Semantic Cohesion for Image Annotation and Retrieval

Hugo Jair Escalante, Luis Enrique Sucar, and Manuel Montes-y-Gómez

Coordinación de Ciencias Computacionales,
Instituto Nacional de Astrofísica, Óptica y Electrónica,
Tonantzintla, Puebla, 72840,
Mexico

{hugojair, esucar, mmontesg}@inaoep.mx

Abstract. We present methods for image annotation and retrieval based on semantic cohesion among terms. On the one hand, we propose a region labeling technique that assigns an image the label that maximizes an estimate of semantic cohesion among candidate labels associated to regions in segmented images. On the other hand, we propose document representation techniques based on semantic cohesion among multimodal terms that compose images. We report experimental results that show the effectiveness of the proposed techniques. Additionally, we describe an extension of a benchmark collection for evaluation of the proposed techniques.

Keywords. Automatic image annotation, region labeling, multimedia image retrieval, ground truth data creation.

Cohesión semántica para la anotación y recuperación de imágenes

Resumen. Presentamos métodos para la anotación y recuperación de imágenes que se basan en la cohesión semántica entre términos. Por un lado, proponemos una técnica para etiquetar regiones que asigna a cada imagen el conjunto de etiquetas que maximiza un estimado de la cohesión semántica entre estas. Por otro lado, proponemos métodos para representar imágenes anotadas que se basan en la cohesión semántica entre términos multimodales que aparecen en las imágenes. Reportamos resultados experimentales que muestran la efectividad de las técnicas propuestas. Adicionalmente, describimos la extensión que realizamos a una colección estándar para la evaluación de los métodos propuestos.

Palabras clave. Anotación automática de imágenes, etiquetado de regiones, recuperación multimodal de imágenes, creación de datos para evaluación.

1 Introduction

Image retrieval has been an active research area for more than two decades now [1, 2]. Despite a substantial advance achieved in this field, most of the reported work focuses on methods that consider a single modality (i.e., either image or text), thus limiting the effectiveness and applicability of two groups of methods: one based on texts and the other based on images. On the one hand, text-based methods are unable to retrieve images that are visually similar to the query image; on the other hand, image-based techniques cannot retrieve relevant images to queries that involve non-visual information (e.g., places, events, or dates).

Due to the above limitations, in the last few years there has been an increasing interest of the scientific community in the development of retrieval techniques that incorporate both visual and textual information [2, 8]. Nevertheless, current multimodal techniques still deal with both sources of information separately, eventually applying a standard information fusion technique [2]; therefore, such methods do not exploit the association among multiple modalities to obtain better representations of images. Furthermore, in many databases, images are not accompanied with any textual information, which further complicates the application of multimodal retrieval

Extended abstract of PhD thesis. Graduated: Hugo Jair Escalante. Advisors: Luis Enrique Sucar and Manuel Montes-y-Gómez. Graduation date: 25/03/2010.
Resumen extendido de tesis doctoral. Graduado: Hugo Jair Escalante. Directores: Luis Enrique Sucar y Manuel Montes y Gómez. Fecha de graduación: 25/03/2010.

methods. In the latter collections, automatic image annotation (AIA) methods are used for assigning text to images because manual labeling of images is a time-consuming and labor extensive task [1].

Both tasks, image annotation and image retrieval, are closely interrelated and hence can be studied jointly. Accordingly, in this paper we faced such problems with the goal of improving the performance of current techniques and overcoming some of their limitations. More specifically, we focused on the region-level AIA task with the goal of giving support to multimodal image retrieval methods that attempt to exploit the redundancy and complementariness of information as provided by labels and text.

We propose methods for the annotation and retrieval of images that are based on *semantic cohesion* modeling; where *semantic cohesion* is defined as the degree of affinity of terms in a document according to their meaning or their use in the context given by other terms that occur in the same document. Intuitively, the greater the degree of semantic cohesion among terms, the higher is the probability that such terms are used together in similar contexts. The rest of this paper summarizes our research and outlines the main findings of our work, for further details we refer the reader to the thesis [3] and the representative publications derived from it [4, 7]. Before presenting our methods, we describe the extension we made to a benchmark collection for allowing the evaluation of region-level AIA methods and of image retrieval methods that consider AIA labels.

2 The SAIAPR TC12 Collection

Due to the lack of a suitable database to evaluate the methods we propose, part of our work included the development of a benchmark image collection. Specifically, we proposed an extension of the IAPR TC12 collection[2], a benchmark data set for the evaluation of image retrieval methods [8]. The extension consisted in manual segmentation and annotation of each image in the IAPR TC12 collection according to predefined rules and by using a hierarchical organization of the vocabulary we defined. The proposed hierarchy is composed of six branches: "*Animal*", "*Humans*", "*Food*", "*Man-made*", "*Landscape*" and "*Other*". Summarizing, a total of 20,000 images have been manually segmented and the resultant 99,535 regions were manually labeled by using a vocabulary of 255 labels; the data derived from our extension is publicly available at the official ImageCLEF website[3]. Our extension has increased the number of applications of the collection and its scope in terms of the tasks that can be evaluated with it [4]; furthermore, the extension has been extremely helpful for the evaluation of the methods we developed, and has attracted the interest from the scientific community. A detailed description of our extension to the IAPR TC12 collection can be found in [4].

3 Semantic Cohesion for Image Annotation

For AIA, we propose an energy-based model that attempts to maximize the semantic cohesion among labels that have been assigned to adjacent regions in segmented images [5, 6]. The model seeks to refine the initial labeling as provided by a multiclass classifier trained with purely visual information. The classifier (which can be built by using diverse learning algorithms) provides candidate labels for every region in an image; next, using information about the association between labels (estimated through co-occurrence statistics), the energy-based models select the best combination of labels that should be assigned to the image. Figure 1 illustrates the proposed energy-based model.to regions; shaded nodes denote the confidence of classifiers in the candidate labels. We consider dependencies between spatially connected regions.

[2] http://imageclef.org/photodata

[3] http://imageclef.org/SIAPRdata

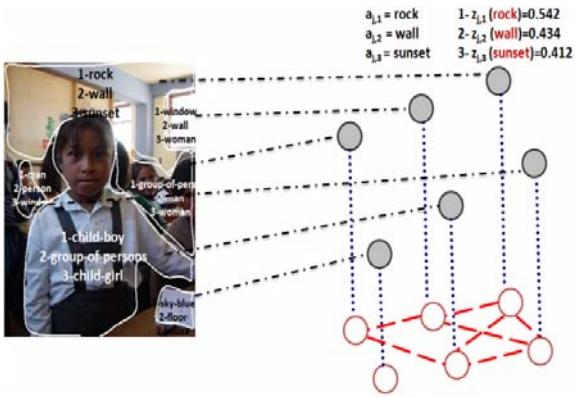

Fig. 1. Energy-based model for AIA. Left: segmented image with candidate labels per region; a relevance weight is associated with each candidate label via a multiclass classifier. Right: graph associated with the image according to the energy-based model. Unshaded nodes represent assignations of labels to regions; shaded nodes denote the confidence of classifiers in the candidate labels. We consider dependencies between spatially connected regions

Table 1. Comparison of labeling accuracy obtained by the initial classifier (column 3), the proposed energy-based model (column 4), and the best reported result for the corresponding image collection (column 2), see [6] for further details

Data set	Ref.	OVA-RF	EBM
C-AN	45.64%	57.90%	**58.97%**
C-AG	50.50%	56.56%	**57.23%**
C-BN	39.50%	46.64%	**48.74%**
C-BG	43.00%	46.08%	**46.87%**
C-CN	42,50%	54.13%	**55.59%**
C-CG	47.50%	52.28%	**52.71%**
SCEF-I	**60.94%**	59.99%	60.35%
SCEF-II	78.73%	81.55%	**82.92%**
MSRC-1	**93.94%**	86.60%	88.82%
MSRC-2	70.50%	70.86%	**76.03%**
VOGEL	71.70%	70.78%	**72.54%**

We report experimental results obtained with the proposed method over several benchmark image collections of heterogeneous characteristics. Table 1 shows the annotation accuracy obtained by our method and the best reported results for each data set (see caption).

Our experimental results show the usefulness of the proposed method: the multiclass classification approach to AIA proved to be very effective (see *OVA-RF*); the energy-based model improved the initial labeling for all of the considered collections (see *EBM,* the difference was statistically significant according the Wilcoxon signed-rank test with 95% of confidence); the proposed method outperformed the best results reported in related works where the authors have used the same collections we did (see *Ref.*).

The main benefits of the proposed method are generality of the approach, easiness of its implementation, its effectiveness and high efficiency. Our work on image annotation with the energy-based model is described in detail in [6].

4 Semantic Cohesion for Image Retrieval

For image retrieval, we propose methods based on semantic cohesion among labels and text to represent multimodal documents. Specifically, we propose two forms of representing images based on distributional term representations (DTRs) that have been widely used in computational linguistics [9]. Under the considered DTRs, each multimodal term (i.e., a label term or a word term) is represented by a vector of statistics of occurrence over the documents in a collection or co-occurrences over other terms in a vocabulary.

In this way, the representation of a term will be influenced by the documents in which it occurs (capturing dependencies between terms and documents) for the document-occurrence representation (DOR) or by the terms it mostly co-occurs with (capturing dependencies between terms) for the term co-occurrence representation (TCOR). Once each term in the multimodal vocabulary is represented through DTRs, documents are represented by the weighted sum of the DTRs of terms that appear in the document. Intuitively, each document is represented by the context associated with the terms that occur in the document; where the context is given by other documents in the collection or other terms in the multimodal vocabulary.

Table 2. Retrieval results by using the proposed representations (rows 3-4), standard techniques (rows 5-7) and unimodal methods (rows 8-9). We report the mean-average precision (MAP), precision (P20) and recall (R20) for 20 documents and for the number of retrieved documents that are relevant (RR)

Topics	ImageCLEF2007				ImageCLEF2008			
Method	**MAP**	**P20**	**R20**	**RR**	**MAP**	**P20**	**R20**	**RR**
Multimodal-DOR	0.2141	0.2425	0.3295	2507	0.1958	0.25	0.3333	1679
Multimodal-TCOR	0.1935	0.2392	0.2951	2379	0.1763	0.2564	0.2893	1632
Late-fusion	0.1348	0.1858	0.1879	1703	0.1126	0.1936	0.1759	1139
Early-fusion	0.189	0.2508	0.2996	2226	0.1565	0.2372	0.2695	1416
IM-relevance feedback	0.1659	0.2142	0.262	1952	0.1326	0.1987	0.2205	1302
Labels-only	0.0587	0.1417	0.1066	1201	0.053	0.141	0.1133	727
Text-only	0.1241	0.1767	0.1694	1424	0.1033	0.1795	0.1534	1014

We report experimental results of the developed techniques on the SAIAPR TC12 collection. Table 2 compares the retrieval performance of our proposals: multimodal DOR and multimodal TCOR, with unimodal (text-only and labels-only) and standard multimodal techniques (late fusion, early fusion and inter-media relevance feedback), over two sets of topics (ImageCLEF2007 and ImageCLEF2008).

Experimental results obtained with the standard methods show that the combination of labels and text can be helpful for improving the performance of unimodal strategies significantly. However, the proposed representations achieve better performance than the standard techniques. The difference in performance is statistically significant for multimodal DOR according to the pair-wise t-student test with 95% of confidence. Furthermore, the content of multimodal images is better represented with our techniques, when compared to unimodal or standard multimodal strategies.

In summary, we provide evidence showing that the combination of labels and text can be very helpful for image retrieval, and we prove that the proposed representations provide an effective solution to the multimodal image retrieval task. Our developments on multimodal image retrieval with distributional term representations are explained in detail in [7].

5 Conclusions and Future Work

We provided experimental evidence which shows that the idea of *semantic cohesion* can be effectively exploited for modeling multimodal information. The proposed methods for image annotation and image retrieval based on such idea obtained superior performance than those reported in related papers; furthermore, our techniques offer additional benefits. Thus, we can conclude that the semantic cohesion modeling, and more specifically, that a modeling based on co-occurrence statistics offers important benefits in terms of effectiveness, efficiency and representation power.

Also, the experimental evidence we provided shows that the combination of labels and text can improve the retrieval performance of unimodal methods, even when standard information fusion techniques are used. Despite the latter is highly intuitive, there are no similar works that attempt to combine text and labels, to the best of our knowledge. As future work, we would like to explore application of the energy-based model to similar problems from structured prediction. We also want to include global image information into the energy function. Concerning image retrieval, we would like to explore the use of multimodal DTRs for combining raw-visual features with textual information and the use of DTRs for generating visual vocabularies for object recognition and image categorization.

Acknowledgements

We are grateful to Nando de Freitas, Aurelio López, Eduardo Morales, J. Francisco Trinidad, and Luis Villaseñor for their valuable comments which helped us to improve our work. Also, we thank the TIA research group at INAOE. This work was supported by CONACyT under Project Grant No. 61335 and Scholarship No. 205834.

References

1. **Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M., & Jordan, M.I. (2003).** Matching words and pictures. *The Journal of Machine Learning Research,* 3, 1107–1135.
2. **Datta, R., Joshi, D., Li, J., & Wang, J.Z. (2008).** Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys,* 40(2), Article 5.
3. **Escalante, H.J. (2010).** *Cohesión semántica para la anotación y recuperación de imágenes*. Tesis de Doctorado, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México.
4. **Escalante, H.J., Hernández, C.A., González, J.A., López-López, A., Montes, M., Morales, E.F., Sucar, L.E., Villaseñor, L., & Grubinger, M. (2010).** The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding*, 114(4), 419–428.
5. **Escalante, H.J., Montes, M., & Sucar, L.E. (2007).** Word co-occurrence and Markov random fields for improving automatic image annotation. *Proceedings of the 18th British Machine Vision Conference*, Warwick, UK, 2, 600–609.
6. **Escalante, H.J., Montes-y-Gómez, M., & Sucar, L.E. (2011).** An energy-based model for region-labeling. *Computer Vision and Image Understanding,* 115(6), 787–803.
7. **Escalante, H.J., Montes, M., & Sucar, E. (2011).** Multimodal indexing based on Semantic cohesion for image retrieval. *Information Retrieval,* DOI: 10.1007/s10791-011-9170-z.
8. **Grubinger, M. (2007)** *Analysis and Evaluation of Visual Information Systems Performance*. PhD thesis, Victoria University, Melbourne, Australia.
9. **Lavelli, A., Sebastiani, F., & Zanoli, R. (2004).** Distributional term representations: an experimental comparison. *Thirteenth ACM International Conference of Information and Knowledge Management (CIKM '04),* Washington D.C., U.S.A, 615–624.

Hugo Jair Escalante obtained his Ph.D. in Computer Science from the Computational Sciences Department of the National Institute of Astrophysics, Optics and Electronics, where he has been a full-time researcher since January 2012. From August 2010 to December 2011, Dr. Escalante had been an assistant professor at the Graduate Program on Systems Engineering of the UANL. He is a Candidate Researcher of the Mexican National System of Researchers (SNI) and author of more than 30 international papers in journals, books, conference proceedings and workshops. Dr. Escalante obtained the "2010 Best Ph.D. Thesis on AI Award" granted by the Mexican Society of Artificial Intelligence for the thesis reported in this paper. His main research interests are machine learning and its applications in natural language processing and high level computer vision.

Luis Enrique Sucar has a Ph.D. in Computing from Imperial College, London, UK, 1992; a M.Sc. in electrical engineering from Stanford University, California, USA, 1982; and a B.Sc. in electronics and communications engineering from ITESM, Monterrey, Mexico, 1980. He is currently the Director of Research at the National Institute for Astrophysics, Optics, and Electronics (INAOE), Puebla, Mexico. He has been an invited professor at the University of British Columbia, Canada; Imperial College, London; and INRIA, France. He has more than 150 publications in journals, books and conference proceedings, and has directed 16 Ph.D. theses. Dr. Sucar is a Senior Member of IEEE, a member of the Mexican National System of Researchers (SNI), and the Mexican Academy of Sciences. He served as the President of the Mexican AI Society and was a member of the Advisory Board of IJCAI. His main research interests are graphical models and probabilistic reasoning, and their applications in computer vision, robotics, and biomedicine.

Manuel Montes-y-Gomez obtained his Ph.D. in Computer Science from the Computing Research Center of the National Polytechnic Institute of Mexico. Currently, he is a full-time researcher at the Computational Sciences Department of the National Institute Astrophysics, Optics and Electronics, Puebla, Mexico. His research is on automatic text processing. He is author of more than 150 international papers in the fields of information retrieval, question answering, information extraction, and text mining. Dr. Montes has been a visiting professor at the Polytechnic University of Valencia (Spain), the University of Geneva (Italy), and the University of Alabama at Birmingham (USA). In addition, Dr. Montes is a member of the Mexican National System of Researchers (SNI), the Mexican AI Society, the Mexican Association for Natural Language Processing, and the International Web Intelligence Consortium.

Article received on 02/06/2010; accepted on 12/01/2012.