Corrección del punto de vista en videoconferencia para garantizar el contacto visual

Jaume Civit y Tomás Montserrat

División de Tecnologías de Vídeo, Telefónica Investigación y Desarrollo
{jaume, tmmora}@tid.es

***Abstract* — In a typical desktop video-conference setup, the camera and the display screen cannot be physically aligned. This problem produces lack of eye contact and substantially degrades the user's experience. Expensive hardware systems using semi-reflective materials are available on the market to solve the eye gazing problem. However, these specialized systems are far away from the mass market. This paper presents an alternative approach using stereo rigs to capture a three-dimensional model of the scene. This information is then used to generate the view from a virtual camera aligned with the conference image the user looks at.**

I. Introducción

La videoconferencia permite la comunicación cara a cara de personas geográficamente distantes mediante la transmisión bidireccional de audio y vídeo. Sin embargo, la expansión de esta tecnología sigue muy por debajo de las expectativas generadas inicialmente. Superados problemas como el coste o el ancho de banda, parece que la principal barrera para una adopción generalizada de la videoconferencia es la falta de contacto visual [1].

En un entorno doméstico típico, la cámara y la pantalla no se pueden alinear físicamente, tal y como se muestra en la figura 1. El usuario mira hacia la imagen del interlocutor remoto mostrada en el monitor, pero no directamente a la cámara desde la cual es observado, por lo que se pierde la impresión de estar mirando a los ojos del interlocutor. Se ha demostrado [3] que si el ángulo de divergencia entre la cámara y la pantalla es superior a cinco grados, la pérdida de contacto visual es apreciable. En un escenario habitual, con en usuario sentado frente al ordenador, el valor de este ángulo se sitúa entre los quince y veinte grados. Esto conlleva efectos psicológicos negativos, dado que la falta de contacto visual, o esquivar la mirada del interlocutor, tiende a asociarse con el engaño [2], por lo que por encima del umbral de divergencia, la información de vídeo pierde su valor comunicativo, pudiendo incluso llegar a incomodar.

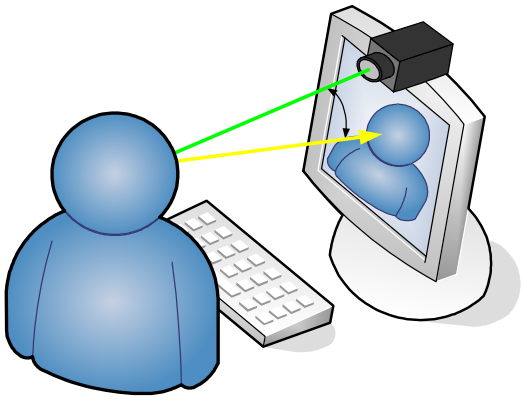

Fig. 1. Sistema de videoconferencia convencional.

Distintos sistemas hardware han sido propuestos para corregir la desviación de la mirada en la videoconferencia. En mayoría ellos se emplean materiales semi-reflectantes para alinear la cámara y la imagen remota. Los productos ofrecidos por la empresa Digital Video Enterprises en su línea de telepresencia son un buen ejemplo de ello [5]. El sistema sugerido por Okada et al. [6] proyecta la imagen sobre una pantalla semitransparente y captura la cara del usuario con una cámara situada detrás de la pantalla. A pesar de su efectividad, el elevado coste de estos sistemas, juntamente con un montaje aparatoso, los ha mantenido alejados del mercado de masas.

El presente sistema se basa en la obtención de una descripción tridimensional de la escena para así generar la imagen correspondiente a una cámara virtual situada en el punto de vista deseado. De esta forma se evita la necesidad de situar físicamente la cámara a la altura de los ojos para garantizar el contacto visual. La información 3D se puede obtener mediante la captura de la escena con dos o más cámaras.

Dados los parámetros de calibración de las cámaras y las correspondencias de puntos entre las distintas vistas, es posible triangular la posición del punto tridimensional que ha originado cada conjunto de correspondencias y, por lo tanto, generar cualquier vista arbitraria. Existen dos aspectos fundamentales para que los resultados de este proceso sean satisfactorios. En primer lugar, el algoritmo de búsqueda de correspondencias debe proporcionar resultados precisos. En segundo lugar, la cámara virtual debe estar situada entre las dos cámaras reales, de modo que cualquier punto que aparezca en la nueva imagen sea visible al menos para una de las cámaras. Si únicamente se emplean dos cámaras, esta última restricción obligará a situarlas en lados opuestos del monitor. Sin embargo, tanto si la orientación es vertical como horizontal, el tamaño del monitor forzará una separación de cámaras elevada, lo cual dificulta en gran medida el proceso de búsqueda de correspondencias. Para solventar este problema, se propone añadir dos cámaras adicionales según el esquema de la figura 2.

This work has been partially developed in the collaborative VISION Project, financed by Spanish CENIT Programme.
This work has been performed in part within the framework of the EU FP7 Project 3DPresence (ICT, 215269).

La menor separación, o línea base, entre los centros ópticos de ambas cámaras permite un buen funcionamiento de la búsqueda de correspondencias entre los pares de cámaras situados en el mismo lado. Mientras que las cámaras situadas en el lado opuesto ayudan a sintetizar correctamente la vista virtual. La figura 3 muestra los cuatro pasos fundamentales en los que se basa el sistema aquí propuesto: calibración, rectificación, búsqueda de correspondencias y síntesis de la nueva vista. Estas etapas serán descritas en los siguientes apartados.

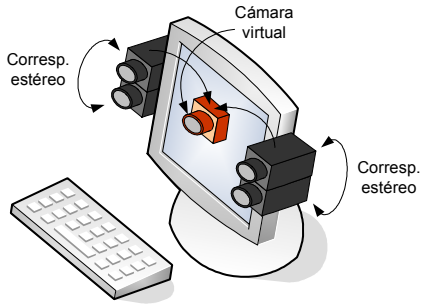

Fig. 2. Ubicación de las cámaras.

En la literatura se pueden hallar también distintas propuestas para solventar el problema del contacto visual empleando algoritmos de visión artificial. Ott et al. [4] propusieron un sistema similar al aquí presentado, pero con una configuración de dos cámaras en lados opuestos. Las limitaciones de la búsqueda de correspondencias provocaban la aparición de artefactos en la vista generada. Pese a emplear un algoritmo de programación dinámica altamente paralelizable, el tiempo de procesado era superior a 50 s por cuadro con el hardware de la época.

El proyecto GazeMaster de Microsoft Research [7] emplea una única cámara para seguir la orientación de la cabeza y los ojos. La síntesis de la vista se lleva a cabo substituyendo los ojos del usuario por unos sintéticos mirando en la dirección deseada. La textura de la cara con la mirada corregida se aplica posteriormente a un modelo rígido de cara 3D que puede ser rotado. Las imágenes generadas tienen un aspecto sintético, similar al de un avatar, probablemente debido al modelo genérico de cara empleado. La substitución de los ojos puede ocasionar también cambios en la expresión facial.

Yang y Zhang [8], utilizan un modelo personalizado de cara ajustado al vídeo mediante el seguimiento de puntos característicos en 3D, para ello emplean dos cámaras montadas en los lados superior e inferior del monitor. En una segunda etapa, los objetos que se hallan fuera del modelo de cara son tratados mediante la búsqueda de correspondencias sobre contornos y puntos característicos. El resultado se emplea para generar una malla 3D con la que se sintetiza la nueva vista. A diferencia del sistema aquí propuesto, la elevada separación de las cámaras únicamente permite realizar una búsqueda de correspondencias fiable basándose en la detección de puntos característicos. La baja densidad del mapa de profundidad resultante exige apoyarse en un conocimiento previo de la escena y en modelos tridimensionales prediseñados.

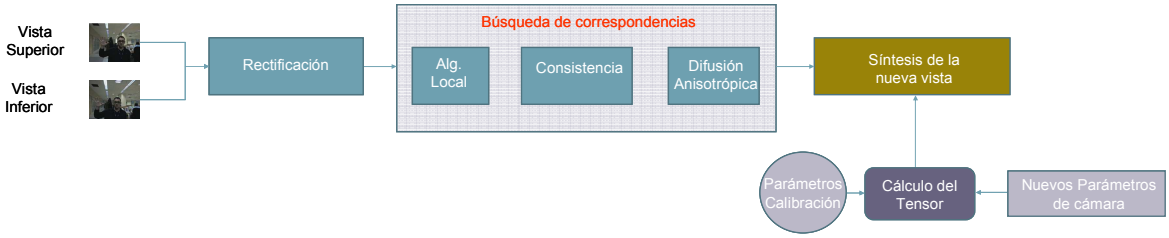

Fig. 3. Diagrama parcial de arquitectura del sistema.

II. CALIBRACIÓN Y RECTIFICACIÓN

El proceso de calibración permite recuperar la posición relativa entre las cámaras, obteniendo la información geométrica global del sistema. Mediante este procedimiento se recuperan los parámetros intrínsecos de las cámaras (distancia focal, centro óptico, etc) propios de cada lente, y los parámetros extrínsecos que relacionan el sistema de referencia de cada cámara, con origen en el centro óptico, respecto un sistema de referencia global (en el espacio 3D) común para todas las cámaras; además de obtener los parámetros de distorsión de las lentes. El algoritmo empleado para esta etapa es el propuesto en [9].

Una vez conocidos los parámetros de calibración es posible explotar las restricciones geométricas conocidas entre puntos de vista, para facilitar etapas posteriores. Los parámetros se manejan en forma de la matriz de proyección, o de cámara, P.

Para cada par de cámaras se introduce una etapa llamada rectificación epipolar. Una vez recuperados los parámetros de calibración se pueden aplicar las restricciones establecidas por la geometría epipolar (geometría de 2 vistas)[11]. La rectificación consiste en generar dos nuevas cámaras con matrices P1' y P2", tales que dado un píxel en una de las imágenes, su píxel correspondiente se encuentre sobre la misma línea en la otra imagen. Para más información consultar [14]. Este proceso simplifica la etapa de búsqueda de correspondencias a una sola dimensión. La rectificación de las imágenes exige realizar también una corrección de las distorsiones radial y tangencial propias de la lente de cada cámara.

III. BÚSQUEDA DE CORRESPONDENCIAS

La etapa de búsqueda de correspondencias tiene como objetivo emparejar aquellos puntos en las imágenes que son proyecciones de un mismo punto tridimensional. Se denomina disparidad al desplazamiento relativo de la posición de un punto respecto a su correspondencia en la imagen de referencia, este valor es inversamente proporcional a la profundidad del punto 3D. El conjunto de todas las disparidades asociadas a cada píxel de una imagen recibe el nombre de mapa de disparidad.

Gracias al proceso de rectificación, se simplifica problema de la búsqueda de correspondencias. Sin embargo, esta sigue siendo una tarea computacionalmente costosa dado que exige lidiar con las ambigüedades ocasionadas por regiones homogéneas, oclusiones, etc. Las disparidades pueden ser halladas mediante distintos métodos, siendo habitual realizar la distinción entre métodos locales y globales, según el tipo de restricciones empleadas. Los métodos locales únicamente hacen uso de la información proporcionada por un pequeño número de píxeles vecinos del píxel de interés. A pesar de que pueden ser muy eficientes, son altamente sensibles a las zonas localmente ambiguas de las imágenes. En el otro extremo, los métodos globales imponen restricciones que afectan a la imagen entera, lo cual hace que sean más robustos que los locales a cambio de un mayor coste computacional. En [13] se puede encontrar una detallada clasificación de los distintos tipos de algoritmos empleados.

La naturaleza del presente sistema exige calcular dos mapas de disparidad independientes a una resolución VGA y una frecuencia de 30 imágenes por segundo. Estos requisitos de tiempo real resultan prohibitivos para los algoritmos de tipo global, cuya paralelización es también más compleja. Por este motivo, se ha optado por un método local con ventana de agregación no adaptativa [13]. Por motivos de eficiencia, la función de coste empleada es la suma de diferencias absolutas (SAD). En cada grupo de estéreo-cámaras se realiza la comprobación de consistencia superior/inferior del mapa de disparidad, lo que permite descartar las asignaciones erróneas fruto de oclusiones o zonas de poca textura. Este proceso da como resultado un mapa de disparidad con huecos ocasionados por las zonas de disparidad no fiable.

Para poder realizar la síntesis de la nueva vista es necesario disponer de un mapa de disparidad denso a cada lado del monitor. Por este motivo, los mapas "consistentes" son sometidos un proceso de post-procesado basado en la difusión anisotrópica propuesta por Perona y Malik [15]. La principal diferencia del filtro empleado respecto al descrito en [15], es el uso de la imagen de referencia del mapa de disparidad para calcular los coeficientes de difusión, en lugar del propio mapa a filtrar. Esta estrategia se basa el la restricción de que las regiones de intensidad homogénea no presentarán cambios bruscos de profundidad. En las primeras iteraciones del filtro se modifican únicamente los valores de los puntos marcados como descartados, los cuales se inicializan con el valor de disparidad hallado en el frame anterior, para reducir el número total de iteraciones por imagen. Las últimas iteraciones sí modifican la totalidad del mapa de disparidad contribuyendo a suavizar el resultado final.

El método de post-procesado empleado permite conservar los contornos de la imagen original, lo que contribuye a obtener resultados foto-consistentes en la reproyección de la nueva vista.

El sistema de búsqueda de correspondencias aquí descrito funciona en tiempo real (VGA@30fps) implementado sobre una FPGA Virtex-4 SX-35 de Xilinx, con una frecuencia de funcionamiento entorno a los 200 MHz. El paralelismo de los algoritmos utilizados hace que sea posible plantear su funcionamiento sobre las nuevas tarjetas gráficas.

En este apartado se plantea como línea futura la mejora de los mapas de disparidad incorporando técnicas de segmentación por color y métodos globales.

IV. REPROYECCIÓN DEL NUEVO PUNTO DE VISTA

La técnica empleada para generar la imagen percibida por la nueva cámara virtual, está dentro del grupo Image Based Rendering [12]. Tal y como su nombre indica, este grupo de algoritmos utiliza únicamente, en el caso ideal, información contenida dentro de las imágenes reales, es decir, aquellas que se capturan a partir de las cámaras físicas. En el caso particular, presentado en este artículo, se utiliza un método basado en transferencia de puntos: los píxeles de las imágenes originales se transfieren a su posición correspondiente en la imagen sintética, con ello se consigue obtener imágenes virtuales completamente realistas (foto-realismo). Para ello se modela cada par de cámaras, situado a cada lado de la pantalla, como un sistema de captura trinocular, formado por las dos cámaras reales y la cámara virtual, en el cual se aprovechan las restricciones de la geometría de tres vistas.

La entidad algebraica que relaciona la posición de los píxeles en las tres imágenes se denomina Tensor Trifocal. Se trata de una matriz 3x3x3 que contiene todas las relaciones geométricas entre las tres vistas, equivalente a la matriz fundamental para un sistema binocular. Dadas tres matrices de cámara P=[I|0] P1=[A|a_i] y P2=[B|b_i], obtenidas durante el proceso de calibración, se puede obtener el tensor de la siguiente forma, expresado en notación tensorial y utilizando la convención de Einstein [11].

$$\mathfrak{T}_i^{jk} = a_i^j b_4^k - a_4^j b_i^k \tag{1}$$

A partir del tensor trifocal se obtienen las relaciones de incidencia entre puntos y líneas de las tres imágenes. La más útil para la configuración planteada en este artículo es:

$$x''^{k} = x^{i} l'_{j} \, \Im_{i}^{jk} \tag{2}$$

Conocida como la incidencia punto-línea-punto [11], el diagrama se puede observar en la figura 4. Según la ecuación anterior, conociendo un punto de la primera imagen y una línea que pase por el punto correspondiente en la segunda imagen, es posible calcular cuál es la posición del píxel correspondiente en la tercera imagen, en este caso la imagen sintética.

Para establecer las correspondencias entre las dos primeras vistas y calcular los parámetros de entrada de la ecuación (2), se recurre a las técnicas de estereoscopía introducidas en el apartado III, que obtienen como resultado un mapa de disparidad que contiene la información de las correspondencias entre píxeles del par estéreo. De este modo se obtiene la línea perpendicular a la línea epipolar de x, en la segunda imagen, que pasa por el punto x'. Para más detalles sobre este algoritmo véase [10].

Este procedimiento se aplica a todos los píxeles de la imagen 1, para generar la imagen sintética en la posición de la cámara virtual, imagen 3. Evidentemente habrá partes de la escena que no son visibles, a la vez desde la cámara virtual y las cámaras reales, por consiguiente, la imagen generada presentará huecos, coincidentes con las oclusiones entre ambas vistas. Para compensar este hecho se aplica el procedimiento anterior a ambos pares de cámaras, según la configuración propuesta en el apartado I, generando dos imágenes sintéticas para el mismo punto de vista, cámara virtual, de este modo, los huecos en cada una de la imágenes serán distintos y no coincidentes, debido a la arquitectura del sistema de captura. Con ambas imágenes sintéticas es posible generar una vista virtual completa y foto-realista.

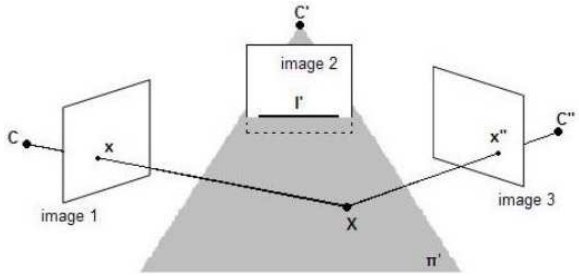

Fig. 4. Incidencia punto-línea-punto [11].

Este algoritmo se implementa en formato de tabla de búsqueda 3D, una tabla con tres entradas: posición del píxel en la imagen de referencia (u,v) en coordenadas imagen y la disparidad del píxel. De este modo las operaciones se realizan una sola vez, consiguiendo ratios de tiempo real. Realmente se producen dos tablas de búsqueda, una para cada par estéreo.

V. Conclusión

En el presente artículo se ha presentado una propuesta para la mejora del servicio de videoconferencia. Mediante la aplicación de técnicas de visión artificial se consigue la corrección del punto de vista del usuario, creando una comunicación más realista.

REFERENCIAS

[1] L. Mhlbach, B. Kellner, A. Prussog, and G. Romahn, "The Importance of Eye Contact in a Videotelephone Service," *Proc. 11th Int'l Symp. Human Factors in Telecomm*., 1985.

[2] Ernst Bekkering, J.P. Shim. "Trust in Videoconferencing." *Communications of the ACM*, Volume 49 Issue 7, July 2006.

[3] R.R. Stokes, "Human Factors and Appearance Design Considerations of the Mod II PICTUREPHONE Station Set," *IEEE Trans. Comm. Technology*, vol. 17, no. 2, Apr. 1969.

[4] M. Ott, J. Lewis, and I. Cox, "Teleconferencing Eye Contact Using a Virtual Camera," *Proc. Conf. Human Factors in Computing Systems*, pp. 119-110, 1993.

[5] Digital Video Enterprises. http://www.dvetelepresence.com. Telepresence systems. 2008

[6] Okada, K.I., Maeda, F., Ichikawa, Y., and Matsushita, Y. "Multiparty videoconferencing at virtual social distance: MAJIC design." *Proc. CSCW '94*, pp.385-395, 1994.

[7] J. Gemmell, C.L. Zitnick, T. Kang, K. Toyama, and S. Seitz, "Gaze-Awareness for Videoconferencing: A Software Approach," *IEEE Multimedia*, vol. 7, no. 4, pp. 26-35, Oct. 2000.

[8] Yang, R., and Zhang, Z.; Eye gaze correction with stereovision for video-teleconferencing. *Microsoft Research, Technical Report*, MSR-TR-2001-119, 2001.

[9] Zhang, Z. "A flexible new technique for camera calibration" *IEEE Transactios on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

[10] Avidan, S. and Shashua. M. "Novel View Synthesis by Cascading Trilinear Tensors". *IEEE Transactions on Visualization and Computer Graphics*, 4(4), pp. 1077-2626, October to December 1998.

[11] Hartley R. I. and A. Zisserman. *Multiple View Geometry in Computer Vision*. CUP, Cambridge, 2000.

[12] Shum, Heung-Yeung, Chan, Shing-Chow, Kang, Sing Bing, *Image-Based Rendering* XX, 408 p. 95 illus., Hardcover4, 2007.

[13] Daniel Scharstein , Richard Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms", *International Journal of Computer Vision*, v.47 n.1-3, p.7-42, April-June 2002

[14] A. Fusiello, E. Trucco, and A.Verri. "A compact algorithm for rectification of stereo pairs". *Machine Vision and Applications*, 12(1):16–22, 2000.

[15] Perona, P.; Malik, J. "Scale-space and edge detection using anisotropic diffusion", *Pattern Analysis and Machine Intelligence, IEEE Transactions on* Volume 12, Issue 7, Page(s):629 – 639, Jul 1990.