 2005, Cao et al., 2005, Vlasic et al., 2005, Sifakis et al., 2006, Pei y Zha, 2007). Se debe destacar que los modelos propuestos y asociados a esta tesis (Melenchón et al., 2002a, 2003a) han perseguido también el máximo nivel de foto y vídeo realismo, sin priorizar ninguno de los dos.

Los trabajos que han priorizado el vídeo realismo han sido los basados en modelos articulatorios, gracias a la explotación de su más fácil control de movimientos; para contemplar la característica de foto realismo, este tipo de modelos ha incluido un mapeo de texturas (a veces, a costa de sacrificar el vídeo realismo, debido a priorizar otros elementos como la facilidad de personalización (Liu et al., 2001)), ya que la síntesis de elementos

faciales utilizando sólo una malla poligonal no se ha conseguido todavía. Se han llegado a conseguir resultados con altos niveles en ambas características, como en los trabajos de Blanz et al. (2003), Cao et al. (2005) y Sifakis et al. (2006), aunque ninguno de éstos posee un nivel de personalización fácil en absoluto.

Los modelos basados en imágenes poseen un nivel muy alto de foto realismo ya que se basan en reproducir apariencias copiadas de la realidad. No obstante, su punto débil es el vídeo realismo, ya que carecen de métodos directos de control de movimientos de los diferentes elementos faciales. No obstante, algunos de estos modelos han mostrado altos niveles de ambas características (Geiger et al., 2003).

Finalmente, se debe comentar que existe un creciente interés en comparar las diferentes propuestas, aunque aún no se encuentre un conjunto normalizado de pruebas cuantitativas que ayuden en las comparaciones. Hasta el momento, los niveles de realismo de la síntesis audiovisual se miden principalmente en términos subjetivos de inteligibilidad (Ouni et al., 2007) y percepción. Los trabajos de Cosatto (2002), Geiger et al. (2003), Beskow (2003) y Cosker et al. (2005) conforman diferentes ejemplos en esta línea.

1.5.1.5. Énfasis

En este trabajo, se utiliza este concepto relacionándolo con la exageración de movimiento o de cambio de apariencia que experimenta un objeto, normalmente asociado a una actividad comunicativa. Por ejemplo, el movimiento que experimentan los labios de una persona cuando reza o cuando canta a viva voz varía sustancialmente entre un nivel de énfasis bajo y uno alto, respectivamente.

El énfasis visual se puede ver como un tipo de exageración. Según Lasseter (1987), la exageración se puede definir como la acentuación de la esencia de una idea a través del diseño y la acción. En el campo de la síntesis de imágenes, el diseño se refiere al aspecto estático y la acción, al movimiento. La exageración se diferencia de la distorsión en la aceptación o negación de la realidad: la primera la lleva al límite, mientras que la segunda la sobrepasa (Redman, 1984).

En el ámbito de caras parlantes, el énfasis visual se ha utilizado en dos ámbitos principales. Desde el punto de vista de la acción, aparece en el estudio de la relación existente entre el aspecto facial visual y la acentuación cuando se habla, ya desde el trabajo de Ekman (1979). Desde el punto de vista del diseño, se ha utilizado para generar caricaturas de imágenes faciales (Rautek et al., 2006), el cual es un punto que queda un poco alejado del ámbito de esta tesis.

La correlación entre los elementos faciales y el acento («peculiar energía, ritmo o entonación con que el hablante se expresa» (Real Academia Española, 2001)), que se puede ver como un tipo especial de énfasis, la enunció Ekman (1979), al sugerir que existía una relación entre el movimiento de las cejas de una persona y la acentuación en su discurso hablado. A partir de esta idea, se ha ido analizando objetivamente este tipo de relación. Cavé et al. (1996) estudió particularmente las variaciones de frecuencia fundamental y

Trabajo	Modelo	Característ.	Guiado por	Facil. pers.	Foto re.
Parke (1972)	3D	Paramétrico	Animador	Muy tediosa	Nulo
Platt y Badler (1981)	3D	Muscular	Animador	Muy tediosa	Nulo
Waters (1987)	3D	Muscular	Animador	Muy tediosa	Nulo
Beier y Neely (1992)	2D	*Morphing*	Animador	Muy tediosa	Alto
Lee et al. (1995)	3D	Muscular +textura	Animador	Muy tediosa	Medio
Ezzat y Poggio (1997)	2D	*Morphing*	Fonética	Tediosa	Alto
Ostermann et al. (1998)	3D	Paramétrico +textura	Animador	Muy tediosa	Medio
Bregler et al. (1997)	2D	Trivisemas	Fonética trivisemas	Muy tediosa	Alto
Cosatto y Graf (1998)	2D	Modular	Fonética trivisemas	Tediosa	Alto
Brooke y Scott (1998)	2D	Trivisemas +malla	Fonética trivisemas	Muy tediosa	Medio*
Brand (1999)	3D	Paramétrico +textura	Voz HMM	Tediosa	Bajo
Cosatto y Graf (2000)	2D	Modular +planos 3D	Fonética trivisemas	Tediosa	Alto
Liu et al. (2001)	3D	Paramétrico +textura	Animador	Media	Medio
Revéret et al. (2000)	3D	Paramétrico +textura	Fonética	Muy tediosa	Medio
Pelachaud et al. (2001)	3D	Muscular	Animador	Muy tediosa	Nulo
Morishima (2001)	3D	Muscular +textura	Voz RNA	Tediosa	Medio
Chen (2001)	3D	Paramétrico	Voz HMM	Tediosa	Nulo
Melenchón et al. (2002a)	2D	PCA	Fonética	Tediosa	Alto*
Ezzat et al. (2002)	2D	MMM	Fonética	Fácil	Alto
Hong et al. (2002)	3D	Paramétrico +textura	Voz RNA	Tediosa	Medio
Melenchón et al. (2003a)	2D	PSFAM +interpolac.	Fonética	Fácil	Alto
Blanz et al. (2003)	3D	Paramétrico +textura	Fonética	Muy tediosa	Alto
Beskow (2003)	3D	Paramétrico	Voz HMM/RNA	Muy tediosa	Nulo
Melenchón et al. (2003b)	2D	PSFAM	Voz est. bayes.	Fácil	Alto
Theobald et al. (2004)	2D	AAM +Malla 3D	Fonética	Tediosa	Alto
Fagel (2004)	2D/3D	Paramétrico	Fonética	Tediosa	Alto
Kuratate (2004)	3D	Paramétrico +textura	Voz lineal	Muy tediosa	Medio
Cosker et al. (2004)	2D	AAM	Voz lineal	Muy tediosa	Alto
Cao et al. (2005)	3D	Paramétrico +textura	Fonética	Muy tediosa	Alto
Choi y Hwang (2005)	3D	Paramétrico +textura	Voz HMM	Fácil	Medio
Gutiérrez-Osuna et al. (2005)	3D	Muscular	Voz k-*nearest*	Muy tediosa	Nulo
Vlasic et al. (2005)	3D	Paramétrico Multilineal	Fonética	Muy Tediosa	Alto
Deng y Neumann (2006)	3D	Paramétrico +isomap	Fonética	Muy tediosa	Nulo
Sifakis et al. (2006)	3D	Muscular +fisemas	Fonética	Muy tediosa	Alto
Pei y Zha (2007)	3D	Paramétrico +text./isom.	Fonética	Muy tediosa	Alto
Propuesta	2D	PSFAM +enfático	Fonética/Voz est. bayes.	Fácil	Alto

Cuadro 1.1: Comparativa general entre los principales esquemas de síntesis audiovisual de cabezas parlantes hasta mediados de 2007 y la presente propuesta. En cursiva se encuentran las propuestas asociadas a esta tesis. (*: niveles de gris)

el movimiento de cejas. Más tarde, Swerts y Krahmer (2004) concluyó que el acento era mucho más perceptible en el canal visual que en el auditivo, al realizar sus estudios para el idioma francés en particular. Scarborough et al. (2006) concretó un subconjunto de los articuladores más relacionados con la acentuación visual. A su vez, Swerts y Krahmer (2006) estudió el impacto que tienen diferentes zonas faciales (superior, inferior, izquierda y derecha) en la representación de la acentuación visual.

Por otro lado, los movimientos faciales influyen en la percepción del estado anímico y el mensaje transmitido. El trabajo de Pourtois et al. (2002) revela que ciertas maneras de hablar pueden producir sensaciones de más o menos realismo en el mensaje transmitido. Además, se ha mostrado que el estado de ánimo se interpreta de forma más realista con expresiones más exageradas de movimiento que con aquellas de duración más larga (Hill et al., 2005).

1.5.1.6. Aplicaciones comerciales relacionadas

Aunque el nivel de madurez alcanzado en el desarrollo de cabezas parlantes aún no se puede decir que sea alto, ya han empezado a aparecer las primeras soluciones comerciales (Ananova Ltd., 2007, Anthropics Technology Ltd., 2007, Artificial Solutions Iberia S.L., 2007, IKEA Ibérica S.A., 2007, Digimask Ltd., 2007, LifeFX Inc., 2007) que explotan económicamente el concepto. Todas utilizan modelos faciales 3D excepto el ayudante de IKEA Ibérica S.A. (2007) implementado por Artificial Solutions Iberia S.L. (2007), que usa uno bidimensional no foto realista.

Las principales ventajas que ofrecen estas soluciones comerciales se basan en su inmensa facilidad de descarga, instalación y/o uso de los sistemas informáticos asociados. Por otro lado, los resultados que ofrecen estos productos son pobres en cuanto a sincronización labial (nulos en el caso de IKEA Ibérica S.A. (2007) y Artificial Solutions Iberia S.L. (2007)), excepto en el caso de Ananova Ltd. (2007), ya que posee un proceso de supervisión en su animación que garantiza movimientos correctos. Las soluciones propuestas por Anthropics Technology Ltd. (2007) y Digimask Ltd. (2007) obtienen un foto realismo bajo pero permiten una interesante facilidad de personalización gracias a una pequeña intervención manual y al uso de modelos tridimensionales genéricos preestablecidos. LifeFX Inc. (2007) ofrece un mejor foto realismo que los anteriores a costa de limitar las posibilidades de personalización a un conjunto limitado y predefinido de caras. Finalmente, comentar que Ananova Ltd. (2007) y Artificial Solutions Iberia S.L. (2007) no ofrecen ningún tipo de facilidad de personalización del aspecto facial obtenido. En el cuadro 1.2 se resumen las características comentadas en este apartado sobre estos sistemas.

1.5.2. Síntesis de lengua de signos

La síntesis de lengua de signos ha tenido una preocupación creciente en la última década desde que Schantz y Poizner (1982) propusiera la primera versión sintética para la lengua de signos americana (LSA) (Stokoe et al., 1965). Esta versión alcanzaba unos

Producto	Modelo	Sincronía labial	Posibilidad de personalización	Foto realismo
Ananova Ltd. (2007)	3D	Buena	No	Nulo
Anthropics Technology Ltd. (2007)	3D	Pobre	Si	Bajo
Artificial Solutions Iberia S.L. (2007)	3D	Nula	No	Nulo
Digimask Ltd. (2007)	3D	Pobre	Sí	Bajo
IKEA Ibérica S.A. (2007)	2D	Nula	No	Nulo
LifeFX Inc. (2007)	3D	Pobre	Limitada	Medio

Cuadro 1.2: Comparativa general entre los principales productos comerciales sobre síntesis audiovisual de cabezas parlantes hasta principios de 2006.

resultados muy limitados debido a la falta de potencia de cálculo de los sistemas de procesamiento contemporáneos. No fue hasta la década de los años noventa que aparecieron nuevas propuestas, que aprovecharon el aumento de las capacidades de procesamiento de los ordenadores. Estas propuestas se dividen en tres grupos: *i*) la generación de modelos tridimensionales articulatorios que emulan los movimientos de las lenguas de signos mediante diversos métodos de interpolación; *ii*) la generación de modelos tridimensionales con movimientos importados a partir de la captura de movimiento de diferentes actores; *iii*) concatenación de secuencias de vídeo que contienen diferentes glosas (o signos textuales). Éstos últimos presentan el mayor realismo posible (ya que son grabaciones de personas reales), aunque es muy complicado conseguir una animación fluida con ellos y presentan muy poca facilidad para añadir nuevos signos (hay que realizar nuevas grabaciones, y las personas pueden haber cambiado su aspecto). Las características son inversas en el primer grupo, mientras que el segundo representa un punto intermedio entre realismo, fluidez en las transiciones entre secuencias y facilidad para añadir nuevos signos (se refiere al lector al cuadro 1.3 para un resumen de estas características).

Tipo	Realismo	Facilidad de concatenación	Escalabilidad
Articulatorios	+	+++	+++
Captura de movimiento	++	++	++
Vídeos pregrabados	+++	+	+

Cuadro 1.3: Diferentes tipos de generación de lengua de signos y sus características principales.

La evolución de la síntesis de la lengua de signos se encuentra muy ligada al desarrollo de su retórica, con menos de un siglo de vida Speers (2001). Para dar una idea de lo reciente que es este desarrollo, sólo hace falta ver que el nacimiento de la retórica asociada a la expresión oral se remonta a la época clásica griega (Ramírez, 2003). No es de extrañar, entonces, que la lengua de signos evolucione también conjuntamente con su representación artificial. En los siguientes subapartados se ofrece una descripción del estado de la cuestión para cada tipo de síntesis teniendo en cuenta el idioma y las novedades aportadas en cuanto a visualización.

Basada en modelos articulatorios La primera propuesta de lengua de signos sintética la propuso Schantz y Poizner (1982) para la LSA mediante una primitiva animación tridimensional de un brazo esquemático a través de la interpolación de los ángulos de las articulaciones. No se producen avances significativos en este ámbito hasta la década de los noventa. Alonso et al. (1995) propuso la síntesis de una mano tridimensional capaz de deletrear palabras en lengua de signos española (LSE), no obstante, aún no incorporaba información gramatical en la misma. El trabajo de Losson y Vannobel (1998) empezó a incluir aspectos sintácticos y a tratar la lengua de signos francesa (LSFr). Veale et al. (1998), en el marco del proyecto *ZARDOZ*, se centró también en aspectos semánticos y empezó a tratar la LSA, la lengua de signos irlandesa (LSI) y la lengua de signos británica (LSB). Utilizando una especificación de alto nivel para el LSFr, Lebourque y Gibet (1999), consigue, con *GESSYCA*, animar las extremidades anteriores de un avatar virtual de modelado simple mediante la especificación de un conjunto ordenado de cuadro clave. El uso de cinemática inversa se puede observar en la propuesta de Zhao et al. (2000), asociada a la traducción de inglés a lengua de signos americana por máquina (TEAM), que propone el primer sistema de traducción automática de inglés a LSA y modela el cuerpo humano entero de forma aproximada. Grieve-Smith (2001) propone el sistema SignSynth para poder representar la LSA en internet utilizando Web3D a través de la notación *Stokoe* (Stokoe et al., 1965). A su vez, el trabajo de Fabian y Francik (2001) es capaz de generar un traductor de texto a lengua de signos polonesa (LSP) mediante un avatar tridimensional de medio cuerpo y la especificación simbólica dada por Suszczańska y Szmal (2001). El proyecto *E-Sign* (Switerslood et al., 2004), heredero de *ViSiCAST* (Marshall y Sáfár, 2002), utiliza interpolación de cuadros clave mediante la especificación *HamNoSys* (Hanke, 2002) para la LSB, la lengua de signos alemana (LSAl) y la lengua de signos holandesa (LSH) e incluye un procedimiento de incorporación de nuevos signos, aprovechando la escalabilidad de este tipo de síntesis. Esta incorporación se simplifica en gran medida en el posterior trabajo de Irving y Foulds (2005) para LSA, que modela el cuerpo entero en tres dimensiones. Papadogiorgaki et al. (2004) propuso la utilización del estándar H-Anim (2007) para el modelado con detalle del cuerpo entero y una interpolación jerárquica de parámetros de animación corporal (BAP)s definida en el estándar MPEG-4 para la lengua de signos griega (LSG), consiguiendo menos coste y más realismo en los movimientos. El uso de la notación *HamNoSys* de Karpouzis et al. (2007) complementa el trabajo anterior, para conseguir una mejor representación de la LSG.

Basada en captura de movimiento Uno de los primeros trabajos en este ámbito lo representa el de Ohki et al. (1994) para la lengua de signos japonesa (LSJ), que utilizaba un *DataGlove* para la captura del movimiento de las manos y los brazos; a pesar de preferir el realismo de los vídeos, los descartaba debido a las dificultades para realizar transiciones suaves al concatenarlos, las cuales sí puede conseguir en la descripción tridimensional del movimiento mediante interpolación. Speers (2001) presentó el *ASL Workbench*, cuya novedad consistía en una interacción con el usuario para obtener el mejor signo cuando existía una duda razonable entre varios candidatos. No obstante, el trabajo más representativo de este tipo de síntesis lo constituye *ViSiCAST* (Marshall y Sáfár, 2002), definido para la LSB, la LSAl y la LSH. Utiliza las especificaciones de HamNoSys para crear un lenguaje XML propio llamado SiGML, con el cual especifica las animaciones a sintetizar mediante

un modelo que sigue el estándar H-Anim (2007). Este último trabajo permite la creación de signos por parte de diferentes personas para aprovechar la mayor escalabilidad frente a la síntesis basada en vídeos.

Basada en vídeos Existen pocas propuestas en este ámbito, empezando por los diccionarios de Sternberg (1994) para LSA y Jaklič et al. (1995) para lengua de signos eslovena (LSEsl), que representan los sintetizadores más sencillos posibles, basados únicamente en la reproducción de vídeos aislados. Neidle et al. (2001) propone una versión más avanzada de diccionario para LSA mediante la búsqueda compleja de glosas, anotadas automáticamente mediante técnicas de visión por ordenador, pero sin intentar suavizar las transiciones entre secuencias de vídeos, provocando una visualización por partes de una frase. El trabajo de Solina et al. (2001), basado en el diccionario de Jaklič et al. (1995), construye frases a través de la concatenación de vídeos intentando reducir al mínimo posible los saltos entre diferentes vídeos mediante la alineación de la posición de los brazos; no obstante, no es capaz de eliminarlos completamente, provocando animaciones espasmódicas.

1.5.3. Cálculo incremental de la SVD

El primer trabajo que introdujo una propuesta de cálculo incremental de la SVD en visión por ordenador fue Murakami y Kumar (1982). Aunque es un algoritmo eficiente basado en descomposición en autovalores (EVD), sólo puede actualizar una imagen vectorizada por iteración (también llamada actualización columna), no tiene en cuenta la información media y posee una potencial inestabilidad numérica; además, sólo obtiene los vectores singulares izquierdos y los valores singulares. Chandrasekaran et al. (1997) propusieron un algoritmo de actualización más estable basándose en el trabajo de Gu y Eisenstat (1993), en el que se ofrece un ajuste directo de la SVD que incluye los vectores singulares derechos; no obstante, es incapaz de actualizar más de una imagen vectorizada por iteración y tampoco considera la información media. Afortunadamente, Hall et al. (2000) la incluyeron en su esquema de cómputo incremental permitiendo, incluso, actualizaciones de múltiples imágenes vectorizadas, o actualización en bloque; este trabajo, basado en EVD, propone un método para unir y separar conjuntos de puntos expresados en diferentes subespacios, lo cual puede ser interpretado como incrementar y decrementar la SVD. Por desgracia, sólo se ofrecen los vectores singulares izquierdos y los valores singulares. Posteriormente, estos autores presentaron otra aproximación (Hall et al., 2002) basada en una actualización directa de la SVD, obteniendo los vectores singulares derechos con una mejor estabilidad numérica; sin embargo, afirmaron que el cálculo decremental de la SVD era imposible de conseguir de forma cerrada con su formulación. Paralelamente, Brand (2002) propuso un algoritmo muy eficiente y estable de cálculo de la SVD incremental con actualización en bloque e incluso señaló una manera de adaptar el autoespacio definido por la SVD de sistemas no estacionarios por medio de valores singulares decayentes; no obstante, no tuvo en cuenta la información media. El trabajo desarrollado por Skočaj y Leonardis (2003) es muy similar al de Hall et al. (2000): usa EVD y contempla la actualización de la información media, pero añade características robustas, aunque sacrificando la actualización en bloque por una actualización columna. Sin contar el presente trabajo (Melenchón et al.,

2004), una de las propuestas más recientes la representa Lim et al. (2005), presentando una nueva alternativa basada en bidiagonalización R de la SVD Golub y Loan (1996); en ésta se contempla la actualización de la información media, la actualización en bloque y la inclusión de un factor de olvido, basado en el trabajo de Levy y Lindenbaum (2000) y siguiendo la idea expresada por Brand (2002). Este tipo de olvido es de tipo decayente y, al igual que el de Brand (2002), nunca se deja de tener en cuenta en su totalidad. Se refiere al lector al cuadro 1.4 para un resumen de los trabajos descritos. Información adicional sobre la historia inicial de la SVD, se puede consultar en el resumen de Stewart (1993).

Propuesta	Método	Actualización	Media
Murakami y Kumar (1982)	EVD	columna	no
Chandrasekaran et al. (1997)	SVD	columna	no
Hall et al. (2000)	EVD	bloque	si
Hall et al. (2002)	SVD	bloque	si
Brand (2002)	SVD	bloque	no
Skočaj y Leonardis (2003)	EVD	columna	si
Lim et al. (2005)	RSVD	bloque	si
Este trabajo	SVD	bloque	si

Cuadro 1.4: Evolución del cómputo de la SVD de forma incremental en el campo de la visión por ordenador. Se muestran las principales características de los distintos métodos propuestos. Destacar los valores singulares decayentes de Brand (2002) y Lim et al. (2005) y las características robustas de Skočaj y Leonardis (2003)..

1.6. Contribución original

El trabajo realizado se titula **síntesis audiovisual realista personalizable** y versa sobre la **creación artificial de secuencias de imágenes lo más indistinguibles posible de la versión real del objeto** que intentan representar. **Guiadas por información simbólica** en general, pueden ir **sincronizadas con un canal de audio**, si existe, para obtener secuencias audiovisuales. Este canal de audio también puede utilizarse para determinar el orden de la secuencia de imágenes, siempre que exista una relación entre los modos auditivo y visual. El **realismo** perseguido se plantea tanto a nivel **estático** (foto realismo) como **dinámico** (vídeo realismo) y el proceso de **personalización**, **no intrusivo** y **lo más automático posible**. Se habla de crear («producir algo de la nada» (Real Academia Española, 2001)) y no de reproducir («ser copia de un original» (Real Academia Española, 2001)) ya que la mayoría de las imágenes obtenidas no son copia de ninguna imagen original previamente observada.

En el ámbito asociado a las **caras parlantes**, la presente propuesta pretende llenar el hueco que existe en el conjunto de soluciones presentadas hasta el momento al intentar obtener un **alto realismo con una fácil personalización** del aspecto completo de la cara, con la posibilidad de animar cada uno de los elementos faciales de la misma.

En cuanto a la síntesis de **lengua de signos**, la propuesta se sitúa en un **punto intermedio entre la síntesis basada en concatenación de vídeos y la paramétrica**, ya que la síntesis se basa en la creación de imágenes realistas previamente no observadas y parametrizadas en un subespacio de baja dimensionalidad, permitiendo la creación de secuencias de vídeo a partir de la especificación de una serie de imágenes clave e interpolando entre ellas dentro del subespacio definido. Además, la mano también disfruta de la fácil personalización asociada a las caras parlantes.

El objetivo de ofrecer un marco único para estos dos tipos de objetos ha permitido ofrecer una **descripción genérica que se puede aplicar potencialmente a cualquier tipo de objeto**, abarcando desde la animación de rotaciones en botellas hasta la simulación de efectos de iluminación. No obstante, no se plantea como objetivo tratar de evaluar el comportamiento para todos y cada uno de los objetos posibles, centrándose principalmente en la síntesis de caras parlantes y lengua de signos.

La **inclusión de efectos de énfasis visual** en la síntesis de secuencias audiovisuales se presenta como novedad en el sentido de incorporar niveles diferentes de gesticulación sostenida, en especial en el campo de las caras parlantes, permitiendo un aumento en el control y el realismo de éstas según el estilo de comunicación que se desee sintetizar.

Aparte de estas contribuciones, se han producido otras más concretas como el desarrollo de un marco de **cálculo composicional de la SVD** y la obtención de los grupos de visemas personales a partir de habla natural y características de bajo nivel (como la información radiométrica de cada píxel), así como la cuantificación numérica de la bondad de predicción entre los modos visual y auditivo en la comunicación hablada. En esta última se ha encontrado que mejorar el nivel de predicción de un modo empeora el del otro y al revés, hecho que se ha bautizado en este trabajo como **incertidumbre audiovisual**.

1.6.1. Trabajo relacionado

La propuesta utiliza un modelo basado en imágenes bidimensionales y viene respaldada por diferentes publicaciones, en las que se muestra un esquema de cómputo incremental (Melenchón et al., 2004, 2005), extendido en (Melenchón y Martínez, 2007), una extracción personalizada de fonemas visuales (Melenchón et al., 2007) y tres modelos: uno para el movimiento de labios y textura (Melenchón et al., 2002a), otro para la generación de transiciones en las apariencias bucales o coarticulación visual (Melenchón et al., 2003a) y otro para la estimación de apariencias visuales (Melenchón et al., 2003b) a partir de voz humana. No obstante, esta tesis extiende el contenido de estas publicaciones. Así, el modelo labial de Melenchón et al. (2002a) se mejora en esta propuesta mediante el modelado de la cara entera. Este nuevo modelo se construye a partir de una secuencia de vídeo de entrenamiento con una persona pronunciando un conjunto equilibrado de visemas y mostrando diferentes expresiones y gestos. El conjunto de visemas se encuentra mediante el uso de algoritmos genéticos paralelos (Cantu-Paz, 2000) y el modelo facial se obtiene automáticamente siguiendo el algoritmo detallado por Melenchón et al. (2005). Éste mejora la actuación de la anterior versión incluida en el trabajo de Melenchón et al. (2003a) ya que no necesita la secuencia entera para comenzar el proceso, consiguiendo un comportamiento

causal, aparte de una drástica reducción en el consumo de recursos de almacenamiento y tiempo de cálculo. La intervención manual requerida en (Melenchón et al., 2003a) se puede eliminar mediante el uso de técnicas de diferenciación de imágenes, aunque sacrificando parte de su precisión. Por otro lado, se obtiene un conjunto de alófonos junto con su información temporal a través de técnicas de segmentación automática de fonemas (Young et al., 2003) aplicada al canal de audio asociado a la secuencia de entrenamiento. Esta información temporal se utiliza para identificar las imágenes correspondientes a cada sonido. El proceso de síntesis puede estar dirigido por información fonética o voz humana. En el primer caso se ha utilizado una aproximación inicial de los grupos de fonemas visuales existentes en castellano basándose en el trabajo de Cosatto (2002), asociado al idioma inglés, y centrándose finalmente en los resultados obtenidos por Melenchón et al. (2007), que intentan emular de forma objetiva las conclusiones extraídas sobre identificación de visemas (ver apartado 1.5.1.2). Respecto a la síntesis basada en voz, se mejora sustancialmente el realismo de los resultados obtenidos por Melenchón et al. (2003b) al combinar el algoritmo de coarticulación visual presentado en el trabajo de Melenchón et al. (2003a) con el modelo de predicción utilizado por Melenchón et al. (2003b).

1.7. Organización

El lector puede encontrar en estas páginas la descripción de todos los procesos involucrados en el esquema propuesto, desde el análisis hasta la síntesis, así como el modelo visual definido, analizando sus características asociadas en cuanto a comportamiento, realismo y facilidad de uso, principalmente. Su distribución se describe a continuación:

- El modelo visual y los aspectos relativos a la representación de la información antes y después de su creación se muestran en el capítulo 2.
- Las técnicas relativas al análisis de la información para crear modelos visuales se ofrecen en el capítulo 3.
- Las relativas a la síntesis de secuencias audiovisuales realistas se discuten en el capítulo 4.
- Los resultados obtenidos por las técnicas de análisis y síntesis, así como la evaluación de sus características se detalla en el capítulo 5.
- Las conclusiones de esta tesis y las líneas de futuro propuestas para su posible continuación se pueden consultar en el capítulo 6.
- Finalmente, los aspectos tangenciales a la línea de discusión se encuentran en los apéndices. Destacar el apéndice A, relacionado con el impacto científico y social que ha tenido la tesis a lo largo de su desarrollo.

Capítulo 2

Representación de la información

La presente tesis doctoral trata sobre síntesis de información audiovisual, la cual necesita de un análisis de la misma, como se verá en el capítulo 3. Antes de explicar cada uno de los dos procesos, se hace necesario definir la representación concreta que ha de tener esta información audiovisual. De hecho, se entiende por representación a la «figura, imagen o idea que sustituye a la realidad» (Real Academia Española, 2001). En el marco propuesto, la "*realidad*" viene expresada por la propia información audiovisual digitalizada, es decir, secuencias de imágenes y de formas de onda discretas, o señales audiovisuales, almacenadas digitalmente como corpus audiovisual. Por otro lado, la "*figura*" que representa esta realidad se conoce en este trabajo como modelo visual y contiene toda la información audiovisual no redundante, o información visual esencial, presente en el corpus (ver figura 2.1). Se entiende por redundancia a la «cierta repetición de la información contenida en un mensaje, que permite, a pesar de la pérdida de una parte de este, reconstruir su contenido» (Real Academia Española, 2001), por lo que la falta de redundancia en el modelo visual no impide la posterior recuperación de las señales audiovisuales.

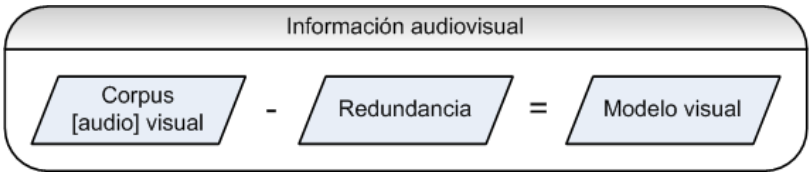


Figura 2.1: La información audiovisual viene dada por las señales audiovisuales, las cuales, pueden representarse mediante un modelo visual extrayéndoles su redundancia.

Dado que la propuesta de síntesis trata los modos visuales y auditivos de forma independiente (ver capítulo 4), se pueden generar secuencias visuales únicamente, sin que haya un canal de sonido asociado. Este hecho provoca que se pueda trabajar sobre un corpus visual, ignorando la parte auditiva. No obstante, el caso contrario no se ofrece ya que en este trabajo la información auditiva juega un papel de apoyo y la visual, el papel principal. A lo largo del capítulo, y por extensión a toda la memoria de tesis, se realizarán las indicaciones oportunas respecto a cuando se excluye y se considera el audio.

2.1. Corpus audiovisual

Antes de comentar el corpus audiovisual empleado, es necesario clarificar el significado de este sintagma. Para hacerlo, hay que tener en cuenta que la expresión consta de dos palabras. El término **corpus** significa «Conjunto lo más extenso y ordenado posible de datos [...] que pueden servir de base a una investigación» (Real Academia Española, 2001). Acompañado del vocablo **audiovisual** («Que se refiere conjuntamente al oído y a la vista, o los emplea a la vez» (Real Academia Española, 2001)), se pretende dar a entender el significado siguiente: **un conjunto lo más extenso y ordenado posible de datos percibidos conjuntamente con el oído y la vista, que pueden servir de base a una investigación**.

Al trabajar con un modelo basado en imágenes se necesita disponer de un conjunto de imágenes de referencia con el fin de poder reproducirlas total, parcial o combinadamente en un instante posterior mediante la construcción del modelo visual. Este hecho implica que cuanto más reales sean las imágenes de referencia, tanto más reales serán las secuencias obtenidas, al tratarse, al fin y al cabo de una reproducción. Los conjuntos de imágenes de referencia más reales posibles consisten en grabaciones de escenas naturales. Por ejemplo, si se desean crear secuencias realistas de caras parlantes mediante el modelo aquí presentado, se deberá partir de grabaciones de personas mientras hablan. Además, dado que el sonido presenta una alta correlación con las observaciones visuales, también es aconsejable incluirla para facilitar tareas posteriores de análisis (ver capítulo 3) y posibilitar el estudio de la relación entre los modos visual y acústico (ver apartado 4.2). La obtención de un corpus audiovisual permite disponer de ambos tipos de información y, además, de forma completamente sincronizada. En el caso en que se deseen crear modelos visuales de objetos sin información auditiva asociada como, por ejemplo, las manos, no será necesario realizar la captura del canal auditivo, obteniendo un corpus visual únicamente.

Las características del corpus se presentan en el apartado 2.1.1 y las guías sobre cómo realizar su registro se encuentran en el apartado 2.1.2. Los apartados 2.1.3 y 2.1.4 muestran detalles asociados a la grabación de dos tipos de objetos concretos: cabezas humanas y objetos no rígidos, respectivamente.

2.1.1. Características del corpus

El corpus a registrar, ya sea visual o audiovisual, sigue el objetivo marcado en el apartado 1.4 de acercar todo lo posible las técnicas desarrolladas a la población. En cuestiones de grabación, es importante que se realice con los dispositivos más comunes y con el menor número de elementos a controlar, en especial, la iluminación.

2.1.1.1. Uso de dispositivos domésticos

El corpus debe poder ser capturado y almacenado con dispositivos domésticos (*web-cam* y ordenador personal (PC)), que son los más extendidos entre la población.

2.1.1.2. Uso de iluminación de ámbito doméstico

El control de la iluminación es complicado y costoso, con lo que permitir una iluminación normal, tanto de interiores, como de exteriores, elimina la necesidad de este control, facilitando la grabación del corpus. Además, este tipo de iluminación no cambia drásticamente en instantes cortos de tiempo. No obstante, hay que tener en cuenta que el resultado de la síntesis compartirá las características concretas de iluminación que se hayan producido en el registro del corpus.

2.1.1.3. Corta duración

La grabación del corpus no puede exigir una duración excesiva, ya que esta provoca cansancio cuando se graba a una persona o parte de ella. Por esta razón, se persigue una duración de la actividad de grabación lo más corta posible.

2.1.2. Registro

El proceso de captura del corpus audiovisual involucra tres elementos básicos y uno opcional. Sin alguno de los elementos básicos no se puede llevar a cabo el registro, mientras que el optativo facilita o habilita procesos posteriores, en función de la información que se desee reproducir en la fase de síntesis (ver capítulo 4). Estos elementos son los siguientes:

- El objeto a capturar.
- La lista de movimientos o cambios de apariencia a efectuar por el objeto.
- Un aparato de adquisición de imágenes.
- Opcionalmente, un aparato de adquisición de audio.

El objeto a capturar debe existir físicamente y debe poder ser perceptible de forma visual. La lista de movimientos y/o cambios de apariencia ha de ser reproducida por el objeto a capturar. Es importante destacar que el resultado de la síntesis tendrá el mismo aspecto que el que se obtenga en el corpus; por ejemplo, si se realizan grabaciones de una persona hablando que tiene un aspecto dormido, la síntesis será de esa persona hablando con el mismo aspecto poco despierto. El aparato de adquisición de imágenes actúa como muestreador del subespacio de apariencia observado; de este modo, la tasa de imágenes por segundo marcará el nivel de detalle con el que el subespacio de apariencia quedará construido. El aparato de adquisición de audio será de utilidad cuando, en la síntesis, sea necesario establecer una relación sincrónica entre imágenes y sonidos.

Recordar que se desea que la captura se pueda realizar a nivel doméstico. Esta premisa básica provoca la aparición de dos requisitos principales, que marcarán buena parte de las líneas de investigación de la tesis:

- La adquisición se realiza utilizando una *webcam* doméstica conectada a un PC estándar (el uso de cámaras profesionales puede implicar el uso de sistemas de iluminación extraordinarios, sobretodo para capturas a alta velocidad).

- Las restricciones en cuanto a las condiciones de iluminación son prácticamente inexistentes, puesto que ha de servir cualquier tipo de iluminación común de interior y exterior.

La utilización de una *webcam* viene motivada por el enfoque doméstico con el que se desea dotar al método desarrollado. Siguiendo con el objetivo de fácil personalización marcado en el apartado 1.6, cualquier persona ha de ser capaz de poder construir el modelo visual que desee, con lo que los recursos necesarios para ello no deben ser fuera de lo comúnmente disponible; puesto que hoy en día los aparatos de adquisición de vídeo más comunes son las cámaras web, se decide obtener la información visual del corpus usando uno de estos aparatos. De este modo, se consiguen dos efectos: *i)* desarrollo y evaluación de los métodos de procesamiento visual con el tipo de datos real a utilizar en un futuro; *ii)* obtención de las primeras instancias de los modelos visuales con datos reales. Finalmente, el método de análisis a desarrollar no ha de requerir condiciones de iluminación especiales puesto que el uso del sistema en un ambiente doméstico únicamente puede llegar a permitir cambios limitados y poco controlados de las mismas.

2.1.2.1. Algoritmo de registro del corpus

El corpus audiovisual a registrar debe seguir un esquema como el de la figura 2.2. En los apartados 2.1.3 y 2.1.4 se muestran dos ejemplos de su aplicación. El primer paso del algoritmo hace referencia a la especificación de las apariencias clave que se deben obtener del objeto, es decir, al conjunto de configuraciones distinguibles entre ellas. A partir de éste, se genera una descripción que produzca una secuencia de transiciones entre todas sus apariencias clave. Finalmente se realiza la grabación del objeto siguiendo la descripción anterior y se obtiene el corpus audiovisual.

```
[C] = Obtención_Corpus(E,B)

    Entrada:
    E: Especificaciones del corpus
    B: Objeto a registrar
    Salida:
    C: Corpus audiovisual asociado al objeto

    1. Identificación de las apariencias clave de E
    2. Obtención de la descripción D del corpus a partir de E
    3. Grabación de B siguiendo D, obteniendo C
```

Figura 2.2: Algoritmo de obtención de corpus audiovisuales.

2.1.3. Grabación de una cara parlante

Cuando el corpus audiovisual describe la apariencia visual de la cara de una persona se debe cumplir adicionalmente que:

- En cada secuencia sólo debe aparecer la cabeza y la parte superior del tronco de una persona, ocupando la mayor parte de la región de las imágenes (más de la mitad).
- La secuencia puede contener discursos y/o gesticulación de la persona que aparece. Los discursos pueden contener información con y/o sin sentido.
- Mientras habla, la cara de la persona sólo puede realizar rotaciones alrededor del eje de visión.
- Cada secuencia debe estar almacenada en un archivo de vídeo, y si es necesaria, con la información auditiva de calidad arbitraria debidamente sincronizada.

Al describir la cara de una persona, se desea obtener la resolución lo más alta posible en la región facial. El análisis debe recoger la apariencia de todos los elementos faciales usados en la comunicación interpersonal. La información sin sentido se incluye para ayudar en el proceso de vocalización por parte del actor. Es importante destacar que el proceso de análisis propuesto supone que la cara es plana mientras habla (ver apartado 3.1.2.2). Esta asunción sólo es válida si la cabeza no asiente ni niega mientras realiza dicha acción. El archivo que contiene la información audiovisual puede provenir de una *webcam* cualquiera; este hecho implica que se debe poder trabajar con diferentes calidades de vídeo y audio. Este último hecho no implica que la calidad final de la síntesis sea constante, sino que dependerá directamente de la calidad del corpus capturado. En el caso en que se requiera un corpus audiovisual sólo para investigar, se puede, adicionalmente, fijar una calidad más alta que la doméstica para el vídeo (720x576 o 100 imágenes por segundo, por ejemplo) y el sonido (como 48KHz y 24 bits por muestra).

2.1.3.1. Variabilidad oral

La cara humana ofrece una gran cantidad de información comunicativa, tanto verbal como no verbal. Este subapartado describe cómo capturar, de forma general y en un corpus audiovisual, toda la variabilidad visual asociada a la comunicación verbal en un tiempo reducido para facilitar su proceso de grabación.

Dado que el elemento de mínimo significado en la comunicación verbal es el fonema, se trabaja sobre un conjunto de fonemas. El idioma escogido es el castellano, debido, en parte, a su bajo número de fonemas, en concreto 24 (Ríos, 1999). No obstante, la información no se transmite físicamente a través de fonemas, que son entidades abstractas, sino a través de sus representaciones físicas: los alófonos, clasificados según características de sonoridad, punto y modo de articulación. En este trabajo se utiliza una parte principal del vasto conjunto de alófonos existentes en castellano (ver cuadro 2.1 y cuadro 2.2).

Aunque los alófonos son «cada una de las variantes que se dan en la pronunciación de un mismo fonema» (Real Academia Española, 2001), la pronunciación repetida de un mismo alófono produce señales sonoras diferentes, aunque perceptualmente indistinguibles entre ellas. Desde un punto de vista más matemático, se podría asociar el concepto de proceso estocástico al alófono y el de instancias concretas a cada una de las señales sonoras producidas al pronunciarlo. Así, cada alófono posee un conjunto de instancias sonoras posibles perceptualmente indistinguibles.

	Anterior	Central	Posterior
cerrada	[i]/[ɪ]		[u]/[ʊ]
media	[e]		[o]
abierta		[a]	

Cuadro 2.1: Clasificación fonética de los alófonos vocálicos / semivocálicos en castellano contemplados en la grabación del corpus audiovisual.

	Bil		Lab		Int		Den		Alv		Pal		Vel	
	Sd	Sn	Sd	Sn	Sd	Sn	Sd	Sn	Sd	Sn	Sd	Sn	Sd	Sn
Oc	[p]	[b]					[t]	[d]					[k]	[g]
Ap		[β]						[ð]						[ɣ]
Fr			[f]		[θ]				[s]	[z]		[ʝ]	[x]	
Na		[m]		[ɱ]						[n]		[ɲ]		[ŋ]
La										[l]		[ʎ]		
Af											[t͡ʃ]			
VS										[ɾ]				
VM										[r]				

Cuadro 2.2: Clasificación fonética de los alófonos consonánticos en castellano contemplados en la grabación del corpus audiovisual. Las abreviaciones indican el modo, punto de articulación y sonoridad del alófono. *Oc*: oclusiva, *Ap*: aproximante oclusiva, *Fr*: fricativa, *Na*: nasal, *La*: lateral, *Af*: africada, *VS*: vibrante simple, *VM*: vibrante múltiple; *Bil.*: bilabial, *Lab.*: labiodental, *Int.*: interdental, *Den.*: dental, *Alv.*: alveolar, *Pal.*: palatal, *Vel.*: velar; *Sd*: sorda; *Sn*: sonora.

El canal auditivo no es el único que se percibe en una comunicación verbal, ya que también existe el visual. Este hecho implica que la percepción del mensaje también debería poderse realizar mediante este canal. En este caso, la entidad abstracta de fonema también posee una representación análoga a la de alófono, que es su representación física visual, conocida desde el año 1968 como **visema** (Fisher, 1968), también llamado anteriormente sonido homófeno o fonema visual. En este caso, la pronunciación, o mejor dicho, la articulación repetida de un mismo visema produce señales visuales diferentes, aunque perceptualmente indistinguibles.

La relación que existe entre alófonos y visemas debería ser idealmente de *uno a uno*, pero, debido a la existencia de oclusiones en el canal perceptivo visual, existe una menor cantidad de visemas, ya que un subconjunto de ellos contienen características visuales ocluidas y, por tanto, no perceptibles, con lo que no pueden usarse para establecer

diferencias entre ellos. Un claro ejemplo es la sonoridad, cuya diferencia es la vibración de las cuerdas vocales, que es un aspecto visualmente no perceptible. No obstante se usan diferentes relaciones entre alófonos y visemas en la bibliografía: *muchos a uno* (Gutiérrez-Osuna et al., 2005) o *muchos a muchos* (Brand, 1999). Esta última opción realiza, en el fondo, dos suposiciones:

1. Los alófonos difícilmente distinguibles se pueden clasificar por grupos de alófonos parecidos. Igualmente, los visemas parecidos se pueden unir en grupos de visemas.

2. Hay ocasiones en que alófonos difícilmente distinguibles de forma auditiva corresponden a visemas fácilmente diferenciables. Esto se traduce en que los grupos de visemas no son simplemente conglomerados de grupos de alófonos. Por ejemplo, y dependiendo del grado de parecido utilizado, los alófonos [m] y [n] pueden llegar a formar parte de un grupo de alófonos porque son acústicamente similares, pero difícilmente formarán parte del mismo grupo de visemas, ya que visualmente son muy diferentes (uno muestra la boca cerrada y el otro, abierta).

En este trabajo se toma la segunda suposición para mantener la generalidad, ya que se puede interpretar que la primera es un caso particular de la segunda cuando los grupos de alófonos están formados únicamente por un alófono cada uno, comportamiento que puede estar asociado a una pronunciación excelente sin ningún tipo de ruido, por ejemplo. La relación existente entre fonemas, alófonos, visemas y grupos de estos dos últimos se puede contemplar en la figura 2.3.

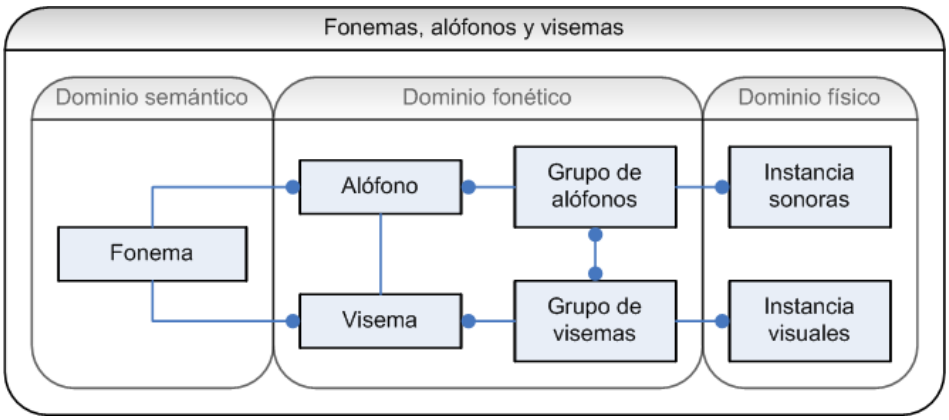


Figura 2.3: Relación entre fonemas, alófonos, visemas, grupos de alófonos, grupos de visemas e instancias sonoras y visuales. El diagrama se divide en tres niveles, desde abstracto a físico, pasando por un dominio fonético intermedio. La cardinalidad de *muchos* se representa con una terminación en círculo, mientras que la de *uno* se representa sin terminación; por ejemplo, un fonema está asociado a varios alófonos.

Las primeras agrupaciones de visemas aparecieron en la década de los setenta (Binnie et al., 1974) y ochenta (Benguerel y Pichora-Fuller, 1982). De hecho, se mostró que se podían obtener diferentes agrupamientos estableciendo diferentes umbrales de similaridad entre visemas (Summerfield, 1987). Muchos de los trabajos de investigación actuales se

basan en estos estudios preliminares; por ejemplo, el estándar MPEG-4 (Tekalp y Ostermann, 2000) ofrece una especificación de visemas que se basa en la ofrecida por Binnie et al. (1974) (ver cuadro 2.3).

Grupos de visemas castellanos siguiendo MPEG-4			
1	[p][b][m][β]	8	[n][l]
2	[f][ɱ]	9	[ɾ][r]
3	[θ][ð]	10	[a]
4	[t][d]	11	[e]
5	[k][g][ɣ][x][ŋ]	12	[i][ɪ]
6	[ʝ][t͡ʃ][ɲ][ʎ]	13	[o]
7	[s][z]	14	[u][ʊ]

Cuadro 2.3: Grupos de visemas castellanos adaptando la clasificación dada por MPEG-4. El grupo 6, todo y corresponder a un punto de articulación palatal, en inglés tiene un aspecto diferente al castellano, sobretodo en la forma de los labios con el alófono [ʃ].

Partiendo de las agrupaciones de visemas anteriores (un resumen de los mismos se puede encontrar en Owens y Blazek (1985)), de las consideraciones dadas en Summerfield (1987) y siguiendo el ejemplo preliminar de Cosatto (2002), se propone una clasificación genérica preliminar en seis niveles para los visemas asociados a la lengua castellana que se puede consultar en el cuadro 2.4. En el primer nivel aparecen todos los alófonos por separado sin ningún tipo de agrupamiento; seguidamente, en el nivel dos, las consonantes pierden su distinción por sonoridad; en el nivel tres, las consonantes se agrupan, además, por punto de articulación, dejando aparte las aproximantes y las líquidas, y las semivocales se agrupan con la vocal más cercana; en el cuarto nivel, se agrupan las aproximantes con sus homólogas sonoras y las líquidas con las alveolares; el penúltimo nivel representa un primer estado de compactación ya que se agrupan los puntos de articulación internos (alveolares, dentales y palatales por una parte y el dental con el interdental por otra, al tener un aspecto muy parecido en habla continua); finalmente, el último nivel agrupa todos los puntos de articulación internos con las vocales anteriores, así como los puntos de articulación externos con el silencio, dejando una posición de boca abierta, otra intermedia, otra casi cerrada y otra cerrada. El rasgo más diferenciador entre visemas se ha tomado como el punto de articulación, teniendo en cuenta la consideración de que los puntos de articulación externos son más fácilmente distinguibles que los internos (Summerfield, 1987). Se puede ver que el nivel cuarto ofrece una clasificación parecida a la definida en el cuadro 2.3, aunque un poco más simplificada. Más adelante (ver apartado 4.2.2) se ofrece una propuesta para obtener los grupos de visemas propios de cada persona.

2.1.3.2. Descripción óptima

Si se desea obtener una descripción de todos los visemas de una persona, no es suficiente con observar uno o varios ejemplos de cada uno de ellos. Las apariencias que puede adoptar una boca al hablar no están únicamente limitadas a los visemas, sino que

Nivel 1 (33)	Nivel 2 (29)	Nivel 3 (17)	Nivel 4 (12)	Nivel 5 (8)	Nivel 6 (4)
[_]	[_]	[_]	[_]	[_]	*SILENCIO*
[p]	*P*	*P*	*P*	*P*	
[b]					
[m]	[m]				
[β]	[β]	[β]			
[f]	[f]	*F*	*F*	*F*	
[ɱ]	[ɱ]				
[a]	[a]	[a]	[a]	[a]	*A*
[u]	[u]	*U*	*U*	*U*	*U*
[ʊ]	[ʊ]				
[o]	[o]	[o]	[o]		
[i]	[i]	*I*	*I*	*I*	*I*
[ɪ]	[ɪ]				
[e]	[e]	[e]	[e]		
[t]	*T*	*T*	*T*	*T*	
[d]					
[ð]	[ð]	[ð]			
[θ]	[θ]	[θ]			
[s]	*S*	*S*	*S*	*S*	
[z]					
[n]	[n]				
[l]	[l]				
[r]	[r]	*R*			
[ɾ]	[ɾ]				
[t͡ʃ]	[t͡ʃ]	*CH*	*CH*		
[ɲ]	[ɲ]				
[ʎ]	[ʎ]				
[j]	[j]				
[k]	*K*	*K*	*K*		
[g]					
[x]	[x]				
[ŋ]	[ŋ]				
[ɣ]	[ɣ]	[ɣ]			

Cuadro 2.4: Conjunto de alófonos en castellano contemplados en la grabación del corpus audiovisual. El número entre paréntesis indica la cantidad de grupos de visemas que hay en ese nivel jerárquico. Los símbolos en mayúsculas indican agrupaciones de los alófonos de la columna anterior que se encuentren desde su misma fila hasta la fila del siguiente símbolo de la misma columna.

existe un amplio abanico resultante de todas las transiciones que pueden suceder entre ellos (ver figura 2.4), y que reciben el nombre de efectos de coarticulación visual.

Figura 2.4: Secuencia real con una transición entre dos alófonos ([β] y [a]), ejemplificando el concepto de coarticulación visual

Como se puede ver en el apartado 4.1, la calidad al reproducir artificialmente un efecto de coarticulación visual es directamente proporcional a la cantidad de transiciones reales contenidas en el corpus audiovisual. Siguiendo esta directiva, es necesario efectuar grabaciones en las que se observen todo tipo de transiciones entre visemas. Un modo de realizarlo puede ser mediante la reproducción de un conjunto de frases como el del cuadro 2.5. Dado que este conjunto tiene una distribución de alófonos similar a la propia del castellano, se dice que es un conjunto de frases balanceado.

No obstante, este tipo de corpus puede resultar demasiado largo y cansado de pronunciar. Se desea observar la máxima variabilidad posible y minimizar la duración del proceso, teniendo en cuenta que las grabaciones serán realizadas por locutores no expertos. Para ello, inicialmente se escogen los visemas definidos en el nivel 4 del cuadro 2.4, que contiene información únicamente sobre el punto de articulación, que es la característica visual más relevante (Summerfield, 1987). De este modo, se reduce drásticamente el número de combinaciones y la duración de las grabaciones a realizar. Utilizando esta agrupación, las coarticulaciones visuales podrán ser generadas a partir de las transiciones entre el silencio, las cinco vocales y los puntos de articulación labial, labio-dental, interdental, alveolar, palatal y velar. En el cuadro 2.6 se muestra la elección de los alófonos específicos que representan cada visema. En particular se han elegido los que no son oclusivos para facilitar la observación del propio visema, dado que los oclusivos son muy cortos.

Los alófonos relacionados con los visemas escogidos (ver cuadro 2.6) se deben agrupar para poder obtener las coarticulaciones visuales deseadas. La agrupación de alófonos forma palabras, que deben ser pronunciadas por un locutor no experto. Es deseable que las palabras tengan un significado nulo, para forzar su pronunciación y obtener, así, una mejor observación de los efectos de coarticulación. Para simplificar la tarea del proceso de registro de estas grabaciones sin sentido, facilitando al locutor el proceso de lectura y pronunciación de las palabras, se determina que éstas sean todas de dos sílabas del estilo consonante-vocal, generando palabras que siguen el modelo de consonate-vocal-consonante-vocal (CVCV).

Finalmente, para poder observar todas las posibles señales visuales (tanto los visemas como sus transiciones), se determina que el conjunto de palabras sea equilibrado en

Frases fonéticamente equilibradas
¿Y para qué sirve un archivo?
La existencia de Internet ha provocado una considerable atención sobre los mecanismos de difusión y análisis de la información, y no sólo los que tienen que ver con tecnologías avanzadas.
Primero, por la pretensión de que la Internet, está ahí precisamente para eso: para facilitar el intercambio de informaciones y en último extremo la creación de conocimiento, entre muy distintos colectivos.
El sistema no cumple exactamente lo que promete, pero por cuestiones prácticas: no hay medios humanos y técnicos para guardar la inmensa masa de publicaciones de todo tipo que llegan en virtud del depósito legal.
¿Cómo sabemos qué es lo que nos hará falta, de entre toda la documentación del pasado?
La culminación de esa tendencia la encontraríamos en la mencionada obra, que reconoce la deseabilidad de una especie de "*mundo al revés*" en el que "*quien mandara en la universidad fuera el bibliotecario, y el profesor fuera un acólito suyo*".
¿Por qué no podría estarse reproduciendo, con buenos frutos, una nueva oralidad? Precisamente los estudios sobre la transmisión oral, no regulada, han crecido explosivamente en los últimos tiempos, como si quisiéramos cerciorarnos de que tales cosas son posibles, y no necesariamente malas.
En último extremo, la historia nos ofrece numerosos casos en los que lo importante no es que una atribución sea inexacta o real sino el hecho de que se produzca la transmisión de una obra.
Yo soy el que escribió esta obra, y esta otra.
Normalmente la misma materialidad de la obra enuncia su género, y promete los correspondientes mecanismos de validación.
Y desde bien pronto la *Web* vio mecanismos de publicación de estos senderos personales, curiosamente autónomos, desprovistos de una obra que los organizara o justificase.
A través de él vemos la *Web* como un sistema inmensamente flexible y reorganizable.

Cuadro 2.5: Conjunto de frases fonéticamente equilibrado usando el registro de corpus audiovisuales de caras parlantes.

cuanto a visemas y transiciones entre ellos. Más concretamente, se desean distribuciones uniformes en cuanto a vocales, consonantes, transiciones entre vocal y consonante, entre vocal y silencio y entre silencio y consonante. Además, se toma la suposición de que una transición observada entre un elemento y otro es idéntica a la transición inversa. En el cuadro 2.7 se pueden observar la cantidad y variedad de elementos de que consta una serie de X palabras.

El número de palabras X se fija para poder obtener distribuciones totalmente uniformes de cada tipo de los cinco tipos de elementos. Para ello, se debe cumplir que cada elemento posea un número natural de repeticiones (ver cuadro 2.7):

$$\frac{2X}{5} = p,\ \frac{X}{3} = q,\ \frac{X}{5} = r,\ \frac{X}{6} = s,\ \frac{X}{10} = t,\ p, q, r, s, t \in \mathbb{N} \tag{2.1}$$

A partir de las condiciones expresadas en (2.1), se puede decir que X debe cumplir las

Articulación	Alófono escogido	Alófonos representados
Vocal *a*	[a]	[a]
Vocal *e*	[e]	[e]
Vocal *i*	[i]	[i][ɪ]
Vocal *o*	[o]	[o]
Vocal *u*	[u]	[u][ʊ]
Labial	[m]	[p][m][b][β]
Labio-dental	[f]	[f][ɱ]
Alveolar	[n]	[n][s][z][l][r][ɾ]
Interdental	[θ]	[θ][ð][t̪][d̪]
Palatal	[ɲ]	[ɲ][j][ʎ][t͡ʃ]
Velar	[x]	[x][k][g][ɣ][ŋ]
Silencio	[_]	[_]

Cuadro 2.6: Visemas escogidos para cada vocal y punto de articulación consonántico.

siguientes condiciones de multiplicidad:

$$2X = \dot{5},\ X = \dot{3},\ X = \dot{5},\ X = \dot{6},\ X = \dot{10}$$

las cuales serán todas ciertas para el mínimo común múltiplo de tres, cinco, seis y diez, que es treinta (2.2).

$$m.c.m.\,(3, 5, 6, 10) = 2 \cdot 3 \cdot 5 = 30 \tag{2.2}$$

En treinta palabras bisílabas del modo CVCV existen 120 grados de libertad, donde cada uno puede tomar entre cinco y seis valores, lo cual da un orden de magnitud en el número de combinaciones posibles de 10^{88}. Por otro lado, dadas las restricciones de que existen 30 posibles sílabas de la forma consonante-vocal, y que cada una puede aparecer sólo tres veces y que las transiciones vocal-consonante también cuentan, el número de soluciones se reduce a un orden de 10^{59}. Aunque esta magnitud puede parecer astronómica, si se compara con el número de combinaciones posibles del orden de 10^{88}, se puede ver que es 10^{29} veces más pequeño, es decir, prácticamente nulo en comparación. Estas magnitudes hacen imposible en la práctica intentar buscar una solución de forma exhaustiva: en media, se deberían hacer del orden de 10^{29} pruebas para encontrar una solución. Buscando un símil físico, esta relación equivaldría a buscar uno hormiga en toda la superficie del planeta Júpiter. Para poder encontrar este conjunto de palabras CVCV en un tiempo razonable se han utilizado los algoritmos genéticos (Goldberg, 1989) debido a la alta dimensionalidad de este problema de búsqueda y sus características no lineales. Concretamente, se ha optado por el uso de algoritmos genéticos paralelos ya que muestran un comportamiento más eficiente y tienen un nivel de convergencia prematura menor (Cantu-Paz, 2000). En el cuadro 2.8 se puede consultar una solución dada por el algoritmo utilizado, cuya descripción concreta se puede consultar en el apéndice B.

Tipo	Total	Variedad	Cantidad	Elementos
vocales	$2X$	5	$2X/5$	[a] [e] [i] [o] [u]
consonantes	$2X$	6	$X/3$	[m] [f] [θ] [n] [ɲ] [x]
transiciones V-silencio	X	5	$X/5$	[a_] [e_] [i_] [o_] [u_]
transiciones C-silencio	X	6	$X/6$	[m_] [f_] [θ_] [n_] [ɲ_] [x_]
transiciones C-V	$3X$	30	$X/10$	[ma] [me] [mi] [mo] [mu] [fa] [fe] [fi] [fo] [fu] [θa] [θe] [θi] [θo] [θu] [na] [ne] [ni] [no] [nu] [ɲa] [ɲe] [ɲi] [ɲo] [ɲu] [xa] [xe] [xi] [xo] [xu]

Cuadro 2.7: Elementos constituyentes de una serie de X palabras de la forma CVCV. Se especifica la cantidad total de cada tipo de elementos, así como el número de posibilidades, o variedad, asociada a cada uno. Las últimas dos columnas contienen la cantidad específica de cada elemento y los elementos específicos del tipo indicado en la primera columna.

Palabras sin sentido
giza, jami, jano, mañu, geñe, nuja, fujo, gemi, mejo, moza, fugi, ñimo, fegi, ñafo, fuma, zuño, noñe, nefa, cemu, ñifi, nazu, moñu, mune, ninu, jozu, ñafe, ñina, cige, fone, zofi

Cuadro 2.8: Una posible lista de treinta palabras sin sentido a pronunciar en el proceso de grabación de un corpus audiovisual de caras parlantes, encontrado por el algoritmo de descripción de corpus utilizado.

2.1.4. Grabación de un objeto no rígido

Si el corpus contiene un objeto que posee movimiento no rígido (como una mano, por ejemplo), se puede tratar éste como si las deformaciones no rígidas fuesen cambios en su apariencia visual.

Las restricciones a cumplir en este caso concreto son las siguientes:

- En cada secuencia sólo deber aparecer el objeto no rígido, el cual no debe desaparecer de la escena.
- La secuencia debe contener toda la variabilidad de deformaciones no rígidas del objeto que se deseen reproducir en una etapa posterior.
- El fondo debe tener un color homogéneo y la iluminación no debe provocar sombras muy marcadas (que no se suelen producir con la iluminación de ámbito doméstico).
- El objeto puede moverse con libertad, pero los movimientos de rotación que no sean en el eje de visión de la cámara se interpretarán como cambios de apariencia.

- Cada secuencia debe estar almacenada en un archivo de vídeo, el cual puede tener información auditiva asociada si el objeto posee características sonoras sincronizadas con su aspecto.

La principal diferencia con respecto al caso del apartado 2.1.3 consiste en que en éste se debe tener en cuenta que el fondo ha de tener un color homogéneo en toda la extensión capturada en la imagen. Como se verá en el apartado 2.2, la región de la imagen que contiene el objeto no puede cambiar de forma. No obstante, si se considera que el objeto no rígido se encuentra enganchado a una pantalla virtual que no cambia de color, el hecho de tener un fondo homogéneo de ese mismo color provoca que se pierda toda referencia al situar la pantalla sobre dicho fondo. Esta consideración facilita el proceso de alineación posterior (ver apartado 3.1). En línea con esta restricción, se permite una libertad de movimientos mayor que en el caso del apartado 2.1.3, aunque con la desventaja de que los movimientos de rotación adicionales, es decir, los que no son alrededor del eje de visión, se interpretan como cambios de apariencia. Este hecho implica que no se puede separar este tipo de rotaciones de sus transformaciones no rígidas, quedando unidas inexorablemente si se producen.

Dado que el tipo de objetos es genérico no se puede definir una variabilidad específica. No obstante, se puede proceder de manera similar al caso del apartado 2.1.3 siguiendo el algoritmo de la figura 2.2. Además, debido a que este tipo de objetos puede no poseer información auditiva intrínseca, en principio no se puede contar con ésta para identificar las apariencias clave a posteriori. No obstante, se puede acompañar su grabación con un discurso hablado que tenga información sincronizada descriptiva para poder extraerlas en la fase de análisis automáticamente con segmentadores de voz, al igual que ocurre con las caras parlantes (ver apartado 3.3.1.3).

2.2. Modelo visual

El término de modelo visual proviene de los vocablos **modelo** y **visual**. El primero significa «esquema teórico, generalmente en forma matemática, de un sistema o de una realidad compleja» (Real Academia Española, 2001), mientras que por el segundo se entiende algo «perteneciente o relativo a la visión» (Real Academia Española, 2001). Su unión se refiere, en este trabajo, al esquema teórico que explica la realidad observada a través de la acción de ver o, dicho con otras palabras, al **contenedor de la información visual esencial de un objeto** extraída de un corpus audiovisual (ver apartado 2.1) en el que aparece éste.

La existencia de un modelo visual se basa en la necesidad de poseer una manera eficiente de representar toda la información contenida en el corpus audiovisual. El modelo visual compacta esta información en un subespacio acotado de apariencia. A partir del modelo presentado, se puede reproducir todo el corpus audiovisual e imágenes no existentes en éste, incluyendo las obtenidas por interpolaciones realistas realizadas entre cualquier par de imágenes originales.

En el apartado 2.2.1 se especifican las características del modelo y en el apartado 2.2.2, sus partes constituyentes.

2.2.1. Características

El modelo visual usado está basado en imágenes bidimensionales. Cada una de las instancias del mismo contiene la información asociada a la apariencia y la dinámica visual de un objeto concreto de forma detallada. El modelo visual es general, compacto, modular y acotado.

2.2.1.1. General

Este modelo se puede aplicar a todo tipo de objetos, mientras éstos no desaparezcan de la imagen y puedan ser capturados sin dificultad. Se define un proceso a seguir para el caso concreto de las caras humanas, que puede tomarse como ejemplo para el resto de objetos.

2.2.1.2. Compacto

El modelo visual es compacto ya que, mediante la utilización de técnicas de reducción de dimensionalidad (Kirby, 2001), como análisis de componentes principales (PCA) Jolliffe (1986), se recogen los principales modos de variación o componentes principales de las señales visuales, logrando una descripción de los datos originales en un subespacio de dimensionalidad mínima.

2.2.1.3. Modular

Es modular porque la región observada se puede segmentar en diferentes regiones, las cuales son tratadas de forma independiente. Actuando de este modo se reducen los modos de variación, requiriendo menos imágenes para recoger toda la variedad en la apariencia.

2.2.1.4. Acotado

El modelo visual usado está paramétricamente acotado ya que se conocen los límites del subespacio definido por las componentes principales (ver apartado 3.2.2). Estos límites pueden ser extraídos también mediante las técnicas de PCA usadas en la obtención de dichas componentes.

2.2.2. Definición

El modelo visual contiene tres tipos de información, dos asociadas al canal visual y una tercera asociada al canal acústico, en el caso de representar un corpus audiovisual. Si el corpus es simplemente visual, el tercer tipo no aparece. La información asociada al canal visual se refiere, por un lado, a la manera de describir la apariencia y, por el otro, al modo en que quedan representadas las transiciones entre las diferentes apariencias posibles. El primero se recoge en el subespacio visual del modelo y el segundo en la dinámica visual. El cuadro 2.9 ofrece un resumen de todos los elementos que se detallarán en los siguientes subapartados.

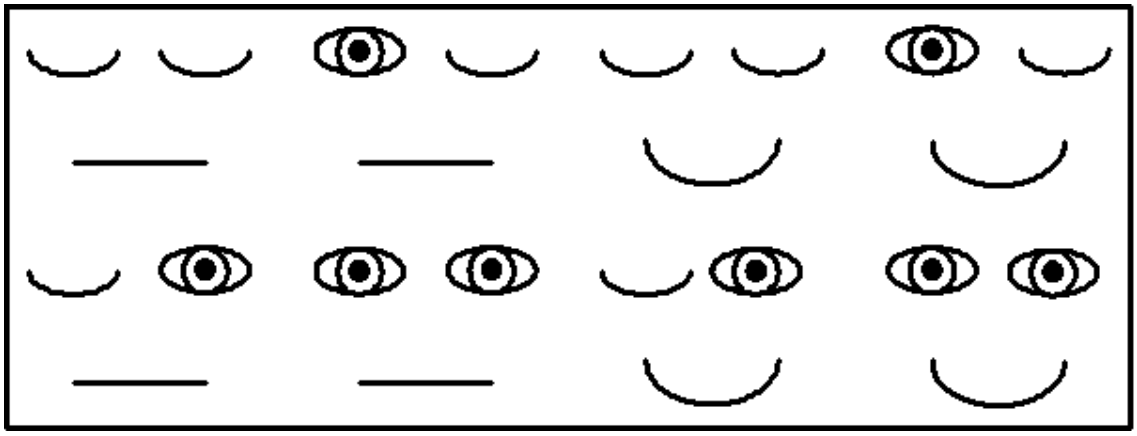
Elementos del modelo visual				
Subespacio visual		Dinámica visual		Modelo acústico
Máscaras	$\mathbf{\Pi}$	Muestreo	$\mathbf{C}^1 \dots \mathbf{C}^L$	CIV
Base	$\mathbf{U}^1 \dots \mathbf{U}^L$	Coarticulación	$\mathbf{G}^1 \dots \mathbf{G}^L$	RVV (RVI)
Límites	$\mathbf{\Sigma}^1 \dots \mathbf{\Sigma}^L$			
Media	$\bar{\mathbf{o}}^1 \dots \bar{\mathbf{o}}^L$			

Cuadro 2.9: Elementos constituyentes del modelo visual. El subespacio visual está formado por una base que lo define en unos límites, un punto medio que lo ubica en el espacio y un conjunto de L imágenes binarias que identifican las L regiones. La dinámica está formada por el muestreo no uniforme del subespacio visual y los grafos de distancias entre todas las muestras. El modelo acústico aparece asociado a corpus audiovisuales y puede contener una de las siguientes opciones: una correspondencia identificador-visema (CIV), una relación voz-visema (RVV) y/o una relación voz-identificador (RVI). Cada identificador es una etiqueta relacionada con un alófono, el visema se encuentra codificado en función del subespacio visual y la voz se encuentra codificada vectorialmente.

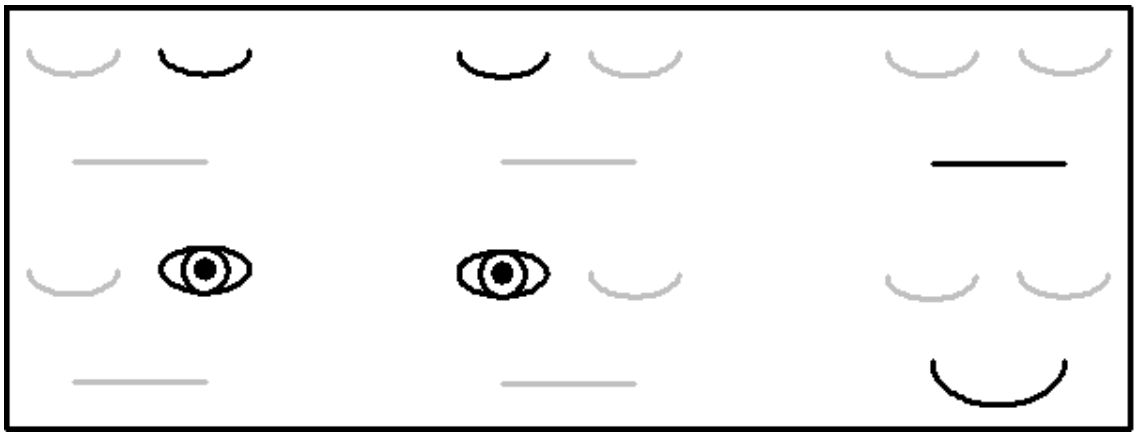
2.2.3. Subespacio visual

Esta parte del modelo visual representa la misma información de apariencia visual del objeto contenido en el corpus, pero de forma lo más compacta posible. Para codificarla, se utiliza la información radiométrica de los píxeles asociados al objeto en sus diferentes configuraciones. Esta información se compacta mediante el uso de transformaciones lineales óptimas como la transformación de Karhunen Loeve (Karhunen, 1947, Loève, 1955), más conocida como PCA (Jolliffe, 1986), la cual ha sido usada extensamente para representar patrones, entre los cuales destacan las imágenes faciales (Kirby, 2001).

Con el objetivo de reducir la variabilidad se propone dividir la imagen en regiones con un cambio de aspecto claramente independiente. Por ejemplo, en el caso de una apariencia facial, se puede optar por separar la región de los ojos de la de la boca (Jebara et al., 1998) (ver figura 2.5).



(a) Versión no modular



(b) Versión modular

Figura 2.5: Reducción de los modos de variación al trabajar con regiones. En este ejemplo esquemático, para representar los (a) ocho modos de variación de la versión no modular, únicamente son necesarios (b) dos modos de variación por región en la versión modular, que acaban ocupando una cuarta parte del espacio.

2.2.3.1. Definición del subespacio de apariencia

Sea $\mathbf{o}_n$ la representación vectorial que expresa la información radiométrica del objeto en la imagen n (ver figura 2.6) y la matriz $\mathbf{O}$, el conjunto de las N imágenes vectorizadas de las consta el corpus, dispuestas en columnas y manteniendo el orden original de grabación. A partir de éstos se puede definir un subespacio engendrado por K vectores generadores, los cuales se encuentran en las columnas de $\mathbf{U}$, conocida como base de apariencia. Sobre ésta se proyectan las imágenes de los objetos obteniendo su representación en vectores de dimensión más pequeña $\mathbf{c}_n$. Tal y como se detalla más adelante, en el apartado 3.2.2, se cumple que:

$$\mathbf{o}_n = \mathbf{U}\mathbf{c}_n + \bar{\mathbf{o}} \tag{2.3}$$

donde $\bar{\mathbf{o}} = N^{-1}\mathbf{O} \cdot \mathbf{1}$, es decir, la media de todas las imágenes recogidas en las columnas de $\mathbf{O}$. De este modo el subespacio de apariencia viene definido por $\mathbf{U}$ y la información media $\bar{\mathbf{o}}$. Este subespacio contiene las componentes principales de todos los vectores centrados $\mathbf{o}_n - \bar{\mathbf{o}}$, está acotado por los valores singulares asociados σ_k (3.72) y es el que mejor compacta su energía sin utilizar ninguna transformación no lineal. Se puede observar un ejemplo más adelante en la figura 3.13, en el que se ha utilizado un subespacio de dimensión diez.

Figura 2.6: Diferentes instancias de la apariencia facial correspondientes a la vista frontal de la cabeza de una persona.

2.2.3.2. División en regiones

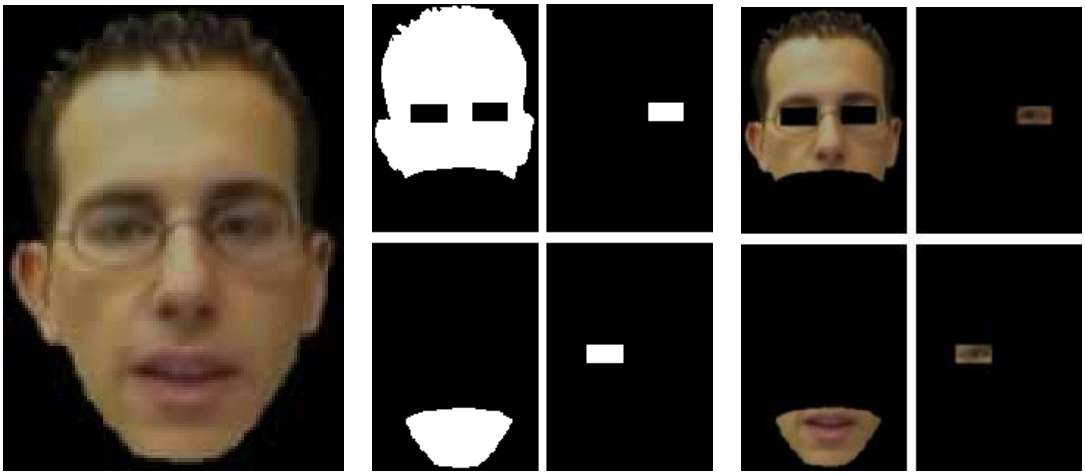
La división del objeto en L regiones se realiza mediante la especificación de las imágenes máscara π^l (ver figura 2.7). Estas imágenes binarias toman el valor unitario para los píxeles que pertenecen a la región que describen, valiendo cero en el resto. La unión de todas las regiones π^l da como resultado el conjunto conexo de píxeles asociados al objeto de interés, denotados por la imagen binaria $\mathbf{\Pi}$. Dado que las diferentes regiones que forman el objeto tienen una intersección nula, se puede decir que su suma equivale a la imagen binaria $\mathbf{\Pi}$ (2.4).

$$\mathbf{\Pi} = \bigcup_{l=1}^{L} \pi^l = \sum_{l=1}^{L} \pi^l \tag{2.4}$$

Como se puede observar, esta definición supone que el objeto no cambia de forma, puesto que cada imagen binaria π^l se define de forma estática. La división en regiones implica la existencia de L vectores $\mathbf{o}_n^l$ por imagen, denotados por el superíndice l, implicando la existencia de L matrices $\mathbf{U}^l$ y $\mathbf{\Sigma}^l$ y L vectores media $\bar{\mathbf{o}}^l$. Para facilitar la lectura, se supondrá una sola máscara hasta la finalización del capítulo, evitando el uso del superíndice al no ser necesario para explicar los conceptos restantes del mismo.

En caso de no necesitar la subdivisión en regiones, $\pi^1 = \mathbf{\Pi}$ y $L = 1$, con lo que siempre será necesaria definir, como mínimo una región. Para realizarlo, se puede consultar el apéndice C para obtener un ejemplo con caras humanas, aunque cada objeto requerirá su conjunto de máscaras concreto.

En el caso de tener un objeto flexible, es decir, con movimiento no rígido asociado, se puede simular una región fija circundante que lo englobe, la cual no cambia de forma y toma las variedades del movimiento no rígido asociado como cambios de apariencia de la



(a) (b) (c)

Figura 2.7: Representación de una imagen en regiones a partir de un conjunto de imágenes máscara: (a) la imagen; (c) las imágenes máscara; (b) las regiones.

región circundante. Para que esta aproximación tenga efecto debe garantizarse uno de los dos efectos siguientes:

- Que el objeto no padece ningún tipo de movimiento rígido, o bien
- Que el fondo posee un color homogéneo, de tal manera que cuando las transformaciones del objeto provocan transformaciones de su región circundante, éstas no son distinguibles observando únicamente el contenido de la región circundante.

En línea con las restricciones del apartado 2.1.4, se utilizará el segundo, que ofrece una mayor facilidad al usuario a la hora de obtener un corpus de este tipo de objetos (ver figura 2.8), ya que no le limita sus movimientos rígidos.

2.2.4. Dinámica visual

Este elemento del modelo visual representa la información asociada a las transiciones posibles entre las diferentes apariencias visuales que existen en el corpus registrado. Además, la dinámica visual contiene todas las apariencias del corpus codificadas en el subespacio visual (apartado 2.2.3). Dado que una transición entre dos apariencias no es más que un conjunto ordenado de otras apariencias, se puede decir que las transiciones son secuencias ordenadas de apariencias, las cuales al proyectarse sobre el subespacio visual se acaban traduciendo en secuencias de vectores o trayectorias de este subespacio. Para generar una transición realista únicamente es necesario realizar una trayectoria dentro del subespacio visual.

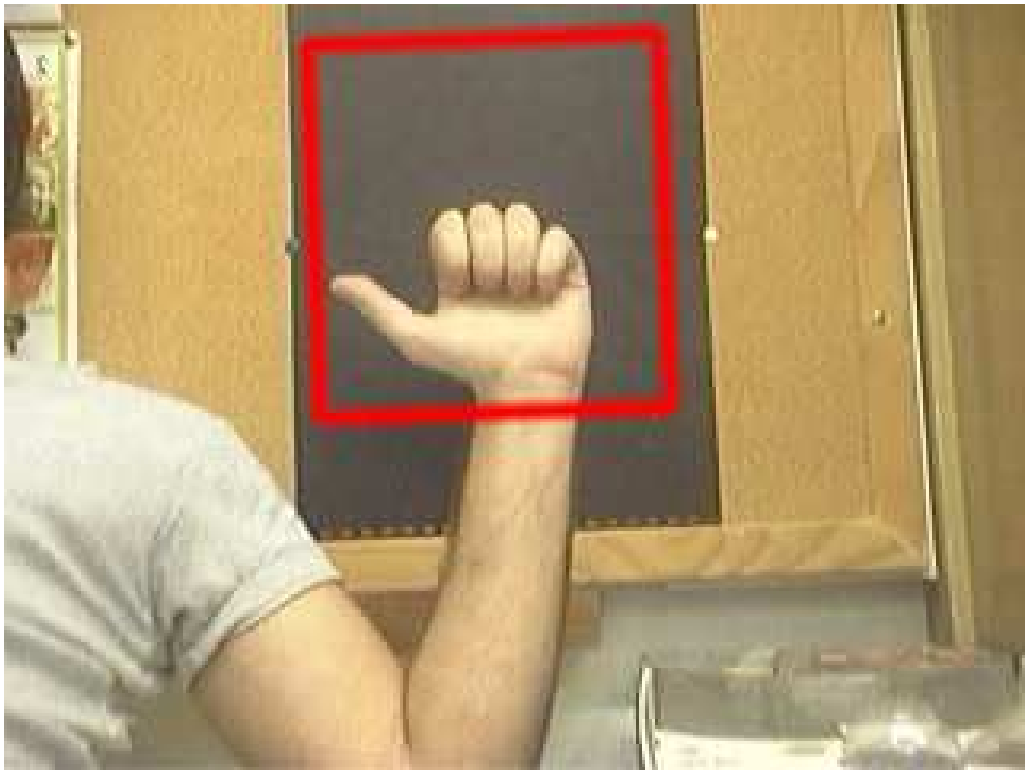
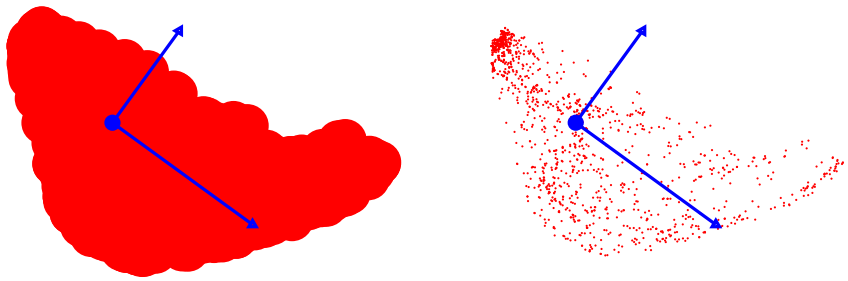


Figura 2.8: Región asociada a un objeto con movimiento no rígido. El fondo homogéneo permite que el objeto se pueda desplazar sin provocar cambios en su apariencia

2.2.4.1. Muestreo del subespacio de apariencia

Para poder determinar los límites reales y precisos del subespacio visual no basta con utilizar sus valores singulares asociados a sus componentes principales ya que éstos sólo garantizan una permanencia dentro del subespacio si éste tiene una distribución de valores gaussiana. En un caso genérico, esta distribución no se puede suponer. En vez de intentar encontrar un modelo descriptivo concreto, se trabaja con los datos existentes en el corpus. Si el corpus fuese infinito, se observarían todas las configuraciones posibles del objeto y, por ende, su subespacio completo. No obstante, dado que el tiempo de grabación es finito, sólo se obtendrá un subconjunto de las mismas, que corresponderán, en general, a un muestreo no uniforme del subespacio visual real. Este conjunto de apariencias está formado por todos los vectores $\mathbf{c}_n$ (2.3), proyecciones de todas las imágenes del corpus, y se denota por la matriz $\mathbf{C}$, que los contiene en sus columnas. En este trabajo, cada una de estas muestras se conoce como unidad visual real. Un dibujo esquemático se puede observar en la figura 2.9



(a) Subespacio teórico (b) Muestreo del subespacio

Figura 2.9: Muestreo del subespacio de apariencia: (a) Subespacio teórico bidimensional que representa la dinámica visual correspondiente a un objeto; (b) muestreo no uniforme de dicho subespacio, donde cada muestra $\mathbf{c}_n$, denotada por un punto rojo, representa la codificación de una imagen observada en el corpus. En ambos casos se pueden observar: en azul, los ejes de coordenadas, los cuales representan los dos vectores ortonormales $\mathbf{u}_1$ y $\mathbf{u}_2$ generadores del subespacio; y como puntos circulares azules grandes, el centroide, que es la media $\bar{\mathbf{o}}$ de todas las apariencias visuales del objeto. Notar también que los vectores $\mathbf{u}_1$ y $\mathbf{u}_2$ se presentan escalados proporcionalmente a la varianza del subespacio en sus respectivas direcciones (σ_1 y σ_2).

2.2.4.2. Grafo de coarticulación

Este elemento contiene las distancias entre todas las unidades visuales reales $\mathbf{c}_n$ o muestras del subespacio visual. Cada distancia concreta se encuentra en una posición de la matriz $\mathbf{G}$:

$$\mathbf{G}(i,j) = d(\mathbf{c}_i, \mathbf{c}_j) = \|\mathbf{c}_1 - \mathbf{c}_2\|_2^\alpha \tag{2.5}$$

donde la función $d(\mathbf{c}_1, \mathbf{c}_2)$ se puede sustituir por cualquier tipo de medida de distancia en general, y por la expresión $\left\|\mathbf{c}_1^l - \mathbf{c}_2^l\right\|_2^\alpha$ en particular para este caso (α es un valor real mayor que 1). Las razones de esta elección se pueden encontrar en el apartado 4.1.3. La distancia entre muestras $\mathbf{c}_n$ (2.5) está directamente relacionada con el parecido entre las apariencias visuales correspondientes. Dos muestras muy parecidas dan lugar a imágenes del mismo parecido:

$$\|\mathbf{o}_i - \mathbf{o}_j\|_2^\alpha = \|\mathbf{U}\mathbf{c}_i + \bar{\mathbf{o}} - \mathbf{U}\mathbf{c}_j - \bar{\mathbf{o}}\|_2^\alpha = \|\mathbf{U}(\mathbf{c}_i - \mathbf{c}_j)\|_2^\alpha = \|\mathbf{c}_i - \mathbf{c}_j\|_2^\alpha \tag{2.6}$$

ya que $\mathbf{U}$ es una matriz de rotación de norma unitaria, como se puede ver en el apartado 3.2.2. No obstante, (2.6) sólo es cierta si $\mathbf{U}$ es cuadrada (tantos vectores base como píxeles tenga el objeto a representar). Si no, en el caso en que $\mathbf{U}$ sólo posea K columnas se produce un error ϵ_k al aproximar $\mathbf{o}_k$ con $\mathbf{U}\mathbf{c}_k + \bar{\mathbf{o}}$, quedando:

$$\|\mathbf{o}_i - \mathbf{o}_j\|_2^\alpha = \|\mathbf{U}\mathbf{c}_i + \epsilon_i + \bar{\mathbf{o}} - \mathbf{U}\mathbf{c}_j - \epsilon_j - \bar{\mathbf{o}}\|_2^\alpha \leq \|\mathbf{c}_i - \mathbf{c}_j\|_2^\alpha + 2(\sigma_{K+1})^\alpha$$

$$\|\mathbf{o}_i - \mathbf{o}_j\|_2^\alpha = \|\mathbf{U}\mathbf{c}_i + \epsilon_i + \bar{\mathbf{o}} - \mathbf{U}\mathbf{c}_j - \epsilon_j - \bar{\mathbf{o}}\|_2^\alpha \geq \|\mathbf{c}_i - \mathbf{c}_j\|_2^\alpha - 2(\sigma_{K+1})^\alpha$$

$$\|\mathbf{o}_i - \mathbf{o}_j\|_2^\alpha = \|\mathbf{c}_i - \mathbf{c}_j\|_2^\alpha + \xi, \quad -2(\sigma_{K+1})^\alpha \leq \xi \leq 2(\sigma_{K+1})^\alpha$$

En este caso, la diferencia con la distancia real viene dada por ξ, que es un valor desconocido que depende de los valores singulares asociados a las K componentes principales del subespacio visual (columnas de $\mathbf{U}$). Notar que cuanto mayor es el valor de K, más pequeño es el margen de error ξ, ya que los valores singulares están ordenados decrecientemente, es decir, $\sigma_K > \sigma_{K+1}$ (ver apartado 3.2.2.

2.2.5. Modelo acústico

Esta componente optativa del modelo visual contiene la información relativa a la sincronía de la apariencia visual del objeto con la información auditiva. En este trabajo, este modelo se aplica a la cara humana y se utiliza en el proceso de síntesis para guiar o dirigir el proceso de generación de información visual. El modelo acústico puede tomar dos representaciones, según sea la fuente de datos que guíe la síntesis:

- Una correspondencia identificador-visema (CIV) entre el conjunto de alófonos (del cuadro 2.4) y un subconjunto de las columnas de $\mathbf{C}$ de la dinámica visual. Esta correspondencia relaciona dos conjuntos de símbolos y puede permitir, por ejemplo, la sincronización con un conversor de texto a habla (TTS) dado.
- Una relación voz-visema (RVV) entre información auditiva y un conjunto mínimo de visemas, representado por un subconjunto de las columnas de $\mathbf{C}^l$. En este caso, se trabaja con la propia información directamente (con los vectores de información, tanto de vídeo como de audio -ver apartado 4.2.3-). Esta aplicación se puede usar, por ejemplo, para estimar visemas a partir de voz.

Existe un caso intermedio que se conoce en este trabajo como relación voz-identificador (RVI), que es un tipo específico de RVV, pero donde se establece una relación entre los vectores de información auditiva o tramas de voz (ver apartado 4.2.3) y un conjunto de símbolos que pueden ser la entrada a un CIV.

2.2.5.1. Correspondencia identificador-visema

La CIV consiste en la construcción de una correspondencia f:

$$f : \mathcal{A} \longmapsto \mathcal{B} \tag{2.7}$$

Por un lado, el conjunto inicial de salida $\mathcal{A}$ está formado por un subconjunto muy particular de los números enteros; estos números concretos son los que usan los conversores TTS para codificar los visemas asociados al texto que están sintetizando; el conjunto $\mathcal{A}$ puede variar según cada conversor TTS. Por otro lado, el conjunto final de llegada $\mathcal{B}$ está constituido por un subconjunto de las columnas de $\mathbf{C}$. Cada una de estas columnas representa un visema clave, asociado a un conjunto de alófonos visualmente similares. Se habla de correspondencia y no de función porque un alófono puede tener asociados diversos visemas (ver apartado 2.1.3.1). Será el algoritmo de síntesis el encargado de decidir qué unidad es

la más adecuada para representar cada alófono durante el transcurso de la generación de información visual (ver apartado 4.1.3.2).

2.2.5.2. Relación voz-visema

La RVV se basa también en una correspondencia del mismo estilo que (2.7). En este caso, el conjunto inicial $\mathcal{A}$ es mucho más amplio, ya que está formado por la información sonora que genera la voz humana, pero codificada vectorialmente (ver apartado 4.2). El conjunto final $\mathcal{B}$ puede ser el propio muestreo $\mathbf{C}$ de la dinámica visual o conjunto de visemas clave, al igual que en el caso de la CIV (o sea, un conjunto de visemas). Para esta segunda opción, la RVV recibe el nombre de RVI.

La forma que adopta la RVV (o RVI) depende del método utilizado y puede ir desde el uso de HMM (Yamamoto et al., 1998) hasta el de redes neuronales Agelfors et al. (1999). Se refiere al lector al apartado 4.2.5 para una descripción del método utilizado y la RVV asociada.

2.3. Cualidades

La representación elegida para la información audiovisual en este trabajo posee características asociadas a cualidades de fiabilidad, flexibilidad, facilidad de uso y coste, que se detallan a continuación y se resumen en el cuadro 2.10.

Cualidad	Características
Fiabilidad	Modelo visual acotado
Flexibilidad	Modelo visual genérico
Facilidad de uso	Uso de dispositivos de adquisición domésticos e iluminación de ámbito doméstico
Coste	Corpus de corta duración, modelo visual modular y compacto

Cuadro 2.10: Cualidades de la representación de la información elegida.

2.3.1. Fiabilidad

La representación escogida para contener la información visual es fiable en cuanto a que el modelo es acotado y las transiciones se limitan al espacio definido por éste, no permitiendo la generación de apariencias irreales.

2.3.2. Flexibilidad

La aplicación del modelo propuesto es genérica, ya que no posee características que lo hagan depender de un tipo de apariencia en particular, al menos, en lo que a información visual se refiere, ya que la información auditiva va estrechamente relacionada a un tipo concreto de objetos.

2.3.3. Facilidad de uso

Partir de la premisa de poder usar dispositivos domésticos de adquisición de secuencias de vídeo es vital para garantizar una mínima facilidad de uso. El requisito de aparatos profesionales anula automáticamente la posibilidad de construir un modelo visual desde casa. Por otro lado, permitir el uso de una iluminación de ámbito doméstico aumenta también la facilidad de su aplicación.

2.3.4. Coste

El hecho de garantizar una mínima duración en la grabación del corpus ayuda a reducir el esfuerzo de grabación del objeto y la cantidad de espacio de almacenamiento necesaria para guardarla. Además, las características de modularidad y compactación del modelo visual ayudan a reducir aún más esta necesidad de espacio, simplificando su tratamiento posterior.

Capítulo 3

Análisis

Como se puede comprobar, la tesis se titula síntesis audiovisual realista personalizable. Aunque la aparición de un apartado de análisis pueda parecer, a priori, paradójica, su existencia en el esquema de trabajo propuesto resulta imprescindible y las razones se pueden encontrar ya en el campo semántico. Según el Diccionario de la Real Academia Española, la palabra **síntesis** significa «Composición de un todo por la reunión de sus partes» (Real Academia Española, 2001). En el presente trabajo, el **todo** consiste en secuencias de vídeo con sonido, es decir, las señales audiovisuales presentes en el corpus audiovisual, con una apariencia visual concreta y cambiante. Las **partes** que componen estas secuencias están incluidas dentro de un conjunto de señales audiovisuales de referencia conocido como el conjunto de entrenamiento. Su análisis o «Distinción y separación de las partes de un todo hasta llegar a conocer sus principios o elementos» (Real Academia Española, 2001) de este último permite obtener dichas **partes**, referidas como información visual esencial y recogidas en el modelo visual. Se debe recordar que, tal como se especifica en el apartado 2.1, el corpus audiovisual puede carecer de la parte auditiva, conociéndose entonces como corpus visual.

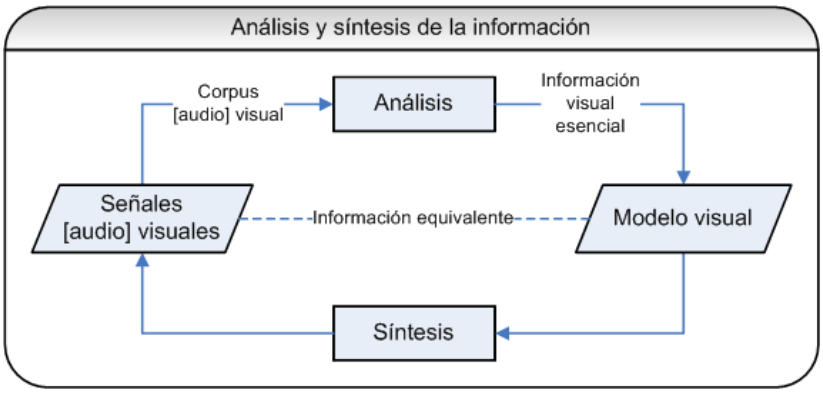


Figura 3.1: Dentro de las posibles señales audiovisuales existentes a estudiar, un subconjunto concreto llamado corpus audiovisual se analiza, obteniendo la descripción de su información visual esencial resumida en el modelo visual obtenido; a partir de éste se pueden crear o sintetizar nuevas señales audiovisuales de la misma naturaleza que el corpus.

En este trabajo, el análisis llevado a cabo permite encontrar una cantidad de información finita y reducida con la cual sintetizar un abanico teóricamente infinito de señales audiovisuales. En otras palabras, y ejemplificando el esquema propuesto con el caso de imágenes faciales, el análisis de un subconjunto reducido de todas las posibles apariencias faciales de una persona ofrece toda la información necesaria para sintetizar (ver capítulo 4) casi cualquier tipo de expresión, gesto o movimiento facial. Sea cual sea la apariencia visual a analizar, el esquema de análisis utiliza cuatro elementos, cuya relación se puede observar en el diagrama de la figura 3.2:

- Un corpus audiovisual, o conjunto de señales audiovisuales que componen el conjunto de entrenamiento o de muestra (apartado 2.1).
- Un modelo visual, que sirve como contenedor de la información visual esencial que se extrae a partir del corpus audiovisual (apartado 2.2).
- Un método de seguimiento para localizar la información de interés en el corpus audiovisual disponible (apartado 3.1).
- Un método de aprendizaje, que extrae y resume la información ya localizada, incluyendo su posterior almacenamiento en el modelo visual especificado (apartado 3.2).

Además, el esquema de análisis propuesto posee un conjunto de propiedades dirigidas que lo hacen fiable, flexible y fácil de usar, entre otras (ver capítulo 3.3).

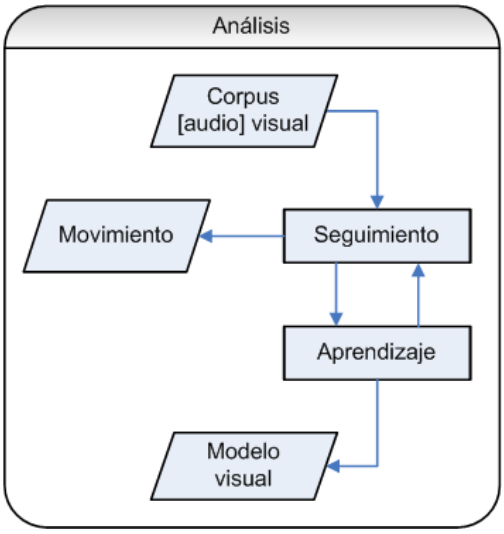


Figura 3.2: Diagrama de bloques del proceso de análisis incluido en el marco propuesto. El seguimiento y el aprendizaje son sus partes constituyentes, mientras que el modelo visual y el corpus (que puede ser audiovisual o no) son su salida y entrada, respectivamente.

3.1. Seguimiento

Se entiende por seguimiento (o en inglés *tracking*) «la acción de ir en busca de [...] algo» (Real Academia Española, 2001); en el ámbito de la visión por ordenador se puede encontrar una definición más adecuada como: «encontrar la localización de un objeto en la escena en cada cuadro de una secuencia de imágenes» (McGraw-Hill Dictionary, 2007). En este trabajo, se emplea el término de seguimiento como **el proceso de encontrar la posición, orientación y escala del objeto u objetos de interés que se hallan en un conjunto ordenado de imágenes**.

El proceso de seguimiento se plantea como un medio para extraer toda la información de movimiento posible de los objetos de interés de un conjunto de imágenes (ver figura 3.3). De este modo, las variaciones de intensidades en las imágenes únicamente podrán ser debidas a cambios en su apariencia. Es de especial interés aplicarlo a las imágenes del corpus audiovisual para facilitar el posterior proceso de aprendizaje (ver apartado 3.2). No obstante, debido a la aplicabilidad a un rango amplio de objetos perseguida en el presente trabajo de investigación, únicamente se extrae el movimiento rígido, ya que los modelos que representan el movimiento no rígido son particulares para cada objeto. Esta suposición comporta la interpretación de los movimientos no rígidos como cambios de apariencia.

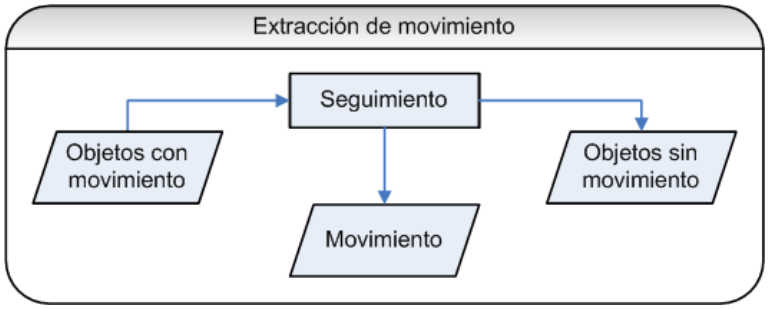


Figura 3.3: El proceso de seguimiento tiene como finalidad extraer todo el movimiento posible de los objetos que aparecen en la secuencia de imágenes.

3.1.1. Características del algoritmo de seguimiento

El seguimiento se realiza sobre las regiones de interés que aparecen en las imágenes obtenidas en el corpus (ver apartado 2.1) para facilitar y mejorar los resultados obtenidos en la fase de aprendizaje (ver apartado 3.2). El seguimiento es general, causal, no intrusivo, automático, sin entrenamiento previo, independiente del fondo y tolera cambios en la apariencia de la región de interés a seguir.

3.1.1.1. General

Es general en el sentido de que puede aplicarse para seguir cualquier tipo de objeto bajo cualquier tipo de condiciones de iluminación, siempre que éstas no cambien brus-

camente entre imágenes consecutivas. Todos los cambios de posición que impliquen una rotación que no sea sobre el eje de visión se interpretan como cambios de apariencia.

3.1.1.2. Causal

El algoritmo es causal ya que la información necesaria para localizar el objeto de interés en una imagen dada sólo requiere información de esa imagen e imágenes anteriores, haciendo posible su implementación en tiempo real.

3.1.1.3. No intrusivo

El seguimiento realizado es no intrusivo porque no es necesario que el objetivo a localizar posea marcas especiales de ningún tipo. El algoritmo propuesto necesita la misma información que el sistema visual humano para operar: la de iluminación que poseen intrínsecamente los diferentes píxeles.

3.1.1.4. Sin intervención manual

El proceso de seguimiento presentado no posee intervención manual porque el algoritmo funciona automáticamente mediante la optimización de una función de error y no necesita ayuda externa para localizar el objeto en cada imagen. El proceso de inicialización se puede realizar también mediante un proceso automático basado en diferenciación de imágenes y morfología matemática (ver apéndice C).

3.1.1.5. Sin entrenamiento previo

No necesita entrenamiento previo puesto que realiza el aprendizaje del modelo simultáneamente con el proceso de seguimiento.

3.1.1.6. Permite cambios de apariencia

El algoritmo permite cambios de apariencia en los objetos a seguir ya que la búsqueda se basa en comparaciones con un subespacio de apariencia, el cual permite resumir de forma altamente eficiente todo un conjunto de apariencias diferentes.

3.1.1.7. Independiente del fondo

El proceso es independiente del fondo de la imagen (entendiéndolo como la región que no pertenece al objeto de interés), aunque sólo en el caso de no existir movimiento no rígido.

3.1.2. Modelo de seguimiento

En este trabajo se pretenden obtener las variaciones en la apariencia de un objeto visual. No obstante, es difícil que este objeto se encuentre totalmente quieto a lo largo de una secuencia, con lo que se hace necesario un proceso de alineamiento de la imagen. Dado que no se conocen a priori las deformaciones que puede padecer el objeto, no se puede distinguir el movimiento rígido del no rígido ni del cambio de apariencia que padece el objeto. Por esta razón, se decide introducir un proceso que elimine automáticamente todo el movimiento rígido que el objeto haya podido efectuar en la grabación del corpus, quedando el movimiento no rígido agrupado con los cambios de apariencia. Es más, también pueden llegar a ser indistinguibles las rotaciones que no se produzcan en el eje de visión de la imagen (ver la figura 5.8 del capítulo 5), con lo que el movimiento rígido sólo se eliminará en dos dimensiones mediante alineación de imágenes. Este pensamiento sigue las conclusiones del trabajo de Tarr y Pinker (1989), que expresa que la representación mental de un objeto se realiza sin tener en cuenta el movimiento rígido que pueda experimentar.

El trabajo de Lucas y Kanade (1981) conforma la base de una gran familia de algoritmos que alinean imágenes (Baker y Matthews, 2004) y representa una de las técnicas más usadas en visión por ordenador hasta el día de hoy. Dadas su flexibilidad y fiabilidad, se elige como método base para desarrollar el seguidor o alineador de imágenes que representará el núcleo de esta sección. Se proponen diferentes extensiones del mismo en el apartado 3.1.2.4 para extenderlo en el marco de este trabajo si el lector lo cree conveniente en sus pruebas y/o aplicaciones. De hecho, el uso del subespacio como referencia y la multirresolución se toman como extensiones básicas en el desarrollo de los algoritmos implementados, basándose en el trabajo de Black y Jepson (1998).

3.1.2.1. Bases

El algoritmo de seguimiento empleado utiliza la suposición de iluminación constante que aparece en la teoría de flujo óptico empleada en Lucas y Kanade (1981) y la suposición de subespacio constante que se utiliza en Black y Jepson (1998). La primera postula que la intensidad de un píxel se mantiene igual al cambiar éste de posición a lo largo del tiempo:

$$I\left(x, y, t\right) = I\left(x + \Delta x, y + \Delta y, t + \Delta t\right) \tag{3.1}$$

donde $\begin{bmatrix}\Delta x & \Delta y\end{bmatrix}^T$ representa el vector de movimiento del píxel y Δt, el incremento de tiempo asociado a tal movimiento. La segunda suposición se basa en que una imagen, tras una transformación geométrica, se puede reconstruir a partir de su proyección en un subespacio concreto, que no cambia:

$$I\left(g\left(\mathbf{x}, \mathbf{a}\right)\right) = \left[\mathbf{U}\mathbf{c}\right]\left(\mathbf{x}\right) \tag{3.2}$$

donde g representa la transformación geométrica aplicada a los píxeles $\mathbf{x}$ de la imagen $\mathbf{I}$ en base a unos parámetros $\mathbf{a}$, la matriz $\mathbf{U}$ contiene la base generadora del subespacio en cuestión y $\mathbf{c}$ es el vector que representa la proyección de la imagen sobre el subespacio.

La descripción del movimiento que aparece en (3.1) como $\begin{bmatrix}\Delta x & \Delta y\end{bmatrix}^T$ y en (3.2) como g se puede expresar de forma unificada del siguiente modo:

$$g(\mathbf{x},\mathbf{a}) = \begin{bmatrix}x+\Delta x\\ y+\Delta y\end{bmatrix} = \begin{bmatrix}x\\ y\end{bmatrix} + \begin{bmatrix}g_x(x,y,\mathbf{a})\\ g_y(x,y,\mathbf{a})\end{bmatrix}$$

La transformación concreta g a utilizar en este trabajo es el modelo de similaridad, que permite traslación, escala y rotación alrededor del eje de visión:

$$g(\mathbf{x},\mathbf{a}) = \begin{bmatrix}x+\Delta x\\ y+\Delta y\end{bmatrix} = \begin{bmatrix}x\\ y\end{bmatrix} + \begin{bmatrix}a^1 & a^2\\ -a^2 & a^1\end{bmatrix}\begin{bmatrix}x-x_c\\ y-y_c\end{bmatrix} + \begin{bmatrix}a^3\\ a^4\end{bmatrix} \tag{3.3}$$

donde x_c e y_c identifican el centro respecto al cual se realiza la transformación g y $\mathbf{a}_n = \begin{bmatrix}a_n^1 & a_n^2 & a_n^3 & a_n^4\end{bmatrix} = [S\,cos(\theta)\ \ S\,sen(\theta)\ \ t_x\ t_y]$ representan los parámetros de transformación, siendo S la escala, θ el ángulo de rotación y t_x y t_y las magnitudes de traslación en las dos dimensiones espaciales.

3.1.2.2. Función objetivo a minimizar

El objetivo principal de este algoritmo consiste en alinear las regiones especificadas por el modelo visual (ver apartado 2.2), que identifican el objeto u objetos de interés. La alineación de dos imágenes se entiende como el proceso de transformar espacialmente una de ellas para hacerla coincidir con la otra. Este proceso se puede efectuar encontrando los parámetros de movimiento $\mathbf{a}_n$ que minimizan la diferencia entre el objeto transformado (o alineado) $\mathbf{o}_n^w$ y el de referencia $\mathbf{o}^{ref}$. Esta diferencia puede expresarse mediante la siguiente medida de error de alineación E, que pasará a ser la función objetivo a minimizar respecto los parámetros de movimiento $\mathbf{a}_n$ en cada imagen I_n:

$$E(\mathbf{a}_n) = \left\|\mathbf{o}^{ref} - \mathbf{o}_n^w\right\|_2^2 \tag{3.4}$$

$$\mathbf{o}^{ref} = vec\left(\mathbf{F}^{ref}, \mathbf{\Pi}\right), \quad \mathbf{o}_n^w = vec(\mathbf{F}_n^w, \mathbf{\Pi}) \tag{3.5}$$

$$\mathbf{F}^{ref} = \sum_{l=1}^{L}\mathbf{R}^{l,ref}, \quad \mathbf{F}_n^w = \sum_{l=1}^{L}\mathbf{R}_n^{l,w} \tag{3.6}$$

$$\mathbf{R}^{l,ref} = I^{ref}\left(g(\mathbf{x},\mathbf{a}_0)\right)\pi^l(\mathbf{x}), \quad \mathbf{R}_n^{l,w} = I_n\left(g(\mathbf{x},\mathbf{a}_n)\right)\pi^l(\mathbf{x}) \tag{3.7}$$

donde $\mathbf{\Pi} = \sum_{l=1}^{L}\pi^l$ (2.4), siendo π^l una imagen con valor diferente de cero para los píxeles no pertenecientes a la región l; I^{ref} y I_n representan la descripción funcional de la imagen de referencia y la enésima; si se desea una descripción matricial de las mismas, se puede utilizar $\mathbf{I}^{ref}$ y $\mathbf{I}_n$, respectivamente; el operador $vec(\mathbf{F},\mathbf{\Pi})$ representa el proceso de vectorización de la matriz $\mathbf{F}$ únicamente en los píxeles de $\mathbf{\Pi}$ diferentes de cero; como se puede ver, las matrices $\mathbf{F}$ están formadas por la suma de matrices $\mathbf{R}^l$, que contienen las diferentes regiones l (definidas por π^l) de la imagen representada por la función bidimensional I; de hecho, la función I y la matriz $\mathbf{F}$ hacen referencia a la misma imagen excepto en el fondo (la región que no pertenece a la unión de las L regiones), ya que la segunda vale cero en los píxeles de

éste. De este modo, es conveniente notar que $vec(\mathbf{F}_n, \mathbf{\Pi}) = vec(\mathbf{I}_n, \mathbf{\Pi})$. La función I^{ref}, una vez transformada según g con unos parámetros arbitrarios $\mathbf{a}_0$, representa la imagen respecto la cual se alinea el resto de imágenes I_n; alternativamente, I^{ref} puede tomarse como la imagen alineada anterior $I_{n-1}\left(g\left(\mathbf{x}, \mathbf{a}_{n-1}\right)\right)$; ambas opciones presentan comportamientos diferentes, que se pueden consultar en Matthews et al. (2003). Las matrices $\mathbf{F}^{ref}$ y $\mathbf{F}_n^w$ representan la imagen de referencia y la enésima alineada, respectivamente, pero sin el fondo; de forma similar ocurre con la región l en las expresiones $\mathbf{R}^{l,ref}$ y $\mathbf{R}_n^{l,w}$; finalmente, el vector de parámetros $\mathbf{a}_n = \begin{bmatrix} a_n^1 & a_n^2 & a_n^3 & a_n^4 \end{bmatrix}$ identifica la transformación de similaridad $g\left(\mathbf{x}, \mathbf{a}_n\right)$ que se aplica sobre la imagen I_n siguiendo el modelo (3.3):

$$g\left(\mathbf{x}, \mathbf{a}_n\right) = \begin{bmatrix} g_x\left(\mathbf{x}, \mathbf{a}_n\right) \\ g_y\left(\mathbf{x}, \mathbf{a}_n\right) \end{bmatrix} = \begin{bmatrix} x_f \\ y_f \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a_n^1 & a_n^2 \\ -a_n^2 & a_n^1 \end{bmatrix} \begin{bmatrix} x - x_c \\ y - y_c \end{bmatrix} + \begin{bmatrix} a_n^3 \\ a_n^4 \end{bmatrix} \tag{3.8}$$

siendo el punto $\left[x_c\ y_c\right]^T$ el centro respecto al cual se realiza la transformación, $\mathbf{x} = \left[x\ y\right]^T$, las coordenadas 2D iniciales de los píxeles de la imagen I_n y $\mathbf{x}_f = \left[x_f\ y_f\right]^T$, las coordenadas finales de la imagen transformada. Como se puede observar, los parámetros $\mathbf{a}_n$ son comunes para todas las regiones $\mathbf{R}_n^{l,w}$ en una misma imagen. El uso del modelo de similaridad (3.3) supone que la imagen es plana y no existen rotaciones en ejes que no sean los de visión.

El proceso de minimización se plantea como encontrar los parámetros $\mathbf{a}_n$ de la transformación geométrica g (3.8) que alinea la imagen I_n con la imagen de referencia I^{ref}, según el contenido visual marcado por las máscaras $\mathbf{\Pi}$. La expresión (3.4) se puede reescribir como sigue, sustituyendo (3.7) en (3.6), ésta en (3.5) y el resultado en (3.4):

$$E\left(\mathbf{a}_n\right) = \sum_{\forall \mathbf{x} \in \mathbf{\Pi}} \left(I^{ref}\left(\mathbf{x}\right) - I_n\left(g\left(\mathbf{x}, \mathbf{a}_n\right)\right)\right)^2 \tag{3.9}$$

es decir, como la suma de las diferencias cuadráticas entre los píxeles de la referencia I^{ref} y la imagen I_n alineada mediante la transformación g en la región definida por $\mathbf{\Pi}$. El valor de E será cero (su mínimo valor) cuando I^{ref} e I_n se hallen totalmente alineadas, tomando la suposición de iluminación constante (3.1). Sin embargo, debido al ruido y a que dicha suposición es únicamente una aproximación de la realidad, no se puede esperar que E alcance dicho valor nulo. De todos modos, dado que una de las restricciones del trabajo (ver apartado 2.1.1) supone iluminación de ámbito doméstico, se puede considerar que entre dos instantes de tiempo cercanos se cumple aún la suposición de iluminación constante (3.1). Aun así, todavía quedan sin compensar los efectos producidos por los diferentes tipos de ruido. Agrupando el efecto del ruido en ν, y tomando un modelo de ruido aditivo, (3.9) queda como:

$$E_r\left(\mathbf{a}_n\right) = \sum_{\forall \mathbf{x} \in \mathbf{\Pi}} \left(I^{ref}\left(\mathbf{x}\right) - I_n\left(g\left(\mathbf{x}, \mathbf{a}_n\right)\right) - \nu\left(\mathbf{x}\right)\right)^2$$

$$\begin{aligned} E_r\left(\mathbf{a}_n\right) &= \left\| \mathbf{o}^{ref} - \mathbf{o}_n^w - \nu \right\|_2^2 = \left\| \mathbf{D}_n - \nu \right\|_2^2 = \\ &= \left(\mathbf{D}_n - \nu\right)^T \left(\mathbf{D}_n - \nu\right) = \mathbf{D}_n^T \mathbf{D}_n - 2\nu^T \mathbf{D}_n + \nu^T \nu \end{aligned}$$

donde $\nu\left(\mathbf{x}\right)$ representa el ruido añadido en el píxel $\mathbf{x}$ y $\mathbf{D}_n = \mathbf{o}^{ref} - \mathbf{o}_n^w$, la diferencia entre la referencia y la región transformada. Sin presencia de ruido (esto es, $\nu = \mathbf{0}$) y usando

(3.4), se cumple que:

$$E_r\left(\mathbf{a}_n\right) = E\left(\mathbf{a}_n\right) = \mathbf{D}_n^T\mathbf{D}_n \tag{3.10}$$

Por otro lado, y bajo la suposición de incorrelación del ruido con la información de las regiones $\mathbf{o}^{ref}$ y $\mathbf{o}_n^w$ (3.11), se puede ver que el valor mínimo de $E\left(\mathbf{a}_n\right)$ se consigue para los mismos valores de $\mathbf{a}_n$ independientemente de la cantidad de ruido (3.12):

$$\nu^T\mathbf{o}^{ref} = 0,\ \nu^T\mathbf{o}_n^w = 0 \ \Rightarrow\ \nu^T\mathbf{D}_n = 0 \tag{3.11}$$

$$\begin{aligned}\min_{\mathbf{a}_n} E_r\left(\mathbf{a}_n\right) &= \min_{\mathbf{a}_n}\left(\mathbf{D}_n^T\mathbf{D}_n - 2\nu^T\mathbf{D}_n + \nu^T\nu\right) = \min_{\mathbf{a}_n}\left(\|\mathbf{D}_n\|_2^2 + \|\nu\|_2^2\right) = \\ &= \|\nu\|_2^2 + \min_{\mathbf{a}_n}\|\mathbf{D}_n\|_2^2 = \|\nu\|_2^2 + \min_{\mathbf{a}_n} E\left(\mathbf{a}_n\right)\end{aligned} \tag{3.12}$$

Así, sin pérdida de generalidad ((3.12) incluye el caso mostrado en (3.10)), se puede decir que la alineación de la imagen I_n se produce con el vector $\mathbf{a}_n$ que minimiza $E\left(\mathbf{a}_n\right)$, independientemente de la cantidad de ruido presente y tomando las suposiciones de iluminación constante y ruido aditivo incorrelacionado con la imagen. En resumen, el objetivo del seguimiento (obtener una alineación de las imágenes) se traduce al final en resolver el siguiente problema de optimización:

$$\min_{\mathbf{a}_n} E\left(\mathbf{a}_n\right) = \min_{\mathbf{a}_n}\sum_{\forall\mathbf{x}\in\mathbf{\Pi}}\left(I^{ref}\left(\mathbf{x}\right) - I_n\left(g\left(\mathbf{x},\mathbf{a}_n\right)\right)\right)^2 \tag{3.13}$$

En el caso en que ν represente otro tipo de ruido correlacionado con $\mathbf{D}_n$, como errores derivados de la aplicación de transformaciones geométricas sobre imágenes, cambios de apariencia en la imagen u oclusiones, la segunda igualdad de la expresión (3.12) pasa a ser una aproximación con error proporcional a $\nu^T\mathbf{D}_n$. La consecuencia directa de este hecho es la interpretación de (3.13) también como una aproximación del problema de seguimiento.

3.1.2.3. Minimización

Una posible estrategia para minimizar $E\left(\mathbf{a}_n\right)$ (3.13) consiste en tomar las derivadas de esta función sobre los parámetros $\mathbf{a}_n$, igualar a 0 todas las ecuaciones y resolver el sistema resultante:

$$\frac{\partial E}{\partial \mathbf{a}_n} = -2\sum_{\forall\mathbf{x}\in\mathbf{\Pi}}\left(I^{ref}\left(\mathbf{x}\right) - I_n\left(g\left(\mathbf{x},\mathbf{a}_n\right)\right)\right)\left(\nabla I_n\left(g\left(\mathbf{x},\mathbf{a}_n\right)\right)\right)^T\begin{bmatrix}\frac{\partial g_x}{\partial \mathbf{a}_n}\left(\mathbf{x},\mathbf{a}_n\right)\\ \frac{\partial g_y}{\partial \mathbf{a}_n}\left(\mathbf{x},\mathbf{a}_n\right)\end{bmatrix} = 0 \tag{3.14}$$

$$\nabla I_n = \begin{bmatrix}\frac{\partial I_n}{\partial x}\\ \frac{\partial I_n}{\partial y}\end{bmatrix} = \begin{bmatrix}I_{x,n}\\ I_{y,n}\end{bmatrix} \tag{3.15}$$

El problema que presenta (3.14) es que no se puede conocer analíticamente el valor de ∇I_n, ya que, generalmente, no se posee la descripción funcional de la imagen I_n. Con

ánimo de simplificar el problema, se aproxima (3.9) mediante el uso del desarrollo de Taylor de primer orden sobre $I_n\left(g\left(\mathbf{x},\mathbf{a}_n\right)\right)$ alrededor del punto $g\left(\mathbf{x},\mathbf{a}_{n-1}\right)$:

$$E\left(\mathbf{a}_n\right) \approx \sum_{\forall \mathbf{x}\in\mathbf{\Pi}} \left(I^{ref}\left(\mathbf{x}\right) - I_n\left(g\left(\mathbf{x},\mathbf{a}_{n-1}\right)\right) - \nabla I_n\left(g\left(\mathbf{x},\mathbf{a}_{n-1}\right)\right)^T \begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix} \right)^2 \tag{3.16}$$

en este caso, Δx y Δy representan la variación de posición del píxel siguiendo el modelo de similaridad (3.3), tomando $\begin{bmatrix}x_c & y_c\end{bmatrix} = \begin{bmatrix}0 & 0\end{bmatrix}$ y $\mathbf{a}_0 = \begin{bmatrix}1 & 0 & x_c & y_c\end{bmatrix}$. Ambas variaciones están relacionadas con las de los parámetros de movimiento $\Delta\mathbf{a}_n$, los cuales se pueden relacionar con $\mathbf{a}_n$ y $\mathbf{a}_{n-1}$ mediante el modelo de movimiento composicional (Baker y Matthews, 2001):

$$g\left(\mathbf{x},\mathbf{a}_n\right) = g\left(g\left(\mathbf{x},\mathbf{a}_{n-1}\right),\Delta\mathbf{a}_n\right)$$

$$\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix} = g\left(\mathbf{x},\mathbf{a}_n\right) - g\left(\mathbf{x},\mathbf{a}_{n-1}\right) = \begin{bmatrix}\Delta a_n^1 & \Delta a_n^2\\ -\Delta a_n^2 & \Delta a_n^1\end{bmatrix}\begin{bmatrix}g_x\left(\mathbf{x},\mathbf{a}_{n-1}\right)\\ g_y\left(\mathbf{x},\mathbf{a}_{n-1}\right)\end{bmatrix} + \begin{bmatrix}\Delta a_n^3\\ \Delta a_n^4\end{bmatrix} \tag{3.17}$$

Como se puede observar en (3.16), la nueva función aproximada depende en realidad de los incrementos de los parámetros de movimiento $\Delta\mathbf{a}_n$ y se conocerá a partir de este punto como la función aproximada del error $\widetilde{E}\left(\Delta\mathbf{a}_n\right)$. Utilizando notación vectorial sobre (3.16):

$$\widetilde{E}\left(\Delta\mathbf{a}_n\right) = \left\| \mathbf{o}^{ref} - \mathbf{o}_n^w - diag\left(\Delta\mathbf{x}\right)\mathbf{o}_{x,n}^w - diag\left(\Delta\mathbf{y}\right)\mathbf{o}_{y,n}^w \right\|_2^2 \tag{3.18}$$

$$\mathbf{o}^{ref} = vec\left(I^{ref}\left(g\left(\mathbf{x},\mathbf{a}_0\right)\right),\mathbf{\Pi}\right) \tag{3.19}$$

$$\mathbf{o}_n^w = vec\left(I_n\left(g\left(\mathbf{x},\mathbf{a}_{n-1}\right)\right),\mathbf{\Pi}\right) \tag{3.20}$$

$$\mathbf{o}_{x,n}^w = vec\left(I_{x,n}\left(g\left(\mathbf{x},\mathbf{a}_{n-1}\right)\right),\mathbf{\Pi}\right) \tag{3.21}$$

$$\mathbf{o}_{y,n}^w = vec\left(I_{y,n}\left(g\left(\mathbf{x},\mathbf{a}_{n-1}\right)\right),\mathbf{\Pi}\right) \tag{3.22}$$

donde la función $diag\left(\cdot\right)$ crea una matriz diagonal con los elementos del vector de entrada y las imágenes $I_{x,n}$ e $I_{y,n}$ son las descritas en el vector gradiente especificado en (3.15). Hay que aclarar que $\mathbf{o}^{ref}$, $\mathbf{o}_n^w$, $\mathbf{o}_{x,n}^w$ y $\mathbf{o}_{y,n}^w$ son vectores columna que sólo contienen los valores de la imagen resultante en los píxeles donde $\mathbf{\Pi}$ es diferente de cero. De modo similar, los vectores $\Delta\mathbf{x}$ y $\Delta\mathbf{y}$ únicamente contienen los valores Δx y Δy (3.17) correspondientes también a las coordenadas $\mathbf{x}$ donde $\mathbf{\Pi}$ es diferente de cero. Si se toma como referencia la imagen anterior I_{n-1}, la identidad (3.19) pasa a ser:

$$\mathbf{o}_n^{ref} = vec\left(I_{n-1}\left(g\left(\mathbf{x},\mathbf{a}_{n-1}\right)\right),\mathbf{\Pi}\right) \tag{3.23}$$

Simplificando la expresión de la igualdad (3.18), se puede reescribir del siguiente modo, utilizando notación vectorial:

$$\widetilde{E}\left(\Delta\mathbf{a}_n\right) = \left\|\mathbf{b}_n - \mathbf{A}_n\cdot\left(\Delta\mathbf{a}_n\right)\right\|_2^2 \tag{3.24}$$

donde:

$$\mathbf{b}_n = \mathbf{o}^{ref} - \mathbf{o}_n^w \tag{3.25}$$

$$\Delta\mathbf{a}_n = \begin{bmatrix} \Delta a_n^1 \\ \Delta a_n^2 \\ \Delta a_n^3 \\ \Delta a_n^4 \end{bmatrix}$$

$$\mathbf{A}_n = \begin{bmatrix} \mathbf{A}_n^1 & \mathbf{A}_n^2 & \mathbf{A}_n^3 & \mathbf{A}_n^4 \end{bmatrix} \tag{3.26}$$

$$\mathbf{A}_n^1 = diag\left(g_x\left(\mathbf{x}, \mathbf{a}_{n-1}\right)\right)\mathbf{o}_{x,n}^w + diag\left(g_y\left(\mathbf{x}, \mathbf{a}_{n-1}\right)\right)\mathbf{o}_{y,n}^w \tag{3.27}$$

$$\mathbf{A}_n^2 = diag\left(g_y\left(\mathbf{x}, \mathbf{a}_{n-1}\right)\right)\mathbf{o}_{x,n}^w - diag\left(g_x\left(\mathbf{x}, \mathbf{a}_{n-1}\right)\right)\mathbf{o}_{y,n}^w \tag{3.28}$$

$$\mathbf{A}_n^3 = \mathbf{o}_{x,n}^w \tag{3.29}$$

$$\mathbf{A}_n^4 = \mathbf{o}_{y,n}^w \tag{3.30}$$

Entonces:

$$\min_{\Delta\mathbf{a}_n} \widetilde{E}\left(\Delta\mathbf{a}_n\right) = \min_{\Delta\mathbf{a}_n} \left\|\mathbf{b}_n - \mathbf{A}_n \cdot \left(\Delta\mathbf{a}_n\right)\right\|_2^2 \tag{3.31}$$

Esta minimización se puede conseguir encontrando la solución (3.32) al sistema sobredeterminado que representa, es decir, encontrando el vector $\Delta\mathbf{a}_n$ más cercano a todas las ecuaciones en el sentido de los mínimos cuadrados (Golub y Loan, 1996):

$$\Delta\mathbf{a}_n = \left(\mathbf{A}_n^T\mathbf{A}_n\right)^{-1}\mathbf{A}_n^T\mathbf{b}_n \tag{3.32}$$

Destacar que la expresión $\left(\mathbf{A}_n^T\mathbf{A}_n\right)^{-1}\mathbf{A}_n^T$ se conoce como la pseudoinversa por la izquierda de $\mathbf{A}_n$ (Golub y Loan, 1996). La solución obtenida $\Delta\mathbf{a}_n$ se debe componer con la transformación acumulada anterior $\mathbf{a}_{n-1}$ para obtener la transformación acumulada actual $\mathbf{a}_n$. Este proceso se puede realizar directamente en el caso del modelo de similaridad utilizado (3.3) mediante el uso de notación homogénea (Bloomenthal y Rokne, 1994):

$$\begin{bmatrix} 1+a_n^1 & a_n^2 & a_n^3 \\ -a_n^2 & 1+a_n^1 & a_n^4 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1+\Delta a_n^1 & \Delta a_n^2 & \Delta a_n^3 \\ -\Delta a_n^2 & 1+\Delta a_n^1 & \Delta a_n^4 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1+a_{n-1}^1 & a_{n-1}^2 & a_{n-1}^3 \\ -a_{n-1}^2 & 1+a_{n-1}^1 & a_{n-1}^4 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.33}$$

No obstante, debido a la aproximación realizada a la función a minimizar (3.16), al cálculo de los gradientes ∇I_n, al uso de la suposición de iluminación constante y a los procesos de interpolación requeridos para obtener $\mathbf{o}^{ref}$, $\mathbf{o}_n^w$, $\mathbf{o}_{x,n}^w$ y $\mathbf{o}_{y,n}^w$, se tiene que la variación de los parámetros de transformación $\Delta\mathbf{a}_n$ únicamente es capaz de compensar movimientos muy pequeños (del orden de 2-4 píxeles aproximadamente, aunque depende del tamaño del objeto de interés y su estructura espacio-frecuencial (Black y Jepson, 1998)). Este hecho implica que en situaciones con movimientos más bruscos es necesario iterar el procedimiento sobre la misma imagen a partir de resultados anteriores. El número de iteraciones depende en alto grado de las características de la imagen y la referencia I^{ref} utilizada. El resumen del proceso a llevar a cabo se muestra en la figura 3.4.

```
[a_n] = Seguimiento_Básico(I_n, a_{n-1}, o^{ref})

    Entrada:
    I_n: Imagen que contiene el objeto a localizar
    a_{n-1}: Localización del objeto en la imagen anterior
    o^{ref}: Vector con los datos de la imagen de referencia
    Salida:
    a_n: Parámetros de localización del objeto en I_n

    1. Construir o_n^w con a_{n-1} según (3.20)
    2. Construir o_{x,n}^w con a_{n-1} según (3.21)
    3. Construir o_{y,n}^w con a_{n-1} según (3.22)
    4. Construir b_n según (3.25) utilizando o^{ref}
    5. Construir A_n según (3.26)
    6. Calcular Δa_n siguiendo (3.32)
    7. Actualizar los parámetros a_{n-1} con Δa_n (3.33)
    8. Volver a 1 hasta fin de iteraciones
    9. a_n = a_{n-1}
```

Figura 3.4: Algoritmo de seguimiento básico.

3.1.2.4. Características adicionales

En el apartado anterior se ha presentado el algoritmo de seguimiento básico (figura 3.4) desarrollado en este trabajo de investigación. No obstante, se pueden incluir múltiples mejoras para dotarlo de mayores prestaciones. Basándose en trabajos anteriores (Black y Jepson, 1998, Hager y Belhumeur, 1998, Baker y Matthews, 2001, Buenaposada et al., 2004), se presentan a continuación un conjunto de extensiones propuestas adaptadas al problema definido aquí, las cuales pueden ser combinadas entre sí según se desee.

Subespacio como referencia El problema de optimización originalmente planteado en el apartado 3.1.2.2 (3.13) contiene una imagen fija como referencia, la cual puede recibir también el nombre de plantilla en la literatura (Matthews et al., 2003). A menudo es necesario realizar el seguimiento sobre una forma que cambia de apariencia, por ejemplo, una cara. Para poder llevarlo a cabo se debería realizar el proceso de seguimiento mediante la comparación con varias plantillas que mostrasen los diferentes aspectos posibles de la forma de interés. Sin embargo, este procedimiento presenta dos desventajas principales:

1. En el seguimiento de formas complejas como una cara, el número posible de aspectos diferentes puede ser desorbitado (del orden de millones o incluso más, dependiendo del nivel de detalle a considerar), haciendo prácticamente imposible la recolección del conjunto completo.

2. El coste computacional aumenta de forma directamente proporcional al número de aspectos diferentes.

En vez de utilizar un conjunto de imágenes de referencia, se propone el uso de un subespacio capaz de generarlas todas (Black y Jepson, 1998) y utilizar la suposición de subespacio constante (3.2). Este subespacio de referencia debe ser capaz de generar únicamente las apariencias deseadas dando al proceso de seguimiento la capacidad de localizar únicamente formas que puedan ser descritas por este subespacio. Aplicando la suposición de subespacio constante sobre la imagen candidata se obtiene la mejor representación de la imagen transformada $\mathbf{o}_n^w$ (3.20) sobre el subespacio de referencia. Utilizando el modelo visual (ver apartado 2.2) con una sola máscara ($\mathbf{\Pi} = \pi^1$), la imagen referencia de la función a minimizar (3.13) pasa, en este caso, a ser la mejor reconstrucción que se puede realizar de $\mathbf{o}_n^w$ con el subespacio $\mathbf{U}$ y la media $\bar{\mathbf{o}}$, que se denotará por $\mathbf{o}_n^{rec}$:

$$\mathbf{o}_n^{ref} = \mathbf{o}_n^{rec} = \mathbf{U}\mathbf{c}_n + \bar{\mathbf{o}} \tag{3.34}$$

$$\mathbf{c}_n = (\mathbf{U})^T (\mathbf{o}_n^w - \bar{\mathbf{o}}) \tag{3.35}$$

En este caso, se puede considerar la imagen de referencia $\mathbf{o}_n^{ref}$ como $\mathbf{o}_n^{rec}$, de modo que se minimice la diferencia entre $\mathbf{o}_n^w$ y su mejor reconstrucción por el subespacio $\mathbf{U}$. Utilizando la suposición de subespacio constante, cuando esta diferencia sea mínima, es de esperar que $\mathbf{o}_n^w$ esté bien representado por el subespacio $\mathbf{U}$, es decir, que se está observando una zona de la imagen que contiene una de las apariencias de interés exclusivamente.

En esta propuesta, la solución mediante linealización (3.32) no se altera, la matriz $\mathbf{A}_n$ (3.26) no sufre ningún tipo de variación (ya que no contiene información alguna acerca del subespacio de referencia) y únicamente hay que tener en cuenta la expresión (3.34) en el momento de definir el residuo $\mathbf{b}_n$ (3.36). El algoritmo de seguimiento es el mismo que el básico (figura 3.4), pero particularizando (3.25) con (3.36) (ver figura 3.5).

$$\mathbf{b}_n = \mathbf{o}_n^{ref} - \mathbf{o}_n^w = \mathbf{o}_n^{rec} - \mathbf{o}_n^w = \mathbf{U}(\mathbf{U})^T (\mathbf{o}_n^w - \bar{\mathbf{o}}) + \bar{\mathbf{o}} - \mathbf{o}_n^w \tag{3.36}$$

Modularidad La posibilidad de contemplar diferentes máscaras para seguir una única forma está asociada a la característica de modularidad del modelo visual comentada en el apartado 2.2.1. Por lo que respecta a la formulación, la referencia $\mathbf{o}^{ref}$ (3.19) se puede expresar como la concatenación (siempre respetando el mismo orden) de las referencias $\mathbf{o}^{ref,l}$, relativas a las diferentes regiones visuales:

$$\mathbf{o}^{ref} = \left[\left(\mathbf{o}^{ref,1}\right)^T \quad \left(\mathbf{o}^{ref,2}\right)^T \quad \dots \quad \left(\mathbf{o}^{ref,L}\right) \right]^T \tag{3.37}$$

$$\mathbf{o}^{ref,l} = vec\left(I^{ref}\left(g\left(\mathbf{x}, \mathbf{a}_0\right)\right), \pi^l \right)$$

Por otro lado, el objeto alineado $\mathbf{o}_n^w$ y sus derivadas espaciales $\mathbf{o}_{x,n}^w$ y $\mathbf{o}_{y,n}^w$ pasan a ser, en este caso:

$$\mathbf{o}_n^w = \left[\left(\mathbf{o}_n^{w,1}\right)^T \quad \left(\mathbf{o}_n^{w,2}\right)^T \quad \dots \quad \left(\mathbf{o}_n^{w,L}\right) \right]^T \tag{3.38}$$

$$\mathbf{o}_n^{w,l} = vec\left(I_n\left(g\left(\mathbf{x}, \mathbf{a}_{n-1}\right)\right), \pi^l \right)$$

```
[a_n] = Seguimiento_Subespacio(I_n, a_{n-1}, M)

    Entrada:
    I_n: Imagen que contiene el objeto a localizar
    a_{n-1}: Localización del objeto en la imagen anterior
    M: Modelo visual, que incluye U y ō
    Salida:
    a_n: Parámetros de localización del objeto en I_n

    1. Construir o_n^w con a_{n-1} según (3.20)
    2. Construir o_{x,n}^w con a_{n-1} según (3.21)
    3. Construir o_{y,n}^w con a_{n-1} según (3.22)
    4. Construir b_n según (3.36) utilizando U y ō
    5. Construir A_n según (3.26)
    6. Calcular Δa_n siguiendo (3.32)
    7. Actualizar los parámetros a_{n-1} con Δa_n (3.33)
    8. Volver a 1 hasta fin de iteraciones
    9. a_n = a_{n-1}
```

Figura 3.5: Algoritmo de seguimiento utilizando un subespacio como referencia.

$$\mathbf{o}_{x,n}^{w} = \left[\left(\mathbf{o}_{x,n}^{w,1}\right)^T \quad \left(\mathbf{o}_{x,n}^{w,2}\right)^T \quad \dots \quad \left(\mathbf{o}_{x,n}^{w,L}\right) \right]^T \tag{3.39}$$

$$\mathbf{o}_{x,n}^{w,l} = vec\left(I_{x,n}\left(g\left(\mathbf{x}, \mathbf{a}_{n-1}\right)\right), \pi^l \right)$$

$$\mathbf{o}_{y,n}^{w} = \left[\left(\mathbf{o}_{y,n}^{w,1}\right)^T \quad \left(\mathbf{o}_{y,n}^{w,2}\right)^T \quad \dots \quad \left(\mathbf{o}_{y,n}^{w,L}\right) \right]^T \tag{3.40}$$

$$\mathbf{o}_{y,n}^{w,l} = vec\left(I_{y,n}\left(g\left(\mathbf{x}, \mathbf{a}_{n-1}\right)\right), \pi^l \right)$$

Hay que notar que $\mathbf{A}_n$ (3.26) no cambia con un reordenamiento adecuado de sus filas, ya que cada una de ellas corresponde a un píxel de la imagen, independientemente de la región π^l a la que pertenezca. De este modo, únicamente hay que utilizar (3.37), (3.38), (3.39) y (3.40) en vez de sus versiones no modulares de la derivación original ((3.19), (3.20), (3.21) y (3.22), respectivamente) en el algoritmo básico de seguimiento (ver figura 3.4) para obtener la versión modular (ver figura 3.6).

Multirresolución El uso de estructuras multirresolutivas (Black y Jepson, 1998) en el proceso de seguimiento se basa en la utilización de la imagen en diferentes resoluciones, mediante la construcción de una pirámide resolutiva sobre la misma. El proceso se inicia sobre la resolución menor y se va refinando el resultado sobre las resoluciones mayores. Este procedimiento presenta varias ventajas:

1. Menor coste computacional global. Debido principalmente a que la mayor parte de iteraciones se dan en la resolución menor.

```
[a_n] = Seguimiento_Modular(I_n, a_{n-1}, o_n^{ref})

    Entrada:
    I_n: Imagen que contiene el objeto a localizar
    a_{n-1}: Localización del objeto en la imagen anterior
    o_n^{ref}: Datos de referencia dispuestos según (3.37)
    Salida:
    a_n: Parámetros de localización del objeto en I_n

    1. Construir o_n^w con a_{n-1} según (3.38)
    2. Construir o_{x,n}^w con a_{n-1} según (3.39)
    3. Construir o_{y,n}^w con a_{n-1} según (3.40)
    4. Construir b_n según (3.25) utilizando o_n^{ref}
    5. Construir A_n según (3.26)
    6. Calcular Δa_n siguiendo (3.32)
    7. Actualizar los parámetros a_{n-1} con Δa_n (3.33)
    9. Volver a 1 hasta fin de iteraciones
    10. a_n = a_{n-1}
```

Figura 3.6: Algoritmo de seguimiento modular.

2. Menor número de iteraciones en los niveles de menor resolución. Se debe a que el desplazamiento de un píxel en la resolución menor equivale al desplazamiento de varios píxeles en la mayor resolución.

3. Menor riesgo de caer en mínimos locales. Debido al menor contenido de alta frecuencia espacial de las imágenes de resolución menor, las superficies definidas por la función de error (3.4) son más suaves.

Aunque, por otro lado, también presenta algunos inconvenientes:

1. Mayor coste en memoria, al tener que almacenar información para cada nivel resolutivo.

2. La información almacenada en cada nivel resolutivo contiene matices diferentes debido a no poder realizar diezmos perfectos sin solapamiento frecuencial sobre la imagen original.

La aplicación de multirresolución sobre el algoritmo de seguimiento propuesto necesita que el modelo visual (ver cuadro 2.9) contenga la información relativa a cada nivel resolutivo. Si se trabaja con R niveles, se multiplicará por este mismo número la cantidad de elementos (ver cuadro 3.1). Además, las imágenes máscaras involucradas en cada nivel resolutivo serán diferentes y se identificarán como:

$$\mathbf{\Pi}^r = \sum_{l=1}^{L} \pi^{l,r}$$

Elementos del modelo visual multirresolutivo			
Subespacio visual		Dinámica visual	Modelo acústico
Máscaras	$\boldsymbol{\Pi}^1 \dots \boldsymbol{\Pi}^R$	Muestreo $\begin{matrix}\mathbf{C}^{1,1} & \dots & \mathbf{C}^{L,1} \\ \vdots & \ddots & \vdots \\ \mathbf{C}^{1,R} & \dots & \mathbf{C}^{L,R}\end{matrix}$	CIV
Base	$\begin{matrix}\mathbf{U}^{1,1} & \dots & \mathbf{U}^{L,1} \\ \vdots & \ddots & \vdots \\ \mathbf{U}^{1,R} & \dots & \mathbf{U}^{L,R}\end{matrix}$	Coarticulación $\begin{matrix}\mathbf{G}^{1,1} & \dots & \mathbf{G}^{L,1} \\ \vdots & \ddots & \vdots \\ \mathbf{G}^{1,R} & \dots & \mathbf{G}^{L,R}\end{matrix}$	RVV (RVI)
Límites	$\begin{matrix}\boldsymbol{\Sigma}^{1,1} & \dots & \boldsymbol{\Sigma}^{L,1} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}^{1,R} & \dots & \boldsymbol{\Sigma}^{L,R}\end{matrix}$		
Media	$\begin{matrix}\bar{\mathbf{o}}^{1,1} & \dots & \bar{\mathbf{o}}^{L,1} \\ \vdots & \ddots & \vdots \\ \bar{\mathbf{o}}^{1,R} & \dots & \bar{\mathbf{o}}^{L,R}\end{matrix}$		

Cuadro 3.1: Elementos que conforman el modelo visual con R niveles resolutivos.

La construcción de una pirámide resolutiva de R niveles sobre una imagen $\mathbf{I}_n$ se denota como $\mathcal{P}(\mathbf{I}_n)$ y produce R imágenes $\mathbf{I}_n^r = \mathcal{D}\{\mathbf{I}_n\}$, donde el operador $\mathcal{D}$ simboliza el proceso de diezmo. Los factores de diezmo se suelen escoger de manera que I^R sea la imagen de menor resolución y I^1 la imagen original.

El algoritmo de seguimiento (ver figura 3.7) es muy parecido a su versión sin multirresolución (ver figura 3.4). Únicamente se debe tener en cuenta que al cambiar de nivel resolutivo, los píxeles muestran desplazamientos diferentes y se debe adaptar el valor de los parámetros de movimiento relativos a la traslación (si se usa un modelo de similaridad, como es el caso). Por ejemplo, a una resolución de la mitad de la original en ambas dimensiones, los valores de traslación serán también la mitad durante el trabajo en esa resolución. Por otro lado, la mayor parte de las iteraciones se realizan en el nivel resolutivo R (el de menor tamaño).

Optimización del algoritmo utilizando una referencia fija El algoritmo presentado en la figura 3.4 presenta dos puntos críticos, en cuanto a coste computacional se refiere: el proceso de interpolación para la construcción del objeto transformado $\mathbf{o}_n^w$ y sus derivadas espaciales $\mathbf{o}_{x,n}^w$ y $\mathbf{o}_{y,n}^w$ en los pasos 2, 3 y 4 y el cálculo de la inversa $\left(\mathbf{A}_n^T\mathbf{A}_n\right)^{-1}$ en el 7 para hallar la solución intermedia $\Delta\mathbf{a}_n$. En los trabajos de Hager y Belhumeur (1998) y Baker y Matthews (2001) se ofrece una solución para reducir la carga del segundo problema en el caso de usar una plantilla como referencia.

La idea consiste en linealizar la función a minimizar (3.13) respecto la referencia I^{ref} en vez de hacerlo con respecto a la imagen de entrada I_n para conseguir que las derivadas

```
[a_n] = Seguimiento_Multirresolución(I_n, a_{n-1}, P(o^{ref}))

    Entrada:
    I_n: Imagen que contiene el objeto a localizar
    a_{n-1}: Localización del objeto en la imagen anterior
    P(o^{ref}): Pirámide resolutiva de la referencia
    Salida:
    a_n: Parámetros de localización del objeto en I_n

    1. Construir P(I_n)
    2. Fijar r = R
    3. Adaptar parámetros de traslación según resolución
    4. [a_{n-1}] = Seguimiento_Básico(I_n^r, a_{n-1}, o^{ref,r})
    5. Si r > 1 hacer r = r - 1 y volver a 3
    6. a_n = a_{n-1}
```

Figura 3.7: Algoritmo de seguimiento multirresolutivo. Como se puede apreciar, se basa en el de seguimiento básico.

contenidas en la matriz $\mathbf{A}_n$ (3.26) sean siempre las mismas (es decir, las de la referencia):

$$E(\mathbf{a}_n) \approx \sum_{\forall \mathbf{x} \in \mathbf{\Pi}} \left(I^{ref}(\mathbf{x}) + \nabla I^{ref}(\mathbf{x})^T \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} - I_n(g(\mathbf{x}, \mathbf{a}_{n-1})) \right)^2 \tag{3.41}$$

$$\nabla I^{ref} = \begin{bmatrix} \frac{\partial I^{ref}}{\partial x} \\ \frac{\partial I^{ref}}{\partial y} \end{bmatrix} = \begin{bmatrix} I_x^{ref} \\ I_y^{ref} \end{bmatrix}$$

donde el movimiento final se consigue mediante la composición inversa (3.42) de los antiguos parámetros de movimiento con los nuevos incrementos, definiendo la variación de los puntos tras la transformación de similaridad (3.17) y minimizando la expresión (3.41).

$$g(\mathbf{x}, \mathbf{a}_n) = g^{-1}(g(\mathbf{x}, \mathbf{a}_{n-1}), \Delta\mathbf{a}_n) \tag{3.42}$$

Utilizando el objeto de referencia (3.19) y el transformado (3.20) y aplicando notación vectorial sobre (3.41), se puede reescribir la función aproximada del error (3.18) como:

$$\widetilde{E}(\Delta\mathbf{a}_n) = \left\| \mathbf{o}^{ref} - \mathbf{o}_n^w + diag(\Delta\mathbf{x})\, \mathbf{o}_x^{ref} + diag(\Delta\mathbf{y})\, \mathbf{o}_y^{ref} \right\|_2^2 \tag{3.43}$$

$$\mathbf{o}_x^{ref} = vec\left(I_x^{ref}(g(\mathbf{x}, \mathbf{a}_0)), \mathbf{\Pi} \right) \tag{3.44}$$

$$\mathbf{o}_y^{ref} = vec\left(I_y^{ref}(g(\mathbf{x}, \mathbf{a}_0)), \mathbf{\Pi} \right) \tag{3.45}$$

donde $\mathbf{o}_x^{ref}$ y $\mathbf{o}_y^{ref}$ se componen únicamente por aquellos elementos cuyas coordenadas $\mathbf{x}$ tienen un valor diferente de cero en $\mathbf{\Pi}$. La expresión de la función aproximada del error en notación vectorial (3.24) se mantiene, haciendo falta únicamente reescribir las expresiones

asociadas a la matriz del mismo ((3.27), (3.28), (3.29) y (3.30)) por (3.46), (3.47), (3.48) y (3.49):

$$\mathbf{A}^1 = diag\left(g_x\left(\mathbf{x}, \mathbf{a}_0\right)\right)\mathbf{o}_x^{ref} + diag\left(g_y\left(\mathbf{x}, \mathbf{a}_0\right)\right)\mathbf{o}_y^{ref} \tag{3.46}$$

$$\mathbf{A}^2 = diag\left(g_y\left(\mathbf{x}, \mathbf{a}_0\right)\right)\mathbf{o}_x^{ref} - diag\left(g_x\left(\mathbf{x}, \mathbf{a}_0\right)\right)\mathbf{o}_y^{ref} \tag{3.47}$$

$$\mathbf{A}^3 = \mathbf{o}_x^{ref} \tag{3.48}$$

$$\mathbf{A}^4 = \mathbf{o}_y^{ref} \tag{3.49}$$

y cambiarle el signo a la matriz $\mathbf{A}$ en los cálculos. Hay que destacar que la matriz $\mathbf{A}$ pierde el subíndice n ya que no depende de la imagen de entrada $\mathbf{I}_n$ puesto que está en función de los datos de la referencia $\mathbf{o}^{ref}$, que se puede suponer fija; de este modo, $\mathbf{A}$ queda constante a lo largo de todo el proceso.

$$\Delta\mathbf{a}_n = \mathbf{M}\mathbf{b}_n \tag{3.50}$$

$$\mathbf{M} = \left(\left(-\mathbf{A}^T\right)\left(-\mathbf{A}\right)\right)^{-1}\left(-\mathbf{A}^T\right) = -\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T \tag{3.51}$$

El proceso de minimización no se modifica respecto al presentado en el caso básico (3.31). Como $\mathbf{A}$ es constante en este caso, se puede precalcular la matriz $\mathbf{M}$ (3.51) y usarla durante todo el proceso para calcular la solución. Además, el proceso de interpolación sólo es necesario para encontrar el objeto transformado $\mathbf{o}_n^w$. Finalmente, para obtener los nuevos parámetros $\mathbf{a}_n$ a partir de $\mathbf{a}_{n-1}$ y $\Delta\mathbf{a}_n$, se debe utilizar una expresión parecida al modelo de similaridad en notación homogénea (3.33), pero teniendo en cuenta que en este caso se usa un esquema composicional inverso (3.42), quedando:

$$\begin{bmatrix} 1+a_n^1 & a_n^2 & a_n^3 \\ -a_n^2 & 1+a_n^1 & a_n^4 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1+\Delta a_n^1 & \Delta a_n^2 & \Delta a_n^3 \\ -\Delta a_n^2 & 1+\Delta a_n^1 & \Delta a_n^4 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1+a_{n-1}^1 & a_{n-1}^2 & a_{n-1}^3 \\ -a_{n-1}^2 & 1+a_{n-1}^1 & a_{n-1}^4 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.52}$$

Optimización del algoritmo utilizando un subespacio como referencia A continuación se ofrece un desarrollo similar al anterior, pero usando un subespacio como referencia, en base al trabajo desarrollado por Buenaposada et al. (2004). No obstante, el lector descubrirá que las ventajas ofrecidas serán inferiores respecto a las del caso anterior, debido a que no se puede precalcular totalmente la matriz $\mathbf{A}$ (3.51) ya que la referencia no es constante, sino que depende de la imagen observada $\mathbf{I}_n$. Se obtiene una mejora en la cantidad de operaciones a realizar (o carga computacional) respecto al caso básico, aunque no así del orden del mismo (o coste). La función objetivo a minimizar sigue siendo la versión linealizada respecto a la referencia (3.41), pero incluyendo, como dicha referencia, la reconstrucción del objeto transformado ($\mathbf{o}_n^w$ en (3.35) aplicado sobre (3.34) para obtener la referencia $\mathbf{o}_n^{ref}$) mediante el subespacio $\mathbf{U}$, la información media $\bar{\mathbf{o}}$ y un vector de parámetros $\mathbf{c}_n$, que depende del instante n. Hasta el final de esta sección y por cuestiones de simplicidad, se seguirá suponiendo una única región de interés, obviando los superíndices asociados a su identificación.

```
[a_n] = Seguimiento_Inverso(I_n, a_{n-1}, o^{ref}, M)

    Entrada:
    I_n: Imagen que contiene el objeto a localizar
    a_{n-1}: Localización del objeto en la imagen anterior
    o^{ref}: Vector con los datos de la imagen de referencia
    M: Matriz precalculada según (3.51)
    Salida:
    a_n: Parámetros de localización del objeto en I_n

    1. Construir o_n^w con a_{n-1} según (3.20)
    2. Construir b_n según (3.25) utilizando o^{ref}
    3. Calcular Δa_n siguiendo (3.50) utilizando M
    4. Actualizar los parámetros a_{n-1} con Δa_n (3.52)
    5. Volver a 1 hasta fin de iteraciones
    6. a_n = a_{n-1}
```

Figura 3.8: Algoritmo de seguimiento con movimiento composicional inverso.

La función a minimizar, en este caso, dependerá de los vectores $\mathbf{c}_n$. El movimiento final se consigue mediante el modelo composicional inverso presentado en (3.42). Utilizando notación vectorial sobre la función de error linealizada respecto a la referencia (3.41) y sabiendo que la referencia $\mathbf{o}_n^{ref}$ es diferente para cada valor de n, se puede obtener una expresión (3.53) similar a la obtenida en el caso inmediatamente anterior (3.43), manteniendo la expresión del objeto transformado $\mathbf{o}_n^w$ (3.20), sustituyendo la expresión de la referencia estática $\mathbf{o}^{ref}$ (3.19) por la reconstrucción de $\mathbf{o}_n^w$ (3.34) y cambiando las expresiones de las derivadas espaciales de la referencia estática de (3.44) y (3.45) por las propuestas en (3.54) y (3.55), que tienen en cuenta el vector de parámetros $\mathbf{c}_n$:

$$\widetilde{E}\left(\Delta\mathbf{a}\right) = \left\| \mathbf{o}_n^{ref} - \mathbf{o}_n^w + diag\left(\Delta\mathbf{x}\right)\mathbf{o}_{x,n}^{ref} + diag\left(\Delta\mathbf{y}\right)\mathbf{o}_{y,n}^{ref} \right\|_2^2 \tag{3.53}$$

$$\mathbf{o}_{x,n}^{ref} = \mathbf{U}_x\mathbf{c}_n + \overline{\mathbf{o}}_x = \begin{bmatrix}\mathbf{U}_x & \overline{\mathbf{o}}_x\end{bmatrix}\begin{bmatrix}\mathbf{c}_n^T & 1\end{bmatrix}^T \tag{3.54}$$

$$\mathbf{o}_{x,n}^{ref} = \mathbf{U}_y\mathbf{c}_n + \overline{\mathbf{o}}_y = \begin{bmatrix}\mathbf{U}_y & \overline{\mathbf{o}}_y\end{bmatrix}\begin{bmatrix}\mathbf{c}_n^T & 1\end{bmatrix}^T \tag{3.55}$$

$$\mathbf{U}_x = \begin{bmatrix}\frac{\partial\mathbf{u}_1}{\partial x} & \cdots & \frac{\partial\mathbf{u}_R}{\partial x}\end{bmatrix}, \quad \mathbf{U}_y = \begin{bmatrix}\frac{\partial\mathbf{u}_1}{\partial y} & \cdots & \frac{\partial\mathbf{u}_R}{\partial y}\end{bmatrix}$$

$$\overline{\mathbf{o}}_x = \frac{\partial\overline{\mathbf{o}}}{\partial x}, \quad \overline{\mathbf{o}}_y = \frac{\partial\overline{\mathbf{o}}}{\partial y}$$

A continuación, se propone la formulación del problema a optimizar en este caso (3.53) siguiendo una notación vectorial (3.56) siguiendo las indicaciones del trabajo de Buenaposada et al. (2004), para obtener una representación de la función aproximada del error $\widetilde{E}$ factorizada en matrices que contengan, por separado, los diferentes coeficientes y expresiones constantes:

$$\widetilde{E}\left(\Delta\mathbf{a}_n, \mathbf{c}_n\right) = \left\|\mathbf{b}_n - \mathbf{A}^u \cdot \mathbf{A}_n^c \cdot \left(\Delta\mathbf{a}_n\right)\right\|_2^2 \tag{3.56}$$

donde:

$$\mathbf{b}_n = \mathbf{o}_n^{ref} - \mathbf{o}_n^{w} \tag{3.57}$$

$$\mathbf{A}^u = \begin{bmatrix}\mathbf{A}^{u,1} & \mathbf{A}^{u,2} & \mathbf{A}^{u,3} & \mathbf{A}^{u,4}\end{bmatrix} \tag{3.58}$$

$$\mathbf{A}^{u,1} = diag\left(g_x\left(\mathbf{x}, \mathbf{a}_{n-1}\right)\right)\begin{bmatrix}\mathbf{U}_x & \overline{\mathbf{o}}_x\end{bmatrix} + diag\left(g_y\left(\mathbf{x}, \mathbf{a}_{n-1}\right)\right)\begin{bmatrix}\mathbf{U}_y & \overline{\mathbf{o}}_y\end{bmatrix}$$

$$\mathbf{A}^{u,2} = diag\left(g_y\left(\mathbf{x}, \mathbf{a}_{n-1}\right)\right)\begin{bmatrix}\mathbf{U}_x & \overline{\mathbf{o}}_x\end{bmatrix} - diag\left(g_x\left(\mathbf{x}, \mathbf{a}_{n-1}\right)\right)\begin{bmatrix}\mathbf{U}_y & \overline{\mathbf{o}}_y\end{bmatrix}$$

$$\mathbf{A}^{u,3} = \begin{bmatrix}\mathbf{U}_x & \overline{\mathbf{o}}_x\end{bmatrix}$$

$$\mathbf{A}^{u,4} = \begin{bmatrix}\mathbf{U}_y & \overline{\mathbf{o}}_y\end{bmatrix}$$

$$\mathbf{A}_n^c = \begin{bmatrix}\mathbf{c}_n & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 1 & 0 & 0 & 0 \\ \mathbf{0} & \mathbf{c}_n & \mathbf{0} & \mathbf{0} \\ 0 & 1 & 0 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{c}_n & \mathbf{0} \\ 0 & 0 & 1 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{c}_n \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.59}$$

sabiendo que $\mathbf{0}$ es un vector columna de ceros de iguales dimensiones que $\mathbf{c}_n$.

El valor mínimo de la función aproximada del error a optimizar (3.56) se puede lograr siguiendo un esquema iterativo, en el que se mantiene $\mathbf{c}_n$ constante para encontrar $\Delta\mathbf{a}_n$ y viceversa (ver figura 3.9). El valor inicial de $\mathbf{c}_n$ viene dado por la reconstrucción del objeto transformada de entrada $\mathbf{o}_n^w$ (3.35), mientras que el de $\Delta\mathbf{a}_n$ se puede encontrar de forma similar al caso de seguimiento básico (3.32):

$$\Delta\mathbf{a}_n = \mathbf{M}_n\mathbf{b}_n \tag{3.60}$$

$$\mathbf{M}_n = \left(\left(\mathbf{A}^u\mathbf{A}_n^c\right)^T\mathbf{A}^u\mathbf{A}_n^c\right)^{-1}\left(\mathbf{A}^u\mathbf{A}_n^c\right)^T \tag{3.61}$$

Como se puede observar, se debe realizar una inversa cada vez, ya que, a diferencia de usar una referencia fija y linealizar respecto a ésta, la matriz $\mathbf{M}_n$ depende de cada imagen y no se puede precalcular. Únicamente se puede obtener con antelación $\left(\mathbf{A}^u\right)^T\mathbf{A}^u$.

Robustez Como se ha presentado en la medida de error de alineación (3.4), la solución al problema de seguimiento se extrae a partir de las diferencias (o residuos) observadas entre la observación $\mathbf{F}_n^w$ y la referencia $\mathbf{o}_n^{ref}$. Si existe alguna que, debido a oclusiones o cambios de iluminación (es decir, que se deje de cumplir la suposición de iluminación constante), deja de ser significativa, no debería tenerse en cuenta en la resolución, calificándola de *outlier*. Utilizando la norma euclídea $\|\cdot\|_2$, los *outliers* producen una desviación significativa respecto a la estimación correcta. Una posible solución para reducir en gran medida este efecto se basa en limitar la importancia de cada elemento (ver figura 3.10).

```
[a_n] = Seguimiento_Inverso_Subespacio(I_n, a_{n-1}, M, A^u)

    Entrada:
    I_n: Imagen que contiene el objeto a localizar
    a_{n-1}: Localización del objeto en la imagen anterior
    M: Modelo visual
    A^u: Precálculo especificado en (3.58)
    Salida:
    a_n: Parámetros de localización del objeto en I_n

    1. Construir o_n^w con a_{n-1} según (3.20)
    2. Obtener c_n según (3.35)
    3. Construir o_n^{ref} con c_n según (3.34)
    4. Construir A^c según (3.59)
    5. Construir b_n según (3.57)
    6. Construir M_n según (3.61)
    7. Calcular Δa_n siguiendo (3.60)
    8. Actualizar los parámetros a_{n-1} con Δa_n (3.52)
    9. Volver a 1 un número determinado de iteraciones
    10. a_n = a_{n-1}
```

Figura 3.9: Algoritmo de seguimiento bajo la suposición de movimiento composicional inverso y utilizando un subespacio como referencia.

La norma euclídea aplicada sobre los residuos e los pondera de forma supralineal. Usando una norma como la propuesta por German y McClure (1987) (3.62), se consigue limitar la importancia de los residuos e para $-\eta > e > \eta$, siendo $\eta = \frac{\sigma}{\sqrt{3}}$ los puntos de inflexión de (3.62), donde σ es una constante arbitraria a fijar.

$$\rho\left(e,\sigma\right) = \frac{e^2}{e^2+\sigma^2} \tag{3.62}$$

$$\psi\left(e,\sigma\right) = \frac{\partial\rho\left(e,\sigma\right)}{\partial e} = \frac{2e\sigma^2}{\left(e^2+\sigma^2\right)^2}$$

Se puede estimar de forma robusta la escala de los residuos utilizando la expresión (3.63) (Rousseeuw y Leroy, 1987). El valor obtenido se encuentra en la misma escala que e, con lo que se puede aplicar directamente en la norma (3.62) como σ.

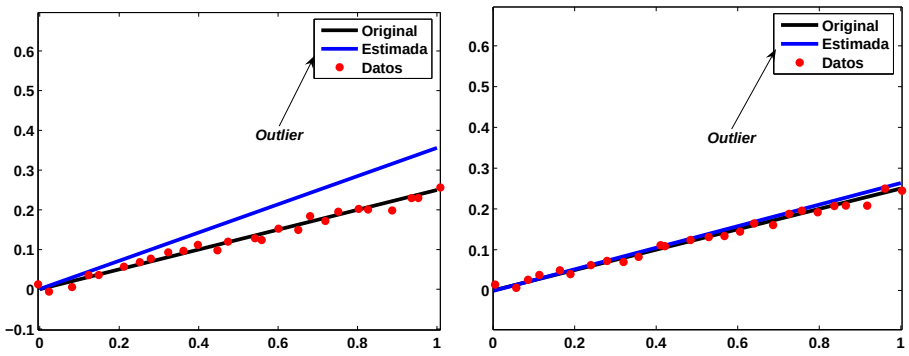
$$\sigma = 1'4826 \cdot \underset{\forall e\in\mathbf{R}}{\text{mediana}}\,|e| \tag{3.63}$$

Aplicando una norma general ϕ sobre el problema de seguimiento planteado a través de la función aproximada del error a minimizar ($\widetilde{E}$) expresada en notación vectorial (3.24):

$$\widetilde{E}\left(\Delta\mathbf{a}_n\right) = \sum_{\forall\mathbf{x}\in\mathbf{\Pi}} \phi\left(e\left(\mathbf{x},\Delta\mathbf{a}_n\right)\right) \tag{3.64}$$

donde $e\left(\mathbf{x},\Delta\mathbf{a}_n\right)$ representan los diferentes elementos del vector definido dentro de la norma euclídea presente en (3.24). Así, si $\phi = e^2$, se obtiene ese mismo planteamiento:

$$\widetilde{E}\left(\Delta\mathbf{a}_n\right) = \sum_{\forall\mathbf{x}\in\mathbf{\Pi}} \left(e\left(\mathbf{x},\Delta\mathbf{a}_n\right)\right)^2 \tag{3.65}$$



(a) Regresión con norma $\|\cdot\|_2$ (b) Regresión con norma robusta

Figura 3.10: Influencia de los *outliers* en la estimación con mínimos cuadrados: (a) sin norma robusta; (b) con norma robusta.

La derivada de la función de error anterior (3.65) respecto a $\Delta\mathbf{a}_n$ da como resultado:

$$\frac{\partial \widetilde{E}}{\partial \Delta\mathbf{a}_n}\left(\Delta\mathbf{a}_n\right) = \sum_{\forall \mathbf{x}\in\mathbf{\Pi}} 2\left(e\left(\mathbf{x},\Delta\mathbf{a}_n\right)\right)\frac{\partial e\left(\mathbf{x},\Delta\mathbf{a}_n\right)}{\partial \Delta\mathbf{a}_n}$$

cuya resolución, al igualar todas las expresiones a 0, se ofrece en (3.32) mediante el uso de la pseudoinversa. Por otro lado, si ϕ se define como la norma robusta a la de German y McClure (1987), es decir, $\rho\left(e,\sigma\right)$ (3.62), la función aproximada del error (3.64) pasa a ser:

$$\widetilde{E}\left(\Delta\mathbf{a}_n\right) = \sum_{\forall \mathbf{x}\in\mathbf{\Pi}} \rho\left(e\left(\mathbf{x},\Delta\mathbf{a}_n\right),\sigma\right)$$

y su derivada respecto a $\Delta\mathbf{a}_n$ se obtiene como:

$$\frac{\partial \widetilde{E}}{\partial \Delta\mathbf{a}_n}\left(\Delta\mathbf{a}_n\right) = \sum_{\forall \mathbf{x}\in\mathbf{\Pi}} \psi\left(e\left(\mathbf{x},\Delta\mathbf{a}_n\right),\sigma\right)\frac{\partial e\left(\mathbf{x},\Delta\mathbf{a}_n\right)}{\partial \Delta\mathbf{a}_n} \tag{3.66}$$

La resolución del sistema que se obtiene de igualar a 0 la expresion anterior (3.66) se puede realizar mediante minimización cuadrática ponderada iterativamente (IRLS) (Holland y Welsch, 1977), que se basa en asignar un valor ω a cada ecuación. Reescribiendo la derivada del error (3.66) como:

$$\frac{\partial \widetilde{E}}{\partial \Delta\mathbf{a}_n}\left(\Delta\mathbf{a}_n\right) = \sum_{\forall \mathbf{x}\in\mathbf{\Pi}} \omega\left(e\left(\mathbf{x},\Delta\mathbf{a}_n\right),\sigma\right)\left(e\left(\mathbf{x},\Delta\mathbf{a}_n\right)\right)\frac{\partial e\left(\mathbf{x},\Delta\mathbf{a}_n\right)}{\partial \Delta\mathbf{a}_n}$$

$$\omega\left(e\left(\mathbf{x},\Delta\mathbf{a}_n\right),\sigma\right) = \frac{\psi\left(e\left(\mathbf{x},\Delta\mathbf{a}_n\right),\sigma\right)}{\left(e\left(\mathbf{x},\Delta\mathbf{a}_n\right)\right)}$$

se puede expresar la función aproximada del error original (3.18) como:

$$\widetilde{E}\left(\Delta\mathbf{a}\right) = \left\|\mathbf{W}_n\left(\mathbf{o}^{ref} - \mathbf{o}_n^w - diag\left(\Delta\mathbf{x}\right)\mathbf{o}_{x,n}^w - diag\left(\Delta\mathbf{y}\right)\mathbf{o}_{y,n}^w\right)\right\|_2^2 \tag{3.67}$$

$$\mathbf{W}_n = diag\left(\omega\left(e\left(\mathbf{x},\Delta\mathbf{a}_n\right),\sigma\right)\right) \tag{3.68}$$

donde $\mathbf{W}_n$ es una matriz diagonal que contiene los diferentes pesos. Finalmente, la solución que minimiza la nueva función de error (3.67) respecto $\Delta\mathbf{a}_n$ viene dada por:

$$\Delta\mathbf{a}_n = \left(\mathbf{W}_n^T\mathbf{A}_n^T\mathbf{A}_n\mathbf{W}_n\right)^{-1}\mathbf{W}_n^T\mathbf{A}_n^T\mathbf{b}_n \tag{3.69}$$

Para una descripción más detallada del algoritmo de seguimiento robusto utilizado, consultar la figura 3.11.

```
[a_n, σ_n] = Seguimiento_Robusto(I_n, a_{n-1}, σ_{n-1}, o^{ref})

    Entrada:
    I_n: Imagen que contiene el objeto a localizar
    a_{n-1}: Localización del objeto en la imagen anterior
    σ_{n-1}: Valor anterior de σ en (3.62)
    o^{ref}: Vector con los datos de la imagen de referencia
    Salida:
    a_n: Parámetros de localización del objeto en I_n
    σ_n: Actualización de σ_{n-1}

    1. a = a_{n-1}
    2. Construir o_n^w con a según (3.20)
    3. Construir o_{x,n}^w con a según (3.21)
    4. Construir o_{y,n}^w con a según (3.22)
    5. Construir b_n según (3.25) y o^{ref}
    6. Construir A_n según (3.26)
    7. σ_n = mín (σ_{n-1}, 1'4826 · mediana |b_n|)
    8. Construir W_n con b_n y σ_n siguiendo (3.68)
    9. Calcular Δa_n siguiendo (3.69)
    10. Actualizar los parámetros a con Δa_n (3.33)
    11. Volver a 2 un número determinado de iteraciones
    12. a_n = a
```

Figura 3.11: Algoritmo de seguimiento con norma robusta de German-McLure.

3.2. Aprendizaje

El aprendizaje se define, en general, como «La adquisición del resultado de averiguar [...] la naturaleza, cualidades y relaciones de las cosas por medio [...] de la experiencia» (Real Academia Española, 2001). En este trabajo, se utiliza concretamente para definir **el conjunto de procesos orientados a construir** (en el sentido de encontrar o averiguar) **un modelo visual** (el cual resume toda la información que se desea contenida en un corpus, ver capítulo 2) **a partir de objetos sin movimiento rígido** (ver figura 3.12), los cuales se pueden conseguir aplicando un proceso de seguimiento como el del apartado 3.1).

El proceso de aprendizaje tiene como objetivo resumir la información de los objetos de interés presentes en una secuencia de imágenes. Se hace necesario este resumen ya que

la información visual es de dimensionalidad muy alta, lo cual hace costoso su tratamiento y almacenaje en un sistema de procesamiento digital. Por suerte, la redundancia presente en este tipo de señales bidimensionales suele ser también muy alta (Kirby, 2001), con lo que se pueden aplicar técnicas de reducción de dimensionalidad sobre ellas, obteniendo una representación compacta de las mismas. No obstante, esta representación viene referida respecto a un marco descriptivo o base. En el presente trabajo se utilizará la transformación lineal conocida como la transformada de Karhunen-Loeve (Karhunen, 1947, Loève, 1955), rebautizada más tarde como análisis de componentes principales (PCA) (Jolliffe, 1986), la cual elimina la redundancia de los datos de forma óptima. Su cálculo se ha particularizado en la técnica de descomposición matricial de SVD (Golub y Loan, 1996).

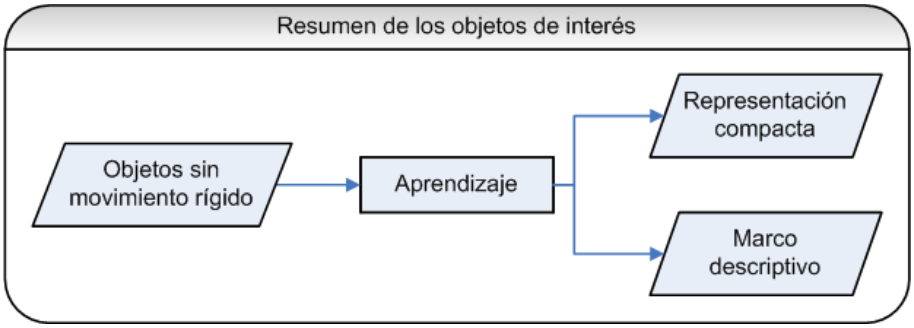


Figura 3.12: El proceso de aprendizaje busca la representación compacta del objeto mediante la especificación de un marco descriptivo óptimo.

3.2.1. Características

El proceso de aprendizaje se realiza sobre las apariencias visuales debidamente localizadas por el seguimiento (apartado 3.1), juntamente con información auditiva asociada al corpus audiovisual (apartado 2.1), si ésta se ha capturado. El aprendizaje es necesario para llevar a cabo el proceso de síntesis (ver capítulo 4) y es general, automático, incremental, causal y simultáneo con el de seguimiento.

3.2.1.1. General

El aprendizaje de la información visual mostrado en este trabajo es general ya que se puede aplicar a cualquier tipo de texturas e incluso de naturaleza de datos.

3.2.1.2. Automático

El proceso de aprendizaje no necesita ningún tipo de intervención manual gracias al uso del proceso de seguimiento del apartado 3.1. La inicialización también se puede llegar a realizar automáticamente mediante diferenciación de imágenes, morfología matemática y segmentadores de voz (ver apéndice C) en el caso de caras parlantes o fijando una región inicial predefinida para el caso de los objetos no rígidos.

3.2.1.3. Incremental

El modo en que se procesa la información es incremental gracias al uso del nuevo algoritmo propuesto para el cálculo incremental de la SVD con actualización de la media (ver apartado 3.2.4). Así, se van añadiendo nuevos datos constantemente, manteniendo una versión completamente actualizada del modelo visual durante todo el proceso y no solamente al final del mismo.

3.2.1.4. Causal

El proceso de actualización del modelo visual se realiza a partir de información presente y pasada, con lo que su implementación en tiempo real es plausible.

3.2.1.5. Simultáneo con el seguimiento

Los procesos de aprendizaje y seguimiento se necesitan mutuamente. El primero necesita al segundo para mantener una versión actualizada de forma precisa del modelo visual en todo momento y el segundo necesita al primero para poder mejorar la calidad de la alineación continuamente. Gracias a las características incrementales y causales de ambos procesos se puede conseguir una simultaneidad entre los mismos, de manera que la salida de uno alimenta la entrada del otro y viceversa (ver figura 3.2).

3.2.2. Representación compacta

La construcción de un modelo visual se realiza a partir de las imágenes alineadas y vectorizadas $\mathbf{o}_n^w$ mostradas en la expresión (3.5) del apartado 3.1. Estos vectores suelen ser de dimensión elevada (del orden de decenas de miles en imágenes de 320×240), con lo que se hace necesario un proceso de reducción de dimensionalidad, tal y como se ha comentado al inicio del apartado 3.2. Utilizando técnicas de análisis de componentes principales (PCA), se pueden encontrar las direcciones o componentes principales que describen la información de la forma más compacta posible (Jolliffe, 1986).

Una forma de obtener el PCA de un conjunto de datos se basa en el uso de la técnica de factorización matricial llamada descomposición en valores singulares (SVD) (Golub y Loan, 1996). Dado un conjunto de N apariencias visuales alineadas $\{\mathbf{o}_1^{w,l}, \mathbf{o}_2^{w,l}, \cdots \mathbf{o}_N^{w,l}\}$, para las L regiones visuales l, se construyen L matrices $\mathbf{O}^l$ poniendo en sus columnas los N vectores $\mathbf{o}_n^{w,l}$ (objetos transformados) de P^l píxeles cada uno (estos vectores reciben la notación simplificada de $\mathbf{o}_n$ en el apartado 2.2.3.1 para facilitar la lectura en dicho punto). Cada matriz $\mathbf{O}^l$, de rango máximo $R = \text{mín}\left(P^l, N\right)$, se puede expresar como el producto de tres matrices mediante la SVD:

$$\mathbf{O}^l_{P^l \times N} = \mathbf{U}^{i,l}_{P^l \times P^l} \mathbf{\Sigma}^{i,l}_{P^l \times N} \left(\mathbf{V}^{i,l}_{N \times N}\right)^T = \mathbf{U}^{i,l}_{P^l \times R} \mathbf{\Sigma}^{i,l}_{R \times R} \left(\mathbf{V}^{i,l}_{N \times R}\right)^T \tag{3.70}$$

donde $\mathbf{U}^{i,l} = [\mathbf{u}^{i,1}\ \mathbf{u}^{i,2}\ \cdots\ \mathbf{u}^{i,R}]$ y $\mathbf{V}^{i,l} = [\mathbf{v}^{i,1}\ \mathbf{v}^{i,2}\ \cdots\ \mathbf{v}^{i,R}]$ son matrices ortonormales que contienen, en sus columnas, los autovectores de $\mathbf{O}^l(\mathbf{O}^l)^T$ (o vectores singulares izquierdos) y de $(\mathbf{O}^l)^T\mathbf{O}^l$ (o vectores singulares derechos), respectivamente. La matriz diagonal $\mathbf{\Sigma}^{i,l} = diag(\sigma_1^{i,l}, \sigma_2^{i,l}, \ldots, \sigma_R^{i,l})$ contiene la raíz cuadrada de los R autovalores de las matrices de covarianzas $\mathbf{O}^l(\mathbf{O}^l)^T$ y $(\mathbf{O}^l)^T\mathbf{O}^l$ (ya que son los mismos) en su diagonal ordenados de mayor a menor; los elementos de la diagonal de $\mathbf{\Sigma}^{i,l}$ también reciben el nombre de valores singulares de $\mathbf{O}^l_{P^l\times N}$ y se denotan por σ_r^l. Las columnas de $\mathbf{U}^{i,l}$ abarcan el espacio columna de $\mathbf{O}^l_{P^l\times N}$; dada la naturaleza de esta matriz, se puede decir que $\mathbf{U}^{i,l}$ es una base generadora del subespacio de apariencia que contiene todos los objetos transformados $\mathbf{o}_n^{w,l}$, también conocidos a lo largo de este trabajo como las apariencias visuales. Las columnas de $\mathbf{V}^{i,l}$ abarcan el espacio fila de $\mathbf{O}^l_{P^l\times N}$, mientras que sus filas representan la misma información que $\mathbf{O}^l_{P^l\times N}$, aunque de forma blanqueada. De hecho, el producto de los valores singulares con los vectores singulares derechos (3.71) deja en cada columna de la matriz resultante $\mathbf{C}^{i,l} = \begin{bmatrix}\mathbf{c}^{i,1} & \mathbf{c}^{i,2} & \cdots & \mathbf{c}^{i,N}\end{bmatrix}$ la proyección de cada apariencia visual $\mathbf{o}_n^{w,l}$ sobre $\mathbf{U}^{i,l}$ (que son las unidades visuales reales del apartado 2.2.4.1).

$$\mathbf{C}^{i,l}_{R\times N} = \Sigma^{i,l}_{P^l\times N}\left(\mathbf{V}^{i,l}_{N\times N}\right)^T \tag{3.71}$$

Además, suponiendo que los vectores $\mathbf{c}^{i,l} = \begin{bmatrix}c_{n,1}^{i,l} & c_{n,2}^{i,l} & \cdots & c_{n,R}^{i,l}\end{bmatrix}^T$ siguen una distribución normal, aproximadamente el 92 % de ellos quedan acotados según un valor proporcional de sus desviaciones estándar, identificadas por los valores singulares $\sigma_r^{i,l}$:

$$-\sqrt{3}\sigma_r^{i,l} < c_{n,r}^{i,l} < \sqrt{3}\sigma_r^{i,l} \tag{3.72}$$

Así, se puede observar que los primeros autovectores de la base de apariencia son los que presentan los límites más amplios, mientras los últimos son los que poseen una menor variabilidad (debido a la ordenación decreciente de $\sigma_r^{i,l}$). De este modo, cada apariencia visual $\mathbf{o}_n^{w,l}$ de P^l píxeles queda representada con R coeficientes $c_{n,r}^{i,l}$. Mediante esta representación de $\mathbf{O}^l$ (3.70), se puede encontrar su mejor aproximación $\breve{\mathbf{O}}^l$ de rango K según la norma $\|\cdot\|_2$ (3.73) quedándose las primeras K columnas de $\mathbf{U}^{i,l}$ y $\mathbf{V}^{i,l}$, junto con los primeros K valores singulares de $\mathbf{\Sigma}^{i,l}$. Cualquier otra matriz de rango K y de dimensiones $P^l \times N$ será menos parecida según esta norma y tendrá un error mayor que el expresado en (3.74).

$$\underset{\text{rank}(\mathbf{B})}{\text{argmín}} \left\|\mathbf{O}^l_{P^l\times N} - \mathbf{B}_{P^l\times N}\right\|_2 = \breve{\mathbf{O}}^l_{P^l\times N} = \mathbf{U}^{i,l}_{P^l\times K}\mathbf{\Sigma^{i,l}}_{K\times K}\left(\mathbf{V}^{i,l}_{N\times K}\right)^T \tag{3.73}$$

$$\left\|\mathbf{O}^l - \breve{\mathbf{O}}^l\right\|_2 = \sigma^{i,l}_{K+1} \tag{3.74}$$

No obstante, el PCA descrito en (3.70) sólo es válido si los vectores $\mathbf{o}_n^{w,l}$ están centrados alrededor del centro de coordenadas, es decir, si su valor medio o centroide $\bar{\mathbf{o}}^l$ (3.76) es $\mathbf{0}$. Esto es debido a que el PCA se realiza sobre datos centrados (de media nula). Así, si el vector $\bar{\mathbf{o}}^l$ no es nulo, debe ser extraído de cada vector $\mathbf{o}_n^{w,l}$, obteniendo una matriz $\bar{\mathbf{O}}^l$ con las apariencias visuales centradas:

$$\bar{\mathbf{O}}^l_{P^l\times N} = \mathbf{O}^l_{P^l\times N} - \bar{\mathbf{o}}^l_{P^l\times 1}\mathbf{1}_{1\times N} \tag{3.75}$$

$$\bar{\mathbf{o}}^l = \frac{1}{N}\sum_{n=1}^{N}\mathbf{o}_n^{w,l} \tag{3.76}$$

Utilizando la mejor aproximación de rango K (3.73) sobre la matriz de datos centrados $\bar{\mathbf{O}}^l$ (3.75) se tiene:

$$\mathbf{O}^l_{P^l\times N} \approx \mathbf{U}^l_{P^l\times K}\mathbf{\Sigma}^l_{K\times K}\left(\mathbf{V}^l_{N\times K}\right)^T + \bar{\mathbf{o}}^l_{P^l\times 1}\mathbf{1}_{1\times N} \tag{3.77}$$

Adicionalmente, se debe destacar que si únicamente se desea reducir la dimensionalidad de los datos mediante la SVD, el hecho de que éstos no estén centrados alrededor de una media dificulta las tareas de clasificación (Hall et al., 2002).

En la figura 3.13 se puede observar la calidad conseguida utilizando un valor de $K = 10$ donde $R = 3612$ y $P = 40662$ (P no lleva superíndice ya que, en este ejemplo, sólo se utiliza una región visual, la cara). El error cometido viene indicado por σ^l_{11}, que se puede consultar en la figura 3.14 y es menor al 95 % de la energía total de la información original. Notar finalmente que en este ejemplo se consigue un importante factor de compresión de más de 4000 a 1.

Este tipo de aprendizaje de la apariencia visual es automático y causal, pero no incremental, ya que necesita todas las imágenes para obtener el marco descriptivo y su representación compacta (ver figura 3.12), por lo que no se puede aplicar simultáneamente con el proceso de seguimiento, sino después de que éste haya alineado todas las imágenes. No obstante, la alineación no podrá ser con respecto al subespacio de apariencia en un principio, dado que no existe. Este hecho conlleva la realización de varias iteraciones de seguimiento-aprendizaje, que ralentizan enormemente el proceso e impiden toda implementación en tiempo real, ya que se establece un esquema no causal. Con tal de conseguir la causalidad conjunta con el seguimiento, se propone un novedoso método de cálculo de PCA mediante combinaciones de la SVD en el apartado 3.2.3, aplicándolo a la construcción del subespacio de apariencia visual en el apartado 3.2.4.

3.2.3. Cálculo dinámico de la SVD

El cálculo de la SVD tiene un coste computacional y de memoria nada despreciables de $\mathcal{O}(p^2q+pq^2)$ y $\mathcal{O}(pq)$, respectivamente, para una matriz $\mathbf{M}_{p\times q}$ y usando un algoritmo de cálculo estándar (Golub y Loan, 1996). Este tipo de algoritmos tiene la desventaja añadida de la necesidad de disponer de la matriz $\mathbf{M}$ para comenzar a realizar los cálculos. En el caso de datos de alta dimensionalidad tales como secuencias de vídeo, este último hecho es vital, ya que pueden fácilmente acabar con los recursos de memoria de la máquina. A talle de ejemplo y para dar una idea de magnitudes, 12 minutos y medio de vídeo a color sin comprimir a una resolución de 320×240 y 25 imágenes por segundo consumirían los cuatro GBytes de RAM de que puede disponer un sistema de proceso estándar de hoy en día independientemente de la velocidad de procesador. Si, además, se une el hecho de que existe un interés creciente por el procesamiento de imágenes y secuencias de vídeo, resulta lógico que los esquemas de cálculo incrementales hayan cobrado importancia durante la última década, ya que no necesitan todos los datos de golpe para inicial los cálculos, sino

(a) Reales

(b) Reconstruidas

Figura 3.13: (a) Tres imágenes reales alineadas. (b) Las mismas tres imágenes reconstruidas por el modelo visual creado.

que los van solicitando poco a poco, siendo capaces de manejar secuencias mucho más largas sin consumir toda la memoria. La mayoría de esfuerzos se han dirigido en el cálculo incremental de la SVD, dando muy poca atención al conjunto de casos parecidos, como la SVD decremental (que experimenta pérdida de información, al revés que su contrapartida incremental) y la adición y sustracción de dimensiones a una o entre varias SVDs existentes. Se ofrecen las diferentes modalidades de cálculo eficientes desarrolladas en los apartados 3.2.3.1, 3.2.3.3, 3.2.3.4, 3.2.3.5, 3.2.3.6 y 3.2.3.7, así como un pequeño análisis del rendimiento teórico de cada una en el apartado 3.2.4.1. El contenido de estos subapartados se basa en una extensión del trabajo de Melenchón y Martínez (2007).

3.2.3.1. Modalidades desarrolladas

El autor ha extendido el trabajo de Brand (2002) en (Melenchón et al., 2004) mediante la proposición de un novedoso algoritmo que utiliza un proceso de reortonormali-

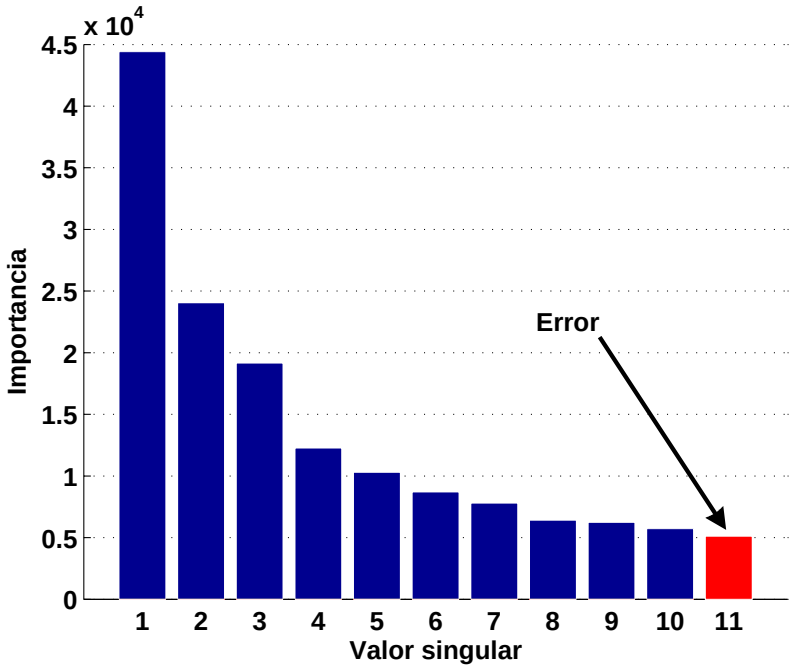


Figura 3.14: Diagrama de barras mostrando los valores singulares de la región visual l, perteneciente a la boca de una persona. En rojo se puede observar el valor singular σ_{11}^{l}.

zación eficiente, incorporando las ideas de actualización en bloque y la actualización de la información media de Hall et al. (2000) en el cálculo de la SVD. En (Melenchón y Martínez, 2007) se presentan diversas modalidades adicionales, las cuales se formalizan en esta tesis a partir de la reortonormalización eficiente explicada en el apartado 3.2.3.2 con cuatro supuestos para calcular SVDs de forma eficiente (ver cuadro 3.2): la incremental del apartado 3.2.3.4, la decremental del apartado 3.2.3.5, la compuesta del apartado 3.2.3.6 y la partida del apartado 3.2.3.7. Además, todos estos métodos tienen en cuenta la información media, ya que se benefician de la actualización eficiente de la media ofrecida en el apartado 3.2.3.3. Todos los métodos ofrecidos presentan una forma cerrada de cálculo y son extremadamente eficientes para SVDs de bajo rango.

3.2.3.2. Reortonormalización eficiente

Los métodos que aparecen en los siguientes cinco subapartados se basan en el proceso de reortonormalización eficiente descrito a continuación. Este proceso se basa en el cálculo eficiente de la SVD del producto matricial $\mathbf{QW}$ expresado en (3.78), donde $\mathbf{W}$ viene descompuesta en valores singulares (3.79) y $\mathbf{Q}^T\mathbf{Q}$ es un matriz identidad. Notar que $\mathbf{U} = \mathbf{QU}_W$.

$$\mathbf{H}_{M\times N} = \mathbf{Q}_{M\times K}\mathbf{W}_{K\times N} = \mathbf{U}_{M\times K}\mathbf{\Sigma}_{K\times K}\mathbf{V}^T_{N\times K} \tag{3.78}$$

$$\mathbf{W}_{K\times N} = (\mathbf{U}_W)_{K\times K}\,\mathbf{\Sigma}_{K\times K}\mathbf{V}^T_{N\times K} \tag{3.79}$$

Propuesta	Incremental	Decremental	Composición	Partición
Murakami y Kumar (1982)	sí	no	no	no
Chandrasekaran et al. (1997)	sí	no	no	no
Hall et al. (2000)	sí	sí	no	no
Hall et al. (2002)	sí	sí[1]	no	no
Brand (2002)	sí[2]	no	no	no
Skočaj y Leonardis (2003)	sí[3]	no	no	no
Lim et al. (2005)	sí[2]	no	no	no
Este trabajo	sí[4]	sí	sí	sí

Cuadro 3.2: Tipos de operaciones posibles entre varias SVDs realizadas eficientemente en diferentes propuestas. [1]: Hall et al. (2002) no ofrece una formulación cerrada para el cálculo de la SVD decremental. [2]: Brand (2002) y Lim et al. (2005) ofrecen la posibilidad de recordar menos las observaciones pasadas al ir incrementando la SVD, aunque sin olvidarlas. [3]: Skočaj y Leonardis (2003) dota de características robustas la SVD incremental. [4]: el presente trabajo permite un olvido total y selectivo de muestras pasadas, así como emular el propuesto por Brand (2002) y Lim et al. (2005).

Para comprobar la validez de (3.78), se muestra que los autovectores y autovalores de $\mathbf{H}^T\mathbf{H}$ y $\mathbf{H}\mathbf{H}^T$ se encuentran en las columnas de $\mathbf{V}$ y $\mathbf{U}$, los primeros, y en la diagonal de $\mathbf{\Sigma}^2$, los segundos. Empezando por la matriz de covarianzas $\mathbf{H}^T\mathbf{H}$:

$$\begin{aligned} \mathbf{H}^T\mathbf{H} &= \left(\mathbf{Q}\mathbf{U}_W\mathbf{\Sigma}\mathbf{V}^T\right)^T \mathbf{Q}\mathbf{U}_W\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}_W^T\mathbf{Q}^T\mathbf{Q}\mathbf{U}_W\mathbf{\Sigma}\mathbf{V}^T \\ \mathbf{H}^T\mathbf{H} &= \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T \\ \mathbf{H}^T\mathbf{H}\mathbf{V} &= \mathbf{V}\mathbf{\Sigma}^2 \end{aligned} \tag{3.80}$$

se puede comprobar que las columnas de $\mathbf{V}$ contienen los autovectores de la matriz $\mathbf{H}^T\mathbf{H}$, y la diagonal de $\mathbf{\Sigma}^2$, sus autovalores. Si se comprueba seguidamente la matriz de covarianzas $\mathbf{H}\mathbf{H}^T$:

$$\begin{aligned} \mathbf{H}\mathbf{H}^T &= \mathbf{Q}\mathbf{U}_W\mathbf{\Sigma}\mathbf{V}^T\left(\mathbf{Q}\mathbf{U}_W\mathbf{\Sigma}\mathbf{V}^T\right)^T = \mathbf{Q}\mathbf{U}_W\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}_W^T\mathbf{Q}^T \\ \mathbf{H}\mathbf{H}^T &= \mathbf{Q}\mathbf{U}_W\mathbf{\Sigma}^2\mathbf{U}_W^T\mathbf{Q}^T \\ \mathbf{H}\mathbf{H}^T\mathbf{Q}\mathbf{U}_W &= \mathbf{Q}\mathbf{U}_W\mathbf{\Sigma}^2 \\ \mathbf{H}\mathbf{H}^T\mathbf{U} &= \mathbf{U}\mathbf{\Sigma}^2 \end{aligned} \tag{3.81}$$

mediante la igualdad $\mathbf{U} = \mathbf{Q}\mathbf{U}_W$ se puede ver que $\mathbf{U}$ contiene, en sus columnas, los autovectores de la matriz $\mathbf{H}\mathbf{H}^T$; en este caso, los autovalores asociados están también en la diagonal de la matriz $\mathbf{\Sigma}^2$.

El caso en que la matriz $\mathbf{Q}^T$ se encuentre postmultiplicando a la SVD de una matriz $\mathbf{X}$ (3.82) se puede modelar como el caso traspuesto de la igualdad (3.78), escogiendo los valores de M y N adecuados. De este modo, al haberse comprobado la validez de (3.78) en (3.80) y (3.81), ésta se puede hacer extensible a (3.82).

$$\mathbf{X}_{S\times K}\mathbf{Q}_{R\times K}^T = \left(\mathbf{Q}_{R\times K}\mathbf{X}_{S\times K}^T\right)^T = \mathbf{H}_{R\times S}^T,\ R = M,\ S = N \tag{3.82}$$

3.2.3.3. Extracción de la media

Dada una matriz $\mathbf{A}$ de rango K y su SVD en (3.83), se desea obtener la columna media de $\mathbf{A}$ (3.84) y la SVD de la matriz centrada $\bar{\mathbf{A}}$ alrededor de esta columna media (3.85):

$$\mathbf{A}_{M\times N} = \mathbf{U}^i_{M\times K}\mathbf{\Sigma}^i_{K\times K}\left(\mathbf{V}^i_{N\times K}\right)^T \tag{3.83}$$

$$\bar{\mathbf{a}}_{M\times 1} = \frac{1}{N}\mathbf{A}_{M\times N}\cdot\mathbf{1}_{N\times 1} \tag{3.84}$$

$$\bar{\mathbf{A}}_{M\times N} = \mathbf{A}_{M\times N} - \bar{\mathbf{a}}_{M\times 1}\cdot\mathbf{1}_{1\times N} = \mathbf{U}_{M\times K}\mathbf{\Sigma}_{K\times K}\mathbf{V}^T_{N\times K} \tag{3.85}$$

Para conseguirlo, primeramente se extrae la fila media de $\mathbf{V}^i$ (3.86), la cual corresponderá a la proyección de $\bar{\mathbf{a}}$ sobre el subespacio $\mathbf{U}^i\mathbf{\Sigma}^i$. Extrayendo la fila media $\bar{\mathbf{v}}$ (3.86) de $\mathbf{V}^i$ (3.87) sobre la SVD inicial (3.83) y su versión centrada (3.85) se obtienen (3.88) y (3.89), respectivamente:

$$\bar{\mathbf{v}}_{1\times K} = \frac{1}{N}\mathbf{1}_N\cdot\mathbf{V}^i_{N\times K} \tag{3.86}$$

$$\bar{\mathbf{V}}^T_{N\times K} = \left(\mathbf{V}^i_{N\times K}\right)^T - \mathbf{1}_{N\times 1}\cdot\bar{\mathbf{v}}^T_{1\times K} \tag{3.87}$$

$$\mathbf{A} = \mathbf{U}^i\mathbf{\Sigma}^i\bar{\mathbf{V}}^T + \mathbf{U}^i\mathbf{\Sigma}^i\bar{\mathbf{v}}^T\cdot\mathbf{1} = \mathbf{U}^i\mathbf{\Sigma}^i\bar{\mathbf{V}}^T + \bar{\mathbf{a}}\cdot\mathbf{1} \tag{3.88}$$

$$\bar{\mathbf{A}} = \mathbf{U}^i\mathbf{\Sigma}^i\bar{\mathbf{V}}^T \tag{3.89}$$

Seguidamente, queda conseguir la SVD centrada (3.85) a partir del producto $\mathbf{U}^i\mathbf{\Sigma}^i\bar{\mathbf{V}}^T$, ya que $\bar{\mathbf{V}}^T$ no tiene porqué ser ortonormal después de la sustracción de $\bar{\mathbf{v}}$ a cada una de sus filas. Realizando la descomposición QR sobre $\bar{\mathbf{V}}$, se obtiene:

$$\bar{\mathbf{V}}_{N\times K} = \mathbf{Q}_{N\times K}\mathbf{R}_{K\times K} \tag{3.90}$$

$$\mathbf{U}^i\mathbf{\Sigma}^i\bar{\mathbf{V}}^T = \mathbf{U}^i\mathbf{\Sigma}^i\mathbf{R}^T\mathbf{Q}^T \tag{3.91}$$

seguidamente, realizando la SVD sobre la matriz $\mathbf{U}^i\mathbf{\Sigma}^i\mathbf{R}^T$, de tamaño $M\times K$, se obtiene la siguiente igualdad:

$$\mathbf{U}^i\mathbf{\Sigma}^i\mathbf{R}^T\mathbf{Q}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}_t^T\mathbf{Q}^T \tag{3.92}$$

y dada la unicidad de la SVD, la demostración del apartado 3.2.3.2 y que $\mathbf{U}\mathbf{\Sigma}\mathbf{V}_t^T\mathbf{Q}^T = \bar{\mathbf{A}}$ siguiendo (3.92), (3.91) y (3.89), se puede concluir que:

$$\mathbf{V} = \mathbf{Q}\mathbf{V}_t \tag{3.93}$$

Una observación debe ser efectuada antes de terminar este apartado. Dado que el proceso descrito se aplica a partir de una matriz $\bar{\mathbf{V}}$ no ortonormal (3.87), si ésta proviniese de una matriz $\mathbf{V}$ que no fuese ortonormal, el método continuaría siendo igualmente válido y eficiente. Este comentario es importante para el uso que se le da a este método en el apartado 3.2.3.5.

El algoritmo para obtener la SVD centrada de $\mathbf{A}$ (3.85) a partir de su SVD inicial (3.83) se detalla en la figura 3.15. El coste asociado del algoritmo se deduce en el cuadro 3.3.

$[\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}, \bar{\mathbf{a}}]$ = Extracción_Media($\mathbf{U}^i, \mathbf{\Sigma}^i, \mathbf{V}^i$)

Entrada:
$\mathbf{U}^i_{M\times K}$: Base del espacio columna de una matriz $\mathbf{A}$
$\mathbf{\Sigma}^i_{K\times K}$: Matriz diagonal con los valores singulares de $\mathbf{A}$
$\mathbf{V}^i_{N\times K}$: Base del espacio fila de $\mathbf{A}$
Salida:
$\bar{\mathbf{a}}_{M\times 1}$: Columna media de $\mathbf{A}$
$\mathbf{U}_{M\times K}$: Base del espacio columna de $\mathbf{A} - \bar{\mathbf{a}}$
$\mathbf{\Sigma}_{K\times K}$: Matriz diagonal con los valores singulares de $\mathbf{A} - \bar{\mathbf{a}}$
$\mathbf{V}_{N\times K}$: Base del espacio fila de $\mathbf{A} - \bar{\mathbf{a}}$

1. Calcular $\bar{\mathbf{v}}$ a partir de $\mathbf{V}^i$ con (3.86)
2. Calcular $\bar{\mathbf{V}} = \mathbf{V}^i - \mathbf{1} \cdot \bar{\mathbf{v}}$ (3.87)
3. Obtener $\bar{\mathbf{a}} = \mathbf{U}^i \mathbf{\Sigma}^i \bar{\mathbf{v}}^T$ (3.88)
4. Descomponer $\bar{\mathbf{V}}$ en $\mathbf{QR}$ según (3.90)
5. Calcular la SVD de $\mathbf{U}^i \mathbf{\Sigma}^i \mathbf{R}^T$ para obtener $\mathbf{U}\mathbf{\Sigma}\mathbf{V}_t^T$ (3.92)
6. Obtener $\mathbf{V}$ con $\mathbf{Q}$ y $\mathbf{V}_t$ según (3.93)

Figura 3.15: Algoritmo de extracción de la información media dada una SVD.

3.2.3.4. SVD incremental

En este apartado se expresa la relación algebraica entre la SVD de una matriz $\mathbf{A}$ expresada en (3.94) y la SVD de la misma matriz a la que se le han añadido columnas adicionales denotadas por la matriz $\mathbf{B}$ en (3.95).

$$\mathbf{A}_{M\times N} = \mathbf{U}^i_{M\times K} \mathbf{\Sigma}^i_{K\times K} \left(\mathbf{V}^i_{N\times K}\right)^T + \bar{\mathbf{a}}^i_{M\times 1} \cdot \mathbf{1}_{1\times N} \tag{3.94}$$

$$\begin{bmatrix}\mathbf{A}_{M\times N} & \mathbf{B}_{M\times R}\end{bmatrix} = \mathbf{U}^f_{M\times K} \mathbf{\Sigma}^f_{K\times K} \left(\mathbf{V}^f_{(N+R)\times K}\right)^T + \bar{\mathbf{a}}^f_{M\times 1} \cdot \mathbf{1}_{1\times (N+R)} \tag{3.95}$$

Para empezar, se utiliza la relación entre la media inicial $\mathbf{a}^i$, la final $\mathbf{a}^f$ y su variación $\Delta\mathbf{a}$ (3.96) y se centran la SVD inicial (3.94) y la que se le han añadido columnas (3.95), quedando (3.97) y (3.98), respectivamente, para usar en posteriores derivaciones.

$$\bar{\mathbf{a}}^f = \bar{\mathbf{a}}^i + \Delta\mathbf{a} \tag{3.96}$$

$$\mathbf{A} - \bar{\mathbf{a}}^i \cdot \mathbf{1} = \mathbf{U}^i \mathbf{\Sigma}^i \left(\mathbf{V}^i\right)^T \tag{3.97}$$

$$\begin{bmatrix}\mathbf{A} & \mathbf{B}\end{bmatrix} - \bar{\mathbf{a}}^i \cdot \mathbf{1} = \mathbf{U}^f \mathbf{\Sigma}^f \left(\mathbf{V}^f\right)^T + \Delta\mathbf{a} \cdot \mathbf{1} \tag{3.98}$$

Sustituyendo (3.97) en (3.98), se obtiene la expresión centrada:

$$\begin{bmatrix}\mathbf{A} - \bar{\mathbf{a}}^i \cdot \mathbf{1} & \mathbf{B} - \bar{\mathbf{a}}^i \cdot \mathbf{1}\end{bmatrix} = \begin{bmatrix}\mathbf{U}^i \mathbf{\Sigma}^i \left(\mathbf{V}^i\right)^T & \mathbf{B} - \bar{\mathbf{a}}^i \cdot \mathbf{1}\end{bmatrix} = \begin{bmatrix}\mathbf{U}^i \mathbf{\Sigma}^i \left(\mathbf{V}^i\right)^T & \bar{\mathbf{B}}\end{bmatrix} \tag{3.99}$$

donde $\bar{\mathbf{B}} = \mathbf{B} - \bar{\mathbf{a}}^i \cdot \mathbf{1}$ y los vectores fila $\mathbf{1}$ tienen el número de columnas adecuadas para poder realizar la resta matricial resultante en cada caso. A continuación se calcula la

Costes del algoritmo de extracción de la media a una SVD		
Operación	Coste computacional	Coste de memoria
$\bar{\mathbf{v}}_{1\times K} = \frac{1}{N}\mathbf{1}_N \cdot \mathbf{V}^i_{N\times K}$	$\mathcal{O}(NK)$	$\mathcal{O}(NK)$
$\bar{\mathbf{V}}_{N\times K} = \mathbf{V}^i_{N\times K} - \mathbf{1}_{N\times 1} \cdot \bar{\mathbf{v}}_{1\times K}$	$\mathcal{O}(NK)$	$\mathcal{O}(NK)$
$\bar{\mathbf{a}}_{M\times 1} = \mathbf{U}^i_{M\times K}\mathbf{\Sigma}^i_{K\times K}\bar{\mathbf{v}}^T_{1\times K}$	$\mathcal{O}\left(MK + K^2\right)$	$\mathcal{O}\left(MK + K^2\right)$
$\mathbf{Q}_{N\times K}\mathbf{R}_{K\times K} = \bar{\mathbf{V}}_{N\times K}$	$\mathcal{O}\left(NK^2\right)$	$\mathcal{O}\left(NK + K^2\right)$
$\mathsf{SVD}\left(\mathbf{U}^i_{M\times K}\mathbf{\Sigma}^i_{K\times K}\mathbf{R}^T_{K\times K}\right)$	$\mathcal{O}\left(MK^2 + K^3\right)$	$\mathcal{O}\left(MK + K^2\right)$
$\mathbf{V}_{N\times K} = \mathbf{Q}_{N\times K}\left(\mathbf{V}_t\right)_{K\times K}$	$\mathcal{O}\left(NK^2\right)$	$\mathcal{O}\left(NK + K^2\right)$
TOTAL	$\mathcal{O}\left(MK^2 + NK^2 + K^3\right)$	$\mathcal{O}\left(MK + NK + K^2\right)$
TOTAL SIMPLIFICADO	$\mathcal{O}\left(NK^2\right)$	$\mathcal{O}(NK)$

Cuadro 3.3: Análisis del coste computacional del algoritmo de extracción de la media a una SVD existente. El total simplificado asume que $M \sim N$, $K \ll M$, $K \ll N$.

descomposición QR de $\begin{bmatrix}\mathbf{U}^i & \bar{\mathbf{B}}\end{bmatrix}$:

$$\begin{bmatrix}\mathbf{U}^i & \bar{\mathbf{B}}\end{bmatrix} = \mathbf{QR} = \mathbf{Q}\begin{bmatrix}\mathbf{R}_U & \mathbf{R}_B\end{bmatrix} \tag{3.100}$$

donde $\mathbf{Q}$ es una matriz ortonormal de $M \times (K+R)$ y $\mathbf{R}$, otra triangular superior de $(K+R)\times(K+R)$; $\mathbf{R}_U$ contiene las primeras K columnas de $\mathbf{R}$ y $\mathbf{R}_B$, el resto. Utilizando la descomposición anterior (3.100), la expresión centrada (3.99) se puede continuar escribiendo como:

$$\left[\mathbf{U}^i\mathbf{\Sigma}^i\left(\mathbf{V}^i\right)^T \quad \bar{\mathbf{B}}\right] = \mathbf{Q}\left[\mathbf{R}_U\mathbf{\Sigma}^i\left(\mathbf{V}^i\right)^T \quad \mathbf{R}_B\right] \tag{3.101}$$

Si ahora se calcula la SVD de la matriz $\left[\mathbf{R}_U\mathbf{\Sigma}^i(\mathbf{V}^i)^T \quad \mathbf{R}_B\right]$, que es de $M\times(K+R)$ se tiene:

$$\mathbf{Q}\left[\mathbf{R}_U\mathbf{\Sigma}^i\left(\mathbf{V}^i\right)^T \quad \mathbf{R}_B\right] = \mathbf{Q}\mathbf{U}_t\mathbf{\Sigma}\mathbf{V}^T \tag{3.102}$$

Dadas (3.101), (3.102) y la demostración del apartado 3.2.3.2, se puede concluir que la expresión (3.103) obtiene la matriz $\mathbf{U}$ correspondiente a la SVD de $\left[\mathbf{U}^i\mathbf{\Sigma}^i(\mathbf{V}^i)^T \quad \bar{\mathbf{B}}\right]$ (3.104) inicialmente buscada .

$$\mathbf{U} = \mathbf{Q}\mathbf{U}_t \tag{3.103}$$

$$\left[\mathbf{U}^i\mathbf{\Sigma}^i\left(\mathbf{V}^i\right)^T \quad \bar{\mathbf{B}}\right] = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{3.104}$$

Dado que las expresiones (3.104) y (3.98) expresan la misma información centrada, se puede decir que:

$$\begin{bmatrix}\mathbf{A} & \mathbf{B}\end{bmatrix} - \bar{\mathbf{a}}^i \cdot \mathbf{1} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{U}^f\mathbf{\Sigma}^f\left(\mathbf{V}^f\right)^T + \Delta\mathbf{a}\cdot\mathbf{1}$$

lo cual implica que $\Delta\mathbf{a}$ se encuentra dentro de $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Para extraerla y obtener $\mathbf{U}^f\mathbf{\Sigma}^f\left(\mathbf{V}^f\right)^T$ se puede seguir la derivación indicada en el apartado 3.2.3.3 tomando $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ como la SVD

inicial de $\mathbf{A}$ en (3.83), $\mathbf{U}^f\mathbf{\Sigma}^f\left(\mathbf{V}^f\right)^T$ como la SVD centrada de $\mathbf{A}$ en (3.85) y la actualización de la información media $\Delta\mathbf{a}$ como $\bar{\mathbf{a}}$ en (3.84).

Una vez realizada la actualización, se deben despreciar las últimas R columnas de $\mathbf{U}^f$ y $\mathbf{V}^f$, así como los últimos R valores singulares de $\mathbf{\Sigma}^f$ si se desea mantener el rango de aproximación de la nueva matriz $\begin{bmatrix}\mathbf{A} & \mathbf{B}\end{bmatrix}$, dando un error de σ^l_{K+1}, suponiendo la $\|\cdot\|_2$.

En la figura 3.16 se puede observar el algoritmo para llevar a cabo una actualización incremental de la SVD y su coste asociado en el cuadro 3.4.

```
[U^f, Σ^f, V^f, ā^f] = SVD_Incremental(U^i, Σ^i, V^i, ā^i, B)

    Entrada:
    U^i_{M×K}, Σ^i_{K×K}, V^i_{N×K}: SVD de los datos iniciales
    ā^i_{M×1}: Media de los datos iniciales
    B_{M×R}: Bloque de actualización
    Salida:
    U^f_{M×K}, Σ^f_{K×K}, V^f_{N×K}: SVD de todos los datos
    ā^f_{M×1}: Media de todos los datos

    1. Calcular B̄ = B − ā · 1 (3.99)
    2. Obtener Q, R_U y R_B a partir de [U^i  B̄] (3.100)
    3. Calcular la SVD de [R_U Σ^i (V^i)^T  R_B] como U_t Σ V^T (3.102)
    4. Calcular U = Q U_t (3.103)
    5. [U^f, Σ^f, V^f, Δa] = Extracción_Media(U, Σ, V)
    6. Obtener ā^f a partir de ā^i y Δa con (3.96)
```

Figura 3.16: Algoritmo de cómputo de la SVD incremental.

Costes del algoritmo de cálculo de la SVD incremental		
Operación	Coste computacional	Coste de memoria
$\bar{\mathbf{B}}_{M\times R} = \mathbf{B}_{M\times R} - \bar{\mathbf{a}}_{M\times 1}\cdot\mathbf{1}_{1\times R}$	$\mathcal{O}(MR)$	$\mathcal{O}(MR)$
$\begin{bmatrix}\mathbf{U}^i & \bar{\mathbf{B}}\end{bmatrix}_{M\times S} = \mathbf{Q}_{M\times S}\begin{bmatrix}\mathbf{R}_U & \mathbf{R}_B\end{bmatrix}_{S\times S}$	$\mathcal{O}\left(MS^2\right)$	$\mathcal{O}\left(MS+S^2\right)$
$\text{SVD}\left(\begin{bmatrix}\mathbf{R}_U\mathbf{\Sigma}^i(\mathbf{V}^i)^T & \mathbf{R}_B\end{bmatrix}_{S\times(N+R)}\right)$	$\mathcal{O}\left((N+R)S^2+NSK+SK^2\right)$	$\mathcal{O}\left((N+R)S+S^2\right)$
$\mathbf{U}_{M\times S} = \mathbf{Q}_{M\times S}\left(\mathbf{U}_t\right)_{S\times S}$	$\mathcal{O}\left(MS^2\right)$	$\mathcal{O}\left(MS+S^2\right)$
$\text{Extracción\_Media}(\mathbf{U}_{M\times S},\mathbf{\Sigma}_{S\times S},\mathbf{V}_{N\times S})$	$\mathcal{O}\left(MS^2+NS^2+S^3\right)$	$\mathcal{O}\left(MS+NS+S^2\right)$
$\bar{\mathbf{a}}^f_{M\times 1} = \bar{\mathbf{a}}^i_{M\times 1} + \Delta\mathbf{a}_{M\times 1}$	$\mathcal{O}(M)$	$\mathcal{O}(M)$
TOTAL	$\mathcal{O}\left(MS^2+NS^2+S^3\right)$	$\mathcal{O}\left(MS+NS+S^2\right)$
TOTAL SIMPLIFICADO	$\mathcal{O}\left(NK^2\right)$	$\mathcal{O}(NK)$

Cuadro 3.4: Análisis del coste computacional del algoritmo de cálculo de la SVD incremental. $S = K + R$ y se usa para clarificar la notación. Notar que se utiliza el algoritmo de actualización de la media en este caso. El total simplificado se obtiene al usar $S = K + R$ y suponer $K \ll M$, $K \ll N$, $R \sim K$ y $M \sim N$.

3.2.3.5. SVD decremental

En este apartado se presenta una solución eficiente al problema de eliminar columnas de una SVD existente preservando la información media. El hecho de eliminar columnas de una matriz de datos puede ser considerado como una acción de olvido radical de esas columnas. Dada una matriz $\mathbf{A}$, su columna media $\bar{\mathbf{a}}$ y la SVD de la matriz $\mathbf{A}$ centrada sobre $\bar{\mathbf{a}}$ (3.105) se desea obtener el mismo tipo de descomposición sobre la matriz $\mathbf{D}$ (3.106), que contiene sólo L columnas de $\mathbf{A}$:

$$\mathbf{A}_{M\times N} = \mathbf{U}_{M\times K}\mathbf{\Sigma}_{K\times K}\mathbf{V}^T_{N\times K} + \bar{\mathbf{a}}_{M\times 1}\cdot\mathbf{1}_{1\times N} \tag{3.105}$$

$$\mathbf{D}_{M\times L} = \mathbf{U}^d_{M\times K}\mathbf{\Sigma}^d_{K\times K}\left(\mathbf{V}^d_{L\times K}\right)^T + \bar{\mathbf{a}}^d_{M\times 1}\mathbf{1}_{1\times L} \tag{3.106}$$

La expresión (3.106) se puede obtener a partir de (3.105) sin necesidad de recalcular $\mathbf{A}$. La matriz $\mathbf{D}$ se puede representar a partir de la SVDcentrada de $\mathbf{A}$ (3.105), pero eliminando las columnas de $\mathbf{V}$ correspondientes a las columnas que se han eliminado de $\mathbf{A}$ para obtener $\mathbf{D}$. Esta operación se puede modelar mediante la siguiente operación de multiplicación matricial:

$$\mathbf{D} = \mathbf{AZ} = \mathbf{U\Sigma V}^T\mathbf{Z} + \bar{\mathbf{a}}\cdot\mathbf{1}_A\cdot\mathbf{Z} = \mathbf{U\Sigma}\tilde{\mathbf{V}}^T + \bar{\mathbf{a}}\cdot\mathbf{1}_D \tag{3.107}$$

donde $\mathbf{Z}$ representa una matriz de $N\times L$ de ceros excepto en las posiciones $Z(v(i), i) = 1$, en las que el vector $\mathbf{v}$ contiene la lista de los índices de columna que se conservan de la matriz $\mathbf{A}$ e $\mathbf{i}$ es un vector de L elementos ordenados de 1 a L con incrementos unitarios. El vector fila $\mathbf{1}_A$ tiene N elementos, mientras que el $\mathbf{1}_D$ tiene sólo L. Dado que la fila media de $\tilde{\mathbf{V}}^T$ no tiene porqué ser ahora nula (sí lo era la de $\mathbf{V}^T$, dadas las condiciones iniciales supuestas), se debe extraer la información media de la misma y reortogonalizar el resultado. Para hacerlo, se puede usar el algoritmo del apartado 3.2.3.3 sobre $\mathbf{U\Sigma}\tilde{\mathbf{V}}^T$, obteniendo:

$$\mathbf{U\Sigma}\tilde{\mathbf{V}}^T = \mathbf{U}^d\mathbf{\Sigma}^d\left(\mathbf{V}^d\right)^T + \Delta\mathbf{a}\cdot\mathbf{1}_D \tag{3.108}$$

Finalmente, partiendo de la matriz $\mathbf{A}$ con las L columnas eliminadas (3.107), se obtiene la actualización de la media como:

$$\bar{\mathbf{a}}^d = \bar{\mathbf{a}} + \Delta\mathbf{a} \tag{3.109}$$

El lector puede encontrar en la figura 3.17 y en el cuadro 3.5, el algoritmo para ejecutar esta operación, así como su análisis de coste computacional y de memoria, respectivamente.

3.2.3.6. SVD compuesta

Dadas las matrices $\mathbf{A}$, $\mathbf{B}$ y sus SVDs (3.110), (3.111), la de la matriz $\mathbf{C} = \begin{bmatrix}\mathbf{A}^T & \mathbf{B}^T\end{bmatrix}^T$ (3.112) se puede obtener de forma eficiente a partir de ellas (3.117) si $M = L + P$ y $K = R + S$. Se supone, además, que $K \ll N$ y $K \ll M$.

$$\mathbf{A}_{L\times N} = \mathbf{U}^A_{L\times R}\mathbf{\Sigma}^A_{R\times R}\left(\mathbf{V}^A_{N\times R}\right)^T + \bar{\mathbf{a}}_{L\times 1}\cdot\mathbf{1}_{1\times N} \tag{3.110}$$

```
[U^d,Σ^d,V^d,ā^d] = SVD_Decremental(U,Σ,V,ā,v)

    Entrada:
    U_{M×K}, Σ_{K×K}, V_{N×K}: SVD de los datos iniciales
    ā_{M×1}: Media de los datos iniciales
    v: Vector que indica las columnas a conservar de V
    Salida:
    U^d_{M×K}, Σ^d_{K×K}, V^d_{L×K}: SVD de los datos finales
    ā_{M×1}: Media de los datos finales

    1. Obtener Ṽ a partir de V utilizando v (3.107)
    2. [U^d,Σ^d,V^d,Δa] = Extracción_Media(U,Σ,Ṽ)
    3. Obtener ā^d a partir de ā y Δa con (3.109)
```

Figura 3.17: Algoritmo de cálculo de la SVD decremental.

Costes del algoritmo de cálculo de la SVD decremental		
Operación	Coste computacional	Coste de memoria
Construir $\tilde{\mathbf{V}}_{L\times K}$ a partir de $\mathbf{V}_{N\times K}$	$\mathcal{O}(NK)$	$\mathcal{O}(NK)$
Extracción_Media($\mathbf{U}_{M\times K}$,$\mathbf{\Sigma}_{K\times K}$,$\tilde{\mathbf{V}}_{L\times K}$)	$\mathcal{O}\left(MK^2+LK^2+K^3\right)$	$\mathcal{O}\left(MK+LK+K^2\right)$
$\bar{\mathbf{a}}^d_{M\times 1} = \bar{\mathbf{a}}^d_{M\times 1} + \Delta\mathbf{a}_{M\times 1}$	$\mathcal{O}(M)$	$\mathcal{O}(M)$
TOTAL	$\mathcal{O}\left(MK^2+NK^2+ +NK+K^3\right)$	$\mathcal{O}\left(MK+NK+K^2\right)$
TOTAL SIMPLIFICADO	$\mathcal{O}\left(NK^2\right)$	$\mathcal{O}(NK)$

Cuadro 3.5: Análisis del coste computacional del algoritmo de SVD decremental. Notar que $L < N$. En el total simplificado se ha supuesto $L \sim N$, $M \sim N$, $K \ll N$ y $K \ll M$.

$$\mathbf{B}_{P\times N} = \mathbf{U}^B_{P\times S}\mathbf{\Sigma}^B_{S\times S}\left(\mathbf{V}^B_{N\times S}\right)^T + \bar{\mathbf{b}}_{P\times 1}\cdot\mathbf{1}_{1\times N} \tag{3.111}$$

$$\mathbf{C}_{M\times N} = \mathbf{U}^C_{M\times K}\mathbf{\Sigma}^C_{K\times K}\left(\mathbf{V}^C_{N\times K}\right)^T + \bar{\mathbf{c}}_{M\times 1}\cdot\mathbf{1}_{1\times N} \tag{3.112}$$

Se realiza la descomposición QR de $\begin{bmatrix}\mathbf{V}^A & \mathbf{V}^B\end{bmatrix}$ en (3.113) y, seguidamente, la SVD de $\begin{bmatrix}(\mathbf{U}^A\mathbf{\Sigma}^A\mathbf{R}^T_A)^T & (\mathbf{U}^B\mathbf{\Sigma}^B\mathbf{R}^T_B)^T\end{bmatrix}^T$ en (3.114) (con un coste menor que calcular la SVD de $[\mathbf{A}^T\mathbf{B}^T]^T$, ya que $K \ll N$).

$$\begin{bmatrix}\mathbf{V}^A & \mathbf{V}^B\end{bmatrix} = \mathbf{Q}\begin{bmatrix}\mathbf{R}_A & \mathbf{R}_B\end{bmatrix} \tag{3.113}$$

$$\begin{bmatrix}\mathbf{U}^A\mathbf{\Sigma}^A\mathbf{R}^T_A \\ \mathbf{U}^B\mathbf{\Sigma}^B\mathbf{R}^T_B\end{bmatrix} = \mathbf{U}^C\mathbf{\Sigma}^C\mathbf{V}^T_t \tag{3.114}$$

Basándose en la demostración del apartado 3.2.3.2, la matriz $\mathbf{V}^C$ se obtiene según (3.115). El vector $\bar{\mathbf{c}}$ se construye como la concatenación vertical de $\bar{\mathbf{a}}$ y $\bar{\mathbf{b}}$.

$$\mathbf{V}^C = \mathbf{Q}\mathbf{V}_t \tag{3.115}$$

$$\bar{\mathbf{c}} = \begin{bmatrix}\bar{\mathbf{a}} \\ \bar{\mathbf{b}}\end{bmatrix} \tag{3.116}$$

Finalmente, el proceso se puede resumir con la siguiente cadena de igualdades:

$$\begin{aligned}\mathbf{C} = \begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix} = \begin{bmatrix}\mathbf{U}^A\mathbf{\Sigma}^A\left(\mathbf{V}^A\right)^T\\ \mathbf{U}^B\mathbf{\Sigma}^B\left(\mathbf{V}^B\right)^T\end{bmatrix} + \begin{bmatrix}\bar{\mathbf{a}}\\ \bar{\mathbf{b}}\end{bmatrix} = \begin{bmatrix}\mathbf{U}^A\mathbf{\Sigma}^A\mathbf{R}_A^T\\ \mathbf{U}^B\mathbf{\Sigma}^B\mathbf{R}_B^T\end{bmatrix}\mathbf{Q}^T + \begin{bmatrix}\bar{\mathbf{a}}\\ \bar{\mathbf{b}}\end{bmatrix} = \\ = \mathbf{U}^C\mathbf{\Sigma}^C\mathbf{V}_t^T\mathbf{Q}^T + \bar{\mathbf{c}} = \mathbf{U}^C\mathbf{\Sigma}^C\left(\mathbf{V}^C\right)^T + \bar{\mathbf{c}}\end{aligned} \tag{3.117}$$

En esta derivación se han compuesto dos matrices en una, aunque se puede extender a múltiples matrices siempre que éstas posean el mismo número de columnas. El la figura 3.18 se puede consultar el algoritmo para la SVD compuesta y en el cuadro 3.6, su coste de ejecución y de consumo de recursos de almacenamiento.

$[\mathbf{U}^f, \mathbf{\Sigma}^f, \mathbf{V}^f, \bar{\mathbf{a}}^f]$ = SVD_Compuesta($\mathbf{U}^1 \dots \mathbf{U}^L, \mathbf{\Sigma}^1 \dots \mathbf{\Sigma}^L, \mathbf{V}^1 \dots \mathbf{V}^L, \bar{\mathbf{a}}^1 \dots \bar{\mathbf{a}}^L$)

Entrada:
$\left.\begin{matrix}\mathbf{U}^1_{M_1\times K_1}\cdots\mathbf{U}^L_{M_L\times K_L}\\ \mathbf{\Sigma}^1_{M_1\times K_1}\cdots\mathbf{\Sigma}^L_{M_L\times K_L}\\ \mathbf{V}^1_{N_1\times K_1}\cdots\mathbf{V}^L_{N\times K_L}\end{matrix}\right\}$: SVDs de los conjuntos de datos
$\bar{\mathbf{a}}^1 \dots \bar{\mathbf{a}}^L$: Medias de los conjuntos de datos
Salida:
$\mathbf{U}^f_{M\times K}$, $\mathbf{\Sigma}^f_{K\times K}$, $\mathbf{V}^f_{L\times K}$: SVD de todos los datos concatenados
$\bar{\mathbf{a}}^f_{M\times 1}$: Media de los datos concatenados

1. Descomposición QR a $\left[\mathbf{V}_1\cdots\mathbf{V}_L\right] = \mathbf{Q}\left[\mathbf{R}_1\cdots\mathbf{R}_L\right]$ siguiendo (3.113)
2. SVD de $\begin{bmatrix}\mathbf{U}^1\mathbf{\Sigma}^1\mathbf{R}_1^T\\ \vdots\\ \mathbf{U}^L\mathbf{\Sigma}^L\mathbf{R}_L^T\end{bmatrix} = \mathbf{U}^f\mathbf{\Sigma}^f\mathbf{V}_t^T$ siguiendo (3.114)
3. Calcular $\mathbf{V}^f = \mathbf{Q}\mathbf{V}_t$ siguiendo (3.115)
4. Obtener $\bar{\mathbf{c}} = \begin{bmatrix}\bar{\mathbf{a}}_1\\ \vdots\\ \bar{\mathbf{a}}_L\end{bmatrix}$

Figura 3.18: Algoritmo de cálculo de la SVD compuesta.

3.2.3.7. SVD partida

Este caso es el opuesto al del apartado 3.2.3.6: dada una SVD (3.112), se desean obtener dos (o más) SVDs más pequeñas (3.110), (3.111), partiendo el subespacio del primero en dos (o más) subespacios de menor dimensionalidad (en este caso $K = R = S$, $K < N$ y $K < M$). Por claridad, se realizará la derivación para el caso de dos grupos, pero se ofrece el algoritmo para un caso genérico de grupos. Inicialmente se divide la matriz $\mathbf{U}^C$ en dos submatrices mediante la agrupación de sus filas en dos conjuntos (3.118). De forma similar se procede con el vector media global $\bar{\mathbf{c}}$ (3.119).

$$\mathbf{U}^C = \begin{bmatrix}\mathbf{U}^A_p\\ \mathbf{U}^B_p\end{bmatrix} \tag{3.118}$$

Costes del algoritmo de cálculo de la SVD compuesta		
Operación	Coste computacional	Coste de memoria
$\left[\mathbf{V}_1 \cdots \mathbf{V}_L\right]_{N\times\mathcal{K}} = \mathbf{Q}_{N\times\mathcal{K}} \left[\mathbf{R}_1 \cdots \mathbf{R}_L\right]_{\mathcal{K}\times\mathcal{K}}$	$\mathcal{O}\left(N\mathcal{K}^2\right)$	$\mathcal{O}\left(N\mathcal{K}\right)$
$\text{SVD}\left(\begin{bmatrix}\mathbf{U}^1\mathbf{\Sigma}^1\mathbf{R}_1^T\\ \vdots \\ \mathbf{U}^L\mathbf{\Sigma}^L\mathbf{R}_L^T\end{bmatrix}_{\mathcal{M}\times\mathcal{K}}\right)$	$\mathcal{O}\left(\mathcal{M}\mathcal{K}^2+\mathcal{K}^3\right)$	$\mathcal{O}\left(\mathcal{M}\mathcal{K}+\mathcal{K}^2\right)$
$\mathbf{V}^f_{N\times\mathcal{K}} = \mathbf{Q}_{N\times\mathcal{K}}\left(\mathbf{V}_t\right)_{\mathcal{K}\times\mathcal{K}}$	$\mathcal{O}\left(N\mathcal{K}^2\right)$	$\mathcal{O}\left(N\mathcal{K}\right)$
$\begin{bmatrix}\bar{\mathbf{a}}_1\\ \vdots \\ \bar{\mathbf{a}}_L\end{bmatrix}_{\mathcal{M}\times 1}$	$\mathcal{O}\left(\mathcal{M}\right)$	$\mathcal{O}\left(\mathcal{M}\right)$
TOTAL	$\mathcal{O}\left(\mathcal{M}\mathcal{K}^2+N\mathcal{K}^2+\mathcal{K}^3\right)$	$\mathcal{O}\left(\mathcal{M}\mathcal{K}+N\mathcal{K}+\mathcal{K}^2\right)$
TOTAL SIMPLIFICADO	$\mathcal{O}\left(N\mathcal{K}^2\right)$	$\mathcal{O}\left(N\mathcal{K}\right)$

Cuadro 3.6: Análisis del coste computacional del algoritmo de SVD compuesta. Por cuestiones de claridad se adopta la siguiente nomenclatura: $\mathcal{M} = \sum_{i=1}^{L} M_i$ y $\mathcal{K} = \sum_{i=1}^{L} K_i$. Se supone, además, que $\mathcal{K} \ll N$, $\mathcal{K} \ll M$ y $\mathcal{M} \sim N$.

$$\bar{\mathbf{c}} = \begin{bmatrix}\bar{\mathbf{a}}\\ \bar{\mathbf{b}}\end{bmatrix} \tag{3.119}$$

Después, para cada submatriz resultante se realiza la descomposición QR de la misma (3.120) seguida de la SVD conjuntamente con $\mathbf{\Sigma}^C$ y $\mathbf{V}^C$ (3.121) (con un coste menor que el que se obtendría al realizar la SVD sobre $\mathbf{U}_p^A\mathbf{\Sigma}^A(\mathbf{V}^A)^T$ y $\mathbf{U}_p^B\mathbf{\Sigma}^B(\mathbf{V}^B)^T$, ya que $K < M$).

$$\mathbf{U}_p^A = \mathbf{Q}^A\mathbf{R}^A, \quad \mathbf{U}_p^B = \mathbf{Q}^B\mathbf{R}^B \tag{3.120}$$

$$\mathbf{R}^A\mathbf{\Sigma}^C(\mathbf{V}^C)^T = \mathbf{U}_t^A\mathbf{\Sigma}^A(\mathbf{V}^A)^T, \quad \mathbf{R}^B\mathbf{\Sigma}^C(\mathbf{V}^C)^T = \mathbf{U}_t^B\mathbf{\Sigma}^B(\mathbf{V}^B)^T \tag{3.121}$$

Finalmente, con apoyo de la demostración dada en el apartado 3.2.3.2, se puede deducir que:

$$\mathbf{U}^A = \mathbf{Q}^A\mathbf{U}_t^A, \quad \mathbf{U}^B = \mathbf{Q}^B\mathbf{U}_t^B \tag{3.122}$$

quedando una derivación final como sigue:

$$\begin{aligned}\mathbf{C} &= \begin{bmatrix}\mathbf{U}_p^A\\ \mathbf{U}_p^B\end{bmatrix}\mathbf{\Sigma}^C\left(\mathbf{V}^C\right)^T + \begin{bmatrix}\bar{\mathbf{a}}\\ \bar{\mathbf{b}}\end{bmatrix} = \begin{bmatrix}\mathbf{Q}^A\mathbf{R}^A\mathbf{\Sigma}^C\left(\mathbf{V}^C\right)^T\\ \mathbf{Q}^B\mathbf{R}^B\mathbf{\Sigma}^C\left(\mathbf{V}^C\right)^T\end{bmatrix} + \begin{bmatrix}\bar{\mathbf{a}}\\ \bar{\mathbf{b}}\end{bmatrix} =\\ &= \begin{bmatrix}\mathbf{Q}^A\mathbf{U}_t^A\mathbf{\Sigma}^A\left(\mathbf{V}^A\right)^T\\ \mathbf{Q}^B\mathbf{U}_t^B\mathbf{\Sigma}^B\left(\mathbf{V}^B\right)^T\end{bmatrix} + \begin{bmatrix}\bar{\mathbf{a}}\\ \bar{\mathbf{b}}\end{bmatrix} = \begin{bmatrix}\mathbf{U}^A\mathbf{\Sigma}^A\left(\mathbf{V}^A\right)^T + \bar{\mathbf{a}}\\ \mathbf{U}^B\mathbf{\Sigma}^B\left(\mathbf{V}^B\right)^T + \bar{\mathbf{b}}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix}\end{aligned} \tag{3.123}$$

Dado que las filas de la matriz $\mathbf{C}$ se pueden reordenar con una matriz de permutación $\mathbf{P}$, obteniendo $\mathbf{C}_r$, también $\mathbf{U}^C$ y $\bar{\mathbf{c}}$ padecen el mismo reordenamiento con $\mathbf{P}$, quedando $\mathbf{U}_r^C$ y $\bar{\mathbf{c}}_r$ (3.124). Por consiguiente, las filas de una matriz $\mathbf{C}$ se pueden partir en los grupos que se desee: primero, se realiza un agrupamiento por reordenación, colocando las filas de

grupos similares juntas unas con otras; después, se puede aplicar el proceso de partición (3.123) a la SVD reordenada.

$$\mathbf{C}_r = \mathbf{PC} = \mathbf{PU}^C\mathbf{\Sigma}^C\left(\mathbf{V}^C\right)^T + \mathbf{P}\bar{\mathbf{c}} = \mathbf{U}_r^C\mathbf{\Sigma}^C\left(\mathbf{V}^C\right)^T + \bar{\mathbf{a}}_r^C \tag{3.124}$$

$[\mathbf{U}^1 \dots \mathbf{U}^L, \mathbf{\Sigma}^1 \dots \mathbf{\Sigma}^L, \mathbf{V}^1 \dots \mathbf{V}^L, \bar{\mathbf{a}}^1 \dots \bar{\mathbf{a}}^L]$ = SVD_Partida($\mathbf{U}^i, \mathbf{\Sigma}^i, \mathbf{V}^i, \bar{\mathbf{a}}^i, \mathbf{P}$)

Entrada:
$\mathbf{U}^i_{M\times K}$, $\mathbf{\Sigma}^i_{K\times K}$, $\mathbf{V}^i_{L\times K}$: SVD de los datos de entrada
$\bar{\mathbf{a}}^i_{M\times 1}$: Media de los datos de entrada
$\mathbf{P}$: Matriz de permutación de filas
Salida:
$\left.\begin{matrix}\mathbf{U}^1_{M_1\times K_1}\dots\mathbf{U}^L_{M_L\times K_L}\\ \mathbf{\Sigma}^1_{M_1\times K_1}\dots\mathbf{\Sigma}^L_{M_L\times K_L}\\ \mathbf{V}^1_{N_1\times K_1}\dots\mathbf{V}^L_{N\times K_L}\end{matrix}\right\}$: SVDs de los conjuntos de datos
$\bar{\mathbf{a}}^1 \dots \bar{\mathbf{a}}^L$: Medias de los conjuntos de datos

1. Reordenar las filas de $\mathbf{U}^i$ y $\bar{\mathbf{a}}^i$ con $\mathbf{P}$ (3.124)
2. Dividir $\mathbf{U}_r$ en L submatrices $\mathbf{U}_p^l$ siguiendo (3.118)
3. Dividir $\bar{\mathbf{a}}_r$ en L vectores $\mathbf{a}_p^l$ siguiendo (3.119)
4. Para $l = 1$ a L descomponer en QR $\mathbf{U}_p^l = \mathbf{Q}^l\mathbf{R}^l$ (3.120)
5. Para $l = 1$ a L, SVD de $\mathbf{R}^l\mathbf{\Sigma}^f(\mathbf{V}^f)^T = \mathbf{U}_t^l\mathbf{\Sigma}^l(\mathbf{V}^l)^T$ (3.121)
6. Para $l = 1$ a L calcular $\mathbf{U}^l = \mathbf{Q}^l\mathbf{U}_t^l$

Figura 3.19: Algoritmo de cálculo de la SVD partida.

Costes del algoritmo de cálculo de la SVD partida			
Operación		Coste computacional	Coste de memoria
Reordenar $\mathbf{U}^i_{M\times K}$ y $\bar{\mathbf{a}}^i_{M\times 1}$		$\mathcal{O}(M)$	$\mathcal{O}(M)$
Dividir $(\mathbf{U}_r)_{M\times K}$		$\mathcal{O}(M)$	$\mathcal{O}(M)$
Dividir $(\bar{\mathbf{a}}_r)_{M\times K}$		$\mathcal{O}(M)$	$\mathcal{O}(M)$
$(\mathbf{U}_p)^l_{M_l\times K} = \mathbf{Q}^l_{M_l\times K}\mathbf{R}^l_{K\times K}$	$(\times L)$	$\mathcal{O}\left(LMK^2\right)$	$\mathcal{O}(MK)$
$\text{SVD}\left(\mathbf{R}^l_{K\times K}\mathbf{\Sigma}^f_{K\times K}(\mathbf{V}^f)^T_{N\times K}\right)$	$(\times L)$	$\mathcal{O}\left(NLK^2 + LK^3\right)$	$\mathcal{O}\left(LK^2 + NLK\right)$
$\mathbf{U}^f_{M_l\times K} = \mathbf{Q}^l_{M_l\times K}\left(\mathbf{U}^l_t\right)_{K\times K}$	$(\times L)$	$\mathcal{O}\left(LMK^2\right)$	$\mathcal{O}(LMK)$
TOTAL		$\mathcal{O}\left(LMK^2+NLK^2+LK^3\right)$	$\mathcal{O}\left(MK+NLK+LK^2\right)$
TOTAL SIMPLIFICADO		$\mathcal{O}\left(NLK^2\right)$	$\mathcal{O}(NLK)$

Cuadro 3.7: Análisis del coste computacional del algoritmo de SVD partida. Tener en cuenta que las matrices $\mathbf{Q}$ y $\mathbf{R}^l$ son temporales y no se guardan en el resultado final. Notar también que $\sum_{l=1}^{L} M_l = M$. En este caso se supone $K \ll N$, $K \ll M_l$, $K \sim L$ y $M \sim N$.

3.2.4. Aprendizaje del modelo mediante SVD incremental

La descomposición en valores singulares teniendo en cuenta la información media (3.77) se puede lograr mediante el uso del algoritmo de SVD incremental propuesto en el

apartado 3.2.3.4. Dado que es un algoritmo iterativo se necesita partir de una inicialización, que consiste en la SVD de las primeras columnas de $\mathbf{O}^l$, habiéndoles restado su columna media inicialmente.

$$\mathbf{O}_0^l = \mathbf{U}_0^l \mathbf{\Sigma}_0^l \left(\mathbf{V}_0^l\right)^T + \bar{\mathbf{o}}_0^l \cdot \mathbf{1} \tag{3.125}$$

donde $\mathbf{1}$ es un vector fila de unos de tantos elementos como número de columnas tiene $\mathbf{O}_0^l$. En el caso especial en que $\mathbf{O}_0^l$ es un vector columna, $\bar{\mathbf{o}}_0^l = \mathbf{O}_0^l$, $\mathbf{U}_0^l$ consta de una única columna de norma unitaria, $\mathbf{\Sigma}_0^l = 0$ y $\mathbf{V}_0^l = 1$. Realizada la inicialización, se aplica el algoritmo de SVD incremental añadiendo nuevas columnas $\mathbf{B}_{i+1}^l$ a $\mathbf{O}_i^l$ para formar $\mathbf{O}_{i+1}^l$. Normalmente se opta por añadir una cantidad constante de columnas R, del mismo orden de magnitud de K: $R \sim K$ (ver apartado 5.2.1).

Si se opta por limitar el rango de la SVD a K se debe comprobar, al finalizar cada iteración i, el número de valores singulares S_i de $\mathbf{\Sigma}_i^l$. Si S_i es mayor que K se desechan las últimas $S_i - K$ columnas de $\mathbf{U}_i^l$ y $\mathbf{V}_i^l$, así como los últimos $S_i - K$ valores singulares de $\mathbf{\Sigma}_i^l$, obteniendo la mejor aproximación de rango K de $\left[\mathbf{U}_{i-1}^l \mathbf{\Sigma}_{i-1}^l (\mathbf{V}_{i-1}^l)^T + \bar{\mathbf{o}}_{i-1}^l \cdot \mathbf{1} \quad \mathbf{B}_i^l\right]$.

$$\left[\mathbf{U}_{i-1}^l \mathbf{\Sigma}_{i-1}^l \left(\mathbf{V}_{i-1}^l\right)^T + \bar{\mathbf{o}}_{i-1}^l \cdot \mathbf{1} \quad \mathbf{B}_i^l\right] \approx \mathbf{U}_i^l \mathbf{\Sigma}_i^l \left(\mathbf{V}_i^l\right)^T + \bar{\mathbf{o}}_i^l \cdot \mathbf{1} \tag{3.126}$$

$$Rango\left(\left[\mathbf{U}_{i-1}^l \mathbf{\Sigma}_{i-1}^l \left(\mathbf{V}_{i-1}^l\right)^T + \bar{\mathbf{o}}_{i-1}^l \cdot \mathbf{1} \quad \mathbf{B}_i^l\right]\right) = S_i$$

$$Rango\left(\mathbf{U}_i^l \mathbf{\Sigma}_i^l \left(\mathbf{V}_i^l\right)^T + \bar{\mathbf{o}}_i^l \cdot \mathbf{1}\right) = K \leq S_i$$

El algoritmo para llevar a cabo este procedimiento se puede ver en la figura 3.20. Aunque $\mathbf{O}^l$ se explicite como parámetro de entrada, hay que aclarar que se pueden ir suministrando columnas de ésta e ir utilizando las SVD intermedias $\mathbf{U}_i^l \mathbf{\Sigma}_i^l (\mathbf{V}_i^l)^T + \bar{\mathbf{o}}_i^l \cdot \mathbf{1}$ para otros usos, como el seguimiento basado en un subespacio del apartado 3.1.2.4. De este modo, a medida que se realiza el seguimiento (o alineación) de la secuencia y se van obteniendo los vectores $\mathbf{o}_n^{w,l}$, se puede ir actualizando el subespacio de apariencia, el cual se utilizará para alinear las siguientes imágenes y así sucesivamente. El detalle del coste computacional del algoritmo de la figura 3.20 se detalla en el cuadro 3.8.

3.2.4.1. Rendimiento teórico

En este apartado se comparan el coste computacional y de memoria asociados al cálculo de la SVD de rango K de forma directa y de forma incremental de una matriz $\mathbf{O}^l$ de dimensiones $P^l \times N$. El método directo tiene un coste computacional de $\mathcal{O}((P^l)^2 N + P^l N^2)$ y de memoria tal como $\mathcal{O}(P^l N)$. Por otro lado, y tal como se puede consultar en el cuadro 3.8, en el caso del cálculo incremental el coste computacional es $\mathcal{O}(N/R(P^l S^2 + N S^2 + S^3))$ y el de memoria, $\mathcal{O}(P^l S + N S + S^2)$, suponiendo un bloque de actualización de tamaño R y donde $S = R + K$. Suponiendo P^l y N de órdenes de magnitud similares y K de órdenes de magnitud inferiores a los de P^l y N, se puede obtener un análisis del comportamiento analizando R:

$[\mathbf{U}^l, \mathbf{\Sigma}^l, \mathbf{V}^l, \bar{\mathbf{o}}^l]$ = SVD_Incremental_Total($\mathbf{O}^l$, K, R, R_0)

Entrada:
$\mathbf{O}^l_{P^l \times N}$: Datos iniciales por columnas
K: Rango de la SVD final
R: Número de columnas del bloque de actualización
R_0: Número de columnas del bloque inicial de datos
Salida:
$\mathbf{U}^l_{P^l \times K}$, $\mathbf{\Sigma}^l_{K \times K}$, $\mathbf{V}^l_{N \times K}$: SVD de $\mathbf{O}^l$
$\bar{\mathbf{o}}_{P^l \times 1}$: Columna media de $\mathbf{O}^l$

1. $\mathbf{B}^l_0 = \left[\mathbf{o}^{w,l}_1 \quad \cdots \quad \mathbf{o}^{w,l}_{R_0}\right]$
2. Descomposición de $\mathbf{B}^l_0 = \mathbf{U}^l_0 \mathbf{\Sigma}^l_0 \left(\mathbf{V}^l_0\right)^T + \bar{\mathbf{o}}^l_0$ (3.125)
3. $i = 1$
4. $\mathbf{B}^l_i = \left[\mathbf{o}^{w,l}_{iR-R+R_0+1} \quad \cdots \quad \mathbf{o}^{w,l}_{iR+R_0}\right]$
5. $[\mathbf{U}^l_i, \mathbf{\Sigma}^l_i, \mathbf{V}^l_i, \bar{\mathbf{o}}^l_i]$ = SVD_Incremental($\mathbf{U}^l_{i-1}, \mathbf{\Sigma}^l_{i-1}, \mathbf{V}^l_{i-1}, \bar{\mathbf{o}}^l_{i-1}, \mathbf{B}^l_i$)
6. Aproximar $\mathbf{U}^l_i \mathbf{\Sigma}^l_i \left(\mathbf{V}^l_i\right)^T$ con rango K si $Rango(\mathbf{\Sigma}^l_i) > K$ (3.126)
7. $i = i + 1$

Si $iR < N$ volver a 4, sino acabar

Figura 3.20: Algoritmo de cálculo de la SVD incremental aplicada al cálculo de la SVD de una matriz cualquiera. Se supone que $\mathbf{O}^l$ tiene un número de columnas múltiplo de R, sino, el último bloque de actualización $\mathbf{B}^l_{div(N,R)+1}$ consta de $mod\,(N, R)$ columnas. $div\,(A, B)$ representa la división entera de A entre B, mientras que su resto viene definido por $mod\,(A, B)$.

1. Si R es de un orden de magnitud similar al de P^l y N, el coste computacional de la aproximación incremental es $\mathcal{O}(N^3)$ independientemente del valor de K, frente al mismo coste de la versión directa (ya que $P^l \sim N$). En cuanto al coste de memoria, el de la versión incremental también es el mismo que el de la directa, es decir $\mathcal{O}(N^2)$.

2. Si R es de órdenes de magnitud inferior a P^l y N y $R \sim K$, el coste computacional de la versión incremental pasa a ser $\mathcal{O}(N^2K)$ frente al $\mathcal{O}(N^3)$ de la directa. El coste en memoria también se reduce a $\mathcal{O}(NK)$ respecto al $\mathcal{O}(N^2)$ de la directa.

3. Si R se escoge tan pequeño como para ser inferior en órdenes de magnitud a K, el coste computacional de la versión incremental aumenta hasta $\mathcal{O}(N^2K^2)$, aunque el de memoria no cambia respecto al caso anterior. Este hecho puede suponer un aumento significativo del coste computacional respecto a los dos casos anteriores dependiendo de los ordenes de magnitud de N y K.

Como conclusión, el valor de R deseable sería el de órdenes de magnitus similares al de K, para asegurar un coste computacional y de memoria más reducidos que los pertenecientes al método de cálculo directo de la SVD. Estos resultados se encuentran resumidos en el cuadro 3.9

Costes del cálculo de la SVD mediante SVD incremental			
Operación		Coste computacional	Coste de memoria
Construcción de $\left(\mathbf{B}_0^l\right)_{P^l \times R_0}$		$\mathcal{O}\left(P^l R_0^2\right)$	$\mathcal{O}\left(P^l R_0\right)$
Construcción de $\mathbf{B}_{P^l \times R}^l$	$\left(\times \frac{N}{R}\right)$	$\mathcal{O}\left(NP^l R\right)$	$\mathcal{O}\left(P^l R\right)$
SVD_Incremental $\left(\mathbf{U}_{P^l \times K}^l, \mathbf{\Sigma}_{K \times K}^l, \mathbf{V}_{iR \times K}^l, \bar{\mathbf{o}}_{P^l \times 1}^l, \mathbf{B}_{P^l \times R}^l\right)$	$\left(\times \frac{N}{R}\right)$	$\mathcal{O}\left(\frac{N}{R}\left(P^l S^2 + \right.\right.$ $\left.\left. + NS^2 + S^3\right)\right)$	$\mathcal{O}\left(P^l S + NS + S^2\right)$
$\mathbf{U}_{P^l \times S}^l \mathbf{\Sigma}_{S \times S}^l \left(\mathbf{V}_{iR \times S}^l\right)^T$ a rango K	$\left(\times \frac{N}{R}\right)$	$\mathcal{O}\left(P^l K + NK + \right.$ $\left. + \frac{N^2}{R} K + K^2\right)$	$\mathcal{O}\left(P^l K + NK + K^2\right)$
TOTAL		$\mathcal{O}\left(\frac{N}{R}\left(P^l S^2 + \right.\right.$ $\left.\left. + NS^2 + S^3\right)\right)$	$\mathcal{O}\left(P^l S + NS + S^2\right)$
TOTAL SIMPLIFICADO		$\mathcal{O}\left(N^2 K\right)$	$\mathcal{O}\left(NK\right)$

Cuadro 3.8: Análisis de los costes del cálculo de la SVD mediante SVD incremental. Se toma $R_0 = R$ para simplificar la notación. Notar que todas las operaciones excepto la primera se realizan N/R veces. El coste en memoria siempre es el máximo entre las diferentes iteraciones, mientras que el coste computacional se acumula entre ellas. Por cuestiones de claridad, se usa $S = K + R$. Para la simplificación de coste dada, se supone que $K \ll P^l$, $K \ll N$, $P^l \sim N$ y $K \sim R$.

SVD Incremental			SVD Directa	
Suposición	Cálculo	Memoria	Cálculo	Memoria
$R \sim N$	$\mathcal{O}\left(N^3\right)$	$\mathcal{O}\left(N^2\right)$		
$R \sim K \ll N$	$\mathcal{O}\left(N^2 K\right)$	$\mathcal{O}\left(NK\right)$	$\mathcal{O}\left(N^3\right)$	$\mathcal{O}\left(N^2\right)$
$R \ll K$	$\mathcal{O}\left(N^2 K^2\right)$	$\mathcal{O}\left(NK\right)$		

Cuadro 3.9: Costes asociados a la SVD directa y a la SVD incremental de rango K de una matriz $\mathbf{A}_{P^l \times N}$. Se supone $P^l \sim N$. En la tabla se muestran los costes para diferentes magnitudes del tamaño R del bloque de actualización de la SVD incremental.

Análisis del error En el caso en que se ignore la información media o se realice su extracción al final el proceso, se puede obtener una medida del error cometido en el aprendizaje del modelo mediante SVD incremental. La medida de error ofrecida corresponde a una cota superior en el caso del algoritmo de la figura 3.20, ya que los datos se encuentran más fielmente descritos por éste último, al realizar un PCA bloque a bloque. Se parte de una matriz expresada en B bloques de R columnas y P filas cada uno $\mathbf{O} = \begin{bmatrix}\mathbf{O}_1 & \cdots & \mathbf{O}_B\end{bmatrix}$. La matriz $\mathbf{O}$ se puede expresar como sigue:

$$\begin{aligned}
\mathbf{O} &= \begin{bmatrix}\mathbf{O}_0 & \mathbf{O}_1 & \mathbf{O}_2 & \cdots & \mathbf{O}_B\end{bmatrix} = \begin{bmatrix}\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T + \mathbf{U}_{r,1}\mathbf{\Sigma}_{r,1}\mathbf{V}_{r,1}^T & \mathbf{O}_2 & \cdots & \mathbf{O}_B\end{bmatrix} = \\
&= \begin{bmatrix}\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T & \mathbf{O}_2 & \cdots & \mathbf{O}_B\end{bmatrix} + \begin{bmatrix}\mathbf{U}_{r,1}\mathbf{\Sigma}_{r,1}\mathbf{V}_{r,1}^T & \mathbf{0} & \cdots & \mathbf{0}\end{bmatrix} = \\
&= \begin{bmatrix}\mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^T & \cdots & \mathbf{O}_B\end{bmatrix} + \begin{bmatrix}\mathbf{U}_{r,2}\mathbf{\Sigma}_{r,2}\mathbf{V}_{r,2}^T & \cdots & \mathbf{0}\end{bmatrix} + \mathbf{R}_1 = \\
&= \mathbf{U}_B\mathbf{\Sigma}_B\mathbf{V}_B^T + \sum_{b=1}^{B}\mathbf{R}_b = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T + \bar{\mathbf{o}}\cdot\mathbf{1} + \mathbf{R}
\end{aligned} \tag{3.127}$$

donde $\mathbf{O}_0 = \mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0^T$ y $\mathbf{T}_b = \begin{bmatrix}\mathbf{U}_{b-1}\mathbf{\Sigma}_{b-1}\mathbf{V}_{b-1}^T & \mathbf{O}_b\end{bmatrix} = \mathbf{U}_b\mathbf{\Sigma}_b\mathbf{V}_b^T + \mathbf{U}_{r,b}\mathbf{\Sigma}_{r,b}\mathbf{V}_{r,b}^T$, siendo $\mathbf{U}_b\mathbf{\Sigma}_b\mathbf{V}_b^T$ la SVD de rango K de $\mathbf{T}_b$ y $\mathbf{U}_{r,b}\mathbf{\Sigma}_{r,b}\mathbf{V}_{r,b}^T$, la SVD de la diferencia $\mathbf{T}_b - \mathbf{U}_b\mathbf{\Sigma}_b\mathbf{V}_b^T$.

$\mathbf{R}_b$ se utiliza para identificar la matriz error al añadir el bloque $\mathbf{O}_b$. El error cometido al aproximar $\mathbf{O}$ por $\mathbf{U}_B\mathbf{\Sigma}_B\mathbf{V}_B^T$ es $\mathbf{R} = \sum_{b=1}^{B}\mathbf{R}_b = \mathbf{U}_R\mathbf{\Sigma}_R\mathbf{V}_R^T$, cuya magnitud se puede expresar como $E = \|\mathbf{R}\|_2^2$, la cual se puede calcular como $E = \left\|\mathbf{R}\mathbf{R}^T\right\|_2$ (3.128).

$$\begin{aligned}&\left\|\mathbf{R}\mathbf{R}^T\right\|_2 = \left\|\mathbf{U}_R\mathbf{\Sigma}_R^2\mathbf{U}_R^T\right\|_2 = \text{máx}\left(diag\left(\mathbf{\Sigma}_R^2\right)\right) = \\ &= \sigma_{R,1}^2 = \left(\text{máx}\left(diag\left(\mathbf{\Sigma}_R\right)\right)\right)^2 = \|\mathbf{R}\|_2^2\end{aligned} \tag{3.128}$$

Donde $diag(\mathbf{A})$ recoge únicamente los elementos de la diagonal de la matriz $\mathbf{A}$. Sustituyendo el valor de $\mathbf{R}$ de la expresión del error cometido por la SVD incremental (3.127) en la propiedad anterior (3.128), se tiene:

$$\begin{aligned}E = \left\|\mathbf{R}\mathbf{R}^T\right\|_2 = \left\|\sum_{b=1}^{B}\mathbf{R}_b\sum_{b=1}^{B}\mathbf{R}_b^T\right\|_2 = \left\|\sum_{b=1}^{B}\mathbf{R}_b\mathbf{R}_b^T + \sum_{\substack{c=1\\c\neq d}}^{B}\sum_{d=1}^{B}\mathbf{R}_c\mathbf{R}_d^T\right\|_2 = \\ = \left\|\sum_{b=1}^{B}\mathbf{R}_b\mathbf{R}_b^T\right\|_2 = \left\|\sum_{b=1}^{B}\mathbf{U}_b\mathbf{\Sigma}_b^2\mathbf{U}_b^T\right\|_2 \leq \sum_{b=1}^{B}\left\|\mathbf{\Sigma}_b^2\right\|_2 = \sum_{b=1}^{B}\sigma_{b,1}^2 = E_{umbral}\end{aligned} \tag{3.129}$$

ya que $\mathbf{R}_c\mathbf{R}_d^T = \mathbf{0}$ para $c \neq d$. La demostración de este hecho se puede realizar a partir de la actualización incremental de la SVD dada por Brand (2002). Sea $\mathbf{R}_d = \mathbf{R}_b$ y $\mathbf{R}_c = \mathbf{R}_{b+k}$, entonces:

$$\begin{aligned}&\mathbf{R}_{b+k}\mathbf{R}_b^T = \begin{bmatrix}\mathbf{U}_{r,b+k}\mathbf{\Sigma}_{r,b+k}\mathbf{V}_{r,b+k}^T & \mathbf{0}_{b+k}\end{bmatrix}\begin{bmatrix}\mathbf{U}_{r,b}\mathbf{\Sigma}_{r,b}\mathbf{V}_{r,b}^T & \mathbf{0}_b\end{bmatrix}^T = \\ &= \mathbf{U}_{r,b+k}\mathbf{\Sigma}_{r,b+k}\mathbf{V}_{r,b+k}^T\begin{bmatrix}\mathbf{U}_{r,b}\mathbf{\Sigma}_{r,b}\mathbf{V}_{r,b}^T & \mathbf{0}_k\end{bmatrix}^T = \mathbf{U}_{r,b+k}\mathbf{\Sigma}_{r,b+k}\mathbf{V}_{r,b+k}^T\begin{bmatrix}\mathbf{V}_{r,b}\\ \mathbf{0}_k^T\end{bmatrix}\mathbf{\Sigma}_{r,b}\mathbf{U}_{r,b}^T\end{aligned}$$

donde $\mathbf{0}_k$ es una matriz que contiene el exceso de ceros de $\mathbf{0}_b$ respecto de $\mathbf{0}_{b+k}$. $\mathbf{V}_{r,b+k}^T$ se puede expresar en función de $\mathbf{V}_b^T$ (3.130) (la cual cumple que $\mathbf{V}_b^T\mathbf{V}_{r,b} = \mathbf{0}$) y depende de las matrices $\mathbf{V}_{G,s}$ y $\mathbf{V}_{R,s}$, que se obtienen a partir del proceso de SVD incremental de Brand (2002) aplicado sobre $\mathbf{U}_s$, $\mathbf{\Sigma}_s$, $\mathbf{V}_s$ y $\mathbf{O}_{s+1}$ (3.131).

$$\begin{aligned}&\mathbf{V}_{r,b+k}^T = \hat{\mathbf{V}}_{R,b+k-1}^T\begin{bmatrix}\mathbf{V}_{b+k-1}^T & \mathbf{0}\\ \mathbf{0} & \mathbf{Id}\end{bmatrix} = \hat{\mathbf{V}}_{R,b+k-1}^T\begin{bmatrix}\hat{\mathbf{V}}_{G,b+k-2}^T\begin{bmatrix}\mathbf{V}_{b+k-2}^T & \mathbf{0}\\ \mathbf{0} & \mathbf{Id}\end{bmatrix} & \mathbf{0}\\ \mathbf{0} & \mathbf{Id}\end{bmatrix} = \\ &= \hat{\mathbf{V}}_{R,b+k-1}^T\begin{bmatrix}\hat{\mathbf{V}}_{G,b+k-2}^T\begin{bmatrix}\hat{\mathbf{V}}_{G,b+k-3}^T\begin{bmatrix}\cdots\begin{bmatrix}\mathbf{V}_b^T & \mathbf{0}\\ \mathbf{0} & \mathbf{Id}\end{bmatrix} & \cdots & \mathbf{0}\\ \vdots & \ddots & \\ \mathbf{0} & & \mathbf{Id}\end{bmatrix} & \mathbf{0}\\ \mathbf{0} & \mathbf{Id}\end{bmatrix} & \mathbf{0}\\ \mathbf{0} & \mathbf{Id}\end{bmatrix}\end{aligned} \tag{3.130}$$

$$\begin{bmatrix}\mathbf{\Sigma}_s & \mathbf{U}_s^T\mathbf{O}_{s+1}\\ \mathbf{0} & Q\left(\mathbf{O}_{s+1} - \mathbf{U}_s\mathbf{U}_s^T\mathbf{O}_{s+1}\right)\end{bmatrix} = \mathbf{U}_{t,s}\mathbf{\Sigma}_{t,s}\begin{bmatrix}\mathbf{V}_{G,s}^T\\ \mathbf{V}_{R,s}^T\end{bmatrix} \tag{3.131}$$

donde $Q(\mathbf{X})$ representa una base ortogonal del subespacio representado por las columnas de $\mathbf{X}$ y $\mathbf{U}_{t,s}$ y $\mathbf{\Sigma}_{t,s}$ representan parte de la SVD expresada en (3.131). Seguidamente, la

matriz $\mathbf{V}_{r,b+k}^T$ se puede reescribir utilizando (3.130) y dos matrices $\mathbf{X}$ e $\mathbf{Y}$, que están en función de $\mathbf{V}_{R,b+k-1}^T$ y las diferentes $\mathbf{V}_{G,s}^T$ (3.132) (y cuyo valor es irrelevante en el cálculo mostrado más adelante en (3.133)).

$$\mathbf{V}_{r,b+k}^T = \begin{bmatrix} \mathbf{X}\mathbf{V}_b^T & \mathbf{0} \\ \mathbf{0} & \mathbf{Y} \end{bmatrix} \tag{3.132}$$

De este modo, el producto $\mathbf{V}_{r,b+k}^T \begin{bmatrix} \mathbf{V}_b^T & \mathbf{0} \end{bmatrix}^T$ se puede expresar como (3.133), dando como resultado una matriz de ceros ya que $\mathbf{V}_b^T \mathbf{V}_{r,b} = \mathbf{0}$.

$$\begin{bmatrix} \mathbf{X}\mathbf{V}_b^T & \mathbf{0} \\ \mathbf{0} & \mathbf{Y} \end{bmatrix} \begin{bmatrix} \mathbf{V}^T \\ \mathbf{0}^T \end{bmatrix} = \mathbf{X}\mathbf{V}_b^T \mathbf{V}_{r,b} = \mathbf{0} \tag{3.133}$$

Finalmente, dado que la extracción de la media sobre la descomposición final $\mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B^T$ se realiza de forma exacta (ver apartado 3.2.3.3), el error producido con ella o sin ella es el mismo que el indicado por E o acotado por E_{umbral} (3.129).

3.2.4.2. Olvido parcial del modelo

En términos generales, la calidad que se puede conseguir en el aprendizaje de un modelo visual se rige por dos factores principales:

1. La complejidad de los datos originales.
2. Los grados de libertad que se permiten al modelo.

Cuanto más sencillos sean de modelar los datos (por ejemplo, cuanto menos varianza tengan) y cuantos más grados de libertad se den al modelo, mejor calidad se conseguirá en la construcción del modelo visual y en su uso para la posterior reproducción de su contenido. Por contra, aumentar su complejidad o disminuir los grados de libertad del modelo (por ejemplo, para reducir su consumo de memoria), tiene un efecto negativo sobre la calidad obtenida. Si no se desea renunciar a la calidad bajo estas condiciones, se puede encontrar una solución haciendo al modelo dependiente del tiempo, de tal manera que sólo represente la información de un margen temporal (reduciendo así la complejidad de los datos). De este modo, en cada instante t, el modelo $\mathcal{M}(t)$ será potencialmente diferente y sólo capaz de representar los datos, por ejemplo, desde $t - t_0$ hasta t, donde t_0 simboliza el margen temporal de observación. En la figura 3.21 se puede observar el ajuste del modelo (un segmento) a una nube de puntos a través del tiempo; si se intenta aproximar globalmente la nube (figura 3.21 (a), (b), (c) y (d)) se acaba obteniendo una aproximación de una determinada calidad; si, por otro lado, el modelo se construye "*olvidando*" los datos más antiguos, es decir, que en cada instante de tiempo se represente la región de datos más recientes de la nube (figura 3.21 (i), (j), (k) y (l)), se logra una calidad global mayor. Este comportamiento se puede obtener mediante el uso de la SVD incremental del apartado 3.2.3.4 y la SVD decremental del apartado 3.2.3.5. Los nuevos puntos se añaden mediante el algoritmo de la figura 3.16; seguidamente se descartan u olvidan la misma cantidad de entre los puntos más antiguos mediante el uso del algoritmo de la figura 3.17.

Otra aproximación que se puede tomar para tener en cuenta el efecto de "*olvido*" es la sugerida por Lim et al. (2005). Ésta se basa en ponderar los valores singulares iniciales $\mathbf{\Sigma}^i$ (3.101) por un factor $\lambda < 1$. El resultado obtenido es un olvido menos drástico que el planteado anteriormente (ver figura 3.21 (e), (f), (g) y (h)).

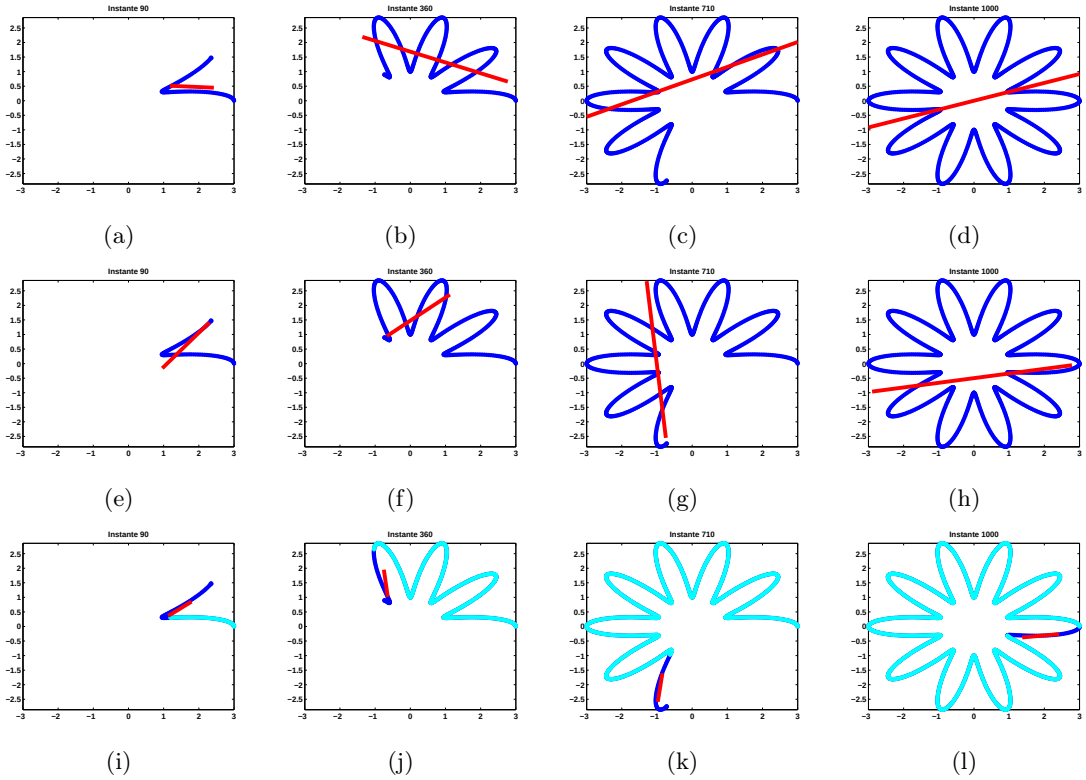


Figura 3.21: Se muestran diferentes instantes de la construcción del modelo que representa la nube de puntos. En azul oscuro se representan los puntos que tienen en cuenta el modelo en cada instante de tiempo y en azul pálido, los puntos que se ignoran por el mismo. La línea roja es la estimación de la dirección principal de los puntos azul oscuro. Las figuras (a), (b), (c) y (d) corresponden al modelo que representa todos los puntos observados a través del tiempo; (e), (f), (g) y (h) pertenecen al modelo que tiene en cuenta también todos los puntos, pero prestando mucho más interés en los puntos más recientes; finalmente, (i), (j), (k) y (l), se derivan del modelo que sólo tiene en cuenta el último conjunto de puntos observados en cada instante.

3.3. Algoritmo general de análisis

El algoritmo de análisis da como salida un modelo visual a partir de un corpus audiovisual (ver apartado 2.1) de entrada y se encuentra explicitado en la figura 3.22 y en la figura 3.2. Este algoritmo posee unas cualidades específicas que se detallan en el apartado 3.3.1, en base a sus partes constituyentes especificadas en los apartades 3.1, 3.2 y 2.2. Los pasos 5 y 7 son optativos, involucran el uso de información de voz y sólo se utilizan en el caso de modelar apariencias faciales.

```
[M] = Análisis(C)

    Entrada:
    C: Corpus audiovisual
    Salida:
    M: Modelo visual extraído del corpus audiovisual

    1. Inicialización de las máscaras (2.2)
    2. Obtención del modelo visual inicial (3.3)
    3. Alineación o seguimiento de un grupo de imágenes (3.1)
    4. Aprendizaje incremental del modelo visual (3.2)
    Volver a 3 hasta que no haya más imágenes
    5. Identificación de los visemas clave [OPTATIVO] (2.2)
    6. Construcción del grafo de coarticulación (2.2)
    7. Construcción del modelo acústico [OPTATIVO] (2.2)
```

Figura 3.22: Algoritmo de análisis. Entre paréntesis se indica la sección donde se explica cada proceso.

El paso 2 del algoritmo, que consiste en una inicialización del modelo visual, se consigue mediante los pasos 1 y 2 del algoritmo de aprendizaje del modelo mediante SVD incremental de la figura 3.20 y los primeros R_0 vectores $\mathbf{o}_n^{w,l}$ se alinean todos respecto a la forma media inicial, que se toma como $\bar{\mathbf{o}}^l = \mathbf{o}_1^{w,l}$.

3.3.1. Cualidades del análisis

Las técnicas descritas a lo largo de este capítulo han sido desarrolladas y utilizadas para dotar al algoritmo de análisis de la figura 3.22 con un cierto grado de fiabilidad, flexibilidad y facilidad de uso, así como para optimizar los recursos empleados y permitir su implementación en tiempo real con una máquina de cálculo adecuada (ver cuadro 3.10). El modelo visual empleado influye en las cualidades del análisis, ya que éste está orientado a construirlo.

3.3.1.1. Fiabilidad

La fiabilidad se muestra en disponer de un modelo visual acotado, en la simultaneidad del seguimiento y el aprendizaje y en la posibilidad de capturar los cambios de apariencia en el método de seguimiento. El primero ayuda a que no se puedan generar apariencias visuales irreales, mientras que el segundo y tercero ofrecen un nivel de adaptabilidad continuo respecto a la secuencia observada, lo cual permite localizar la apariencia de la cara de forma más precisa adaptándose a los cambios de apariencia que ésta pueda sufrir y aprenderlas de forma más adecuada gracias al mejor proceso de alineación o seguimiento.

Cualidad	Características
Fiabilidad	Modelo visual acotado, seguimiento y aprendizaje simultáneos y posibilidad de cambios de apariencia en el seguimiento
Flexibilidad	Aplicación generalizada del seguimiento y aprendizaje e independencia del fondo
Facilidad de uso	Seguimiento y aprendizaje automáticos, inexistencia de entrenamiento previo y seguimiento no intrusivo
Coste	Modularidad, compactación y cálculo incremental del modelo visual
Tiempo real	Causalidad del seguimiento y el aprendizaje

Cuadro 3.10: Cualidades del algoritmo de análisis y características asociadas del seguimiento y el aprendizaje.

3.3.1.2. Flexibilidad

El proceso de análisis posee características que aportan adaptabilidad, referidas a las condiciones externas y de aplicabilidad. Esta característica va muy ligada a la facilidad de uso, ya que cuanto más flexible es un algoritmo, menos restricciones se deben cumplir. Permitir cualquier fondo en determinados contextos (ver apartado 3.1.1.7) es un ejemplo de esto último. La generalidad de aplicación de los algoritmos de seguimiento y aprendizaje a cualquier tipo de objeto en una imagen ofrecen un grado alto de flexibilidad del algoritmo de análisis respecto a las apariencias visuales. Desgraciadamente, la dificultad del modelado de rotaciones tridimensionales no permite extenderla a una generalidad de movimiento, aunque si éstos se interpretan como cambios en la apariencia del objeto, pueden quedar recogidos dentro de la flexibilidad aportada para las apariencias visuales.

3.3.1.3. Facilidad de uso

Ésta es una de la características principales que persigue el algoritmo, ya que se desea que el trabajo aquí realizado se pueda utilizar de forma sencilla y sin la necesidad de intervención manual. Los elementos que facilitan el uso del algoritmo son la automatización de los procesos de seguimiento y aprendizaje, la inexistencia de entrenamiento previo y la no intrusividad del seguimiento.

En el algoritmo general de análisis (figura 3.22), los pasos 2, 3, 4, 6 y 7 no requieren interacción con el usuario ya que se pueden realizar de forma totalmente automática. Los pasos 2, 6 y 7 se encuentran descritos en el apartado 2.2, el 3, en el apartado 3.1 y, finalmente, el paso 4 se explica con detalle en el apartado 3.2. La interacción humana es recomendable en los pasos 1 y 5. Existen métodos para conseguir eliminar esta necesidad y obtener procesos totalmente automatizados. No obstante, estos métodos dependen fuertemente de las condiciones externas y el tipo de objeto a seguir, lo cual dificulta enormemente el hallazgo de un método automático general.

La obtención del conjunto de máscaras **Π** (paso 1 del algoritmo de la figura 3.22), clave para el análisis, no es más que un problema de segmentación de imágenes, del cual no se conoce aún su solución genérica (Micusik y Hanbury, 2005), con lo que se deben fijar restricciones para poder obtener resultados satisfactorios. El lector puede encontrar la descripción de una posible solución a este problema en el caso de regiones faciales en el apéndice C.

La identificación de visemas clave (paso 5 del algoritmo de la figura 3.22), de gran importancia particularmente para la síntesis de apariencias faciales, como se puede ver en el capítulo 4, se traduce en un problema de segmentación de voz, ya que gracias a la información temporal que produce este proceso se pueden determinar las correspondencias identificador-visema necesarias para la síntesis. Dado que se conoce el texto reproducido, el problema de segmentación simplemente debe marcar las transiciones entre los diferentes sonidos, sin preocuparse de reconocer el mensaje; este hecho permite el uso de segmentadores de voz generales (Young et al., 2003) con un grado de acierto aceptable, lo cual reduce ampliamente el grado de interacción humana necesaria. No obstante, y bajo la consideración de uso de cámaras web domésticas, aparece un problema técnico debido a la leve desincronización del contenido generado por los canales visuales y auditivos de las señales que capturan. Este hecho dificulta mucho la automatización del paso 6 ya que la desincronización es variable y desconocida. Si se desea eliminar la interacción humana en este paso es necesaria la utilización de un sistema de captura profesional. Bajo esta última observación, y únicamente con un objetivo puramente de investigación, se han añadido grabaciones profesionales en la captura del corpus de ejemplo comentado en el capítulo 5.

En el cuadro 3.11 se puede observar un resumen de la interacción humana necesaria en el algoritmo de análisis propuesto. Asimismo, se presenta su origen y cómo intentar evitarla. Notar que ningún paso posee una interacción humana estrictamente necesaria.

Paso	Interacción	Origen	Forma de evitarla
(1) Inicialización	Recomendable	Teórico	Restricciones adicionales
(2) Modelado	Nula		
(3) Seguimiento	Nula		
(4) Aprendizaje	Nula		
(5) Identificación	Recomendable	Técnico	Sistema profesional de captura
(6) Coarticulación	Nula		
(7) Acústica	Nula		

Cuadro 3.11: Grado de interacción humana asociada en cada paso del algoritmo de análisis y, si existe, origen y forma de evitarla.

3.3.1.4. Coste

Se persigue un coste computacional y un consumo de memoria lo más bajos posibles para aumentar el rendimiento de la implementación del algoritmo. La modularidad y

compactación del modelo visual, así como el método de cálculo incremental para la SVD usado en el aprendizaje facilitan la consecución de este objetivo.

3.3.1.5. Tiempo real

Pese a que las implementaciones actuales del algoritmo de análisis no consiguen un rendimiento de tiempo real (se llega hasta diez imágenes por segundo, aproximadamente), el algoritmo lo posibilita gracias a la causalidad de los procesos de seguimiento y aprendizaje. El tiempo real ofrece un nivel mucho más alto de interactividad con el usuario, revirtiendo positivamente en su facilidad de uso, ya que se permiten realimentaciones hacia el usuario mientras se produce la interacción con el sistema. Esta característica tiene su importancia en aplicaciones que van más allá del marco de trabajo propuesto pero que utilizan el mismo algoritmo de análisis.

Capítulo 4

Síntesis

Si se observan las definiciones de síntesis y análisis ofrecidas en el capítulo 3 y dentro del ámbito de este trabajo, el lector las encontrará totalmente complementarias: la síntesis persigue la unión, mientras que el análisis se plantea la disgregación; la síntesis tiene como objetivo ser observada, mientras que el análisis observa cierta información objetivo; la síntesis pretende producir señales audiovisuales a partir de un modelo visual, mientras que el análisis parte del mismo tipo de señales para generar dicho modelo (ver figura 3.1). Como se puede ver, el punto de unión de ambos procesos es, precisamente, el modelo visual (detallado en el apartado 2.2), ya que ambos lo usan; se podría decir que son procesos que existen en una misma dirección, pero con sentidos opuestos.

Las señales audiovisuales que se pueden generar, o sintetizar, han de venir guiadas o determinadas por algún tipo de información ya que el modelo visual actúa únicamente de contenedor de la información visual esencial, con todas las características descritas en el apartado 2.2.1. Una manera de guiar el proceso de síntesis es mediante información auditiva: si se conoce su relación con la información que contiene el modelo visual se puede establecer un vínculo y sintetizar señales audiovisuales sincronizadas a partir de esta información auditiva. Esta relación se recoge en el modelo acústico (ver apartado 2.2.5) y dada su existencia, se puede incorporar la señal auditiva con la secuencia de imágenes producidas por el modelo visual, obteniendo, así, señales audiovisuales. Si la relación audiovisual no existe, se generan únicamente señales visuales carentes de audio a partir de otro tipo de información.

El esquema de síntesis presentado consta de dos elementos básicos (ver figura 4.1):

- Un modelo visual, o contenedor de la información visual esencial, fuente de información para las señales audiovisuales (o únicamente visuales) generadas artificialmente (apartado 2.2). Destacar la importancia del modelo acústico (apartado 2.2.5) cuando se dispone de información auditiva.

- Un método constructor de imágenes con la posibilidad de generación de transiciones suaves entre ellas para mejorar la calidad de la síntesis (apartado 4.1).

Adicionalmente, se presenta el caso especial de la síntesis de cabezas parlantes, dado su alto interés mostrado en el capítulo 1. En este caso se trabaja con la información de habla humana, que es bimodal (posee dos modos: auditivo y visual), con lo que el modelo acústico contiene información que relaciona las imágenes con la fonética mediante:

- Un método de conversión fonética del texto o fuente sonora para permitir el reconocimiento de los sonidos que se deben pronunciar, incluyendo información temporal de los mismos (apartado 4.2).

El esquema de síntesis propuesto posee una serie de propiedades, de forma similar al de análisis, que se pueden encontrar en el apartado 4.3.

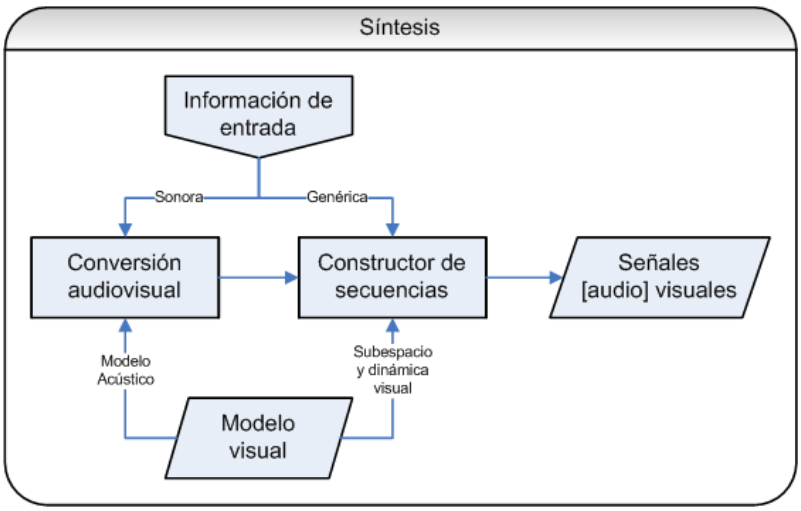


Figura 4.1: Diagrama de bloques del proceso de síntesis asociado al trabajo presentado. Está constituido por un constructor de secuencias visuales y, en el caso de disponer de información auditiva de entrada, por un módulo de conversión sonora. Éstos utilizan el modelo visual para obtener las señales audiovisuales como salida.

4.1. Síntesis visual

En esta sección se describe el proceso de creación del modo visual de las señales audiovisuales objeto del presente trabajo de investigación. Primeramente, se comentan las características del proceso de síntesis propuesto en el apartado 4.1.1. Se presenta después, en el apartado 4.1.2, el procedimiento para generar las imágenes a partir de los datos contenidos en un modelo visual cualquiera. Finalmente, el apartado 4.1.3 explica el proceso utilizado para generar las interpolaciones en las transiciones que se encuentran en las secuencias de imágenes sintetizadas.

4.1.1. Características de la síntesis visual

La síntesis visual parte de una descripción vectorial de la secuencia de imágenes a producir y es capaz de conseguirla sin intervención del usuario y de forma realista. Destacar que el método usado es suficientemente genérico como para poder ser utilizado con cualquier tipo de objeto que presente una forma rígida, o bien que haya sido grabado sobre un fondo uniforme (ver apartado 2.1).

4.1.1.1. Sin intervención del usuario

El método de síntesis presentado en este apartado obtiene el resultado automáticamente a partir de la especificación de un conjunto de vectores, que se pueden conseguir, por ejemplo, a partir de texto o voz siguiendo el apartado 4.2, también de forma automática.

4.1.1.2. Realista

La secuencia de imágenes producida es visualmente parecida a la natural, tanto a nivel estático, o de fotograma (foto realismo), como a nivel dinámico, o de secuencia (vídeo realismo). Se refiere al usuario al capítulo 5 para una consulta de los resultados obtenidos.

4.1.1.3. Genérico

El método desarrollado se puede aplicar a cualquier objeto del cual se haya obtenido un modelo visual a partir del análisis (ver capítulo 3) del corpus audiovisual (ver apartado 2.1) que lo representa.

4.1.2. Generación de imágenes

El proceso para la generación de una imagen usa el subespacio de apariencia del modelo visual y se compone de dos fases: obtención de la información de textura (ver apartado 4.1.2.1) y ubicación de ésta (ver apartado 4.1.2.2). En la primera fase se utilizan la base de apariencia $\mathbf{U}^l$, sus valores singulares asociados $\mathbf{\Sigma}^l$ y la información media $\bar{\mathbf{o}}^l$ del modelo visual para cada región, mientras que en la segunda se hace uso de las imágenes máscara $\mathbf{\Pi}$ (ver cuadro 2.9).

4.1.2.1. Obtención de la información de textura

El proceso de obtención de la textura utiliza la base de apariencia $\mathbf{U}^l$, la información media $\bar{\mathbf{o}}^l$ y los valores singulares $\mathbf{\Sigma}^l$, obtenidas en el proceso de construcción del modelo visual (ver apartados 2.2 y 3.2.2). La base $\mathbf{U}^l$ puede generar cualquier vector de textura

$\mathbf{f}^l = [f_1^l,\ f_2^l,\ \ldots,\ f_{P^l}^l]^T$ mediante la especificación de un vector de apariencia $\mathbf{d}^l$; los elementos de este vector están limitados por la diagonal de $\mathbf{\Sigma}^l$ del modelo visual de forma similar a la indicada en (3.72) y representan la proyección de $\mathbf{f}^l$ sobre $\mathbf{U}^l$.

$$\mathbf{f}^l = \mathbf{U}^l\mathbf{d}^l + \bar{\mathbf{o}}^l \tag{4.1}$$

Se puede observar que (4.1) es una expresión muy parecida a la reconstrucción indicada en (2.3) cambiando $\mathbf{c}$ por $\mathbf{d}^l$ y $\mathbf{o}$ por $\mathbf{f}^l$ aunque se utiliza con diferentes fines según el sentido de $\mathbf{c}$ o $\mathbf{d}^l$. En este apartado su uso va orientado a generar imágenes que pueden o no haber sido observadas en el corpus audiovisual utilizado para generar el modelo visual con el que se trabaja. En el apartado 2.2.3.1, la expresión únicamente va ligada a imágenes observadas en el corpus audiovisual. Dicho de otro modo, en este apartado el vector $\mathbf{d}$ no tiene porqué coincidir con ninguna de las unidades visuales reales de la dinámica visual denotadas por $\mathbf{C}$ (ver apartado 2.2.4.1), con lo que el vector $\mathbf{f}^l$ no tiene porqué coincidir con ninguno de los vectores $\mathbf{o}$, que aparecen en el corpus (ver apartado 2.2.3.1).

La obtención del vector $\mathbf{f}^l$ a partir de $\mathbf{U}^l$, $\bar{\mathbf{o}}^l$, $\mathbf{\Sigma}^l$ y $\mathbf{d}^l$ representa el procedimiento inverso explicado en el apartado 3.2, donde, a partir de las imágenes del corpus (debidamente alineadas), representadas como $\mathbf{o}^l$, se obtienen los elementos $\mathbf{U}^l$, $\bar{\mathbf{o}}^l$, $\mathbf{\Sigma}^l$ y $\mathbf{C}^l$. Este fenómeno complementario se introduce en el inicio de este capítulo, donde se presenta la síntesis como proceso inverso al análisis.

4.1.2.2. Composición de la apariencia

Una vez conocidos los vectores de textura $\mathbf{f}^l$ para los L elementos faciales, es necesario un proceso que distribuya su contenido de forma coherente en la imagen. La información de localización necesaria se encuentra en las imágenes máscara $\mathbf{\Pi}$ del modelo visual. El proceso que ubica la textura no es más que una especie de procedimiento inverso de la operación *vec* (ver apartado 3.1.2.2) aplicado a cada región. Este procedimiento inverso se basa en colocar cada elemento f_i^l en cada coordenada $\left(x_i^l, y_i^l\right)$ (ver figura 4.2), las cuales corresponden a los píxeles diferentes de cero de las imágenes máscara π^l. El orden se construye realizando una exploración por columnas de las imágenes máscara.

$$\mathbf{R}^l = vec^{-1}\left(\mathbf{f}^l, \pi^l\right)$$

En este caso y al igual que en el apartado 3.1.2.2, $\mathbf{R}^l$ es una imagen que contiene la información de textura $\mathbf{f}^l$ en los píxeles de la región definida por π^l y en el resto, ceros.

Si el modelo visual consta de una única máscara ($L = 1$) la imagen final se obtiene como $\mathbf{I} = \mathbf{R}^1 = \mathbf{R}$. En el caso de existir más de una máscara y dadas sus características (no se solapan y su suma produce la máscara global que define el objeto entero), se puede pensar en construir la imagen final como:

$$\mathbf{I} = \sum_{l=1}^{L} \mathbf{R}^l \tag{4.2}$$

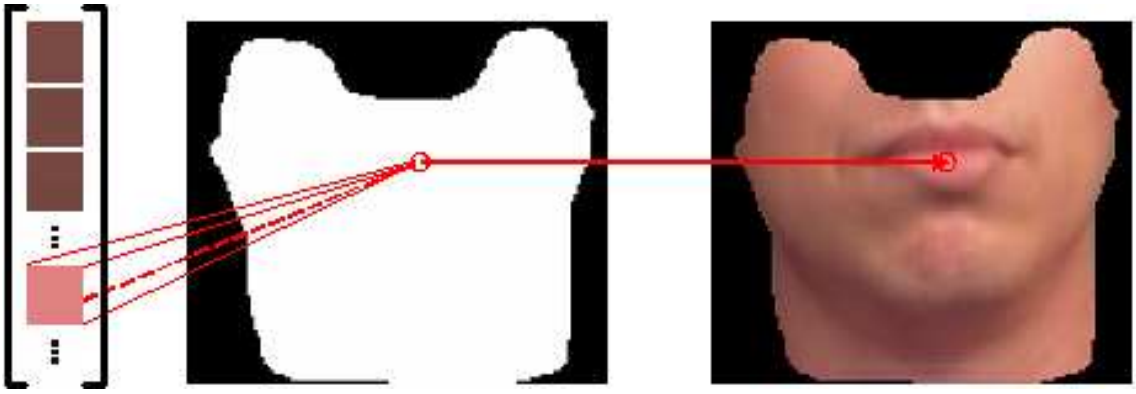


Figura 4.2: Proceso de construcción de una región (imagen derecha) a partir de su información de textura (esquema izquierdo) y su imagen máscara (imagen central) asociada.

No obstante, este procedimiento puede provocar la aparición de falsos contornos en los píxeles situados en las fronteras entre regiones debida, principalmente, a leves diferencias de iluminación. Dado que el método presentado no posee ningún control sobre la iluminación se pueden eliminar estos falsos contornos mediante la inclusión de efectos de suavizado en las transiciones entre regiones. Se define un conjunto de pesos $w^l(x, y)$ para cada píxel (x, y) de cada máscara l, cuyo valor es inversamente proporcional a la distancia con el centro de la región asociada si se encuentra cerca de la frontera, en caso contrario vale uno para la máscara que posea el píxel y cero para el resto.

$$I(x, y) = \sum_{l=1}^{L} w^l(x, y) R^l(x, y) \tag{4.3}$$

Por ejemplo, si el píxel se encuentra en la frontera entre dos regiones, su valor final se calcula como la media de los valores propuestos por cada región; por otro lado, si el píxel se encuentra suficientemente lejos de cualquier frontera obtendrá el valor asociado únicamente a la región a la que pertenece. De hecho, las imágenes máscara se pueden igualar a $\mathbf{w}^l = \pi^l$, conllevando que las regiones sean un poco más grandes que las inicialmente definidas por π^l, lo cual implica que se solapen en las fronteras, aunque su suma sigue siendo $\mathbf{\Pi}$ debido al valor de los pesos en las mismas. En la figura 4.3 se puede ver la diferencia entre aplicar suavizado (4.3) o no (4.2).

Destacar el efecto adicional de función de dispersión puntual (PSF) que se aplica a los píxeles que se encuentran en el límite definido por la máscara global $\mathbf{\Pi}$ mediante un filtro de plantilla gaussiana.

4.1.3. Generación de secuencias

En el apartado 4.1.2 se ha ofrecido un método para construir imágenes sintéticas a partir de la especificación de vectores de apariencia $\mathbf{d}^l$. Si se desea generar una secuencia de imágenes con cambios de textura, una posible solución consiste en especificar L conjuntos

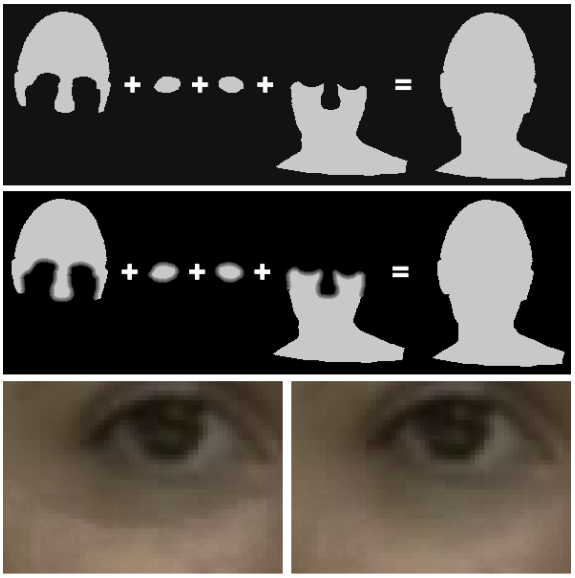


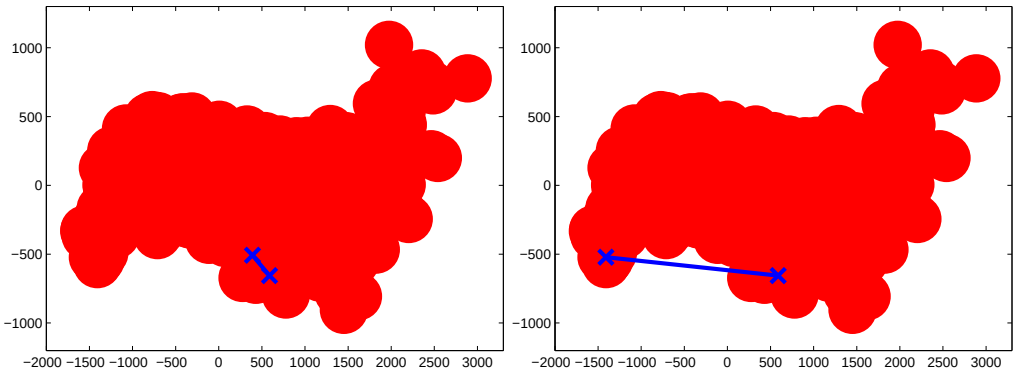
Figura 4.3: Unión de regiones. Fila superior: conjunto de máscaras sin pesos. Fila central: conjunto de máscaras que incorporan pesos para el suavizado de las fronteras al unir regiones. Fila inferior: la imagen de la izquierda se obtiene utilizando el conjunto de máscaras sin pesos, mientras que la de la derecha usa el conjunto de máscaras con pesos; notar como los falsos contornos de la imagen de la izquierda no aparecen en la de la derecha, partiendo de la misma información de textura.

$\mathcal{D}^l$ ordenados de estos vectores:

$$\mathcal{D}^l = \left\{ \mathbf{d}_1^l,\ \mathbf{d}_2^l,\ \ldots,\ \mathbf{d}_M^l \right\} \tag{4.4}$$

Para que la secuencia sea realista (tanto a nivel de fotograma -foto realista- como a nivel de transiciones entre los mismos -vídeo realista-) es necesario que los vectores de apariencia pertenezcan a la dinámica visual (ver apartado 2.2.4) utilizada, es decir, que sean lo más parecidos posible a alguna de las unidades visuales reales (2.6). De este modo, todas las imágenes $\mathbf{I}_m$ obtenidas tendrán el aspecto del objeto a partir del cual se obtuvo el modelo visual utilizado. Una primera manera de cumplir esta restricción es limitar los posibles valores de $\mathbf{d}_m^l$ al conjunto de los vectores de apariencia real denotado por $\mathbf{C}^l$. No obstante, y dado que las columnas de $\mathbf{C}^l$ son sólo un muestreo no uniforme del subespacio engendrado por la dinámica visual, se pueden intentar realizar interpolaciones entre estas columnas para obtener muestras adicionales, con el objetivo de crear transiciones menos bruscas, generando, así, efectos de coarticulación visual. Sin embargo, sólo se pueden tomar como válidas aquellas nuevas muestras que pertenezcan a dicho subespacio. Este hecho provoca que el resultado de la interpolación lineal entre dos unidades visuales reales sólo sea válido cuando éstos sean parecidos o cercanos (2.6) (ver figura 4.4), ya que este tipo de interpolación no garantiza que los valores obtenidos se encuentren dentro del subespacio visual. En el apartado 4.1.3.1 se propone un método de **interpolación no lineal de alta dimensionalidad** Melenchón et al. (2003a) para poder obtener interpolaciones realistas a partir de cualquier unidad visual real del modelo.

Las secuencias $\mathcal{D}^l$ pueden determinarse directamente o indirectamente a través de



(a) Vectores cercanos (b) Vectores alejados

Figura 4.4: Interpolación lineal entre dos unidades visuales reales de la dinámica visual de un objeto para obtener un nuevo vector de apariencia. La dinámica visual está representada por la región roja. Se muestran los casos de interpolación entre dos vectores cercanos (a) y entre dos vectores alejados (b). Notar como los vectores alejados tienen más probabilidad de atravesar zonas que se encuentran fuera del subespacio visual del objeto al utilizar una interpolación lineal.

otro tipo de información inicial. En este último caso se hace necesario una correspondencia entre la información inicial y las unidades visuales reales, utilizando después el proceso de interpolación del apartado 4.1.3.1. Puede ocurrir que esta correspondencia sea de uno a muchos (en vez de una simple aplicación), con lo que se debe escoger entre diversas unidades visuales reales para cada unidad de información de entrada. El apartado 4.1.3.2 ofrece una propuesta para realizar la selección de la mejor unidad visual real en base a sus vecinas siguiendo un esquema parecido al de concatenación de unidades propuesto para síntesis de habla (Hunt y Black, 1996), en el que también se basó Cosatto (2002).

Adicionalmente, se puede incluir el efecto de énfasis visual a la secuencia sintetizada. Este efecto está relacionado con la exageración de movimiento o cambio de apariencia que experimenta un objeto, normalmente asociado a una actividad comunicativa. Por ejemplo, el movimiento que experimentan los labios de una persona cuando reza o cuando canta a viva voz varía sustancialmente entre un nivel de énfasis bajo y uno alto, respectivamente. En el apartado 4.1.3.3 se propone la generación de énfasis visual utilizando también el algoritmo del apartado 4.1.3.2.

4.1.3.1. Interpolación no lineal de alta dimensionalidad

Sea $\mathcal{I} = \{\mathbf{I}_1 \dots \mathbf{I}_M\}$ la textura asociada a la región de un objeto en un conjunto de M imágenes correspondientes a la transición entre dos unidades visuales reales $\mathbf{c}_i^l$ y $\mathbf{c}_j^l$ del subespacio de apariencia, donde $\mathbf{I}_1 = \mathbf{U}^l\mathbf{c}_i^l + \bar{\mathbf{o}}^l$ y $\mathbf{I}_M = \mathbf{U}^l\mathbf{c}_j^l + \bar{\mathbf{o}}^l$. Esta transición se considera vídeo realista en este trabajo si se cumplen dos condiciones:

- Los pares consecutivos de imágenes $\mathbf{I}_m$ y $\mathbf{I}_{m+1}$ son similares: cuanto más similares, más vídeo realista será la transición.

- Cada imagen $\mathbf{I}_m$ debe ser foto realista, es decir, debe quedar dentro del subespacio de apariencia definido por $\mathbf{U}^r$. De forma más restrictiva, su proyección $\mathbf{c}_m^r$ debe ser similar a alguna unidad visual real.

Ambas condiciones se pueden traducir conjuntamente en encontrar la trayectoria geodésica $\mathcal{T}$ de $\mathbf{c}_i^l$ hasta $\mathbf{c}_j^l$ dentro del subespacio de apariencia $\mathbf{U}^l$.

La matriz $\mathbf{U}^r$ representa una descripción genérica del subespacio de apariencia de la región r y se obtiene a través de la descomposición en valores singulares (SVD) (ver apartado 3.2.4), la cual asume distribuciones gaussianas en los datos. El subespacio real es, de hecho, desconocido. No obstante, se dispone del muestreo no uniforme del mismo en la matriz $\mathbf{\Sigma}^l(\mathbf{V}^l)^T = \mathbf{C}^l$ (ver apartado 2.2.4.1), que son las proyecciones de las imágenes originales. Esta información se puede utilizar para encontrar aproximaciones de las trayectorias geodésicas deseadas. La idea consiste en encontrar un camino de $\mathbf{c}_i^r$ a $\mathbf{c}_j^r$ que, al menos, esté cerca de alguna de estas muestras, es decir, de alguna unidad visual real. Cuanto más cerca se encuentre esta trayectoria $\mathcal{T}$, más probable es que se encuentre dentro del subespacio real. Una manera de conseguir $\mathcal{T}$ consiste en que ésta pase a través de unidades visuales reales, de este modo se consigue que $\mathcal{T}$ no esté lejos de todas en ninguna situación.

Sea $\mathbf{G}^l$ un grafo conectado donde las unidades visuales reales son los nodos y las distancias entre ellos se encuentran en los arcos entre los nodos (4.5). Sea también $\left\|\mathbf{c}_k^l - \mathbf{c}_l^l\right\|_2^\beta$ la medida de distancia entre nodos, donde $\|\cdot\|_2$ es la norma euclídea.

$$\mathbf{G}^l(\mathbf{c}_k^l, \mathbf{c}_l^l) = \left\|\mathbf{c}_k^l - \mathbf{c}_l^l\right\|_2^\beta \tag{4.5}$$

Cuando $\beta = 1$, el camino más corto entre nodos es una línea recta. Cuando $\beta > 1$, el camino más corto es probable que contenga unidades visuales reales adicionales intermedias (ver figura 4.5) porque los saltos grandes se penalizan debido a la potencia β. De hecho, esta penalización aumenta al aumentar β. En este trabajo, se escoge un valor de dos para β con el objetivo de obtener la opción de menor coste computacional. Una vez definido el grafo y la medida de distancia, el camino más corto entre dos nodos se obtiene mediante el algoritmo de Dijkstra (1959).

Una vez obtenido el camino más corto o trayectoria geodésica $\mathcal{T}$ entre $\mathbf{c}_i^l$ y $\mathbf{c}_j^l$, también se dispone de la distancia que las separa, que se conocerá en este trabajo como distancia geodésica. Para obtener las imágenes de la transición $\mathcal{I}$, la trayectoria $\mathcal{T}$ debe ser muestreada en M puntos $\hat{\mathbf{c}}_k^r$. Cada uno de éstos se obtiene a partir de una interpolación lineal de las dos unidades visuales reales más cercanas. Dado que el salto entre ellas ha de ser pequeño debido al uso de $\beta > 1$, las interpolaciones intermedias $\hat{\mathbf{c}}_k^r$ tienen una probabilidad alta de pertenecer al subespacio de apariencia real, consiguiendo imágenes foto realistas parecidas entre ellas. Estas interpolaciones intermedias reciben el nombre de unidades visuales virtuales, ya que no han sido observadas en el corpus inicial. De hecho, es muy probable que todas las imágenes $\mathbf{I}_m$ estén relacionadas con unidades visuales virtuales excepto la primera y la última (ver figura 4.6).

Existe la opción de realizar el muestreo de $\mathcal{T}$ de forma no uniforme, concentrando más puntos en los extremos para dar una sensación de mayor naturalidad Maestri (1996).

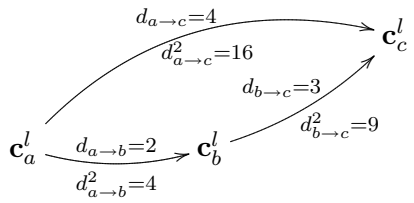


Figura 4.5: Distancia entre unidades visuales reales según diferentes métricas. La distancia euclídea se representa por d, mientras que d^2 identifica su cuadrado. Se puede comprobar que $d_{a\to b}+d_{b\to c} = 5 > 4 = d_{a\to c}$ y $d^2_{a\to b} + d^2_{b\to c} = 13 < 16 = d^2_{a\to c}$. En el primero, la distancia mínima se obtiene realizando un salto grande en vez de dos pequeños (que implican pasar por un nodo intermedio). En el segundo, ocurre el caso contrario.

Este efecto se puede conseguir mediante la aplicación de *splines* Bartels et al. (2006) o también, por ejemplo, mediante una función sigmoide (ver figura 4.7). Sea la trayectoria geodésica entre $\mathbf{c}^l_i$ y $\mathbf{c}^l_j$, suponiendo un muestreo de la misma en M puntos y definiendo $\mathbf{c}^l_i = \hat{\mathbf{c}}^r_0$ en el inicio y $\mathbf{c}^l_j = \hat{\mathbf{c}}^r_M$ en el tiempo M, la unidad visual virtual $\hat{\mathbf{c}}^r_m$ ha de situarse en el tiempo $T(m)$:

$$T(m) = A\left(\frac{1}{1+e^{\beta\left(-\frac{m}{M-1}+\frac{1}{2}\right)}} - \frac{1}{1+e^{\frac{\beta}{2}}}\right) \tag{4.6}$$

donde la constante A se calcula como $A = \text{cosech}(0{,}5\beta) + \text{cotanh}(0{,}5\beta)$ y la pendiente en el punto de inflexión se especifica mediante su ángulo α (que cumple que $\beta = tan(\alpha)/A$), que toma un valor de $\pi/3$ en este trabajo. Valores menores o mayores provocan movimientos con falta o exceso de naturalidad, con lo que pueden parecer demasiado forzados.

4.1.3.2. Algoritmo de selección de unidades visuales

En el apartado 4.1.3.1 se ha mostrado como encontrar el camino más corto entre dos unidades visuales reales $\mathbf{c}^r_0$ y $\mathbf{c}^r_P$ restringiendo la búsqueda dentro del subespacio visual. No obstante, puede que se desee pasar a través de unidades visuales reales intermedias $\mathbf{c}^r_p$. En este caso, el resultado es la concatenación de las trayectorias entre cada par $\mathbf{c}^r_p$ y $\mathbf{c}^r_{p+1}$. Se pueden definir diferentes valores de α o parametrizaciones de los splines para cada subtrayectoria, de manera que se disponga de empalmes más o menos marcados, según convenga en la animación. Notar que el muestreo final se realiza en la trayectoria global y no en cada subtrayectoria por separado.

Cuando las unidades visuales reales vienen definidas de forma concreta no existe, por consiguiente, ningún tipo de selección de las mismas. Éste aparece necesariamente cuando se dispone de un conjunto de unidades visuales reales candidatas $\mathbf{\Psi}^r_p$ para cada uno de los $\mathbf{c}^r_p$ que definen la trayectoria final. Uno de los primeros trabajos que aplican la selección de unidades es el de Hunt y Black (1996), aunque en el ámbito de la síntesis de voz. Este trabajo busca la secuencia de unidades óptima respecto a dos sentidos: i) ser lo

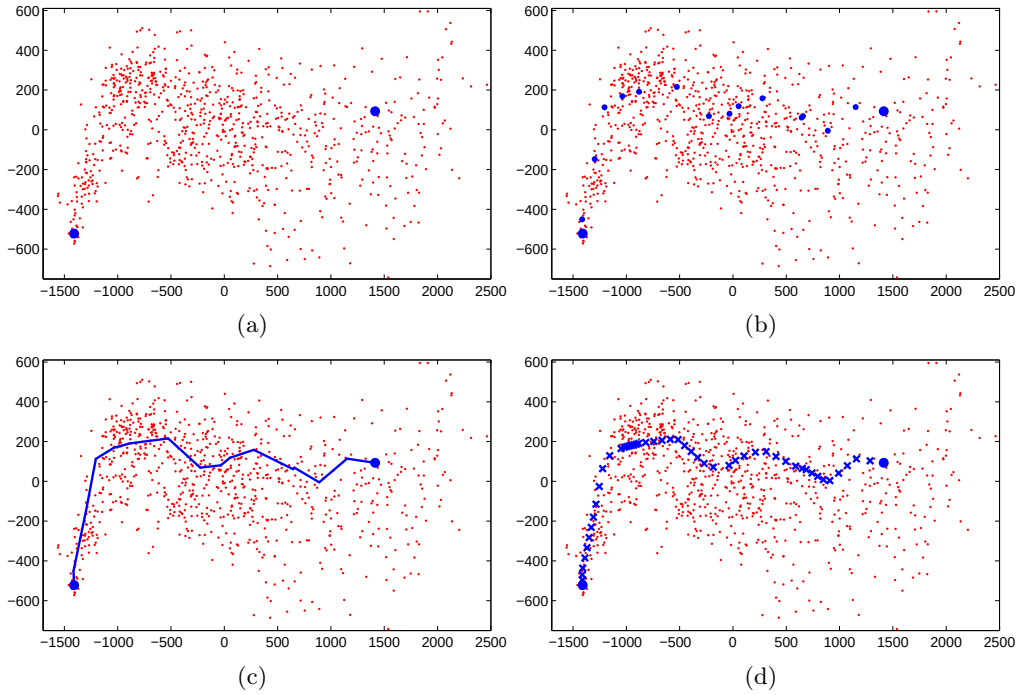


Figura 4.6: Proceso de interpolación no lineal de alta dimensionalidad. Dadas las unidades visuales reales a interpolar (a), se obtiene el camino más corto (b) y la trayectoria asociada (c). El resultado final se obtiene muestreando esta última (d).

más parecidas posibles a sus vecinas y *ii*) adecuarse a la información que dirige la síntesis; el primero se mide mediante el coste de concatenación y el segundo, mediante el coste de objetivo. En esta aproximación se sigue la misma filosofía, aunque únicamente se utiliza el coste de concatenación; el de objetivo se simplifica con la determinación a priori de los diferentes Ψ_p^r. A modo de ejemplo, en el caso de síntesis de caras parlantes, cada uno de estos grupos podría estar constituido por las unidades visuales reales asociadas a un grupo de visemas visualmente difícil de distinguir.

Dado el conjunto de unidades visuales reales candidatas actual $\mathbf{\Psi}_p^r$, la siguiente $\mathbf{\Psi}_{p+1}^r$ y la unidad visual real seleccionada anteriormente $\mathbf{c}_{p-1}^r$, se puede definir el error de concatenación para $\mathbf{c}_p^r$ como:

$$E_c\left(\mathbf{c}_p^r, \mathbf{c}_{p+1}^r\right) = D\left(\mathbf{c}_{p-1}, \mathbf{c}_p\right) + D\left(\mathbf{c}_p, \mathbf{c}_{p+1}\right)$$

donde $D(a, b)$ representa una aproximación de la distancia geodésica entre las unidades visuales reales a y b, que se puede encontrar mediante el algoritmo de interpolación presentado en el apartado 4.1.3.1. Se debe notar que $\mathbf{c}_p^r$ y $\mathbf{c}_{p+1}^r$ pueden tomar diferentes valores mientras que $\mathbf{c}_{p-1}^r$ viene fijado con anterioridad. Tomando el mínimo valor de E_c para cada $\mathbf{c}_p^r$ posible, se obtiene un vector de costes de concatenación, que se identificará por $\mathcal{E}$. El índice de su estadístico de primer orden se puede usar para encontrar el valor de $\mathbf{c}_p^r$ que minimiza $\mathcal{E}$ y, por lo tanto, también E_c. Para implementar este procedimiento se puede usar el algoritmo de Viterbi (1967) de profundidad dos (ver figura 4.8).

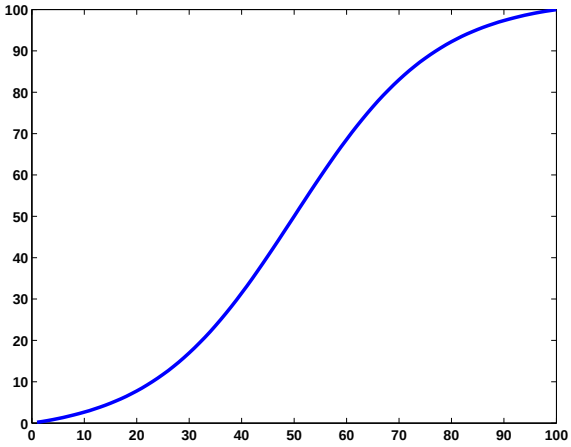


Figura 4.7: Ejemplo de la función sigmoide empleada en la interpolación no lineal de alta dimensionalidad para conseguir mayor naturalidad de cambios en la apariencia.

4.1.3.3. Énfasis visual

En el apartado anterior (4.1.3.2) se ha utilizado un estadístico de primer orden de $\mathcal{E}$ para encontrar la unidad visual real más parecida a sus dos vecinas. Si se utiliza un estadístico de otro orden, se obtiene otra unidad visual real de menor similitud, lo cual implica que se produce una secuencia de unidades visuales reales más diferentes entre ellas. De hecho, cuanto mayor es el orden del estadístico, mayor es la diferencia. Este comportamiento se puede utilizar para forzar transiciones entre unidades visuales reales muy diferentes, obteniendo un tipo de transiciones visuales exageradas o enfatizadas. Éstas contienen la misma secuencia de $\mathbf{\Psi}_p^r$, pero la representan mediante unidades que se diferencian entre ellas (ver figura 4.9).

La secuencia sintética con menos variación se obtiene con un estadístico de primer orden. En este caso, las imágenes resultantes son todo lo más parecidas posibles entre ellas a nivel temporal. Si se desea dotar a la animación de más énfasis o exageración en los cambios de apariencia, se debe usar un estadístico de orden superior. De este modo, se puede controlar la cantidad de énfasis visual a través de un valor escalar. Normalizando los órdenes de los estadísticos, se asocia el valor $\alpha = 0$ al primero y $\alpha = 1$ al máximo, para cada uno de los $\mathbf{\Psi}_p^r$.

4.2. Conversión fonética

En el caso de tratar con caras parlantes, el estudio del proceso de síntesis se centra en el elemento visual de la boca, principalmente, dado que es el principal elemento facial responsable de la producción de los sonidos en el habla humana. Para poder realizar la síntesis de nuevo contenido audiovisual con habla humana utilizando un modelo visual asociado a la cara de éstos, resulta imprescindible disponer de un elemento **director** que guíe el proceso. El término director significa «que guía, mostrando o dando las señas de

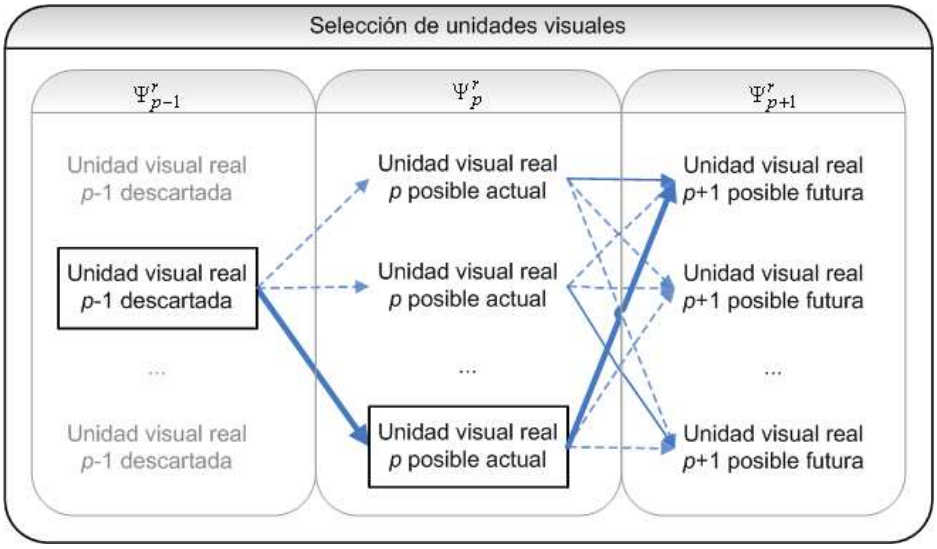


Figura 4.8: Proceso de selección de la unidad visual real $\mathbf{c}^r_p$, dada la especificación de dos conjuntos de unidades visuales reales candidatas $\mathbf{\Psi}_p$ y $\mathbf{\Psi}_{p+1}$. La unidad $\mathbf{c}^r_{p-1}$ ya ha sido seleccionada en el paso anterior, descartando el resto de posibilidades. La unidad visual real $\mathbf{c}^r_{p+1}$ más cercana a cada una de las $\mathbf{c}^r_p$ se marca con una línea sólida y el camino más corto entre $\mathbf{c}^r_{p-1}$ y $\mathbf{c}^r_{p+1}$ se muestra con una línea sólida más gruesa. Notar que las unidades $\mathbf{c}^r_{p+1}$ se usan pero no se selecciona ninguna hasta el próximo paso.

un camino» (Real Academia Española, 2001); así, se podría interpretar que a partir de las **señas** se puede recorrer el **camino** al cual hacen referencia. En este trabajo, el **camino** se representa como el resultado del proceso de síntesis, las **señas**, como cierta información fonética inicial (aunque podría corresponder a otro tipo de información en el caso de tratar la síntesis de otros elementos faciales distintos a la boca o, incluso, otros objetos) y la conversión fonética, como el elemento **director** descrito.

En este trabajo se han supuesto dos posibles fuentes de información auditiva (ver figura 4.1): *i*) fonética y *ii*) voz. El primer origen da nombre a la síntesis guiada por fonética (apartado 4.2.4), mientras que el segundo caracteriza la síntesis guiada por voz (apartado 4.2.5). La síntesis guiada por fonética parte de una cadena de caracteres (en código ASCII, por ejemplo) y extrae la información fonética utilizando un conversor texto a habla (TTS) (Guaus y Iriondo, 2000, Scansoft Inc., 2005, AT&T Corp., 2007). El segundo tipo utiliza la forma de onda generada por una voz (humana o incluso sintética) y, usando métodos predictivos, obtiene la descripción de la información visual, directa o indirectamente, según el método empleado. Sea como fuere, ambas técnicas buscan encontrar un conjunto ordenado de unidades visuales reales a partir de las cuales generar la animación final mediante los procesos detallados en el apartado 4.1.

4.2.1. Características de la conversión fonética

Se utilizan los modelos acústicos (apartado 2.2.5) para poner en correspondencia información fonética y visual, obteniendo procesos automáticos, sincronizados y genéricos.

Figura 4.9: Ejemplos de énfasis visual. Cuatro imágenes que muestran la pronunciación del mismo alófono [i] con diferentes niveles de énfasis α de 0,0, 0,3, 0,7 y 1,0, de izquierda a derecha.

4.2.1.1. Automática

La conversión fonética no precisa ninguna intervención humana para su funcionamiento, una vez se dispone de un modelo acústico. El proceso es capaz de decidir la salida a partir de cualquier entrada.

4.2.1.2. Bajo coste computacional

Los procesos de conversión fonética son capaces de procesar la información rápidamente. En el caso de síntesis guiada por fonética es constante y en el caso de estar guiada por voz, el proceso de predicción de los métodos estudiados tiene un coste computacional muy reducido (no así el de su construcción, que se realizan a priori y una sola vez por modelo acústico).

4.2.1.3. Genérica

Los procesos desarrollados para la conversión fonética pueden ser utilizados para convertir otros tipos de información (sonora o no) ya que no están ligadas a las características especiales que tienen la voz y la cara humana.

4.2.2. Agrupación visémica personalizada

Un elemento importante a considerar al realizar el proceso de conversión de información sonora a visual es la agrupación de los datos empleada y, más concretamente, los grupos de visemas elegidos (ver apartado 2.1.3.1). Estos grupos pueden variar entre personas, ya que no todo el mundo habla con los mismos movimientos bucales (Johnson et al., 1993). A través de los diferentes estudios en este ámbito, que se pueden consultar en Owens y Blazek (1985), se pueden extraer diferentes conclusiones:

- Se han utilizado procesos empíricos para encontrar los grupos de visemas óptimos, lo cual implica un nivel de esfuerzo y tediosidad elevados.
- Se han considerado combinaciones artificiales de sonidos y no habla natural en los experimentos.
- Los grupos de visemas se distinguen, principalmente, por los articuladores externos (bilabiales, labiodentales e interdentales). Los diferentes estudios coinciden en tres grupos de visemas asociados a estos articuladores: [p], [β], [m]; [f] y [ɱ]; y [θ]. A su vez, muestran divergencias sobre los grupos de visemas asociados al resto de articuladores (dentales, alveolares, palatales y velares).
- No se ha personalizado el estudio a personas concretas.

Se propone un nuevo método objetivo de agrupamiento de visemas mediante el uso de la distancia de Bhattacharyya (Fukunaga, 1990) con el objetivo de evitar las dificultades de testeo de los trabajos previos, permitir el uso de habla natural en los mismos y ofrecer una vía para encontrar los visemas propios de cada individuo. Para conseguirlo, se utiliza la notación vectorial de las unidades visuales reales y la segmentación de éstos en base al canal auditivo asociado.

Una forma de definir las clases es mediante alguno de los niveles del cuadro 2.4. Sin embargo este cuadro está construido a partir de conceptos generales y no particulares de los datos en sí, con lo que puede no ser la mejor clasificación posible para una persona concreta, en el sentido de que puede haber clases que pudieran estar juntas debido a la manera de hablar de esa persona y otras que debieran estar separadas. Una manera de crear un conjunto de clases en base a los datos particulares es mediante el algoritmo de *K-means* Moon (1999). Sin embargo, las clases finales no tienen porqué estar asociadas a ningún grupo de sonidos en particular, ya que no se incluye ningún tipo de información fonética en dicho algoritmo. Dada esta situación, se propone una tercera manera de agrupar los datos teniendo en cuenta su distribución y su información fonética asociada.

El método de agrupación propuesto se basa en simular los experimentos llevados a cabo por personas reales en trabajos preliminares tales como los de Fisher (1968), Binnie et al. (1974), Walden et al. (1977), Benguerel y Pichora-Fuller (1982), Owens y Blazek (1985). La suposición que manejaban estos trabajos era la de que los grupos resultantes eran visualmente distinguibles entre ellos pero no dentro de los mismos, implicando la práctica indivisibilidad de cada grupo. Simplificando sus procedimientos, determinados conjuntos de personas eran testadas para usarse como sistemas expertos de clasificación de los diferentes alófonos en grupos de visemas.

Se propone sustituir las medidas expertas dadas por las personas implicadas por una medida objetiva calculada por una máquina. Para ello, es necesario disponer de una codificación numérica de la información visual y una medida de distancia asociada a esta codificación. La primera parte de las unidades visuales reales almacenadas en la dinámica visual e identificadas por $\mathbf{C}^r$ (en este caso particular r identifica la región de la boca). Éstas se agrupan por alófonos y se caracterizan mediante 32 gaussianas multidimensionales (ya que se consideran 32 alófonos: 7 vocálicos y 25 consonánticos). En lo que a medida

de distancia se refiere, se utiliza la distancia de Bhattacharyya, que sirve para medir la distancia que separa dos distribuciones gaussianas.

El proceso de segmentación asociado a la agrupación por alófonos se traduce en encontrar las etiquetas temporales t_n de cada uno de los alófonos pronunciados. Esta tarea se puede conseguir automáticamente mediante el uso de las herramientas HTK (Young et al., 2003). A partir de t_n se pueden obtener los índices n de las imágenes asociadas como $n = \text{round}(t_n * f_{ps})$, donde f_{ps} identifica las imágenes por segundo de la secuencia de vídeo. Desafortunadamente, debido a pequeñas asincronías dadas por los sistemas de adquisición domésticos, puede ser necesaria una supervisión manual del canal de vídeo para evitarlas.

4.2.2.1. Distancia entre grupos

Para encontrar la distancia entre diferentes subconjuntos de unidades visuales reales asociadas al mismo alófono, también llamados conjuntos visuales (y que corresponderían a los $\mathbf{\Psi}_p^r$ del apartado 4.1.3.2, siendo r el identificador de la región de la imagen correspondiente a la boca), se modelan mediante una distribución normal multidimensional $\mathcal{N}_g$, es decir, mediante un vector de medias μ_g y una matriz de covarianzas $\mathbf{K}_g$. Si $\mathbf{C}_g$ es la matriz en cuyas columnas se disponen los N_g vectores de un conjunto visual g y se define $\mathbf{1}$ como un vector columna de N_g unos, entonces:

$$\mu_g = \frac{1}{N_g}\mathbf{C}_g \cdot \mathbf{1}$$

$$\frac{1}{N_g}\left(\mathbf{C}_g - \mu_g \cdot \mathbf{1}^T\right)\left(\mathbf{C}_g - \mu_g \cdot \mathbf{1}^T\right)^T$$

Seguidamente, se puede encontrar una aproximación de la distancia entre dos conjuntos visuales k y l como la distancia de Bhattacharyya (4.7), que da una medida de lo alejadas que estan dos distribuciones normales entre sí (Fukunaga, 1990).

$$\begin{aligned} B\left(\mathcal{N}_k, \mathcal{N}_l\right) &= \frac{1}{8}B_1 + \frac{1}{2}B_2 \\ B_1 &= \left(\mu_k - \mu_l\right)^T \left(\frac{\mathbf{\Sigma}_k + \mathbf{\Sigma}_l}{2}\right)^{-1} \left(\mu_k - \mu_l\right) \\ B_2 &= \ln \frac{\left|\frac{\mathbf{\Sigma}_r + \mathbf{\Sigma}_l}{2}\right|}{\sqrt{\left|\mathbf{\Sigma}_k\right|\left|\mathbf{\Sigma}_l\right|}} \end{aligned}$$

De este modo, se puede construir un grafo con las distancias entre todos los conjuntos visuales representado por $\mathbf{H}$, un ejemplo del cual se puede observar en la figura 4.10.

$$\mathbf{H}(k, l) = B(\mathcal{N}_k, \mathcal{N}_l) \tag{4.7}$$

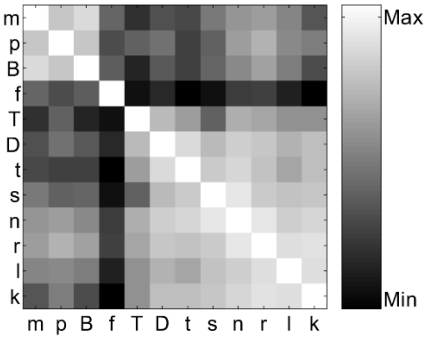


Figura 4.10: Grafo de distancia **H** entre conjuntos visuales en notación ASCII siguiendo el estándar propuesto por SAMPA (Wells, 1997). El color blanco está relacionado con la similitud máxima, mientras que el negro, con la mínima.

4.2.2.2. Agrupar conjuntos similares

Tomando el grafo de distancias (4.7), la similitud de un conjunto visual p al resto viene representada por la columna p de la matriz $\mathbf{H}_v$, identificada por $\mathbf{h}^p$. Dados dos conjuntos visuales k y l con sus respectivas columnas $\mathbf{h}^k$ y $\mathbf{h}^l$ (4.8), si $\mathbf{h}^k$ es similar o diferente a $\mathbf{h}^p$ (4.9), entonces $\mathbf{h}^l$ y $\mathbf{h}^p$ comparten la misma relación (4.10)(4.11) si $\mathbf{h}^k$ y $\mathbf{h}^l$ son similares (4.8).

$$\left\|\mathbf{h}^k - \mathbf{h}^l\right\|_2 = \left\|\mathbf{h}^k - \mathbf{h}^l\right\|_2 = \epsilon \tag{4.8}$$

$$\left\|\mathbf{h}^k - \mathbf{h}^p\right\|_2 = \left\|\mathbf{h}^p - \mathbf{h}^k\right\|_2 = S \tag{4.9}$$

$$\left\|\mathbf{h}^l - \mathbf{h}^p\right\|_2 \leq \left\|\mathbf{h}^l - \mathbf{h}^k\right\|_2 + \left\|\mathbf{h}^k - \mathbf{h}^p\right\|_2 \Rightarrow \left\|\mathbf{h}^l - \mathbf{h}^p\right\|_2 \leq S + \epsilon \tag{4.10}$$

$$\left\|\mathbf{h}^k - \mathbf{h}^p\right\|_2 \leq \left\|\mathbf{h}^k - \mathbf{h}^l\right\|_2 + \left\|\mathbf{h}^l - \mathbf{h}^p\right\|_2 \Rightarrow \left\|\mathbf{h}^l - \mathbf{h}^p\right\|_2 \geq S - \epsilon \tag{4.11}$$

Donde $\epsilon \ll \left\|\mathbf{h}^k\right\|_2$. Consecuentemente, si se desea distinguir entre G visemas únicamente, se pueden agrupar las columnas de $\mathbf{D}$ en G macrogrupos, uniendo conjuntos visuales similares mediante la aplicación de un algoritmo como *K-means* (Moon, 1999). Los macrogrupos conformarían, en este caso, los nuevos conjuntos visuales $\mathbf{\Psi}_p^r$. Cada uno de ellos puede llamarse visemas, siguiendo la definición original dada por Fisher (1968), ya que poseen unidades visuales parecidas entre ellas y diferentes entre conjuntos. En la figura 4.11 se pueden observar tres conjuntos visuales a lo largo de cuatro dimensiones, donde dos de ellos son muy similares (los correspondientes a [m] y [β]).

Para resaltar aún más las similitudes, se puede construir otra matriz de distancias nueva $\mathbf{H}_1$ que contenga el producto escalar entre todas las columnas normalizadas de $\mathbf{H}$, cogidas dos a dos:

$$\mathbf{H}_1(i,j) = \frac{\left(\mathbf{h}^k\right)^T \left(\mathbf{h}^l\right)^T}{\left\|\mathbf{h}^k\right\|_2 \left\|\mathbf{h}^l\right\|_2} \tag{4.12}$$

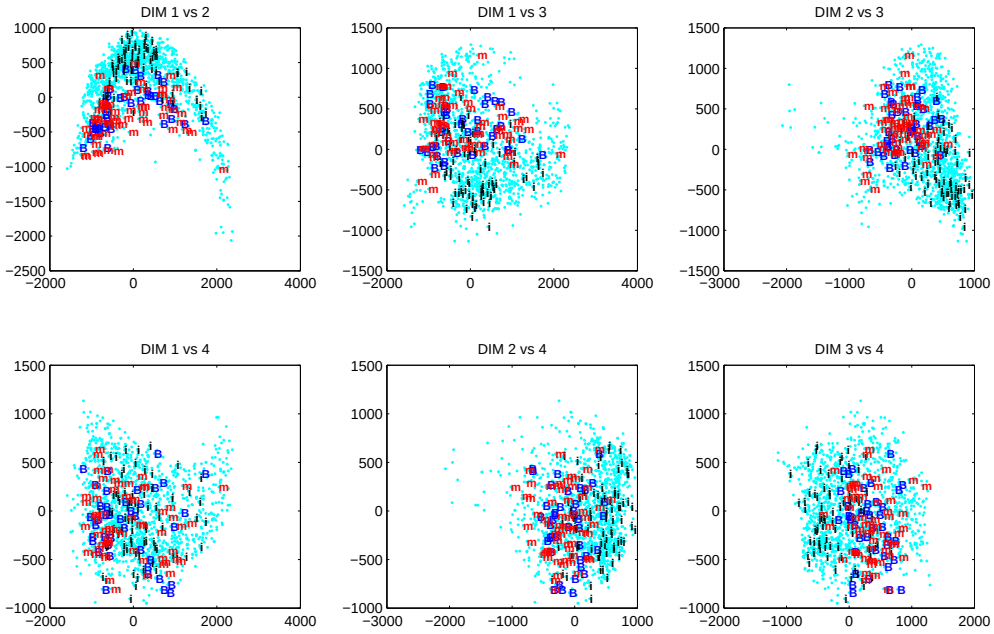


Figura 4.11: Ejemplos de tres conjuntos visuales a lo largo de cuatro dimensiones mostradas dos a dos. Los mostrados son los correspondientes a los alófonos [i], [m] y [β]. Notar como los dos últimos están superpuestos en todas las gráficas, dando a entender que son de un aspecto visual muy parecido, al menos, en las cuatro dimensiones mostradas.

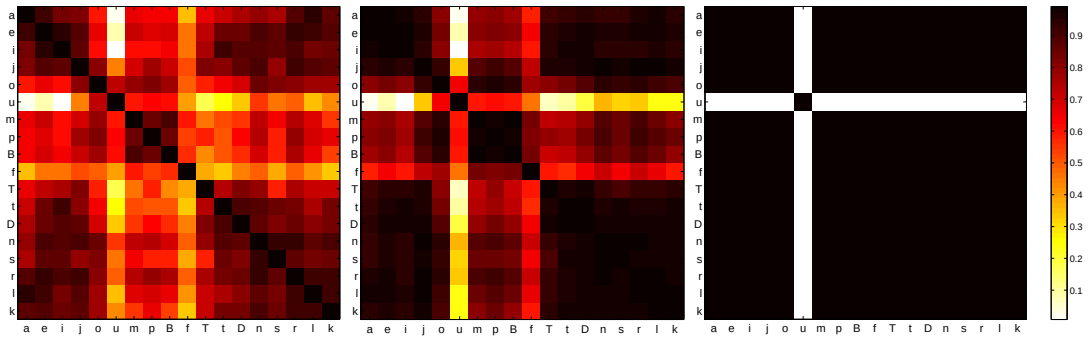
Los valores de la matriz D_1^V ocupan todo el rango de cero a uno (ya que $\mathbf{H}(k,l) > 0 \; \forall \; k,l$), y distinguen así entre más parecido (uno) y menos (cero). De hecho, la identidad (4.12) se puede aplicar recursivamente (4.13) definiendo $\mathbf{h}_i^k$ como las columnas de $\mathbf{H}_i$. Cuantas más veces se aplique, más binaria tiende a ser la agrupación final (ver figura 4.12)

$$\mathbf{H}_{i+1}(i,j) = \frac{\left(\mathbf{h}_i^k\right)^T \left(\mathbf{h}_i^l\right)^T}{\left\|\mathbf{h}_i^k\right\|_2 \left\|\mathbf{h}_i^l\right\|_2} \tag{4.13}$$

En el capítulo 5 se muestran los resultados de aplicar este tipo de agrupamiento sobre diferentes personas.

4.2.3. Codificación de la información auditiva

El objetivo principal del apartado 4.2 es relacionar un contenido acústico con otro visual, pero siempre, bajo el punto de vista de la percepción humana. En última instancia, el resultado será considerado correcto si la relación conseguida entre voz e imagen es similar a la que se observaría en una persona. Se conoce que la percepción acústica realizada por el ser humano se basa en características frecuenciales (Huang et al., 2001). Si en la búsqueda de la relación objetivo de este apartado se representa la información sonora desde otro punto de vista que no sea el frecuencial, puede que ciertas características, que resultan importantes para la percepción realizada por el oído, queden atenuadas o escondidas, dificultando la búsqueda de la relación correcta. Como conclusión, es interesante utilizar



(a) (b) (c)

Figura 4.12: Diferencia visual entre varios conjuntos visuales midiendo diferencias entre ellos recursivamente. (a) distancias iniciales; (b) distancias de distancias; y (c) distancias calculadas recursivamente diez veces. Notar que la gradación de colores va desde el blanco (el menos parecido) al negro (el más parecido) pasando por una gama de amarillos, naranjas y rojos.

una representación de los datos similar a la que utiliza el oído para, al menos, partir del mismo punto de vista sobre la información. Además, las características más importantes se encuentran en el modelado del tracto vocal ya que, en última instancia, se pretende relacionar este tipo de información con la información visual, la cual tiene una relación directa con la posición que adopta el tubo vocal al hablar (Yehia y Itakura, 1994).

Dado que la señal de voz es un tipo de señal estocástica con una cierta estacionariedad (es decir y probabilísticamente hablando, que mantiene su comportamiento durante instantes inmediatos de tiempo, pero que lo cambia a lo largo de éste), su caracterización frecuencial sufrirá una evolución a lo largo del tiempo. Debido a este motivo, se realiza un ventaneo sobre ésta. La longitud de la ventana a utilizar es de veinte milisegundos (para tener una buena localización temporal de las variaciones) y se usa superposición de diez milisegundos, con el objetivo de tener una descripción detallada de la evolución temporal de las características frecuenciales de la señal (y limitando la duración mínima de los fonemas a alrededor de diez milisegundos, más que suficiente para un ritmo normal de habla). Cada ventana se conoce por el nombre de trama de voz y se representa mediante algún tipo de parametrización basada en características frecuenciales.

Existen muchos tipos de parametrizaciones para la señal de voz basadas en la frecuencia: coeficientes de predicción lineal (LPC), cepstrum, coeficientes de reflexión (RC), ratios del logaritmo de las áreas (LAR), frecuencias espectrales lineales (LSF), coeficientes cesptrum de escala frecuencial Mel (MFCC), ... La comunidad científica ha usado mayoritariamente el conjunto de MFCC, sobre todo en aplicaciones de reconocimiento de voz gracias a su compacta representación (Davis y Mermelstein, 1980), aunque los LSF y derivados han mostrado su interés en codificación de voz (Huang et al., 2001), como el desarrollo del estándar GSM 06.90 version 7.2.1 (1998). En un principio, se contemplan:

- LPC, ya que ofrecen un modelo todo polos del tracto vocal.
- Cepstrum, ya que ofrecen también otra manera de representar el tracto vocal me-

diante la cuefrencia.

- RC y LAR, debido a que están directamente relacionados con la representación en modelo de tubos del tracto vocal.
- LSF, ya que ofrecen la localización de los formantes principales asociados al tracto vocal.
- MFCC, ya que representan también el tracto vocal, aunque desde un punto de vista frecuencial y basado en la escala Mel (Stevens y Volkman, 1940).

Dado un conjunto de tramas dispuestas en la matriz $\mathbf{T}^{\mathcal{S}} = \left[\mathbf{t}_1^{\mathcal{S}}, \ldots, \mathbf{t}_N^{\mathcal{S}}\right]$ correspondiente a un corpus balanceado (por ejemplo, el de la cuadro 2.5), se parametriza cada una de ellas según Q parámetros $\mathcal{P}$, obteniendo $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_N]$. Mediante la segmentación expresada en el apartado 4.2.2 se pueden obtener también los conjunto de tramas asociadas a un mismo alófono y caracterizarlos mediante un distribuciones normales $\mathcal{N}_p^{\mathcal{S}}$ cada uno de ellos. Seguidamente se puede encontrar la distancia de Bhattacharyya (4.7) entre ellos e incluso utilizar métodos de estimación bayesiana para comprobar la bondad de clasificación que ofrecen, dada una parametrización. En la figura 4.13 se pueden observar los resultados que miden esta bondad para los diferentes espacios paramétricos utilizados. Se puede observar que los mejores resultados se obtienen con los MFCC y los LSF, destacando los primeros sobre las consonantes y los segundos sobre las vocales. Se ha decidido tomar la representación LSF por dos motivos: el primero se basa en que las vocales tienen un alto grado de importancia en la apariencia visual, debido a las diferentes aperturas bucales y a su elevada influencia en la apariencia de los visemas consonánticos; el segundo consiste en que la parametrización en LSF se utiliza en la codificación de la voz en telefonía (GSM 06.90 version 7.2.1, 1998), de tal manera que utilizarla permitiría aprovechar dicha codificación (evitando así cualquier reparametrización) al obtener los datos de voz a partir del canal telefónico.

4.2.4. Síntesis guiada por fonética

Una manera de especificar el contenido de la señal audiovisual que se desea crear es mediante la transcripción fonética de un texto escrito. Para tratar con este tipo de información se utilizan los conversores TTS (Huang et al., 2001). Este tipo de herramientas son también procesos de síntesis, aunque solamente de voz y suelen incorporar un conjunto de reglas sintácticas y semánticas que traducen el texto de entrada a su transcripción fonética, entre más detalles (para más información, consultar el trabajo de Allen et al. (1987)). Su objetivo final consiste en generar una forma de onda correspondiente a la pronunciación de un texto escrito, incluyendo la prosodia, tono de voz y velocidades indicadas. Adicionalmente, también se están realizando estudios para incorporar características emocionales en la síntesis de voz, como los de Iriondo et al. (2004).

Hasta este punto, se pueden ver los conversores TTS como las herramientas a utilizar para obtener el canal auditivo de la síntesis audiovisual propuesta en este trabajo. No obstante, queda pendiente la sincronización de la información de ambos canales. Esta

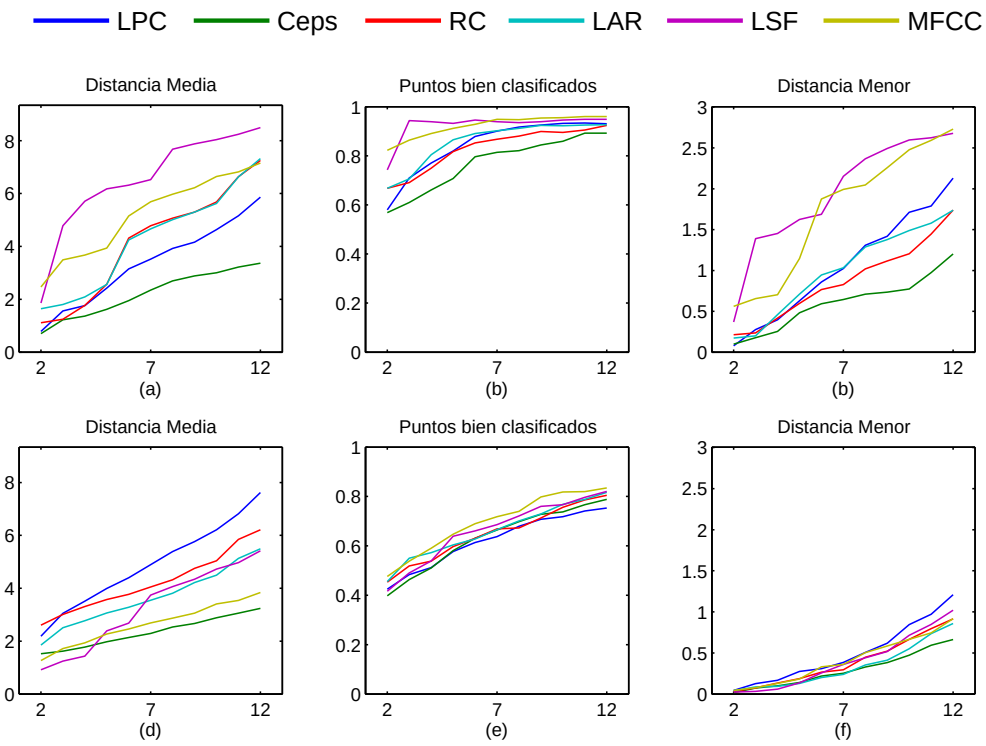


Figura 4.13: Bondad de la separabilidad de seis espacios paramétricos al aumentar su dimensionalidad de 2 a 12 parámetros. Las gráficas (a), (b) y (c) corresponden a la separabilidad de sonidos vocálicos. Las marcadas como (d), (e) y (f) pertenecen a la de los sonidos consonánticos. (a) y (d) ofrecen la distancia media entre sonidos; (c) y (f), la mínima; finalmente, (b) y (e) ofrecen el tanto por uno de tramas bien clasificadas mediante estimación bayesiana utilizando una distribución normal por clase. Se puede observar que los mejores resultados corresponden al espacio paramétrico LSF y MFCC, y a los sonido vocálicos.

sincronía se puede obtener mediante la captura de la información fonética que producen los conversores TTS en su funcionamiento en etapas intermedias. Estos sistemas necesitan conocer el sonido a sintetizar, así como su duración, para poder crear la señal acústica final. Mediante el conocimiento de estos datos (el sonido y su duración) también es posible la creación de la señal visual correspondiente (ver apartado 4.1). Además, existen algunos tipos de conversores TTS (como los que cumplen las especificaciones de SAPI® de Microsoft®) que notifican este tipo de información, simplificando la implementación del proceso de síntesis de caras parlantes.

Conocida la codificación de la lista de sonidos a sintetizar junto con su duración, se puede utilizar la correspondencia identificador-visema (CIV) (apartado 2.2.5.1) para obtener la secuencia de imágenes de boca codificadas (apartado 4.1.2) utilizando la generación de secuencias comentada en el apartado 4.1.3 (figura 4.14). Además, cuando el proceso de notificación de sonidos se realiza en tiempo real (esto es, mientras se reproduce la forma de onda), se puede generar la componente visual simultáneamente y sin necesidad de la información de duración si ésta es suficientemente rápida (del orden de menos de 40ms).

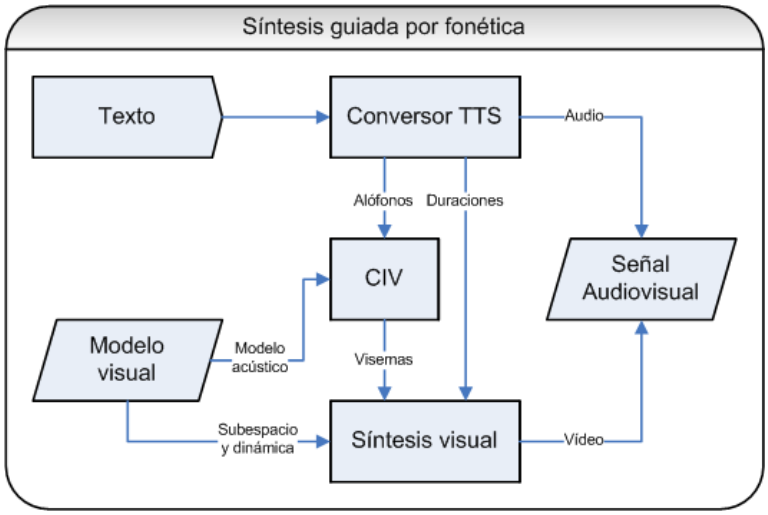


Figura 4.14: Diagrama de la síntesis de caras parlantes guiada por información fonética y utilizando un conversor TTS

4.2.5. Síntesis guiada por voz

Alternativamente al apartado 4.2.4, se puede determinar el contenido de la señal audiovisual mediante voz: «Sonido que el aire expelido de los pulmones produce al salir de la laringe, haciendo que vibren las cuerdas vocales» (Real Academia Española, 2001). Hay que tener en cuenta que este tipo de especificación de la información es más susceptible a errores ya que no se dispone de la información real a codificar (como sí que ocurre en el caso anterior), sino de una versión sonora de la misma. Sin embargo, no es necesario sintetizar el canal sonoro, ya que se puede utilizar directamente la señal de voz original.

A diferencia de la síntesis guiada por fonética, la guiada por voz trabaja con una relación voz-visema (RVV) (ver apartado 2.2.5.2), compuesta por un proceso de estimación de visemas a partir de la información sonora previamente ventaneada y parametrizada (ver figura 4.15). Dependiendo del proceso de estimación, se puede usar también una CIV, en cuyo caso, la RVV recibe el nombre de relación voz-identificador (RVI). En cualquier caso, la información sonora de entrada se acaba transformando en realizaciones concretas de visemas o unidades visuales reales, las cuales se interpolan siguiendo el proceso indicado en el apartado 4.1.3.

La información sonora de entrada debe ser cuantificada de algún modo para poder actuar como entrada en un proceso digital. Por esta razón, se aplica el proceso de ventaneo y parametrización explicados en el apartado 4.2.3.

Se presenta un método de estimación lineal de los parámetros (apartado 4.2.5.1) seguido por otro basado en estimadores bayesianos (apartado 4.2.5.2). Finalmente, se utilizan estos últimos para presentar los curiosos efectos sobre la incertidumbre audiovisual (apartado 4.2.6) que se obtienen al tener un único canal en cuenta a la hora de realizar estimaciones sobre el otro.

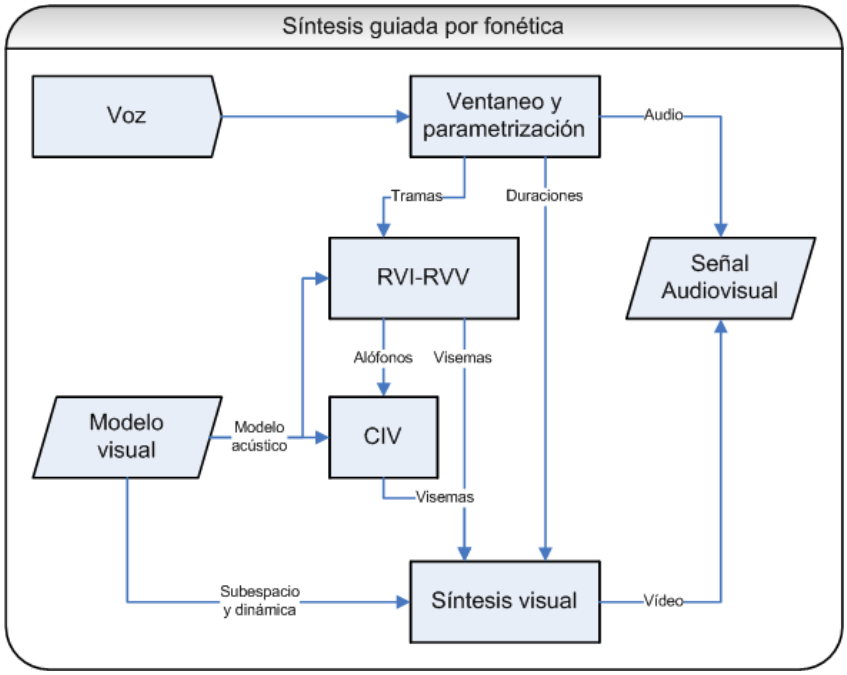


Figura 4.15: Diagrama de la síntesis de caras parlantes guiada por voz mediante una RVV o combinación de RVI y CIV.

4.2.5.1. Estimación directa lineal

El objetivo de este tipo de estimación es poder obtener la codificación de la apariencia bucal a partir de la parametrización de una trama de voz mediante operaciones lineales. En este caso, las unidades visuales reales son producidas directamente por este estimador de visemas mediante una RVV, evitando el uso de cualquier CIV (ver figura 4.15).

La estimación directa lineal involucra tres elementos: *i*) la dinámica visual de la región de la boca, es decir, la matriz $\mathbf{C}$, cuyas columnas son las unidades visuales reales; *ii*) una matriz $\mathbf{T}$, cuyas columnas representan la parametrización de las tramas de voz correspondientes a las unidades anteriores; y *iii*) una matriz $\mathbf{E}$, que expresa la mejor relación lineal según mínimos cuadrados entre las dos anteriores. $\mathbf{E}$ representa la RVV en este método, y relaciona $\mathbf{C}$ y $\mathbf{T}$ como sigue:

$$\mathbf{C} = \mathbf{ET} \tag{4.14}$$

Para encontrar $\mathbf{E}$, simplemente hay que utilizar la definición de inversa por la derecha de $\mathbf{T}$ (Moon, 1999):

$$\mathbf{E} = \mathbf{C}\mathbf{T}^T \left(\mathbf{T}\mathbf{T}^T\right)^{-1} \tag{4.15}$$

Una vez conocida $\mathbf{E}$, se puede obtener una estimación de unidades visuales reales $\mathbf{c}_n$ a partir de una trama de voz parametrizada $\mathbf{t}_n$ como:

$$\mathbf{c}_n = \mathbf{E}\mathbf{t}_n \tag{4.16}$$

Existe también otro procedimiento equivalente para encontrar la matriz $\mathbf{E}$, que consiste en la concatenación vertical de las matrices $\mathbf{C}$ y $\mathbf{T}$, formando una matriz $\mathbf{M}$ con columnas que contienen información visual y acústica del mismo instante temporal. A continuación se realiza la SVD de $\mathbf{M}$, obteniendo $\mathbf{U}_M$, $\mathbf{\Sigma}_M$ y $\mathbf{V}_M$:

$$\begin{bmatrix}\mathbf{T}\\ \mathbf{C}\end{bmatrix} = \mathbf{M} = \mathbf{U}_M\mathbf{\Sigma}_M\mathbf{V}_M^T = \begin{bmatrix}\mathbf{U}_V\\ \mathbf{U}_C\end{bmatrix}\mathbf{\Sigma}_M\mathbf{V}_M^T = \begin{bmatrix}\mathbf{U}_V\mathbf{\Sigma}_M\mathbf{V}_M^T\\ \mathbf{U}_C\mathbf{\Sigma}_M\mathbf{V}_M^T\end{bmatrix} \tag{4.17}$$

donde las matrices $\mathbf{U}_T$ y $\mathbf{U}_C$ corresponden, respectivamente, a las bases relativas a los parámetros acústicos y las unidades visuales reales. Para estimar una unidad visual real $\mathbf{c}_n$, se busca primero la expresión del vector de parámetros acústicos $\mathbf{v}_n$ asociado, según la base $\mathbf{U}_V$ (4.18) y, a continuación, se utiliza $\mathbf{v}_n$ para obtener la unidad visual real. Este procedimiento se puede resumir en un valor de $\mathbf{E}$ concreto (4.20), que se puedea aplicar directamente a toda trama $\mathbf{t}_n$.

$$\mathbf{v}_n = \mathbf{U}_V^T\mathbf{t}_n \tag{4.18}$$

$$\mathbf{c}_n = \mathbf{U}_C\mathbf{v}_n = \mathbf{U}_C\mathbf{U}_V^T\mathbf{t}_n = \mathbf{E}\mathbf{t}_n \tag{4.19}$$

$$\mathbf{E} = \mathbf{U}_C\mathbf{U}_V^T \tag{4.20}$$

Ambos procesos ofrecen, en realidad, la misma solución. Sustituyendo (4.17) sobre la estimación lineal (4.15), se obtiene el mismo resultado para $\mathbf{E}$ (4.21) que a partir del método que utiliza la SVD (4.20).

$$\begin{aligned}\mathbf{E} &= \mathbf{C}\mathbf{T}^T\left(\mathbf{T}\mathbf{T}^T\right)^{-1} = \mathbf{U}_C\mathbf{\Sigma}_M\mathbf{V}_M^T\mathbf{V_M}\mathbf{\Sigma}_M\mathbf{U}_V^T\left(\mathbf{U}_V\mathbf{\Sigma}_M^2\mathbf{U}_V^T\right)^{-1} = \\ &= \mathbf{U}_C\mathbf{\Sigma}^2\mathbf{U}_V^T\mathbf{U}_V\mathbf{\Sigma}_M^{-2}\mathbf{U}_V^T = \mathbf{U}_C\mathbf{U}_V^T\end{aligned} \tag{4.21}$$

4.2.5.2. Estimación bayesiana

Cuando resulta difícil modelar un comportamiento mediante una expresión matemática, una posible solución se encuentra en la teoría de probabilidades. La idea es asociar una medida de probabilidad al efecto que se desea predecir o estimar a partir de unas observaciones. En este trabajo se ofrece la aplicación de estimación bayesiana al problema de obtener la información visual a partir de la auditiva. Si no se conoce ningún tipo de información a priori sobre este efecto, la estimación se basa comúnmente en el principio de máxima verosimilitud (ML); si esta información se encuentra disponible, se utiliza el máximo a posteriori (MAP). Este tipo de técnicas es genérica, ya que no depende del orígen ni la naturaleza de la información. Por otro lado, se necesitan funciones de densidad de probabilidad, normalmente de desconocidas, con lo que se deben aproximar a partir de los propios datos disponibles.

En el caso que ocupa en este trabajo, las observaciones son las tramas de voz parametrizadas $\mathbf{t}_n$. El efecto a predecir sería, en teoría, la unidad visual real asociada a la región de la boca; sin embargo, dado que en la práctica el número de datos disponibles es limitado, se trabaja con agrupaciones de vectores o clases como elementos a estimar,

denotados por la variable aleatoria discreta θ. Cada clase θ_p tiene asociada una función de probabilidad $p(\mathbf{t}_n|\theta_p)$ y está asociada a un conjunto de unidades visuales reales $\mathbf{\Psi}_p$. El proceso de estimación es el RVI, mientras que la relación entre θ_p y $\mathbf{\Psi}_p$ es una CIV, en el contexto de esta propuesta. Para aproximar las funciones de densidad de probabilidad, se propone utilizar una gaussiana multidimensional $\mathcal{N}_p$ calculada al igual que en apartado 4.2.2.1. La información a priori sobre el efecto se denota por la función de probabilidad $p(\theta = \vartheta)$ y puede estar o no disponible.

La estimación puede venir representada a través del uso del principio de ML, que se basa en encontrar el valor de θ que maximiza la probabilidad de generar los valores observados $\mathbf{t}_n$, obteniendo $\hat{\theta}_{\mathrm{ML}}$ (y, por consiguiente, $\mathbf{\Psi}_{\mathrm{ML}}$):

$$\hat{\theta}_{\mathrm{ML}} = \max_{\theta} p(\mathbf{t}_n|\theta)$$

Si se dispone de la información a priori sobre θ, utilizando el teorema de Bayes (Moon, 1999), se puede intentar maximizar la probabilidad a posteriori $p(\theta|\mathbf{t})$ para obtener $\mathbf{\Psi}_{\mathrm{MAP}}$ a través de encontrar $\hat{\theta}_{\mathrm{MAP}}$ (4.22). Esta expresión se conoce como MAP.

$$\hat{\theta}_{\mathrm{MAP}} = \max_{\theta} p(\theta|\mathbf{t}) = \max_{\theta} \frac{p(\mathbf{t}|\theta)\, p(\theta)}{p(\mathbf{t})} = \max_{\theta} p(\mathbf{t}|\theta)\, p(\theta) \tag{4.22}$$

En el caso particular en que sólo exista un candidato $\mathbf{c}$ por conjunto $\mathbf{\Psi}_p$, se puede plantear el problema como la minimización de la función de error (4.23), con solución (4.24).

$$\hat{\theta} = \min_{\tilde{\theta}} E\left[\left\|\tilde{\theta} - \theta_k\right\|_2^2\right] \tag{4.23}$$

$$\mathbf{c}_{\mathrm{MMSE}} = E[\theta] = \sum_{\forall k} \mathbf{c}_k p(\theta_k|\mathbf{t}) \tag{4.24}$$

La esperanza se realiza sobre las unidades visuales reales asociadas a cada clase, obteniendo el mínimo error cuadrático medio (MMSE) $\mathbf{c}_{\mathrm{MMSE}}$. Utilizando el teorema de Bayes y la ley de probabilidad total (4.25) sobre (4.24) se puede expresar la solución en función de las densidades de probabilidad y las probabilidades a priori (4.26).

$$p(\mathbf{t}) = \sum_{\forall k} p(\mathbf{t}|\theta_k)\, p(\theta_k) \tag{4.25}$$

$$\mathbf{c}_{\mathrm{MMSE}} = \sum_{\forall k} \mathbf{c}_k \frac{p(\mathbf{t}|\theta_k)\, p(\theta_k)}{p(\mathbf{t})} = \frac{1}{\sum_{\forall k} p(\mathbf{t}|\theta_k)\, p(\theta_k)} \sum_{\forall k} \mathbf{c}_k p(\mathbf{t}|\theta_k)\, p(\theta_k) \tag{4.26}$$

En este caso, no se distingue entre un conjunto limitado de clases, ya que se obtiene la unidad visual real directamente en $\mathbf{c}_{\mathrm{MMSE}}$, con lo que la RVV consta únicamente de un estimador de visemas y se evita el uso de ninguna CIV. Sin embargo, este método tiene el problema de generar puntos no incluidos dentro de la dinámica visual descrita por el modelo visual ya que no la tiene en cuenta en su tratamiento. Por esta razón, este método no se utiliza en este trabajo, aunque no se descarta adaptarlo en un futuro ya que, sin tener en cuenta la dinámica visual, ofrece los mejores resultados (Melenchón et al., 2003b).

4.2.6. Incertidumbre Audiovisual

La estimación realizada en el apartado 4.2.5.2 utiliza la agrupación en conjuntos de unidades visuales reales parecidas o visemas. No obstante, dado que la estimación se realiza a partir de la información de audio, ésta tiene limitaciones ya que la información auditiva se encuentra agrupada por un criterio visual; por ejemplo, distinguir entre los visemas asociados a los alófonos [m] y [n] será complicado ya que éstos son muy parecidos acústicamente. No obstante, si la información auditiva estuviese agrupada o clasificada siguiendo un criterio acústico, la estimación mejoraría necesariamente. Se propone comprobar este hecho mediante cuatro agrupaciones diferentes:

1. Agrupar la información visual mediante la matriz de similaridad del apartado 4.2.2.
2. Agrupar la información auditiva mediante la matriz de similaridad del apartado 4.2.2, pero construida a partir de la distancia de Bhattacharyya entre las distribuciones normales asociadas a las tramas de voz.
3. Utilizar la agrupación ofrecida por un *K-means* sobre la información visual.
4. Utilizar la agrupación ofrecida por un *K-means* sobre la información aúditiva.

Cada una de las agrupaciones (ver cuadro 4.1) se puede utilizar para construir las funciones de densidad de probabilidad $p_i\left(\mathbf{t}_n|\theta_p\right)$ (ver apartado 5.1.5.4). Cada tipo de agrupación será capaz de estimar la información visual identificada por las diferentes clases θ_p a partir del audio en diferente medida. Se puede intentar realizar también la estimación contraria, es decir, definir las distintas $p_j\left(\mathbf{c}_n|\theta_{\text{audio},p}\right)$ para estimar la información auditiva, identificada por las clases $\theta_{\text{audio},p}$, a partir de la visual.

Agrupación	Matriz de similaridad		*K-means*	
por	de audio	de vídeo	de audio	de vídeo
Audio	Agrupación 1	Agrupación 2	Agrupación 3	Agrupación 4
Vídeo	Agrupación 5	Agrupación 6	Agrupación 7	Agrupación 8

Cuadro 4.1: Agrupaciones obtenidas sobre los datos de audio y vídeo. Los grupos pueden estar definidos mediante el resultado del algoritmo *K-means* sobre el audio o el vídeo o mediante el uso de matrices de similaridad entre las de unidades visuales reales inicialmente clasificadas por alófonos.

Al comparar el porcentaje de vectores $\mathbf{t}_n$ y $\mathbf{c}_n$ correctamente clasificados se obtienen resultados interesantes: *i*) aumentar la calidad de estimación de un tipo de información hace disminuir la del otro; *ii*) la media geométrica de porcentajes de clasificación se mantiene aproximadamente constante a través de los diferentes agrupamientos para una misma persona. Se remite al lector al capítulo 5 para más información. Este resultado va totalmente en la línea marcada en trabajos como los de Chen y Rao (1998), donde expresa que la información dada por un sólo modo es limitada y para incrementarla, se deben utilizar ambos simultáneamente.

4.3. Algoritmo general de síntesis

En este apartado se presenta el algoritmo general de síntesis audiovisual (ver figura 4.16) y se detallan las propiedades del mismo. Dado un modelo visual y una serie ordenada de elementos que se puedan asociar a los diferentes conjuntos de unidades visuales reales $\mathbf{\Psi}^l$ para las L regiones, se puede construir una secuencia audiovisual o simplemente visual si no existe información auditiva asociada. En el caso específico de sintetizar caras parlantes, la información que guía la síntesis puede ser fonética o auditiva.

```
[S] = Síntesis(M, A)

    Entrada:
    M: Modelo visual
    A: Información de guía
    Salida:
    S: Secuencia sintética

    1. Conversión de A a secuencia de Ψ_p^l
    2. Construcción de trayectorias T^l a partir de Ψ_p^l
    3. Muestreo de T^l en unidades visuales virtuales ĉ_m^l
    4. Construcción de las imágenes I_m a partir de ĉ_m^l y M
    5. Concatenación de I_m
    6. Incorporación del canal auditivo si existe
```

Figura 4.16: Algoritmo general de síntesis.

Hay que destacar que ningún paso del algoritmo de la figura 4.16 requiere interacción humana y el sistema es capaz de producir el resultado por sí mismo (ver cuadro 4.2). Un caso especial del paso 1 se encuentra descrito en el apartado 4.2, los pasos 5 y 6 son autoexplicativos y el resto (2, 3 y 4) están detallados en el apartado 4.1.

Paso	Interacción
(1) Conversión	Nula
(2) Trayectorias	Nula
(3) Muestreo	Nula
(4) Imágenes	Nula
(5) Secuencia	Nula
(6) Sonido	Nula

Cuadro 4.2: Requisitos de interacción humana en cada paso del algoritmo general de síntesis.

4.3.1. Cualidades de la síntesis

El proceso de síntesis presentado en este capítulo se caracteriza por su facilidad de uso, realismo y flexibilidad, cualidades aportadas por el uso de las técnicas vistas en el apartado 4.1 y en el apartado 4.2 (ver cuadro 4.3).

Cualidad	Características
Flexibilidad	Conversión fonética y síntesis genéricas
Facilidad de uso	Conversión fonética automática y síntesis sin intervención del usuario
Realismo	Uso de un modelo visual acotado, conversión fonética sincronizada y síntesis realista

Cuadro 4.3: Cualidades del algoritmo de síntesis y sus características asociadas.

4.3.1.1. Flexibilidad

Gracias a no crear una dependencia especial de los métodos de conversión fonética y síntesis con ningún tipo de imágenes, todas las técnicas de síntesis desarrolladas se pueden aplicar a cualquier tipo de imágenes que sea capaz de representar el modelo visual sin necesidad de cambiar su funcionamiento en absoluto. Este hecho no prescinde de la posibilidad de aplicar mejoras concretas a un caso preciso (como el mostrado sobre las caras parlantes, por ejemplo).

4.3.1.2. Facilidad de uso

La síntesis de secuencias es totalmente automática, dado un modelo visual y una secuencia de conjuntos de unidades visuales reales. En el caso de cabezas parlantes, la carencia de intervención manual en los procesos de conversión fonética y síntesis de imágenes faciales dejan todo el trabajo para la máquina de cálculo utilizada.

4.3.1.3. Realismo

El esquema presentado está plenamente orientado a conseguir realismo estático y dinámico en la síntesis. El primero, gracias a utilizar un proceso de unión de regiones de forma gradual, simular PSFs e incluir imágenes originales y el segundo, gracias a usar un algoritmo de interpolación que se mantiene dentro del subespacio visual definido por el modelo. En el caso de cabezas parlantes, la sincronía aportada por los sistemas de conversión fonética también es un punto que favorece esta característica.

Capítulo 5

Resultados

En los tres capítulos anteriores (2, 3 y 4) se han presentado las estructuras, técnicas y métodos desarrollados en este trabajo de investigación, con unas cualidades determinadas. En este capítulo se muestra una discusión de todas ellas en base a una serie de experimentos, con sus respectivos resultados.

El lector puede encontrar ejemplos relativos a la creación de corpus (ver apartado 2.1), la construcción de modelos visuales (ver apartado 2.2), los algoritmos de seguimiento (ver apartado 3.1) y aprendizaje (ver apartado 3.2) y la síntesis de secuencias (ver apartado 4.1) audiovisuales. Dado su especial interés, se ofrecen también ejemplos relacionados con la síntesis de caras parlantes, incluyendo síntesis a partir de texto y voz, cambios de expresiones y una evaluación de los efectos de énfasis visual (ver apartado 4.1.3.3). También se incluyen ejemplos de síntesis de lenguaje de signos, además de ejemplos puntuales con otro tipo de objetos para mostrar las capacidades del esquema desarrollado.

Debido a las limitaciones inherentes del formato de presentación de la memoria, se invita al lector a visitar la página personal del autor (hasta el momento de la impresión de este documento se encuentra en `http://www.salleurl.edu/~jmelen`), donde puede encontrar información multimedia adicional relacionada con este trabajo de investigación, así como otra información general sobre el autor del mismo.

5.1. Flexibilidad, fiabilidad y facilidad de uso

La aplicación de los métodos de análisis y síntesis a objetos tan diferentes como una cara, una mano, una botella o efectos de iluminación (ver apartados 5.1.3, 5.1.4 y 5.1.5) ofrece una idea general de su flexibilidad de aplicación a diferentes tipos de objetos. Además, las condiciones de grabación de cada corpus han sido heterogéneas en cuanto al fondo y las condiciones de iluminación utilizadas, sin afectar al éxito del análisis y de la síntesis. Por otro lado, la síntesis efectuada se ha podido realizar a partir de información de voz, texto o símbolos, dando así una idea de la variedad de entradas posibles.

La facilidad de uso de los algoritmos desarrollados ha permitido el uso de dispositivos domésticos, no tener que cuidar ningún tipo de iluminación en concreto, independizarse del fondo de imagen utilizado y no necesitar ningún tipo de marcadores ni imágenes previas del objeto de interés. Estos hechos facilitan el proceso de captura del corpus, que es la entrada del esquema desarrollado. Se han desarrollado prototipos de aplicaciones informáticas orientadas a usuarios no expertos como la de *PREVIS II* presentada en el apéndice D. Además, la no intervención humana presentada a lo largo de los procesos existentes en los capítulos 3 y 4 y experimentada en la generación, entre otras, de las imágenes de los apartados 5.1.4 y 5.1.5, representan una muestra del gran automatismo que posee el esquema presentado.

El esquema de análisis presentado ha permitido crear con éxito y sin necesidad de repeticiones más de cuarenta modelos visuales a partir de diferentes corpus relativos a un abanico de objetos muy variado. Esto ha sido posible, concretamente, gracias al algoritmo combinado de seguimiento y aprendizaje (ver apartado 3.3), con su inherente capacidad para recoger dinámicamente los cambios de apariencia de los objetos de interés, y a la acotación del modelo, que restringe la variedad de apariencias dentro de un margen razonable.

5.1.1. Grabación de corpus

En esta primera sección de resultados se muestran diferentes ejemplos de corpus registrados. La mayoría de ellos son ejemplos relativos a caras parlantes, debido a la atención especial que se le presta a este tipo de objetos, aunque también se adjuntan otros de diferentes características.

Siguiendo las instrucciones para la grabación de corpus dadas en el apartado 2.1, por un lado se han gravado 27 secuencias audiovisuales a una resolución de 320 por 240 píxeles, tres por cada uno de nueve individuos. Debido, sobretodo, a los procesos de investigación relacionados con la síntesis a partir de voz, estas grabaciones se han efectuado en un estudio de grabación profesional, registrando el sonido a 48KHz y 24 bits para disponer de una alta calidad en la información acústica. En la figura 5.1 se pueden observar las nueve personas incluidas en este corpus audiovisual en un momento de su grabación. Por otro lado, también se han realizado grabaciones de corpus en entorno doméstico, unos a una resolución de 720 por 578 píxeles y sonido capturado a 16KHz y 16 bits y otros a 320 por 240 píxeles e igual calidad de sonido (ver figura 5.2). Estas últimas condiciones son asequibles fácilmente con cualquier aparato de adquisición de vídeo típico en la actualidad.

También se han capturado cuatro corpus en el que aparece una mano realizando diferentes movimientos siguiendo el lenguaje de signos dactilológico español (Ministerio de Educación y Ciencia, 1990). En este caso la grabación ha sido realizada sin sonido y a una velocidad de veinticinco cuadros por segundo y una resolución de 320×240 píxeles. En otro caso se ha realizado una captura a cien imágenes por segundo y una resolución más reducida de 220×212; esta alta velocidad ha sido posible gracias al uso de aparatos de adquisición profesionales no presentes en un entorno doméstico. Su uso ha venido motivado por la necesidad de investigar estos símbolos en concreto con otras finalidades diferentes a

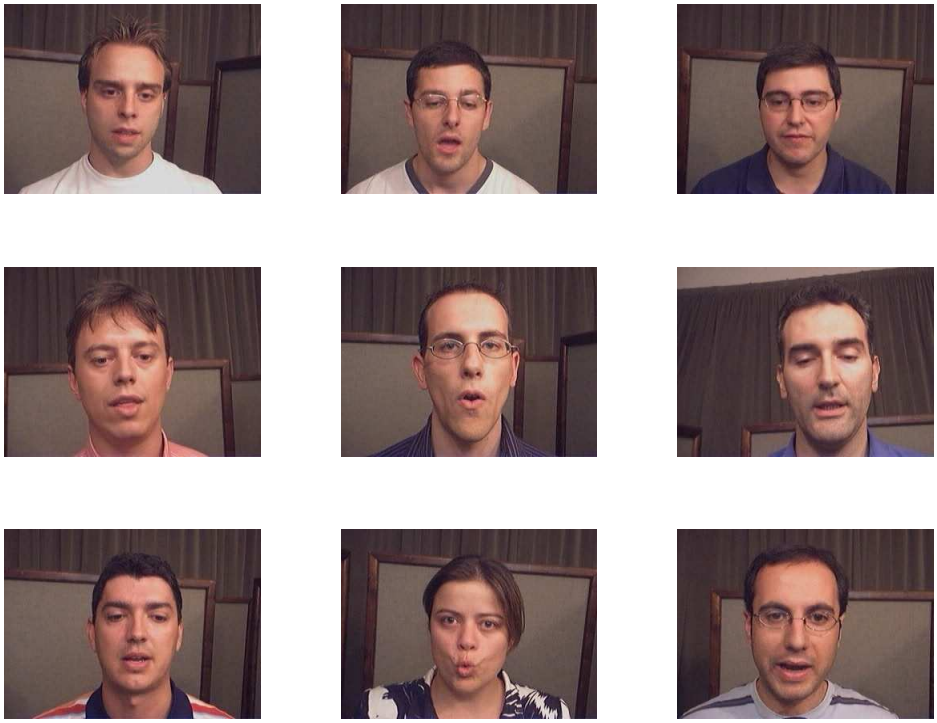


Figura 5.1: Los nueve individuos presentes en un corpus audiovisual registrado en un estudio profesional.

las presentadas en el apartado 1.4. Imágenes relativas a este corpus y su proceso de captura se pueden observar en la figura 5.3.

5.1.2. Creación de modelos

A partir de los corpus grabados se generan cada uno de los modelos visuales. Éstos son capaces de representar la apariencia que se ha observado en cada corpus. En la figura 5.4 se pueden observar las representaciones dadas por el modelo para cada una de las imágenes de la figura 5.1. Notar la gran similitud entre ellas y el efecto de normalización de tamaño que contiene el propio modelo, fruto del proceso de alineación (ver apartado 3.1) realizado sobre los corpus.

Del mismo modo, en la figura 5.5 se muestra la representación dada por el modelo visual en el caso de contener la información relativa al corpus registrado con lenguaje de signos (ver apartado 5.1.1).

Figura 5.2: Imágenes del corpus doméstico registrado sobre personas.

5.1.3. Algoritmo de análisis

En la construcción de los diferentes modelos, se ha comprobado indirectamente la efectividad del algoritmo general de análisis (ver apartado 3.3) mediante la observación de su aplicación sobre diferentes ejemplos en condiciones diversas de iluminación y entorno. El algoritmo se ha utilizado sobre más de cien corpus diferentes, resultando exitoso en todos.

El algoritmo de seguimiento concreto utilizado para la construcción de modelos visuales se basa en el de la figura 3.4, con las siguientes extensiones (de entre las presentadas en el apartado 3.1.2.4):

- Usa un subespacio como referencia (las matrices $\mathbf{U}^l$ del modelo visual concretamente).
- Utiliza subespacios modulares (uno para cada región l).
- Posee multirresolución.

Utilizar una única imagen de referencia en vez de un subespacio de apariencia ofrece peores resultados (Matthews et al., 2003) en algoritmos basados en el trabajo de Lucas y Kanade (1981). Así, el hecho de conocer en cada momento el subespacio de apariencia $\mathbf{U}^l$ de cada objeto aumenta las posibilidades de realizar una localización correcta del mismo. La modularidad y la multirresolución se incluyen para reducir el consumo de espacio y facilitar la localización de los objetos, respectivamente. La optimización del algoritmo de seguimiento no se ha implementado en la aplicación global aunque sí se ha probado por separado, obteniendo una reducción en la carga computacional del 40 % aproximadamente. La robustez también se ha testado por separado pero no se ha incluido a la hora de generar los modelos visuales ya que no aparecen *outliers* importantes en los corpus registrados. Su uso provocaría un aumento de la carga computacional sin necesidad. En las figuras 5.6 y 5.7 se observan diferentes ejemplos del resultado del seguimiento sobre varios objetos.

(a)

(b)

Figura 5.3: Grabación de un corpus de lenguaje de signos. A la derecha de los dos ejemplos se pueden observar diferentes cuadros de la secuencia registrada y a la izquierda, el entorno de grabación. El uso de aparatos específicos en el ejemplo (b) inferior viene motivado en este caso por razones ajenas a este trabajo.

5.1.4. Síntesis a partir de un modelo no facial

El esquema de síntesis presentado es capaz de reproducir secuencias de imágenes parecidas a las contenidas en el corpus registrado (ver apartado 4.1). Si se observan objetos como los mostrados en el proceso de seguimiento en la figura 5.7, se pueden sintetizar diferentes secuencias especificando únicamente el índice de unas pocas unidades visuales reales y dejando que el algoritmo de síntesis (ver apartado 4.3) realice las interpolaciones entre ellos (ver figura 5.8).

En el caso de sintetizar secuencias de lenguaje de signos, se puede disponer de la información simbólica del mismo modo del que se dispone de la información fonética en el caso de caras parlantes guiadas por fonética (ver apartado 4.2.4). Se asocia cada símbolo a un conjunto de unidades visuales reales $\mathbf{\Psi}$ (construyendo, por tanto, una especie de CIV, pero sustituyendo el término identificador por símbolo y visema por gesto de mano). De este modo, una secuencia de símbolos se traduce en una secuencia de diferentes $\mathbf{\Psi}$ y ésta, a su vez, en secuencias de imágenes interpoladas (ver apartado 4.1.3.1) aplicando el algoritmo

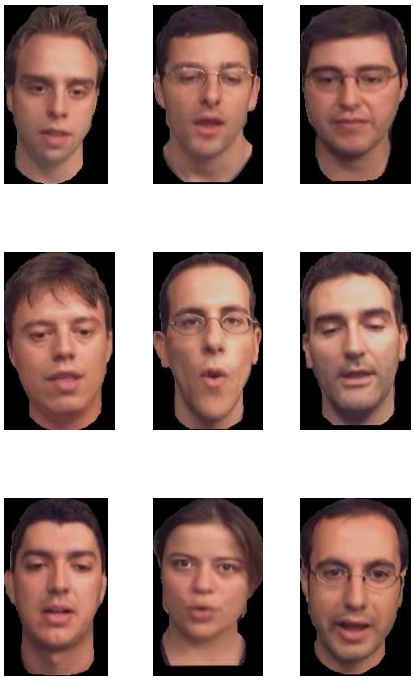


Figura 5.4: Imágenes generadas por diferentes modelos visuales relativos a caras humanas.

de selección de unidades visuales (ver apartado 4.1.3.2). En la figura 5.9 se puede observar un ejemplo de la síntesis de lenguaje de signos obtenida.

5.1.5. Síntesis a partir de un modelo facial

En este apartado se ofrecen resultados relativos a la síntesis de modelos visuales aplicados a caras humanas mientras hablan y gesticulan. En el apartado 5.1.5.1 se ofrecen ejemplos sobre el control de los diferentes elementos visuales en general y en el apartado 5.1.5.2 se muestra el control del énfasis en la síntesis de la zona de la boca. También se muestran los resultados relativos a la obtención personalizada de visemas en el apartado 5.1.5.3 y los efectos de incertidumbre audiovisual en el 5.1.5.4. Finalmente, se ofrecen diferentes ejemplos de los resultados obtenidos en la síntesis guiada por fonética y voz (apartados 5.1.5.5 y 5.1.5.6).

5.1.5.1. Síntesis de elementos faciales

En este apartado se muestra el control en la generación de gestos sobre los diferentes elementos faciales asociados a una cara almacenada en un modelo visual. El control independiente de cada uno de los elementos faciales se realiza gracias a la característica de modularidad del modelo visual (ver apartado 2.2.1). En el caso que se muestra en la figura 5.10 se han considerando cuatro elementos faciales: la boca juntamente con el maxilar

Figura 5.5: Imágenes generadas por un modelo visual construido a partir de la grabación de una mano comunicándose en lenguaje de signos.

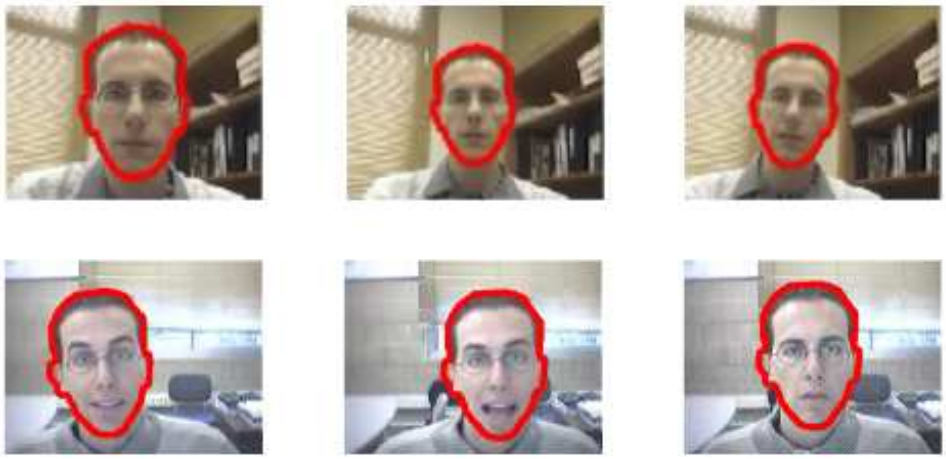


Figura 5.6: Diferentes ejemplos de seguimiento de caras. Para poder comprobar visualmente su comportamiento, se ha coloreado en rojo el borde del objeto que se desea seguir, en este caso, caras. Las imágenes superiores contienen un movimiento suave, mientras que las inferiores poseen movimientos más bruscos.

inferior, el ojo izquierdo, el derecho y el resto de la cara. Cada elemento facial se puede sintetizar por separado como se puede apreciar en las subfiguras 5.10(b), 5.10(c) y 5.10(d), en las que se arquea una ceja, se guiña un ojo y se sonríe de forma leve, respectivamente. También se puede realizar la síntesis conjunta de todos los elementos, como se puede comprobar en la subfigura 5.10(a).

En la figura 5.11 se muestran tres transiciones entre diferentes estados de ánimo en las que únicamente se ha especificado los puntos de partida y fin mediante dos índices de imagen, de modo similar que en los ejemplos de las figuras 5.8 y 5.9.

5.1.5.2. Síntesis enfática

En este apartado se muestran las pruebas realizadas para comprobar si los efectos de énfasis visual propuestos en el apartado 4.1.3.3 son distinguibles entre ellos y si ejercen influencia en la percepción del realismo en el mensaje.

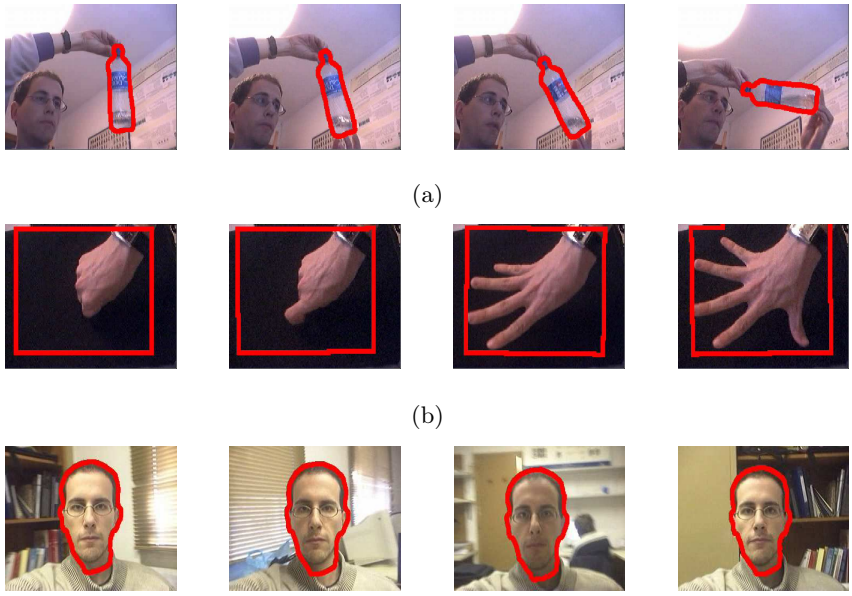


Figura 5.7: Ejemplos de seguimiento de tres objetos: una botella (a), una mano (b) y una cara con fuertes cambios de iluminación (c).

Se han realizado dos pruebas que constan de una serie de características comunes:

- Ha sido realizada por 94 evaluadores no expertos con edades comprendidas entre 14 y 56 años y una distribución equitativa de sexos.
- Se han mostrado secuencias audiovisuales con la cara de una persona (siempre la misma) a 25 imágenes por segundo, de un tamaño de 336×256 píxeles, una frecuencia de muestreo de sonido de 8000 Hz y 16 bits por muestra.
- La duración de los vídeos es de alrededor de 10 segundos cada uno.
- Se han introducido puntos de control para medir la consistencia de cada evaluador. Siguiendo esta medida, se ha tenido en cuenta el primer 50 % de los evaluadores ordenados de mejor a peor consistencia.
- Cada test ha proporcionado una MOS.

La primera prueba analizó la correlación percibida entre el énfasis visual percibido y sintetizado (ver α en la apartado 4.1.3.3). La segunda pretendía conocer si existía algún tipo de influencia entre el énfasis visual y dos estilos concretos de comunicación audiovisual mediante la cara humana: el habla y el canto.

Prueba de nivel de énfasis En esta prueba se pidió a los evaluadores que midieran el nivel de énfasis visual dado por el movimiento de los labios de una cara sintetizada artificialmente mientras pronunciaba una frase de 9 segundos de duración. Algunas imágenes

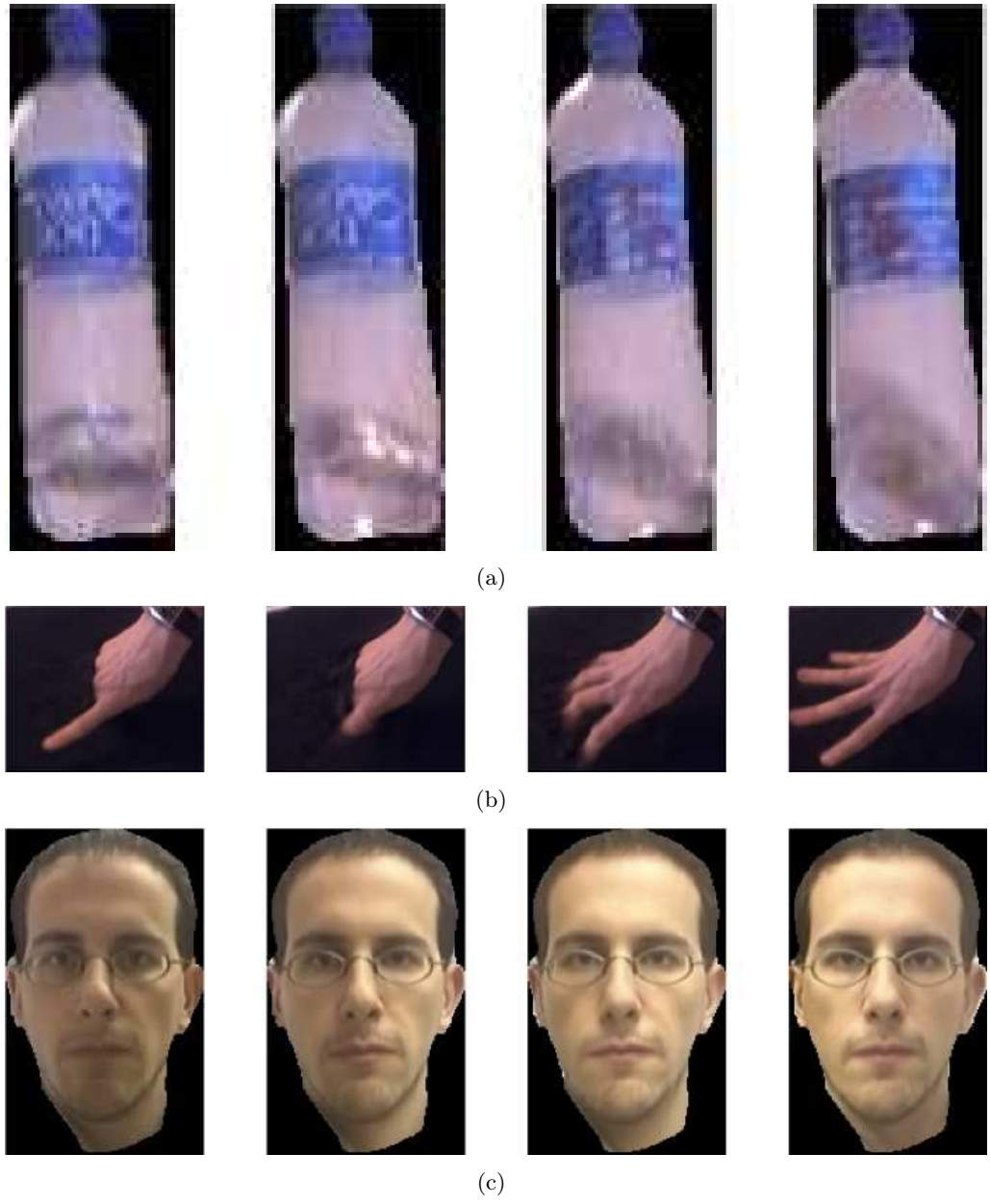


Figura 5.8: Ejemplos de síntesis de cambios de apariencia en una botella (a) y una mano (b) y generación de cambios de iluminación en una cara (c), a partir de sus respectivos modelos previamente aprendidos y la especificación de dos unidades visuales reales en cada caso (la inicial y la final).

de este test se pueden observar en la figura 4.9 del apartado 4.1.3.3. Los valores posibles abarcaban el rango desde uno (menor énfasis) hasta 5 (mayor énfasis). En esta prueba se mostraron inicialmente a cada evaluador dos vídeos con valores de énfasis extremos para intentar normalizar los límites en cada evaluación. El contenido de la prueba constó de cuatro vídeos diferentes con valores de énfasis (α) de 0,0 0,3, 0,7 y 1,0. Cada vídeo se repe-

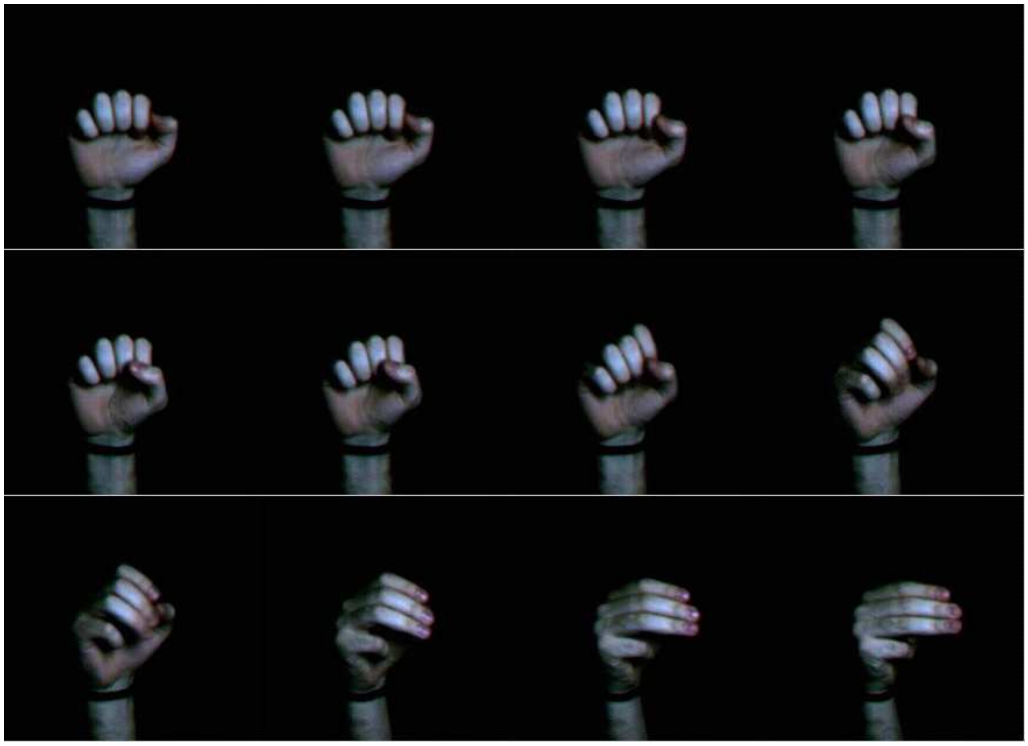


Figura 5.9: De izquierda a derecha y de arriba a abajo: doce imágenes de la síntesis de la secuencia correspondiente a la transición *ab* realizada en lengua dactilológica española a partir de dos unidades visuales reales, la de inicio (superior izquierda) y la final (inferior derecha).

tía dos veces y estaban desordenados aleatoriamente. El coeficiente de correlación cruzada de Spearman revela una relación estadísticamente significativa del 99 %. Los resultados específicos de esta prueba se pueden consultar en el cuadro 5.1 y en el diagrama de cajas (o en inglés *boxplot*) de la figura 5.12.

Prueba de estilo La comprobación de la influencia de los efectos de énfasis visual sobre la percepción de caras parlantes fue realizada en esta segunda prueba. Se mostraron seis vídeos diferentes (repetidos dos veces cada uno) variando el estilo de comunicación y el nivel de énfasis. Los estilos fueron hablar y cantar y los niveles de énfasis fueron 0,0, 0,5 y 1,0. El cuadro 5.1 revela las puntuaciones obtenidas para cada estilo y grado de énfasis. En el caso del estilo cantado, el contraste de Kruskal-Wallis[1] revela diferencias estadísticamente no significativas ($p < 0{,}62$) (Sheskin, 2007) entre las evaluaciones asociadas a los diferentes niveles. En cambio, en el caso del estilo hablado, se ha encontrado que la distribución de los resultados asociados al nivel de énfasis 0,0 está más inclinada hacia valores altos que las de los otros dos niveles con un nivel de significación estadística del 99 % (según el test de Kolmogorov-Smirnov para dos conjuntos mostrales); por otro lado, también se revelan diferencias estadísticamente significativas (con el mismo tipo de test de hipótesis) entre los niveles 0,5 y 1,0 ($p > 0{,}98$), siendo el segundo menor que el primero. A partir de estos

[1]Tipo de ANOVA no paramétrico que se utiliza con distribuciones similares, sin los requisitos de normalidad de la versión paramétrica (Devore, 2000).

(a)

(b)

(c)

(d)

Figura 5.10: Control de síntesis de los diferentes elementos faciales contenidos en un modelo visual que representa una cara humana. La secuencias (b), (c) y (d) contienen gesticulaciones de diferentes elementos faciales, que son un arqueo de ceja, un guiño y una sonrisa. En (a) se puede observar la unión simultánea de todos estos movimientos.

resultados y los de la figura 5.13 y el cuadro 5.1 se puede interpretar que al hablar se percibe más realismo con un nivel bajo de énfasis, mientras que al cantar no hay ningún nivel de énfasis preferido. Este resultado concuerda con las conclusiones de Pourtois et al. (2002): el realismo percibido depende de la manera en que se habla. En este caso, hablar con diferentes niveles de énfasis produce percepciones de realismo diferentes. El hecho de que

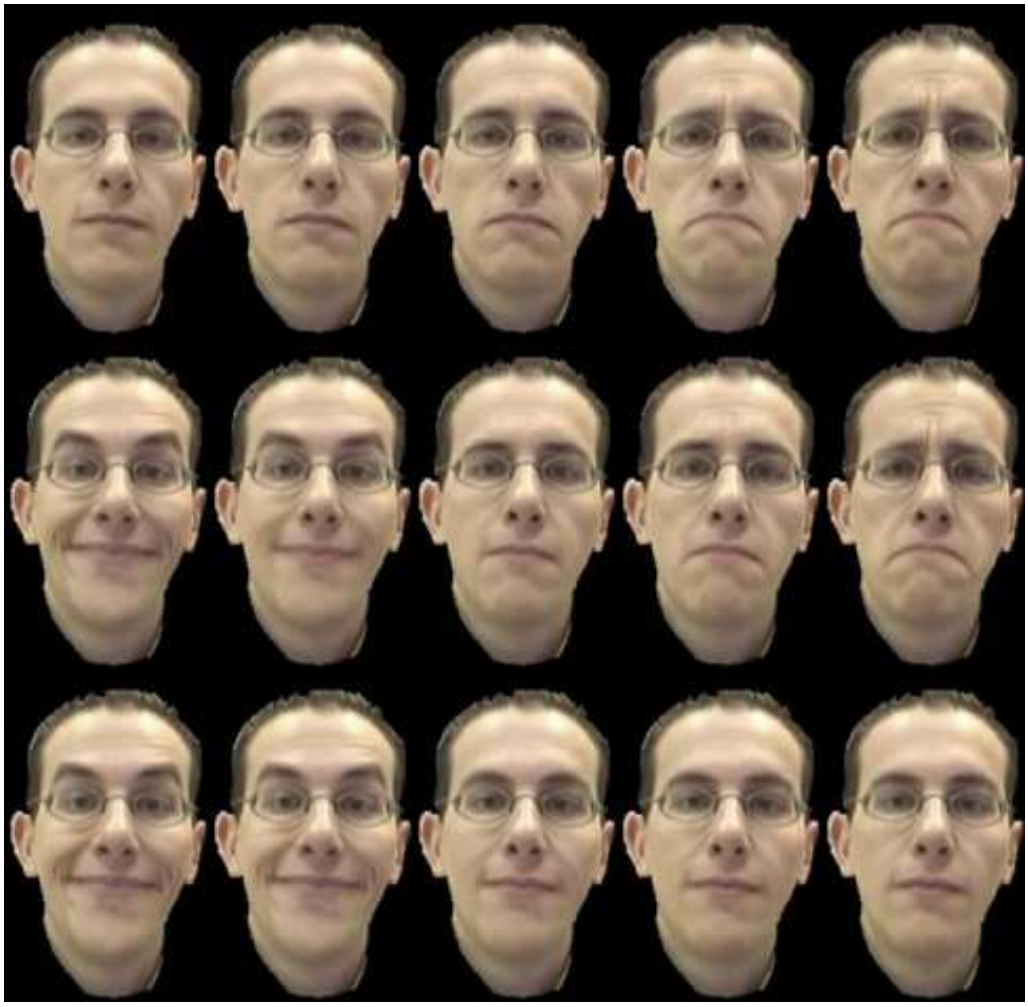


Figura 5.11: Diferentes secuencias sintéticas relativas a cambios de estados de ánimo entre alegría, seriedad y tristeza. Su construcción sólo necesita la especificación de la imagen inicial y la final en cada caso.

al cantar no se produzcan estos efectos va asociado al cambio del estilo en la comunicación; los evaluadores aceptan niveles más altos de énfasis dado que en el canto se puede llegar a gesticular más que en el habla. En resumen, se puede decir que existe una interacción entre el énfasis y el estilo de comunicación.

5.1.5.3. Obtención personalizada de visemas

Se ha aplicado el proceso de extracción personalizada de visemas (ver apartado 4.2.2) asociados a consonantes en tres corpus audiovisuales de tres minutos cada uno pronunciando el conjunto de las doce frases con sentido, indicadas en el cuadro 2.5 del apartado 2.2.2. Han sido grabados a 25 cuadros por segundo con una resolución de 320 por 240, una frecuencia de muestreo del canal de audio de 16000 Hz y 16 bits por muestra.

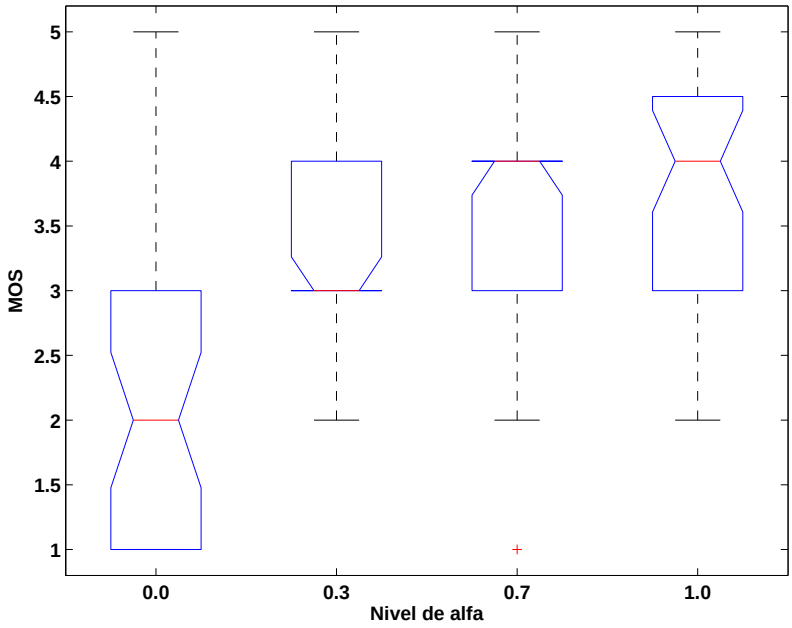


Figura 5.12: Diagrama de cajas asociado a la MOS de cuatro vídeos con niveles de énfasis diferentes.

Primera prueba		Segunda prueba		
			Canto	Habla
α	MOS	α	MOS	MOS
0,0	2,14	0,0	3,61	3,63
0,3	3,31	0,5	3,45	3,35
0,7	3,64	1,0	3,56	2,99
1,0	3,78			

Cuadro 5.1: Resultados del MOS de las dos pruebas de énfasis.

Debido a la propia distribución de alófonos castellanos y al tamaño del conjunto de frases, algunos de los alófonos no han aparecido un número suficiente de veces (en este caso 12) para conseguir una correcta estimación de los parámetros necesarios para realizar la agrupación visémica personalizada (en concreto, matrices de covarianza). Los alófonos tenidos en cuenta se muestran en el cuadro 5.2 y aparecen entre 19 y 199 veces cada uno. Los alófonos descartados han sido: [c], [b], [g], [ɱ], [j], [ɲ], [ŋ], [c], [z], [ʎ], [ɾ], [ɣ], [x], [d].

Alófonos a agrupar por visemas
[f], [θ], [p], [β], [ð], [k], [m], [t], [l], [r], [n], [s]

Cuadro 5.2: Alófonos considerados en el experimento de agrupamiento visémico personalizado.

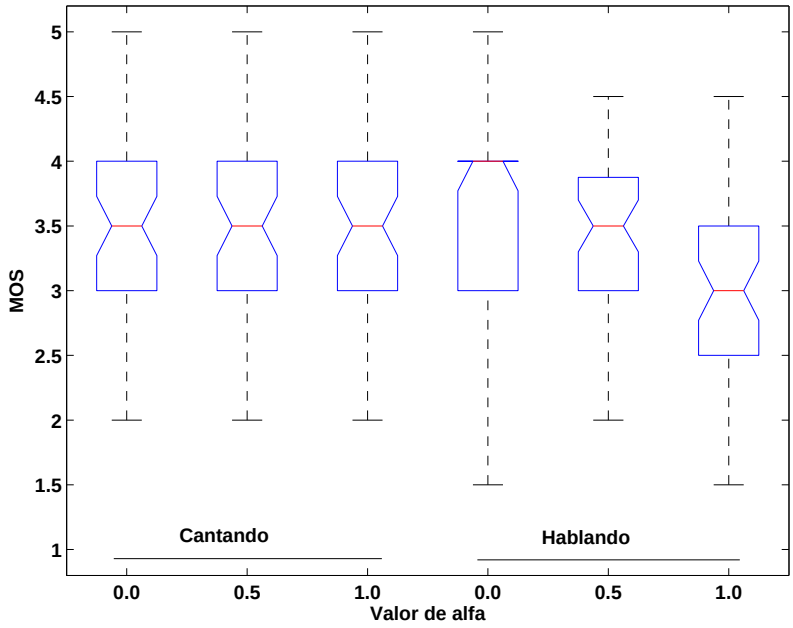


Figura 5.13: Diagrama de cajas asociado al MOS de seis vídeos con tres niveles de énfasis (0,0, 0,5 y 1,0) y dos estilos de comunicación diferentes (cantar y hablar).

Los visemas obtenidos a partir de los alófonos del cuadro 5.2 se resumen en el cuadro 5.3 para cada individuo. Se ofrece el resultado para seis grupos para poder comparar más directamente con las agrupaciones realizadas anteriormente (Owens y Blazek, 1985). Hay que destacar que para cada individuo se obtienen los mismos visemas asociados a articuladores externos, los cuales consisten en la agrupación de los alófonos [m], [p], [β], el asociado a [f] y el correspondiente a [θ]. No obstante existe un débil acuerdo en los otros tres grupos, relacionados con articuladores internos del tracto vocal.

Visema	Sujeto 1	Sujeto 2	Sujeto 3
1	[m], [p], [β]	[m], [p], [β]	[m], [p], [β]
2	[θ]	[θ]	[θ]
3	[f]	[f]	[f]
4	[ð], [t]	[ð], [t], [s]	[ð], [k]
5	[n], [r]	[n], [l], [k]	[n], [r], [l]
6	[s], [l], [k]	[r]	[t], [s]

Cuadro 5.3: Visemas obtenidos para tres personas agrupando los alófonos en seis conjuntos.

Los resultados obtenidos concuerdan con las conclusiones extraídas en trabajos anteriores (Owens y Blazek, 1985) en dos aspectos: *i*) existe coincidencia en los grupos de alófonos que involucran articuladores externos (salvando las diferencias idiomáticas entre el castellano y el inglés): visema [m], [p], [β], visema [f] y visema [θ]; *ii*) existe poco consenso en la clasificación de grupos de alófonos asociados a articuladores externos. Estos resulta-

dos refuerzan las conclusiones de este trabajo y los anteriores y sugiere una nueva manera objetiva de encontrar los visemas propios de cada persona a partir de habla natural.

5.1.5.4. Incertidumbre audiovisual

Siguiendo el tipo de agrupaciones marcadas en el cuadro 4.1 del apartado 4.2.6 se ha realizado una medida de su bondad de clasificación de información auditiva o visual al usar un rango de grupos desde 5 hasta 14. Sus valores medios a través de todos los grupos y las tres personas estudiadas se muestran en el cuadro 5.4. Como se puede ver, la mejora de clasificación para un tipo de información se traduce en un empeoramiento de la misma pero para el otro tipo de información. Además, la media geométrica entre la medida de bondad asociada a audio y la de vídeo se mantiene similar a través de las columnas, con un 7 % de diferencia entre el valor máximo y el mínimo.

Agrupación	Matriz de similaridad		*K-means*	
por	de audio	de vídeo	de audio	de vídeo
Audio	0,682	0,630	0,916	0,326
Vídeo	0,386	0,443	0,331	0,923
Media geométrica	0,513	0,528	0,549	0,551

Cuadro 5.4: Bondad de clasificación de información de audio y de vídeo para cada tipo de agrupamiento. La media geométrica de las columnas se muestra en la última fila.

Este resultado muestra numéricamente la idea expresada en trabajos previos sobre las limitaciones de trabajar con un modo (audio o vídeo) únicamente. En particular, el menor rendimiento por parte del agrupamiento por vídeo utilizando matrices de similaridad se puede explicar por el hecho de que se basa en información inicialmente segmentada por audio (ver apartado 4.2.2). Notar que con las agrupaciones basadas en *K-means*, las dos opciones son casi simétricas. De hecho, en este último caso, se obtiene un análisis de varianza (ANOVA) de $F(1, 250) = 0{,}13$, $p = 0{,}7203$, cumpliendo las asunciones de independencia, normalidad e igualdad de varianzas, lo cual indica que no se puede suponer (con estos datos) una diferencia estadísticamente significativa en las medias geométricas obtenidas. En el caso de agrupar mediante matrices de similaridad, la revisión manual del vídeo (debido a las asincronías provocadas por los aparatos de adquisición) provoca unas pequeñas variaciones particulares pero desconocidas, lo cual dificulta su análisis estadístico con el debido rigor. No obstante, se puede ver que la diferencia, aunque significativa, no es sustancialmente grande (menor del 7 %) respecto a la agrupación realizada con *K-means*.

5.1.5.5. Síntesis guiada por fonética

Las imágenes mostradas en la figura 5.14 pertenecen a cuadros de una secuencia audiovisual generada a partir de la especificación de texto. Un conversor de texto a habla

(TTS) transforma este texto en información fonética, y ésta, a su vez, se transforma a visemas mediante el uso de una correspondencia identificador-visema (CIV), construida con la ayuda de un segmentador de alófonos. Se ha probado la generación de secuencias que hablen en diferentes idiomas, como castellano, catalán e inglés, con éxito.

Figura 5.14: Secuencia de imágenes sintéticas creadas a partir de la entrada de texto especificado como cadena de caracteres. En este caso se muestra la síntesis de la palabra *ángel*

La calidad de la síntesis por texto ha sido evaluada en comparación con la conseguida por la síntesis a partir de voz apartado 5.1.5.6, en las pruebas asociadas al realismo (ver apartado 5.3), obteniendo los mejores resultados. Este hecho se debe a que el error de estimación de la información visual de salida a partir del identificador es nulo, ya que viene inequívocamente determinado por la propia CIV.

5.1.5.6. Síntesis guiada por voz

Los métodos utilizados para generar ejemplos de secuencias sintetizadas a partir de voz han sido los basados en teoría de la estimación (ver apartado 4.2.5.2). En la figura 5.15 se muestran ejemplos de buenas y malas estimaciones de apariencias visuales utilizando este tipo de métodos.

Las técnicas de estimación bayesianas (ver apartado 4.2.5.2), en comparación con las técnicas de síntesis dirigida por texto (ver apartado 5.1.5.5), muestran una mayor dificultad a la hora de intentar predecir exactamente la apariencia visual a partir de la información acústica. Una posible justificación a esta mayor dificultad se puede encontrar en el trabajo de Yehia et al. (1998). En dicho trabajo se concluye que la relación entre la información acústica, representada con parámetros derivados de los LSF, y la información visual utilizando un modelo articulatorio es lineal sólo en un 65 % aproximadamente, quedando un 35 % de información relacionada de forma no lineal (si es que lo está plenamente). Realizando un estudio similar al propuesto por Yehia et al. (1998) y utilizando la información auditiva (ver apartado 4.2.3) y la dinámica visual del modelo (ver apartado 2.2.4) descrita en el apartado 4.2.5, se obtiene como resultado que ambas informaciones utilizadas en el presente trabajo de investigación están relacionadas linealmente alrededor de un 33 %. De este dato se puede interpretar que la relación entre la información de audio y la visual, representada mediante el modelo basado en imágenes propuesto, es más complicada que

(a)

(b)

Figura 5.15: Imágenes creadas a partir de la información de habla incluida en una señal de voz. En este caso se está pronunciando la secuencia de alófonos [aɲxel]. En (a) se puede ver un ejemplo de estimación correcta, utilizando la técnica MAP de la estimación bayesiana y en (b), una errónea, realizada mediante la variante MMSE del mismo tipo de estimación.

utilizando un modelo articulatorio como el de Yehia et al. (1998).

5.2. Coste

La reducción de dimensionalidad aportada por el PCA que incorpora el modelo visual (ver apartado 3.2) se muestra como factor clave para reducir el coste de almacenamiento asociado al modelo y el de procesamiento que se encuentra en la síntesis. Además, el esquema incremental propuesto (ver apartado 3.2.4) elimina los problemas de costes computacional y de memoria asociados al cálculo de la SVD para obtener el PCA. En el caso particular de las caras parlantes, se ha ofrecido un método para optimizar la descripción de corpus, reduciendo al máximo el coste de creación de un nuevo modelo visual.

La velocidad de proceso conseguida en el algoritmo combinado de seguimiento y aprendizaje (ver apartado 3.2.4) ha sido, en media, de unos tres segundos por imagen utilizando un procesador Intel® Pentium® IV a 1 GHz con 768 MB de RAM con un código

en MATLAB® no optimizado en tiempo de ejecución. En las aplicaciones presentadas en el apéndice D se llega a una tasa de síntesis de diez imágenes por segundo con resoluciones de 320×240 píxeles. Ésta puede llegar a ser mucho mayor (más de veinticinco imágenes por segundo) si se precarga un conjunto limitado de imágenes y se prescinde de los efectos de coarticulación visual de coste cuadrático con las medidas de la imagen, como se ofrece en la versión interactiva de la aplicación *PREVIS II* (apéndice D.2). En cuanto al uso de memoria, para una resolución de 320×240 píxeles, se usan unos 25MB de RAM permitiendo coarticulación y unos 35MB precargando 12 imágenes (correspondientes a 12 visemas diferentes). El coste más importante es el asociado a la búsqueda de las distancias más cortas en el grafo de coarticulación, que es cúbico respecto a la cantidad de imágenes de la secuencia del corpus. Su impacto se ve reducido gracias a la posibilidad de poder precalcular los caminos de interés, consiguiendo un coste constante en tiempo de síntesis.

En el apartado 5.2.1 se ofrece un análisis detallado del coste y precisión de cálculo obtenidos por el novedoso esquema incremental (propuesto en el apartado 3.2.4), donde se puede observar las ventajas, principalmente en el ahorro de memoria, que lleva asociadas.

5.2.1. Algoritmo de aprendizaje

El objetivo principal de este apartado consiste en analizar la precisión conseguida en la obtención del subespacio y la dinámica visuales a partir del cálculo incremental de la SVD especificada en el apartado 3.2.4 y comprobar la coherencia del coste computacional teórico (ver apartado 3.2.4.1) con el real. Para ello, se ha partido de una selección aleatoria de 34 de los corpus del apartado 5.1.1 y se han realizado diferentes pruebas variando el valor R del bloque de actualización y el rango K de la SVD en todos ellos.

La figura 5.16 muestra el tiempo de ejecución asociado, mientras que el error cometido se encuentra en la figura 5.17. En esta segunda figura se muestran seis gráficas de error, el cometido por la aproximación incremental (5.1) en la subfigura 5.17(a), el valor mínimo teóricamente posible (5.2) en la 5.17(b), el error relativo entre los dos anteriores (5.3) en la subfigura 5.17(c), el límite superior (5.4) (basado en (3.129)) en la 5.17(d) y las diferencias relativas (5.5) y (5.6) entre éste y (5.1) y (5.2) en las subfiguras 5.17(e) y 5.17(f), respectivamente.

$$E^i(R,K) = \frac{\left\| \mathbf{O} - \mathbf{U}^i_{R,K} \mathbf{\Sigma}^i_{R,K} \left(\mathbf{V}^i_{R,K} \right)^T - \bar{\mathbf{m}}^i_{R,K} \right\|_2}{\|\mathbf{O}\|_2} \tag{5.1}$$

$$E^b(K) = \frac{\left\| \mathbf{O} - \mathbf{U}^b_K \mathbf{\Sigma}^b_K \left(\mathbf{V}^b_K \right)^T - \bar{\mathbf{m}}^b_K \right\|_2}{\|\mathbf{O}\|_2} \tag{5.2}$$

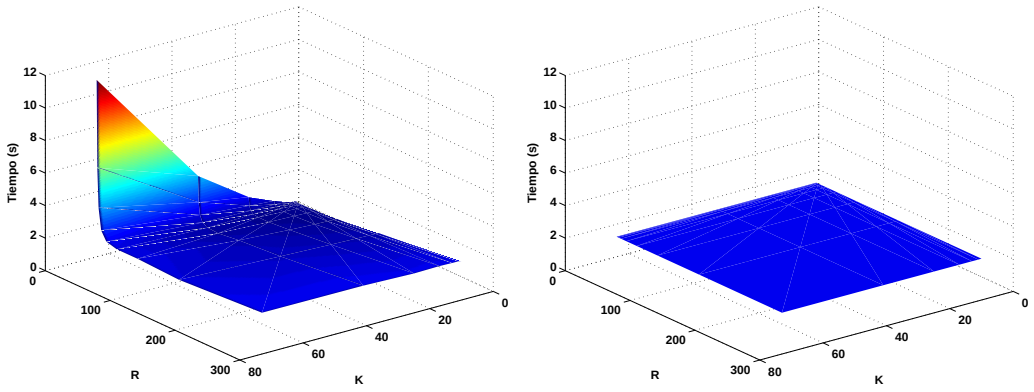
$$E^r(R,K) = \frac{E^i(R,K) - E^b(K)}{E^b(K)} \tag{5.3}$$

$$E^c(R,K) = \frac{E_{umbral}(R,K)}{\|\mathbf{O}\|_2} \tag{5.4}$$

$$E^{ri}\left(R,K\right)=\frac{E^{c}\left(R,K\right)-E^{i}\left(R,K\right)}{E^{i}\left(R,K\right)} \tag{5.5}$$

$$E^{rb}\left(R,K\right)=\frac{E^{c}\left(R,K\right)-E^{b}\left(K\right)}{E^{b}\left(K\right)} \tag{5.6}$$

donde $\mathbf{U}^{i}_{R,K}\mathbf{\Sigma}^{i}_{R,K}(\mathbf{V}^{i}_{R,K})^{T}+\bar{\mathbf{m}}^{i}_{R,K}$ es la descomposición obtenida por el algoritmo de cálculo incremental de la SVD, la expresión $\mathbf{U}^{b}_{K}\mathbf{\Sigma}_{b}(\mathbf{V}^{b}_{K})^{T}+\bar{\mathbf{m}}^{b}_{K}$ es la teóricamente correcta (y, por tanto, asociada a un error mínimo para un rango k concreto), y $E_{umbral}(R,K)$ es la cota superior del error (ver apartado 3.2.4.1) de cada caso concreto.



(a) (b)

Figura 5.16: Carga computacional para diferentes valores del bloque de actualización y rango de la matriz aproximadora (igual al número de valores singulares contemplados). (a) coste computacional de la SVD incremental; (b) coste computacional del cálculo de la SVD teórica. Este último caso es constante dado que se calcula el rango completo y después se descartan las últimas columnas que procedan según el rango seleccionado.

Por un lado, el tiempo de ejecución es proporcional a los costes del cuadro 3.9 del apartado 3.2.4.1. Éste es máximo para valores pequeños de R y elevados de K y mínimo cuando R y K son pequeños y similares. Cuando R es grande, el valor de K no tiene efecto; de hecho, en las pruebas realizadas no se revela estadísticamente significativo para órdenes de magnitud similares de R y N (este último valor, el número de imágenes de las secuencias). Por otro lado, el error cometido en la aproximación tiene un comportamiento muy parecido al límite inferior posible. No obstante, si se observa la gráfica del error relativo se descubre que éste es grande para valores pequeños del bloque de actualización (excepto para valores de rango pequeños). Estas observaciones se pueden resumir en tres puntos:

1. Mantener un error relativo y absoluto reducidos es computacionalmente menos costoso para una aproximación de rango (K) bajo con un bloque de actualización pequeño (R) también pequeño.

2. Tomando como referencia el punto 1, aumentar el rango de la aproximación o el bloque de actualización provoca un incremento en la carga computacional con una reducción leve de ambos errores. Si el aumento es igual para ambos valores, el incremento de la carga es el más reducido posible.

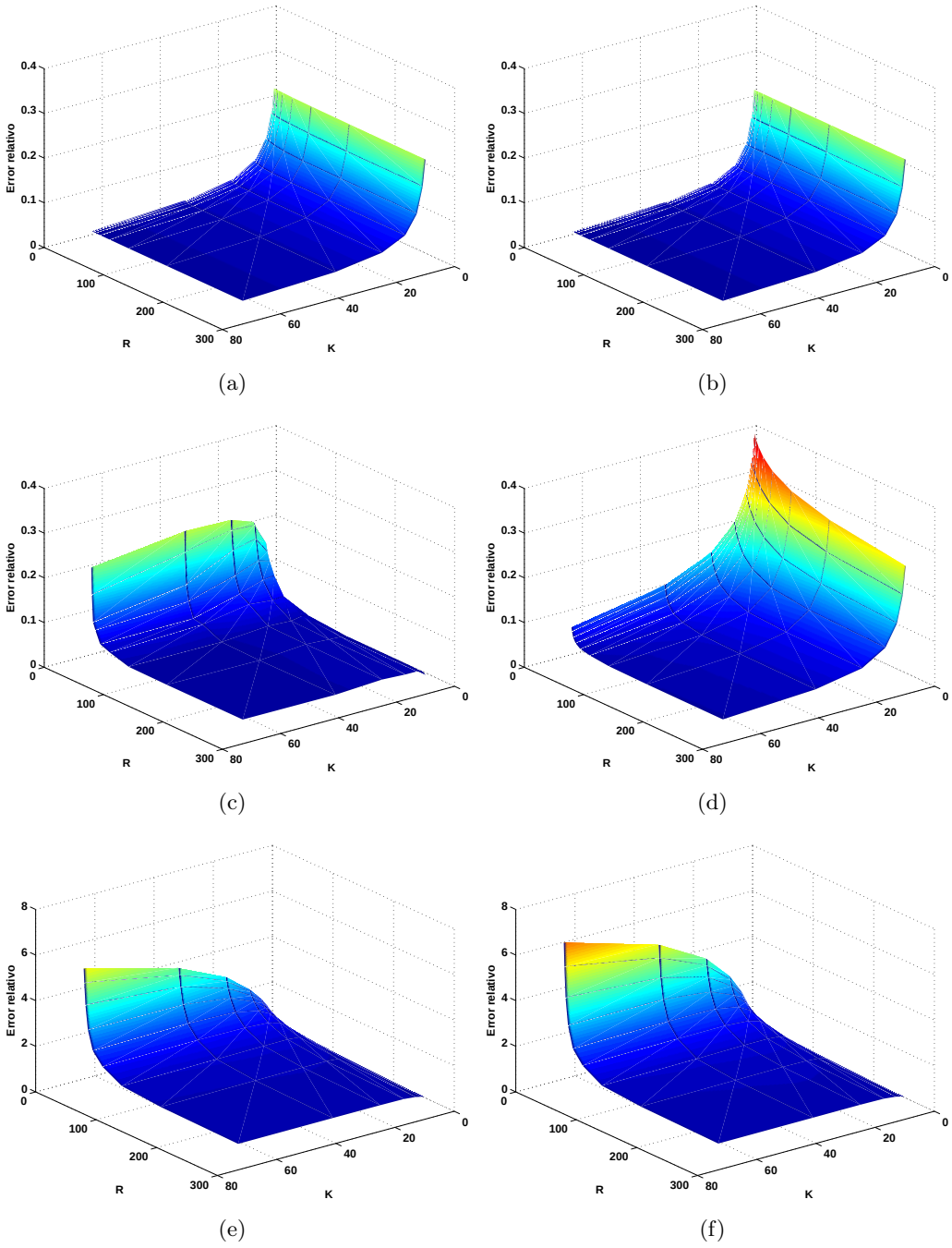


Figura 5.17: Precisión conseguida en el cálculo de la SVD con extracción de la media: (a) utilizando el algoritmo propuesto de SVD incremental con actualización de la media; (b) mediante la extracción de la media y posterior cálculo de la SVD de la matriz de diferencias (que supone una cota inferior de (a)). La medida de error relativo entre (a) y (b) se muestra en (c). En (d) se ofrece la cota superior del error, así como la diferencia relativa entre esta cota y el error real cometido mediante la aproximación incremental en (e) y sin ella (f).

3. Cuanto mayor es el tamaño del bloque de actualización (R), más se parecen los resultados del cálculo incremental y no incremental así como sus cargas computacionales.

Respecto a la cota del error, se puede ver que se vuelve más ajustada al aumentar R, principalmente. El aumento del rango K provoca también un ligero aumento de su precisión.

Seguidamente se procede a comparar, utilizando la misma máquina, la carga computacional y el uso de memoria en el cálculo de la SVD incremental y la directa para valores bajos de K y R, concretamente $K = R = 10$. En la figura 5.18 se puede observar una gráfica que relaciona ambos comportamientos.

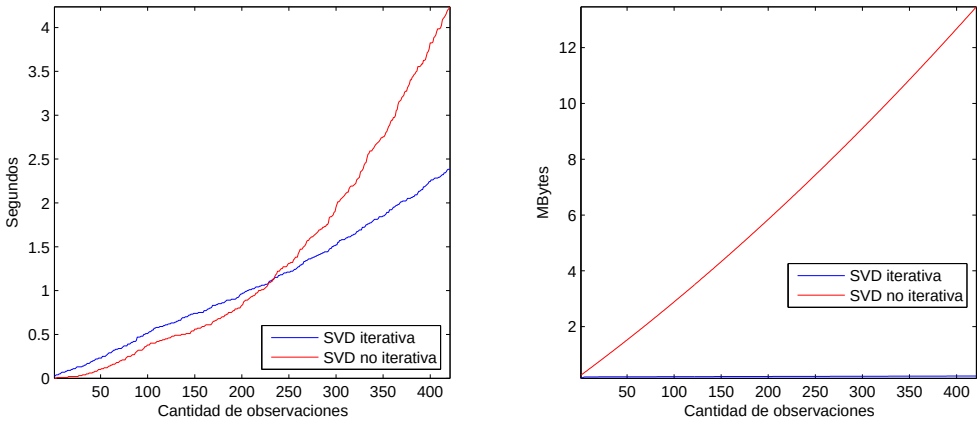


Figura 5.18: Carga computacional (a) y requisitos de memoria (b) asociados al cómputo de SVDs variando el número de observaciones o columnas de la matriz de entrada. En línea azul aparece el comportamiento del algoritmo iterativo de cálculo de la SVD con actualización de la media y en color rojo, el del no iterativo. Notar la gran diferencia de consumo de memoria conforme aumenta el número de observaciones.

A partir de un determinado número de imágenes (233 en el caso de la figura 5.18), el cálculo directo de la SVD es más lento que el incremental. El consumo de memoria es el resultado que más llama la atención: la versión directa presenta un consumo de memoria de un orden muy superior al de la versión incremental; mientras que, para 421 imágenes, la incremental consume alrededor de 0,23 MBytes, la directa necesita 13,46. El error asociado a la versión incremental ($E^i\,(10, 10)$) es de 0,0544 y el de la versión no incremental ($E^b\,(10)$) es de 0,0518. Finalmente, notar que los datos asociados a la gráfica de consumo de memoria revelan un crecimiento de 0,078 KBytes por imagen en la versión incremental frente a los 39,25 KBytes por imagen en la directa.

5.2.2. Tiempo real

Se ha posibilitado el análisis en tiempo real gracias a dotar de causalidad al método de análisis de corpus (ver apartados 3.1.1.2 y 3.2.1.4). Ligado también a la flexibilidad, fiabilidad y facilidad de uso es posible efectuar este seguimiento sin el conocimiento de la apariencia del objeto a localizar. Cambiando únicamente el origen de los datos, desde un

archivo a una fuente de vídeo, y manteniendo el resto del algoritmo intacto se ha conseguido efectuar el seguimiento y aprendizaje simultáneo de objetos inicialmente desconocidos. La construcción incremental de los subespacios de apariencia se ha mostrado de gran utilidad para conseguir este comportamiento. El seguimiento conseguido ha sido de unas diez imágenes por segundo sobre código MATLAB® y algunas imágenes se pueden observar en la figura 5.19.

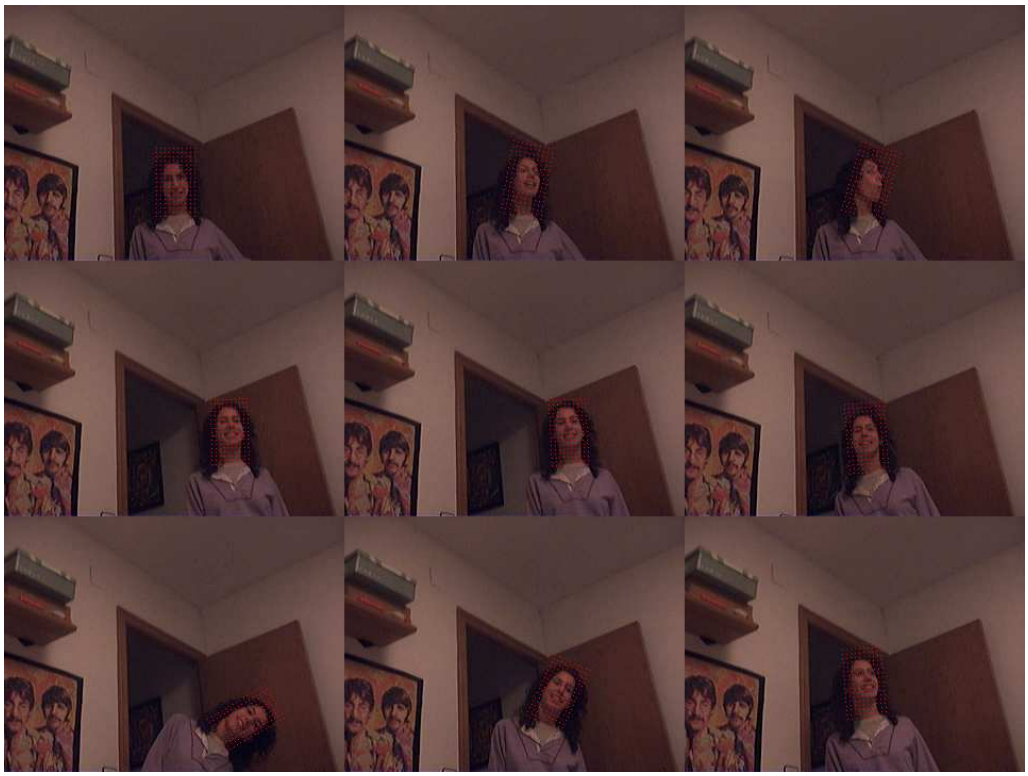


Figura 5.19: Seguimiento de una cara en tiempo real. La región localizada se representa por medio de una nube de puntos rojos.

No obstante, el algoritmo propuesto para la selección de unidades visuales (ver apartado 4.1.3.2) es no causal, ya que necesita conocer el conjunto de unidades visuales futuras. Sin embargo, dado que sólo se aplica una profundidad de dos niveles, se puede transformar en causal retardando la generación de salida una cantidad igual a la duración asociada a una unidad visual.

5.3. Realismo de la síntesis

Una primera forma de evaluar el realismo conseguido por el esquema de síntesis propuesta puede ser mediante una medida de diferencia entre la referencia ideal y el resultado real, utilizando, por ejemplo, técnicas basadas en relación entre el pico y ruido de la señal (PSNR) o error cuadrático medio (MSE). Sin embargo, dado que los objetos a sintetizar tienen como propósito ser evaluados como reales por personas, estas técnicas

no son de utilidad ya que no están correlacionadas con la percepción humana (Li et al., 2004). Este hecho implica una dificultad añadida en la evaluación de la naturalidad de los resultados, ya que no existen métodos numéricos asociados. Un ejemplo de este hecho se puede observar en la figura 5.20, donde, comparando con una cara real, una cara imposible tiene un MSE de 34,61 y una producida por el modelo visual propuesto posee un MSE de 118,07. En este caso, la percepción y la medida objetiva dan respuestas contrarias.

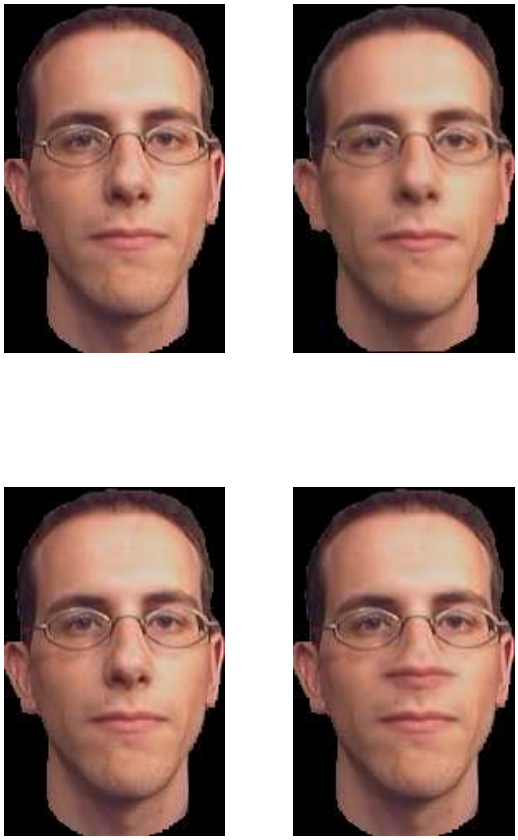


Figura 5.20: Comparación entre una cara real, una sintética y otra imposible. La cara real se encuentra repetida en las dos filas de la columna izquierda para facilitar la comparación. La cara sintética se encuentra en la posición superior derecha y la imposible está situada en la posición inferior derecha. El MSE entre la cara real y la sintética es de 118,07, mientras que el MSE entre la real y la imposible es mucho menor, concretamente de 34,61.

De este modo, la evaluación del realismo ha de realizarse mediante medidas dadas por la percepción humana. Se ha optado por recoger las impresiones de un conjunto de personas frente a distintas demostraciones que muestran las capacidades del esquema de síntesis presentado, para poder evaluar la calidad obtenida en cuanto a realismo. Además, se ha evaluado si la inclusión del esquema de selección de unidades visuales reales y los efectos de suavizado en las transiciones entre regiones obtienen alguna mejora significativa.

Las pruebas de realismo se han efectuado sobre imágenes faciales, debido a la focalización que posee el trabajo sobre este tipo de objetos y a que el ser humano está

especializado en reconocerlos, con lo que su capacidad crítica es mayor que con cualquier otro tipo de imágenes.

Se han realizado pruebas que evalúan la capacidad de representación estática y dinámica del modelo, es decir, su foto realismo y su vídeo realismo, respectivamente. Además, se han comparado los resultados obtenidos al guiar la síntesis de una cara parlante por fonética y por voz, ésta última mediante dos métodos basados en la estimación bayesiana, el MAP y el MMSE. Se ha escogido la resolución de 320×240, ya que se ha considerado un tamaño ampliamente extendido en el ámbito doméstico.

5.3.1. Evaluación del foto realismo

Se han realizado dos pruebas para evaluar esta característica. La primera se basa en sintetizar secuencias audiovisuales siguiendo la misma dinámica definida en el corpus del cual son originales y compararlas con éstas. La segunda se basa en sintetizar imágenes estáticas individuales nuevas, comparándolas con sus versiones reales más parecidas y con otras sintéticas sin los efectos de suavizado (ver apartado 4.1.2.2), estas últimas para evaluar el incremento de realismo que ofrecen estos efectos.

5.3.1.1. Indistinguibilidad respecto a la versión real

Se ofrecieron ocho secuencias de dos segundos de duración cada una, mostradas aleatoriamente y por duplicado a cada uno de los 32 evaluadores asociados a la prueba, escogidos entre una población de entre 21 y 45 años. Cada secuencia constaba de una cara, que podía ser original o sintética, habiendo un 50 % de versiones reales y otro 50 % de sintéticas. El evaluador debía intentar determinar si cada una de ellas era real, sintética o indistinguible. No se permitió repetir la observación de ninguna secuencia. El resultado de esta prueba se resume en un 30 % de respuestas indistinguibles y un 48 % de respuestas correctas. Hay que destacar también el alto grado de inconsistencia que tuvieron los evaluadores en las repeticiones de las secuencias. Se considera inconsistencia al hecho de que un evaluador tenga preferencias diferentes sobre una misma evaluación que se repite a lo largo del test. El hecho de que haya habido un alto número de evaluadores que una vez escogen la versión real y otra vez, la sintética, da a entender un grado de indistinguibilidad aún más alto, ya que estos evaluadores creen que están escogiendo la correcta en cada momento (porque si dudasen, escogerían la opción de indistinguibilidad).

5.3.1.2. Puntuación media diferencial

En esta prueba se han mostrado seis imágenes estáticas entre reales y sintéticas de forma aleatoria y repetida a 94 personas con edades comprendidas entre 14 y 56 años y una representación equitativa de ambos sexos. Las imágenes sintéticas podían contener los efectos de suavizado entre regiones o no. En este caso se pedía una puntuación del grado de realismo de las fotografías entre uno (mínimo) y cinco (máximo), obteniendo un MOS al

final de la prueba para cada imagen. No obstante, dado que no se ofrecía ninguna referencia de máximo ni mínimo en este caso, se optó por dar una puntuación diferencial media de opinión (DMOS) con respecto a la imagen original para cada evaluador. El DMOS entre la imagen original y la sintética con efectos de suavizado fue de $0{,}37 \pm 0{,}16$, mientras que con ausencia de efectos de suavizado fue de $0{,}65 \pm 0{,}30$. El test de Kolmogorov-Smirnov para diferentes conjuntos mostrales y el de Kruskal-Wallis[2] revelan diferencias estadísticamente significativas (en un 95 %) (Sheskin, 2007) entre la puntuación real y la de las imágenes sintéticas sin efectos de suavizado. No obstante, en el caso de comparar las reales con las suavizadas, se obtiene un grado de significación menor del 90 %. Por referencia, los vídeos originales fueron evaluados con una puntuación de $3{,}52 \pm 0{,}32$.

5.3.1.3. Discusión

El foto realismo conseguido por el método de síntesis propuesto se considera muy alto dado el elevado nivel de indistinguibilidad, aciertos e inconsistencias de la primera prueba y la similaridad estadística con los resultados obtenidos para las imágenes reales. Esta valoración da a entender que las imágenes producidas por el modelo visual se encuentran dentro (o lo suficientemente cercanas) del subespacio de apariencia real del objeto, o lo que es lo mismo, que el modelo visual contiene una representación adecuada del mismo. Destacar también que los efectos de suavizado en las transiciones entre regiones se revela como un aspecto que mejora la calidad de la síntesis.

5.3.2. Evaluación del vídeo realismo

Se han realizado dos pruebas para evaluar esta cualidad. La primera compara el resultado de la síntesis con transiciones reales, mientras que la segunda se basa en estudiar la calidad vídeo realista conseguida y evaluar las mejoras conseguidas por el algoritmo de seleccion (ver apartado 4.1.3.2).

5.3.2.1. Indistinguibilidad respecto a la versión real

Esta prueba buscaba evaluar el nivel de vídeo realismo conseguido al sintetizar expresiones mediante la comparación entre una cara real y otra sintética realizando los mismos movimientos. Parte de la secuencia se puede observar en la figura 5.11. Cada evaluador observó seis secuencias una única vez cada una. En cada una de ellas se mostró un par de caras realizando un cambio de expresión facial. Una de ellas era real y la otra, sintética. El evaluador debía intentar determinar cuál era la cara real o si era indistinguible. De forma parecida a la primera prueba de foto realismo (ver apartado 5.3.1), los evaluadores fueron altamente inconsistentes en sus respuestas y se obtuvo tan sólo un 52 % de aciertos y un 26 % de indecisiones. En este caso la inconsistencia se puede interpretar del mismo

[2]Tipo de ANOVA no paramétrico que se utiliza con distribuciones similares, sin los requisitos de normalidad de la versión paramétrica (Devore, 2000).

modo que en el apartado 5.3.1.1. Esta prueba se realizó sobre 317 personas de entre 12 y 63 años (con equidad entre sexos) en *La Pedrera* de Barcelona en el año 2004 durante el acontecimiento de *Vive la Ciencia Contemporánea*.

5.3.2.2. Puntuación media diferencial

Realizada sobre la misma población que en el tercer test de foto realismo (ver apartado 5.3.1), esta prueba ha incluído la muestra de seis vídeos de personas hablando, reales y sintéticos, incluyendo los efectos de selección de unidades visuales reales (ver apartado 4.1.3.2) en la mitad de estos últimos. Los vídeos se han mostrado repetidos para medir la consistencia de los evaluadores. En este caso también se obtuvo un DMOS entre secuencias originales y sintéticas, ya que no se ofrecía ningún conjunto de vídeos de referencia para no influir a los evaluadores. El DMOS entre los vídeos originales y los sintéticos sin selección fue de $1{,}81 \pm 0{,}28$, mientras que con selección fue de $0{,}48 \pm 0{,}21$. Utilizando la prueba de Kolmogorov-Smirnov para dos cojuntos mostrales distintos y el contraste de Kruskal-Wallis, se tiene que la puntuación obtenida por las evaluaciones con los vídeos sintéticos están por debajo de la de los vídeos originales con una significación estadística (Sheskin, 2007) del 99 % en el caso de sin selección y del 95 % en el caso de con selección. Para tener una referencia de puntuaciones absolutas, la de las secuencias reales fue de de $3{,}63 \pm 0{,}28$.

5.3.2.3. Discusión

El nivel de vídeo realismo logrado por el esquema de síntesis propuesto ha sido también elevado dados los elevados índices de indistiguibilidad e inconsistencia y el porcentaje de aciertos de la primera prueba, así como la reducida diferencia de puntuación (alrededor de un 10 % menor) obtenida respecto a la versión real. Esta última medida revela una cierta distinguibilidad estadística muy ajustada, lo cual da a entender que el foto realismo conseguido parece ser de mayor importancia que el vídeo realismo, aunque sin desmejorar este último. Destacar, además, la mejora sustancial que aporta el algoritmo de selección (ver apartado 4.1.3.2) sobre la síntesis de secuencias.

El elevado vídeo realismo obtenido significa que el modo de representar las transiciones visuales, mediante el método de selección (apartado 4.1.3.2) y el algoritmo de interpolación (apartado 4.1.3.1) propuestos, consigue emular las trayectorias reales de éstas sobre el subespacio visual almacenado en el modelo.

5.3.3. Evaluación de la síntesis por voz

En este caso se ha evaluado la percepción de la síntesis por voz, sobretodo su componente vídeo realista, comparándola con la síntesis por fonética, cuyos resultados se han mostrado en el subapartado anterior. Se realizaron dos pruebas, una de preferencia entre

síntesis guiada por fonética y voz y otra de valoración global de la calidad de la síntesis guiada por voz, con los métodos MAP y MMSE.

5.3.3.1. Comparación con síntesis por fonética

Esta prueba ha sido realizada sobre la misma población que la primera asociada al foto realismo. Se han comparado tres métodos de síntesis: el dirigido por fonética, el dirigido por voz y basado en un estimador MAP y otro basado en un estimador MMSE (ver apartado 4.2.5.2). Se construyeron diversas secuencias mostrando dos caras en cada una de ellas. En cada secuencia se sintetizaba la misma pronunciación pero con dos métodos diferentes, manteniendo una distribución uniforme en términos de combinación de técnicas: síntesis guiada por fonética (SGPF)-MAP, SGPF-MMSE y MAP-MMSE. Una vez construidas todas las secuencias, se mostraron ocho de éstas a los evaluadores (con repeticiones para descartar evaluadores inconsistentes), garantizando la aparición de, como mínimo, dos repeticiones de cada combinación. Cada evaluador pudo ver las secuencias tantas veces como desease hasta poder decidir cuál era la opción que veía más realista en cada una de ellas (ya que sólo existían opciones de respuesta de derecha o izquierda). Un 98 % de las evaluaciones consistentes eligieron como mejor técnica a la síntesis guiada por fonética frente a las otras dos. La comparación entre las técnicas de MAP y MMSE dió como resultado una preferencia del 83 % del primero frente al segundo.

5.3.3.2. Comparación entre MMSE y MAP

Esta segunda prueba tomó las mismas secuencias que la primera pero pidió a los evaluadores una puntuación sobre el realismo de la secuencia entre uno (mínimo realismo) y cinco (máximo). En este caso se presentaban parejas de secuencias donde la de la derecha siempre era la real (hecho conocido por los evaluadores) y la de la izquierda era una sintetizada mediante una de las tres técnicas presentes en la primera prueba. Al igual que en la segunda prueba de foto realismo (ver apartado 5.3.1), se pidió al usuario que si no era capaz de distinguir ambas imágenes, diese una puntuación máxima y sino, una puntuación mínima a la secuencia (o secuencias) que creía que se veía peor. Los resultados de puntuación revelaron un $3{,}47 \pm 0{,}24$ para la SGPF, $2{,}25 \pm 0{,}31$ para el MAP y $1{,}38 \pm 0{,}17$ para el MMSE. Utilizando el test de Kolmogorov-Smirnov para diferentes conjuntos mostrales, se puede afirmar que el SGPF obtiene una puntuación significativamente superior al MAP, el cual obtiene, a su vez, una puntuación significativamente superior al MMSE, ambas con un grado de significación estadística mayor del 99 % (Sheskin, 2007). Se puede observar que estos resultados se encuentran totalmente correlacionados con los obtenidos en la primera prueba.

5.3.3.3. Discusión

Estas pruebas han revelado la inherente dificultad que tienen los métodos de síntesis por voz para obtener un nivel de realismo equivalente a los guiados por fonética. El

algoritmo subyacente es el mismo y la única diferencia está en la determinación del orden y la identidad de las apariencias clave que van guiando la secuencia sintética. Así, el menor realismo obtenido por la síntesis guiada por voz se achaca a una identificación incorrecta de las apariencias clave, la cual viene plenamente determinada en el caso de la síntesis guiada por fonética. Particularmente, se cree que el peor rendimiento del método de MMSE viene dado principalmente por el hecho de que no tiene en cuenta la dinámica visual del objeto.

5.3.4. Interpretación

Las pruebas de foto y vídeo realismo revelan diferencias leves respecto a las secuencias originales (no significativas respecto al foto realismo) al utilizar el algoritmo descrito de síntesis. Por otro lado, la síntesis guiada por fonética ofrece resultados mucho mejores que la guiada por voz, lo cual tiene sentido según lo explicitado en el apartado 4.2.5.

Capítulo 6

Conclusiones

Las conclusiones sobre el trabajo se presentan organizadas en cinco apartados. Un primero sobre el marco propuesto en general, (apartado 6.1), otros tres, que analizan un poco más en detalle el modelo visual presentado (apartado 6.2) y las partes de análisis (apartado 6.3) y síntesis expuestas (apartado 6.4) y un último apartado que expone una propuesta de varias líneas de futuro que se pueden extraer a partir de los hechos, conceptos y procedimientos desarrollados en estas páginas.

6.1. Generales

Se ha propuesto un marco único para la creación o síntesis de secuencias realistas de caras parlantes y lengua de signos a partir de la especificación de un conjunto ordenado de símbolos, como puede ser el texto. El primer caso puede ir acompañado de voz, obteniendo secuencias audiovisuales totalmente sincronizadas y, además, puede utilizarse ésta para generarlas, aunque a costa de obtener una menor sincronía. La información necesaria para el proceso de síntesis se encuentra almacenada de forma óptima en lo que se conoce en este trabajo como modelo visual. Su obtención se realiza a través del resumen o análisis de una corta secuencia de ejemplo que contenga toda la variabilidad que se desee reproducir posteriormente.

El esquema propuesto es de uso sencillo ya que no es intrusivo, es capaz de funcionar con transductores domésticos, no requiere restricciones de iluminación y la intervención manual necesaria es nula en la síntesis y muy reducida en el análisis, llegando a ser evitable con el uso de aparatos profesionales de captura. También ofrece una elevada flexibilidad, ya que es aplicable a todo tipo de objetos que se mantengan dentro de la región observada en la secuencia de ejemplo. Asimismo, el análisis es altamente fiable, ya que ha sido capaz de construir el modelo a partir de más de un centenar de secuencias, y la síntesis ha mostrado su elevado realismo en las evaluaciones realizadas. Todas estas características facilitan la creación de secuencias personalizadas de objetos arbitrarios a elegir por el potencial usuario del marco descrito.

En el ámbito particular de las cabezas parlantes, se ha propuesto una nueva manera de definir el énfasis visual mediante estadísticos de orden, se ha ofrecido un sistema de elección objetivo y personalizado de visemas y se han cuantificado los efectos de incertidumbre audiovisual entre tres sujetos. Además, se ha proporcionado un nuevo marco común para la partición, la composición y el cálculo incremental y decremental de la descomposición en valores singulares (SVD) teniendo en cuenta la información media.

En cuanto a los costes de memoria y tiempo conseguidos se ha podido aplicar el análisis a secuencias muy largas (más de 12 minutos a 25 imágenes por segundo) con un consumo muy reducido de memoria (50 MB); el tiempo de cálculo ha sido alrededor de tres segundos por imagen (con una resolución de 320 por 240) en un Pentium IV a un GHz. Se han desarrollado librerías en lenguaje *C* y en *MATLAB* que implementan las principales técnicas presentes en el trabajo, así como diferentes aplicaciones, tanto por línea de comandos como interactivas, pasando por diversas aplicaciones a través de internet. Estas aplicaciones muestran la aplicabilidad de las técnicas expuestas en este trabajo de investigación, aportándole una visión de utilidad real para con la sociedad. Gracias a ellas se han producido numerosas aportaciones en numerosos proyectos, tanto públicos como privados, y ha obtenido un importante impacto en la sociedad a través de los diferentes medios de comunicación y eventos culturales, científicos y sociales en los que ha participado.

6.2. El modelo

El modelo propuesto contiene información acerca de la apariencia del objeto representado, así como un muestreo no uniforme de todas sus posibles configuraciones, que se conoce en este trabajo como la dinámica del objeto. La información relativa a la sincronía audiovisual (si existe información de audio asociada) se encuentra almacenada en el modelo acústico, que forma parte del modelo visual, y permite establecer relaciones entre información fonética y/o auditiva con la dinámica del objeto.

El modelo desarrollado es compacto, modular y acotado, gracias al uso de la reducción de dimensionalidad aportada por las técnicas de análisis de componentes principales (PCA) utilizadas y la segmentación del objeto en regiones independientes. Estas características han contribuido a la fiabilidad, flexibilidad y coste del sistema (ver apartados 2.3 y 3.3.1), así como en el realismo del mismo, tal y como se puede ver en el apartado 4.3.1

6.3. Análisis

El análisis propuesto en este trabajo construye un modelo a partir de un corpus dado con una mínima (o nula, utilizando transductores de alta calidad) interacción humana y sin ningún tipo de marcadores ni elementos intrusivos. La construcción del modelo se realiza mediante un algoritmo de seguimiento y aprendizaje simultáneos, aportando causalidad al proceso. Además, este algoritmo se puede aplicar a cualquier tipo de objeto. Su única limitación consiste en que la región que contiene el objeto no puede cambiar de forma.

Se ha propuesto una descripción para grabar corpus dedicados a generar modelos visuales. Utilizando algoritmos genéticos paralelos, se ha definido una herramienta para el diseño de descripciones de corpus lo más pequeños posibles y con una distribución equitativa de los sonidos de interés, de manera que se puedan grabar corpus para la creación de modelos visuales que hablen en cualquier idioma. El cuadro 2.8 muestra un resultado ofrecido por esta herramienta para los sonidos castellanos principales.

La grabación del corpus no necesita seguir un gran número de restricciones, y la más fuerte tiene que ver con los movimientos del objeto, ya que es deseable que no contengan rotaciones laterales ni verticales. Este hecho es debido a que estos movimientos son interpretados como cambios de apariencia por el modelo. El sonido y el vídeo se pueden grabar con una simple cámara web doméstica. No obstante, y sólo para realizar la investigación de los métodos de predicción de imagen a partir de voz, se han realizado varias grabaciones en un estudio profesional de grabación para tener una fuente de datos acústicas con la mejor relación señal-ruido posible.

El proceso de seguimiento se ha planteado como un problema de optimización, que se resuelve de forma iterativa debido a la inclusión de una linealización para simplificarlo. Este proceso se ha mostrado extensible y se han ofrecido diferentes posibilidades de ampliación, de las cuales se han escogido la multiresolución y los subespacios modulares como referencia por motivos de velocidad, fiabilidad y consumo de recursos. La robustez no se ha utilizado, ya que aumentaba el coste computacional y no aportaba ningún beneficio adicional sobre el comportamiento obtenido en el esquema propuesto.

El aprendizaje del modelo se ha realizado de forma dinámica gracias al desarrollo de un novedoso método de cálculo de la SVD incremental con actualización de la información media. Se ha ofrecido un estudio del rendimiento teórico de esta nueva propuesta, acompañándolo de un estudio del error de aproximación cometido, comparándolo con cálculos reales y comprobando su perfecta concordancia. Destacar que se ha introducido también el concepto de SVD decremental, que ofrece unas posibilidades de *olvido* de los datos que puede ser muy útil en futuros trabajos.

El esquema de análisis propuesto es fiable, flexible, muy fácil de utilizar y tiene un consumo de recursos relativamente bajo comparado con otros sistemas que se basan en características similares pero necesitan el corpus completo para operar (Ezzat et al., 2002, Cosatto, 2002, Cosker et al., 2004). Además, la causalidad conseguida por la simultaneidad de los procesos de seguimiento y aprendizaje permite el comportamiento en tiempo real, según la máquina en que se implemente, la resolución requerida y la habilidad del programador.

6.4. Síntesis

El proceso de síntesis se ha presentado como la creación de secuencias audiovisuales a partir de un modelo visual y guiada por la conversión de información fonética o sonora en información visual.

La creación de secuencias de imágenes se ha utilizado principalmente con objetos del tipo facial y manual debido a la motivación principal del trabajo. En el primer caso, se ha utilizado para reproducir los movimientos de los labios de una persona, aunque también se ha mostrado su uso en generación de gestos y cambios de expresividad. Además, se ha presentado un nuevo método de interpolación no lineal de alta dimensionalidad (apartado 4.1.3.1) para mejorar el nivel de vídeo realismo conseguido en la creación de secuencias, así como un algoritmo de selección de visemas con control del énfasis visual en la síntesis. Las evaluaciones realizadas han revelado un alto realismo en la síntesis y una elevada correlación entre la especificación y la percepción del nivel de énfasis visualenfatización visual en una secuencia. En el caso de caras parlantes, el nivel de énfasis en la zona de la boca afecta a la percepción del realismo obtenido cuando se transmite un mensaje en diferentes estilos de locución, revelándose así como un efecto a tener en cuenta en la síntesis.

La calidad conseguida en la síntesis está directamente relacionada con la representación del objeto de interés por el modelo. Éste, a su vez, depende de la precisión contemplada en la etapa de análisis. De hecho, al ser la síntesis el proceso inverso del análisis (los subprocesos son funciones inversas, como la función vec^{-1}), aquello que ignore el análisis, nunca podrá ser obtenido por la síntesis.

En el caso de existir una sincronía audiovisual, como en el caso de las caras parlantes, la síntesis puede ir guiada por dos tipos de información: fonética o voz. La segunda ofrece peores resultados que la primera ya que la relación entre la voz y la imagen es mucho más complicada que en el caso de utilizar información fonética. La estimación por voz se ha llevado a cabo mediante técnicas de estimación bayesiana y se ha aprovechado este marco para desarrollar un método de cuantificación de los efectos de incertidumbre audiovisual, los cuales dicen que dado un agrupamiento de datos audiovisuales, cuanto mejor separados quedan en un modo, peor quedan en el otro. Los resultados obtenidos revelan una constancia del valor de incertidumbre audiovisual a través de las agrupaciones probadas, así como a través de las diferentes apariencias faciales utilizadas.

El esquema de síntesis propuesto es flexible, muy fácil de utilizar y obtiene resultados realistas con un consumo de memoria y tiempo proporcional a la resolución de la imagen.

6.5. Líneas de futuro

El presente trabajo de investigación se encuentra dentro de la punta del iceberg de una temática muy joven a la que cada día se le encuentran más interrogantes por contestar. En este trabajo se ha mostrado que es posible recrear secuencias visuales de aspecto realista mediante un proceso de interpolación no lineal por selección de unidades visuales a partir de la información resumida de una corta secuencia de ejemplo. No obstante, y a partir de los conceptos y procedimientos desarrollados, surgen nuevas preguntas que incitan a trabajar para obtener las correspondientes respuestas:

- El proceso de interpolación planteado es independiente de la fuente de datos, con lo que se puede pensar en aplicarlo a otros tipos, como la voz. ¿Qué resultados ofrecería?

- Siguiendo la reflexión anterior, también se puede pensar en la aplicación del modelo visual y del proceso de síntesis propuestos para animar caras parlantes tridimensionales. ¿Qué se obtendría del uso de la información tridimensional de los vértices del modelo 3D con el esquema propuesto? ¿Se podría pensar en incorporar como texturas el propio resultado de este trabajo sobre dicho modelo?

- Se ha planteado una solución para deletrear palabras con las manos que se ha clasificado como una mezcla de dos modelos existentes en la bibliografía. ¿Se puede interpretar como un nuevo modelo para este tipo de lenguaje? Se puede intentar extenderlo al resto del cuerpo o incluso adaptar partes del algoritmo de interpolación al tipo de modelos basados en captura de movimiento, tomando como observaciones las coordenadas tridimensionales de los brazos, cuerpo y cabeza.

- Si se logra una implementación del análisis a suficiente velocidad se obtiene la definición de la apariencia de objetos que no se conocían a tiempo real. ¿Puede facilitar algún tipo de decisiones en aplicaciones que se basan en la observación del entorno? ¿La inclusión del olvido puede ayudar en esta tarea?

- Sobre el control del énfasis visual obtenido, ¿se puede pensar en extenderlo sobre otros elementos faciales y adjuntarle información prosódica del habla para obtener un control de la acentuación visual, asociada a la comunicación no verbal?

- La SVD partida y composicional se pueden ver como técnicas que agrupan y separan regiones de la imagen que evolucionan a lo largo del tiempo. ¿Puede ser de utilidad para mejorar el actual estado de la técnica en segmentación, al incluir implícitamente la evolución temporal de las regiones en baja dimensionalidad?

- La incertidumbre audiovisual, ¿revela algún aspecto más profundo? ¿O simplemente constata el hecho conocido de que "*el todo es más que la unión de las partes*" en la comunicación interpersonal humana?

Estas preguntas representan un ejemplo de las diferentes vías que se pueden desprender de la presente tesis doctoral para permitir el inicio o la continuación de la actividad investigadora en el ámbito de las tecnologías de la información.

Bibliografía

Agelfors, E., Beskow, J., Granström, B., .Lundeberg, M, Salvi, G., Spens, K.E. y Öhman, T. (1999). "Synthetic Visual Speech Driven from Auditory Speech". En: *Proceedings of Auditory Visual Speech Processing*, pp. 123–127. Santa Cruz, CA, EUA.

Allen, J., Hunnicutt, M.S., Klatt, D.H., A., R.C. y Pisoni, D.B. (1987). *From text to speech: the MITalk system.* Cambridge University Press, New York, NY, USA. ISBN 0-521-30641-8.

Alonso, F., de Antonio, A., Fuertes, J.L. y Montes, C. (1995). "Teaching Communication Skills to Hearing-Impaired Children". *IEEE MultiMedia*, **2(4)**, pp. 55–67. ISSN 1070-986X.

Ananova Ltd. (2007). "Ananova".
`http://www.ananova.com`

André, E. (2000). "The Generation of Multimedia Presentations". En: R. Dale, H. Moisl y H. Somers (Eds.), *A Handbook of Natural Language Processing: techniques and applications for the processing of language as text*, pp. 305–327. Marcel Dekker Inc.

André, E. (2003). "Natural Language in Multimedia/Multimodal Systems". En: R. Mitkov (Ed.), *Handbook of Computational Linguistics*, pp. 650–669. Oxford University Press.

Anthropics Technology Ltd. (2007). "FaceStore".
`http://www.anthropics.com`

Artificial Solutions Iberia S.L. (2007). "Anna".
`http://www.artificial-solutions.com`

AT&T Corp. (2007). "AT&T Natural Voices - Products and Services".
`http://www.naturalvoices.att.com/products/tts_data.html`

Bailly, G. (2001). "Audiovisual Speech Synthesis". En: *Proceeding of ESCA Tutorial and Research Workshop on Speech Synthesis*, pp. 1–10. Pertshire, Escocia.

Baker, S. y Matthews, I. (2001). "Equivalence and Efficiency of Image Alignment Algorithms". En: *Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition*, volumen 1, pp. 1090–1097. Kauai, HI, EUA.

Baker, S. y Matthews, I. (2004). "Lucas-Kanade 20 Years On: A Unifying Framework". *International Journal of Computer Vision*, **56(3)**, pp. 221 – 255. ISSN 0920-5691.

Bartels, R.H., Beatty, J.C. y Barsky, B.A. (2006). *An Introduction to Splines for Use in Computer Graphics and Geometric Modeling*. Morgan Kaufmann Publishers, Inc.. ISBN 1-558-60400-6.

Beier, T. y Neely, S. (1992). "Feature-Based Image Metamorphosis". En: *Proceedings of the annual conference on Computer graphics and interactice techniques*, pp. 35–42. Chicago, IL, EUA.

Benguerel, A.P. y Pichora-Fuller, M. (1982). "Coarticulation effects in lipreading". *Journal of Speech and Hearing Research*, **25**, pp. 600–607. ISSN 1092-4388.

Beskow, J. (1995). "Rule-based Visual Speech Synthesis". En: *Proceedings of the European Conference on Speech Communication and Technology*, pp. 299–302. Madrid, España.

Beskow, J. (2003). *Talking Heads: Models and Applications for Multimodal Speech Synthesis*. Tesis doctoral, Department of Speech, Music and Haring, KTH, Estocolmo, Suecia.

Binnie, C., Montgomery, A. y Jackson, P. (1974). "Auditory and Visual Contributions to the Perception of Consonants". *Journal of Speech and Hearing Research*, **17**, pp. 619–630. ISSN 1092-4388.

Black, M.J. y Jepson, A. (1998). "EigenTracking: Robust matching and tracking of articulated objects using a view-based representation". *International Journal of Computer Vision*, **26(1)**, pp. 63–84. ISSN 0920-5691.

Blanz, V., Basso, C., Poggio, T. y Vetter, T. (2003). "Reanimating Faces in Images and Video". *Computer Graphics Forum*, **22(3)**, pp. 641–650. ISSN 0167-7055.

Blanz, V. y Vetter, T. (1999). "A morphable model for the synthesis of 3D faces". En: *Proceedings of the annual conference on Computer graphics and interactive techniques*, pp. 187–194. Nueva York, NY, EUA.

Bloomenthal, J. y Rokne, J. (1994). "Homogeneous Coordinates". *The Visual Computer: International Journal of Computer Graphics*, **11(1)**, pp. 15–26. ISSN 0178-2789.

Brand, M. (2002). "Incremental Singular Value Decomposition of Uncertain Data with Missing Values". En: *Proceedings of European Conference on Computer Vision*, pp. 707–ff. Copenhague, Dinamarca.

Brand, Matthew (1999). "Voice puppetry". En: *Proceedings of the annual conference on Computer graphics and interactive techniques*, pp. 21–28. Nueva York, NY, EUA.

Bregler, C., Covell, M. y Slaney, M. (1997). "Video Rewrite: Driving Visual Speech with Audio". En: *Proceedings of the annual conference on Computer graphics and interactive techniques*, pp. 353–360. Los Ángeles, CA, EUA.

Brooke, N. y Scott, S. (1998). "Two- and three-dimensional audio-visual speech synthesis". En: *Proceedings of Auditory-Visual Speech Processing*, pp. 213–218. Terrigal, Australia.

Buenaposada, J.M., Muñoz, E. y Baumela, L. (2004). "Efficient appearance-based tracking". En: *Proceedings of IEEE Workshop on Nonrigid and Articulated Motion*, pp. 6–6. Washington, DC, EUA.

Cantu-Paz, Erick (2000). *Efficient and Accurate Parallel Genetic Algorithms*. Kluwer Academic Publishers. ISBN 0-792-37221-2.

Cao, Y., Tien, W.C., Faloutsos, P. y Pighin, F. (2005). "Expressive speech-driven facial animation". *ACM Transactions on Graphics*, **24(4)**, pp. 1283–1302. ISSN 0730-0301.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F. y Espesser, R. (1996). "About the relationship between eyebrow movements and F_0 variations". En: *International Conference on Spoken Language Processing*, pp. 2175–2179.

Chandrasekaran, S., Manjunath, B., Wang, Y., Winkeler, J. y Zhang, H. (1997). "An Eigenspace Update Algorithm for Image Analysis". *Graphical Models and Image Processing*, **59(5)**, pp. 321–332. ISSN 1077-3169.

Chandrasiri, N.P., Naemura, T., Ishizuka, M., Harashima, H. y Barakonyi, I. (2004). "Internet Communication Using Real-Time Facial Expression Analysis and Synthesis". *IEEE MultiMedia*, **11(3)**, pp. 20–29. ISSN 1070-986X.

Chang, Y.J. y Ezzat, T. (2005). "Transferable videorealistic speech animation". En: *Proceedings of the ACM SIGGRAPH/Eurographics symposium on Computer animation*, pp. 143–151. Los Angeles, CA, EUA.

Chen, T. (2001). "Audiovisual Speech Processing: Lip Reading and Lip Synchronization". *IEEE Signal Processing Magazine*, **18(1)**, pp. 9–31. ISSN 1053-5888.

Chen, T. y Rao, R. (1998). "Audio-visual integration in mulimodal communication". *Proceedings of the IEEE*, **86(5)**, pp. 837–852. ISSN 0018-9219.

Choi, K.H. y Hwang, J.N. (2005). "Automatic Creation of a Talking Head from a Video Sequence". *IEEE Transactions on Multimedia*, **7(4)**, pp. 628–637. ISSN 1520-9210.

Cohen, M.M. y Massaro, D.W. (1993). "Modeling Coarticulation in Synthetic Visual Speech". En: *Models and Techniques in Computer Animation*, pp. 139–156. Springer-Verlag. ISBN 3-540-70124-9.

Cootes, T., Edwards, G. y Taylor, C. (1998). "Active Appearance Models". En: *Proceedings of European Conference on Computer Vision*, pp. 581–695. Freiburg, Alemania.

Cosatto, E. (2002). *Sample-Based Talking-Head Synthesis*. Tesis doctoral, Swiss Federal Institute of Technology, Lausanne, Suiza.

Cosatto, E. y Graf, H.P. (1998). "Sample-Based Synhtesis of Photo-Realistic Talking Heads". En: *Computer Animation*, pp. 103–110. Philadelphia, PA, EUA.

Cosatto, E. y Graf, H.P. (2000). "Photo-Realistic Talking-Heads from Image Samples". *IEEE Transactions on Multimedia*, **2(3)**, pp. 152–163. ISSN 1520-9210.

Cosker, D., Marshall, D., Rosin, P. y Hicks, Y. (2004). "Speech-driven facial animation using a hierarchical model". *IEE Proceedings on Vision, Image and Signal Processing*, **151(4)**, pp. 314–321. ISSN 1350-245X.

Cosker, D., Paddock, S., Marshall, D., R., P.L. y Rushton, S. (2005). "Toward Perceptually Realistic Talking Heads: Models, Methods, and McGurk". *ACM Transactions on Applied Perception*, **2(3)**, pp. 270–285. ISSN 1544-3558.

Cotton, J. (1935). "Normal 'visual-hearing"'. *Science*, **82**, pp. 582–593. ISSN 0036-8075.

Davis, S. y Mermelstein, P. (1980). "Comparison of Parametric Representations for Monosyllable Word Recognition in Continuous Spoken Sentences". *IEEE Transactions on Acoustics, Speech and Signal Processing*, **28(4)**, pp. 357–366. ISSN 0096-3518.

Deng, Z. y Neumann, U. (2006). "eFASE: expressive facial animation synthesis and editing with phoneme-isomap controls". En: *Proceedings of the ACM SIGGRAPH/Eurographics symposium on Computer animation*, pp. 251–260. Aire-la-Ville, Suiza.

Devore, J.L. (2000). *Probabilidad y estadística para ingeniería y ciencias.* Duxbury, Brooks/Cole, 5ª edición. ISBN 0-534-37281-3.

Digimask Ltd. (2007). "Digimask".
`http://www.digimask.com/`

Dijkstra, E. (1959). "A note on two problems in connexion with graphs". *Numerische Mathematik*, **1**, pp. 269–271. ISSN 0029-599X.

Ekman, P. (1979). *About brows: emotional and conversational signals.* pp. 169–248. Cambridge University Press.

Ezzat, T., Geiger, G. y Poggio, T. (2002). "Trainable Videorealistic Speech Animation". En: *Proceedings of the annual conference on Computer graphics and interactive techniques*, pp. 225–228. San Antonio, TX, EUA.

Ezzat, T. y Poggio, T. (1996). "Facial analysis and synthesis using image-based models". En: *International Conference on Automatic Face and Gesture Recognition*, pp. 116–121. Kellington, VT, EUA.

Ezzat, T. y Poggio, T. (1997). "Videorealistic Talking Faces: A Morphing Approach". En: *Proceedings of Auditory Visual Speech Processing*, pp. 26–27. Rodas, Grecia.

Fabian, P. y Francik, J. (2001). "Synthesis and Presentation of the Polish Sign Language Gestures". En: *Proceedings of International Conference on Applied Mathematics and Informatics at Universities*, pp. 190–197. Gabcikovo, Eslovaquia.

Fagel, S. (2004). "Video-realistic Synthetic Speech with a Parametric Visual Speech Synthesizer". En: *Proceedings of International Conference on Spoken Language Processing*, pp. 2033–2036. Isla Jeju, Korea.

Finn, K. (1986). *An Investigation of Visible Lip Information to be use in Automatic Speech Recognition.* Tesis doctoral, Georgetown University, Washington, DC, EUA.

Fisher, C. (1968). "Confusions among visually perceived consonants". *Journal of Speech and Hearing Research*, **11**, pp. 796–804. ISSN 1092-4388.

Forner, A. (1999). *La Comunicación No Verbal*. Grao, Spain. ISBN 8-485-72960-9.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. USA: Academic Press, EUA. ISBN 0-122-69851-7.

Geiger, G., Ezzat, T. y Poggio, T. (2003). "Perceptual Evaluation of Video-Realistic Speech". *Informe técnico CBCL Paper number 224/ AI Memo number 2003-003*, Massachusetts Institute of Technology.

German, S. y McClure, D. (1987). "Statistical methods for tomographic image reconstruction". *Bulletin of the International Statistical Institute*, **52**, pp. 4–5. ISSN 0074-8609.

Goldberg, David E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, EUA. ISBN 0-201-15767-5.

Goldschen, A.J. (1993). *Continuous Automatic Speech Recognition by Lipreading*. Tesis doctoral, George Washington University.

Golub, G.H. y Loan, C.F. Van (1996). *Matrix Computations*. The Johns Hopkins University Press. ISBN 0-801-85413-8.

Grieve-Smith, Angus B. (2001). "SignSynth: A Sign Language Synthesis Application Using Web3D and Perl". En: *International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction*, pp. 134–145. Londres, Reino Unido.

GSM 06.90 version 7.2.1 (1998). "Digital cellular telecommunications system (Phase 2+); Adaptive Multi-Rate speech transcoding".

Gu, M. y Eisenstat, S.C. (1993). "A Stable and Fast Algorithm for Updating the Singular Value Decomposition". *Informe técnico YALEU/DCS/RR-966*, Yale University, New Haven, CT, EUA.

Guaus, R. y Iriondo, I. (2000). "Diphone based Unit Selection for Catalan Text-to-Speech Synthesis". En: *Workshop on Text, Speech and Dialogue*, pp. 277–282. Brno, República Checa.

Gutiérrez-Osuna, R., Kakumanu, P., Espósito, A., García, O.N., Bojórquez, A., Castello, J. y Rudomin, I. (2005). "Speech-Driven Facial Animation with Realistic Dynamics". *IEEE Transactions on Multimedia*, **7**, pp. 33–42. ISSN 1520-9210.

H-Anim (2007). "Humanoid Animation ISO/IEC 19774".
`www.h-anim.org`

Hager, G.D. y Belhumeur, P.N. (1998). "Efficient Region Tracking With Parametric Models of Geometry and Illumination". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20(10)**, pp. 1025–1039. ISSN 0162-8828.

Hall, P. M., Marshall, D. R. y Martin, R. (2002). "Adding and subtracting eigenspaces with eigenvalue decomposition and singular value decomposition". *Image and Vision Computing*, **20(13-14)**, pp. 1009–1016. ISSN 0262-8856.

Hall, P.M., Marshall, A.D. y Martin, R.R. (2000). "Merging and. Splitting Eigenspace Models". *Transactions on Pattern Analysis and Machine Intelligence*, **22(9)**, pp. 1042–1049. ISSN 0162-8828.

Hanke, T. (2002). "Interface Definitions, ViSiCAST Deliverable D5-1". *Informe técnico*, Insitute of German Sign Language and Communication of the Deaf, University of Hamburg, Germany.

Hill, H., Troje, N. y Johnston, A. (2005). "Range- and domain-specific exaggeration of facial speech". *Journal of Vision*, **5(10)**, pp. 793–807. ISSN 1534-7362.

Holland, P.W. y Welsch, R.E. (1977). "Robust Regression Using Iteratively Reweighted Least-Squares". *Communications in Statistics: Theory and Methods*, **A6**, pp. 813–827. ISSN 0361-0926.

Hong, P., Wen, Z. y Huang, T.S. (2002). "Real-Time Speech-Driven Face Animation With Expressions Using Neural Networks". *IEEE Transactions on neural networks*, **13(4)**, pp. 916–927. ISSN 1045-9227.

Huang, Xuedong, Hon, Hsiao-Wuen y Reddy, Raj (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Prentice Hall PTR. ISBN 0-130-22616-5.

Huenerfauth, M. (2006). *Generating American Sign Language Classifier Predicates for English-to-ASL Machine Translation.* Tesis doctoral, University of Pennsylvania, Philadelphia, PA, EUA.

Hunt, A. y Black, A. (1996). "Unit selection in a concatenative speech synthesis system using a large speech database". En: *Proceedings of ICASSP*, volumen 1, pp. 373–376. Atlanta, GA, EUA.

IKEA Ibérica S.A. (2007). "Pregúntale a Anna".
`www.ikea.com/ms/es_ES`

Iriondo, I., Alías, F., Melenchón, J. y Llorca, M.A. (2004). "Modeling and Synthesizing Emotional Speech for Catalan Text-to-Speech Synthesis". *Lecture Notes in Computer Science Tutorial and Research Workshop on Affective Dialog Systems*, **3068(1)**, pp. 197–208. ISSN 0302-9743.

Irving, Amanda y Foulds, Richard (2005). "A parametric approach to sign language synthesis". En: *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*, pp. 212–213. Nueva York, NY, EUA.

Jaklič, A., Vodopivec, D, Komac, V. y Gašperič, M. (1995). "Multimedia learning tools for the hearing impaired". En: *World Conference on Educational Multimedia and Hypermedia ED-MEDIA*, pp. 354–359. Graz, Austria.

Jebara, T., Russell, K. y Pentland, A. (1998). "Mixtures of eigenfeatures for real-time structure from texture". En: *International Conference on Computer Vision*, pp. 128–135. Bombay, India.

Johnson, K., Ladefoged, P. y Lindau, M. (1993). "Individual differences in vowel production". *Journal of the Acoustical Society of America*, **94(2)**, pp. 701–714. ISSN 0001-4966.

Jolliffe, I.T. (1986). *Principal Component Analysis*. New York Springler-Verlag. ISBN 0-387-95442-2.

Karhunen, H. (1947). "Uber lineare methoden in der wahrscheinlich-keitsrechnung". *Annales Academiae Scientiarium Fennicae, A1: Mathematica-Physica*, **37(1)**, pp. 3–79. ISSN 1239-629X.

Karpouzis, K., Caridakis, G., Fotinea, S.E. y Efthimiou, E. (2007). "Educational resources and implementation of a Greek sign language synthesis architecture". *Computers and Education*, **49(1)**, pp. 54–74. ISSN 0360-1315.

Kirby, M. (2001). *Geometric Data analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns*. John Wiley & Sons Inc., New York. ISBN 0-471-23929-1.

Kuratate, T. (2004). *Talking Head Animation System Driven by Facial Motion Mapping and a 3D Face Database*. Tesis doctoral, Nara Institute of Science and Technology, Nara, Japón.

Kuratate, T., Garcia, F., Yehia, H. y Vatikoitis-Bateson, E. (1997). "Facial animation from 3D Kinematics". En: *Acoustical Society of Japan*, pp. 323–324. Sapporo, Japón.

Lasseter, J. (1987). "Principles of Animation as Applied to 3D Character Animation". *Computer Graphics*, **21**, pp. 35–44. ISSN 0097-8930.

Lebourque, T. y Gibet, S. (1999). "High Level Specification and Control of Communication Gestures: The GESSYCA System". En: *Proceedings of the Computer Animation*, pp. 24–37. Washington, DC, USA.

Lee, Y., Terzopoulus, D. y Waters, K. (1995). "Realistic Modeling for Facial Animations". En: *Proceedings of the annual conference on Computer graphics and interactive techniques*, pp. 55–62. Los Angeles, CA, EUA.

Levy, A. y Lindenbaum, M. (2000). "Sequential Karhunen-Loeve basis extraction and its application to images". *IEEE Transactions on Image Processing*, **9(1)**, pp. 1371–1374. ISSN 1057-7149.

Li, J., Chen, G., Chi, Z. y Lu, C. (2004). "Image Coding Quality Assessment Using Fuzzy Integrals With a Three-Component Image Model". *IEEE Transactions on Fuzzy Systems*, **12(1)**, pp. 99–106. ISSN 1063-6706.

LifeFX Inc. (2007). "LifeFX".
`http://www.lifefxtechnologies.com/`

Lim, J., Ross, D. y ans M.H. Yang, R.S. Lin (2005). "Incremental Learning for Visual Tracking". En: *Conference on Neural Information Processing Systems*, pp. 793–800. Vancouver, Canadá.

Liu, Z., Zhang, Z., Jacobs, C. y Cohen, M. (2001). "Rapid Modeling of Animated Faces from Video". *Journal of Visualization and Computer Animation*, **12(4)**, pp. 227–240. ISSN 1049-8907.

Losson, O. y Vannobel, J.M. (1998). "Sign language formal description and synthesis". *International Journal of Virtual Reality*, **3(4)**, pp. 27–35. ISSN 1081-1451.

Loève, M. (1955). *Probability Theory*. Van Nostrand, NJ, EUA. ISBN 0-387-90210-4.

Lucas, B. y Kanade, T. (1981). "An iterative image registration technique with an application to stereo vision". En: *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 674–679. Vancouver, Canadá.

Maestri, G. (1996). *Digital Character Animation*. New Riders Publishing. ISBN 1-562-05559-3.

Marshall, I. y Sáfár, É. (2002). "Sign language generation using HPSG". En: *Proceedings of International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 105–114. Santa Cruz, CA, EUA.

Massaro, D.W. (2001). "Auditory Visual Speech Processing". En: *Proceedings of the European Conference on Speech Communication and Technology*, pp. 1–4. Aalborg, Dinamarca.

Massaro, D.W., Beskow, J., Cohen, M.M., Fry, C.L. y Rodriguez, T. (1999). "Picture My Voice: Audio to Visual Synthesis using Artificial Neural Networks". En: *Proceedings of Auditory Visual Speech Processing*, pp. 133–138. Santa Cruz, CA, EUA.

Matthews, I., Ishikawa, T. y Baker, S. (2003). "The Template Update Problem". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26(6)**, pp. 810–815. ISSN 0162-8828.

McGraw-Hill Dictionary (2007). "computer vision".
`http://www.answers.com/topic/computer-vision`

McGurk, H. y McDonald, J. (1976). "Hearing Lips and Seeing Voices". *Nature*, **264**, pp. 746–748. ISSN 0028-0836.

Melenchón, J., Simó, J., Cobo, G. y Martínez, E. (2007). "Objective Viseme Extraction and Audiovisual Uncertainty: Estimation Limits between Auditory and Visual Modes". En: *Proceedings of the International Conference on Auditory Visual Speech Processing*, Hilvarenbeek, Holanda. A aparecer.

Melenchón, J., Alías, F. y Iriondo, I. (2002a). "PREVIS: A Person-Specific Realistic Virtual Speaker". En: *Proceedings of International Conference on Multimedia and Expo*, pp. 461 – 464. Lausanne, Switzerland.

Melenchón, J., de la Torre, F., Iriondo, I., Alías, F., Martínez, E. y Vicent, Ll. (2003a). "Text to Visual Synthesis with Appearance Models". En: *Proceedings of International Conference of Image Processing*, pp. 237–240. Barcelona, España.

Melenchón, J., Iriondo, I. y Alías, F. (2002b). "Modelo 2D Parametrizado Basado en Imágenes Reales Orientado a Síntesis de Cabezas Parlantes". En: *XVII Simposium Nacional de la Unión Científica Internacional de Radio (URSI)*, pp. 383–384. Alcalá de Henares, España.

Melenchón, J., Iriondo, I. y Meler, L. (2005). "Simultaneous and Causal Appearance Learning and Tracking". *Electronic Letters on Computer Vision and Artificial Intelligence*, **3(5)**, pp. 44–54. ISSN 1577-5097.

Melenchón, J., Iriondo, I., Socoró, J.C., Martínez, E. y Meler, L. (2003b). "Lip Animation of a Personalized Facial Model from Auditory Speech". En: *Proceedings of International Symposium on Signal Processing and Information Technology*, pp. 255 – 258. Darmstat, Alemania.

Melenchón, J. y Martínez, E. (2007). "Efficiently Downdating, Composing and Splitting Singular Value Decompositions Preserving the Mean Information". *Lecture Notes on Computer Science*, **4478**. A aparecer.

Melenchón, J., Meler, L. y Iriondo, I. (2004). "On-the-fly training". *Lecture Notes in Computer Science*, **3179**, pp. 146–153. ISSN 0302-9743. 2nd best paper award of International Workshop of Articulated Motion and Deformable Objects.

Microsoft® (2007). "Microsoft SAPI 5".
`http://www.microsoft.com/speech`

Micusik, B. y Hanbury, A. (2005). "Semi-Automatic Segmentation of Textured Images". En: *Proceedings of the Computer Vision Winter Workshop*, Zell an der Pram, Austria.

Ministerio de Educación y Ciencia (1990). "Alfabeto dactilológico".

Moon, T. (1999). *Mathematical Methods & Algorithms for Signal Processing.* Addison Wesley Publishing Company. ISBN 0-201-36186-8.

Morishima, S. (2001). "Face analysis and Synthesis". *IEEE Signal Processing Magazine*, **18(3)**, pp. 26–34. ISSN 1053-5888.

Murakami, H. y Kumar, V. (1982). "Efficient calculation of primary images from a set of images". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **4(5)**, pp. 511–515. ISSN 0162-8828.

Neidle, C., Sclaroff, S. y Athitsos, V. (2001). "SignStream: A tool for linguistic and computer vision research on visual-gestural language data". *Behavior Research Methods, Instruments and Computers*, **33(3)**, pp. 311–320. ISSN 0743-3808.

Ohki, M., Sagawa, H., Sakiyama, T., Oohira, E., Ikeda, H. y Fujisawa, H. (1994). "Pattern recognition and synthesis for sign language translation system". En: *Proceedings of the first annual ACM conference on Assistive technologies*, pp. 1–8. Nueva York, NY, EUA.

Ostermann, J., Chen, L.S. y Huang, T.S. (1998). "Animated Talking Head with Personalized 3D Head Model". *Journal of VLSI Signal Processing System*, **20(1-2)**, pp. 97–105. ISSN 0922-5773.

Ostermann, J. y Weissenfeld, A. (2004). "Talking Faces - Technologies and Applications". En: *Proceedings of the International Conference on Pattern Recognition*, pp. 826–833. Washington, DC, EUA.

Ouni, S., Cohen, M.M., Ishak, H. y Massaro, D.W. (2007). "Visual Contribution to Speech Perception: Measuring the Intelligibility of Animated Talking Heads". *EURASIP Journal on Audio, Speech, and Music Processing*, pp. Article ID 47891, 12 pages. ISSN 1687-4714.

Owens, E. y Blazek, B. (1985). "Visemes Observed by Hearing-Impaired and Normal-Hearing Adult Viewers". *Journal of Speech and Hearing Research*, **28**, pp. 381–393. ISSN 1092-4388.

Papadogiorgaki, M., Grammalidis, N., Makris, L., Sarris, N. y Strintzis, M. G. (2004). "VSigns - A Virtual Sign Synthesis Web Tool". En: *Proceedings of COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication*, Tesalónica, Grecia.

Parke, F.I. (1972). "Computer generated animation of faces". En: *Proceedings of the ACM annual conference*, pp. 451–457. New York, NY, EUA.

Pei, Y. y Zha, H. (2007). "Transferring of Speech Movements from Video to 3D Face Space". *IEEE Transactions on Visualization and Computer Graphics*, **13(1)**, pp. 58–69. ISSN 1077-2626.

Pelachaud, C., Magno-Caldognetto, E., Zmarich, C. y Cost, P. (2001). "Modeling an Italian Head". En: *Proceedings of the International Conference on Auditory Visual Speech Processing*, pp. 72–77. Scheelsminde, Dinamarca.

Platt, S.M. y Badler, N.I. (1981). "Animating facial expressions". En: *Proceedings of the annual conference on Computer graphics and interactive techniques*, pp. 245–252. New York, NY, EUA.

Pourtois, G., Debatisse, D., Despland, P.A. y de Gelder, B. (2002). "Facial expressions modulate the time course of long latency auditory brain potentials". *Cognitive brain research*, **14(1)**, pp. 99–105. ISSN 0926-6410.

Radovan, M. y Pretorius, L. (2006). "Facial animation in a nutshell: past, present and future". En: *Proceedings of the 2006 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries*, pp. 71–79. Somerset West, South Africa.

Ramírez, J.L. (2003). "La retórica, pórtico de la ciencia". *Elementos*, **10(50)**, pp. 3–7. ISSN 0187-9073.

Rautek, P., Viola, I. y Gröller, M.E. (2006). "Caricaturistic Visualization". *IEEE Transactions on Visualization and Computer Graphics*, **12(5)**, pp. 1085–1092. ISSN 1077-2626.

Real Academia Española (2001). *Diccionario de la Lengua Española (22a edición).* Real Academia Española. ISBN 8-423-96813-8.

Redman, L. (1984). *How To Draw Caricatures.* McGraw-Hill. ISBN 0-809-25685-1.

Revéret, L., Bailly, G. y Badin, P. (2000). "MOTHER: A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation". En: *Proceedings of International Conference of Speech and Language Processing*, Beijing, China.

Robbe-Reiter, S., Carbonell, N. y Dauchy, P. (2000). "Expression constraints in multimodal human-computer interaction". En: *Intelligent User Interfaces*, pp. 225–228. New Orleans, Louisiana, United States.

Ríos, A. (1999). *La transcripción fonética automática del diccionario electrónico de formas simples flexivas del Español: Estudio fonológico en el léxico.* volumen 4.

Rousseeuw, P. y Leroy, A. (1987). *Robust Regression and Outlier Detection.* John Wiley and Sons. ISBN 0-471-85233-3.

Scansoft Inc. (2005). "RealSpeak: Expressive, Natural, Multi-Lingual TTS". `www.scansoft.com/speechworks/realspeak/`

Scarborough, R., Keating, P., Baroni, M., Cho, T., Mattys, S., Alwan, A., Auer, E. y Bernstein, L.E. (2006). "Optical Cues to the Visual Perception of Lexical and Phrasal Stress in English". En: *International Conference on Speech Prosody*, pp. 217–220. Dresden, Alemania.

Schantz, M. y Poizner, H. (1982). "A computer program to synthesize American Sign Language". *Behavior Research Methods and Instrumentation*, **14(5)**, pp. 467–474. ISSN 0005-7878.

Sheskin, D.J. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures.* Chapman & Hall/CRC. ISBN 1-584-88814-8.

Sifakis, E., Selle, A., Robinson-Mosher, A. y Fedkiw, R. (2006). "Simulating speech with a physics-based facial muscle model". En: *Proceedings of the ACM SIGGRAPH/Eurographics symposium on Computer animation*, pp. 261–270.

Skočaj, D. y Leonardis, A. (2003). "Weighted and robust incremental method for subspace learning". En: *Proceedings of International Conference of Computer Vision*, volumen 2, pp. 1494–1501. Niza, Francia.

Solina, F., Krapež, S., Jaklič, A. y Komac, V. (2001). *Multimedia Dictionary and Synthesis of Sign Language.* pp. 268–281. Idea Group Publishing. ISBN 1-930-70800-9.

Speers, A.L. (2001). *Representation of American Sign Language for Machine Translation.* Tesis doctoral, Georgetown University, Washington, DC, EUA.

Stephanidis, C. (1999). "Designing for all in the Information Society: Challenges towards universal access in the information age". *Informe técnico ERCIM ICS Report*, Institute of Computer Science, Heraklion, Grecia.

Sternberg, M.L.A. (1994). *The American Sign Language Dictionary on CD-ROM.* Harper Collins, New York.

Stevens, S.S. y Volkman, J. (1940). "The Relation of Pitch to Frequency". *Journal of Psichology*, pp. 329–353. ISSN 1577-7057.

Stewart, G.W. (1993). "On the early history of the singular value decomposition". *Society for Industrial and Applied Mathematics Review*, pp. 551–566. ISSN 0368-4245.

Stokoe, W., Casterline, D. y Croneberg, C. (1965). *A Dictionary of American Sign Language on Linguistic Principles.* Linstok Press.

Sumby, W. y Pollack, I. (1954). "Visual Contribution to Speech Intelligibility in Noise". *The Journal of the Acoustical Society of America*, pp. 212–215. ISSN 0001-4966.

Summerfield, A.Q. (1987). *Hearing by Eye: The psychology of lip-reading.* capítulo Some preliminaries to a comprehensive account of audio-visual speech perception, pp. 3–51. Lawrence Erlbaum Associates. ISBN 0-863-77038-X.

Suszczańska, N. y Szmal, P. (2001). "Machine Translation from Written Polish to the Sign Language in a Symbolic Form". En: *First International Conference on Applied Mathematics and Informatics at Universities*, pp. 90–97. Gabèikovo, Eslovaquia.

Swerts, M. y Krahmer, E. (2004). "Congruent and incongruent audiovisual cues to prominence". En: *Speech Prosody Conference*, pp. 69–72. Nara, Japón.

Swerts, M. y Krahmer, E. (2006). "The importance of different facial areas for signalling visual prominence". En: *Proceedings of Interspeech*, Pittsburg, PA, EUA. Paper 1289.

Switerslood, I., Verlinden, M., Ros, J. y van der Schoot, S. (2004). "Synthetic Signing for the Deaf: eSIGN". En: *Conference and Workshop on Assistive Technologies for Vision and Hearing Impairment*, Granada, Spain.

Tarr, M.J. y Pinker, S. (1989). "Mental rotation and orientation-dependence in shape recognition". *Cognitive Psychology*, pp. 233–282. ISSN 0010-0285.

Tekalp, A.M. y Ostermann, J. (2000). "Face and 2-D Mesh Animation in MPEG-4". *Signal Processing: Image Communication*, **15**, pp. 387–421. ISSN 0923-5965.

Theobald, B., Cawley, G., Matthews, I. y Bangham, J. (2004). "Near videorealistic synthetic talking faces: Implementation and Evaluation". *Speech Communication*, **44(1-4)**, pp. 127–140. ISSN 0167-6393.

Tolkien, J.R.R. (1954). *The Lord of the Rings.* Ballantine Books, New York. ISBN 0-618-00222-7.

Torre, F. y Black, M. (2002). "Robust parameterized component analysis: Theory and applications to 2d facial modeling". En: *Proceedings of European Conference of Computer Vision*, pp. 654–669. Copenhague, Dinamarca.

Veale, T., Conway, A. y Collins, B. (1998). "The Challenges of Cross-Modal Translation: English-to-Sign-Language Translation in the Zardoz System". *Machine Translation*, **13(1)**, pp. 81–106. ISSN 0922-6567.

Viterbi, A.J. (1967). "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm". *IEEE Transactions on Information Theory*, **13**, pp. 260–269. ISSN 0018-9448.

Vlasic, D., Brand, M., Pfister, H. y Popović, J. (2005). "Face transfer with multilinear models". *ACM Transactions on Graphics*, **24(3)**, pp. 426–433. ISSN 0730-0301.

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K. y Jones, C.J. (1977). "Effects of training on the visual recognition of consonants". *Journal of Speech and Hearing Research*, **20**, pp. 130–145. ISSN 1092-4388.

Waters, K. (1987). "A muscle model for animation three-dimensional facial expression". En: *Proceedings of the annual conference on Computer graphics and interactive techniques*, pp. 17–24. New York, NY, EUA.

Web3D (2007). "Web3D Consortium - Royalty Free, Open Standards for Real-Time 3D Communication".
`http://www.web3d.org/`

Wells, J.C. (1997). *SAMPA computer readable phonetic alphabet.* Gibbon, D., Moore, R. and Winski, R. (eds.), Berlin and New York: Mouton de Gruyter.

Yamamoto, E., Nakamura, S. y Shikano, K. (1998). "Lip movement synthesis from speech based on Hidden Markov Models". *Speech Communication*, **26(1-2)**, pp. 105–115. ISSN 0167-6393.

Yehia, H. y Itakura, F. (1994). "Determination of Human Vocaltract Dynamic Geometry from Formant Trajectories Using Spatial and Temporal Fourier Analysis". En: *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pp. 477–480. Adelaide, Australia.

Yehia, H., Rubin, P. y Vatikiotis-Bateson, E. (1998). "Quantitative Association of Vocaltract and Facial Behaviour". *Speech Communication*, **26(1-2)**, pp. 23–43. ISSN 0167-6393.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. y Woodland, P. (2003). *The HTK Book (for HTK Version 3.2.1).* Cambridge University Engineering Department.

Zhang, Z., Liu, Z., Adler, D., Cohen, M.F., Hanson, E. y Shan, Y. (2004). "Robust and Rapid Generation of Animated Faces from Video Images: A Model-Based Modeling Approach". *International Journal of Computer Vision*, **58(2)**, pp. 93–119. ISSN 0920-5691.

Zhao, L., Kipper, K., Schuler, W., Vogler, C., Badler, N.I. y Palmer, M. (2000). "A Machine Translation System from English to American Sign Language". En: *Proceedings of the*

Conference of the Association for Machine Translation in the Americas on Envisioning Machine Translation in the Information Future, pp. 54–67. Londres, Reino Unido.

Zhao, W., Chellappa, R., Phillips, P. J. y Rosenfeld, A. (2003). "Face recognition: A literature survey". *ACM Computer Survey*, **35(4)**, pp. 399–458. ISSN 0360-0300.

Índice alfabético

Apéndice A

Aportaciones

En este capítulo se presentan las aportaciones del trabajo a la comunidad científica y a la sociedad en general. Se describen las publicaciones realizadas, comentándolas y ofreciendo las relaciones oportunas con lo expuesto en capítulos anteriores. También se mencionan los proyectos, empresas y eventos relacionados que muestran el uso de los métodos aquí expuestos para ofrecer un punto de vista adicional sobre la aplicación práctica del presente proyecto.

A.1. Publicaciones científicas

El presente trabajo de investigación ha proporcionado diferentes aportaciones de interés para la comunidad científica. De hecho, las principales ideas en las que se sustenta la actividad investigadora aquí presentada se encuentran expuestas en diferentes congresos y revistas de ámbito nacional e internacional.

A.1.1. Internacionales

El impacto que ha tenido el contenido del trabajo de investigación sobre la comunidad científica, hasta el momento, se puede resumir en las siguientes publicaciones:

1. J. Melenchón, F. Alías, I. Iriondo, *PREVIS: A Person-Specific Realistic Virtual Speaker*, Proc. of IEEE international Conference on Multimedia and Expo (ICME) [CD-ROM], Lausanne (SUIZA), Agosto 2002, ISBN 0-7803-7304-9.

2. J. Melenchón, F. De la Torre, I. Iriondo, F. Alías, E. Martínez, Ll. Vicent, *Text to visual synthesis with appearance models*, Proc. of IEEE International Conference on Image Processing (ICIP), vol. 1, n. 1, pp: 237-240, Barcelona (ESPAÑA), Setiembre 2003, ISBN 0-7803-7750-8.

3. J. Melenchón, I. Iriondo, J.C. Socoró, E. Martínez, L. Meler, *Lip Animation of a personalizad Facial Model from Auditory Speech*, Proc. of IEEE Internacional Symposium on Signal Processing and Information Technology (ISSPIT) [CD-ROM], Darmstat (ALEMANIA), Diciembre 2003, ISBN 0-7803-8293-5.

4. J. Melenchón, L. Meler, I. Iriondo, *On-the-fly Training*, Lecture Notes in Computer Science, Third International Workshop, AMDO 2004, vol. 3179, n.1, pp: 146-153, Palma de Mallorca (ESPAÑA), Setiembre 2004, ISSN 0302-9743.

5. J. Melenchón, I. Iriondo, L. Meler, *Simultaneous and Causal Appearance Learning and Tracking*, Electronic Letters on Computer Vision and Image Analysis Special Issue on Articulated Motion and Deformable Objects, ELCVIA 2005, vol. 5, n. 3, pp: 44-54 , Barcelona (ESPAÑA), Abril 2005, ISSN 1577-5097.

6. J. Melenchón, E. Martínez, *Efficiently Downdating, Composing and Splitting Singular Value Decompositions Preserving the Mean Information*, a aparecer en la Iberian Conference on Pattern Recognition and Image Analysis, en J.Martí et al. (Eds.): IbPRIA 2007, Part II, LNCS 4478, pp: 436-443, Girona (ESPAÑA), Junio 2007.

7. J. Melenchón, J. Simó, G. Cobo, E. Martínez, *Objective Viseme Extraction and Audiovisual Uncertainty: Estimation Limits between Auditory and Visual Modes*, a aparecer en la International Conference on Auditory-Visual Speech Processing, Hilvarenbeek (HOLANDA), Agosto 2007.

La publicación número 1 está relacionada con los métodos de análisis por autoespacios detallados en el capítulo 3. En esta comunicación se presenta una versión primitiva del esquema de síntesis y análisis audiovisual finalmente desarrollado. El modelo visual, en este trabajo, consiste en representar las diferentes apariencias de los labios utilizando información de forma y textura y almacenar las apariencias faciales como imágenes pregrabadas. Aunque se puede crear el modelo visual a partir de cualquier cara, el proceso es mucho más tedioso que el presentado en el capítulo 3, debido a la mayor cantidad de segmentación manual necesaria, y sólo es capaz de tratar imágenes en niveles de gris. Por otro lado, se introduce el método de síntesis guiada por fonética presentada en el apartado 4.2.4 mediante el uso de motores que implementan las interfaces de SAPI® de Microsoft®.

La comunicación número 2 muestra una versión inicial del algoritmo de interpolación no lineal de alta dimensionalidad, explicado con detalle en el apartado 4.1.3.1, aplicado a la síntesis de caras parlantes. También introduce el uso de modelos para representar la cara incorporándoles información de color, que derivarán finalmente en el modelo visual utilizado en este trabajo y explicado en el apartado 2.2.

La publicación 3 contiene el primer estudio preliminar sobre la síntesis guiada por voz que aparece en el apartado 4.2.5 utilizando estimación bayesiana. El trabajo utiliza modelos visuales de la cara con información de color y vislumbra la necesidad de aumentar la longitud de las secuencias a analizar para mejorar el modelo construido (esta necesidad desembocará en el desarrollo de las siguientes publicaciones presentadas). Además, muestra un conjunto reducido de pruebas orientadas a medir el aumento de inteligibilidad aportado

por el canal visual. A diferencia de las pruebas realizadas en el presente trabajo de investigación para evaluar el realismo de la síntesis (apdo. 5.3), en las de esta comunicación no se evalúa su naturalidad y tampoco se usan efectos de coarticulación conjuntamente con el método de estimación MAP, con lo que el hecho de que el método MMSE obtenga los mejores resultados de inteligibilidad, bajo esas condiciones, no tiene porqué ofrecer ninguna contradicción.

El algoritmo de seguimiento y aprendizaje simultáneo explicados en el apartado 3 toma como base la publicación número 4. Premiada como la segunda mejor publicación presentada en el evento asociado, esta comunicación presenta el *On-the-fly Training Algorithm*, el algoritmo de cálculo incremental de la SVD con actualización de la media que conforma una base inicial para del desarrollo del apartado 3.2.3.4, aplicándolo, además, al seguimiento de caras y construcción de modelos visuales. La publicación número 5 contiene una versión más aproximada del algoritmo de análisis de este trabajo, juntamente con una primera evaluación más detallada del coste y precisión conseguidos por el algoritmo de cálculo incremental de la SVD con actualización de la media.

Los aspectos sobre combinaciones de SVDs explicados en el apartado 3.2.3 se presentan en la publicación número 6, en concreto, las versiones decremental (ver apartado 3.2.3.5), compuesta (ver apartado 3.2.3.6) y partida (ver apartado 3.2.3.7).

Finalmente, los aspectos sobre obtención personalizada de conjuntos de visemas del apartado 4.2.2, así como la incertidumbre audiovisual presentada en el apartado 4.2.6 se ofrecen en la publicación 7.

A.1.2. Nacionales

También se han realizado otras publicaciones secundarias de ámbito nacional relacionadas con el trabajo de investigación presentado. Aclarar únicamente que las que tienen como primer autor a J. Melenchón poseen una relación más estrecha con la actividad investigadora que el resto:

1. J. Melenchón, I. Iriondo, F. Alías, *Modelo 2D Parametrizado Basado en Imágenes Reales Orientado a Síntesis de Cabezas Parlantes*, XVII Simposium Nacional de la Unión Científica Internacional de Radio (URSI), vol. 1, n. 1, pp: 383-384,Alcalá de Henares, Setiembre 2002, ISBN 84-8138-517-4.

2. J. Gonzalvo, J.A. Morán, J. Melenchón, *Detección del Ángulo de Llegada con un Array Microfónico*, XVIII Simposium Nacional de la Unión Científica Internacional de Radio (URSI) [CD-ROM], La Coruña, Setiembre 2003, ISBN 84-9749-081-9.

3. A. Arañó, J. Melenchón, *Compresión en Tiempo Real para Síntesis Audiovisual*, XIX Simposium Nacional de la Unión Científica Internacional de Radio (URSI) [CD-ROM], Barcelona, Setiembre 2004, ISBN: 84-688-7736-0.

4. J. Melenchón, *Síntesis Multimodal Realista Personalizable*, I Congreso Español de Informática (CEDI) [CD-ROM], Granada, Setiembre 2005, ISBN: 84-9732-447-1.

5. X. Sevillano, J. Melenchón, J.C. Socoró, *Análisis y Síntesis Audiovisual para Interfaces Multimodales Ordenador-Persona*, VII Congreso Internacional de Interacción Persona-Ordenador, noviembre 2006, ISBN: 84-690-1613-X.

La primera publicación contiene una breve descripción del modelo facial utilizado en Melenchón et al. (2002a), cuya filosofía de reducción de dimensionalidad comparte el modelo visual presentado en este trabajo. El modelo presentado en esta comunicación soporta movimientos de cabeza, expresiones faciales y diferentes apariencias de labios; éstos últimos se representan usando PCA sobre su textura y forma por separado.

La publicación 2 presenta una comparación entre diferentes métodos de estimación del ángulo de llegada de información acústica, aplicando el mejor método evaluado en una aplicación de locución virtual. El resultado final consiste en una cara parlante que mira en la dirección en que le habla el usuario.

La comunicación número 3 presenta un nuevo método de organización de los libros de códigos asociados a una cuantificación vectorial, utilizándolo en la compresión del canal visual producido por una aplicación de síntesis audiovisual. Además, se ofrecen resultados experimentales de su aplicabilidad en la plataforma de telefonía móvil.

La publicación 4 presenta el marco de trabajo para síntesis audiovisual realista de caras parlantes que se encuentra detallado en el presente trabajo de investigación. El trabajo cita una encuesta realizada a diferentes personas en el evento *Viu la ciència contemporània* organizado en *La Pedreda* de Barcelona por el *Parc Tecnològic de la UB* (ver apartado 5.3.2).

Finalmente, la publicación número 5 representa un primer acercamiento a la síntesis audiovisual simultánea de audio y vídeo. Se ofrece una comparación de los resultados obtenidos al crear secuencias audiovisuales codificando conjuntamente el audio y el vídeo mediante técnicas de SVD incremental desarrolladas en esta tesis y factorización de matrices no negativas (NMF). No obstante, estos resultados no pertencen estrictamente a la línea de trabajo desarrollado en la presente tesis.

A.2. Proyectos de investigación asociados

El trabajo de investigación desarrollado ha estado implicado en diferentes proyectos de ámbito privado y público, tanto de investigación como de explotación. En general, los proyectos con financiación pública versan sobre investigación y, en algunos casos, también sobre el desarrollo de un prototipo. Las colaboraciones con empresas privadas (con o sin financiación) han estado principalmente más ligadas al ámbito de la explotación de las técnicas desarrolladas en este trabajo. La existencia de proyectos con financiación pública relativos a la investigación pone de manifiesto el interés del estado en avanzar sobre las materias expuestas en los anteriores capítulos; por otro lado, encontrar empresas interesadas en la investigación llevada a cabo descubre una inquietud empresarial sobre la aplicación a la realidad de los métodos expuestos, para acercarlos de este modo a la sociedad.

A.2.1. Proyectos con financiación pública

Este tipo de proyectos tienen como objetivo principal realizar actividades investigadoras sobre la materia de síntesis audiovisual facial personalizable con el objetivo de mejorar la calidad, la facilidad de personalización, la fiabilidad y la simplicidad en el uso de futuras aplicaciones que implementen las funcionalidades asociadas.

A.2.1.1. Detección de estados incipientes de somnolencia en conductores de vehículos a motor

Este proyecto de investigación obtiene financiación a partir de la MCyCIT DPI2002-02279 y el PROFIT FIT-1101100, dentro del periodo 2000-2005. Se basa en el desarrollo de un prototipo (juntamente con sus técnicas asociadas) no intrusivo de seguimiento e interpretación de la apariencia de los ojos para evaluar el estado de somnolencia de la persona asociada. La colaboración se basa en el uso de diferentes modelos para representar la apariencia ocular, así como la utilización de técnicas de descomposición en valores singulares (SVD) para trabajar con subespacios de apariencia con el objetivo de mejorar la compactación de la información utilizada.

A.2.1.2. Desarrollo de un locutor virtual

Este proyecto ha obtenido financiación en 2002 a través del programa PROFIT del Ministerio de Industria, Turismo y Comercio con código FIT-150500-2002-410. Su finalidad consiste en el desarrollo y validación del prototipo de una nueva interfaz hombre-máquina multimedia, basada en un locutor de apariencia real con la capacidad de reproducir texto de forma automática, sincronizando la voz con la gesticulación facial. Los estudios relativos al modelo visual utilizado y la unión de éste con un sistema de texto a habla (TTS) son los puntos de confluencia de este proyecto con la tesis presentada.

A.2.1.3. Locutor virtual para interacción natural

La empresa VIDA Software (ver apartado A.2.3.3) y *Enginyeria La Salle* han trabajado en el año 2004, y mediante el desarrollo de este proyecto identificado por FIT-340100-2004-20, en la integración de la tecnología de Interacción Natural de VIDA Software con la Tecnología facial de La Salle. Por un lado, la tecnología de VIDA Software permite la interacción multimodal mediante voz, vista y tacto con aplicaciones móviles. Por otro, la tecnología visual, basada en el contenido de esta tesis, permite generar animaciones foto realistas faciales que son claramente complementarias a la tecnología de VIDA Software.

A.2.1.4. Videófono

Este proyecto se realiza gracias a la ayuda pública FIT-350300-2004-44 concedida en 2004 y se basa en el diseño y validación de una nueva interfaz natural para sistemas de información, capaz de convertir automáticamente la voz de salida del sistema en un mensaje audiovisual sensible a las condiciones del entorno. Es aquí donde se presenta el modelo visual, aplicándolo a caras humanas parlantes, juntamente con una versión más primitiva de los algoritmos de análisis y síntesis presentados en este trabajo.

A.2.2. IntegraTV-4all

El proyecto titulado IntegraTV-4all obtiene la financiación pública identificada por FIT-350301-2004-2, se lleva a cabo a lo largo de 2005 y se puede resumir como: servicios adaptados de ocio, información y tele-trabajo a través de la televisión en hoteles accesibles, con funcionalidades avanzadas de visión y habla asistida para facilitar la estancia a huéspedes con discapacidades sensoriales. La colaboración se basa en la introducción de una cabeza parlante de fácil personalizaciónpersonalizable en la interfaz desarrollada. El uso del esquema de síntesis y análisis presentado en este trabajo se hace palpable en esta parte del proyecto.

A.2.2.1. SAVE

El proyecto de Síntesis AudioVisual Expresiva (SAVE) se financia públicamente (TEC2006-08043/TCM) para trabajar aspectos de creación de caras parlantes capaces de transmitir estados de ánimo mientras hablan. Dado que es un proyecto con fin en el año 2009, su relación con el contenido de esta tesis se basa en las definiciones de corpus audiovisuales y el proceso de captura que aparece en la planificación del año 2007. En el resto del proyecto se pretenden utilizar los algoritmos de análisis y síntesis sobre los diferentes elementos faciales para dotar de expresividad a los avatares producidos.

A.2.3. Colaboraciones con empresas

Este tipo de actividades persiguen el aprovechamiento de los avances introducidos en el presente proyecto de investigación para presentarlos ante la sociedad mediante su implementación y explotación comercial.

A.2.3.1. EPSON

En el segundo semestre de 2002 se lleva a cabo una colaboración financiada con esta multinacional que se basa en la construcción de dos modelos visuales personales a dos individuos de la empresa (entre ellos, un alto directivo comercial) para su uso en diferentes actos sociales.

A.2.3.2. Artificial Solutions

Esta empresa sueca, especializada en sistemas de diálogo, utilizó los primeros resultados del esquema de síntesis audiovisual propuesto para generar diferentes agentes de apariencia realista con los cuales mantener un diálogo activo. Los resultados se pudieron experimentar en el evento APROP 2003 (ver apartado A.3.2.2).

A.2.3.3. VIDA Software

Esta pequeña empresa de reciente creación y gran empuje utiliza los conocimientos sobre síntesis y análisis audiovisual presentes en este trabajo de investigación para aplicarlos al desarrollo de una interfaz hombre máquina muy novedosa aplicada a dispositivos móviles. Las ayudas citadas en los apartados A.2.1.3 y A.2.1.4 se solicitaron conjuntamente con esta empresa.

A.2.3.4. Panoptics Produccions

La colaboración con esta empresa consiste también en el uso del esquema de síntesis audiovisual presentado en un robot llamado *TVMòbil*. Está formado por un ordenador con pantalla y altavoces montados sobre una plataforma móvil motorizada que se desplaza de forma autónoma gracias al uso de una batería. El desplazamiento puede ser por radio control, preprogramado o independiente mediante el uso de sensores. Se desea proporcionar la apariencia visual realista de una persona para que aparezca por la pantalla del robot, y aumentar así la sensación de que el aparato es un ser independiente que se mueve y habla como una persona humana.

A.2.3.5. TMT Factory

Esta empresa es una ingeniería de software y sistemas especializada en el desarrollo e implementación de proyectos tecnológicos a medida. En su proyecto de investigación titulado *IntegraTV for all*, desean incorporar una interfaz más natural basada en un locutor virtual de apariencia realista. Para conseguirlo, han decidido aprovechar el esquema de análisis y síntesis presentado en esta memoria (entre otros elementos) a través del proyecto citado en el apartado A.2.2.

A.2.3.6. Enginyeria i Arquitectura La Salle

Esta centenaria escuela de ingeniería ha mostrado también un gran interés en el desarrollo del contenido de este trabajo de investigación asociándolo a la creación de caras parlantes. Mediante diferentes programas de financiación propios, desde el año 2001 y hasta el 2005, ha apoyado toda la investigación realizada en este trabajo y ha aprovechado

sus resultados para mostrarlos en diferentes actividades y demostraciones. Los proyectos internos concedidos que han apoyado la investigación realizada han sido:

- *Personalització LV*, PGR-PR2002-03, 2001-2002.
- *Locutor Virtual Personalizable*, PGR-PR2002-03, 2002-2003.
- *Smart and Adaptive System-Human Interaction through Multimodal Interfaces*, PGR-PR200302, 2003-2004.
- *Processament Multimodal per interacció natural*, PGR-PR200402, 2004-2005.

A.3. Impacto social

El trabajo de investigación presentado ha estado en contacto directo con la sociedad desde sus inicios a través de los medios de comunicación y diferentes eventos sociales abiertos al público.

A.3.1. Aparición en medios de comunicación

La actividad investigadora asociada a este trabajo ha tenido eco en diferentes medios de comunicación desde el año 2000 hasta la actualidad. Diferentes cadenas de televisión como *Tele 5*, *TVE 1*, *BTV*, *TV3* y *C33* (también llamado *K3*) y de radio como *COM Ràdio*, han querido hacer llegar a la población la idea que existe detrás de este proyecto de investigación así como transmitir sus usos potenciales.

A.3.2. Participación en eventos

Los resultados obtenidos por este trabajo de investigación, tanto parciales como totales, han sido presentados en diferentes eventos con el objetivo de acercar la tecnología desarrollada a la sociedad. Estas actividades han enriquecido enormemente el trabajo y han sido claves para el logro de los objetivos planteados, ya que se han obtenido valiosas y muy útiles opiniones sobre el funcionamiento, aplicaciones y mejoras relacionadas con el trabajo llevado a cabo.

A.3.2.1. *Showroom*

Más que un evento, esta actividad consiste en una sala de demostraciones ubicada en *Enginyeria i Arquitectura La Salle*, que se encuentra abierta al público general. Además, se hacen diversas visitas guiadas por la misma, coincidiendo con diferentes actos orientados a dar a conocer la universidad al público general. Este trabajo de investigación se encuentra representado por una demostración que consiste en un locutor virtual que convierte la

entrada de texto en una cara parlante. La demostración se encuentra en el *Showroom* desde su creación en el año 2003.

A.3.2.2. *A-PROP* 2002-2003

Las jornadas *A-PROP* se realizan en *Enginyeria i Arquitectura La Salle* y consisten básicamente en mostrar el uso de nuevas tecnologías en diferentes ámbitos, acercando la investigación, el mundo empresarial y el particular entre ellos.

En las jornadas de 2002, tituladas *Impacte de la tecnologia sobre el territori*, se presentó una demostración con nombre *PREVIS (Person-specific REalistic VIrtual Speaker)*, que representa la implementación del prototipo explicado en Melenchón et al. (2002a) usando el modelo presentado en Melenchón et al. (2002b). Estas jornadas tuvieron un importante impacto en la sociedad a través de medios televisivos y radiofónicos.

Las jornadas *A-PROP* de 2003 versaron sobre el ocio en el siglo XXI y llevaron por título: *e-Entertainment: Oci pel segle XXI*. La demostración propuesta consistió en *L'interlocutor virtual*, una versión más avanzada de la presentada el año anterior, con la inclusión de color, un personaje femenino adicional y la capacidad de mantener una conversación a partir de las entradas de texto del usuario. Estas jornadas pusieron de manifiesto la colaboración con la empresa *Artificial Solutions* (apdo. A.2.3.2) al combinar su sistema de diálogo con el esquema de síntesis desarrollado hasta aquel momento.

A.3.2.3. El *Bus Tecnològic* de *Enginyeria La Salle*

Esta actividad, organizada por *Enginyeria i Arquitectura La Salle* durante el curso 2003-2004, acerca la investigación realizada en la universidad a las escuelas de Cataluña. El trabajo de investigación presentado ha contribuido a formar parte de esta iniciativa mediante la inclusión de una demostración del locutor virtual que acerca a los jóvenes estudiantes los resultados de un proceso de investigación. La demostración consistía en una aplicación interactiva basada en la presentada en Melenchón et al. (2002a), utilizando un modelo similar al de Melenchón et al. (2002b) con la adición de color.

A.3.2.4. *Viu la Ciència Contemporània* 2005

Esta iniciativa, promovida por el *Parc Científic* de la *Universidad de Barcelona* durante la primavera del año 2005, tuvo lugar en el conocido edificio de *La Pedrera* en Barcelona y se destinó a la población de esta ciudad y sus alrededores. Mostró durante unos días diferentes ejemplos de actividades científicas muy diversas, desde estudios sobre incendios hasta crecimiento de hongos, y en las que participó este trabajo de investigación a través de la aplicación de demostración *Desktop PREVIS II* (ver apdo. D.2). Se enseñaron diferentes resultados obtenidos por el actual esquema de síntesis y análisis audiovisual de caras parlantes, obteniendo una oportunidad adicional para contrastar opiniones con

los diversos visitantes. El impacto mediático de la investigación realizada también fue importante en este caso, sobre todo a través de programas y entrevistas en directo, tanto televisivas como radiofónicas.

A.3.2.5. *Dia de la Ciència a les Escoles* 2005-2006

La última actividad realizada en contacto directo con la sociedad tuvo lugar el 15 de noviembre de 2006. A las 12 horas de ese día y de forma totalmente simultánea, setenta y un científicos transmitieron su experiencia investigadora en setenta y una escuelas de bachillerato y formación profesional repartidas por toda Cataluña. La síntesis audiovisual personalizable asociada a la creación de caras parlantes presentada en este trabajo fue protagonista en una de ellas, concretamente en uno de los institutos de l'Hospitalet de Llobregat, en el que asistieron profesores y alumnos de los últimos cursos de bachillerato de entre toda la ciudad. Esta experiencia también se realizó un año antes, concretamente a las 10 horas del 10 de noviembre de 2005 en El Vendrell, donde asisitieron profesores y alumnos de entre toda la comarca del Baix Penedès de Tarragona.

Apéndice B

Descripción de corpus de caras

Para determinar un conjunto de treinta palabras bisílabas de la forma CVCV con los elementos posibles indicados en el cuadro 2.7 se ha utilizado un algoritmo genético paralelo (Cantu-Paz, 2000).

Sea $\mathbf{n}_{V\rightarrow sil}$ un vector de N_V elementos que contiene el número de veces que aparece cada transición vocal-silencio. Se define también $\mathbf{n}_{sil\rightarrow C}$ como un vector de N_C elementos que contiene el número de ocurrencias de cada transición silencio-consonante. Además, se considera un tercer vector $\mathbf{n}_{V\leftrightarrow C}$ de $N_V N_C$ elementos que almacena las transiciones vocal-consonante o consonante-vocal (ignorando el orden). En el ejemplo dado en el apartado 2.1.3.2, $N_V = 5$ y $N_C = 6$. Dados estos tres vectores, se presenta la función objetivo a minimizar como:

$$\mathcal{F} = \frac{n_V + n_C}{N_V + N_C + N_V\sigma_f + N_C\sigma_s + N_V N_C\sigma_t} \tag{B.1}$$

donde el número de vocales y consonantes diferentes viene representado por n_V y n_C, respectivamente; σ_f es la varianza del vector $\mathbf{n}_{V\rightarrow sil}$, σ_s lo es del vector $\mathbf{n}_{sil\rightarrow C}$ y σ_t, de $\mathbf{n}_{V\leftrightarrow C}$. Se han utilizado las constantes N_V, N_C y $N_V N_C$ para representar las sumas de diferencias de cuadrados a partir de cada una de las tres varianzas, respectivamente. Finalmente, se ha usado la constante de valor $N_V + N_C$ en el denominador para acotar el margen de valores de $\mathcal{F}$ entre cero y uno.

La solución concreta para treinta palabras CVCV se puede conseguir, de hecho, sin el uso de los algoritmos genéticos. No obstante, cualquier pequeña variación en el número de palabras, su formato, o el número de elementos implicados, puede comportar el replanteamiento del problema. Para evitar esta situación se ha elegido el uso de los algoritmos genéticos y no otro tipo de técnica no lineal debido a su modularidad. Éstos permiten cambiar el tipo de palabras y su cantidad modificando únicamente la parte de codificación que pasa de palabras a los tres vectores $\mathbf{n}_{V\rightarrow sil}$, $\mathbf{n}_{sil\rightarrow C}$ y $\mathbf{n}_{V\leftrightarrow C}$, sin tener que replantear todo el problema de nuevo.

Los conjuntos de palabras similares a las del cuadro 2.8 del apartado 2.1.3.2 se han obtenido minimizando (B.1) y siguiendo las sugerencias de Cantu-Paz (2000). La configu-

ración final ha consistido en una red de cuatro islas totalmente interconectadas entre ellas con una tasa de migración de 0,01, población total de 2400 individuos y 300 bits por cada uno. Se ha utilizado elitismo para preservar la mejor solución temporal y se ha utilizado cruce multipunto con selección basada en ruleta. Las probabilidades de emparejamiento y mutación han sido de 1 y 0,001, respectivamente. El algoritmo encuentra una solución óptima con una probabilidad del 99 % en dos horas de procesamiento utilizando dos MB de RAM en un pentium IV a 3,2 GHz.

Apéndice C

Segmentación automática

Para realizar el seguimiento y aprendizaje del modelo visual presentado en el apartado 2.2 se necesitan especificar las regiones de la primera imagen de la secuencia que contienen los elementos faciales de interés. Este proceso se conoce en este trabajo con el nombre de segmentación de máscaras. Por otro lado, para construir la correspondencia identificador-visema (CIV) (ver apdo. 2.2.5) se requiere localizar la posición temporal exacta en la que ocurren los diferentes alófonos pronunciados por el usuario en el corpus audiovisual. Este segundo proceso recibe el nombre de segmentación de voz en este trabajo.

La segmentación de las máscaras en el canal de vídeo y de la voz en el canal auditivo se puede realizar manual o automáticamente. La primera exige un esfuerzo de interacción al usuario que no tiene la segunda, aunque posee resultados más exactos.

La segmentación automática de máscaras consta de un proceso propio basado en diferenciación de imágenes y sus proyecciones. Por otro lado, la segmentación automática de voz se basa en unas herramientas desarrolladas por la Universidad de Cambridge Young et al. (2003) basadas en la construcción de modelos ocultos de Markov (HMM).

C.1. Segmentación automática de máscaras

Esta segmentación se realiza mediante un proceso propio basado en diferencias de imágenes y análisis de proyecciones. El algoritmo propuesto para la extracción automática de máscaras obtiene cuatro máscaras a partir de la imagen facial del actor: una cuadrada para cada ojo, otra que contiene la boca, el maxilar inferior y la zona observada de tronco superior y hombros, y una máscara final que recoge el resto de la zona facial (desde la boca hacia arriba exceptuando las regiones oculares). Para conseguirlo, se propone un esquema de interacción (apdo. C.1.1), que tiene unas ciertas restricciones (apdo. C.1.2), y lleva asociado un procesamiento concreto (apdo. C.1.3).

C.1.1. Esquema de interacción

La segmentación automática se ayuda de una interacción sencilla con el actor, el cual debe:

Paso 1: No estar presente en el inicio de la grabación.

Paso 2: Aparecer y quedarse quieto mirando a la cámara.

Paso 3: Parpadear primero y después pronunciar su nombre (o cualquier otra cosa).

La primera acción sirve para obtener una imagen del fondo que aparece en la secuencia con el fin de facilitar la extracción de la máscara global utilizando diferencias de imágenes. La segunda acción se utiliza para conocer la imagen desde la cual extraer las futuras máscaras y obtener la máscara general (ver fig. C.1). El tercer paso sirve para diferenciar los elementos faciales correspondientes a los ojos, la boca y el resto de la cara (ver fig. C.1). Una vez concluidas estas tres acciones, el actor puede empezar a gravar la secuencia de palabras y frases para generar el corpus audiovisual como se especifica en el apartado 2.1.

C.1.2. Restricciones asociadas

Para que el algoritmo propuesto pueda obtener unos resultados correctos se necesitan cumplir una serie de condiciones en la grabación de la secuencia. Éstas son las siguientes:

- **Inmovilidad de la cámara**. Se impone para no introducir movimiento adicional. El algoritmo supone que lo que se mueve es el objeto al cual extraerle las máscaras. Si se mueve la cámara, se mueve la imagen y el algoritmo interpretará toda la imagen como una gran máscara global.

- **Constancia en el fondo**. Aunque el fondo puede ser aleatorio, éste no debe cambiar para no confundir al algoritmo siguiendo los mismos principios que en la condición anterior.

- **Silencio ambiental**. Dado que el algoritmo necesita saber cuando empieza a hablar el actor, es necesario eliminar todo el ruido de entorno posible, ya que el algoritmo supone que se empieza a hablar cuando la potencia de la señal sonora supera un cierto umbral.

- **Interacción específica con el actor**. Las instrucciones dadas en el apartado C.1.1 ya son en sí mismas una condición para que el algoritmo detallado en el apartado C.1.3 funcione correctamente. En la figura C.1 se pueden observar imágenes correspondientes a los tres pasos comentados en el esquema de interacción.

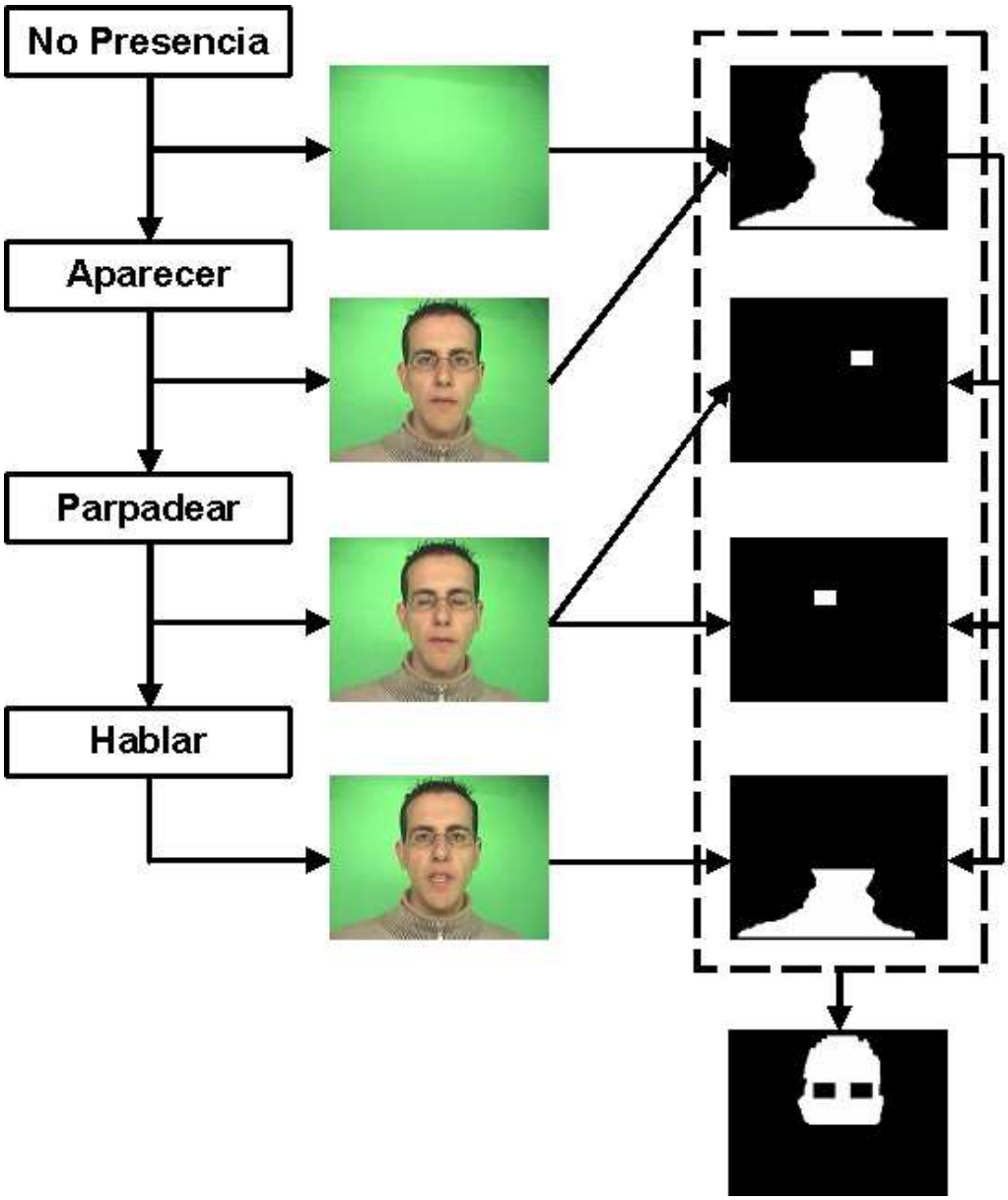


Figura C.1: Pasos en la interacción requerida para extraer las máscaras de una cara (izquierda), mostrando las imágenes ejemplo usadas en cada paso (centro) y las máscaras finalmente obtenidas (derecha).

C.1.3. Algoritmo asociado

En este apartado se detallan las operaciones a realizar en cada uno de los tres pasos de interacción expuestos en el apartado C.1.1.

Paso 1

En este paso se recoge la primera imagen y se utiliza como fondo de referencia, llamándole I^{fondo}. Adicionalmente, se pueden coger las dos o tres primeras y realizar un promediado para reducir el ruido de la imagen de fondo de referencia. La idea base consiste en que al realizar la operación la diferencia entre una imagen cualquiera y la de fondo (C.1) se obtiene una imagen binaria I_m que identifica los píxeles correspondientes a objetos de nueva aparición en la secuencia.

$$I_n^m(x,y) = \begin{cases} 1 & \text{si } \sum_{c=1}^{3} \left| I_n(x,y,c) - I^{fondo}(x,y,c) \right| > \tau \\ 0 & \text{resto} \end{cases} \tag{C.1}$$

donde τ identifica un umbral que se mueve en el margen dinámico de los valores de los píxeles de las imágenes (típicamente de 0 a 255), n representa el número de imagen dentro de la secuencia y el índice c del sumatorio identifica el número de canal, siendo 1 para el rojo, 2 para el verde y 3 para el azul. Dado que $I^{fondo} = I_1$, siempre se cumplirá que $I_1^m(x,y) = 0$ para todo x e y.

Paso 2

El resultado final que busca este paso es el índice de imagen a partir de la cual se extrae la máscara global de toda la cara y a partir de la cual se obtienen las máscaras finales (que se realiza en el siguiente paso). El índice resultado de este paso corresponde al número de imagen de la secuencia que contiene la cara quieta, después de aparecer y quedarse mirando a la cámara. En este proceso se utiliza una medida de diferencia similar a la que aparece en la expresión (C.1):

$$d(n,m) = \sum_{\forall x,y} \sum_{c=1}^{3} |I_n(x,y,c) - I_m(x,y,c)|$$

Este proceso se basa en el hecho de que la cara no puede aparecer instantáneamente en la región central de la imagen, sino que debe aparecer por algún lado e ir trasladándose progresivamente hasta el centro de la misma, momento en el que se quedará quieta. Se define Qe como el número de imagen en la que empieza a aparecer el actor y Qs como el número de imagen en la que éste se queda quieto. Es de esperar que $d(n-1,n)$ sea muy pequeño para $n < Qe$ y $n > Qs$. Si se construye el siguiente vector:

$$v(n) = \begin{cases} 1 & d(n-1,n) > \rho \\ 0 & \text{resto} \end{cases}$$

donde ρ representa un valor de umbral determinado, los valores de v serán de 0 para $n < Qe$ y $n > Qs$ y 1 para el resto. Este hecho se debe a que mientras el actor se desplaza por la región de la imagen, las diferencias entre imágenes consecutivas serán apreciables. Cuando el actor no ha aparecido aún, o se encuentra ya quieto, estas diferencias son muy pequeñas comparadas con las anteriores. Aumentar el valor de ρ implica relajar la condición de entrada y salida, al permitir más error entre imágenes consecutivas; reducir el valor de ρ implica lo contrario. Dado que d es mayor para $n > Qs$ que para $n < Qe$ (debido a que una persona es incapaz de estar completamente quieta) reducir demasiado el valor de ρ puede provocar que nunca se considere que la persona esté quieta o, incluso, que entra a partir de la segunda imagen (debido al ruido presente en la imagen).

Una vez obtenido el número de imagen Qs, en la que el actor se ha quedado quieto, se realiza una diferencia entre ésta (I_{Qs}) y la imagen de fondo, binarizando el resultado (C.1). Aplicando operadores morfológicos, se puede llegar a obtener la máscara global (ver fig. C.1), que se denominará $\mathbf{\Pi}$.

Paso 3

En este paso se obtienen las máscaras correspondientes a los ojos, boca y cabeza. La información para realizar esta segmentación se encuentra en las imágenes desde Qs hasta la que contiene el final de la pronunciación del nombre el actor, conocida como la imagen Qa. En las imágenes desde Qs hasta Qa se supone que el actor está quieto, parpadea y pronuncia su nombre. De este modo, dentro de la región delimitada por la máscara global $\mathbf{\Pi}$ habrá dos únicos movimientos: los ojos y la boca. Aprovechando la información de ambos movimientos se pueden establecer tres puntos de corte verticales para encontrar las regiones de ojos y boca (ver fig. C.2):

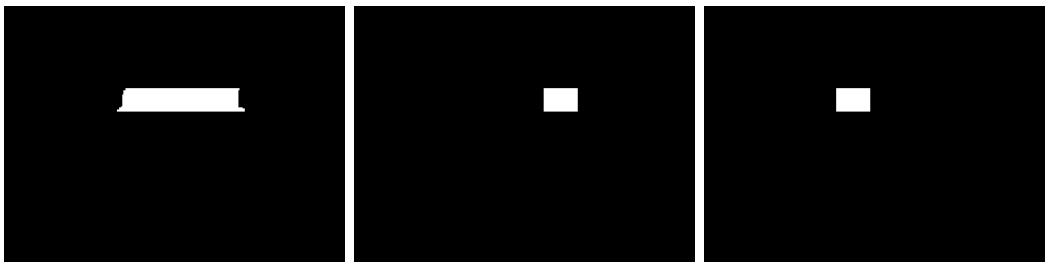
- Límite superior de los ojos
- Límite inferior de los ojos
- Límite superior de la boca

No se establece un límite inferior para la boca debido al deseo de incorporar en una misma región la boca, la barbilla y el cuello.

Para encontrar los tres puntos de corte anteriores se analiza la varianza de las proyecciones de ángulo $\pi/2$ (suma de los valores de las filas en una imagen) a lo largo de todas las imágenes entre Qs y Qa. Aquellas filas que posean mayor varianza de valores serán las que presenten movimientos. La aplicación de filtros de mediana en el análisis de las varianzas puede ayudar en la detección de estos límites. Una vez detectados, se puede realizar el mismo proceso para la región delimitada por los ojos y la máscara global desde la imagen Qs hasta la Qa para encontrar los puntos de corte horizontales delimitadores de los dos ojos y así extraer las máscaras correspondientes (ver fig. C.3).

(a) Máscara global (b) Región de los ojos (c) Región de la boca

Figura C.2: Aplicación de los límites verticales encontrados en el paso 3 para extraer la región de los ojos (b) y de la boca (c) a partir de la máscara global (a).



(a) Región de los ojos (b) Ojo izquierdo (c) Ojo derecho

Figura C.3: Aplicación de los límites horizontales encontrados en el paso 3 para extraer la región del ojo izquierdo (b) y del derecho (c) a partir de la región de los ojos (a).

La máscara de la parte superior de la cabeza se puede obtener restando las máscaras de los ojos (fig. C.3(b) y fig. C.3(c)) y la de la boca (fig. C.2(c)) a la máscara global (fig. C.2(a)).

C.2. Segmentación automática de voz

La segmentación automática de voz se realiza utilizando el paquete de *software* de libre distribución llamado *HTK* Young et al. (2003) y desarrollado en la *Universidad de Cambridge*. Para su correcto funcionamiento, necesita la siguiente información:

- El canal de audio a segmentar
- La descripción de la secuencia de alófonos pronunciados
- Los modelos ocultos de Markov (HMM) asociados a la voz utilizada

El canal de audio debe estar muestreado a 16KHz y 16 bits por muestra para su correcto funcionamiento. Este hecho puede implicar un cambio de la frecuencia de muestreo según sean las condiciones de grabación del corpus audiovisual.

Aunque el paquete *HTK* es capaz de segmentar un canal de audio solamente con la información de forma de onda, la precisión de los resultados aumenta considerablemente si se adjunta la transcripción fonética, o serie de alófonos pronunciados, asociada. Dadas las características del corpus audiovisual presentado en el apartado 2.1, la transcripción fonética es siempre la misma: los actores están forzados a decir el mismo conjunto de palabras y frases.

El paquete *HTK* también necesita los HMM que rigen su comportamiento segmentador. Estos modelos se deben obtener mediante un entrenamiento con señales de voz previamente segmentada manualmente. El caso óptimo consiste en utilizar los HMM construidos a partir de la voz a segmentar, pero como se dispone de la trascripción fonética, si se utilizan otros que describan una voz parecida también se obtiene un buen resultado. En este trabajo se han aprovechado unos HMM ya existentes en la *Sección de Teoría de la Señal* que fueron desarrollados por el *Área de Tecnologías del Habla* para otros estudios. No obstante, estos modelos son de voz masculina, con lo que no funcionarán correctamente con voces femeninas.

El comportamiento presentado por el *HTK* va en función de las dos características siguientes:

- El margen dinámico
- La relación señal-ruido

Cuanto más se aproveche el margen dinámico sin producir saturación y mayor sea la relación señal-ruido, mayor probabilidad de conseguir una segmentación correcta por parte del *HTK*.

Los resultados presentados por el *HTK* son marcas de inicio y fin de alófono en unidades de 100ns. A partir de esta información se puede extraer el centro de cada uno y determinar la imagen asociada de la secuencia audiovisual correspondiente.

Apéndice D

Aplicaciones desarrolladas

El último apartado de este capítulo está dedicado a las aplicaciones desarrolladas para mostrar las capacidades de los algoritmos desarrollados en este trabajo en los capítulos 3 y 4 a través de su uso en caras parlantes. Estas aplicaciones comparten el mismo nombre de *PREVIS II* (Person specific REalistic VIrtual Speaker) debido a que el de *PREVIS* o *PREVIS I* identifican una serie de aplicaciones asociadas a un sistema de locución virtual Melenchón et al. (2002a) anterior al presentado, que ha servido de base para el desarrollo de esta tesis.

D.1. Librería *PREVIS II*

Se ha desarrollado un conjunto de librerías de enlace dinámico (algunas de ellas independientes de plataforma) con el objetivo de simplificar el posterior desarrollo de aplicaciones que implementen el esquema de síntesis audiovisual descrito aquí. Las librerías desarrolladas están interrelacionadas entre ellas (ver fig. D.1) y son las siguientes:

- *slwin32.dll*: Encapsula la conversión de texto a voz sintética (TTS) utilizando las librerías SAPI® de Microsoft®.
- *alwin32.dll*: Encapsula la generación de secuencias audiovisuales en diferentes formatos.
- *coartlib.dll*: Encapsula el algoritmo de interpolación no lineal de alta dimensionalidad presentado en el apartado 4.1.3.1.
- *seglib.dll*: Encapsula el algoritmo de seguimiento explicado en el apartado 3.1 con las características de subespacio como referencia, subespacios modulares y multirresolución.
- *svdlib.dll*: Encapsula el algoritmo de aprendizaje utilizando el cómputo incremental de la SVD con actualización de la media (ver apartado 3.2.4).

- *locutlib.dll*: Encapsula funcionalidades de alto nivel para guiar la síntesis a través de las librerías anteriores (funciones de CIV y RVV).

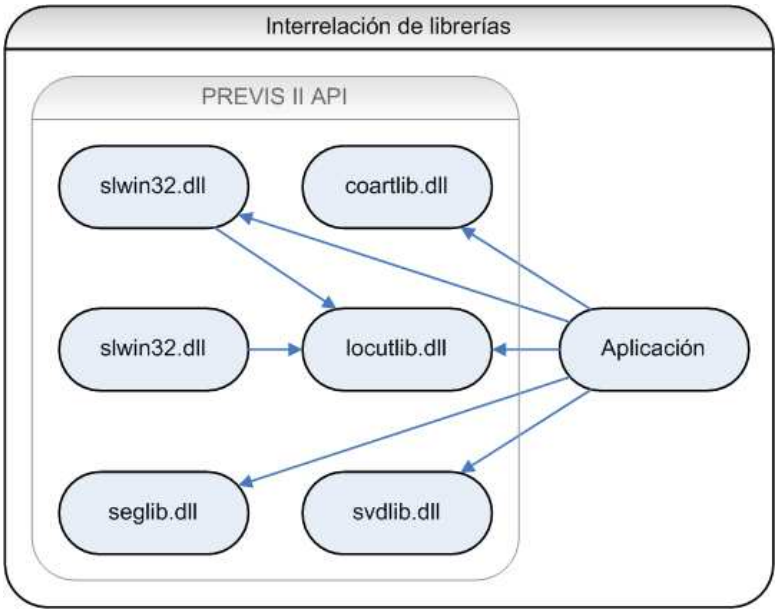
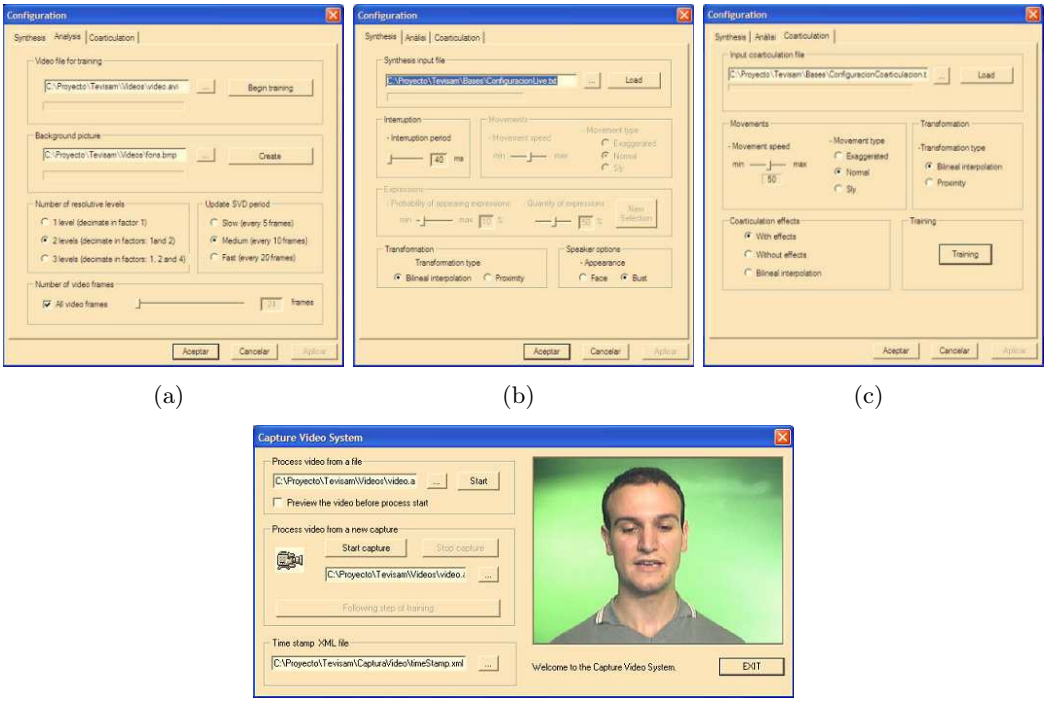


Figura D.1: Interrelación entre las diferentes librerías de enlace dinámico desarrolladas para crear aplicaciones de síntesis audiovisual de caras parlantes.

D.2. *Desktop PREVIS II*

Se ha desarrollado un ejecutable que incluye la mayor parte de las funcionalidades del esquema de síntesis audiovisual propuesto aplicándolo a caras parlantes (incluye todas las prestaciones excepto la síntesis por selección de visemas). El ejecutable ha padecido una importante evolución a partir de una primera versión muy primitiva realizada en el año 2001 Melenchón et al. (2002a), que se basaba en un modelo desfasado, utilizaba únicamente niveles de gris y no poseía ningún tipo de proceso de personalizaciónpersonalizable automatizado. Esta aplicación ha ido plasmando a lo largo del tiempo los progresos realizados en el presente trabajo de investigación. Cuando se implementó por primera vez el modelo visual propuesto en este trabajo con una cara humana, la aplicación adoptó inicialmente el nombre de *TEVISAM*, aunque se decidió no cambiarlo para no olvidar la aplicación original de la cual provenía: *PREVIS*.

En la figura D.2 se pueden observar diferentes capturas de pantalla pertenecientes a las diferentes funcionalidades de los distintos módulos que componen la aplicación *Desktop PREVIS II*: *i*) registro de corpus audiovisual; *ii*) construcción de modelos faciales, mediante el algoritmo de seguimiento y aprendizaje simultáneos; *iii*) síntesis de cabezas parlantes en tiempo real guiadas por conversores TTS; *iv*) generación de vídeos con cabezas parlantes incluyendo efectos de coarticulación visual.



(a) (b) (c)

(d)

Figura D.2: Diferentes capturas de la aplicación correspondientes a los diferentes módulos implementados: (a) análisis; (b) síntesis; (c) generación de efectos de coarticulación; y (d) registro de corpus audiovisual.

Esta aplicación utiliza la librería *PREVIS II* descrita en el apartado D.1 y es capaz de generar modelos visuales de la cara de una persona de forma totalmente automática mediante la segmentación de máscaras y voz comentada en el apéndice C, aunque obtienen una precisión menor que con una segmentación manual. Se permiten también diferentes puntos de entrada y salida de información a fichero, para no tener que realizar todos los pasos previos según se desee crear una nueva cara parlante desde cero, refinar una existente o crear nuevas secuencias a partir de un modelo disponible.

En la figura D.3 se puede observar la síntesis de la palabra catalana *camió* con un modelo generado a partir de un corpus audiovisual muy reducido. Para aumentar la sensación de realismo, se añaden movimientos aleatorios para evitar que el personaje se quede quieto Maestri (1996) y se añaden cuello y hombros para evitar el efecto fantasmagórico que produce un cabeza flotante.

D.3. *PREVIS II* en línea

Utilizando las librerías del apartado D.1, también se ha desarrollado una aplicación demostrativa por Internet para la generación de caras parlantes (ver figura D.4), aunque sólo para el idioma inglés. La demostración es accesible desde las siguientes direcciones

Figura D.3: Imágenes de la secuencia audiovisual que contiene la síntesis de la palabra catalana *camió* producida por la aplicación *Desktop PREVIS II*.

web en el momento de impresión de esta memoria de tesis:

- `http://cepheus.salleurl.edu/www/formulari.html`
- `http://www.salleurl.edu/~jmelen/demovtts.html`
- `http://www.salleurl.edu/eng/elsDCTS/tsenyal/english/tsenyal_previsOnLine.htm`

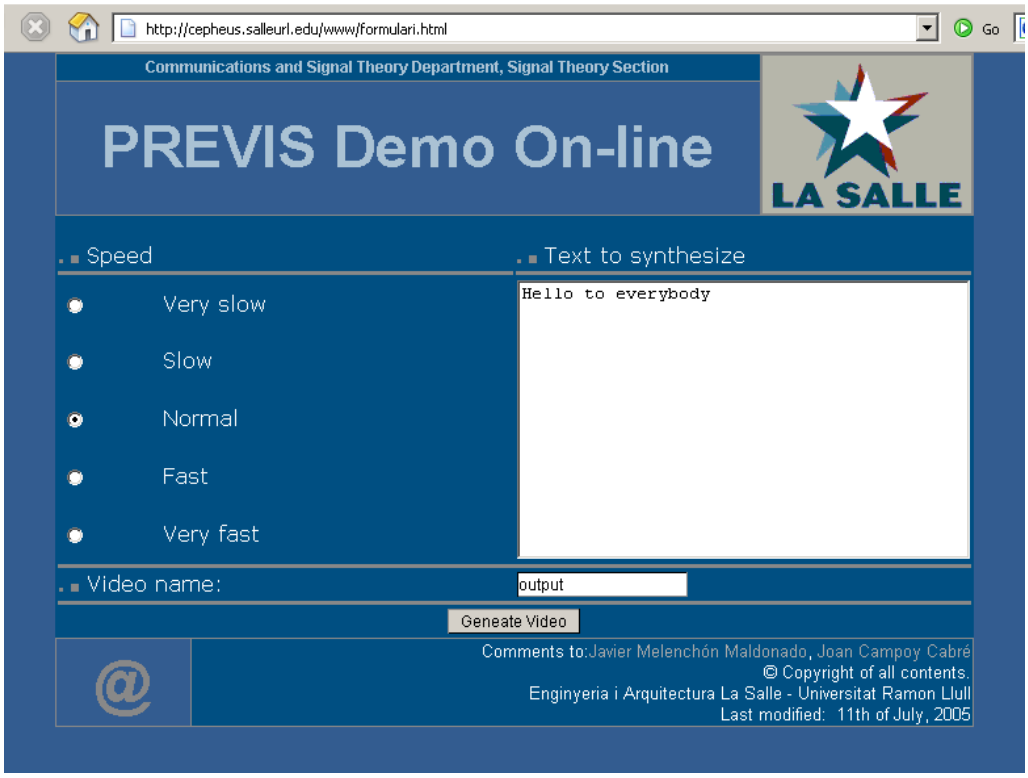


Figura D.4: Página de presentación de la aplicación *PREVIS II* en línea accesible desde `http://cepheus.salleurl.edu/www/formulari.html`

Esta aplicación se ha implementado como un servicio de Microsoft® Windows® con el objetivo de evitar su inicialización manual. Siguiendo con esta filosofía de actuación, también se ha implementado una rutina que reinicia el servicio cuando éste falla sin importar el motivo.

Universitat Ramon Llull

Aquesta Tesi Doctoral ha estat defensada el dia ____ d _____ de 2007

al Centre _____

de la Universitat Ramon Llull

davant el Tribunal format pels Doctors sotasignants, havent obtingut la qualificació:

President/a

_____

Vocal

_____

Vocal

_____

Vocal

_____

Secretari/ària

_____

Doctorand/a
Javier Melenchón Maldonado

C. Claravall, 1-3
08022 Barcelona
Tel. 936 022 200
Fax 936 022 249
E-mail: urlsc@sec.url.es
www.url.es