SEGUIMIENTO AUDIOVISUAL DE LOCUTOR USANDO UN FILTRO DE PARTÍCULAS EXTENDIDO CON PROCESO DE CLASIFICACIÓN

F. Sanabria-Macías[1], J. Macías-Guarasa[2], M. Marrón-Romera[2], D. Pizarro[2] y E. Marañón-Reyes[1]

[1]Grupo de Procesamiento de Voz, CENPIS – Universidad de Oriente, Santiago de Cuba – Cuba

[2]Grupo GEINTRA - Departamento de Electrónica – Universidad de Alcalá, Alcalá de Henares – España

RESUMEN

- Se describe el diseño, implementación y evaluación de un sistema de seguimiento de locutores usando fusión audiovisual
- Un bloque de audio detecta regiones con actividad a partir de una búsqueda por intersección de sectores y el algoritmo Steered Response Power
- Un bloque de vídeo detecta rostros en cada cámara, con Viola & Jones, y los proyecta sobre un plano
- Un filtro de partículas extendido realiza el seguimiento de los datos fusionados
- El sistema ha sido evaluado usando la base de datos AV16.3 con resultados prometedores

INTRODUCCIÓN

- **Espacios Inteligentes:**
  - ⇨Entornos dotados de un conjunto de sistemas sensoriales, de comunicación, y de cómputo transparentes e imperceptibles a los usuario
  - ⇨Perciben el entorno y cooperan entre sí para ayudar en la interacción con los usuarios
  - ⇨La información es extraída con un conjunto de sensores ubicados en el entorno, fundamentalmente cámaras de vídeo y agrupaciones de micrófonos (*arrays*)
- En este contexto se busca la detección, localización y seguimiento de los ocupantes del entorno
- Los métodos que realizan seguimiento de personas combinando información de varias fuentes se denominan de seguimiento multimodal

SEGUIMIENTO AUDIOVISUAL

- **Audio**
  - ⇨Localización basada en diferencias de tiempos de llegada de la voz a los micrófonos
  - ⇨Steered Response Power (SRP): evalúa actividad acústica en localizaciones específicas, orientando el patrón de directividad del array (*beamforming*)
  - ⇨Desventaja: ↑ precisión ⇒ ↑ densidad de localizaciones ⇒ ↑ costo computacional
  - ⇨Alternativa: detección basada en sectores
- **Video**
  - ⇨Detección: detección de rostros en 2D (color, apariencia, etc.)
  - ⇨Visual Hull: Proyección y combinación de detecciones por cámara a 3D
- **Fusión audiovisual**
  - ⇨Orientados a Sistema vs. Orientados a Modelo
- **Novedades de la propuesta**
  - ⇨Detección y localización conjunta + SRP
  - ⇨Filtro de partículas extendido con proceso de clasificación (XPFCP) en contexto de seguimiento audiovisual

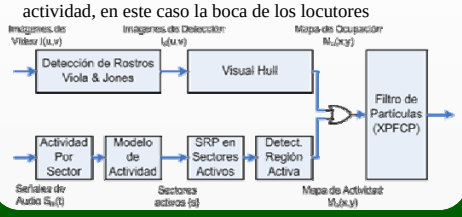
PROPUESTA DESARROLLADA

1. Esquema General:

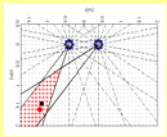
El sistema combina dos mapas (grid), uno de ocupación y otro de actividad sonora en un plano:

⇨El grid de ocupación se genera con la información visual mientras que el de actividad se obtiene a partir de las señales de los micrófonos

⇨La altura del plano es constante y se selecciona de modo que coincida aproximadamente con la de la fuente de actividad, en este caso la boca de los locutores



2. Grid de actividad acústica:



⇨Detección basada en sectores esféricos y centrados en cada array

⇨SAM SPARSE MEAN, evalúa índice de actividad en el volumen del sector a partir de una métrica de fase

⇨Umbral fijo para detectar sectores activos

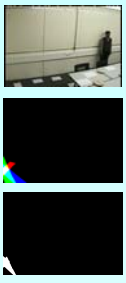
⇨En regiones de "Intersección de sectores activos" del plano de actividad se realiza una búsqueda puntual del máximo de actividad por dos métodos:

- Búsqueda exhaustiva con SRP
- Minimización de métrica de fase

⇨Crecimiento de regiones alrededor de los máximos

3. Grid de actividad visual

- ⇨ Se aplica el algoritmo Viola & Jones a cada imagen por cámaras
- ⇨ Los rostros detectados en cada imagen son proyectados mediante homografía, al plano de ocupación
- ⇨ El resultado es la unión de las intersecciones dos a dos entre las detecciones de cada cámara



4. Fusión audiovisual y XPFCP:

- ⇨ OR-lógico de ambos grids de actividad
- ⇨ XPFCP filtra los datos fusionados
  - Clasificación de las medidas de entrada
  - Clasificación de las partículas
  - Centroide de las clases de partículas definen la posición de los usuarios

CONFIGURACIÓN EXPERIMENTAL

Base de datos

⇨ AV16.3:

- 3 secuencias de vídeo a 25 fps
- 2 arrays circulares de 8 micrófonos, con frecuencia de muestreo 16kHz

⇨ Secuencias seleccionadas

secuencia	duración	modalidad
seq01-1p-0000	217	ST
seq02-1p-0000	189	ST
seq03-1p-0000	242	ST
seq11-1p-0100	30	MV
seq15-1p-0100	35	MV

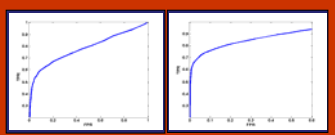
Métricas de evaluación

- ⇨ ***Pcor***: porcentaje de tramas activas con un error inferior a 50cm.
- ⇨ **Error promedio de localización**: Promedio de los errores de localización con respecto a la posición etiquetada manualmente [mm]
- ⇨ **Tasa de borrados**: Falsos negativos, ventanas acústicamente activas no detectadas como tales
- ⇨ **TPR**: Tasa de verdaderos positivos, calculada como el porcentaje de tramas con actividad de voz detectados como activos
- ⇨ **FPR**: Tasa de falsos positivos, calculada como el porcentaje de tramas sin actividad de voz detectados como activos

RESULTADOS Y DISCUSIÓN

Evaluación de detección por sectores

- Como detector y detector-localizador de voz
  - ⇨Curva ROC no presenta buenas prestaciones
- Comportamiento similar con y sin intersección de sectores



Evaluación del bloque de localización puntual

- SBD+SRP superior en localización, a costa de aumento en tasa de borrados con respecto a SRP

	SBD+SCG	SBD+SRP	SRP
Pcor[%]	76	96	79
Error promedio [mm]	524	161	478
Tasa borrados [%]	33	33	0

Evaluación del sistema de seguimiento

- AV supera significativamente a Audio, no así al seguimiento con Vídeo

	Audio	Vídeo	A+V
Pcor[%]	91	100	99
Error promedio [mm]	263	171	170
Tasa borrados [%]	80	33	31

- SBD falla en la detección de inicio y fin de tramos de voz.
  Posibles soluciones:
  - ⇨ combinar métrica SSM con otras características propias de la voz.
  - ⇨ Umbral adaptativo
- Localización 2D no modela variaciones de altura de un mismo locutor y entre locutores.
- Fusión "lógica" de audio y vídeo, no es suficiente para modelar la relación AV
  Alternativas:
  - ⇨ Pesado de importancia de las medidas

CONCLUSIONES Y LÍNEAS FUTURAS

1. Método de seguimiento audiovisual con propuestas de:
   - "Intersección de sectores activos" de múltiples arrays, ⇒ reducción mayor del espacio de búsqueda
   - Uso por primera vez del XPFCP en un contexto de fusión audiovisual
2. Resultados AV superior a audio, similar a vídeo, debido a alta tasa de borrados en audio
3. Modelo de fusión mejorable
4. Localización 3D en versiones futuras

SAAEI'2011 – Seminario Anual de Automática, Electrónica Industrial e Instrumentación – Badajoz, 6-8 de julio de 2011