UNIVERSIDAD DE ALCALÁ

Escuela Politécnica Superior

Departamento de Electrónica

Estudios de Doctorado

PROYECTO DE TESIS DOCTORAL

"Localización multimodal de personas en espacios inteligentes"

Autor: Jose Velasco Cerpa
Directores: Javier Macías Guarasa
Daniel Pizarro Pérez

Septiembre 2011

1. Introducción

El presente proyecto de Tesis se enmarca dentro de la actividad del grupo de investigación GEINTRA (Grupo de Ingeniería Electrónica Aplicada a Espacios Inteligentes y Transporte) del Departamento de Electrónica de la Universidad de Alcalá. La actividad investigadora del grupo se relaciona, entre otros aspectos, con el análisis multimodal de diferentes tecnología de sensores y su aplicación a la robótica móvil, los espacios inteligentes y los interfaces hombre-máquina.

La Tesis participa directamente en la temática desarrollada por el grupo de investigación en varios proyectos competitivos. Destacan el proyecto VISNÚ (TIN2009-08984), donde se abordan cuestiones relacionadas con la localización de personas y robots móviles en espacios inteligentes y la interacción humano-máquina y el proyecto SDTEAM-UAH (TIN2008-06856-C05-05), dedicado al análisis de la señal de audio de agrupaciones de micrófonos para la localización de personas. Ambos proyectos tratan una linea de investigación común centrada en lo que se denomina "espacio inteligente", que consiste en un entorno dotado con sensores capaces de analizar automáticamente la escena con el objetivo de aportar un servicio a los usuarios. En calidad de sensores para acometer dicha tarea, el grupo de investigación ha desarrollado investigación en diversas tecnologías: audio, ultrasonidos, infrarrojos y sensores de visión. El paradigma de los espacios inteligentes tiene un interés notable tanto a nivel científico como social.

Dentro del ámbito donde el grupo GEINTRA desarrolla su investigación, la presente Tesis se centra en el uso de técnicas de audio y visión artificial para la localización de personas con precisión dentro de un entorno dotado con cámaras y agrupaciones o "arrays" de micrófonos. Tanto las cámaras como los micrófonos se encuentran en posiciones fijas y conocidas del entorno y están, en un principio sincronizados en su adquisición.

El objetivo de la Tesis es encontrar un marco óptimo para realizar la fusión de la información proporcionada por sensores de tecnología muy dispar, esto es, audio y vídeo. Existen multitud de retos tecnológicos y teóricos en esta disciplina, y la presente Tesis supone un paso natural dentro de las lineas de investigación del grupo GEINTRA, donde no existen todavía trabajos dentro del grupo que aborden el enfoque multimodal (audio y vídeo) a la localización de personas dentro del entorno.

2. Revisión de conocimientos

El análisis multimodal para la localización de personas es un tema que ha recibido un gran interés dentro de la comunidad científica. Existen aportaciones individuales para cada tecnología que son relevantes para la presente Tesis y también aportaciones que comparten el objetivo multimodal. Su análisis detallado permite enmarcar las aportaciones planteadas para la Tesis.

2.1. Sistemas de localización basados en audio

El sistema basado en audio debe ser capaz de estimar el número de hablantes activos en cada instante de tiempo (detección), así como su posición instantánea (localización). La señal de audio es por naturaleza una señal intermitente y por tanto no se deben realizar

estimaciones durante los períodos de silencio. Los detectores de voz tradicionales (Voice Activity Detectors, VAD) emplean características individuales del canal para calcular las métricas relevantes, como por ejemplo niveles de energía, cruces por cero, combinadas con reglas de clasificación basadas en umbrales fijos o recalculados en los períodos de silencio. Esta forma de actuar hace que presenten problemas en entornos con baja SNR y especialmente con ruidos no estacionarios. Otros métodos alternativos consideran la correlación cruzada entre canales [Lathoud and Magimai-Doss, 2005], o estrategias como la presentada en [Armani et al., 2003] basada en la Cross-Power Spectrum (CSP) de la señal capturada. Una de las alternativas más prometedoras para resolver este problema es la presentada por investigadores del IDIAP en [Lathoud and Odobez, 2007] y denominada short-term clustering, la cual no necesita conocer el número de objetivos a detectar, aplicando técnicas de agrupación online no supervisada. En cuanto al problema de la localización, los algoritmos que emplean agrupaciones (arrays) de micrófonos, pueden ser clasificados en tres grandes estrategias [McCowan, 2001]:

Las que emplean Time Differences of Arrival (TDOA) [Varma, 2002], que calculan en primer lugar las diferencias de retardo de las señales recibidas y luego estiman los posibles lugares geométricos que cumplen dichas restricciones temporales. Su principal ventaja es el reducido coste computacional a costa de unas limitadas prestaciones en entornos fuertemente reverberantes o con bajas SNRs. Estrategias de mejora incluyen el uso de modificaciones en la función de cálculo de la correlación (filtrados PHAT, por ejemplo, como se describe en [DiBiase, 2000]).

Las basadas en High-resolution Spectral Estimation, que se apoyan en la explotación de las propiedades de la Cross-Sensor (spatial) Covariance Matriz (CSCM) del array, a partir de la cual son capaces de dividirla en dos subespacios, uno que contiene la señal de habla y otro el ruido. Su principal inconveniente es que no es posible su implementación cuando el número de llegadas incorreladas supera al número de sensores del array (condición típica en entornos realistas).

Las que utilizan Steered Response Power (SRP), que hacen uso de determinadas técnicas de conformación de haz (beamforming), de modo que el array puede ser utilizado para escanear una serie de localizaciones espaciales conocidas [DiBiase, 2000]. Dos grandes ventajas de esta estrategia son la posibilidad de aceptar un número variable y múltiple de targets simultáneos [Wax and Kailath, 1983] y su rendimiento, muy superior a las basadas en TDOA. Su principal inconveniente es la alta carga computacional, compensable en parte con aproximaciones de tipo jerárquico como la descrita en [Zotkin and Duraiswami, 2004].

2.2. Sistemas de localización basados en vídeo

El uso de cámaras fijas ubicadas en el entorno, como fuente de información precisa para el análisis de un escena, plantea numerosos e importantes problemas específicos de la visión artificial. (deformación de perspectiva, cambios de iluminación, oclusiones, etc.).

La localización de personas mediante cámaras es uno de los pasos claves para realizar el análisis de una escena compleja. En la literatura se ha hecho tradicionalmente un gran esfuerzo en solucionar el problema desde un punto de vista de una sola cámara. En [Yilmaz et al., 2006] se puede encontrar un extenso estudio sobre las principales aportaciones en este tema. Uno de los principales problemas que presenta la localización de personas es

la ausencia de restricciones simples sobre su apariencia y geometría. Existen soluciones que permiten, bajo ciertas restricciones, detectar ciertas partes del cuerpo humanos, como puede ser la cara o las manos [Viola and Jones, 2004, Smith et al., 2008]. La viabilidad de estas alternativas dependen de la resolución de la cámara y el grado de oclusiones que existen en la escena.

Los métodos que utilizan una sola cámara están lejos de poder analizar escenas complejas, especialmente aquellas donde el número de personas en el entorno es grande, y que en la literatura anglosajona se conocen como "crowded scenes" o escenas densamente pobladas. En este tipo de escenarios, los enfoques multi-cámara permiten mejorar enormemente los sistemas de detección y localización de personas, pudiendo proporcionar medidas métricas de la posición gracias a las restricciones que aporta la geometría de múltiples cámaras. Ejemplos notables de este tipo de técnicas son [Mikic et al., 1998,Fleuret et al., 2008,Khan and Shah, 2008]. En muchos de estos métodos se adopta un enfoque que discretiza el espacio de posible ocupación de las personas. Este tipo de enfoque está relacionado con las técnicas de Visual Hull o Shape From Silhouttes. La definición de ocupación para cada una de las celdas o vóxeles que discretizan el espacio de búsqueda puede atribuirse a un sistema de detección de fondo [Mikic et al., 1998, Fleuret et al., 2008, Pizarro et al., 2008] o a un criterio de ocupación mas complejo y basado en distribuciones de probabilidad, como el mostrado en [Khan and Shah, 2008, Utasi and Benedek, 2011].

Estas técnicas muestran resultados espectaculares en escenas complejas. El principal problema radica en el hecho de que dependen en gran medida de modelos del fondo de la escena y del número de cámaras utilizado. En configuraciones reales es difícil y costoso asegurar que toda la zona de interés es vista por un conjunto elevado de cámaras. Los modelos de fondo son también un problema no solucionado. En [Oh et al., 2011] se muestra una comparación de los métodos más utilizados para modelar el fondo de una escena y las problemáticas asociadas.

2.3. Sistemas de localización multimodal

La localización de personas en entornos multimodales (audio y video) ha recibido, al igual que sus homólogos en audio y video, una gran atención de la comunidad científica [Pingali et al., 1999, Hershey and Movellan, 2000, Cutler and Davis, 2000, Pavlovic et al., 2000, Fisher et al., 2001, Aarabi and Zaky, 2001, Chen and Rui, 2004, Checka et al., 2004]. Ambas tecnologías son esencialmente complementarias, ofreciendo claras motivaciones para su fusión. Los trabajos existentes se clasifican fundamentalemente atendiendo a tres criterios: nivel de objetivos (una o múltiples personas), enfoque de detección y seguimiento y la configuración y número de sensores.

Existen numerosos trabajos clásicos que abordan el problema de seguir una sola persona, que o bien se encuentra sola [Cutler and Davis, 2000, Pavlovic et al., 2000, Aarabi and Zaky, 2001] o es el hablante activo en ese momento dentro de un grupo de posibles [Pingali et al., 1999, Hershey and Movellan, 2000, Fisher et al., 2001]. Recientemente se ha abordado el problema más completo, que consiste en seguir de manera continua y a partir de audio y video la posición y actividad de un conjunto variable de personas [Cutler et al., 2002, Siracusa et al., 2003, Busso et al., 2005]. Estos enfoques tienen como antecendentes la gran actividad investigadora que se ha realizado por separado en audio y video anteriormente comentada.

En cuanto al enfoque utilizado en detección y seguimiento destacan los métodos probabilísticos, denominados métodos generativos, y que en general se basan en técnicas bayesianas como los filtros de partículas [Checka et al., 2004, Chen and Rui, 2004, Asoh et al., 2004]. Este tipo de métodos ofrecen un marco de trabajo para realizar el seguimiento de múltiples personas y la fusión de múltiples fuentes de información de una forma natural. Las alternativas existentes varían en el nivel de sofisticación pero siempre bajo un marco de trabajo bayesiano.

La configuración de sensores varía en los trabajos anteriormente comentados en función de los objetivos y la resolución requerida en la localización. Existen numerosas bases de datos de acceso público para investigación con diferentes configuraciones audiovisuales. Destaca por ejemplo la base de datos creada por el consorcio AMI (Augmented Multi-Party Interaction) [Carletta et al., 2006], cuya configuración de sensores incluye dos "arrays" de 8 micrófonos circulares y 3 cámaras rodeando la habitación.

Los métodos existentes en la literatura no están exentos de carencias y problemas sin resolver. La localización de personas en entornos multimodales es todavía una disciplina emergente que depende en gran medida de soluciones a muchos problemas no resueltos en el campo de visión y audio.

El grupo GEINTRA posee una sala dotada con más de 6 cámaras y 4 "arrays" de 4 micrófonos dispuestos en una configuración en forma de "T".

3. Objetivos de la Tésis

El objetivo de la Tesis es investigar técnicas de optimización moderna (optimización convexa y optimización L_1 [Donoho, 2006] [Baraniuk, 2007]) en su aplicación a la localización multimodal (audio y video) de personas en espacios inteligentes.

Mediante dichas herramientas es posible abordar el problema de localización desde una perspectiva novedosa, donde la posición de cada uno de las personas resulta de la solución de un problema de optimización, generalmente convexo. Mediante este enfoque es posible integrar de forma natural un modelado físico de cada sensor involucrado (agrupaciones de micrófonos y cámaras de video) en el proceso de localización. El problema planteado es generalmente convexo, lo que admite una solución global para resolverlo, lo que es una ventaja añadida con respecto a las técnicas existentes.

En la Tesis se esperan conseguir resultados cercanos o superiores a los sistemas considerados referencia en el estado del arte.

Del mismo modo, este tipo de enfoque basado en técnicas L_1 permite abordar problemas más ambiciosos como puede ser la calibración automática de redes de sensores o la reducción del número de ellos sin por ello sufrir una pérdida sustancial en los resultados.

4. Hipótesis de trabajo, material y método de estudio

4.1. Hipótesis de trabajo

Con el objetivo de cumplir los objetivos planteados para la presente Tesis se plantea la siguiente metodología de trabajo, dividida en las siguientes fases principales:

- Estudio de los sistemas sensoriales, tanto de video como de audio, así como de las técnicas y modelos físicos que son necesarios para plantear una localización métrica de una persona en una escena tridimensional.
- Estudio de técnicas de optimización convexa y técnicas L_1 y su uso en aplicaciones como el "Compressed Sensing".
- Planteamiento del problema de la localización como un problema de optimización L_1 convexo.
- Establecer una metodología de experimentación basada en el uso de bases de datos multimodales de uso generalizado en la comunidad científica y envío de artículos con los resultados en los foros que se consideren de mayor impacto.
- Generar un demostrador en el laboratorio del grupo de investigación.

4.2. Material

Para alcanzar los objetivos de la investigación es necesario el siguiente material hardware y software:

4.2.1. Material hardware

- Computadores de alto rendimiento para cálculo intensivo.
- Cámaras a color de baja resolución capaces de capturar a 25 imágenes por segundo y sistemas de sincronización.
- Agrupaciones de micrófonos y etapas de amplificación y adquisición de alta velocidad.
- Soportes omnidireccionales para las cámaras y "arrays" de micrófonos.
- Conjunto de cables de red ethernet y FireWire.
- Cableado de audio de alta calidad entre los "arrays" de micrófonos y las etapas de amplificación y adquisición.

4.2.2. Material software

- Sistema Operativo Linux
- Librerías C para visualización en 3D y procesamiento de imágenes.
- Librerías C para el procesamiento de la señal de audio proveniente de los micrófonos.
- Herramientas software matemáticas para análisis de imágenes y calibración de cámaras.

4.3. Planificación temporal

Para elaborar esta Tesis se han distribuido cuatro años planificados de la siguiente forma:

- Estudio del estado del arte: 6 meses.
- Desarrollo de los algoritmos: 12 meses.
- Verificación experimental: 12 meses.
- Redacción de tesis: 6 meses.

5. Conclusiones

Esta Tesis tiene como objetivo abordar el problema de localización de personas en entornos dotados con sensores de audio y video. El problema resulta de absoluta novedad y tiene un gran interés científico y social. Aunque existe un gran volumen de trabajo científico ya publicado sobre el tema existen numerosos problemas sin una solución considerada madura. En la Tesis se propone un enfoque novedoso basado en técnicas modernas de optimización, que permiten incluir complejos modelos físicos de cada sensor y solucionar de manera global el problema de la localización.

Referencias

[Aarabi and Zaky, 2001] Aarabi, P. and Zaky, S. (2001). Robust sound localization using multi-source audiovisual information fusion. *Information Fusion*, 2(3):209–223.

[Armani et al., 2003] Armani, L., Matassoni, M., Omologo, M., and Svaizer, P. (2003). Use of a csp-based voice activity detector for distant-talking asr. In *Eighth European Conference on Speech Communication and Technology*.

[Asoh et al., 2004] Asoh, H., Asano, F., Yoshimura, T., Yamamoto, K., Motomura, Y., Ichimura, N., Hara, I., and Ogata, J. (2004). An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion. In *Proc. Fusion*, pages 805–812. Citeseer.

[Baraniuk, 2007] Baraniuk, R. (2007). Compressive sensing. *Lecture notes in IEEE Signal Processing magazine*, 24(4):118–120.

[Busso et al., 2005] Busso, C., Hernanz, S., Chu, C., Kwon, S., Lee, S., Georgiou, P., Cohen, I., and Narayanan, S. (2005). Smart room: participant and speaker localization and identification. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 2, pages ii–1117. IEEE.

[Carletta et al., 2006] Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al. (2006). The ami meeting corpus: A pre-announcement. *Machine Learning for Multimodal Interaction*, pages 28–39.

[Checka et al., 2004] Checka, N., Wilson, K., Siracusa, M., and Darrell, T. (2004). Multiple person and speaker activity tracking with a particle filter. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 5, pages V–881. IEEE.

[Chen and Rui, 2004] Chen, Y. and Rui, Y. (2004). Real-time speaker tracking using particle filter sensor fusion. *Proceedings of the IEEE*, 92(3):485–494.

[Cutler and Davis, 2000] Cutler, R. and Davis, L. (2000). Look who's talking: Speaker detection using video and audio correlation. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 3, pages 1589–1592. IEEE.

[Cutler et al., 2002] Cutler, R., Rui, Y., Gupta, A., Cadiz, J., Tashev, I., He, L., Colburn, A., Zhang, Z., Liu, Z., and Silverberg, S. (2002). Distributed meetings: A meeting capture and broadcasting system. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 503–512. ACM.

[DiBiase, 2000] DiBiase, J. (2000). *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. PhD thesis, Citeseer.

[Donoho, 2006] Donoho, D. (2006). For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829.

[Fisher et al., 2001] Fisher, J., Darrell, T., Freeman, W., and Viola, P. (2001). Learning joint statistical models for audio-visual fusion and segregation. *Advances in neural information processing systems*, pages 772–778.

[Fleuret et al., 2008] Fleuret, F., Berclaz, J., Lengagne, R., and Fua, P. (2008). Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 267–282.

[Hershey and Movellan, 2000] Hershey, J. and Movellan, J. (2000). Audio-vision: Using audio-visual synchrony to locate sounds. In *Advances in Neural Information Processing Systems 12*. Citeseer.

[Khan and Shah, 2008] Khan, S. and Shah, M. (2008). Tracking multiple occluding people by localizing on multiple scene planes. *IEEE transactions on pattern analysis and machine intelligence*, pages 505–519.

[Lathoud and Magimai-Doss, 2005] Lathoud, G. and Magimai-Doss, M. (2005). A sector-based, frequency-domain approach to detection and localization of multiple speakers. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 3, pages iii–265. IEEE.

[Lathoud and Odobez, 2007] Lathoud, G. and Odobez, J. (2007). Short-term spatio–temporal clustering applied to multiple moving speakers. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(5):1696–1710.

[McCowan, 2001] McCowan, I. (2001). Robust speech recognition using microphone arrays.

[Mikic et al., 1998] Mikic, I., Santini, S., and Jain, R. (1998). Video processing and integration from multiple cameras. In *Proceedings of the 1998 Image Understanding Workshop, Morgan-Kaufman, San Francisco*, volume 6. Citeseer.

[Oh et al., 2011] Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C., Lee, J., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al. (2011). A large-scale benchmark dataset for event recognition in surveillance video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 6.

[Pavlovic et al., 2000] Pavlovic, V., Rehg, J., Garg, A., and Huang, T. (2000). Multimodal speaker detection using error feedback dynamic bayesian networks. In *cvpr*, page 2034. Published by the IEEE Computer Society.

[Pingali et al., 1999] Pingali, G., Tunali, G., and Carlbom, I. (1999). Audio-visual tracking for natural interactivity. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 373–382. ACM.

[Pizarro et al., 2008] Pizarro, D., Marron, M., Peon, D., Mazo, M., Garcia, J., Sotelo, M., and Santiso, E. (2008). Robot and obstacles localization and tracking with an external camera ring. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 516–521. IEEE.

[Siracusa et al., 2003] Siracusa, M., Morency, L., Wilson, K., Fisher, J., and Darrell, T. (2003). A multi-modal approach for determining speaker location and focus. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 77–80. ACM.

[Smith et al., 2008] Smith, K., Ba, S., Odobez, J., and Gatica-Perez, D. (2008). Tracking the visual focus of attention for a varying number of wandering people. *IEEE transactions on pattern analysis and machine intelligence*, pages 1212–1229.

[Utasi and Benedek, 2011] Utasi, A. and Benedek, C. (2011). A 3-d marked point process model for multi-view people detection. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3385–3392. IEEE.

[Varma, 2002] Varma, K. (2002). *Time-delay-estimate based direction-of-arrival estimation for speech in reverberant environments*. PhD thesis, Citeseer.

[Viola and Jones, 2004] Viola, P. and Jones, M. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.

[Wax and Kailath, 1983] Wax, M. and Kailath, T. (1983). Optimum localization of multiple sources by passive arrays. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 31(5):1210–1217.

[Yilmaz et al., 2006] Yilmaz, A., Javed, O., and Shah, M. (2006). Object tracking: A survey. *Acm Computing Surveys (CSUR)*, 38(4):13.

[Zotkin and Duraiswami, 2004] Zotkin, D. and Duraiswami, R. (2004). Accelerated speech source localization via a hierarchical search of steered response power. *Speech and Audio Processing, IEEE Transactions on*, 12(5):499–508.