

¡200 Petabytes a 100 Megadólares! Surá – Clúster para el almacenamiento y procesamiento de datos científicos del PRIS-Lab

Francisco Siles Canales,
Dr. rer. nat. (in fieri)
fsiles@eie.ucr.ac.cr

Ing. Saúl Calderón Ramírez
saul1917@gmail.com

Ing. Marco Villalta Fallas
mvillalta@eie.ucr.ac.cr



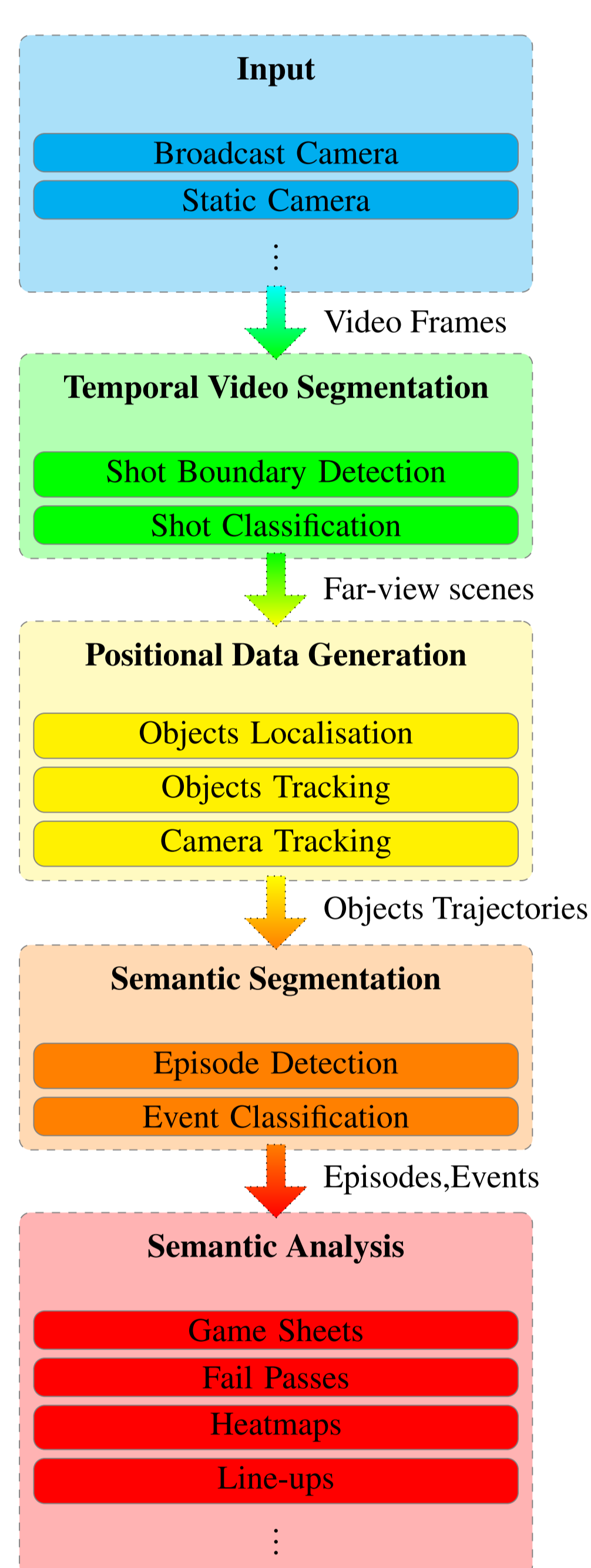
Pattern Recognition and Intelligent Systems Laboratory

Universidad de Costa Rica | Facultad de Ingeniería | Escuela de Ingeniería Eléctrica
<http://pris.eie.ucr.ac.cr>, Apartado Postal: 115012060 UCR, San José, Costa Rica

Introducción

El desarrollo de sistemas computacionales más poderosos en las últimas décadas ha facultado el dar soporte a nuevas aplicaciones, en particular aquellas relacionadas con el análisis de actividades humanas [1-4]. En el caso del deporte, el fútbol ha sido un foco de atención para muchos investigadores interdisciplinarios, debido a ser uno de los deportes más ricos, complejos y diseminados practicado actualmente. Para solventar, de una forma objetiva y eficiente, las necesidades de información acerca del juego de estos investigadores, se requiere investigar y desarrollar sistemas computacionales automatizados [5-20].

El proyecto ACE-Football del PRIS-LAB consiste en desarrollar un sistema inteligente para el análisis automatizado de fútbol. El sistema implementa módulos de percepción para generar las trayectorias seguidas por los jugadores durante el juego; y módulos cognitivos para modelar las acciones, eventos y episodios a partir de las trayectorias generadas. El proyecto de investigación **Rastreo automatizado de jugadores de fútbol a partir de señales de televisión**, inscrito en la Vicerrectoría de Investigación de la UCR, se concentra en investigar la etapa de percepción.



Problema

El sistema ACE-Football utiliza videos de televisión, por ser un mecanismo asequible y barato, en contraposición a los sistemas dedicados con cámaras estereoscópicas instalados en algunos estadios europeos, con costos superiores a 100 mil euros. Videos de los mundiales de fútbol, por ejemplo, se utilizan para validar los algoritmos desarrollados, y son almacenados en formatos sin compresión debido a la necesidad de acceder en tiempo real, a cualquier imagen del video en cualquier instante de tiempo (acceso aleatorio). Además de los videos originales, los videos de salida procesados, con información estadística, táctica y estratégica adicionada, deben ser almacenados también, al igual que los datos científicos de posiciones, acciones, episodios y demás. Lo anterior representa un serio problema por cuanto el volumen de datos por preservar es masivo.

Por ejemplo, un juego de 90 minutos, a 30 imágenes por segundo (estándar estadounidense de televisión), con una resolución por imagen de 1920x1080 píxeles, donde cada píxel a color codifica 3 Bytes, requiere aproximadamente de 1TB (un terabyte) de almacenamiento. Así, para almacenar los siguientes 4 mundiales, cada uno con 64 juegos, donde cada video de entrada producirá al menos un video de salida procesado, y se desea un respaldo de al menos una copia (RAID 10), el gran total es de aproximadamente 1PB (un petabyte). Empresas de almacenamiento masivo [21], por un sistema de estas dimensiones cobraría aproximadamente 300 mil dólares. Por otro lado, este análisis, es muy conservador, pues ignora la necesidad de almacenar otros torneos: Mundial Femenino, Mundial Juvenil, *Champions League*, *Bundesliga*. Además, actualmente están aprobados por la Unión de Telecomunicaciones Internacional (ITU) [22], resoluciones de imagen de 3840x2160 y 7680x4320. En resumen, se requerirá en el futuro cercano 200PB aproximadamente, lo cual representaría una inversión imposible de ¡100 millones de dólares! o (100 Megadólares), lo cual por si fuera poco, es imposible de construir con la tecnología actual de discos duros magnéticos-mecánicos.

Referencias: <http://pris.eie.ucr.ac.cr>

Propuesta

Claramente es imprescindible utilizar mecanismos de compresión de video para reducir la capacidad de almacenamiento requerida, sin eliminar la característica de acceso aleatorio necesaria. Para dar soporte a la aplicación científica presente, y proveer las imágenes a la velocidad requerida, es necesario contar con procesamiento de alto rendimiento para descomprimir el video y aplicar los algoritmos de percepción, produciendo en tiempo real los resultados esperados. Por ello, se implementó **SURA**, un clúster Beowulf en Linux, cuyos componentes fueron donados a la Escuela de Ingeniería Eléctrica por Intel de Costa Rica. El clúster consiste en un nodo maestro y tres esclavos. El nodo maestro es un servidor Dell Power Edge T710 con 2 procesadores Quad-Core E5560, con 2 discos duros SATA de 500GB a 7200rpm. Cada nodo esclavo consiste en una Dell Vostro 430 con procesador Core i5 750, con 1 disco duro SATA de 250GB a 7200rpm y una NVIDIA GeForce G310.

Resultados

Al comprimir los videos utilizando un codec h264 en formato avi, se redujo la memoria necesaria por juego a solamente 60GB de almacenamiento, así para almacenar los mundiales deseados se requerirá de aproximadamente 60TB, unas mil veces menos.

En cuanto a la capacidad de procesamiento, se implementó una de las transformaciones utilizadas en ACE-Football: la conversión entre espacios de color de RGB a HSV (ver figura), y se probó en distintas resoluciones de imágenes, tanto en **SURA**, como en una plataforma de referencia con un CPU Core i7 2600 utilizando una GPU Geforce GTX760ACX con 1152 CUDA Cores. Los resultados se muestran en el cuadro, donde se aprecia una ventaja de tiempo de ejecución del orden de 10, favorable para **SURA**. El promedio de procesamiento por imagen en **SURA** es de 0,30 ms, lo cual corresponde a unas 100 veces menos que el mínimo tiempo requerido para desplegar 30 imágenes por segundo a una resolución de 1920x1080, es decir se alcanza la velocidad necesaria de procesamiento de las imágenes.



Resolución (pel x pel)	Tiempo (ms)	
	Referencia	SURA
256x192*	0,97	0,28
720x576	0,41	0,28
1280x720	0,75	0,27
1920x1080	1,76	0,32
2560x1440*	2,97	0,36
Promedio	1,37	0,30

* corresponde a las figuras

Conclusiones

La implementación de **SURA** permitió reducir las necesidades de almacenamiento, utilizando mecanismos de compresión de video, sin afectar negativamente el acceso aleatorio a los datos de los videos. Además, comparado con el sistema de referencia, **SURA** logró un mejor desempeño en una de las operaciones típicas utilizadas en uno de los algoritmos del proyecto, con una ventaja del orden de 10.

Actualmente **SURA** está siendo actualizada, utilizando Lustre como sistema de archivos distribuido, y Git como mecanismo de control de versiones para el desarrollo de las aplicaciones. También se está desarrollando una librería híbrida CPU/GPU de reconocimiento de patrones para aprovechar al máximo el hardware disponible. Actualizaciones de hardware se llevan a cabo también.