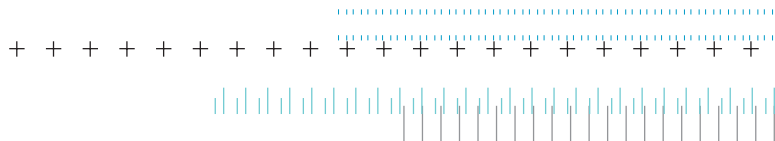


EL PAPEL NOTABLEMENTE RELEVANTE DEL RECONOCIMIENTO DE FORMAS Y LA VISIÓN POR COMPUTADOR EN EL DESARROLLO DE ESTAS TECNOLOGÍAS INTERACTIVAS MULTIMODALES E INTERFACES YA SE PREVIÓ A PRINCIPIOS DE LOS SETENTA, AUNQUE SÓLO UNA PEQUEÑA PARTE DE SU ENORME POTENCIAL HA SIDO EXPLOTADO HASTA LA FECHA



# SISTEMAS DE INTERACCIÓN MULTIMODAL: TÉCNICAS Y APLICACIONES

Elsa Cubel - Gestora de Proyectos del grupo PRHLT del Instituto Tecnológico de Informática

## 1. Retos y oportunidades de la Interacción Multimodal en Reconocimiento de Formas y Visión por Computador

El objetivo tradicional del Reconocimiento de Formas (RF) y la Visión por Computador (VC) es el desarrollo de sistemas totalmente automáticos. Sin embargo, la total automatización se manifiesta elusiva o impropcedente en muchas aplicaciones en las que se espera que la tecnología se use para asistir a los agentes humanos y no para tratar de suplantarlos.

La demanda social e industrial en tecnologías Interactivas Multimodales (IM) para el desarrollo de interfaces hombre-máquina avanzadas ha crecido considerablemente en la última década. Aunque el papel notablemente relevante de RF y VC en el desarrollo de estas tecnologías IM e interfaces ya se previó a principios de los setenta, sólo una pequeña parte de su enorme potencial ha sido explotado hasta la fecha.

La explotación de este potencial conlleva diversos retos y oportunidades de investigación para adaptar la tecnología existente de RF y VC a los entornos dinámicos y cambiantes de los sistemas interactivos.

En particular, identificamos los siguientes **retos y oportunidades** de I+D:

1. **Realimentación:** La información derivada de cada acción interactiva es útil para mejorar las prestaciones directamente,
2. **Adaptación:** La interactividad hace que sea posible y ventajoso realizar procesos de Aprendizaje Adaptativo que ajusten el sistema a las particularidades de la tarea y/o las preferencias del usuario,
3. **Multimodalidad:** Es intrínseca a la interactividad y conlleva mejoras del rendimiento y usabilidad de los sistemas.

Para entender estas ideas vamos a revisar una tarea a la que se pueden incorporar las técnicas anteriormente citadas. Concretamente, vamos a describir un sistema para la **transcripción interactiva de la voz** (ver Figura 1).

Las aplicaciones de estos sistemas van desde el proporcionar subtítulos de programas de televisión o transcripciones de programas de radio

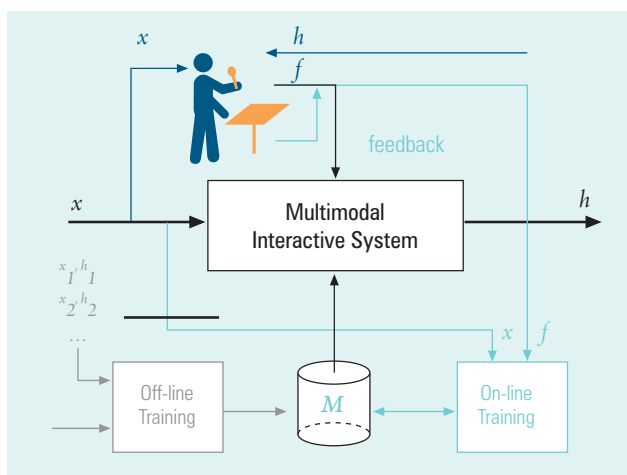


Figura 1.: Diagrama de un sistema interactivo multimodal

a personas con discapacidad auditiva, hasta el uso de estas transcripciones para realizar búsquedas textuales de contenidos, pasando por la transcripción de conferencias, charlas, sesiones judiciales, etc.

Desde el punto de vista del operario el sistema funciona de la siguiente forma:

1. El sistema escucha la señal de audio (posiblemente saltándose los fragmentos sin voz) y propone una posible transcripción.
2. El operario la examina y corrige el primer error que encuentra.
3. Como consecuencia de esta corrección, el sistema reconsidera la transcripción anterior y propone una nueva que muy probablemente corregirá algunos de los errores que vendrán a continuación.
4. Este proceso se repite hasta que el usuario está satisfecho con la transcripción.

Por el hecho de conocer la primera corrección el sistema adquiere nueva información, ahora sabe que el fragmento de texto transcrito hasta donde se produjo el error es correcto. Con esta información puede calcular con mejor precisión cuál será la transcripción del

siguiente fragmento. Además, ahora tiene más información sobre la tarea concreta en la que se está trabajando, ahora tiene nuevos datos acerca de cómo el locutor pronuncia algunas palabras. Esta información puede usarse para mejorar el modelo de pronunciación y así disminuir el número de futuros errores. Por último, en aras a mejorar la facilidad de uso del sistema, y dependiendo del entorno de trabajo del usuario, puede pensarse en alternativas al teclado o ratón para proponer las correcciones: voz, gestos, etc.

Lo que pretendemos es mostrar cómo las técnicas existentes de RF y VC pueden evolucionar de manera natural para permitir el desarrollo de sistemas interactivos multimodales avanzados, que materialicen la sinergia natural entre personas y máquinas de la que durante largo tiempo se ha estado hablando.

Para desarrollar las técnicas comentadas, el grupo Pattern Recognition and Human Language Technologies (PRHLT) del Instituto Tecnológico de Informática - ITI, participa en dos programas que fomentan la investigación de calidad. Por una parte, es el grupo coordinador de un proyecto enmarcado dentro del programa Consolider Ingenio 2010<sup>1</sup> perteneciente al Ministerio de Ciencia e Innovación. Por otra parte, ha obtenido financiación para realizar una actuación dentro del programa Prometeo<sup>2</sup> de la Generalitat Valenciana.

## 2. El programa Consolider Ingenio 2010<sup>1</sup>

Los proyectos Consolider son actuaciones de carácter estratégico para la financiación de actividades científicas de alto nivel que promuevan un avance significativo en el estado del conocimiento o que establezca nuevas líneas de investigación originales y actualizadas situadas en la frontera del conocimiento. Además, el programa Consolider Ingenio 2010 persigue conseguir la excelencia investigadora aumentando la cooperación entre investigadores y formando grandes grupos de investigación. El grupo PRHLT del ITI coordina el proyecto Multimodal Interaction in Pattern Recognition and Computer Vision (MIPRCV) que es una actuación evaluada positivamente en la convocatoria de 2007 del programa Consolider Ingenio 2010. Toda la información referente a MIPRCV se puede encontrar en su web <http://miprcv.iti.upv.es>.

MIPRCV establece un programa de investigación básica y aplicada de cinco años para estudiar los retos y oportunidades científico-técnicas que conlleva el situar RF y VC bajo el paradigma de Interacción Multimodal entre personas y máquinas.

En este contexto, MIPRCV también considera el desarrollo e implementación de prototipos y sistemas reales para un amplio abanico de aplicaciones importantes de interacción multimodal en los campos de visión por computador, robótica y procesado de audio, voz y lenguaje humano, entre las que destacamos:

- Recuperación de imágenes, texto, música, contenidos multimedia.
- Reconocimiento de caras.
- Traducción.
- Transcripción de música, imágenes de textos y texto manuscrito.
- Seguimiento de personas y reconocimiento de acciones humanas.
- Restauración de imágenes.

- Diagnóstico médico.
- Análisis de la estructura de documentos.
- Conducción de vehículos.
- Robótica ubicua.
- Sistemas de diálogo.

El consorcio de investigación que se ha constituido para alcanzar este objetivo está constituido por más de 100 científicos e ingenieros altamente cualificados pertenecientes a siete grupos de investigación y diez instituciones públicas de investigación. Esta plantilla incluye numerosos investigadores de reconocido prestigio internacional en los campos de *Reconocimiento de Formas, Aprendizaje Automático, Procesamiento de Imágenes, Visión por Computador, Procesamiento del Habla y del Lenguaje y Robótica*. Una ventaja añadida de este consorcio es la gran cohesión existente entre sus miembros a través de su pertenencia a la *Asociación Española de Reconocimiento de Formas y Análisis de Imágenes*, a diversas redes nacionales relacionadas con el área, a la participación conjunta en proyectos coordinados de investigación que han dado lugar a un gran número de publicaciones conjuntas, y asimismo en la colaboración de profesores de unos grupos en los programas de doctorados de otros del consorcio.

Los 7 grupos de investigación que componen el consorcio, así como sus principales líneas de trabajo son:

**Computer Vision (CV-CVC).** Sistemas avanzados de asistencia a la conducción, visión robótica, análisis de secuencias de vídeo.

**Computer Vision and Digital Signal Processing Group (CVDSP-UJI).** Minería de datos, análisis del comportamiento humano y biometría, teledetección, análisis de imágenes médicas, realidad mixta y aumentada.

**Computer Vision & Learning (CVL-UGR).** Restauración de imágenes, super-resolución, reconocimiento de acciones humanas, modelos gráficos probabilísticos, recuperación de información, computación evolutiva.

**Pattern Recognition (PR-CVC).** Reconocimiento de objetos (en especial rostros humanos), análisis de imágenes médicas, análisis de documentos, análisis de textura y color.

**Pattern Recognition and Artificial Intelligence (PRAI-UA).** Reconocimiento estadístico y estructural de formas, aprendizaje automático. Tareas de música por computador, como transcripción de audio a partitura, análisis automático de música, extracción de metadatos a partir de audio y partitura, composición algorítmica, etc.

**Pattern Recognition and Human Language Technology (PRHLT-ITI).** Traducción automática y asistida. Reconocimiento de voz incluyendo comprensión y diálogo. Visión por computador: imágenes de texto, gestos, imagen médica, etc. Aprendizaje automático, minería de datos, recuperación de imágenes y biometría.

**Robotics and Perception (RP-IRI).** Robótica ubicua, sistemas de percepción, robots móviles y métodos geométricos en robótica.

### 2.1. Aplicaciones consideradas en MIPRCV

A continuación, se presenta una descripción de las aplicaciones más destacables que se están llevando a cabo por el consorcio de MIPRCV.

**Transcripción (de voz, música, imágenes de texto) y traducción interactiva.** Para la obtención de transcripciones de alta calidad es necesaria la revisión de un experto. Sin embargo, esta tarea es ineficiente cuando se realiza mediante post-edición, es decir, corrigiendo de forma no interactiva la transcripción completa devuelta por el sistema. El paradigma interactivo permite una aproximación más efectiva en la que el experto humano interactúa con el sistema validando segmentos de texto y corrigiendo errores. Esta parte del texto validada por el experto es utilizada por el sistema para mejorar su decisión en la siguiente sugerencia. Las correcciones del usuario pueden ser introducidas, como es habitual, por teclado y ratón, pero otras modalidades de interacción más sofisticadas como lápiz electrónico, seguimiento de la mirada, o reconocimiento de voz son posibles. Estas ideas también son aplicables a la traducción automática y la transcripción de música.

**Recuperación interactiva de contenidos multimedia.** En la recuperación de contenidos convencional, dada una colección de contenidos multimedia, al hacer una consulta el sistema busca en la colección los contenidos que más se asemejan a la consulta realizada. Muchas veces la información recuperada con una consulta no cubre las expectativas del usuario, debido en parte por la propia falta de información en la consulta realizada. Sin embargo, si se utiliza el paradigma interactivo, el usuario puede proporcionar retroalimentación relevante sobre la adecuación de la información recuperada. Este problema general puede ser aplicado a diferentes tareas: reconocimiento automático de caras para vigilancia asistida por ordenador, recuperación interactiva de imágenes, texto o música, etc.

**Fusión interactiva de imágenes y clasificación multimodal.** En fusión y super-resolución, imágenes de poca resolución se combinan para obtener otras de mayor resolución. La información de las imágenes originales puede ser suministrada de diferentes formas, de manera que dará lugar a la aparición de problemas interactivos multimodales de gran interés. Una vez que las imágenes han sido reconstruidas, éstas pueden usarse en tareas de clasificación multimodal, cuyo rendimiento puede mejorar sustancialmente por el uso de la información multimodal (como por ejemplo audio y video).

**Análisis de la composición de documentos asistido por ordenador.** En este tipo de aplicaciones, el sistema tiene que determinar los bloques que contienen la información relevante del documento. En el paradigma de interacción multimodal, cada punto corregido por un usuario es una información que el sistema puede reutilizar para mejorar la selección realizada.

**Diagnóstico asistido por imágenes médicas.** El diagnóstico mediante imagen médica es actualmente un procedimiento usual en Medicina. Típicamente, el sistema calcula automáticamente informaciones que pueden ayudar en el diagnóstico. Estas informaciones se añaden a las imágenes originales por ejemplo en forma de falso color. En un planteamiento interactivo, el sistema obtiene datos de retroalimentación derivados del conocimiento experto del médico que analiza la imagen. De esta forma, el sistema puede mejorar sus predicciones actuales y futuras en situaciones similares.

**Sistemas de diálogo multimodales.** Los sistemas clásicos de reconocimiento de voz pueden ser mejorados con la incorporación de nuevas fuentes de entrada, como por ejemplo un teclado, imágenes de caras, observación del movimiento de los labios o pantallas táctiles.



La combinación de todas estas fuentes de información permite una interpretación más fiable de las intenciones de un usuario y puede ayudar a mejorar las respuestas de un sistema de diálogo.

### Seguimiento cooperativo y reconocimiento de acciones humanas.

Estos sistemas analizan la actividad humana capturada a partir de cámaras, y explican los comportamientos humanos a través de textos de lenguaje natural y entornos virtuales. La videovigilancia es una de las aplicaciones más interesantes de esta tecnología.

**Sistemas avanzados de asistencia a la conducción.** El objetivo de estos sistemas es la prevención de accidentes o la reducción de sus consecuencias cuando son inevitables. Los accidentes de tráfico son actualmente una causa importante de mortalidad en los países desarrollados con 10 millones de personas por año involucradas en los mismos. Los costes hospitalarios, daños a la propiedad, y otros costes relacionados, conllevan un 2 por ciento del producto interior bruto en el mundo. Algunos ejemplos de sistemas avanzados de asistencia a la conducción son: control de velocidad de cruceo adaptativo, sistemas de alertas por cambio involuntario de carril, sistemas de detección de peatones, etc.

**Robótica ubicua.** Muchos de los trabajos en este área se centran en la naturaleza multimodal de las señales involucradas y en los interfaces de comunicación persona-máquina. Bajo el paradigma de interacción multimodal, ambos aspectos están perfectamente integrados, junto con la adaptación on-line (activa) del sistema a la tarea a realizar y a las preferencias del usuario.

## 3. El programa Prometeo<sup>2</sup>

El programa Prometeo, puesto en marcha en 2008 desde la Dirección General de Política Científica de la Generalitat Valenciana, se enmarca en un sistema de ayudas que persigue la identificación de grupos de I+D de excelencia en la Comunidad Valenciana. La finalidad de este programa es promover la investigación de calidad y favorecer el máximo nivel de excelencia de los grupos de investigación con una trayectoria acreditada y solvente dentro de la comunidad científica nacional e internacional, mediante la realización de acciones singulares de I+D.

Una de estas acciones singulares de I+D concedidas por el programa Prometeo ha recaído en el grupo PRHLT del ITI. La acción lleva por título *Adaptive Learning and Multimodality in Pattern Recognition* (ALMPR) y su principal objetivo es el desarrollo de una nueva generación de interfaces interactivos y multimodales (IIMs) basados en RF. Esta acción complementa algunos aspectos de MIPRCV.

Concretamente los objetivos científico/tecnológicos que se pretenden conseguir son: diseñar modelos generales y algoritmos de RF para IIM y aprendizaje adaptativo (AA), desarrollar prototipos para diferentes aplicaciones y diseñar metodologías para la evaluación de IIMs.

En cuanto a transferencia tecnológica, se pretende aumentar el número de publicaciones en revistas y congresos, aumentar el esfuerzo en la preparación de proyectos europeos, aumentar la colaboración con empresas privadas y complementar proyectos nacionales e internacionales con investigación sobre AA e IIMs.

Las líneas de investigación que se van a seguir en ALMPR son:

- Aprendizaje Adaptativo y Multimodalidad para RF

- Transcripción de Voz Interactiva
- Reconocimiento Interactivo de Texto Manuscrito
- Traducción Automática Interactiva
- Recuperación Interactiva de Imágenes por Relevancia

Se puede encontrar más información sobre ALMPR en la página web <http://almater.iti.upv.es/>.

## 4. Aplicaciones de IM desarrolladas en el ITI

A continuación, se exponen las aplicaciones que se están llevando a cabo por el grupo PRHLT del ITI en las que se utilizan las técnicas de IM presentadas en este artículo. Estas aplicaciones se están desarrollando a través de MIPRCV y ALMPR.

### 4.1. Aplicaciones de Procesamiento de Lenguaje Natural

Tradicionalmente, las aplicaciones de *Procesamiento del Lenguaje Natural* (PLN) se han centrado en sistemas totalmente automáticos. Sin embargo, dado que los resultados que obtienen estos sistemas distan mucho de ser perfectos, es imposible reemplazar a los expertos. Típicamente, los usuarios expertos utilizan los sistemas automáticos de la siguiente forma: en primer lugar, el sistema genera automáticamente una respuesta; seguidamente, el usuario revisa esta respuesta, que deberá reescribir completamente para conseguir que tenga una calidad aceptable. Este paso de post-edición resulta totalmente ineficiente e incomodo para el usuario. Como alternativa a este paso de post-edición, se proponen soluciones basadas en IM. El usuario introduce realimentación en el sistema (realiza correcciones) a través de la utilización de un lápiz electrónico sobre una pantalla táctil y/o a través del teclado y ratón convencionales. Solamente corrige aquello que no está bien (y no toda la respuesta generada).

En esta aproximación, el sistema automático y el usuario cooperan para obtener la solución correcta. Las correcciones hechas por el usuario mejoran la precisión del sistema, mientras que la multimodalidad incrementa la ergonomía del sistema y la aceptación por parte del usuario. La IM está planteada de forma que las correcciones hechas por el usuario y el resultado obtenido por el sistema automático se ayuden mutuamente para optimizar el resultado final y la usabilidad del sistema.

Dentro del área de PLN, el grupo PRHLT está trabajando en diversas aplicaciones. En este apartado se comentan las más destacadas.

**Transcripción Interactiva de Texto Manuscrito.** Existen millones de imágenes de texto (antiguo) que esperan su transcripción. El principal inconveniente que hay para que no se lleve a cabo esta tarea es que la transcripción manual requiere expertos altamente cualificados (paleógrafos) y muy caros. Por otra parte, los sistemas de OCR convencionales no son adecuados para el reconocimiento de texto manuscrito en cursiva. Es por ello que proponemos una solución basada en IM en la que el usuario asista al sistema para conseguir transcripciones de calidad.

Los resultados obtenidos hasta el momento utilizando esta técnica muestran una reducción de esfuerzo entre el 60-80% respecto a la transcripción manual y una reducción de entre el 10-30% respecto a la post-edición (ver Figura 2).

Una demostración del sistema de transcripción de texto manuscrito que se está desarrollando se puede encontrar en la página web <http://catti.iti.upv.es>.

**Traducción Automática de Textos Interactiva.** Existe una creciente demanda de traducciones de calidad en la sociedad y los traductores automáticos distan mucho de ofrecer soluciones óptimas. Es por ello que se está desarrollando un sistema de IM con el objetivo de mejorar estas traducciones (ver Figura 3).

Una demostración de este sistema se puede encontrar en la página web <http://cat.iti.upv.es>.

**Transcripción Interactiva de Voz.** Otra de las aplicaciones de IM que se está llevando a cabo es la de transcripción interactiva de voz. Como ya se ha comentado en este artículo, existen numerosos usos de este tipo de sistemas (proporcionar subtítulos, transcripción de conferencias, charlas, etc.) y será muy interesante disponer de un sistema de IM que ayude a que esta transcripción sea óptima (ver Figura 4).

#### 4.2. Aplicaciones de Procesamiento de Imágenes

Otro campo de aplicaciones en el que se está trabajando aplicando técnicas de IM es el procesamiento de imágenes. Lo que se pretende en esta área es mejorar la calidad de las imágenes o facilitar la búsqueda de información en ellas. A continuación, se detalla una de estas aplicaciones.

**Recuperación Interactiva de Imágenes.** La precisión de los sistemas de recuperación de imágenes basadas en contexto ha mejorado durante los últimos años. Sin embargo, esta mejora todavía no es suficiente para obtener resultados óptimos. Por otra parte, el usuario todavía prefiere realizar consultas en forma de texto en lugar de proporcionar una imagen como ejemplo de los resultados deseados. Este inconveniente añade complejidad a la hora de asociar el concepto semántico del texto con el contenido de las imágenes que se quieren recuperar.

La precisión de estos sistemas se puede incrementar utilizando la supervisión del usuario utilizando técnicas de IM. En este sentido se pretenden alcanzar dos objetivos. En primer lugar, se va a diseñar un sistema de recuperación de imágenes basadas en contexto que mejore los resultados actuales definiendo nuevos descriptores de imágenes, nuevas funciones de distancia y utilizando consultas en forma de texto. En segundo lugar, se utilizará el conocimiento aportado por el usuario para mejorar de forma interactiva el resultado de la búsqueda de la imagen utilizando realimentación por relevancia.

#### Artículos sobre IM publicados en revistas internacionales

A continuación, se presenta un listado de los artículos más destacados que hemos publicado en diferentes revistas internacionales y en los que se detallan las técnicas de IM que se han presentado en este artículo.

- Título: *Statistical approaches to computer-assisted translation.*  
Revista: *Computational Linguistics*
- Título: *Human interaction for high quality machine translation.*  
Revista: *Communications of the ACM*
- Título: *Multimodal Interactive Transcription of Text Images.*  
Revista: *Pattern Recognition.* ■



Figura 2: Prototipo de Transcripción Interactiva de Texto Manuscrito

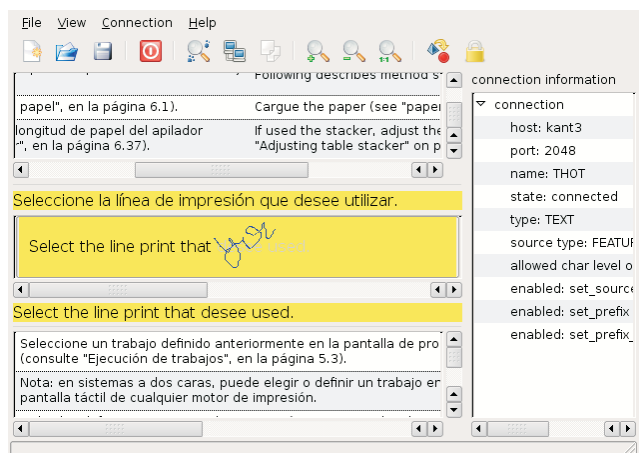


Figura 3: Prototipo de Traducción Automática de Textos Interactiva



Figura 4: Prototipo de Transcripción Interactiva de Voz

[1]Proyecto Consolider MIPRCV: CSD2007-00018  
[2]Proyecto Prometeo ALMPR: PROMETEO/2009/014