Semantic Expansion of Auto-Generated Scene Descriptions to Solve Robotic Tasks

Marco A. Gutiérrez
RoboLab, University of Extremadura, Cáceres, Spain
Email: marcog@unex.es

Rafael E. Banchs
HLT Dept., I2R, A*STAR, Singapore
Email: rembanchs@i2r.a-star.edu.sg

***Abstract*—When a robot is facing object description based tasks, such as "*bring me something to drink water*", it has to semantically relate the concepts on the task with the objects it is able to find. This work expands the semantic scope of words in automatically generated scene descriptions and a given task in order to find a proper match for the robot task. An encoder-decoder pipeline that unifies joint image-text embedding models with multimodal neural language models is used to generate scene descriptions. Then the semantics of those descriptions are extended through word vectors. We improve our previous work by expanding the dimension of the object description by adding the option of negating characteristics of the searched object. Finally we show that we are able to find objects that are in the scene and where not directly referred in the task or labeled by the robot using different words.**

***Index Terms*—object search, semantics, deep neural networks, robotics vision**

I. Introduction

Object searching tasks for robots in unknown environments remains a challenge for the robotics research community. Being able to make a robot successfully perform full pick and place tasks is one of the main objectives of many roboticists. One of the key parts of it is the ability of the robot to properly match the information regarding the object to find with the objects found in in the scenes surrounding it. Properly performing the match between the information and the objects found can help robots perform these tasks in a more natural way.

Numerous ways exist for object detection and recognition on scene images. Deep neural network based image labeling and, lately, image description generation are on the state of the art for scene images understanding. Big advances have been done in this field as recent works like [1]-[3] prove. These works make use of deep neural networks in order to produce somewhat accurate descriptions of the image scene. However these captions are usually short and, even though in the case they provide accurate descriptions, they do not fully express all the information that is contained in the scene. On top of that, same things can sometimes be expressed with different words or people may differ on how they call something that shows up on a certain scene, specially since words can have multiple degrees of similarity [4]. For these reasons we can not solely rely on the words on these descriptions as atomic units that give us all the information we need to match a certain object query. In order to extend the information contained on these generated sentences semantic relations between words can be exploited.

There is a wide range of research works in the field of the analysis of words semantics relations. Works such as [5]-[8] are just an example of some of the most common works in this research area. Some approaches exploit manually created ontologies or taxonomies like WordNet [9] or Freebase [10]. As stated in [11], these works are ontologies that are manually created and maintained in order to provide a means for establishing semantic relations between words and because of that sometimes its further development can be very costly. In consequence, only a determined domains have a suitable ontology, limiting the applicability of similarity measures based on one of them. On the other hand word vectors are a good and fast way to capture semantic relations between words [12], specially when trained over big corpus containing a large amount of words. This makes them easily trainable in the needed semantic scope so the info better matches the application. In our design we decided to handle semantic relations between words by measuring the distance among the word vector representation of those words.

The system presented here weights the semantic relations between a description based search task issued to the robot and scene automatically generated descriptions in order to improve the possibilities to find an object matching the user needs. Our system is even able to handle descriptions that include negations, such as "*find an animal that does not bark*". First, the task is analyzed using the Natural Language Toolkit (NLTK) [13] in order to select the key words on it and differentiate negative requirements from positive ones. The neural network encoder-decoder pipeline described in [14] is used in order to generate captions that describe scenes

Manuscript received August 17, 2015; revised December 28, 2015.

doi: 10.18178/ijmerr.5.2.109-114

from images. Then pre-trained word vectors helps finding semantic similarities between words using the skip-gram model described in [12]. A similarity weight is calculated using the cosine distance in the vector space between the selected words from the descriptive task and the ones in the image generated descriptions. Results are sorted by their calculated similarity weight, the best ones would be the ones with the highest similarity value. This process allows the expansion of the semantic domain of the words on the image generated captions. The system is able to find things that are not explicitly noted in the description sentences. Even in the case of querying for something that is not on the image dataset, the output will still be semantically more relevant than a random ordering of the images.

II. System Design

The goal of our system is to look for scenes that contain the information that is described in the task the robot is given. It accepts descriptive tasks in the form of "*get me something to drink water*" or even with negative parts like "*bring me an instrument with no strings*". When the robot receives a task the system semantically analyzes the sentence detecting the main positive words and the negative ones from the description. It obtains the word vector representations of these words and calculates an average of these vectors by weighing appropriately the negative vectors and the positive ones. Also, as shown in Fig. 1, the system contains a multimodal encoder-decoder pipeline that generates the descriptions for the scenes. The system generates five description candidates for each scene. An average of the cosine distances between the vector representations of adjectives and nouns from these descriptions and the vector representing the description is calculated for each of the scenes. Finally the system provides a rank of best matching scenes according to their weight value. The selected scenes contain the objects that are most semantically related to the words on the description contained in the task.

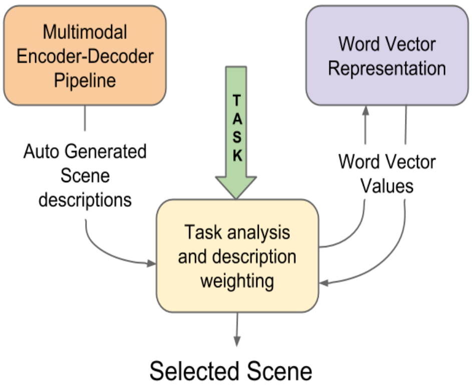

Multimodal Encoder-Decoder Pipeline
Word Vector Representation
TASK
Auto Generated Scene descriptions
Word Vector Values
Task analysis and description weighting
Selected Scene

Figure 1. System's architecture

A. Multimodal Encoder-Decoder Pipeline

The system contains an encoder-decoder pipeline that automatically generates descriptions for the scenes. The encoder (see Fig. 2) is learned with a joint image-sentence embedding where sentences are encoded using long short-term memory (LSTM) recurrent neural networks [15]. Image features from the top layer of a deep convolutional network trained from the ImageNet classification task [16] are projected into the embedding space for the LSTM hidden states. A pairwise ranking loss is minimized in order to learn to rank images and their descriptions.

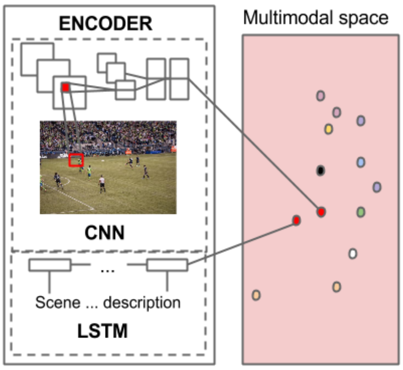


Figure 2. The deep convolutional network (CNN) and long short-term memory recurrent network (LSTM) encoder. It is in charge of learning a joint image-sentence embedding.

As Fig. 3 shows, for decoding, the Structure-Content Neural Language Model (SC-NLM) described in [14] is used, which takes into account the content in the sentences. It is a multiplicative neural language model where the attribute vector is an additive function of the embeddings. These embeddings are conditioned on the embedding vector for the description computed with the LSTM. Allowing the system to make use of large amounts or monolingual text to improve the quality of the language model. Since the embedding vectors share a joint space with the image embeddings, the SC-NLM can also be conditioned on image embeddings after the model has been trained.

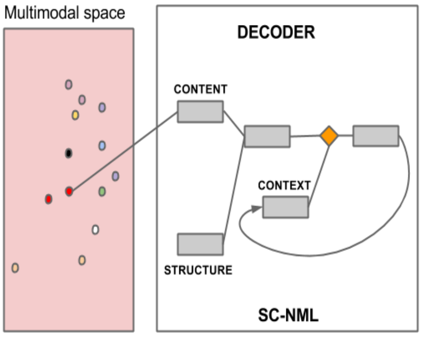


Figure 3. Structure-content neural language model decoder in charge of generating words for the scene description one at a time.

The final output of this pipeline generates are the top five most reliable descriptions for a scene. This is run for each one of the scenes that are in the dataset.

B. Word Semantics Relationships

In order to measure the semantic relationships between words we selected a neural network based tool, since they perform better than Latent Semantic Analysis (LSA) [17] for preserving linear irregularities among words and in terms of computational cost when trained over large datasets [4], [18]. We use an improved version of the Skip-gram model [12] to find word representations that predict the surrounding words in a corpus. Our system was trained using the negative sampling [19] technique instead of the hierarchical softmax, so it tries to differentiate data from noise by means of logistic regression. Semantic relations on the words of the training data are encoded in a word vector space. The semantic relation between words is measured by the cosine distance between their vector representations. These semantic relationships are used to extend the scene descriptions word meanings when the robot search task is being performed.

C. Word Matching System

This is the module in charge of making the semantic matching and evaluation between the task and the scene descriptions. As the robot receives the task this module analyzes it using NLTK. The positive main words are separated from the negative ones. For this, a basic syntactic analysis of the task is performed along with some basic regular expression matching techniques. Prepositions, pronouns and determinants are ignored as they we do not consider them relevant to the task. Firstly the average vector representing the task description is obtained by summing up the vector representations of the positive words and subtracting the ones from the negative words. This leaves us with a vector representation that will represent the semantic of the description of the task. Then the cosine distance of this vector with the words on the scene description is obtained and an average distance for all of them is finally calculated. This would be the weight of the scene for a specific description in a task, and in consequence a representation of how semantically similar they are.

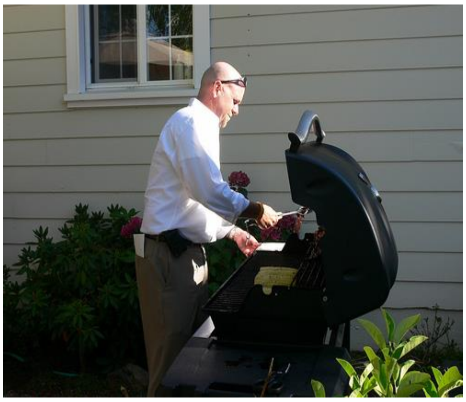
Finally when all scene weights are computed for the given task they are ranked by their weight value. The output of the system will show the scenes with the highest weight, as those are the ones that are supposed to have a higher semantic similarity with the description on the task. Since they are semantically similar, they should be describing similar things.

III. Experiments

In order to perform the experiments the LSTM encoder and SC-NLM decoder of the pipeline described in Section II-A have been trained on sentences from both Flickr30K [20] and Microsoft COCO [21]. We have selected randomly a subset of 1000 images from Flickr30K to use them as a dataset for the scene description generation. These are the ones being considered for a possible selected scene and final match with the task description. The vector space word representation have been trained on a Google News data subset containing about 100 billion words. And the final word vector model contains 300-dimensional vectors for 3 million words and phrases.

Since a manual interpretation of the contents of an image will always be open to criticism of subjectivity [22], there is a high difficulty of quantitatively evaluate the retrieval effectiveness of our approach. However we will perform a manual evaluation on the output to provide an approximated quantitative evaluation, providing besides the visual output as a support of our experiment results.

For evaluation purposes we have tested the system against a direct word to word matching approach. On this direct matching approach we will select the positive and negative words in the same way we do in our system. Then for the positive words we will add a value of *1* to the overall scene weight if the word in the task description appears on any the scene generated descriptions, otherwise *0* will be added. For the negative words we will subtract *1* if the there is a match between the negative word and any word from the scene generated descriptions. The same way as with the positive ones nothing will be subtracted if the word is not found in the descriptions. This measures basically the number of words shared among the description in the task and those from the scene minus the negative words they share. Finally these computed weights would represent the similarity between the scene and the description on the task, the higher the value the more similar they are supposed to be.



a man in a black apron is working on a grill.
a man wearing a black shirt is cooking.
a man with cooking on the ground with his machine.
a young man in a black shirt is cooking on a large grill.
a man is in his left hand.

Figure 4. Top result of the task "*find me a barbecue pit*". Note that the words "*barbecue pit*" do not appear in the generated captions but probably due to the high semantic relation with the word "*grill*" (0.583 cosine distance) we are able to find it.

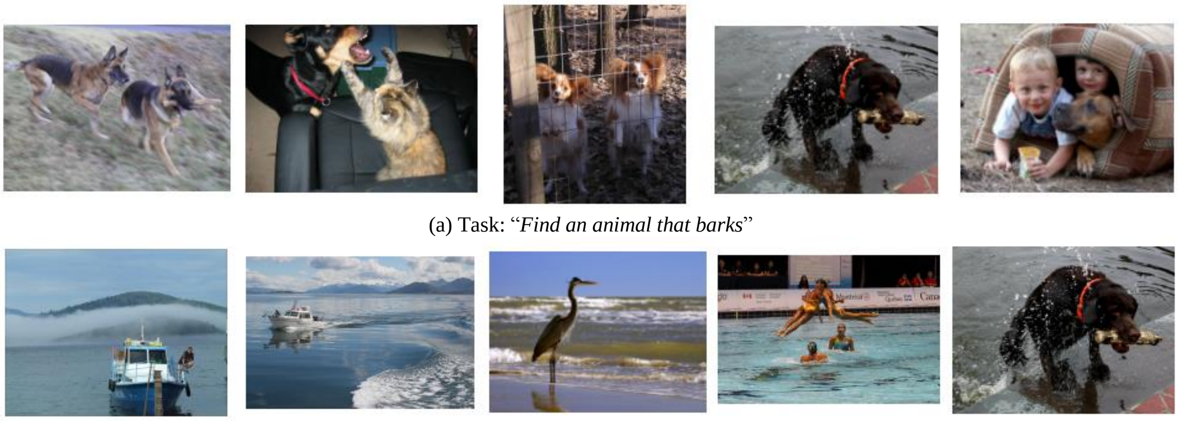
For the quantitative results we have evaluated the top five results for six search tasks on both approaches manually giving a score of 1 for correct matches, 0.5 to partially correct matches and 0 to totally wrong matches. We obtained a total score of 25 for our system against a score of 10.5 for the direct match approach, showing the benefits of our semantic expansion approach. We show here a visual excerpt of the obtained results and add some comments on them for a more specific evaluation.[1]

Fig. 4 shows the top result, a basic example of the robot process of the task "find me a barbecue pit". Not

[1] Due to space limits we can only show some results here, for a wider overview of all the ones used in the evaluation please refer to: http://magutierrez.com/description-based-tasks.

any of the scene generated descriptions show the word “barbecue” among their results. The reason why the system is able to select that picture is due to the high semantic relation between the words “barbecue” and “grill” (cosine distance of 0.583 of their correspondent word vectors representations). In the same way Fig. 5 shows results on different descriptive tasks that had no words for direct matching so the result on the alternative is a total random selection of scenes. However on those pictures our system is still able to provide us with a scene that can match the description from the task. These examples show that even though the description generation system is not reflecting the same words as in the task description we can still match them due to their existing semantic relation represented in the word vector space.

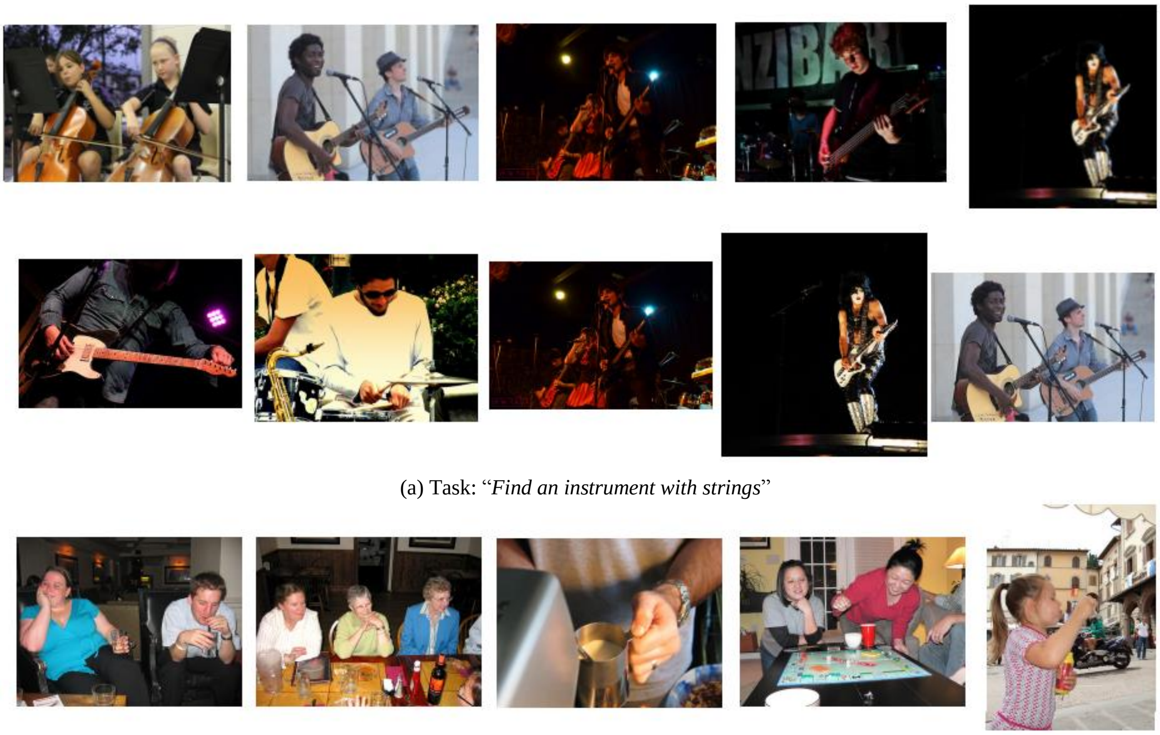
In Fig. 5 we show the results from the two options tested, our system and the direct match approach. Results are displayed ordered by similarity score from left to right, keeping the results from our solution on the top row, while the lower one corresponds to the direct match approach output. On the first task (Fig. 5a) our system is able to relate the words “*strings*” and “*instrument*” with names of instruments with strings. However on the direct approach only the word instrument is matched from the scene descriptions so even the fact that some guitars are shown as a result is pure coincidence as it could have been any other instrument. On the second task (Fig. 5b) the system gets more confused. However is still better as it can relate the word “*drink*” with some liquids or drinking situations. Even though, in this case it is not a great result it is still better than a totally unrelated scene such as some of the results on the direct match approach. Some of the errors here are also due to some errors in the automatically generated scene description.



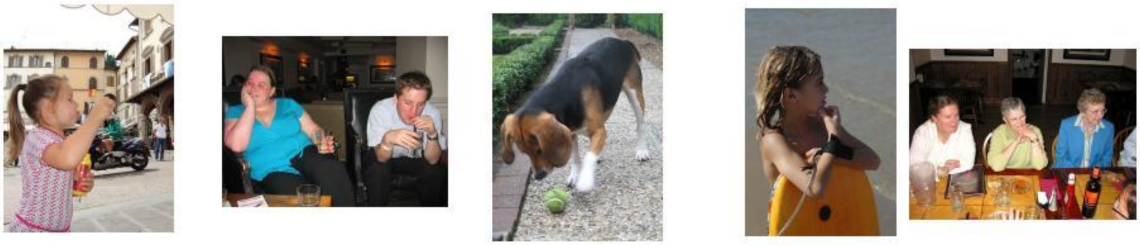
(a) Task: “*Find an animal that barks*”

(b) Task: “*Find something that floats in water*”

Figure 5. Tasks that had none words in common with any of the scene generated descriptions



(a) Task: “*Find an instrument with strings*”



(b) Task: "*Bring me a drink with alcohol*"

Figure 6. First row from left to right of each tasks are our system results compared to the direct match approach on the second row

Fig. 7 uses the novel introduced negation part on the description from the task. This example task is "*find a sport with no ball*". The first results are good as the system relates some sport with the word ball and its able to discriminate them. It gets some confusion though and there is clearly room for improvement. However we found out that we can take the weighted distance value as a reference on how much we can trust the result since when bad results are obtained this weighted distance value is usually very low. Please refer to the online results in order to take a better look at the insights of the word matches.

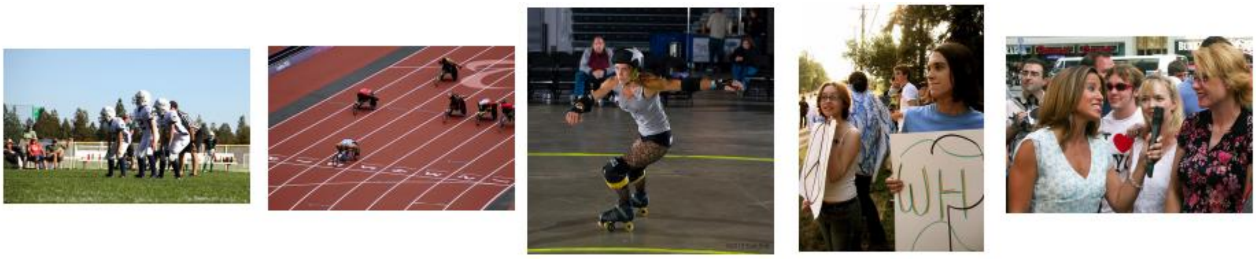


Figure 7. Task*:* "*find a sport with no ball*"

IV. Conclusions and Future Work

Our system processes tasks issued to a robot to search and find objects by its description and look for its matches from different scenes. We use word representations in vector spaces to expand the semantic scope of the descriptions and improve the matching between them. It has been proved that our system is able to properly obtain scene relations to a certain descriptive task using the semantic relations between the descriptions and the robot search task. The system can even provide meaningful results when queried with words that don't even directly appear on the scene descriptions. On the other hand some results might not be accurate enough sometimes due to not very accurate semantic relations and other times due to errors on the scene descriptions. Therefore there is room for improvement on both sides. We could take into account the value of the cosine distance and discard the results when values are too low, as we observed that low values always correspond to very bad matches. Also new deep learning techniques can be applied for the scene description generation in order to improve this part of the system. Dynamically selecting the most important parts of the description on the task can provide an improvement as we can give them a different weighs on the semantic matching algorithm. Other word semantic relation techniques can be tested in order to look for a better semantic matching between the task and the scene description.

Acknowledgment

This work has been conducted as part of an A*STAR Research Attachment Programme (ARAP) at the Human Language Technology Department of Institute for Infocomm Research, Singapore.

References

[1] X. He, R. Srivastava, J. Gao, and L. Deng, "Joint learning of distributed representations for images and texts," arXiv preprint arXiv: 1504.03083, 2015.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," arXiv preprint arXiv: 1411.4555, 2014.

[3] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," arXiv preprint arXiv: 1502.03044, 2015.

[4] T. Mikolov, W. T. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *HLT-NAACL*, June 2013, pp. 746-751.

[5] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *JAsIs*, vol. 41, no. 6, pp. 391-4071990.

[6] M. Sahlgren, "The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces," *Institutionen for Lingvistik*, 2006.

[7] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," arXiv preprint arXiv: 1309.4168, 2013.

[8] N. J. Van Eck, L. Waltman, and J. van den Berg, "A novel algorithm for visualizing concept associations," in *Proc. Sixteenth International Workshop on Database and Expert Systems Applications*, August 2005, pp. 405-409.

[9] G. A. Miller, “WordNet: A lexical database for English,” *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995.

[10] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, “Freebase: A collaboratively created graph database for structuring human knowledge,” in *Proc. ACM SIGMOD International Conference on Management of Data*, June 2008, pp. 1247-1250.

[11] ACMs. Christoph, L. O. F. I., Measuring Semantic Similarity and Relatedness with Distributional and Knowledge-Based Approaches, 2016.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv: 1301.3781, 2013.

[13] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O'Reilly Media, Inc. 2009.

[14] R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” arXiv preprint arXiv: 1411.2539, 2014.

[15] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735-178, 1997.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems*, 2012, pp. 1097-1105.

[17] S. T. Dumais, “Latent semantic analysis,” *Annual Review of Information Science and Technology*, vol. 38, no. 1, pp. 188-230, 2004.

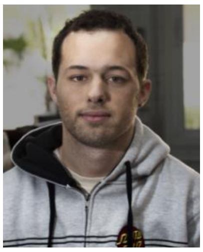
[18] A. Zhila, W. T. Yih, C. Meek, G. Zweig, and T. Mikolov, “Combining heterogeneous models for measuring relational similarity,” in *HLT-NAACL*, 2013, pp. 1000-1009.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Advances in Neural Information Processing Systems*, 2013, pp. 3111-3119.

[20] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” arXiv preprint arXiv:1505.04870, 2015.

[21] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, *et al*., “Microsoft COCO: Common objects in context,” in *Computer Vision–ECCV*, Springer International Publishing, 2014, pp. 740-755).

[22] H. Besser, “Visual access to visual images: The UC berkeley image database project,” *Library Trends*, vol. 38, no. 4, pp. 787-798, 1990.



Marco A. Gutiérrez is a PhD student in cognitive vision for robotics systems at the Robotics and Artificial Vision Laboratory (RoboLab) from the University of Extremadura, Spain since 2011. He is currently holding an A*STAR Research Attachment Programme (ARAP) scholarship in the Human Language Technology Department at I2R, A*STAR, Singapore. He obtained the highest evaluation possible (above expectations on all fields) on his internship at KUKA Laboratories GmbH, Augsburg, Germany in 2012. Recently his team (Ursus) was awarded with Best Team for Functionality Benchmark on Object Perception and Speech Understanding at the Rocking Robot Challenge 2014 in Tolouse, France. He has contributed to several open-source robotics and computer vision related projects such as RoboComp and the Point Cloud Library even as organization administrator and mentor (respectively) for several editions of the Google Summer of Code programme (2013, 2014 and 2015). His recent areas of research include cognitive vision, deep neural networks, multimodal systems and word semantics. He is organizer of the \textit{Workshop on Multimodal Semantics for Robotics Systems} and Advisory Committee for the \textit{The Path to Success: Failures in rEal Robots} Workshop that will take place as part of next IROS 2015 conference in Hamburg, Germany.

Rafel E. Banchs (M’14) is currently a Research Scientist at the Institute for Infocomm Research in Singapore. He received his Ph.D. in Electrical Engineering from the University of Texas at Austin in 1998. He was awarded a Ramon y Cajal fellowship from the Spanish Ministry of Education and Science from 2004 to 2009. His recent areas of research include Machine Translation, Information Retrieval, Cross-language Information Retrieval and Dialogue Systems. More specifically, he has been working on the application of vector space models along with linear and non-linear projection techniques to improve the quality of statistical machine translation and cross-language information retrieval systems. He has served as co-organizer of the 2nd TC-STAR Work-shop on Speech to Speech Translation 2006; the First International Workshop on Content Analysis in the Web2.0 (CAW2) at WWWW 2009, the ESIRMT-HyTra Joint Workshop at EACL’12, the CREDISLAS workshop at LREC’12, the Special Session “Rediscovering 50 Years of Discoveries” at ACL’12, HyTra-2 and HyTra-3 workshops at ACL’13 and EACL’14, and NII-Shonan Meeting Seminar 059 on “The Future of Human-Robot Spoken Dialogue: from Information Services to Virtual Assistants”. He has also served as area chair for IJCNLP’11, general co-chair of AIRS’13 and PC chair of IALP’14