# Labelling Images without Classifiers

Theodore Boyd and Lewis D. Griffin

Computer Science, University College London, London WC1E 6BT, UK
{t.boyd,l.griffin}@cs.ucl.ac.uk

**Abstract.** Verifying that freight containers contain the items listed on their manifests from X-ray scans is a suitable problem for computer vision. However standard techniques do not cope well with the huge numbers of possible categories of cargo, nor with the scarcity of training data for many of these categories. The previously proposed `MIRRORING` algorithm potentially offers a way to deal with such a problem. The algorithm is based on the Appearance Hypothesis that "words with similar contexts tend to have referents with similar appearance" which allows all training images to be relevant to all labels, by exploiting the full range of this relationship (including dissimilar and intermediate relationships). Previously, this algorithm has only been demonstrated when applied to labelling categories of object, represented by a set of 50 images, with a single label. In this work, we demonstrate the algorithm operating on single images with multiple labels.

**Keywords:** multi-label, classification, semantics, machine learning, image processing, security science

## 1 Introduction

At ports of entry (POEs) into a country, governments usually require the inspection of incoming and outgoing goods. At many POEs, freight passing through is imaged by an X-ray scanner like those used for passenger hand luggage but larger and using higher energies, with an X-ray generating linear accelerator on one side and a vertical linear array of detectors on the other [1]. For the high-throughput of a POE, automated analysis of the image would reduce costs and interruptions to the flow of commerce. Such automated analysis could include:

- Detection and localisation of specific objects of interest (eg: high value items such as vehicles, weapons, narcotics and people) [2].
- Detection of anomalous or unexpected images requiring human inspection.
- Manifest verification: ascertaining whether the contents of the container agree with the legal declaration of its contents in type and amount [3].

This process is currently performed by customs staff at the POE using a combination of targeted image and direct visual inspection. The current paper reports work towards an eventual goal of automating the image-based part of manifest verification.

To develop an automated computer-based manifest verification system we need to overcome two domain-specific problems which prevent usage of standard computer vision recognition methods:

1. There is a large number of categories of object possible: the Harmonised Coding system (an internationally agreed list which defines manifest codes and categories) has $10^6$ possible codes [4].
2. It is not possible to collect adequately sized training sets for each category. Indeed many categories we may encounter have no training examples. This is due to the technology being fairly new, the lack of organised databases and the legal and other difficulties in gaining access to these data.

These two problems lead us to focus on an approach different from the standard one-classifier-per-category, instead we hope to exploit the similarities in appearance between categories. This approach was discussed and implemented successfully for object categories appearing in photographic images by Griffin et al. [5]. We wish to discover whether we can apply these ideas to the current problem. In this present work we build on Griffin et al. [5] by taking two steps towards the full manifest verification problem:

1. Our inputs are single images rather than categories of object (represented by a set of 50 images in Griffin et al.).
2. We use multiple labels per image rather than a single label per category.

To test our algorithms, we use the multi-label classification problem proposed at the 2013 ICML Workshop on Representation Learning [6] and publicly available as a competition on the Kaggle service [7]. The Kaggle challenge training dataset consists of 100,000 photographic images paired with unstructured sets of labels (word bags). The challenge is to learn the relation between images and word bags, so that on a disjoint test set the correct word bag, from a choice of two, can be picked for each image (Fig. 1).

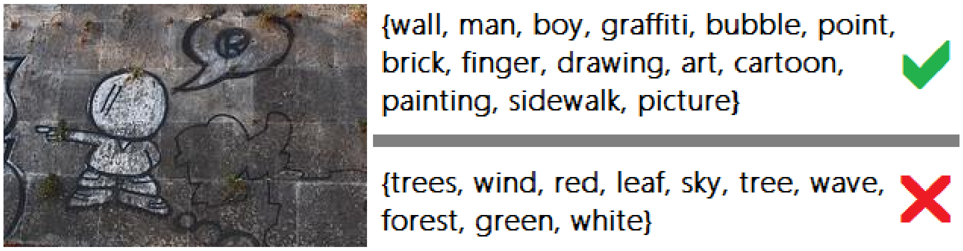

{wall, man, boy, graffiti, bubble, point, brick, finger, drawing, art, cartoon, painting, sidewalk, picture} ✓
{trees, wind, red, leaf, sky, tree, wave, forest, green, white} ✗

Fig. 1: An example image with the correct and a random incorrect bag of words.

Relating back to freight scanning, the Kaggle images are analogous to freight scans and the word bags to individual manifest files containing multiple codes.

## 2 Object Classification in Computer Vision

Now we briefly discuss possible traditional object classification methods that could be applied to the problem.

In an overview of multi-label classification [8], Tsoumakas and Katakis explain that multi-label classification, previously limited in its scope to text categorisation and medical diagnosis, is now being used in fields as disparate as protein function classification, music categorisation and semantic scene classification. Our example in Fig. 1 contains a chalk drawing of a boy on a wall. While it would be a fair training example for each of its contained classes individually, it is best described by the multi-label bag {walk, boy, drawing, . . . }, making it suited to the multi-label classification problem.

Another technique that can be used is multiple-instance learning, whereby classification is trained using both positive and negative instances [9]. Negative instances, say Maron and Lozano-Pérez, are those bags of words in which none of the labels are present in the image — therefore, non-matching words are permitted in the positive examples.

There are numerous such general methods for performing multi-label learning and they have been previously summarised thoroughly [10]. We shall consider mainly what is a good baseline method for multi-label classification: training one classifier per label in a "one versus all" fashion, also known as "binary relevance" [11][12]. This method involves producing as outputs predictions for all of the labels for an unseen input whose binary classifiers return a positive match. This is an effective technique and will be the one we shall employ due to its clarity and meaningful foundation. Alternatives include training a classifier for every possible label combination or using dependencies between labels in the form of *classifier chains*. These multi-label learning methods can be categorised into three major categories [10] [8]. The first is dubbed algorithm adaptation and involves modifying an existing machine learning algorithm for the multi-label problem such as Multi-Label k-Nearest Neighbour (ML-kNN) [13]. The second is problem transformation and works by changing the multi-label problem into a combination of multiple single-label problems which can then be tackled by existing algorithms. Finally, a third idea from Madjarov et al. is that of ensemble methods. They are a hybrid technique that build upon algorithm adaptation and problem transformation methods [10], [8] by either training classifiers on complete subsets of labels (as in RAKEL [14]) or by using additional information such as label dependencies (as in Ensembles of Classifier Chains (ECC) [12]).

## 3 Distributional Learning of Appearance

Due to the two problems discussed in Section 1, the above described methods are unsuitable for our particular problem. Instead we take a different approach. The major concept used, which has previously been demonstrated [5], is a correlation between word and appearance similarity. Exploitation of this correlation has been proposed as underlying the ability of children to correctly extrapolate labels for more object categories than they have explicitly been taught. Correlation is expressed by the Appearance Hypothesis that "words with similar contexts tend to have referents with similar appearance" [5]. The Appearance Hypothesis derives from a broader, older hypothesis: the Symbol Interdependency Hypothesis that relationships embodied in the world can be found in language (among others, [15]). Using the Appearance Hypothesis, we no longer need to rely on class-specific training sets, instead all images from all classes can contribute to the recognition of each label, which potentially deals with the two problems for manifest verification that we identified.

To make use of the Appearance Hypothesis requires a distance measure between words based on their usage patterns and an image-based distance between appearances [16]. The measure of distance between words is termed distributional similarity. An explicit implementation of it is the Correlated Occurrence Analogue to Lexical Semantics (COALS) algorithm [17]. COALS constructs a semantic space of words based on statistical analysis of a corpus of documents in the applicable language. We use the British National Corpus which comprises 97 million words in large collection of written and spoken texts [18]. COALS computes a score for how often a word $x$ appears in the neighbourhood of a word $y$, taking account of the frequency of both. These scores are assembled into a distributional vector for each word $y$. For words $x$ we use the $14,000$ most common (non-stop) words in the corpus. For words $y$ we use all the labels present in the Kaggle dataset. Distributional vectors between words are compared by correlation, rescaled from $[-1,1]$ to $[1,0]$ so that they are like distances.

For appearance similarity we need image descriptors that are suitable for a set of images which are diverse in context and layout, and that have some traction on semantic context. Histogram methods are suitable for this. There are many to chose from (eg: quantised SIFT [19] and Histogram of Oriented Gradients (HOG) [20]) but following Griffin et al. [5], we use oriented Basic Image Features (oBIFs) and Basic Colour histograms [21] [22].

oBIFs are computed by convolving the input image with a set of six derivative of Gaussian filters [21]. From filter responses, a pixel-by-pixel classification of the image into seven symmetry types is computed (roughly: flat, sloped, minimum, maximum, dark line, light line and saddle point). Quantised orientations are also computed with the calculation depending on the symmetry type. Accounting for symmetry type and orientation yields 23 possible pixel labels. oBIFs are calculated using filters of two different scales, yielding $529 = 23^2$ possible pixel codes. The frequencies of these codes are tallied and normalised into a 529-bin histogram.

Colour histograms are calculated by binning RGB values according to a partition of the colour cube into 11 Basic Colour categories [23], [24]. A histogram of these eleven bins forms the colour feature.

Appearance similarity, or distance, is then defined as $arccos(h_i, h_j)$ between two square-rooted normalised histograms $h_i$ and $h_j$ in column vector form.

Griffin et al. [5] also describe a `MIRRORING` algorithm that uses word and appearance distances to assess the compatibility of a label and an image by comparison to a reference set of labelled images. Specifically, the better a label $w$ is for an image $I$, the better the correlation between (i) the image distances from $I$ to the reference set of images, and (ii) the word distances from $w$ to the labels of the reference set. This is illustrated in Fig. 2.

## 4 Experiments

Using the approach described in Section 3, the experiments that are reported in this paper are as follows:

1. Baseline `MIRRORING` implementation. We test whether the algorithm is able to deliver above chance performance even though we apply it to individual images, rather than sets of 50, labelled with a bag of words, rather than a single term. We assess the effect of reference set size, and the performance of oBIF- or colour-based image distance alone or their combination.
2. Comparison of different ways to define build bag-bag distances from word-word distances.

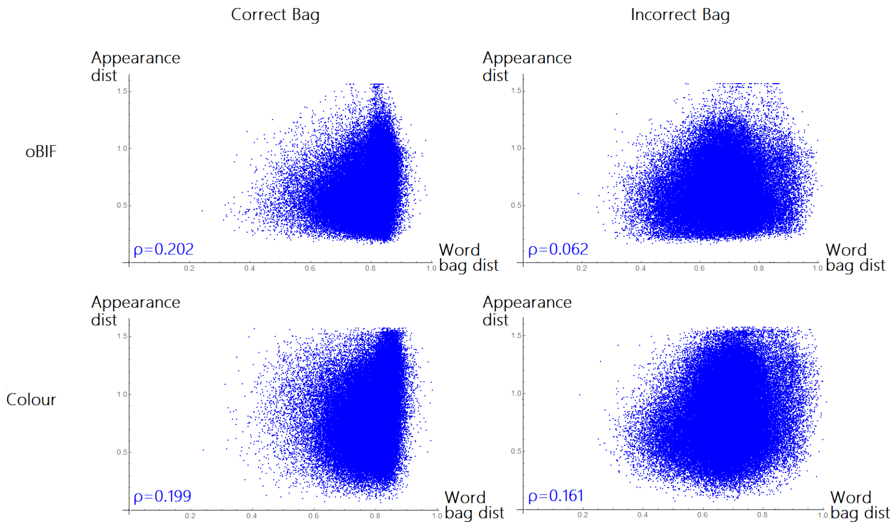


Fig. 2: This figure illustrates the process of choosing the correct word bag for the example in Fig. 1. Scatter plots show appearance distance (vertical) versus word bag distance (horizontal) between the query image and the reference set of 100,000 images. The top row is for oBIF-based appearance distance, the bottom row for colour. The left column is when the correct word bag is paired with the query image, the right the incorrect. In this case, the correct word bag is identified whether colour or oBIFs are used because the plots on the left have larger correlations ($\rho$) than those on the right.

3. Assessment of whether more-frequently occurring words are more useful labels than less-frequently occurring ones.
4. Investigation of whether semantically deep (eg: “tree”, “leaf”) words are more useful labels than semantically shallow (eg: “solid”, “organism”) ones.
5. Assessment of whether the effectiveness of a reference set is determined by composition as well as size.

## 5 Results and Discussion

### 5.1 Baseline

The `MIRRORING` algorithm was run against the data provided in the Kaggle dataset. Results are plotted in Fig. 3 for different sized reference sets ($k$), and using oBIFs, colour or their combination. Each point of the plotted curves is performance assessed over $10^4$ trials. Each trial used a different, randomly chosen, test image paired with its correct word bag and a randomly-chosen incorrect word bag. The reference set for each trial was randomly chosen and did not include the test image or the image for the random bag.

Inspection of the results reveals that performance is greater than chance performance (50%) for $k > 2$, rising with $k$ but plateauing around $k = 4096$. Colour performs better than pure oBIFs for all $k$, and the hybrid approach performs the same as colour alone. The maximum performance, achieved at $k = 4096$ is 73.7% using colour histograms. These results contrast with those

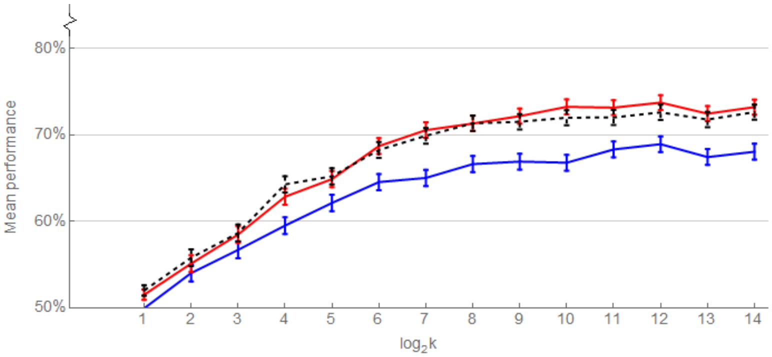


Fig. 3: Performance (vertical) of the `MIRRORING` algorithm as a function of reference set size $k$ (horizontal). Error bars are 95% confidence intervals. Blue line uses oBIF histograms only, red line uses colour histograms only and the black dashed line is a hybrid combination method.

in [5] where results plateaued around 660 categories, attaining a score of 77% for colour, 81% for oBIFs and 84% for their combination.

Since the Kaggle competition involved foul play (cheating using the Hungarian algorithm) [6], there is no other baseline against which to compare our work except that of random chance of picking the correct bag (50%).

### 5.2 Variant Word Bag Distances

The distances between two bags of words should be based on the distances between pairs of words, one from each bag, but there are several ways to do this. A plausible scheme is to use one function ($f_{inner}$) to compute word-bag distances from word-word distances, and a second function ($f_{outer}$) to compute a bag-bag distance from the word-bag distances. We have experimented with using minimum, mean, median and maximum to be these two functions.

The results in Table 1 shows the average performance of the classification algorithm using different variants of the $f_{inner}$ and $f_{outer}$ functions and reference sets of $k = 32$ images. The table shows that the best performing methods are the "mean of minimums" or "mean of means". All other results in this paper (including the baseline results) were computed using the "mean of minimums" method.

### 5.3 Variant Word Weightings: Frequency

The frequency at which different words appear in a corpus varies over several orders of magnitude [25]. Plausibly, high frequency words may be more tightly bound to the image appearance than low frequency or vice-versa. We assessed this by replacing bags with a single word from the bag either randomly chosen, or the most frequent or least frequent.

For a $k = 32$ reference set, the performance using the full bags was 62.1%; using a single random word this dropped, as expected, to 56.1%. The scores using the most frequent (55.7%) and least frequent words (56.2%) were not statistically

| $f_{outer}$ \ $f_{inner}$ | Min | Mean | Median | Max |
|---|---|---|---|---|
| Min | 58.7% | 60.3% | 58.2% | 58.8% |
| Mean | **62.1%** | **62.1%** | 60.6% | 57.6% |
| Median | 60.4% | 61.7% | 60.1% | 57.5% |
| Max | 58.9% | 58.1% | 57.3% | 53.3% |

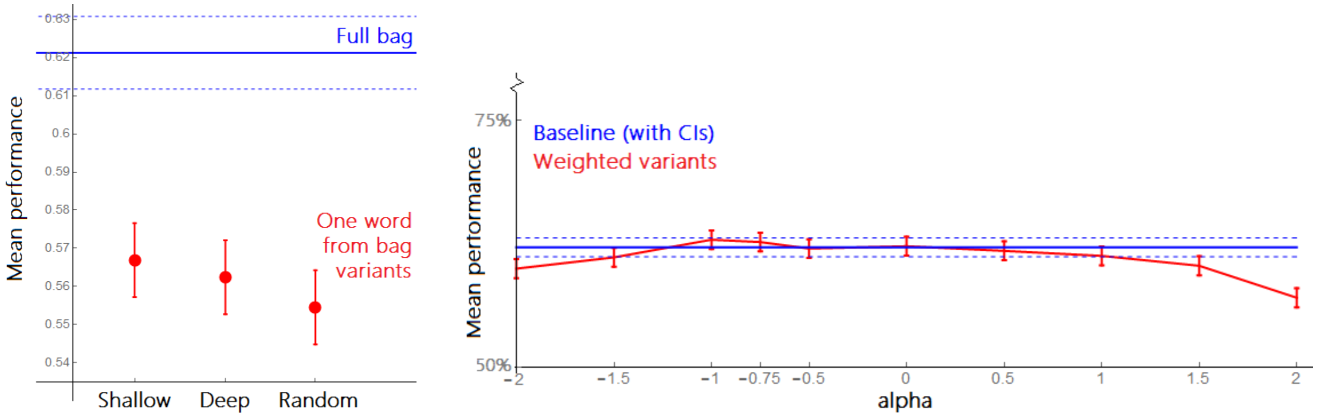
Table 1: Mean performance over $10^4$ trials, for $k = 32$ and varying word bag distance methods.

significantly different from the random word, so we conclude that it is not useful to weight word distances according to frequency.

### 5.4 Variant Word Weightings: Semantic Depth

Semantic depth measures the specificness of words. Low depth words are broad categories (eg: animal), deeper words are more specific (eg: squirrel). Like frequency, semantic depth is a plausible feature to use to weight words in distance calculations [26]. We measured semantic depth using path distances from the word 'entity' in the WordNet hyponym/hypernym lattice [27]. We first assessed semantic depth using the same random single word method used for frequency.

For a $k = 32$ reference set, and the same baseline full bag performance of 62.1%, using a single random word, this dropped to 55.5%. Then instead using the semantically shallowest word we achieved an increase to 56.7%, while using the deepest word a value of 56.2%. The results shown in Fig. 4(a) support this possible slight improvement for shallow words compared to random ones.



(a) (b)

Fig. 4: (a) Comparing a bag against (i) the semantically shallowest, (ii) semantically deepest, or (iii) a random word in the second bag each time. (b) Varying word weightings by $depth^{\alpha}$. Baseline performance for the same trials but with no weighting shown by straight line, with 95% confidence intervals in dashed blue.

We further investigated this by using full bag distances, but weighting words according to $depth^{\alpha}$ for $\alpha \in [-2, 2]$. The results shown in Fig. 4(b) show the best effect for $\alpha = -1$, which corresponds to giving greater weight to shallower words.

### 5.5 Optimal Subsets

We assessed whether all sets of $k$ reference images gave equal performance or whether superior sets could be found. The distribution of performance scores for random $k = 32$ reference sets was found to be approximately normally distributed with $\mu = 62.0\%$ and $\sigma = 1.91\%$, greater than the expected standard variation from $10^4$ trials of 0.49%, indicating that there is performance variation which is dependant on subset choice. So we may be able to improve performance by finding effective subsets. We "hunt" for an effective subset starting from a random subset by repeatedly swapping in new random elements and seeing if performance improves. We perform 256 swaps at each different $k$.

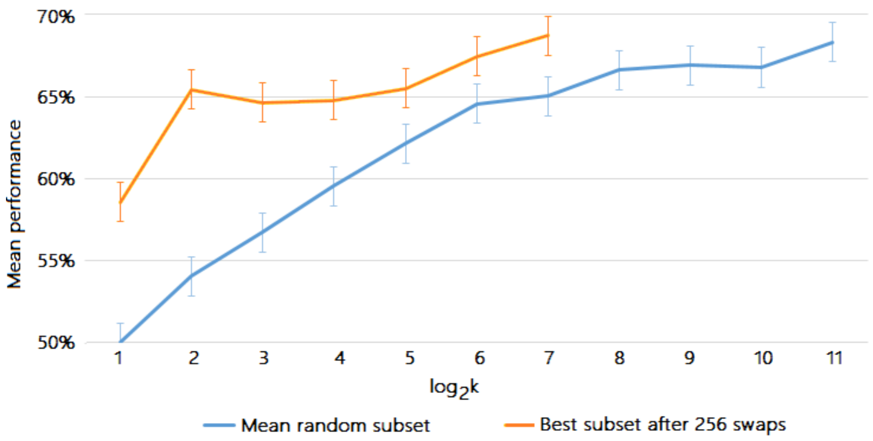


Fig. 5: Performance for random reference sets (blue) compared to effective reference sets discovered by hunting (orange).

Fig. 5 shows that we were able to achieve considerably improved performance by using effective rather than random reference sets. Though the margin between them appears to decrease with increasing $k$.

## 6 Conclusion

We have demonstrated that the `MIRRORING` algorithm achieves above chance performance for the Kaggle dataset. Compared to the demonstration of `MIRRORING` in [5], our experiment used:

1. Single images rather than sets of 50 images.
2. Multiple labels per image rather than single labels per set of 50 images.

Similarly to [5], we observe increasing performance with the number of reference sets, with an eventual plateau. In contrast to [5], we observed better performance for colour than for oBIF or hybrid. This may be due to the labels having been generated freely in response to the images rather than the images being assembled for a particular label.

In other experiments we determined that “mean of minimums” was the best way to compute bag-bag distances from word-word distances, that frequency weighting was ineffective, that semantically shallow words are slightly more bound to image appearance than deep, and that effective reference sets can be found that are better than random.

Overall, these results show that it is possible to use the Appearance Hypothesis to correctly select a labelling for an image containing *multiple* classes rather than just one and using only single reference images rather than larger sets.

Relating to the manifest verification problem, this approach shows it is possible to verify whether a bag of words labels a complex scene without having specific classifiers for each and without necessarily having any examples of any individual object. By equating a bag of words to a list of Harmonised Codes within a manifest file, and an X-ray scan to a complex image scene, we have built firm grounding for being able to quantify the labelling match of a manifest file for a cargo scan and thus notice mismatched labelling as is required by the original practical scenario.

## References

1. Thomas W. Rogers, J. Ollier, Edward J. Morton, and Lewis D. Griffin. Reduction of Wobble Artefacts in Images from Mobile Transmission X-ray Vehicle Scanners. *IEEE Imaging Systems and Techniques*, 2014.
2. Nicolas Jaccard, Thomas W. Rogers, and Lewis D. Griffin. Automated detection of cars in transmission X-ray images of freight containers. *IEEE Advanced Video and Signal Based Surveillance*, 2014.
3. Jarosaw Tuszynski, Justin T. Briggs, and John Kaufhold. A method for automatic manifest verification of container cargo using radiography images. *Journal of Transportation Security*, 6(4):339–356, July 2013.
4. Foreign Trade On-Line (FTOL). Harmonized System Codes, 2014.
5. Lewis D. Griffin, M. Husni Wahab, and Andrew J. Newell. Distributional learning of appearance. *PLOS ONE*, 8(2):e58074, January 2013.
6. Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in Representation Learning: A Report on Three Machine Learning Contests. In *International Conference on Neural Information Processing (3)*, pages 117–124. 2013.
7. Kaggle.com. Challenges in Representation Learning: Multi-modal Learning, 2013.
8. Grigorios Tsoumakas and Ioannis Katakis. Multi-Label Classification. *International Journal of Data Warehousing and Mining*, 3(3):1–13, January 2007.
9. Oded Maron and Tomás Lozano-Pérez. A Framework for Multiple Instance Learning. In *Advances in Neural Information Processing Systems (10)*, pages 570–576, Cambridge, MA, 1998. MIT Press.
10. Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104, September 2012.

11. Krishnakumar Balasubramanian and Guy Lebanon. The Landmark Selection Method for Multiple Output Prediction. In John Langford and Joelle Pineau, editors, *International Conference on Machine Learning (ICML)*, ICML '12, pages 983–990, New York, NY, USA, July 2012. Omnipress.
12. Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Journal of Machine Learning*, 85(3):333–359, June 2011.
13. Min-Ling Zhang and Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Journal of Pattern Recognition*, 40(7):2038–2048, July 2007.
14. Grigorios Tsoumakas and Ioannis Vlahavas. Random k-Labelsets: An Ensemble Method for Multilabel Classification. In *European Conference on Machine Learning (ECML)*, pages 406–417. Springer-Verlag, 2007.
15. Max M. Louwerse. Symbol Interdependency in Symbolic and Embodied Cognition. *Topics in Cognitive Science*, 3(2):273–302, April 2011.
16. Magnus Sahlgren. The Distributional Hypothesis. From context to meaning: Distributional models of the lexicon in linguistics and cognitive science. *Rivista di Linguistica (Italian Journal of Linguistics, special iss.)*, 20(1), 2008.
17. Douglas L. T. Rohde, Laura M. Gonnerman, and David C. Plaut. An improved model of semantic similarity based on lexical co-occurence. *Communications of the ACM*, 8:627–633, 2006.
18. Oxford University Computing Services pp. the BNC Consortium. The British National Corpus, version 3 (BNC XML Edition), 2007.
19. David G. Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1150–1157 (2). IEEE, 1999.
20. Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE, 2005.
21. Lewis D. Griffin, Martin Lillholm, Michael Crosier, and Justus van Sande. Basic Image Features (BIFs) Arising from Approximate Symmetry Type. In *Scale Space and Variational Methods in Computer Vision (SSVM)*, pages 343–355, Voss, Norway, 2009.
22. Michael Crosier and Lewis D Griffin. Using Basic Image Features for Texture Classification. *International Journal of Computer Vision*, 88(3):447–460, January 2010.
23. Brent Berlin and Paul Kay. *Basic Color Terms: their Universality and Evolution.* University of California Press, Berkeley and Los Angeles, 1969.
24. Lewis D. Griffin. Optimality of the basic colour categories for classification. *Journal of the Royal Society: Interface*, 3(6):71–85, February 2006.
25. George Kingsley Zipf. The psycho-biology of language, an introduction to dynamic philology, 1935.
26. Alistair Kennedy and Stan Szpakowicz. Disambiguating Hypernym Relations for Roget's Thesaurus. In Václav Matoušek and Pavel Mautner, editors, *Lecture Notes in Computer Science (4629)*, pages 66–75. Citeseer, Springer Berlin Heidelberg, 2007.
27. Princeton University Cognitive Science Laboratory. WordNet 3.0, 2006.