Semantic Levels of Domain-Independent Commonsense Knowledgebase for Visual Indexing and Retrieval Applications

Amjad Altadmri, Amr Ahmed, and Haytham Mohtasseb

School of Computer Science, University of Lincoln, Lincoln, UK,
{atadmri, aahmed, hmohtasseb}@ lincoln.ac.uk

Abstract. Building intelligent tools for searching, indexing and retrieval applications is needed to congregate the rapidly increasing amount of visual data. This raised the need for building and maintaining ontologies and knowledgebases to support textual semantic representation of visual contents, which is an important block in these applications. This paper proposes a commonsense knowledgebase that forms the link between the visual world and its semantic textual representation. This domain-independent knowledge is provided at different levels of semantics by a fully automated engine that analyses, fuses and integrates previous commonsense knowledgebases. This knowledgebase satisfies the levels of semantic by adding two new levels: temporal event scenarios and psycholinguistic understanding. Statistical properties and an experiment evaluation, show coherency and effectiveness of the proposed knowledgebase in providing the knowledge needed for wide-domain visual applications.

Keywords: Commonsense Knowledgebase, Multimedia Mining, Semantic Levels, Ontology Development, Multimedia Indexing and Retrieval.

1 Introduction

The huge amount of video collections on the internet increased the demand on intelligent mining and analyses tools. This motivated the work on video understanding applications, like semantic video annotation, rating, indexing and retrieval. Work in this area aims to fill the *"semantic gap"*, which is the difference between low-level visual features and human's perception. Many approaches tried to establish a textual semantic representation of the visual data in order to tackle this issue. For achieving this aim, these approaches either build a domain specific *"Ontology"*, or utilized existing commonsense knowledgebases. Commonsense is the information and facts that are expected to be commonly known by ordinary people, including real-life events and popular facts. Other approaches build knowledgebases to tackle one level of semantic such as objects, events or concepts. The high demand of these applications emphasizes the need for generic knowledge on multi semantic levels, which is our inspiration in this work.

In this paper, a commonsense knowledgebase that forms the link between the visual world and its semantic textual representation is introduced. It provides knowledge at the objects, events and scenes level, besides higher-levels

of semantics by adding temporal events, scenarios and psycholinguistic annotation. This is achieved by utilizing knowledge and strong functionalities of three of the largest knowledgebases, WordNet [1], ConceptNet [2] and Linguistic Inquiry Words Count (LIWC) [3], trying to fulfil special requests of this area. This has been achieved by, first, merging different predicates' types from these different knowledgebases. Then, eliminating non-visually-related information. Finally, fusing the resulted nodes into one unified structure. This knowledgebase will is named as Visual Semantic Levels Net (VSL Net). Quantitative analysis and an automatic video annotation experiment show the effectiveness and comprehensiveness of the proposed enhancements.

The rest of the paper is organized as follows: Section 2 discusses related work in ontology for visual applications. Section 3 explains the proposed contributions, while the experiments and evaluation are described in section 4. The paper concludes in section 5, where future work is also suggested.

2   State of the Art

This section states the key work in video annotation and visual retrieval systems, and the most related work in textual semantic knowledge engineering.

In visual applications some approaches tries to use ontology to detect visual concepts. For example, in [4], the ontology is integrated for semantic pooling of positive examples from ontologically neighboring concepts. However, other approaches directly include visual knowledge in domain-specific ontologies, in a form of low-level visual descriptors, to perform semantic annotation [5]. Although most of these methods depend on rules created by domain experts, they are subject to some inconsistency inherited from the differences of the involved humans' culture, mood, personality, as well as the specified topic. In addition, they become almost less efficient in wider domains. In wider domains, common-sense knowledgebases have recently received more attention to solve annotation issues. For example, in [6], a user, supported by WordNet, creates a visual concept for a group of images. Then ConceptNet is used to calculate the distance between the concepts. Moreover, in our previous work [7], a full automated framework for semantic video annotation in wide-domain has been presented based on using WordNet and ConceptNet separately.

A lot of taxonomies exist in visual literature regarding the semantic levels identification. One of the most used is presented in [8]. According to that taxonomy, levels of understanding are:

1. *low-level*: where visual features can be decided directly, for example searching for "red image".
2. *object and action level*: where the properties of objects and actions can be learned as one or more visual features like color, texture and motion.
3. *Event level*: where the objects/actions information is fused with the environment data to find the semantic event based on previous cognitive knowledge.
4. *Scenario (or story)*: is the cognitive output about a number of events combined to achieve a purpose. Moving from the previous level to this higher
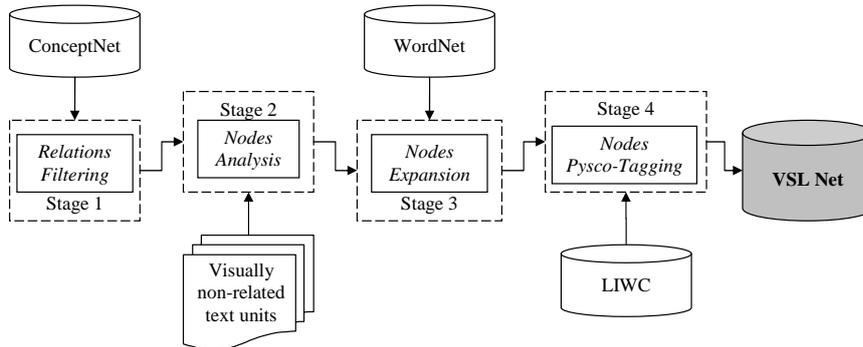
Fig. 1. Visual Semantic Levels knowledgebase (VSL Net) building framework

level is almost non-dependant on visual features directly, but it rather depends on knowledge of merging the events.

5. *Psychological category*: it contains a whole scene abstraction based on information ranging across different levels. For example, *body* and *sport* depend on object recognition whereas *sad* or *happy* may relate to gesture recognition. In contrast, *Achievement* is much more based on the context.

   The proposed knowledgebase in this paper, VSL Net, provides real-life commonsense knowledge ranging from second till fifth levels of semantics. The construction engine is fully automated extracting and merging the useful knowledge for the visual domain on these different semantic levels.

3   Proposed Knowledgebase

This section explains the contribution of this paper. It includes summarization of the two semantic levels proposed in VisualNet, the addition of two new levels of semantic, and the final structure. Figure 1 shows the stages of building:

3.1   VisualNet

This subsection summarizes our previous knowledgebase (VisualNet), which provides knowledge needed for the second and third levels of semantics. Building process was divided into three stages.

   First, relationships filtering operation was performed on ConceptNet to extract needed edges. This depended on the relation type, contents of nodes, and the parameters of the relation. Relationship-types that are useful in the objects' and events' levels in visual applications field have been selected; namely *"capableOf"*, *"usedFor"* and *"locationAt"*. Despite of these relations-types forms only 3 out of 24, their edges occupy the third of the ConceptNet. Dissenting, special

case, uncertain relations and misspelled nodes are discarded. This resulted in a skimmed version of ConceptNet that contains 150K out of 480K edges.

Second, each node was tagged as *noun* or *verb*. Then, it has been analyzed to obtain the core phrase that is matching its type. Complex-sentence concepts that had more than one meaning phrase have been divided.

Third, the resulted nodes were expanded via several steps based on WordNet Synonyms. The words in each node have been returned to the original form, and then each node was replaced by the best synonym set that suits the meaning of the relation. Finally, all resulted matched relations were merged. In the following subsections, the operations of adding the other semantic levels are explained.

3.2   From Events to Scenarios

In VisualNet, only *"capableOf"*, *"usedFor"* and *"locationAt"* relationships have been included. This is due to that they form the first level of semantic by connecting objects to events and objects/events to scenes. To achieve higher level of semantic representation, the ontology (or the knowledgebase) has to contain the relationships among events of a studied context. As this work aims to provide a tool for domain-independent visual applications, the context has to be based on real-life knowledge. Therefore, *"HasFirstSubevent"*, *"HasLastSubevent"* and *"HasSubevent"* relation-types from ConceptNet are included. The first two are almost partial sub-events of the connected concept, thus they have been fused in one relation, emph"P Subevent". While the third is mainly independent one, so it will be *"Subevent"* relations. Other relation types in ConceptNet are analyzed and found not as useful as the utilized ones in visual applications.

3.3   Emotions and Psychological Categories

To provide the knowledge needed to extract possible emotions presents in a scene, either images or videos, LIWC has been utilized to annotate current nodes. Each node in the Net will be assigned tags that contain the psycholinguistic categories (annotations) and their relevance degree. This is performed following the steps detailed in the following subsections.

Initial Annotation The phrases in each concept node are tokenized to allow for inducing the matching LIWC categories and their associated frequencies. In each node, common categories are congregated raising their frequencies. Thus, at this stage, each concept is linked to a list of psycholinguistic categories with their matching frequencies. This could produce some words that do not have matched categories. To address that, the tags of synonym sets in the current concept nodes are utilized as clarifies in the next subsection.

Synonyms Expansion In this stage, the issues with words that don't have matching categorization and others with ambiguity holding too many categories will be solved. This is performed through making benefits from the synonym
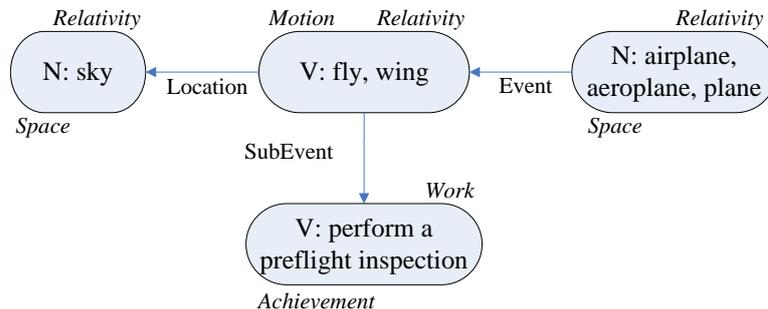
Fig. 2. VSL knowledgebase structure. Annotation tags are attached to the nodes.

nodes that are attached to each concept. For each synonyms set, the annotations of synonyms are obtained. Intersection is performed among these tags then these annotations are considered as annotation to the missed word with a frequency of '1'. Words with many categories will also be consolidated by getting the categories of its equal meaning words in the concept context.

Contextualization In this stage, each concept is annotated with a list of psycholinguistic categories and their matching frequencies. This step filters the psycho-annotations based on the representing words, maintaining only the annotations which suit the context. This performed based on the resulted frequencies applying a threshold on the number and top frequencies of the annotation items list. For example, the concept *"a scream of freedom"* has contrastive annotations before contextualization, *Neg.Emotion* and *Pos.Emotion*, resulted from its component words. Moreover, it got *Hearing* which is related to one of the words, but it is out of the context for this concept. This contextualization step ends up with *Affective* annotation which represents the context of concept.

Figure 2 illustrates the new structure of the constructed knowledgebase. The psycholinguistic tags are sketched attached to their matched nodes.

4   Experiments, Results and Evaluation

In this section, the proposed knowledgebase's statistical properties, in addition to an example application, subsection 4.1, are presented.

Comparing the statistics, it can be seen that in SVL the number of nodes and relationships, 118K and 138k, is reduced, compared to the same relations types in ConceptNet, 136K and 154K, respectively. This is due to deleting the unnecessary nodes/relations and merging the matched ones. However, Interdependency, which is the average number of out- and in-bound edges that are connected to each node, equation 1, has risen from 2.27 to 2.35. This factor is very important and more representative as it justifies coherency and usefulness of the proposed VSL Net in the semantic inference. This is because it reflects

the nodes connectivity degree. In other words, it reveals the number of semantic connections that could be found starting from one word. Thus, the proposed knowledgebase nodes have more interdependency and less ambiguity, in spite of it has lower number of nodes, which shows that it has managed to merge the advantages of the utilized knowledgebases.

$$I = \frac{\sum_{i=1}^{N}(R_i^{in} + R_i^{out})}{N} \tag{1}$$

Where: I is the Interdependency ratio, N is the total nodes' number, R_i^{in} and R_i^{out} are the inbound and outbound relationships' numbers for the node(i).

4.1   Semantic Video Annotation Experiment

In this experiment, the proposed knowledgebase is examined in an application for domain-independent video annotations. This experiment is trying to help filling the semantic gap between the low-level visual features and high level semantics. The input is a *query video* to be annotated (e.g. for indexing or retrieval purposes), and the output is an annotation that semantically represents its contents on different levels of semantics. The algorithm extracts a simple spatiotemporal signature of the query video to be compared against a limited dataset of pre-annotated videos' signatures. The annotations of the closely-matching videos are utilized to activate nodes in the proposed knowledgebase to extract the semantic relationships and commonality between them. A final verification function on the activated nodes will activate higher levels of semantic on high scored activated nodes which are connected via semantic relationships, eliminating isolated nodes. Those resulted active nodes will form the annotation for the query video.

The experiments are performed on a standard dataset for action recognition and video information retrieval; namely UCF Actions [9]. This dataset is composed of 1600 video clips containing various activities involving humans and animals in different indoor and outdoor events. These challenging videos contain a considerable range of variations including types of objects and events in addition to size, appearance, shape, viewpoint and motion of objects. To evaluate the results, the performance has been benchmarked against a baseline which resulted from extracting the annotations from similar pre-annotated videos, without using the proposed knowledgebase. For experimental purposes, 275 video clips have been randomly taken as a test set, covering multiple events across the dataset. The rest are used as the pre-annotated dataset. The results are evaluated against the baseline as an annotation problem, where each annotation is divided into its basic terms, to be compared to the ground truth pool. For evaluation, the performance of our method against the baseline is compared using the precision over numbers of ranked files, which is a TRECVID standard measure [10].

Figure 3 presents the resulting $P(r)$ for the framework against the baseline over various cut-off ranks. The figure shows that the proposed framework consistently outperforms the baselines. We obtain precision of 0.80 at the first annotation, dropping to 0.50 at the ninth one; while the baseline starts at 0.74 for the first annotation, dropping to 0.41 at the ninth.
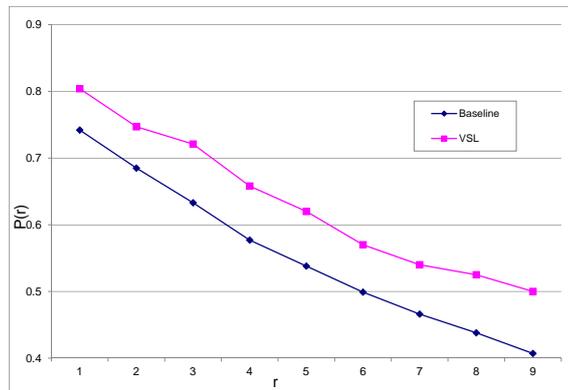
Fig. 3. Precision at a given cut-off rank for UCF Actions dataset experiment. Our proposed framework shows considerable improvement over the baseline all over the curve. Even the number of false alarms increases as more output items considered, however, items with high rank are more important.

Figure 4 illustrates a snapshot of qualitative evaluation of query videos' tested. For each example, it shows a sample of the produced annotations, where each annotation is labeled with either a *noun* or *verb* and associated with its weight. The fourth column gives a sample of the composed annotations, while the last one lists the psycholinguistic categories.

5   Conclusion

In this paper, a commonsense knowledgebase for different high-level semantics visual domain applications is introduced. This knowledgebase is automatically built by carefully and intelligently merging contents and functionalities from non-domain-specific well-known commonsense knowledgebases; namely Word-Net, ConceptNet and LIWC. The automatic engine enables this knowledgebase to be developed, updated, maintained and automatically synchronized with future enhancements on the original ones.

Statistical properties shows that the proposed knowledgebase manages to merge advantages of the Utilized knowledgebases. That is because in spite of it has lower number of nodes, its nodes have more interdependency and less ambiguity. An experiment on one candidate application, which is semantic video annotation for indexing purposes, based on the proposed knowledgebase has been demonstrated. The quantitative and qualitative evaluations of this experiment are represented to illustrate effectiveness and usefulness of this knowledgebase in visual applications. Both evaluations demonstrate coherency, strength and usefulness of the proposed knowledgebase.

The proposed knowledgebase opens new research directions towards multiple wider semantic video and image applications. In addition, some enhancements

Video	Annotations	Weights	Composed Annotation	Categories
	manoeuvre (v)	1		
	gambling (n)	1		
	toy (n)	1	volleyball play	Leisure
	play (v)	1		
	volleyball (n)	0.99	leap gambling	*Nil*
	leap (v)	0.01		
	horseback (n)	1		
	ride (v)	1	horseback ride	Relativity
	spring (v)	0.14		
	domestic dog (n)	0.02	domestic dog chase	Home
	chase (v)	0.02		

Fig. 4. Snapshot of the output of the framework represented as part of speech and composite connections as sentence cores. Underlined results are the correct ones.

on the net are under investigation. One enhancement is to add visual features nodes to make a direct connection between the input and the database, solving the first level of semantics. Other enhancements could be implemented, such as, a post- sentence formalization to produce more connected sentences.

References

1. Fellbaum, C.: WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press (1998)
2. Liu, H., Singh, P.: Conceptnet: a practical commonsense reasoning tool-kit. BT Technology Journal **22**(4) (2004) 211–226
3. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count-liwc2001. Mahway : Lawrence Erlbaum Associates (2001)
4. Zhu, S., Ngo, C.W., Jiang, Y.G.: On the pooling of positive examples with ontology for visual concept learning. In: Proceedings of the 19th ACM international conference on Multimedia, ACM (2011) 1045–1048
5. Dalakleidi, K., Dasiopoulou, S., Stoilos, G., Tzouvaras, V., Stamou, G., Kompatsiaris, Y.: Semantic representation of multimedia content. Knowledge-driven multimedia information extraction and ontology evolution **6050** (2011) 18–49
6. Shevade, B., Sundaram, H.: A visual annotation framework using common-sensical and linguistic relationships for semantic media retrieval. Adaptive Multimedia Retrieval: User, Context, and Feedback **3877** (2006) 251–265
7. Altadmri, A., Ahmed, A.: Automatic semantic video annotation in wide domain videos based on similarity and commonsense knowledgebases. In: The IEEE International Conference on Signal and Image Processing Applications. (2009) 74–79
8. Zhang, Y.: Semantic-Based Visual Information Retrieval. IRM Press (2006)
9. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos in the wild. In: Computer Vision and Pattern Recognition. (2009) 1996–2003
10. Over, P., Awad, G., Fiscus, J., Antonishek, B., Michel, M., Smeaton, A.F., Kraaij, W., Qunot, G.: Trecvid 2010: An overview of the goals, tasks, data, evaluation mechanisms, and metrics. In: TRECVid 2010. (2011) 1–34