MPEG-4: UNA NUEVA REPRESENTACIÓN DE LA INFORMACIÓN AUDIO-VISUAL

Josep Ramon Casas, Ferran Marqués y Philippe Salembier

Este es el primer artículo de una serie en la cual pretendemos dar una visión de cuál es el trabajo que en el entorno de MPEG-4 se está realizando en el Grup de Processament d'Imatge del Departament de Teoria del Senyal i Comunicacions de la Universitat Politécnica de Catalunya. En este primer artículo, presentamos qué es MPEG-4 y cuáles son sus objetivos principales, así como la situación de desarrollo actual del estándar. En futuros artículos comentaremos más detalladamente cómo se enmarca nuestro trabajo en este ámbito.

1. Introducción

La extensión de los servicios de comunicaciones, informáticos y telemáticos en las Últimas décadas ha alcanzado un protagonismo relevante en la mayoría de los ámbitos profesionales. La creciente disponibilidad de potentes ordenadores personales y de canales de comunicación avanzados sugiere la visión de un mundo en el que cualquier tipo de información fluye libremente entre una variedad de sistemas diferentes. No obstante, a pesar del espectacular progreso de los dispositivos de almacenamiento masivo y de las prestaciones de los sistemas digitales de comunicación de datos, la demanda de mayor ancho de banda de transmisión y de mayor capacidad de almacenamiento continúa superando las posibilidades de las tecnologías disponibles.

JOSEP RAMON CASAS, FERRAN MARQUÉS y PHILIPPE SALEMBIER, *son profesores del grupo de Imagen del T.S.C en la ETSETB*

El desarrollo de aplicaciones que realizan un empleo exhaustivo de la capacidad de datos de los sistemas actuales, como las relacionadas con sistemas de audio y video en tiempo real, y la utilización creciente de canales de ancho de banda limitado, como radioenlaces terrestres y satélites de comunicaciones, no solamente ha mantenido la necesidad de encontrar sistemas más eficientes de representar la información sino que han hecho de los sistemas de codificación y compresión de datos un aspecto esencial de la tecnología digital de comunicaciones y almacenamiento de datos.

• *Sistemas de comunicación que incluyen información visual*

En particular, el papel fundamental que desempeñan las señales visuales en nuestro entorno cultural está siendo integrado en este nuevo mundo de las tecnologías de la información. Las aplicaciones de imagen y video digital requieren elevadas velocidades de transmisión, enormes capacidades de almacenamiento y equipos rápidos de tratamiento de estas señales (si las imágenes se manipulan en su forma original). Imágenes típicas de televisión digital, por ejemplo, generan velocidades de transmisión y tratamiento que exceden los 100 Mbits/s.

La emergencia de nuevos sistemas de comunicación visual plantea el problema de cómo comprimir esta vasta cantidad de información en soportes de capacidad limitada, ya sea para la transmisión o para el almacenamiento. Como ejemplo, se pueden citar sistemas de comunicación visual en canales desde 20 Mbit/s [1] hasta 64 kbit/s [2].

Merece especial atención la creciente utilización de imágenes en sistemas de comunicación de muy baja capacidad de datos. En las redes internacionales de comunicación entre ordenadores, especialmente en Internet, cualquier usuario de un ordenador personal abonado a un servicio telemático por vía telefónica y provisto de una tarjeta de conexión a la red (un modem), puede acceder a innumerables fuentes de información de todo tipo, publicaciones electrónicas que se editan con regularidad, servicios de mensajería electrónica (e-mail) y hasta servicios de comunicación interactiva mediante voz e imagen en tiempo real [3, 4 y 5].

El contenido en información visual de estos sistemas de comunicación de datos aumenta sin cesar. Originalmente, con la posibilidad de acceder a imágenes fijas en publicaciones electrónicas. Después, mediante programas de libre utilización que permiten visualizar secuencias de imágenes en movimiento en el propio ordenador personal a partir de información contenida en una base de datos remota. Dichas secuencias contienen a menudo gráficos e imágenes generadas artificialmente junto con imágenes adquiridas de escenas naturales.

La comunidad internacional se esfuerza por estandarizar a tiempo técnicas de comunicación mediante sistemas de codificación de señales de audio y video que racionalicen y aumenten la eficacia de los sistemas empleados en la práctica [6, 7, 8 y 9]. Dichas técnicas deberían permitir el intercambio *universal* de la información entre dos sistemas cualesquiera conectados a una red de comunicación.

• *Representación eficiente y manipulación de contenidos*

En el actual camino hacia nuevos estándar de codificación, se contempla precisamente la convergencia de los ámbitos de video digital, gráficos por computador y sistemas de animación de imágenes sintéticas [10]. Para la representación eficiente (codificación) y la manipulación de las señales que intervienen en los sistemas de comunicación visual se requiere el empleo de técnicas avanzadas de análisis capaces de acceder a los objetos contenidos en las escenas [11]. La representación de la información audiovisual orientada a los objetos permite al usuario del sistema de codificación combinar y manipular de un modo flexible y eficiente «objetos» audiovisuales, representados al elevado nivel de abstracción requerido por las aplicaciones de interés.

2. ¿Qué es MPEG?

MPEG son las iniciales de *Motion Pictures Experts Group*, un grupo de trabajo dependiente de la ISO (Organización Internacional de Estándars) encargado del estudio y desarrollo de técnicas estándar de comunicación visual. La existencia de un estándar es una necesidad evidente en cualquier sistema de comunicación. Si no existe una norma común en la que se hayan puesto de acuerdo previamente los diferentes actores de un sistema de comunicación, difícilmente podrán intercambiar información de modo inteligible. La norma común (el estándar) es más importante, si cabe, cuando el sistema de codificación tiene cierto grado de complejidad (como es el caso de los sistemas de comunicación de imagen). En dicho caso, puede ser muy costoso disponer de decodificadores de tipo *multisistema*.

El desarrollo de aplicaciones como las relacionadas con sistemas de audio y video en tiempo real,... han hecho de los sistemas de codificación y compresión de datos un aspecto esencial de la tecnología digital de comunicaciones y almacenamiento de datos.

Los estudios en el grupo MPEG dieron lugar a la primera propuesta provisional (*draft*) para un estándar de codificación de video digital (MPEG-1) en 1988. En aquel momento se denominó simplemente MPEG, para diferenciarlo de la norma JPEG de codificación de imágenes fijas. MPEG-1 se basó en la norma de codificación para imágenes de videoteléfono y videoconferencia H.261 del CCITT, cuyos primeros borradores se habían escrito cuatro años antes, en 1984. Ambas normas tienen importantes similaridades, lo cual no es de extrañar dado que una parte importante de los investigadores de MPEG también habían trabajado con el CCITT en la elaboración de H.261. La diferencia fundamental entre H.261 y MPEG-1 se encuentra en las aplicaciones para las que se desarrollaron ambas. A continuación se ofrece una breve descripción de los objetivos de éstas y de las propuestas posteriores:

• *CCITT-H.261: el estándar para videotelefonía y videoconferencia digital*

La norma H.261 se desarrolló pensando en aplicaciones de videoconferencia. Las características particulares de esta aplicación, como son: la necesidad de realizar la codificación y decodificación en tiempo real, las secuencias de imágenes con poco movimiento según un modelo predeterminado (primeros planos tipo cabeza y hombros con un fondo simple y estático), etc. determinaron la especificidad del sistema de codificación. Por ejemplo, el fondo sobre el que se encuentra el interlocutor habitualmente no varía con el tiempo (o varía muy poco), por lo tanto, resulta redundante transmitir con precisión esta información en todas las imágenes de la secuencia.

• *MPEG-1: películas digitales para video doméstico*

MPEG-1 se orientó a un conjunto más genérico de secuencias de imágenes. La intención original fue la grabación de películas de video en Compact Disc para uso doméstico. La restricción fundamental era la velocidad máxima admisible por este soporte digital: 1,5 Mbit/s. Otras carac-

terísticas importantes relacionadas con esta aplicación son que no se requiere que el proceso de codificación sea en tiempo real (lógicamente, sí se requiere en el proceso de decodificación y presentación de la secuencia almacenada en el soporte) y que la calidad exigible será del orden (o presumiblemente superior) a la de las grabaciones analógicas en cintas de video VHS. Los sistemas de codificación basados en la norma MPEG-1 están en el mercado desde hace varios años; concretamente desde 1990, fecha en que se determinó la versión actual de la norma [7].

Así una secuencia audiovisual será descrita en términos de los distintos objetos que la componen y de la evolución temporal de éstos

• *MPEG-2: la norma de televisión digital*

Pronto se vio que la aplicación a la que estaba orientada MPEG-1 era demasiado restringida. La explosión de los sistemas interactivos de comunicación de audio y video digital dejaba obsoletas las especificaciones del sistema de codificación para soporte en CD definidas tan sólo dos años antes. En 1990, mientras se terminaba la definición de la norma MPEG-1, nacían dos nuevas propuestas, esta vez más ambiciosas, para el desarrollo de una norma común de codificación de video digital: MPEG-2 y MPEG-3. La intención era la estadarización de la norma definitiva de televisión digital.

MPEG-2 aumentaba la velocidad de salida del codificador hasta 10 Mbit/s, permitiendo la codificación de imágenes y sonido con calidad de radiodifusión de televisión (significativamente mayor que la de video doméstico). En 1993 se publicaban las primeras versiones [8] que, con pocas variaciones, resultarían en el estándar definitivo.

• *MPEG-3: una intención frustrada*

MPEG-3 se estableció en 1990 como un grupo de trabajo paralelo a MPEG-2. Su objetivo era definir un estándar de codificación de imagen y video digital para velocidades de salida del codificador hasta 60 Mbit/s. En un principio parecían velocidades adecuadas para la calidad requerida en los futuros sistemas de televisión de alta definición.

Llegados a este punto es preciso resaltar que en el sector de la electrónica de consumo existían desde hacía casi una década (desde 1982, concretamente), importantes luchas por adquirir posiciones de privilegio en lo que se preveía que iba a ser un impulso mucho más importante para el sector de la electrónica de consumo que la aparición de los sistemas de televisión en color o del CD: la televisión de alta definición. En las discusiones sobre el tipo de sistema a estandarizar habían intervenido desde las mismas empresas, hasta los gobiernos de distintos países con intereses en el sector, productoras de cine, entidades científicas y culturales, etc. No nos extenderemos en este tema puesto que ha generado interminables discusiones hasta la fecha. El lector interesado podrá encontrar en [12], por ejemplo, una extensa descripción desde el punto de vista histórico.

Pero la tecnología superó las expectativas. El sistema de codificación que se estaba desarrollando para MPEG-2 era lo bastante genérico como para incluir calidades de imagen de alta definición, incluso a velocidades de tan sólo 20 Mbit/s. En este sentido, la propuesta americana para televisión de alta definición [1] enfatiza la compatibilidad con MPEG-2. Poco después de su creación, el grupo de trabajo para televisión de alta definición, MPEG-3, fue absorbido por MPEG-2.

• *MPEG-4: y ahora, ¿qué?*

De nuevo la evolución de los sistemas de comunicación audiovisual están dejando obsoletos los sistemas de codificación de video recientemente definidos. En este artículo pretendemos, precisamente, describir el sistema de codificación MPEG-4. La clave del progreso se encuentra, en este caso, en lo que MPEG-4 ha venido en llamar las *funcionalidades* del sistema de codificación.

• *MPEG: comunicación audiovisual*

Para concluir este apartado, y antes de describir MPEG-4, resulta interesante realizar la siguiente observación: aunque normalmente se suele relacionar el trabajo de MPEG con la estandarización en el mundo del video, se debe destacar que MPEG ha desarrollado también importantes trabajos en el ámbito del audio. De esta manera MPEG ha recibido el premio Emmy 96 por su trabajo en codificación realizado en los estándares MPEG-1 y MPEG-2.

3. Objetivos de MPEG-4

El trabajo inicial de MPEG-4 se enfocó sobre la idea de generar un espacio común entre los mundos, históricamente separados, de las telecomunicaciones, la industria del cine y los ordenadores. Las discusiones generadas del contacto de estas tres comunidades han clarificado el significado real de esta primera idea. Así, la definición de los objetivos de MPEG-4 ha ido perfilándose mediante continuas variaciones

introducidas por las distintas necesidades y las diferentes capacidades de cada uno de estos mundos. En este sentido, MPEG-4 suele autodenominarse *a moving target*. En estos momentos, el proyecto de MPEG-4 es mucho más ambicioso que la idea inicial: MPEG-4 persigue la creación de una representación de la información audio-visual que se acerque a la percepción natural que el ser humano tiene de una escena. Se busca una descripción basada, no ya en las distintas imágenes (o bloques de imagen) y señales de audio que componen la secuencia audiovisual, sino en los *objetos* contenidos en la escena.

Figura 1.-Ejemplo de composición de VOs obtenidos separadamente

Una descripción basada en el contenido (*content-based description*) ha de permitir la interactuación sobre los distintos objetos de la secuencia, tanto si provienen de escenas reales como si han sido creados sintéticamente. Así, una secuencia audiovisual será descrita en términos de los distintos objetos que la componen y de la evolución temporal de éstos. El conjunto de esta información para cada objeto se denomina VO (del inglés *Video Object*). En la Figura 1 se muestra un ejemplo de escena sencilla que se puede describir mediante un par de VOs: la presentadora a lo largo de la secuencia y el fondo estático. En este ejemplo, los objetos se han filmado separadamente y, por tanto, se tiene directamente los dos VOs. Por supuesto, se puede suponer una o varias señales de audio independientes para cada VO.

La división de la secuencia en VOs permite definir un conjunto de funcionalidades nuevas en un sistema de codificación. El concepto de *funcionalidad* es una de las aportaciones más novedosas de MPEG-4. Entre las distintas funcionalidades basadas en el contenido, se puede destacar las siguientes:

MPEG-4 persigue la creación de una representación de la información audio-visual que se acerque a la percepción natural que el ser humano tiene de una escena

• *Interacción basada en el contenido:*

1. Acceso a bases de datos de video a partir de su contenido: indexación de las secuencias de video en base a conceptos semánticos. Ejemplo de aplicación: búsqueda de personajes específicos o selección automática de imágenes por temas (noticiarios, deportes, etc).

2. Manipulación del tren de datos para la edición de secuencias: eliminación o cambio de partes del tren de datos que de lugar a cambios de los objetos representados en la escena. Ejemplo de aplicación: inserción de subtítulos o publicidad específica.

3. Codificación híbrida de secuencias con imágenes naturales y sintéticas: combinación de datos provenientes de fuentes naturales o sintéticas mediante herramientas de codificación específicas para cada tipo de dato. Ejemplo de aplicación: creación de vídeojuegos o inserción de gráficos en la escena.

• *Compresión:*

1. Mejora de la eficiencia de codificación: codificación selectiva de unos objetos frente a otros, ya sea variando el número de imágenes que se codifica de cada objeto o la calidad espacial que se requiere para cada uno. Ejemplo de aplicación: transmisión reduciendo la calidad del fondo con respecto al primer plano (deportes, videotelefonía) o de un detalle de la imagen con respecto al resto (telemedicina).

2. Codificación de múltiples fuentes de datos concurrentes: codificación de distintas vis-

tas de la escena aprovechando la redundancia que existe entre ellas, así como el hecho de disponer de modelos tridimensionales de los objetos presentes en ella. Ejemplo de aplicación: codificación de secuencias tridimensionales para aplicaciones de realidad virtual.

• *Acceso Universal:*

1. Robustez frente a canales con presencia de ruido: protección selectiva de unos objetos presentes en la escena frente a otros. Ejemplo de aplicación: videotelefonía móvil.

2. Escalabilidad basada en el contenido: capacidad de crear dos trenes de datos representando la misma información a distintos niveles de calidad de tal manera que el primer tren dé una calidad media mientras que el segundo tren, ofrezca un incremento de calidad al añadirlo al primero. Ejemplo de aplicación: transmisión por el mismo canal a receptores con distintas velocidades.

Además de las funcionalidades de codificación comentadas anteriormente, MPEG-4 persigue una idea básica que es la posibilidad de reutilizar los datos extraidos de una secuencia en otras escenas futuras. De esta manera, además de la capacidad de editar las secuencias al nivel de tren de datos, también se podrá editar al nivel de imagen haciendo composición de los distintos VOs que se han generado o detectado en secuencias previas.

Este concepto de reutilización se comenta en el siguiente ejemplo. En la Figura 2, se presenta un caso de secuencia más complicado que el de la Figura 1 ya que los distintos objetos no se han filmado separadamente. En este caso, se han definido manualmente tres VOs: el conjunto de presentadores, el fondo y la escena en el monitor. Al conseguir aislar uno de estos VOs, se puede componer una nueva secuencia utilizando imágenes o VOs provenientes de otras escenas. Esto proceso (generación de VO y combinación posterior), que normalmente se realiza o bien manualmente o bien mediante grabación separada, se pretende realizar ahora de manera automática.

Figura 2.-Generación manual de VOs y combinación para crear una nueva secuencia

MPEG-4 pretende extender el concepto de reutilización de objetos de un modo parecido a los módulos de los algoritmos de codificación y decodificación. Con esto, MPEG-4 busca generar algo más que un estándar de codificación que puede quedarse obsoleto rápidamente. Lo que pretende es dar lugar a un estándar que sea flexible y así introducir posibles mejoras técnicas que aparezcan en el futuro. Para ello se ha desarrollado el denominado MSDL (del inglés MPEG-4 Syntatic Description Language): un lenguaje de alto nivel que permitirá la descripción de algoritmos completos de codificación, decodificación y composición de forma modular.

Mediante este lenguaje se podrán introducir nuevos herramientas directamente en el estándar. Un ejemplo de este tipo de evolución se podría dar en el caso de proponerse una nueva técnica de estimación de movimiento dentro de la codificación. Esta nueva herramienta de estimación de movimiento podría reemplazar a la anterior directamente, haciendo evolucionar de manera suave el estándar. Además, el uso de este lenguaje ha de permitir la creación de nuevos algoritmos mediante la conexión de distintas herramientas. De esta manera, un sistema transmisor puede seleccionar el algoritmo que desee (siempre dentro de lo permitido por el estándar) para codificar los datos audiovisuales a transmitir y ponerse en contacto con el receptor, testear si el receptor tiene todo el conjunto de herramientas de decodificación necesarias para construir el algoritmo seleccionado y, en caso contrario, enviar los módulos que falten. De esta manera, el receptor puede «aprender» nuevas herramientas y configurar nuevos decodificadores.

4. Estado actual de MPEG-4

Las distintas partes que formarán el estándar de codifica-

ción MPEG-4 están en distintas fases de evolución. El algoritmo de codificación de video ha sido desarrollado a lo largo de este año a base de introducir mejoras mediante experimentos (Core Experiments) propuestos sobre un algoritmo de base (Verification Model). En este mes (Noviembre 1996), tendrá lugar una nueva reunión de MPEG en la cual se fijará el primer algoritmo de trabajo oficial (Working Draft) y se restringirá el número de experimentos de mejora a realizar. Por su parte, el algoritmo básico que debe generar el estándar de audio se debe establecer también en esta reunión, así como la primera propuesta completa de MSDL.

La parte del estándar relacionada con la codificación conjunta de VOs naturales y sintéticos está suscitando un gran interés. Este es un campo actualmente muy abierto, en el cual técnicas

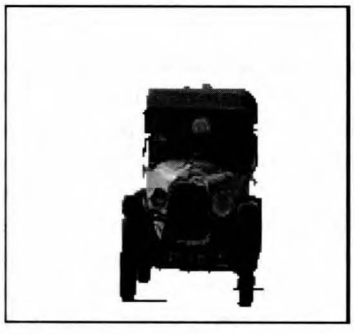

Figura 3.1- Generación automática de VOs

provinientes de ambos ámbitos deben optimizarse conjuntamente para generar un algoritmo común (por ejemplo, métodos de codificación y síntesis de texturas).

5. ¿Qué queda por hacer?

Figura 3.2- Generación automática de VOs

En Noviembre de 1998, el estándar internacional debe estar concluido. Hasta este momento, se debe desarrollar o mejorar todo el conjunto de técnicas de codificación comentadas en el apartado anterior. Sin embargo, es preciso decir que éste no es el único trabajo a realizar para concretar el estándar. Para que, una vez definido el estándar, éste pueda dar servicio hace falta una parte de trabajo adicional muy importante. Este trabajo se refiere a la creación de técnicas de análisis de secuencias que sean capaces de segmentar la escena y su evolución temporal; es decir, crear métodos de generación de VOs.

Las técnicas de análisis han de ser capaces de segmentar una secuencia, mediante interacción humana y/o de forma totalmente automática, en objetos con significado semántico. De esta manera, se deben crear métodos capaces de detectar la presencia de un objeto determinado y extraer su forma y posición correctamente. En los ejemplos presentados en la Figura 3 se muestra el resultado de segmentar dos imágenes distintas con el propósito de hallar objetos diferentes en cada una de ellas.

Referencias

[1] K. Challapali et al. *The Grand Alliance System for US HDTV. Proceedings of the IEEE*, 83(2):158—174, February 1995.

[2] H. Li, A. Lundmark, and R. Forchheimer. *Image sequence coding at very low bit-rates: a review. IEEE Transactions on Image Processing*, 3(5):589—609, September 1994.

[3] S. Casner and S. Deering. *First IETF Internet audiocast. ACM SIGComm Computer Communications*, pages 92—97, July 1992.

[4] D. W. Lin, C.-T. Chen, and T. R. Hsing. *Video on phone lines: technology and applications. Proceedings of the IEEE*, 83(2):175—193, February 1995.

[5] Cornell University. *CU-seeme home page*. Internet address: http://cu-seeme.cornell.edu.

[6] G. K. Wallace. *The JPEG still picture compression standard. Communications of the ACM*, 34(4):30—44, April 1991.

[7] D. Le Gall. *The MPEG video compression algorithm. Signal Processing: Image Communications*, 4:129—140, April 1992.

[8] ISO-IEC CD 13818 Information Technology. *Generic coding of moving pictures and associated audio (MPEG-2). Technical report, Motion Picture Expert Group*, November 1993.

[9] MPEG. *MPEG-4 Proposal Package Description (PPD). Technical Report ISO/IEC JTC1/SC29/WG11, MPEG*, July 1995.

[10] R. Forchheimer and T. Kronander. *Image coding: from waveforms to animation. IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(12):2008—2023, December 1989.

[11] F. Pereira. *MPEG4: a new challenge for the representation of audio-visual information. In Picture Coding Symposium*, pages 7—16, Melbourne VI, March 1996.

[12] M.I. Krivocheev. *The first twenty years of HDTV: 1972-1992. SMPTE Journal*, October 1993.