Improving Multi-Modal Representations Using Image Dispersion: Why Less is Sometimes More

Douwe Kiela*, Felix Hill*, Anna Korhonen and Stephen Clark
University of Cambridge
Computer Laboratory
{douwe.kiela|felix.hill|anna.korhonen|stephen.clark}@cl.cam.ac.uk

Abstract

Models that learn semantic representations from both linguistic and perceptual input outperform text-only models in many contexts and better reflect human concept acquisition. However, experiments suggest that while the inclusion of perceptual input improves representations of certain concepts, it degrades the representations of others. We propose an unsupervised method to determine whether to include perceptual input for a concept, and show that it significantly improves the ability of multi-modal models to learn and represent word meanings. The method relies solely on image data, and can be applied to a variety of other NLP tasks.

1 Introduction

Multi-modal models that learn semantic concept representations from both linguistic and perceptual input were originally motivated by parallels with human concept acquisition, and evidence that many concepts are *grounded* in the perceptual system (Barsalou et al., 2003). Such models extract information about the perceptible characteristics of words from data collected in property norming experiments (Roller and Schulte im Walde, 2013; Silberer and Lapata, 2012) or directly from 'raw' data sources such as images (Feng and Lapata, 2010; Bruni et al., 2012). This input is combined with information from linguistic corpora to produce enhanced representations of concept meaning. Multi-modal models outperform language-only models on a range of tasks, including modelling conceptual association and predicting compositionality (Bruni et al., 2012; Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013).

Despite these results, the advantage of multi-modal over linguistic-only models has only been demonstrated on concrete concepts, such as *chocolate* or *cheeseburger*, as opposed to abstract concepts such as such as *guilt* or *obesity*. Indeed, experiments indicate that while the addition of perceptual input is generally beneficial for representations of concrete concepts (Hill et al., 2013a; Bruni et al., 2014), it can in fact be detrimental to representations of abstract concepts (Hill et al., 2013a). Further, while the theoretical importance of the perceptual modalities to concrete representations is well known, evidence suggests this is not the case for more abstract concepts (Paivio, 1990; Hill et al., 2013b). Indeed, perhaps the most influential characterization of the abstract/concrete distinction, the Dual Coding Theory (Paivio, 1990), posits that concrete representations are encoded in both the linguistic and perceptual modalities whereas abstract concepts are encoded only in the linguistic modality.

Existing multi-modal architectures generally extract and process all the information from their specified sources of perceptual input. Since perceptual data sources typically contain information about both abstract and concrete concepts, such information is included for both concept types. The potential effect of this design decision on performance is significant because the vast majority of meaning-bearing words in everyday language correspond to abstract concepts. For instance, 72% of word tokens in the British National Corpus (Leech et al., 1994) were rated by contributors to the University of South Florida dataset (USF) (Nelson et al., 2004) as more abstract than the noun *war*, a concept that many would consider quite abstract.

In light of these considerations, we propose a novel algorithm for approximating conceptual concreteness. Multi-modal models in which perceptual input is filtered according to our algorithm learn higher-quality semantic representations than previous approaches, resulting in a significant performance improvement of up to 17% in captur-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 835–841,
Baltimore, Maryland, USA, June 23-25 2014.

ing the semantic similarity of concepts. Further, our algorithm constitutes the first means of quantifying conceptual concreteness that does not rely on labor-intensive experimental studies or annotators. Finally, we demonstrate the application of this unsupervised concreteness metric to the semantic classification of adjective-noun pairs, an existing NLP task to which concreteness data has proved valuable previously.

2 Experimental Approach

Our experiments focus on multi-modal models that extract their perceptual input automatically from images. Image-based models more naturally mirror the process of human concept acquisition than those whose input derives from experimental datasets or expert annotation. They are also more scalable since high-quality tagged images are freely available in several web-scale image datasets.

We use *Google Images* as our image source, and extract the first n image results for each concept word. It has been shown that images from Google yield higher-quality representations than comparable sources such as *Flickr* (Bergsma and Goebel, 2011). Other potential sources, such as ImageNet (Deng et al., 2009) or the ESP Game Dataset (Von Ahn and Dabbish, 2004), either do not contain images for abstract concepts or do not contain sufficient images for the concepts in our evaluation sets.

2.1 Image Dispersion-Based Filtering

Following the motivation outlined in Section 1, we aim to distinguish visual input corresponding to concrete concepts from visual input corresponding to abstract concepts. Our algorithm is motivated by the intuition that the diversity of images returned for a particular concept depends on its concreteness (see Figure 1). Specifically, we anticipate greater congruence or similarity among a set of images for, say, *elephant* than among images for *happiness*. By exploiting this connection, the method approximates the concreteness of concepts, and provides a basis to filter the corresponding perceptual information.

Formally, we propose a measure, *image dispersion* d of a concept word w, defined as the average pairwise cosine distance between all the image representations $\{\vec{w_1} \dots \vec{w_n}\}$ in the set of images for that concept:

$$d(w) = \frac{1}{2n(n-1)} \sum_{i<j\leq n} 1 - \frac{\vec{w_i} \cdot \vec{w_j}}{|\vec{w_i}||\vec{w_j}|} \tag{1}$$

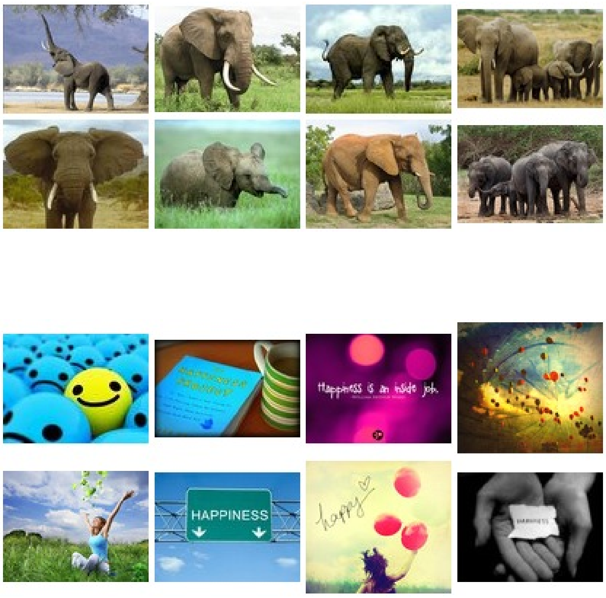

Figure 1: Example images for a concrete (*elephant* – little diversity, low dispersion) and an abstract concept (*happiness* – greater diversity, high dispersion).

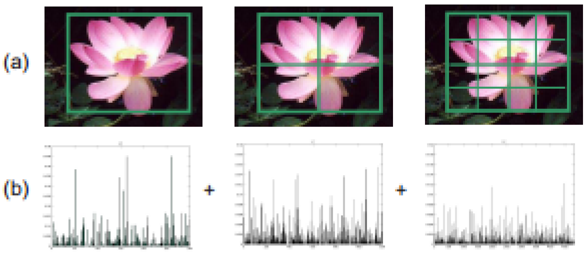

Figure 2: Computation of PHOW descriptors using dense SIFT for levels $l = 0$ to $l = 2$ and the corresponding histogram representations (Bosch et al., 2007).

We use an average pairwise distance-based metric because this emphasizes the total variation more than e.g. the mean distance from the centroid. In all experiments we set $n = 50$.

Generating Visual Representations Visual vector representations for each image were obtained using the well-known *bag of visual words* (BoVW) approach (Sivic and Zisserman, 2003). BoVW obtains a vector representation for an image by mapping each of its local descriptors to a cluster histogram using a standard clustering algorithm such as k-means.

Previous NLP-related work uses *SIFT* (Feng and Lapata, 2010; Bruni et al., 2012) or *SURF* (Roller and Schulte im Walde, 2013) descriptors for identifying points of interest in an image, quantified by 128-dimensional local descriptors. We apply *Pyramid Histogram Of visual Words* (PHOW) descriptors, which are particularly well-suited for object categorization, a key component of image similarity and thus dispersion (Bosch et al., 2007). PHOW is roughly equivalent to running SIFT on a dense grid of locations at a fixed scale and orientation and at multiple scales (see Fig 2), but is both more efficient and more accurate than regular (dense) SIFT approaches (Bosch et al., 2007). We resize the images in our dataset to 100x100 pixels and compute PHOW descriptors using *VLFeat* (Vedaldi and Fulkerson, 2008).

The descriptors for the images were subsequently clustered using mini-batch k-means (Sculley, 2010) with $k = 50$ to obtain histograms of visual words, yielding 50-dimensional visual vectors for each of the images.

Generating Linguistic Representations We extract continuous vector representations (also of 50 dimensions) for concepts using the continuous log-linear skipgram model of Mikolov et al. (2013a), trained on the 100M word British National Corpus (Leech et al., 1994). This model learns high quality lexical semantic representations based on the distributional properties of words in text, and has been shown to outperform simple distributional models on applications such as semantic composition and analogical mapping (Mikolov et al., 2013b).

2.2 Evaluation Gold-standards

We evaluate models by measuring the Spearman correlation of model output with two well-known gold-standards reflecting semantic proximity – a standard measure for evaluating the quality of representations (see e.g. Agirre et al. (2009)).

To test the ability of our model to capture concept similarity, we measure correlations with WordSim353 (Finkelstein et al., 2001), a selection of 353 concept pairs together with a similarity rating provided by human annotators. WordSim has been used as a benchmark for distributional semantic models in numerous studies (see e.g. (Huang et al., 2012; Bruni et al., 2012)).

As a complementary gold-standard, we use the University of South Florida Norms (USF) (Nelson et al., 2004). This dataset contains scores for *free association*, an experimental measure of cognitive association, between over 40,000 concept pairs. The USF norms have been used in many previous studies to evaluate semantic representations (Andrews et al., 2009; Feng and Lapata, 2010; Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013). The USF evaluation set is particularly appropriate in the present context because concepts in the dataset are also rated for conceptual concreteness by at least 10 human annotators.

We create a representative evaluation set of USF pairs as follows. We randomly sample 100 concepts from the upper quartile and 100 concepts from the lower quartile of a list of all USF concepts ranked by concreteness. We denote these sets C, for *concrete*, and A for abstract respectively. We then extract all pairs (w_1, w_2) in the USF dataset such that both w_1 and w_2 are in $A \cup C$. This yields an evaluation set of 903 pairs, of which 304 are such that $w_1, w_2 \in C$ and 317 are such that $w_1, w_2 \in A$.

The images used in our experiments and the evaluation gold-standards can be downloaded from `http://www.cl.cam.ac.uk/~dk427/dispersion.html`.

3 Improving Multi-Modal Representations

We apply *image dispersion-based filtering* as follows: if both concepts in an evaluation pair have an image dispersion below a given threshold, both the linguistic and the visual representations are included. If not, in accordance with the Dual Coding Theory of human concept processing (Paivio, 1990), only the linguistic representation is used. For both datasets, we set the threshold as the median image dispersion, although performance could in principle be improved by adjusting this parameter. We compare dispersion filtered representations with linguistic, perceptual and standard multi-modal representations (concatenated linguistic and perceptual representations). Similarity between concept pairs is calculated using cosine similarity.

As Figure 3 shows, dispersion-filtered multi-modal representations significantly outperform

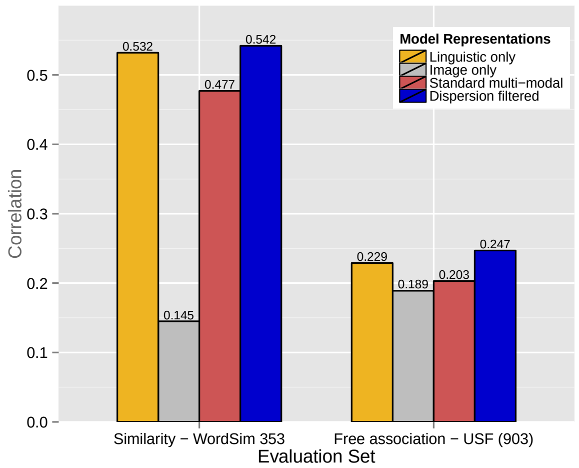

Figure 3: Performance of conventional multi-modal (visual input included for all concepts) vs. image dispersion-based filtering models (visual input only for concepts classified as concrete) on the two evaluation gold-standards.

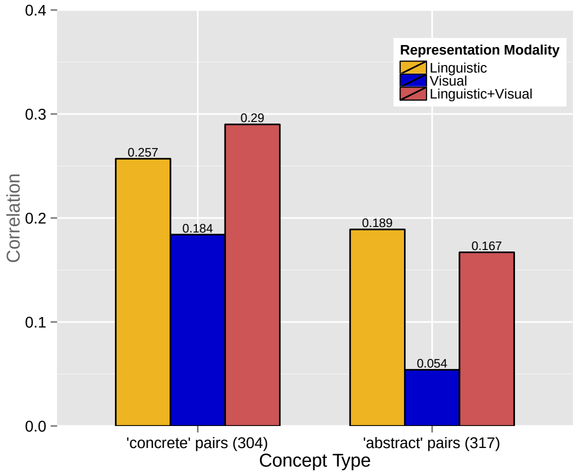

Figure 4: Visual input is valuable for representing concepts that are classified as concrete by the image dispersion algorithm, but not so for concepts classified as abstract. All correlations are with the USF gold-standard.

standard multi-modal representations on both evaluation datasets. We observe a 17% increase in Spearman correlation on WordSim353 and a 22% increase on the USF norms. Based on the correlation comparison method of Steiger (1980), both represent significant improvements (WordSim353, $t = 2.42$, $p < 0.05$; USF, $t = 1.86$, $p < 0.1$). In both cases, models with the dispersion-based filter also outperform the purely linguistic model, which is not the case for other multi-modal approaches that evaluate on WordSim353 (e.g. Bruni et al. (2012)).

4 Concreteness and Image Dispersion

The filtering approach described thus far improves multi-modal representations because image dispersion provides a means to distinguish concrete concepts from more abstract concepts. Since research has demonstrated the applicability of concreteness to a range of other NLP tasks (Turney et al., 2011; Kwong, 2008), it is important to examine the connection between image dispersion and concreteness in more detail.

4.1 Quantifying Concreteness

To evaluate the effectiveness of image dispersion as a proxy for concreteness we evaluated our algorithm on a binary classification task based on the set of 100 concrete and 100 abstract concepts $A \cup C$ introduced in Section 2. By classifying concepts with image dispersion below the median as concrete and concepts above this threshold as abstract we achieved an abstract-concrete prediction accuracy of 81%.

While well-understood intuitively, concreteness is not a formally defined notion. Quantities such as the USF concreteness score depend on the subjective judgement of raters and the particular annotation guidelines. According to the Dual Coding Theory, however, concrete concepts are precisely those with a salient perceptual representation. As illustrated in Figure 4, our binary classification conforms to this characterization. The importance of the visual modality is significantly greater when evaluating on pairs for which both concepts are classified as concrete than on pairs of two abstract concepts.

Image dispersion is also an effective predictor of concreteness on samples for which the abstract/concrete distinction is less clear. On a different set of 200 concepts extracted by random sampling from the USF dataset stratified by concreteness rating (including concepts across the concreteness spectrum), we observed a high correlation between abstractness and dispersion (Spearman $\rho = 0.61, p < 0.001$). On this more diverse sample, which reflects the range of concepts typically found in linguistic corpora, image dispersion is a particularly useful diagnostic for identifying

Concept	Image Dispersion	Conc. (USF)
shirt	.488	6.05
bed	.495	5.91
knife	.560	6.08
dress	.578	6.59
car	.580	6.35
ego	1.000	1.93
nonsense	.999	1.90
memory	.999	1.78
potential	.997	1.90
know	.996	2.70

Table 1: Concepts with highest and lowest image dispersion scores in our evaluation set, and concreteness ratings from the USF dataset.

the very abstract or very concrete concepts. As Table 1 illustrates, the concepts with the lowest dispersion in this sample are, without exception, highly concrete, and the concepts of highest dispersion are clearly very abstract.

It should be noted that all previous approaches to the automatic measurement of concreteness rely on annotator ratings, dictionaries or manually-constructed resources. Kwong (2008) proposes a method based on the presence of hard-coded phrasal features in dictionary entries corresponding to each concept. By contrast, Sánchez et al. (2011) present an approach based on the position of word senses corresponding to each concept in the WordNet ontology (Fellbaum, 1999). Turney et al. (2011) propose a method that extends a large set of concreteness ratings similar to those in the USF dataset. The Turney et al. algorithm quantifies the concreteness of concepts that lack such a rating based on their proximity to rated concepts in a semantic vector space. In contrast to each of these approaches, the image dispersion approach requires no hand-coded resources. It is therefore more scalable, and instantly applicable to a wide range of languages.

4.2 Classifying Adjective-Noun Pairs

Finally, we explored whether image dispersion can be applied to specific NLP tasks as an effective proxy for concreteness. Turney et al. (2011) showed that concreteness is applicable to the classification of adjective-noun modification as either literal or non-literal. By applying a logistic regression with noun concreteness as the predictor variable, Turney et al. achieved a classification accuracy of 79% on this task. This model relies on significant supervision in the form of over 4,000 human lexical concreteness ratings.[1] Applying image dispersion in place of concreteness in an identical classifier on the same dataset, our entirely unsupervised approach achieves an accuracy of 63%. This is a notable improvement on the largest-class baseline of 55%.

5 Conclusions

We presented a novel method, image dispersion-based filtering, that improves multi-modal representations by approximating conceptual concreteness from images and filtering model input. The results clearly show that including more perceptual input in multi-modal models is not always better. Motivated by this fact, our approach provides an intuitive and straightforward metric to determine whether or not to include such information.

In addition to improving multi-modal representations, we have shown the applicability of the image dispersion metric to several other tasks. To our knowledge, our algorithm constitutes the first unsupervised method for quantifying conceptual concreteness as applied to NLP, although it does, of course, rely on the Google Images retrieval algorithm. Moreover, we presented a method to classify adjective-noun pairs according to modification type that exploits the link between image dispersion and concreteness. It is striking that this apparently linguistic problem can be addressed solely using the raw data encoded in images.

In future work, we will investigate the precise quantity of perceptual information to be included for best performance, as well as the optimal filtering threshold. In addition, we will explore whether the application of image data, and the interaction between images and language, can yield improvements on other tasks in semantic processing and representation.

Acknowledgments

DK is supported by EPSRC grant EP/I037512/1. FH is supported by St John's College, Cambridge. AK is supported by The Royal Society. SC is supported by ERC Starting Grant DisCoTex (306920) and EPSRC grant EP/I037512/1. We thank the anonymous reviewers for their helpful comments.

[1] The MRC Psycholinguistics concreteness ratings (Coltheart, 1981) used by Turney et al. (2011) are a subset of those included in the USF dataset.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 19–27, Boulder, Colorado.

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological review*, 116(3):463.

Lawrence W Barsalou, W Kyle Simmons, Aron K Barbey, and Christine D Wilson. 2003. Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, 7(2):84–91.

Shane Bergsma and Randy Goebel. 2011. Using visual information to predict lexical preference. In *RANLP*, pages 399–405.

Anna Bosch, Andrew Zisserman, and Xavier Munoz. 2007. Image classification using random forests and ferns. In *Proceedings of ICCV*.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Max Coltheart. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4):497–505.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

Christiane Fellbaum. 1999. *WordNet*. Wiley Online Library.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Felix Hill, Douwe Kiela, and Anna Korhonen. 2013a. Concreteness and corpora: A theoretical and practical analysis. *CMCL 2013*.

Felix Hill, Anna Korhonen, and Christian Bentz. 2013b. A quantitative empirical analysis of the abstract/concrete distinction. *Cognitive science*, 38(1).

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Oi Yee Kwong. 2008. A preliminary study on the impact of lexical concreteness on word sense disambiguation. In *PACLIC*, pages 235–244.

Geoffrey Leech, Roger Garside, and Michael Bryant. 1994. Claws4: the tagging of the british national corpus. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 622–628. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of International Conference of Learning Representations*, Scottsdale, Arizona, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Allan Paivio. 1990. *Mental representations: A dual coding approach*. Oxford University Press.

Stephen Roller and Sabine Schulte im Walde. 2013. A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157, Seattle, Washington, USA, October. Association for Computational Linguistics.

David Sánchez, Montserrat Batet, and David Isern. 2011. Ontology-based information content computation. *Knowledge-Based Systems*, 24(2):297–303.

D Sculley. 2010. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178. ACM.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods* *in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433. Association for Computational Linguistics.

J. Sivic and A. Zisserman. 2003. Video Google: a text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, volume 2, pages 1470–1477, Oct.

James H Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245.

Peter D Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing*, pages 680–690.

A. Vedaldi and B. Fulkerson. 2008. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`.

Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM.