# Semantic Indexing of Wearable Camera Images: Kids'Cam Concepts

Alan F. Smeaton[1], Kevin McGuinness[1], Cathal Gurrin[1], Jiang Zhou[1],
Noel E. O'Connor[1], Peng Wang[2], Brian Davis[3], Lucas Azevedo[3], Andre Freitas[4]
Louise Signal[5], Moira Smith[5], James Stanley[5], Michelle Barr[5], Tim Chambers[5],
Cliona Ní Mhurchu[6]
[1]Insight Centre for Data Analytics, Dublin City University, Glasnevin, Dublin 9, Ireland
[2]Dept. of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
[3]Insight Centre for Data Analytics, National University of Ireland, Galway, Ireland
[4]Department of Computer Science and Mathematics, Unversity of Passau,Passau,Germany
[5]Health Promotion and Policy Research Unit, University of Otago, Wellington, New Zealand
[6]National Institute for Health Innovation, University of Auckland, Auckland, New Zealand
Alan.Smeaton@DCU.ie

## ABSTRACT

In order to provide content-based search on image media, including images and video, they are typically accessed based on manual or automatically assigned concepts or tags, or sometimes based on image-image similarity depending on the use case. While great progress has been made in very recent years in automatic concept detection using machine learning, we are still left with a mis-match between the semantics of the concepts we can automatically detect, and the semantics of the words used in a user's query, for example. In this paper we report on a large collection of images from wearable cameras gathered as part of the Kids'Cam project, which have been both manually annotated from a vocabulary of 83 concepts, and automatically annotated from a vocabulary of 1,000 concepts. This collection allows us to explore issues around how language, in the form of two distinct concept vocabularies or spaces, one manually assigned and thus forming a ground-truth, is used to represent images, in our case taken using wearable cameras. It also allows us to discuss, in general terms, issues around mis-match of concepts in visual media, which derive from language mismatches. We report the data processing we have completed on this collection and some of our initial experimentation in mapping across the two language vocabularies.

## 1. INTRODUCTION

Natural language, whether written or spoken, as a means of communication is fraught with complexities because it contains ambiguity at all levels of linguistic analysis. At the lexical level, words cam be ambiguous and at the syntactic level, sentence-phrase structure can also be ambiguous. Finding the correct semantic interpretation is challenging while discourse level tasks such as pronoun resolution remain difficult. Vocabulary limitations may also apply when we want to communicate something for which there is not a word or phrase that exactly describes the message. All this is of particular concern when we use computers to automate an information seeking task, when we have to formulate an information need as a query made up of a few keywords, with no structure applied to the query.

When it comes to using computers to help us search for image data, the problems of language, especially vocabulary limitations, are exacerbated and the current default way to represent images is as a series of tags or concepts, often assigned manually. Whilst this may be useful for smaller, niche applications in restricted domains it is clearly not scalable. This scalability concern can be addressed via automatic assignment of tags, or even image captions, and this is now starting to become possible using computer vision and machine learning techniques, as we shall see later.

In this paper we examine some of the issues and limitations of representing images through natural language. We use a dataset of over 1.5M images taken by 169 subjects using wearable cameras. We describe how these have been manually annotated from a small, closed vocabulary of only 83 concepts. We then describe how we automatically processed these images by running over 1,000 individual concept detectors on each of the images. The resulting initial assignment of concepts to images was then refined using a training-free refinement method which availed of concept-assignment patterns across the dataset to "smooth" the assignments by adding consistency, guided by co-occurrence patterns of concept weights across the images, as well as external information such as date/time and GPS location of the image capture, and the ID of the subject who captured the image. This creates a sizeable dataset which has manual annotation groundtruth in one 83-dimension concept space, and automatic annotation in a 1,000-dimension concept space, and with this we can now explore some important image-language mapping questions.

The paper is organised as follows. In the next section we present a brief overview of vision media analytics, how concepts and concept weights are automatically assigned to still


*iV&L-MM'16, October 16 2016, Amsterdam, Netherlands*

DOI: http://dx.doi.org/10.1145/2983563.2983566

images or video. Section 3 then introduces the Kids'Cam project which created and supplied the dataset and in section 4 we describe the data processing we carried out on the dataset. In section 5 we then present some of our preliminary work on mapping images into two very different concept spaces

## 2. VISION MEDIA ANALYTICS

The idea of automatically assigning semantic concepts or tags to an image or video has been the subject of research for decades but more progress has been made within the last few years than in those previous decades [1]. The incorporation of deep learning into the process, coupled with the emergence of huge searchable image resources and training data means that automatic tagging of images is now offered by many websites like Aylien, IMAGGA, and others, as a commodity tagging service. For example, Figure 1 shows a picture of a London bus, visually quite distinctive with its shape and colours, and Table 1 shows the top-ranked concept tags assigned to the image by one of the cloud-based tagging services, Aylien. Large web companies like Google and Facebook are now incorporating such image tagging in some of their services like Google Photos.


Figure 1: Image of a London Bus (courtesy of Wikipedia)

While these developments are welcome, one of the problems that remain is the restrictive nature of the tagging vocabulary and how this maps to users as they try to formulate queries in order to carry out an image search. An alternative approach is taken in [7] where semantic concepts are detected in images on-the-fly, at the time of a user's query, but this is computationally expensive and so not scalable, except to smaller scale collections.

Almost all the work on concept detection is based on concept detectors not working together, but what about indexing by more than one concept at a time, something like bus-road, or road-sky-tree in Figure 1? While independent concept detection results are a valuable resource, many image retrieval scenarios require something more complex and beyond a single concept. Examples of concept pairs could be *computer-screen* combined with *telephone*, or *airplane* combined with *clouds*. Rather than combining the output of detectors at the time of a query, there is an idea for detecting the simultaneous occurrence of pairs of unrelated concepts

| Tag | Confidence |
|---|---|
| bus | 0.880 |
| trolleybus | 0.667 |
| conveyance | 0.649 |
| public transport | 0.637 |
| fire engine | 0.564 |
| truck | 0.499 |
| motor vehicle | 0.338 |
| wheeled vehicle | 0.112 |
| transportation | 0.098 |
| transport | 0.097 |
| car | 0.091 |
| travel | 0.089 |
| road | 0.088 |
| vehicle | 0.086 |
| automobile | 0.082 |

Table 1: Tags automatically assigned to the image in Figure 1, tags courtesy of http://www.aylien.ie/

in an image, where both concepts have to be observed simultaneously. While in theory this is attractive, when tried together in the 2012 and 2013 editions of TRECVid [15], most concept pairs did not work [1] though some like *Government Leader* combined with *Flag* did perform OK, but there is still a need to do something else.

In summary, what we can say is that the current trajectory of work for concept development for images needs a re-alignment or a course correction because it does not avail of all of the information sources available when detecting concepts. What makes search difficult is the multitude of ways of phrasing something in natural language with our language subtleties and this is especially so when we use natural language to search for images as we have to describe those images in an artificial language. So if all that computer vision can offer us is indexing by some fixed, possibly closed set of concepts, albeit rising to 000's of them and refined in some way, then we need some new thinking on this.

Our long-term approach is to index images by some closed concept set, and then allow a user to search based on whatever way they want to phrase a query, which is what we are accustomed to, and then map the vocabulary used in the query text to the vocabulary used in the concept indexing. This sounds feasible, and it even allows a fusion of concept-based retrieval (query text mapped to concept vocabulary) with image-based similarity (query text used to train a classifier based on a sample of positive images harvested from an external sources).

Before we achieve that though, we need to examine how a single set of images can be mapped to two different vocabularies and the interactions between those vocabularies. In this paper we look at images taken from wearable cameras and in the next section we describe the dataset we have used.

## 3. THE KIDS'CAM PROJECT

Child obesity is a significant public health concern internationally [19], including in New Zealand [9]. In 2014/15, 10.8% of New Zealand children aged 2-14 years were obese, and a further 21.7% were overweight. This situation places New Zealand children as the third most overweight or obese in the OECD [10]. There is unequivocal evidence that mar-

keting of energy-dense and nutrient-poor foods and non-alcoholic beverages is a key causal factor of child obesity [19]. As such, the World Health Organisation Commission on Ending Childhood Obesity recommends reducing children's exposure to, and the power of, such marketing [19]. To date, the available evidence on children's exposure to unhealthy food marketing has focused on single media or settings. Children's total exposure to such marketing across the multiple media and settings they encounter daily does not appear to have been quantified. Methods to collect data on children's exposure to unhealthy food marketing have largely relied on observation by researchers, or recall of parents. Wearable cameras can now provide a means of collecting objective data.

Kids'Cam was a cross-sectional observational study that aimed to determine the frequency, nature and duration of children's exposure to food and non-alcoholic beverage marketing. Ethical approval was obtained from the University of Otago Human Ethics Committee (Health) (13/220) to study any aspect of the world children live in and their interaction with it that was of public health interest, including the food available.

From July 2014 to June 2015, 169 randomly-selected children (11-13 years) were recruited from 16 randomly-selected schools in the Wellington region of New Zealand. The children were asked to wear an Autographer wearable camera and carry a GPS recorder all day for four days (two week and two weekend). Images were automatically captured every 7 seconds, and the children's location recorded every 5 seconds. 1.5 million images and 2.5 million GPS coordinates were recorded, and subsequently linked.

Bespoke software developed by ourselves was used to apply annotations to the images as illustrated in Figure 2. Each image was manually examined for the presence of food or non-alcoholic beverage marketing and food availability, and annotated according in a three-level, tree-branch-leaf configuration. An annotation schedule (with an 83 concept vocabulary) developed by the Kids'Cam team was used to guide image annotation for *setting > marketing medium and availability > food product category*. Marketing images were only annotated if 50% or more of a brand name or logo was clearly identified by the annotators. Annotators were tested for reliability before beginning annotation, requiring 90% concurrence with a model dataset. The majority of the data collected was usable and could be manually annotated.

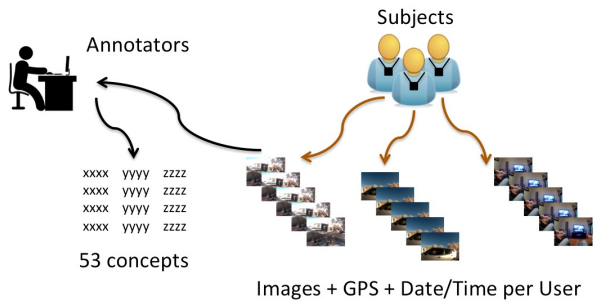


Figure 2: Kids'Cam data capture and annotation

Figures 3, 4 and 5 show examples of Kids'Cam images with accompanying manual annotations.

The problems we address in this paper about automatically annotating images are even more difficult than working


Figure 3: Manual Annotation: Shop front > sign > sugary drinks and juices

Figure 4: Manual Annotation: Convenience store indoors > in-store marketing > convenience store

Figure 5: Manual Annotation: School > sign > fast food

with other forms of image data because lifelog image data is notoriously difficult to process automatically [5] because of its nature. Some of the images are blurred because the subject is moving at the time of capture, some feature the subject's hands occluding what should be in the frame, there are issues associated with lighting conditions, etc. all of which make it a challenging application for computer vision. Despite this, we were able to annotate all the images manually which means we should be able to get some way further down the road of automatic annotation. As an additional impetus to doing this we have found that the Kids'Cam dataset has been used to investigate other aspects of children's lives beyond food and non-alcoholic beverage marketing, including exposure to alcohol marketing, food availability and after-school screen time. This means that there is now a need to develop more efficient and scalable methods of annotating these images for the presence of many different semantic concepts in a scalable way, aiming at thousands of concepts.

In the next section of the paper we describe how we have processed this image collection to achieve such concept indexing.

# 4. PROCESSING THE KIDS'CAM DATA

The data processing we carried out on the Kids'Cam dataset is best described diagrammatically in Figure 6 which summarises the process we followed. The numbers in that figure are used as an index into the following description of the processing:

1. As described in Section 3 of this paper, a group of 169 children from New Zealand each wore a wearable camera for a period of 4 days and ...

2. ... as a result, each subject generated tens of thousands of images which form the image dataset we use.

3. As well as the wearable camera, children also carried an external GPS unit, so each image has the usual meta-data of date and time, as well as GPS location.

4. A small group of manual annotators then tagged each image with a set of tags ...

5. ... from a vocabulary of 83 concepts, each concept being related to the topic of food advertising and ...

6. ... this led to a set of manual tags for each of the 1.5M images in the collection.

7. In a separate process, we analysed the content of each image using a deep convolutional neural network (CNN) to automatically apply semantic tagging to each of the 1.5 million Kids'Cam images. Specifically, we used the VGG-16 network [14], a very deep CNN with 138 million parameters in 16 parameter layers (13 convolutional and 3 fully connected). The network was trained on 1,000 object classes using 1.2 million images from the ImageNet large scale visual recognition challenge [13].

8. The trained model was then used to predict object class probabilities for each target image. The images were warped to a resolution of $224 \times 224$ prior to processing. For computational reasons, we did not use any test time data augmentation (flips or crops). Images were processed in batches of size 64, with the entire collection taking approximately 4 days to process using a NVIDIA GTX Titan X GPU. The processing pipeline was implemented in Python using the MXNet deep learning library [3].

9. Once these concept tags had been assigned, in previous work described elsewhere [17, 18], we had developed a technique called *training free refinement* (TFR), described later in this paper.

10. The new set of tag probabilities (aaaa', bbbb', etc.) shown in Figure 6 generated from the TFR process, are drawn from a vocabulary of 1,000 concepts. The net result of all this processing is that ...

11. ... we have a collection of over 1.5M images, each of which is assigned a set of terms from ...

12. ... a small vocabulary of only 83 concepts, but manually assigned and thus we can assume to be at least 90% reliable, and ...

13. ... a larger vocabulary of 1,000 concepts, automatically assigned and then refined, but still not likely to have all of them correct.

Similar to a (natural) language, semantic concepts can provide a natural way to describe and index vision content which is close to human expectations. In addition, a concept can be represented as text tags, facilitating structural management and linguistic organization in describing users' searching intentions. For example, we can describe Figure 1 with concepts "Bus", "Road", "Sky", etc. to approximate the actual content, which is "a London bus running on the road during daytime." While our (natural) language usually combines different concepts or tags in order to convey the semantics of visual content, the automatic concept detections are not carried out taking such semantic correlation in mind. Current typical concept detection requires a classifier for each concept without considering inter-concept relationships or dependencies. This is counter-intuitive as many concept pairs are often semantically related and dependent and thus will co-occur rather than occur independently.

It is not hard to imagine that the pairwise correlations of concepts increases exponentially as the number of concepts increases. The modelling of correlations between concepts in the classification phase suffers from high computational complexity as experienced in *multi-label training* [11, 20]. Besides how to measure the correlations semantically and to flexibly adapt to the evolution of concept lexicons, there are even more challenges. To alleviate such challenges, we turn to detection refinement in our previous work [17, 18] to improve the one-per-class concept detections by post-processing detection scores obtained from individual detectors, allowing independent and specialised classification techniques to be leveraged for each concept. In *training-free refinement* (TFR), introduced by Wang *et al.* in [17], the relationships between concepts including co-occurrence and re-occurrence relationships, as well as local neighbourhood information, are utilised to refine an initial set of tag probabilities. The method has already been evaluated on a dataset of wearable camera images which is similar to the Kids'Cam data capture as used in this paper.

In the Kids'Cam project, external contextual data, including GPS locations, have potential in further improving concept tagging performance. For example, two images taken in approximately the same location should have similar tags, even if the date/time and the subject wearing the camera are different. Such geographic distance between images can be leveraged as another source of similarity measure in precisely localising the neighbourhoods in similarity-based propagation as used by Wang *et al.* in [17]. However, in our previous work we did not apply TFR to such a large collection containing the kind of richer contextual information provided by Kids'Cam.

At the end of this data processing, in addition to an automatic annotation from 1,000 concepts which are then refined, we also have a manually-assigned ground truth, which is almost like a second description of each image, or more correctly it is a description in a second, smaller and more restricted, vocabulary. This provides an invaluable resource

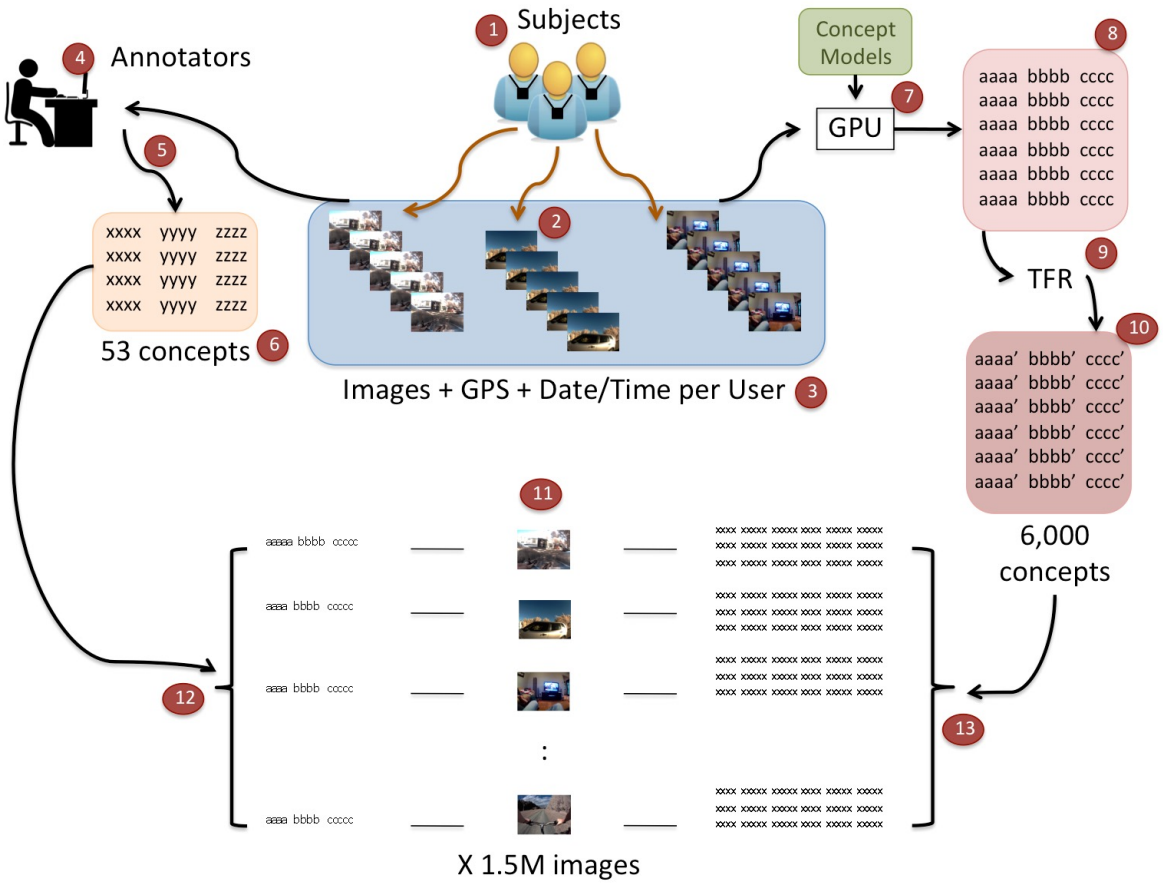


Figure 6: Workflow of image data capture, annotation and analysis

for exploring issues related to language and vision, which we will do in the next section.

## 5. LANGUAGE AND VISION OPPORTUNITIES

One of the intriguing possibilities that the creation of this dataset offers us, is to adjust the configuration of an automatically-generated concept space, populated with images. In our case we have 1.5M images automatically mapped into a concept space of 1,000 concepts, but we do not know anything about the accuracy of the mapping into the concept space unless we do some manual assessments of concepts to images, and we are not going to do that on any real scale. We also have our 1.5M images mapped to an 83-dimensional concept space, manually, and correctly. This is summarised in Figure 7 where 2 images (as a subset of the 1.5M) are mapped into a 2-dimensional space with values (x1, y1) and (x2, y2) respectively in Figure 7 (a), and the same images are mapped into a 3-dimensional space with values (a1, b1, c1) and (a2, b2, c2) in Figure 7 (b).

Given that the mapping in Figure 7 (a) is correct, because it is manually assigned, and the mapping of the same images, 1.5M of them, in Figure 7 (b) is not fully correct, even after our training-free refinement process is applied as described in section 4, can we adjust the values of concept weights in Figure 7 (b), anchored and pivoting around the 1.5M images?

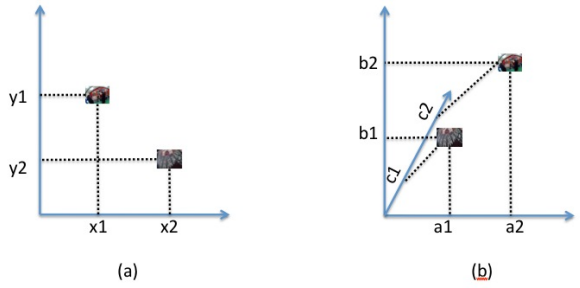


Figure 7: Same images mapped into two different concept spaces

This would be a variant of latent semantic indexing as used in information retrieval, where a document-term matrix is mapped to a matric reduced in dimensionality through a singular value decomposition, an idea first introduced in 1990 by [4]. In our case we want to create a mapping between a weighted image-concept matrix of low dimensionality and a weighted image-concept matrix of much higher dimensionality, and then to adjust that mapped matrix, in the context of the original, automatically-assigned concept weights.

This is shown conceptually in Figure 8 where the same images in Figure 8 (b) and Figure 8 (c), with green and with red highlight circles respectively, exist but the original concept weight assignment in Figure 8 (b), is adjusted as

a result of the correct assignment of (different) concepts in Figure 8 (a).

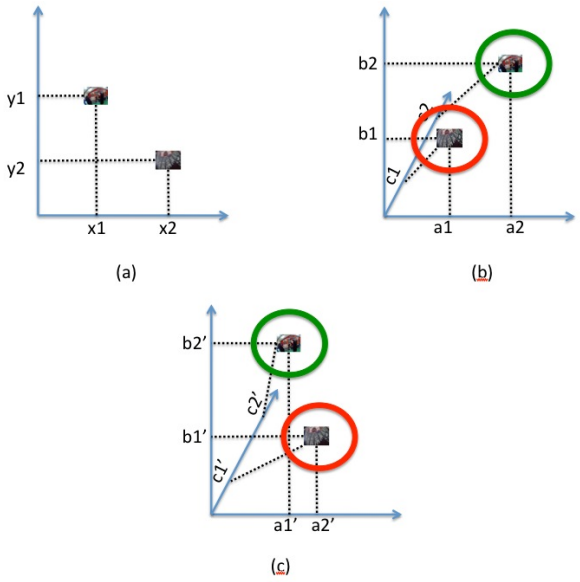


Figure 8: Images mapped into three concept spaces

By starting with a mapping of an image into two separate vocabulary spaces, we are comparing where that image has ended up in one concept space with where it has ended up in another. For a single image that does not mean much, but because we are able to do this for 1.5M we should have statistical advantage here. Semantic similarity among concepts has been previously done at an ontology level and is static. In the approach above we are data-driven and we can do this mapping 1.5M times.

Distributional semantics is a corpus driven approach to computational semantics and it is used to understand how meaning and relatedness arises in sentences. It is based on the distributional hypothesis which claims that co-occurring words in similar contexts tend to have similar meaning [6, 16]. We have developed a Distributional Semantics Infrastructure, DINFRA [2], which is a framework that offers a suite of DSMs (distributional semantic models), including Latent Semantic Analysis (LSA) mentioned earlier.

Distributional Semantics assumes that every word pair occurring in the same context has a certain relatedness, which can be enforced or diminished by its frequency, i.e., the more frequently a pair of words appear in the same context, the stronger the relatedness between them and vice-versa. Thus, any relation type is valid when measuring how related a pair is.

Another important concept in language processing, commonly mistaken with semantic relatedness, is semantic similarity which is a more restricted "special case" of semantic relatedness i.e. while cars and gasoline would seem to be more closely related than, say, cars and bicycles, but the latter pair are certainly more similar, in that *cars* and *bicycles* can be captured as taxonomic *IS_A* relations of *vehicle*. [12, p. 1]

Using the *word2vec* [8] model contained in DINFRA, we can map all the terms in a vocabulary to an n-dimensional vector space. We can obtain a relatedness score among terms

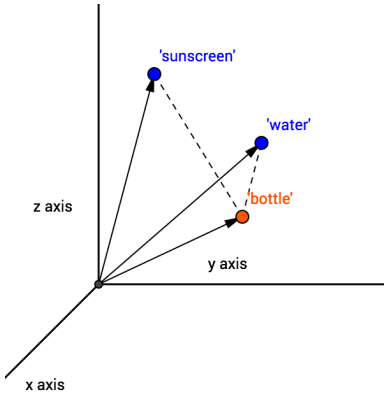


Figure 9: Visual representation of a possible 3-dimensional space with word tokens represented as vectors

by measuring the inversely proportional distance between two terms (Figure 9 gives a visual example). For each of the 1.5M images in the Kids'Cam collection, we can evaluate the relatedness score between the human annotated label and the automatically assigned tags with higher confidences. By doing so, we rely on the accuracy of the manual labels in order to enhance the results of the automatic visual recognition model.

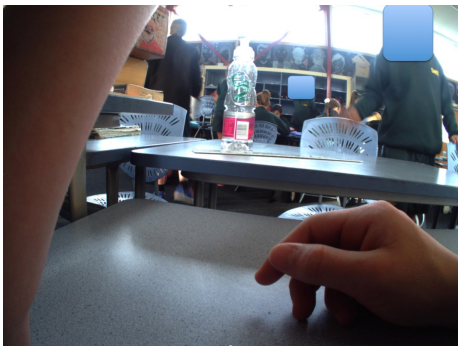

Figure 10: Manual Annotation: School > availability > drink bottle

From Figure 10, the top automatically-assigned tags and their respective confidences are listed in Table 2. In the rightmost two columns we can also see the relatedness score between each automatically assigned tag and in this example, the manual annotation *'drink bottle'* as well as the new confidence, obtained after the processing.

For instance, although the first tag, 'sunscreen, sunblock, sun blocker', has obtained a higher confidence score assigned by the visual recognition model, even after training-free refinement, because it is less related to the manual label the post-processed confidence is lower than the one received by the 'water bottle', which was closely related to the 'drink bottle'.

The methodology described above is a post-process applied to the outcome of the visual recognition model. In current work we are incorporating the highly accurate semantic information from the manual labels into the model by using one of DINFRA's DSMs to map words to a vector space, which will be used as an input parameter.

| Tag | Confidence | Relatedness | New Conf. |
|---|---|---|---|
| 'sunscreen, sunblock, sun blocker' | 0.077 | 0.1905 | 0.0147 |
| 'water bottle' | 0.040 | 1.0 | 0.0401 |
| 'beer glass' | 0.039 | 0.6367 | 0.0254 |
| 'bathing cap, swimming cap' | 0.037 | 0.3461 | 0.0129 |
| 'sunglass' | 0.032 | 0.3367 | 0.0109 |
| 'restaurant, eating house, eating place, eatery' | 0.030 | 0.6445 | 0.0198 |
| 'crossword puzzle, crossword' | 0.029 | 0.1663 | 0.0050 |
| 'goblet' | 0.025 | 0.3245 | 0.0083 |
| 'lampshade, lamp shade' | 0.020 | 0.3508 | 0.0074 |
| 'swimming trunks, bathing trunks' | 0.017 | 0.3461 | 0.0061 |

Table 2: Tags assigned to image in Figure 10

# 6. CONCLUSIONS

Natural language is notoriously difficult as a medium for tagging or describing image data, like photos or videos, and it is even more difficult as a medium for formulating an information need when we have to develop queries. Since the automatic detection of semantic concepts in image data based on using predefined models has made so much progress in recent years we are now seeing the problem of vocabulary or concept space mis-match, between the vocabulary used in indexing images and the vocabulary used in formulating queries.

To explore issues related to how language and vision interact, we introduced a dataset of over 1.5M images taken by children using wearable cameras. This has been annotated manually and automatically using two different vocabularies., one of 83 words and the other containing 1,000 concepts. By pivoting around the representations of images in the two vocabularies we are able to map the vocabulary spaces onto each other, and since one of these is manually assigned and thus correct, we can compensate for errors in automatic assignment of concept tags. To illustrate this we included a worked example of the set of tags for one of the images and we are presently applying this process to the remainder, which will, in aggregate, allow us to compare the two vocabulary concept spaces.

While this work reflects an advance being made in automatic description of visual content it does have a limitation in that it leverages co-occurrences across concepts to estimate the confidence of a concept being present in an image while in practice this will not work for instances where we have concepts that unexpectedly occur together. So co-occurrence may offer a silver bullet in the *majority* of instances, it won't always work.

Ultimately the impact of our work will be to exploit any manual assignment of tags or descriptors to images, something which is currently not done and which is necessary as we start to see automatic tagging of images becoming more mainstream.

The KidsâĂŹCam methodology enables automated, objective observation of childrenâĂŹs lived experience. To our knowledge, this is the first study to objectively research food availability and marketing from a childâĂŹs perspective. While manual image annotation was successful, as noted earlier it was time consuming and will not scale to large datasets. The option of crowdsourcing the annotation task is not an option here because of the nuber of images and because of data privacy restrictions and ethics issues. The automated image recognition explored here has considerable potential for efficient, scalable, comprehensive analysis. As such, it will likely contribute to better knowledge and therefore more effective public health action.

In future work we plan to do some analysis of the reliability and accuracy of our method by choosing a subset of the 1,000 concepts and manually annotating them with multiple annotators in order to get high reliability. This will allow us to investigate the most successful aspects of our method as well as to carry out a failure analysis of when it does not work.

## Acknowledgments

This publication has resulted from research partly funded by Science Foundation Ireland under grant number SFI/12/RC /2289 (Insight Centre), the National Natural Science Foundation of China under Grant Nos. 61272231, 61472204, 61502264, Beijing Key Laboratory of Networked Multimedia. Kids'Cam is a project in the DIET Programme funded by a Health Research Council of New Zealand Programme Grant (13/724). Thanks to the children who participated in Kids'Cam, and to their parents and schools for giving the Kids'Cam team permission to work with them.

# 7. REFERENCES

[1] G. Awad, C. G. M. Snoek, A. F. Smeaton, and G. Quénot. TRECVid Semantic Indexing of Video: A 6-Year Retrospective. *ITE Transactions on Media Technology and Applications*, pages 1–22, 2016. (in press).

[2] S. Barzegar, J. E. Sales, A. Freitas, S. Handschuh, and B. Davis. Dinfra: A one stop shop for computing multilingual semantic relatedness. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 1027–1028, New York, NY, USA, 2015. ACM.

[3] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391, 1990.

[5] C. Gurrin, A. F. Smeaton, and A. R. Doherty. Lifelogging: Personal big data. *Foundations and Trends in Information Retrieval*, 8(1):1–125, 2014.

[6] Z. S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954.

[7] K. McGuinness, R. Aly, K. Chatfield, O. Parkhi, R. Arandjelovic, M. Douze, M. Kemman, M. Kleppe, P. Van Der Kreeft, K. Macquarrie, et al. The axes research video search system. In *IEEE ICASSP-International Conference on Acoustics, Speech and Signal Processing*, pages 4–9, 2014.

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.

[9] Ministry of Health. New Zealand Health Survey. Annual update of key findings 2014/15. Wellington: Ministry of Health. http://www.health.govt.nz/publication/annual-update-key-results-2014-15-new-zealand-health-survey. Accessed Mar 15, 2016.

[10] OECD. Obesity Update. http://www.oecd.org/els/health-systems/Obesity-Update-2014.pdf. Accessed Oct 3, 2015.

[11] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, M. Wang, and H.-J. Zhang. Correlative multilabel video annotation with temporal kernels. *ACM Trans. Multimedia Comput. Commun. Appl.*, 5(1):3:1–3:27, Oct. 2008.

[12] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *CoRR*, abs/1105.5444, 2011.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[15] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[16] P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, Jan. 2010.

[17] P. Wang, L. Sun, S. Yang, and A. F. Smeaton. Towards training-free refinement for semantic indexing of visual media. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part I*, pages 251–263, Cham, 2016. Springer International Publishing.

[18] P. Wang, L. Sun, S. Yang, A. F. Smeaton, and C. Gurrin. Characterizing everyday activities from visual lifelogs based on enhancing concept representation. *Computer Vision and Image Understanding*, 148:181–192, 2016. Special issue on Assistive Computer Vision and Robotics: Assistive Solutions for Mobility, Communication and HMI.

[19] WHO. Report of the Commission on Ending Childhood Obesity. Geneva: World Health Organization. http://apps.who.int.wmezproxy.wnmeds.ac.nz/iris/bitstream/10665/204176/1/9789241510066_eng.pdf. Accessed Dec 18, 2015.

[20] X. Xue, W. Zhang, J. Zhang, B. Wu, J. Fan, and Y. Lu. Correlative multi-label multi-instance image annotation. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 651–658, Nov 2011.