Unsupervised Grounding of Textual Descriptions of Object Features and Actions in Video

Muhannad Alomari, Eris Chinellato, Yiannis Gatsoulis, David C. Hogg and **Anthony G. Cohn**
School of Computing, University of Leeds, UK
{scmara, E.Chinellato, Y.Gatsoulis, D.C.Hogg, A.G.Cohn}@leeds.ac.uk

Abstract

We propose a novel method for learning visual concepts and their correspondence to the words of a natural language. The concepts and correspondences are jointly inferred from video clips depicting simple actions involving multiple objects, together with corresponding natural language commands that would elicit these actions. Individual objects are first detected, together with quantitative measurements of their colour, shape, location and motion. Visual concepts emerge from the co-occurrence of regions within a measurement space and words of the language. The method is evaluated on a set of videos generated automatically using computer graphics from a database of initial and goal configurations of objects. Each video is annotated with multiple commands in natural language obtained from human annotators using crowd sourcing.

Introduction

Learning linguistic and visual concepts from videos and textual descriptions without having a pre-defined set of representations is a challenging yet important task. For example, humans are born without the knowledge of how many representations for directions there are in the world, or how they are described in natural language. In some situations, it is better to use the 4 directions representation (*front, right, left, back*), in others, one can use the 8 directions (*front, front right, right, etc.*). Humans are capable of learning these different representations of directions, and at the same time learn the words that describe them in natural language, without having these concepts pre-programmed into their brains. Such learning ability makes us more capable of operating in different situations, hence the importance of this task.

We exemplify our approach by showing how unsupervised learning of concepts in the following feature spaces is possible: colours, shapes, locations, and actions. For example, in the sentence "*pick up the blue block*" our aim is to learn that the phrase *pick up* is a concept in the actions feature space, the word *blue* is a concept in the colour feature space and the word *block* is a concept in the shape feature space. A key challenge of this task arises from the fact that the system does not know how many concepts there are to learn in each feature space. To avoid this dilemma, researchers have used constraints to simplify the setting sufficiently to enable learning to take place. For example, they only presented the system with a single concept to learn at a time (e.g. a single object in the scene), such that it always knows which concept is to be learned (Roy *et al.* 1999, Steels *et al.* 2002, Kumar *et al.* 2014, Parde *et al.* 2015). In other cases, certain hard-coded knowledge was provided initially (e.g. the colours, or directions), which the system used as fundamental basis to expand its knowledge (Siskind 1996, Dominey *et al.* 2005, Sridhar *et al.* 2010, Dubba *et al.* 2014, Yu *et al.* 2015).

In this paper, we present a novel system that uses a more relaxed set of assumptions, and does not use any hard-coded knowledge in any feature space. Yet we show that our system is still capable of learning about language and vision, from linguistic and visual inputs, such as video sequences with textual descriptions. The main goal is to learn natural language words and their representation in visual domains (e.g. the word *blue* is represented by a subset of the colour feature space). We will refer to the words that have visual representations as ***concrete linguistic concepts*** (e.g. the word *blue* has a representation in the colour space, therefore, *blue* is a concrete linguistic concept). We will refer to these visual representations as ***visual concepts*** (e.g. the blue colour in the colour feature space is a visual concept). Finally, we will use the term ***groundings*** to refer to the connections between the different *linguistic concepts* and *visual concepts*. The word 'concrete' in the *concrete linguistic concepts* is used to distinguish it from abstract concepts (e.g. *love, hate, real numbers*). In this paper we will focus on learning concrete concepts only, so we will omit 'concrete' in the sequel since no confusion will arise.

Connecting Language and Vision Framework

The architecture of the learning framework can be summarised in the following steps (i) the system receives linguistic and visual inputs, a video and a sentence describing it, (ii) the inputs are used to generate candidates for both *visual concepts* and *linguistic concepts*, (iii) these candidates are used to build all possible hypotheses that might ground language and vision, (iv) the system tests the hypotheses, and uses the accepted hypotheses to learn about language and vision. Steps (i,ii) are discussed in §Visual-Linguistic

Representation of the World, and steps (iii,iv) are discussed in §Connecting Language and Vision.

Assumptions

We assume no hard-coded knowledge is given in any feature space, but we make a number of assumptions that help the system in attaining its ambitious goal to connect language and vision. In order to focus on the learning and grounding issues rather than on basic vision processing, we assume that our system is capable of distinguishing and tracking objects in the world, and is capable of computing the basic perceptual properties: colour, shape, and location. We also assume (at least for the location feature), that the camera is static (so location values refer to the same position across frames). Also, since it will be helpful to segment each video into a number of intervals, we make the assumption that whether the values in an object feature space are changing or not provide suitable segmentation points.

Visual-Linguistic Representation of the World

The system receives as inputs (i) a video sequence (with objects already tracked), and (ii) a sentence describing the video, with upper case letters changed to lower case and punctuation characters removed (as these would not be explicitly present in spoken language). Both inputs are represented in a way that allows for efficient and incremental learning as will be discussed in the following sections.

Linguistic Input Representation

The linguistic inputs are represented as n-grams. An n-gram is a sequence of n consecutive words. These n-grams are extracted from the input sentence, and are used as candidates for *linguistic concepts* as shown in Fig. 1; the term 'candidates' is used to indicate that these n-grams have not yet been connected to a *visual concept*.

Visual Input Representation

The system receives a video sequence as a visual input, from which it extracts visual information about the different (i) objects (colours, shapes, and locations), and (ii) actions (intervals). A mixture of Gaussian models is used to abstract and represent the information from the different objects' features, and a number of intervals are used to represent the actions (as shown in the *visual concepts* in Fig. 1). We will discuss these two representations further below.

Objects' Representation: Three pre-defined visual features are computed for the objects in the input video: colours, shapes, and locations (These features are just examples; many further features could be added easily). It is assumed that the system has no pre-given knowledge in any of these feature spaces. The features we use are as follows:

1. $colour : Object \rightarrow [0, 360) \times [0, 1] \times [0, 1]$; colour($o$) gives a hue, saturation, and value (HSV) colour value per pixel.
2. $shape : Object \rightarrow R^h$; shape($o$) gives a histogram of oriented gradients (HOG) values, with a size of 7200 (30 by 30 pixels with 8 directions per pixel).
3. $loc : Object \rightarrow R \times R \times R$; loc($o$) gives an x,y,z coordinate location with respect to the system's base frame.

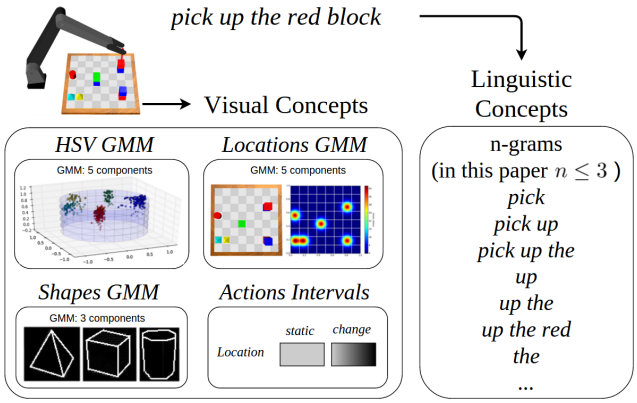

Figure 1: The video sequence and description are used to extract all possible visual and linguistic concept candidates.

Actions' Representations: In this paper, we say that an action has happened if there is a change in the state of an object (e.g. the action *pick up* is defined by a change in the location feature of an object). The changes in a video are represented using intervals. We divide the video sequence into intervals based on whether an object feature is changing or not; the changes are measured across all the features for an object. For example, if an object was static and then started moving, we segment the video at this point; the first segment for the interval during which the object is static, and the second during which it is moving. The system uses these intervals as a way to represent the different actions that happened in a video. This representation can be applicable to a wide variety of basic actions (verbs), such as (*pick up, put down, move, place, shift, take, remove*) which are manifested by changes in the location feature, (*paint*) by a change in the colour feature, or (*cut*) by a change in the shape feature. In this paper, we aim to learn (i) the representations of the different actions in vision, and (ii) the words that describe these different actions in natural language.

Connecting Language and Vision

To learn how to connect the candidate linguistic and visual concepts, we use an approach similar to that presented in Hebbian theory, which can be summarized by the phrase: "*Cells that fire together, wire together*" (Schatz 1992). This idea can be translated into our system as: candidate visual and linguistic concepts that appear together, are connected together. As an example, the 1-gram *red* and the colour *red* will appear consistently together more often than the 1-gram *red* and the colour *green*. Based on this idea, the system uses the inputs (linguistic descriptions and videos) to find the candidate concepts with the strongest associations.

The system will create one to many mapping between each candidate linguistic concept and candidate visual concepts (e.g. the 1-gram *red* is associated with a high probability to a subset of the colour space associated with the

colour red, and other visual concepts). In order to find out which of these mappings are correct, the system tests the validity of each one of them. The correct mappings will then be used to learn about language and vision. The validation and learning procedures are done in an incremental way by processing each video and description individually. These procedures can be described by the following three steps: (i) Compute the strength of the association links between the candidate visual and linguistic concepts, (ii) test the validity of the visual-linguistic associations using the language-vision matching test described bellow, and finally (iii) use the concepts that pass the test to update the system's knowledge about language and vision.

1) Associating Candidate Concepts

To determine which candidate concepts should be connected together, we follow the frequentist approach (Everitt and Skrondal 2002). We keep track of the frequency at which each candidate visual and linguistic concept appears individually in all videos, and the frequency with which the two appear together. The system uses these frequencies to compute the conditional probabilities that associate each candidate linguistic concept with a candidate visual concept.

In our incremental learning process, the system is introduced to new candidate concepts over time (e.g. new n-grams, colours, actions, etc). When this happens, new concepts that are seen for the first time should be created, and the ones that have been seen before should have their frequencies updated. In order to update or create new candidate visual concepts (Gaussian component), we use an Incremental Gaussian Mixture Model (IGMM) approach (Song and Wang 2005) to merge or create new candidate concepts.

In order to find which candidate concepts have the highest association between them, we use their frequencies to compute the conditional probabilities between them. The conditional probability between each pair of candidates represents the strength of associating these pairs together. The computation of this conditional probability is shown in Eq. 1, where l is a candidate linguistic concept, v is a candidate visual concept, F_l is the frequency at which l appeared in all the videos processed so far, F_{vl} is the frequency of seeing both l and v together in the same video in all videos so far. This probability function is computed between every pair of candidate linguistic and visual concepts.

$$P(v|l) = \frac{F_{vl}}{F_l} \tag{1}$$

2) Language-Vision Matching Test (LVMT)

For each new video, once the frequencies of the candidate concepts have been updated, the system tests which of the candidates are correct. At this stage, the system has a set of the strongest associations between candidate concepts. In the absence of a supervising teacher, the system needs to validate the correctness of these associations in an unsupervised way using a Language Vision Matching Test (LVMT) which we have developed for this purpose. This is done by comparing the input video with multiple synthesized virtual videos, where each virtual video reflects a different association. For example, if the system does not know what the 1-gram *red* means, but it found that it has 3 potential strong associations with the colours red, blue and the location bottom left corner. The system generates 3 virtual videos that reflect these associations as shown in Fig. 2 and checks which of these virtual videos match the input video. Since substituting the 1-gram *red* with the colour red leads to a match with the input video, the system accepts the grounding (1-gram *red* $\leftrightarrow$ red colour Gaussian component) as the correct grounding.

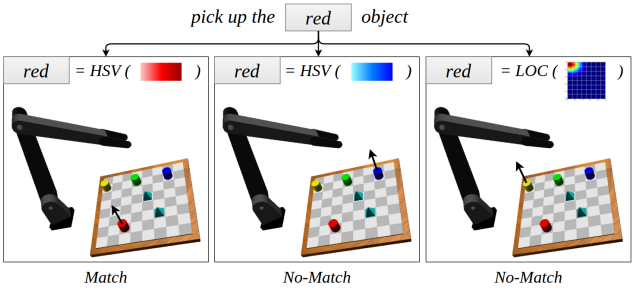

Figure 2: Generating different virtual videos from different groundings of the 1-gram *red* for the LVMT.

Experimental Validation

To validate our system, we used the *Train Robots* dataset which was designed to develop systems capable of understanding verbal spatial commands described in a natural way (Dukes 2013). Non-expert users were asked to annotate appropriate commands to 1000 pairs of different scenes. Each scene pair is represented by an initial and desired goal configuration; we automatically animated these to produce videos. 7752 commands were collected using Amazon Mechanical Turk describing the 1000 scenes. We also translated all the commands from English to Arabic, particular care was taken on not to alter any command or change any mistakes in any of them. An example of the dataset if shown in Fig. 3. The original dataset along with the videos and translated commands can be found at http://doi.org/10.5518/32.

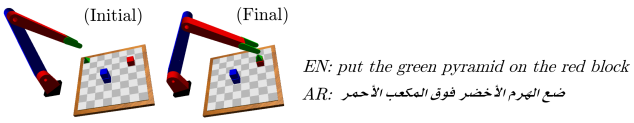

Figure 3: An Example from the *Train Robots* dataset, the Arabic sentence is translated from the English one.

Evaluation and Results

We evaluated the performance of our system based on its ability to acquire correct *visual-linguistic groundings*. The *Train Robots* dataset contains 20 different visual concepts expressible in our chosen feature spaces (e.g. the colour blue, the shape cube, etc) which our system managed to learn all of them correctly. It also has 71 English and 91 Arabic linguistic concepts (which map to visual features in our

chosen feature spaces), from which the system managed to learn 62 (87.3%) and 80 (87.9%) concepts respectively. Table 1 shows these results in more detail and some examples of the learned concepts are shown in Fig. 4.

Grounding Visual and Linguistic Concepts				
	English		*Arabic*	
Visual Features	*Linguistic*	*Visual*	*Linguistic*	*Visual*
Colours	15/16	8/8	30/31	8/8
Shapes	18/18	4/4	22/24	4/4
Locations	17/17	5/5	16/16	5/5
Actions	12/20	3/3	12/20	3/3

Table 1: The results of grounding visual and linguistic concepts in both English and Arabic. The numbers in each column (A/B) mean A correctly acquired concepts out of B available concepts.

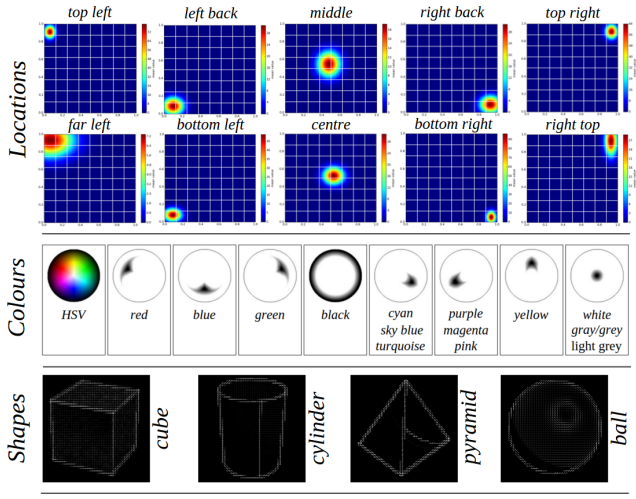

Figure 4: A sample of the learned English *Linguist Concepts* and their perceptual representations (*Visual Concepts*). The different images show the visual concepts learned, and the words next to each image show the linguistic concept associated with that visual concept.

Conclusion and Future Work

We have demonstrated for the first time that a system can simultaneously learn about object features and actions by connecting language and vision. The segmentation of videos based on feature space changes corresponding to actions is also a key contribution of the paper, acting as an intermediary representation between the continuous perceptual space, and the purely symbolic linguistic structures. We plan to extend our system to learn: (1) relations between objects such as (distance, direction); (2) grammar rules that govern the sentence structure; (3) collective words such as (tower, pile); (4) comparative and superlative relations such as (further, furthest); and (5) higher arity relations such as (between).

Acknowledgments

We thank colleagues in the School of Computing Robotics lab and in the STRANDS project consortium (http://strands-project.eu) for their valuable comments. We also acknowledge the financial support provided by EU FP7 project 600623 (STRANDS).

References

Dominey, P. F., and Boucher, J.-D. 2005. Developmental Stages of Perception and Language Acquisition in a Perceptually Grounded Robot. *Cognitive Systems Research* 6(3):243–259.

Dubba, K. S.; De Oliveira, M. R.; Lim, G. H.; Kasaei, H.; Lopes, L. S.; Tomé, A.; and Cohn, A. G. 2014. Grounding Language in Perception for Scene Conceptualization in Autonomous Robots. In *Qualitative Representations for Robots: Papers from the AAAI Spring Symposium. Technical report*, 26–33.

Dukes, K. 2013. Train Robots: A Dataset for Natural Language Human-Robot Spatial Interaction through Verbal Commands. In *International Conference on Social Robotics (ICSR). Embodied Communication of Goals and Intentions Workshop*.

Everitt, B. S., and Skrondal, A. 2002. The Cambridge dictionary of statistics. *University of Cambride Press: Cambridge*.

Kumar, S.; Dhiman, V.; and Corso, J. J. 2014. Learning Compositional Sparse Models of Bimodal Percepts. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Parde, N.; Hair, A.; Papakostas, M.; Tsiakas, K.; Dagioglou, M.; Karkaletsis, V.; and Nielsen, R. D. 2015. Grounding the Meaning of Words through Vision and Interactive Gameplay. *Proc IJCAI 2015*.

Roy, D.; Schiele, B.; and Pentland, A. 1999. Learning Audio-Visual Associations using Mutual Information. In *Integration of Speech and Image Understanding, 1999. Proceedings*, 147–163. IEEE.

Schatz, C. J. 1992. The Developing Brain. *Scientific American* 267(3):60–67.

Siskind, J. M. 1996. A Computational Study of Cross-Situational Techniques for Learning Word-to-Meaning Mappings. *Cognition* 61(1):39–91.

Song, M., and Wang, H. 2005. Highly Efficient Incremental Estimation of Gaussian Mixture Models for Online Data Stream Clustering. In *Defense and Security*, 174–183. International Society for Optics and Photonics.

Sridhar, M.; Cohn, A. G.; and Hogg, D. C. 2010. Unsupervised Learning of Event Classes from Video. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 1631–1638. AAAI Press.

Steels, L., and Kaplan, F. 2002. Aibo's First Words: The Social Learning of Language and Meaning. *Evolution of Communication* 4(1):3–32.

Yu, H.; Siddharth, N.; Barbu, A.; and Siskind, J. M. 2015. A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video. *Journal of Artificial Intelligence Research* 601–713.