VisualNet: Commonsense Knowledgebase for Video and Image Indexing and Retrieval Application

Amjad A.Altadmri
Computing Department
University of Lincoln, UK
atadmri@lincoln.ac.uk

Amr A.Ahmed
Computing Department
University of Lincoln, UK
aahmed@lincoln.ac.uk

***Abstract*—The rapidly increasing amount of video collections, available on the web or via broadcasting, motivated research towards building intelligent tools for searching, rating, indexing and retrieval purposes. Establishing a semantic representation of visual data, mainly in textual form, is one of the important tasks. The time needed for building and maintaining Ontologies and knowledge, especially for wide domain, and the efforts for integrating several approaches emphasize the need for unified generic commonsense knowledgebase for visual applications.**

In this paper, we propose a novel commonsense knowledgebase that forms the link between the visual world and its semantic textual representation. We refer to it as *"VisualNet"*. *VisualNet* is obtained by our fully automated engine that constructs a new unified structure concluding the knowledge from two commonsense knowledgebases, namely WordNet and ConceptNet. This knowledge is extracted by performing analysis operations on WordNet and ConceptNet contents, and then only useful knowledge in visual domain applications is considered. Moreover, this automatic engine enables this knowledgebase to be developed, updated and maintained automatically, synchronized with any future enhancement on WordNet or ConceptNet.

Statistical properties of the proposed knowledgebase, in addition to an evaluation of a sample application results, show coherency and effectiveness of the proposed knowledgebase and its automatic engine.

I. Introduction

The amount of video collections available has tremendously grown for various reasons including the availability of inexpensive hand held digital cameras, popularity of web-based video sharing websites and the huge number of broadcasting channels. As a result, the need for intelligent mining and management tools for these data became crucial. All 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. A number of approaches try to establish a semantic representation of visual data in textual form to tackle this issue. For achieving this aim, these approaches either build a domain specific *"Ontology"*, which refers to the theoretical representation model in knowledge systems [1], or utilized existing commonsense knowledgebases. The more increase of these applications emphasize the need for standardization of semantic tools used, which was our inspiration for a novel commonsense knowledgebase for visual applications.

In this paper, a novel commonsense knowledgebase that forms the link between the visual world and its semantic textual representation is proposed. We refer to it as *"VisualNet"*. This knowledgebase is built by a fully automated engine that performs analysis operations on both nodes and relationships levels on both WordNet[2] and ConceptNet[3], then a new unified structure is constructed containing only useful knowledge in the visual domain.

In addition to that, this automatic engine reduces the time and efforts needed for developing and maintaining such knowledgebase, as it can automatically update the *"VisualNet"* synchronized with future enhancement in WordNet or ConceptNet.

Quantitative analysis shows effectiveness and comprehensiveness of the proposed knowledgebase' representation and how it manages to merge the advantages of both WordNet and ConceptNet. Another experiment on video enhancing annotation shows that results based on this knowledgebase outperform results utilizing either WordNet or ConceptNet, individually.

The rest of the paper is organized as follows: In section II, related previous work is discussed. A comparison between WordNet and ConceptNet, in term of the visual data contents, is presented in section III. Our VisualNet structure and the proposed automatic building process are presented in section IV, while the experiments, results and evaluation are described in section V. The paper is finally concluded in section VI, where future work is also suggested.

II. Previous work

This section focuses mainly on the key work of video annotation and semantic retrieval systems.

Some approaches tried to use Ontology to detect visual concepts. For example, in [4], Ontology was built by learning concepts' relationships based on analyzing co-occurrences between concepts. Other direction was to use association mining techniques to indicate the existence of high-level concept from simultaneously existence of other concepts, as an attempt to enhance accuracy of semantic concepts detection [5].

Other approaches have directly included visual knowledge in multimedia domain-specific Ontology, in a form of low-level visual descriptors for concept instances, to perform semantic annotation [6].

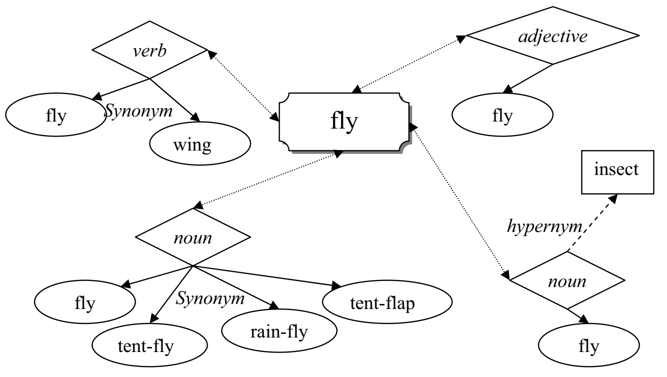


Fig. 1: An example of tree built for one tag based on WordNet.

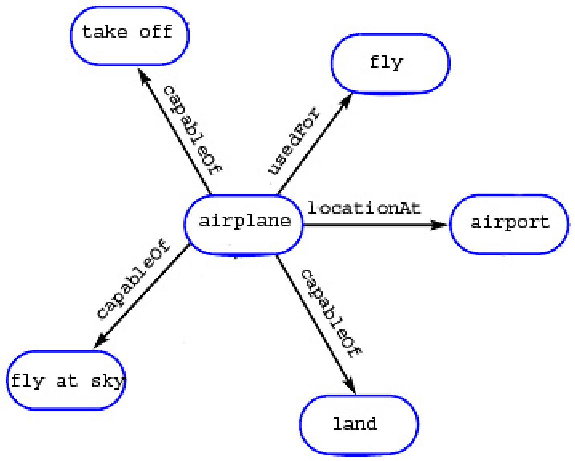


Fig. 2: A snapshot of ConceptNet relationships.

As these methods almost depend on rules that are created by domain experts, they are subject to some inconsistency inherited from variations of the involved humans' culture, mood, personality, as well as the specific topic. In addition to that, they become almost less efficient in wider domains.

Research in text mining area manages to build considerable commonsense knowledgebases. The Commonsense is the information and facts that are expected to be commonly known by ordinary people. WordNet [2], Cyc [7] and ConceptNet [3] are considered to be the widest commonsense knowledgebases currently in use.

In semantic video applications area, commonsense knowledgebases have recently received some attention to solve annotation issues, by finding related concepts. In [8] concepts' relationships are learned, in public video databases, using ConceptNet's *"get context"* functionality.

WordNet [2] has been utilized in many applications in this area, especially for semi-automated annotation approaches, to find similar meaning annotations. In [9], 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. On the other hand, some researchers in text retrieval area merge results, obtained individually, from the ConceptNet and the WordNet to achieve better query expanding [10].

In summary, as current research in visual semantic indexing and retrieval on wide domain increases, it needs establishing knowledge suits its nature for different applications. For that purpose, we try to provide a solution by presenting a novel automatic knowledgebase for visual domain. This knowledgebase utilizes strong functionalities of two of largest knowledgebases, WordNet and the ConceptNet, trying to fulfill special requests of this domain.

III. ConceptNet vs. WordNet

In this section, a brief introduction to the utilized commonsense knowledgebases is presented, and then surface and deep comparison are described.

A. WordNet

WordNet is a very rich non-domain-specific knowledgebase of lexical units. Each one of these units consists of several synonym words. This knowledgebase gained wide popularity and usage due to its ease of use and wideness of trusted laboratories entered information[2]. In addition to that it has rich abstraction taxonomies. Figure 1 shows an example of a tree resulted by selecting synonym sets for the word *"fly"* and their hypernym sets. WordNet is very effective if we search for the relationship among words that have similar meaning, generalization or specialization. For example, it identifies the relationship between *"test"* and *"exam"* as equal meaning words. On the other hand, it is less able to link the different elements of real life scenes like the relationship between *"airplane"* and *"sky"*.

B. ConceptNet

ConceptNet[3] is currently considered to be the largest commonsense knowledgebase [11], [3]. It contains about 480,000 relationships resulted from analyzing free text entered via web pages by hundreds of contributors. Its nodes consist of concepts that each one of them is a part of a sentence that expresses a meaning. ConceptNet is a very rich knowledgebase for several aspects: First, it contains a huge number of assertions and nodes. Second, it has a wide range of information. Finally, it has various types of relationships including descriptions parameters. Figure 2 presents a snapshot of ConceptNet that includes useful relationships in visual field. In contrast to what mentioned in WordNet, ConceptNet is very useful in describing real life scenes, but it is weak in identifying the exact relation between related meaning words.

IV. VisualNet

In this section, the proposed VisualNet automatic construction framework, structure and properties are described. First, the reasons for selecting ConceptNet and WordNet together to form the core of VisualNet are explained:

- Both nets are general-purpose, which serves our purpose in dealing with wide-domain videos.
- Both nets have natural language form.
- Both nets have semantic relational structure.

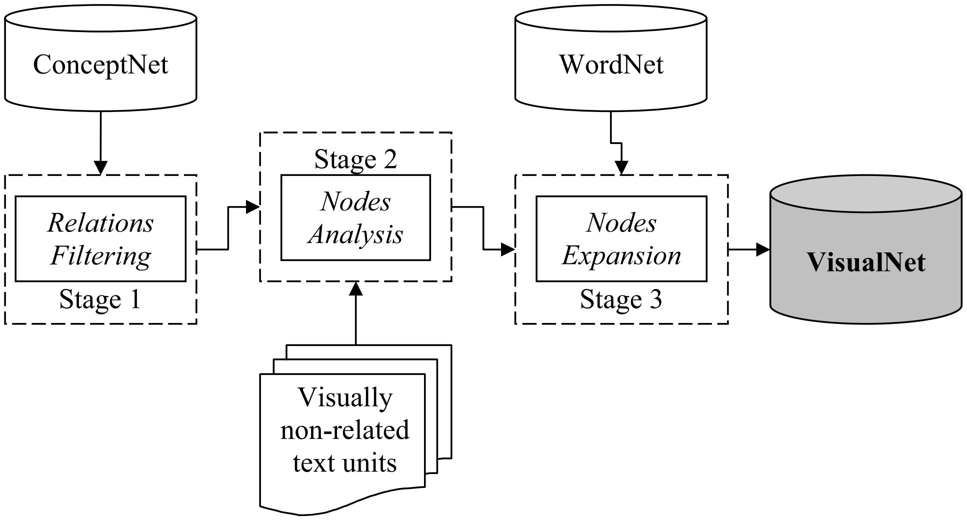


Fig. 3: VisualNet Building Framework

- While ConceptNet nodes mainly address everyday life, WordNet focuses mainly on formal taxonomies. For example: while in ConceptNet *"dog isA pet"*, in WordNet *"dog isA mammal"*.
- There is no connection between sentence parts in WordNet, but in contrast, ConceptNet has relationships that connect objects to their events, and objects to their locations.
- *"Synsets"* relationship in WordNet gives almost equal meaning words with little amount of abstraction, which is useful in many situations in our processing. But in contrast, ConceptNet's *"isA"* relationship is a mixture between abstraction and equality and sometimes just a property of a node. It's therefore neither symmetric relationship, to be considered as synonym, nor fully asymmetric, to be considered as abstraction.

Construction of new knowledgebase for visual applications is needed for:

- The main issue is the difference between natures of text related to visual field applications and traditional text mining applications. This is mainly because analysis in traditional applications is performed on full meaning sentences or even on a full integrated document. But in visual field, video clips/images are annotated usually with a semi sentence, simple title or just few separated tags.
- The nature of description of visual scenes affects the structure of the Net also. As in VisualNet the meaning is constructed via connecting the three parts of the visual world, Objects, Events and Locations.
- Existing of extra un-needed words in ConceptNet nodes, so VisualNet is more efficient version as the comparison with nodes in any application will be performed directly without the need of processing the nodes each time. This clearing operation is also leading to higher score relationships as matched core nodes are merged, which achieve more efficiency also besides the increasing of certainty of relationships.
- Not all nodes or even relationships' types are needed in the visual domain, which results a lighter version.
- Unifying the process so that no need to deal with two separated layers, ConceptNet and WordNet.
- The new representation of the net, as we change the nodes structure and the relationships. VisualNet nodes have similar structure of WordNet nodes, so the meaning of phrase is explained by its synonyms. However, VisualNet relationships have similar structure of ConceptNet, so it has wider relationships types and each relationship holds fields express its weight in the real life.
- VisualNet has better performance than using both nets together.

Figure 3 shows the stages of building the VisualNet, which is divided into three stages as follows:

A. STAGE 1: Relations Filtering

The result of this stage can be considered as a skimmed version of ConceptNet, as only useful and needed relationships are extracted. The decision is made depending on type of the relation, contents of the nodes, and the parameters of the relation. First, only affirmative relations are taken, as dissenting relations, like *"airplane doesn't drink coffee"*, do not add much as refusal relations. Secondly, special case relations that contain information on the level of names like "jack *capableOf* ride a car" and misspelled ones are discarded. In addition to that, uncertain relations, which have no agreement about their validity among contributors, are also removed.

After that, the relationships' types that have usefulness in the visual data field are selected. These relationships are: *"capableOf"*, *"usedFor"*, *"locationAt"*, *"isA"*. These relationships occupies about the third of ConceptNet in spite of there are 24 relationship type which shows their importance. Both the *"capableOf"* and the *"usedFor"* relations are merged into one relation called the *"event"*. However, for the reasons mentioned before, about the *"isA"* relation in ConceptNet, it will be replaced by the same relation from WordNet in the next stages.

As a result, this filtering reduces the number of relations from 480,000 to 150,000. And the resulted Net contains *"event"* and *"locationAt"* relations types only. In the next stage, an analysis operation is performed on nodes to extract the core.

B. STAGE 2: Nodes Analysis

Given the skimmed version relatively of ConceptNet that contains only the useful relationships visually, this stage's main purpose is to analyze all nodes to obtain their cores and delete any extra words. Then, to format the analyzed nodes in more comparable way and merge matched resulted ones.

Although in many mining applications, it is useful to use ConceptNet nodes directly to get related concepts, but in visual applications, it is more useful to obtain the core of these concepts. That is mainly to achieve the most efficient results from merging ConceptNet and WordNet, and because the aim is to form annotations by connecting three parts-of-speech that represent the visual elements of objects, locations, and events.

First of all, each node will be tagged as a *"noun"* or *"verb"* according to its rule in its ingoing and outgoing relationships. Then it is analyzed to get the core phrase matching its type. Hence, the extra words are deleted, or if the concept holds more than one meaning, the node is split as depicted in figures 4a and 4b.

This is done by tagging each node's words using Stanford parser [12], and deleting non-useful parts of sentence in visual field. The non-useful parts in this area vary from some prepositions and stop words to some adjectives and adverbs. For example "fast" is a useful adjective in visual area because it holds a meaning related to motion, but "better" is not as it does not reflect of low-level visual features in an agreed way. Then a split operation will be applied to divide some complex nodes into parts and establish new relationships.

In the example shown in figure 4b, it is clear how the Net started to turn to another kind of graph as not all nodes are equal any more. But some nodes form a root of a local tree. Hence, each dependant relation can not be considered alone in the inference process. i.e. in the same example, two relation can be concluded *"airplane can fly"* and *"airplane can fly at sky"*, but clearly *fly at sky* is non-sense.

Although the output of this stage is fully extracted from ConceptNet but it forms a new representation on the nodes level and the whole net. As nodes formatted in comparable way to WordNet, the merge with WordNet nodes became possible for ambiguity reduction. In the next stage, the current nodes are replaced with WordNet nodes semantically.

C. STAGE 3: Nodes Expansion

At the end of previous stage, a new knowledgebase is fully extracted from ConceptNet. The structure is changed a little as some nodes have their own local trees. The nodes consist of the simplest speech units and the scores for relations are more certain resulted from mixing the matched relations. However, as nodes consist of simplest part of speech without context, they still hold some ambiguity. For example, a node that contains the word *"spring"* only can not be known if it means the beautiful season of the year, a source of water or a metal device. WordNet is utilized in our work to tackle this issue.

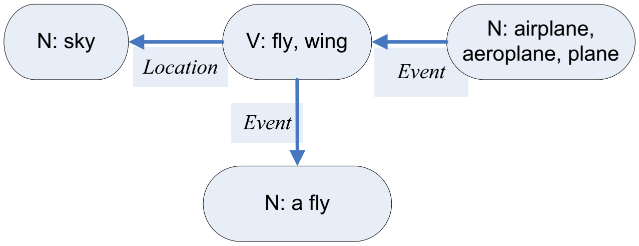


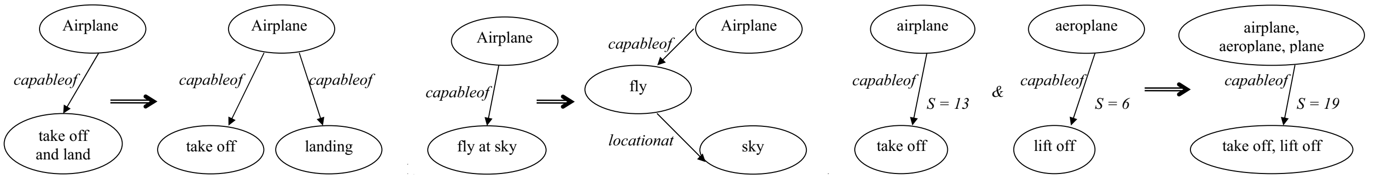
Fig. 5: A snapshot of VisualNet structure

As mentioned before, WordNet consists of units each of which contains a group of equal meaning words (*"synonyms"*). In this layer, previous resulted nodes are extended to be explained by merging with these units as follows. First, each node is tokenized to words then each of these words is returned to the knowledgebase form then the sentence is formatted again. For example, *"coffee shops", "taking off"* becomes *" coffee shop", "take off "*. Second, each resulted node is replaced by the best synonym set that suits the meaning of the relation. The selection of the best matching meaning is explained in details below. Finally, all resulted matched relations are merged so that they gain higher certainty score. An example of these operations is depicted in Figure 4c.

The selection of best matching meaning, in the second step, has three cases as follows:

1) The resulted sentence has one synonym set that matches its type, so it is selected.
2) The resulted sentence has more than one synonym set, so all of them will be hold temporarily. Then in the final merging step the intersection will be taken because it explains the meaning of the sentence
3) If the resulted sentence has no synonyms, it will be analyzed and the main part will be taken recursively and the operation will be repeated. If the analysis reaches to one word level without any synonym, the whole node will be deleted.

This stage not only increases the score of existing relations and decreases the nodes ambiguity, but it also adds new nodes/relationships. This is a result of the fact that, WordNet synonym sets contain the similar meaning parts of speech regardless the variety of names used for the same or similar objects (e.g. car, automobile), the way of describing the event or action (e.g. speed up, accelerate, gain speed), or different spelling in various versions of the language (e.g. aeroplane in British, airplane in American English).



(a) Example of nodes analysis.

(b) Example of nodes analysis.

(c) Example of replacing stage(2) nodes by WordNet nodes and the merge operation

Fig. 4: Nodes Analysis Examples.

TABLE I: Comparing Nets statistics

	ConceptNet	VisualNet
Nodes	136145	117918
Relationships	154322	138463
Interdependency	2.27	2.35

V. Experiments, results and evaluation

In table I, the statistics about the three nets are shown. Interdependency, which formulated in equation 1, represents the average number of outbound and inbound edges (relationships) connected to each node. This factor is very important as it shows how much the node is explained by related nodes and how much new information can be concluded starting from one node.

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

Where: I is the Interdependency ratio, N is the total number of nodes, R_i^{in} and R_i^{out} are the number of inward and outward relationships, respectively, for the node(i).

Comparing those statistics, it can be seen that in VisualNet the number of nodes and relationships is reduced, compared to same relations types in ConceptNet. This is due to deleting the unnecessary nodes/relations and merging the matched ones. Interdependency is more important as it justifies coherency and usefulness of the proposed VisualNet in the semantic inference as explained.

Enhance annotations application experiment:

The aim of this experiment is to automatically enhance annotations for manually tagged web-based video clips for indexing and retrieval purposes. This experiment was performed on random wide-domain video clips from the *vimeo.com* website, which is a personal contributed video website. In 627 randomly selected video clips containing 6058 tags, each annotation tag is usually consists of one word or a small incomplete phrase. To achieve this enhancement, existing initial tags are spelling checked, then each tag activates the matching node in VisualNet. As this node represents a root of a local tree, all children nodes, which represent the equal meaning synonyms and abstraction taxonomy hyponyms also, will be activated.

This tree is very rich comparing to the initial annotations' entries, and users can search for concept using abstraction. For example, although people do not tend to annotate a clip that has a *"car"* using the word *"vehicle"*, it is still highly expected that searching for *"vehicle"* should return all videos containing cars. This is achieved using our expansion tree, in contrast to the difficulty of achieving that through the initial tags.

As a result, it is clear that the expansion will highly increase the number of relevant tags. But it is clear that not all of these tags are valid, because each tag takes all possible meaning for all possible parts of speech. Hence, the next step is to validate those candidate annotation tags. This is done by activating only relationships which have two active nodes in their sides. As a result, new output annotations will be formed from the activated relationships and their nodes. The results were evaluated using Retrieval degree, Enrichment ratio and Diversity.

A. Retrieval degree

For retrieval purposes, the average number of video clips that correctly retrieved for a query phrase is calculated. Initially it was 1.70 video per query, but, using our framework, the average has been increased to 5.31 video per query, figure 6a.

B. Enrichment ratio

Tagging ratio, which is the average number of tags per video, and Enrichment ratio, which is the percentage of tagging ratio increase after enhancing annotation, formulas are explained in equations 2 and 3 respectively.

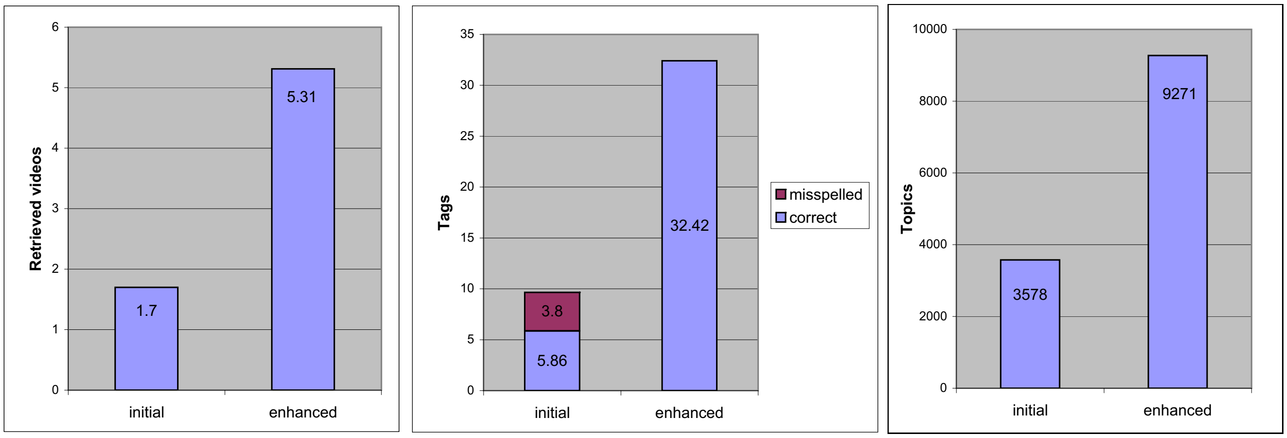
$$T = \frac{\sum_{i=1}^{N}(C_i + M_i)}{N} \tag{2}$$

Where: T is the Tagging ratio, N is the total number of videos, C_i and M_i are the number of Correct and Misspelled tags in video(i).

$$E = T2 \ \ / \ \ T1 \tag{3}$$

Where: E is the Enrichment ratio, T_1 and T_2 are the Tagging ratio before and after enhancement respectively.

As tagging ratio has risen from 9.66 tags per video clips in the dataset to 32.42 tags after annotations' enhancing, enrichment ratio has achieved a considerable degree about 336%. This is although 3.80 misspelled tags per video were removed. Figure 6b depicts the ratio of initial correct and misspelled tags to the resulted correct spelling tags.



(a) Retrieval degree.

(b) Enrichment ratio.

(c) Diversity.

Fig. 6: Results evaluation.

C. Diversity

The Diversity of annotations express the different topics exist in the dataset. It has been raised in a noticeable degree also from 3578 different tags in the first stage to 9271. This diversity achieves 260% increase in the topics indexed. Figure 6c demonstrate this increasing of all differentiated tags.

These results show that searching for a video over enhanced tags outperforms searching using the original tags. In addition to that, annotation enhanced by VisualNet outperforms both those enhanced by WordNet or ConceptNet individually, in terms of tags enrichment ability, Concept Diversity and most importantly retrieval performance.

VI. Conclusion

In this paper, we introduced a novel commonsense knowledgebase, *"VisualNet"*, for high-level semantic visual domain applications. This knowledgebase is automatically built by carefully and intelligently merging contents and functionalities from two non-domain-specific wide-known knowledgebases in text mining applications; namely WordNet and ConceptNet. The automatic engine enables this knowledgebase to be developed, updated, maintained and automatically synchronized with future enhancements of WordNet and ConceptNet.

Statistical properties of the three knowledgebases shows that the proposed knowledgebase manages to merge advantages of both WordNet and ConceptNet. That is because in spite of it has lower number of nodes, its nodes have more interdependency and less ambiguity.

An experiment on one possible application, which is video annotation enhancement for indexing purposes, based on the proposed knowledgebase has been demonstrated. The quantitative evaluation of this experiment is represented by tags enrichment ability, Concept Diversity and the most importantly retrieval performance. This evaluation illustrates effectiveness and usefulness of this knowledgebase in visual applications.

Hence, both evaluations demonstrate coherency, strength and usefulness of the proposed VisualNet knowledgebase.

The proposed knowledgebase opens search towards multiple wider semantic video and image applications. In addition to that, some enhancements on the net are under investigation. One enhancement is to automatically classify the non-related words to visual field, and another is an automatic correction of misspelled words using the net to select the most related word from the correction candidates.

References

[1] B. Chandrasekaran, J. Josephson, and V. Benjamins, “What are ontologies, and why do we need them?” *IEEE Intelligent Systems and their applications*, vol. 14, no. 1, pp. 20–26, 1999.

[2] C. Fellbaum, *WordNet: an electronic lexical database*. Cambridge, Mass: MIT Press, 1998.

[3] H. Liu and P. Singh, “Conceptnet a practical commonsense reasoning tool-kit,” *BT Technology Journal*, vol. 22, no. 4, pp. 211–226, 2004.

[4] A. G. Hauptmann, M. Y. Chen, M. Christel, W. H. Lin, and J. Yang, “A hybrid approach to improving semantic extraction of news video,” in *International Conference on Semantic Computing, 2007. ICSC 2007.*, 2007, pp. 79–86.

[5] K. H. Liu, M. F. Weng, C. Y. Tseng, Y. Y. Chuang, and M. S. Chen, “Association and temporal rule mining for post-filtering of semantic concept detection in video,” *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 240–251, 2008.

[6] A. D. Bagdanov, M. Bertini, A. D. Bimbo, G. Serra, and C. Torniai, “Semantic annotation and retrieval of video events using multimedia ontologies,” in *International Conference on Semantic Computing*, 2007, pp. 713–720.

[7] D. B. Lenat, “Cyc: A large-scale investment in knowledge infrastructure,” *Communications of the ACM*, vol. 38, no. 11, pp. 33–38, 1995.

[8] P. Yuan, B. Zhang, and J. Li, “Semantic concept learning through massive internet video mining,” in *IEEE International Conference on Data Mining Workshops*, 2008, pp. 847–853.

[9] B. Shevade and H. Sundaram, “A visual annotation framework using common-sensical and linguistic relationships for semantic media retrieval,” *LECTURE NOTES IN COMPUTER SCIENCE*, vol. 3877, p. 251, 2006.

[10] M. H. Hsu, M. F. Tsai, and H. H. Chen, “Query expansion with conceptnet and wordnet: An intrinsic comparison,” *Lecture Notes in Computer Science*, vol. 4182, pp. 1–13, 2006.

[11] M. Hsu, M. Tsai, and H. Chen, “Combining WordNet and ConceptNet for automatic query expansion: a learning approach,” in *Asia Information Retrieval Symposium*, vol. 4993, 2008, pp. 213–224.

[12] S. N. Group, “The stanford nlp log-linear part of speech tagger.” [Online]. Available: http://nlp.stanford.edu/software/tagger.shtml