Datos no Estructurados No Textuales: Desarrollo de Nuevas Tecnologías

J. Fernandez, N. Miranda, R. Guerrero, F. Piccoli
LIDIC- Universidad Nacional de San Luis
Ejército de los Andes 950
Tel: 02652 420823, San Luis, Argentina
{jmfer, ncmiran, rag, mpiccoli}@unsl.edu.ar

Resumen

Cuando una persona recibe estímulos sensoriales de tipo visual o auditivo reacciona realizando una asociación y reconocimiento en forma natural, como consecuencia de la información que los estímulos le brindan. Durante los últimos años, el avance de los medios digitales y la proliferación de su uso ha generado la necesidad del desarrollo de herramientas que permitan la eficiente representación, procesamiento y administración (acceso y recuperación) de información de contenido multimedial. En este contexto, la información almacenada principalmente en forma de audio, imagen y video se ha convertido en la principal materia prima utilizada por los sistemas computacionales para la transmisión de información en forma rápida y eficiente, principialmente aquella relacionada con la toma de decisiones y la resolución de problemas de índole general. Dadas sus particularidades, dicha información es catalogada como información no estructurada y su administración y manipulación requiere de la definición de nuevos procesos y métodos que faciliten y agilicen el uso de la misma. Esta propuesta de trabajo establece los lineamientos a seguir con la intención de redefinir nuevas tecnologías para el procesamiento de información no estructurada que permita la incorporación de la misma en procesos generales de resolución de problemas o toma de decisiones.

Palabras Claves: Procesamiento de Señales, Procesamiento de Imágenes, Visión por Computadora, Computación Gráfica, Vector Característica, Computación Paralela.

Contexto

Esta propuesta de trabajo se lleva a cabo dentro de la línea de Investigación "Procesamiento de Información Multimedia" del proyecto "Nuevas Tecnologías para un tratamiento integral de Datos Multimedia". Este proyecto es desarrollado en el ámbito del Laboratorio de Investigación y Desarrollo en Inteligengia Computacional (LIDIC) de la Universidad Nacional de San Luis.

1. Introducción

La administración de datos no estructurados es uno de los mayores problemas aún no resueltos en la industria de la tecnología de la información. La principal razón radica en que las herramientas y técnicas existentes han sido desarrolladas para el tratamiento de información estructurada pero fallan al momento de procesar información no estructurada.

No es un secreto que gran cantidad de la información que manejan actualmente las empresas se encuentra organizada en archivos y documentos no estructurados. Debido a ello, un gran número de organizaciones han tomado conciencia de que la consolidación, acceso y procesamiento de datos no estructurados es un factor importante para la optimización y el análisis de los procesos empresariales.

Cuando se hace mención a datos no estructurados o información no estructurada se hace referencia a información de computadora que no tiene un modelo formal de dato que la represente o que, en caso de poseer uno, no es fácilmente utilizable por un programa de computadora. El término distingue este tipo de información de aquella que se encuentra almacenada en bases de datos y organizada a través de campos.

En particular, el entorno de la Web ha propiciado la ocurrencia de situaciones en donde es necesario el tratamiento de información, la cual se encuentra compuesta por datos que:

- Poseen una estructura formal definida, no obstante su estructura no es útil para el desarrollo de ciertas tareas de procesamiento por lo que deberían ser categorizados como no estructurados.
- Si bien no tienen una estructura formalmente definida, ésta se encuentra implícita. Desde el punto de vista humano, dicha estructura puede ser inferida a partir de la simple observación (imagen o video), la escucha (relato o canción), o la lectura (texto) reconociendo interrelaciones, morfologías, sintáxis, etc..
- En forma individual no poseen una estructura definida, sin embargo ellos suelen encontrarse empaquetados en objetos tales como archivos o documentos los cuales tienen una estructura explícita (documentos multimediales, páginas web, etc.).

En consecuencia, todo software que pretenda automatizar la inferencia de información debería crear estructuras aptas para ser procesadas en forma automática por una máquina, donde se explote la estructura lingüistica, auditiva y visual inherente a todas las formas de comunicación humana. Es decir, debe aumentarse la flexibilidad de los sistemas los cuales deberían extraer significado de los datos no estructurados, para luego utilizarlo en la identificación de interrelaciones y la administración de los mismos.

En la actualidad existen algoritmos y bibliotecas que intentan extraer información de los datos estructurados, yendo desde consultas XPath y analizadores de protocolos hasta el procesamiento de lenguaje natural (Natural Language Processing) y Visión por Computadora. Dependiendo del análisis final a realizar a los datos será el tipo de algoritmos utilizados, no obstante al momento de definir un sistema, todos comparten las mismas características. Un sistema para el procesamiento de datos no estructurados de tiempo real requiere estar conformado básicamente por cuatro partes:

- Objetos de Datos no Estructurados: El procesamiento de los datos comienza con la representación de los mismos. Una plataforma debería poseer un mecanismo eficiente y robusto de representación de los objetos, en conjunto con, entre otros, la administración de memoria.
- Lenguaje Extensible: La mayor parte del procesamiento de los datos no estructurados involucra algoritmos especializados. El lenguaje debe ser extensible de modo que algoritmos con dominio específico puedan coexistir con aquellos de funcionalidad genérica. Dado que los algoritmos se encuentran mayormente disponbiles a través de bibliotecas, es importante que las API soporten lenguajes de uso corriente.
- Herramientas de autoría que consideren el uso de datos no estructurados: Las herramientas de autoría son una parte crítica de cualquier sistema moderno, y más aún es importante que dichas herramientas sean diseñadas para el procesamiento datos no estructurados en tiempo real.

- Capacidades de Clustering: Toda aplicación de procesamiento de datos no estructurado implica el uso intensivo de recursos. La distribución de la carga entre un gran número de servidores es una técnica de escalado crítica así como también el uso de multi-threads en un único servidor.

En función a lo planteado, las investigaciones para el tratamiento de la problemática a abordar podrían organizarse acorde con cuatro grandes objetivos:

- *La obtención de una representación robusta para cada objeto no estructurado.* Intentando el reconocimiento y clasificación en forma automática para la resolución de problemas tradicionalmente complejos tales como la reconstrucción de escenarios 3D y el seguimiento de objetos en movimiento, entre otros [3, 15, 1, 2, 7, 9, 12, 13, 14, 16, 23]. Así como también la definición de nuevas estrategias de almacenamiento y recuperación desde grandes repositorios de almacenamiento [41, 42].
- *La transferencia de los logros obtenidos con los diferentes datos no estructurados.* Intentando utilizar los métodos o técnicas desarrollados con las nuevas representaciones en otros tipos de objetos [4, 6, 10, 11, 20, 21, 22, 24, 27, 28, 29, 33, 34, 35, 37, 38, 39].
- *El abordaje de problemas multimedia.* Donde la especificación de una única representación general para todos los tipos de datos no estructurados habilita el tratamiento de información multimedia en forma sistematizada con mayor probabilidad de éxito que los métodos actuales [5, 36, 40, 43, 44, 45].
- *El uso de técnicas computacionales de alto desempeño*, aplicadas al proceso de Adquisición, pre-procesamiento y análisis de datos no estructurados [7, 17, 18, 19, 25, 26, 30, 31, 32].

En la siguiente sección se detallan algunas líneas de trabajo específicas a abordar según los cuatro objetivos generales detallados.

2. Líneas de Investigación y Desarrollo

Se propone profundizar los estudios asociados con la adquisición, pre-procesamiento y análisis de datos no estructurales no textuales, como así también tratamiento de tendientes a mejorar la transmisión de información mediante el análisis y/o optimización del proceso de producción/obtención de información de contenido a partir de datos no estructurados (audio, imágenes, video, gráficas y texto) mediante el uso de métodos no convencionales.

El trabajo con datos no estructurados para la transmisión de información involucra tres áreas claramente definidas: Procesamiento de Señales, Visión por Computadora y Computación Gráfica; así como también de la interacción de éstas con otras áreas, tales como la Minería de Datos, Recuperación de Información, Computación de alto rendimiento, entre otras. En consecuencia, las líneas específicas a seguir son:

- *El tratamiento de imágenes para la segmentación bajo criterios perceptuales.* Esta línea propone determinar el impacto de la segmentación de imágenes, acorde con criterios perceptuales, en la categorización de las imágenes en clases semánticas y/o geométricas.
- *El reconocimiento y reconstrucción de escenarios 3D.* Como consecuencia de la línea anterior, a partir de la correcta identificación y categorización de los objetos interactuantes en una imagen así como de sus interrelaciones, es deseable la reconstrucción del escenario completo capturado en una imagen o colección de ellas.
- *El tratamiento de Streamings de Video para el seguimiento de objetos en*

movimiento. Dada la gran importancia del procesamiento de transmisiones de canales de TV, bases de datos de video compartido (Google, You Tube), seguridad a través de circuitos cerrados de video, derechos de autor, entre otros, se pretende concentrar el trabajo en el análisis y procesamiento de video en tiempo real.

- *La determinación de Huellas Digitales Robustas de Señales.* El objetivo consiste en desarrollar un método robusto para la determinación de huellas digitales de señales (imagen, streamings de audio o video), que permitan identificar e individualizar una señal a pesar de las distorsiones naturales (compresión, codificación analógica, entre otros) y ataques maliciosos (adición de logo, distorsión geométrica, cortes en la señal, entre otros) en forma eficiente.

- *El tratamiento de Streamings de Sonido para el reconocimiento robusto de segmentos de audio.* La identificación de objetos de audio a partir de segmentos es necesario en diferentes áreas tales como el monitoreo de canales de sonido, detección de duplicaciones, plagios, rotulado automático (MP3 modernos), consulta por ejemplos y filtrados en redes p2p, entre otros.

- *La adaptación de algoritmos efectivos para clustering de documentos en problemas de clustering y segmentación de imágenes.* Esta línea se focalizará en la transferencia de la experiencia lograda en el clustering de documentos cortos a tareas similares con otros tipos de datos no estructurados como por ejemplo clustering de imágenes y segmentación de imágenes.

Como puede observarse, desde un punto de vista computacional, los datos no estructurados y la información contenida en ellos son un nexo que habilita a la resolución de una variedad de problemas en forma colaborativa entre áreas, permitiendo una realimentación de conocimiento y enriquecimiento hacia nuevas líneas de trabajo.

3. Resultados obtenidos / esperados

Dentro de los trabajos desarrollados en el ámbito de las distintas líneas de investigación propuestas se pueden destacar los relacionados al desarrollo de un Sistema de Minería de Imágenes (SMI), principalmente a la etapa de pre-procesamiento, transformación y extracción de características relacionadas al subsistema Qué? asociado al Sistema Visual Humano (Color, Forma y Textura) y al subsistema Dónde? (Punto de Fuga y Gradiente de Texturas). Dado que un SMI demanda un alto costo de recursos y procesamiento, es necesario la búsqueda de técnicas alternativas que permitan reducir los mencionados costos, se propusieron diferentes optimizaciones aplicando modelos de computación paralela propuestas en la tesis doctoral de una integrante del grupo responsable.

La generación de vectores de características robustos depende en gran medida de la correcta selección de la información esencial de los objetos (audio, imagen, video), a ser resumida en dichos descriptores. El área de No Fotorealismo es un área con gran riqueza de métodos y técnicas que contribuyen en la determinación de identificadores robustos, considerando no sólo las características propias del objeto multimedia evaluado (como color, forma, etc.), sino también las posibles distorciones (geométricas, de luz, de calidad, entre otras) a las que pueda ser sometido. Los trabajos desarrollados han sido orientados no sólo en esta dirección, sino que además se pretende lograr la universalidad de la representación (independencia del tipo de objeto multimedia). En la actualidad, los conocimientos adquiridos han sido de gran aporte para el abordaje de nuevas investigaciones relacionadas a la extracción de características alternativas de la información de audio e imágenes sobre arquitecturas GPU-

CPU.

4. Formación de Recursos Humanos

Como resultado de las investigaciones se cuenta con una tesis de maestría concluída y dos tesis doctorales y dos de maestría en desarrollo; así como también varios trabajos de fin de carrera de la Licenciatura en Ciencias de la Computación.

Además las investigaciones se encuadran en el marco de un proyecto dentro del Programa de Promoción de la Universidad Argentina para el Fortalecimiento de Redes Interuniversitarias III en los que participa nuestra universidad junto con las universidades Michoacana (México) y de Zaragoza (España).

Referencias

[1] Blaschke, T. and Hay, G., "Object-oriented image analysis and scale-space: Theory and methods for modeling and evaluating multiscale landscape structure". International Archives of Photogrammetry and Remote Sensing 34: 22-29 (2001).

[2] A.L. Callahan and Dmitry and B. Goldgof and Ph. D and Ph. D and Thomas A. Sanocki and Melanie A. Sutton, "Function from visual analysis and physical interaction: a methodology for recognition of generic classes of objects", Journal on Image and Vision Computing, Vol 16, pp 745-763, 1998.

[3] A. Camarena-Ibarrola, E. Chavez, "On Musical Performances Identification, Entropy and String Matching", MICAI 2006.

[4] P. Carbonetto, N. de Freitas, and K. Barnard, "A statistical model for general contextual object recognition", in Proc. ECCV, 2004.

[5] (CUDA)-IEEE International Conference on Multimedia and Expo 2008 ? Pp 697:700 ? April 2008.

[6] Crockett, et al.,"A Method for Characterizing and Identifying Audio Based on Auditory Scene Analysis", AES Convention Paper 6416, presentedat the 118.sup.th Convention May 28-32, 2005, Barcelona, Spain. cited by other.

[7] Dixon, S.: Live tracking of musical performances using on-line time warping. Proc of the 8th Int Conf on Digital Audio Effects (DAFx?05) (2005)

[8] A. Ess, B. Leibe, and L. Van Gool. "Depth and appearance for mobile scene analysis". In ICCV, 2007.

[9] R. Fergus, P. Perona, and A. Zisserman. "Object class recognition by unsupervised scale-invariant learning". In IEEE Conference on Computer Vision and Pattern Recognition, 2003.

[10] M.A. Fischler and R.C. Bolles. "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography". Communications of the ACM, 24(6):381?395, June 1981.

[11] G. Kim, C. Faloutsos, and M. Hebert, "Unsupervised Modeling and Recognition of Object Categories with Combination of Visual Contents and Geometric Similarity Links", ACM International Conference on Multimedia Information Retrieval (ACM MIR), October, 2008.

[12] Hay, G.J., Marceau, D. J., Dube, P., and Bouchard, A. "A Multiscale Framework for Landscape Analysis: Object-Specific Analysis and Upscaling". Landscape Ecology 16 (6): 471-490 (2001).

[13] A. Hoiem, A. Efros, and M. Hebert, "Geometric context from a single image". In ICCV, 2005.

[14] Hoiem, A.A. Efros, and M. Hebert, "Closing the Loop on Scene Interpretation", In CVPR 2008.

[15] Ibarrola, A.C., Chavez, E.. "A robust, entropy-based audiofingerprint". IEEE, July 2006.

[16] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. "Dynamic 3d scene analysis from a moving vehicle". In CVPR, 2007.

[17] Liu Yang, Rong Jin, Caroline Pantofaru, and Rahul Sukthankar, "Discriminative Cluster Refinement: Improving Object Category Recognition Given Limited Training Data", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June, 2007.

[18] B. Moghaddam, "Principal Manifolds and Probabilistic Subspaces for Visual Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, Issue 6, June 2002, pp. 780-788.

[19] Pech-Pacheco, José L., álvarez-Borrego, Josué, Cristóbal, Gabriel, Keil, Matthias S., "Automatic object identification irrespective of geometric changes", International Society for Optical Engineering (SPIE), Vol 42(2): 551-559 (2003).

[20] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. "Objects in context". In ICCV, 2007.

[21] A. Saxena, M. Sun, and A. Y. Ng. "Learning 3-d scene structure from a single still image". In ICCV 3dRR-07, 2007.

[22] H. Schneiderman, "Learning a restricted bayesian network for object detection," in Proc. CVPR, 2004.

[23] Torres M., "Is there any hope for face recognition?", Proc. of the 5th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2004, 21-23 April 2004, Lisboa, Portugal.

[24] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "Toward Objective Evaluation of Image Segmentation Algorithms", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, No. 6, June, 2007, pp. 929-944.

[25] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in Proc. ICCV, 2003.

[26] P. Viola and M. J. Jones, "Robust real-time face detection," IJCV, vol. 57, no. 2, 2004.

[27] Xiaoxing Li; Mori, G.; Hao Zhang, Expression-Invariant Face Recognition with Expression Classification, The 3rd Canadian Conference on Computer and Robot Vision, Volume , Issue , 07-09 June 2006 Page(s): 77 ? 77, Digital Object Identifier:10.1FER03/CRV.2006.34. 2006.

[28] W. Zhang, B. Yu, D. Samaras, and G. Zelinsky. "Object class recognition using multiple layer boosting with heterogeneous features". In IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[29] W.Y. Zhao, R. Chellappa, "Image-based Face Recognition: Issues and Methods, Image Recognition and Classification", Ed. B. Javidi, M. Dekker, 2002, pp. 375-402.

[30] J Fernández, N. Miranda, R. Guerrero, F. Piccoli, "Image identification thru Multilevels parallelism", In 8th International Information and Telecommunication Technologies Symposium (I2TS´2009), IEEE. Pp. 314-328, ISBN: 978-85-89264-11-2, Florianopolis, DF, Brazil, December 2009.

[31] Fernández J., Miranda N., Guerrero R., Piccoli F., "Multi-level Paralelism in Image Identification", In XVIII Congreso sobre Métodos Numéricos y sus Aplicaciones, (ENIEF 2009) Vol 28. Pp: 227-240. ISSN 1666-6070. Univ. Del Centro de la Prov. de Buenos Aires. Tandil ? Bs. As. Noviembre 2009.

[32] Fernández J., Miranda N., Guerrero R., Piccoli F. "A Distributed Computing for an Image Processing Function Set". In XVII Congreso sobre Métodos Numéricos y sus Aplicaciones, (ENIEF 2008), Pp: 2895-2906. ISSN 1666-6070. Univ. Nac. de San Luis. San Luis. Noviembre 10, 2008.

[33] Fernández J., Miranda N., Guerrero R., Piccoli F. "Driving to a Fast IMS Feature Vector Computing". In 14to Congreso Argentino de Ciencias de la Computación (CACIC 2008), Pp 1993-2004, ISBN 978-987-24611-0-2, Univ. Nac. de Chilecito, La Rioja, Argentina. Octubre 1, 2008.

[34] Adachi M. and Shibata T., “Image representation algorithm featuring human perception of similarity for hardware recognition systems”, In Proc. of the Int. Conf. On Artificial Intelligence (IC-AI?2001), volume 1, pages 229?234. CSREA Press, Las Vegas, Nevada, USA, 2001. ISSN 1-892512-78-5.

[35] Cantoni V., Cantoni V., Lombardi L., Porta M., and Sicard N., “Vanishing point detection: Repre sentation analysis and new approaches”. In Proceedings of the 11 th International Conference on Image Analysis & Processing, pages 26?28. 2001.

[36] Dong W., Zhou N., and Paul J.C., “Perspective-aware texture analysis and synthesis”, In Vis. Comput., 24(7):515?523, 2008. ISSN 0178-2789.

[37] Ilonen J., Kamarainen J., Paalanen P., Hamouz M., Kittler J., and Kalviainen H. “Image feature localization by multiple hypothesis testing of gabor features”. 17(3):311?325, 2008.

[38] Pan Q., Min-Gui Z., De-Long Z., Yong-Mei C., and Hong-Cai Z, “Face recognition based on singular-value feature vectors”. In Optical engineering, volume 42, pages 2368?2374. Society of Photo-Optical Instrumentation Engineers, Bellingham, 2003. ISSN 0091-3286.

[39] Seo K.S., Lee J.H., and Choi H.M. “An efficient detection of vanishing points using inverted coordinates image space”. Pattern Recogn. Lett., 27(2):102?108, 2006. ISSN 0167-8655. doi:http://dx.doi.org/10.1016/j.patrec.2005.07.011.

[40] Serre T., Wolf L., and Poggio T. “Object recognition with features inspired by visual cortex”. In IEEE CSC on CVPR. 2005.

[41] Lee K. and Street W., “Automatic feature mining for personalized digital image retrieval”, In ACM, editor, Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD 2001), pages 38?43. ACM, San Francisco, USA, 2001.

[42] Tan K., Ooi B., and Yee C., “An evaluation of color-spatial retrieval techniques for large image databases”. Multimedia Tools and Applications, 14(1):55?78, 2001.

[43] Marquis-Bolduc M., Deschênes F., and Pan W., “Combining apparent motion and perspective as visual cues for content-based camera motion indexing”. Pattern Recogn., 41(2):445?457, 2008. ISSN 0031-3203. doi:http://dx.doi.org/10.1016/j.patcog.2007.06.021.

[44] Rosin P.L., “Training cellular automata for imageprocessing”. In IEEE Transactions on Image Processing, volume 15, pages 2076?2087. 2006.

[45] Wang C., Yu-bin Y., Wu-jun L., and Shifu C., “Image texture representation and retrieval based on power spectral histograms”. In Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004). 2004.