

Image annotation with high-level words using generalised attributes

QIANYU ZHANG

Abstract

The emergence of social media sharing communities has led to the need for accurate context-based image retrieval methods, which can be accomplished by an automatic annotation system. The ability to annotate high-level context-based words is necessary for such a system; however, it is not well researched due to the inherent difficulty caused by the semantic gap. This thesis identifies a set of high-level words that are frequently used by users to describe images, with a baseline system constructed using linear classifiers. The concept of ‘generalised attributes’ is then proposed and used to improve prediction by bridging the gap between image features and high-level words. The generalised attribute ‘anchor feature’ proposed, together with the ‘total distance’ feature selection method, leads to optimal performance. The resulting system yields not only an improvement in statistical accuracy over the baseline, but also a huge improvement in the quality and relevance of images retrieved in image retrieval and tags predicted in tag recommendation.

Introduction

As photo-sharing web communities like Flickr and Instagram become increasingly popular, the number of images and their associated tags increases at an incredible rate. Flickr alone has 1.6 million images uploaded per day. However, the imprecision and noisiness in user labelling makes accurate context-based image retrieval and management difficult. Given the large numbers of images, expert manual labelling is infeasible. Automatic tagging or automatic image annotation (AIA) (Brahmi and Ziou 2004), which predicts the list of tags associated with images, has become an important research topic. ‘Tag ranking’ (Liu et al. 2009), which ranks the tags according to their relevance, is also becoming an active research focus. The combination of the two techniques yields a ranked list of tags relevant to images, effectively facilitating image retrieval and image management.

Out of the large research efforts devoted to AIA, little research has considered the prediction of abstract keywords, which are hard to detect directly from low-level visual features due to the ‘semantic gap’ (Smeulders et al. 2000). However high-level words are often more desirable to users; therefore, the ability to predict these words plays an important role in achieving practical and accurate context-based image retrieval. This work targets such high-level words and proposes methods to accomplish their detection and prediction.

This work facilitates the following:

- Efficient and accurate image retrieval. As the ranked list of words we predict are sets of high-level words that are often used to describe images by users, this leads to retrieval results that are closer to users’ real needs.
- Automatic analysis of the images and recommend tags for images. This would assist the uploading of images, which is a usual behaviour in daily lives.
- Increased ease of photo management; supports easy browsing and organising of images based on their visual contexts.

Background

High-level words

Words can be partitioned into visualisable versus non-visualisable words. Examples of visualisable versus non-visualisable tags are presented in Figure 1. Visualisable words are related to image content and are objective: they tend to refer to specific objects. On the other hand, non-visualisable words are related to image context and are subjective, such as words describing the aesthetic, sentiment or scene of an image.



Figure 1: Examples of visualisable and non-visualisable words, coloured in red and blue respectively.

Source: Author’s analysis and NUS-WIDE Dataset.

Image annotation

The rapid increase in the volume of images being uploaded motivates research into efficient image retrieval and organisation techniques. Images are often annotated for efficient retrieval in current context-based image retrieval systems, known as annotation-based image retrieval (ABIR). Manual labelling is extremely time consuming and is subject to the users' subjective judgement. Automatic image annotation (AIA) can be employed instead. Current AIA approaches include discriminative methods (such as classification-based methods) and generative methods (such as methods based on probabilistic models and graph models).

Classification methods transform the image annotation problem into an image classification problem by treating images as data samples and tags as class labels. Each word might have several dictionary senses that are visually distinct, which is known as the 'visual polysemous' property. Therefore, image annotation can be transformed into a multi-class classification task, which can be solved using support vector machines (SVMs) or multi-label learning algorithms (Lu et al. 2009) and multi-instance learning algorithms (Zhou and Zhang 2006). The co-occurrence probability between regional image features and concepts can be estimated to produce annotations for images. This can be achieved using either the machine translation-based model or cross-media relevance model. The machine translation model (Duygulu et al. 2002) views the annotation keywords and features as two different languages that describe the same image, and translates between those two languages. On the other hand, the cross-media relevance model (Jeon et al. 2003) uses blobs to represent the semantic contents of images, where blobs are formed using discretised feature clusters. Graph models (Tong et al. 2006) treat every image and every keyword as a graph node, and the relations between them as edges; label information can then be propagated from labelled images to unlabelled images.

Tag ranking

In AIA, tags are annotated in random order, however, an unranked list of tags reveals no information about the relevancy and importance of the tags. Research (Liu et al. 2009) has shown that only 8 per cent of the most relevant tags are ranked first in the list of tags associated with Flickr images. Tag ranking can be achieved using probability density estimation followed by a random 'walk' over the tag similarity graph (Liu et al. 2009). A modified approach (Agrawal et al. 2011) segments the image and associates the objects to the tags in the tag list. Based on this approach, Kennedy et al. (2006) proposed a tag ranking algorithm based on tag clustering using a tag-pair weight matrix; further refinement can be achieved by performing a tag-pair semantic similarity extraction.

Image attributes

Apart from directly predicting words from images, the prediction of higher-level words requires more than just low-level image features. Higher-level features are introduced as image attributes that can be considered as the visual cues of parts of the image. Liu et al. (2009) introduce a list of attributes for the purpose of predicting aesthetics and allure. A generative probabilistic model can be used to learn discriminative features and estimate their distributions (Ferrari and Zisserman 2008). Images can also be described using attributes of the objects in the images (Farhadi et al. 2009), including semantic (shape, part and material) and discriminative attributes. Using discriminative features improves performance in attribute detection compared to using whole features. Research done on animal detection (Lampert et al. 2009) used a set of animal-specific attributes, such as the colour of the animal, whether it has fur, or if it appears with water. Common associating attribute terms for each object class are mined from Flickr image descriptions (Kulkarni et al. 2011), resulting in a list of 21 visual attributes consisting of colour, texture, shape, material descriptors, and general appearance.

Performance measures

For both image annotation and image retrieval, we have a set of ground truth labels that are the tags manually created by users who uploaded the image. For the image annotation task, every predicted tag is labelled either relevant or non-relevant to the image, according to its existence in the ground truth tag list. For the tag-based image retrieval task, every retrieved image is labelled either relevant or non-relevant to a target word, according to the existence of the target tag in the tag list of the retrieved image.

To evaluate the performance of our proposed image annotation system and tag-based image retrieval system, the *precision@n* measure is used.

Precision@n (P@n)

Precision defines the fraction of retrieved items that are relevant.

The precision for image retrieval tasks can be defined in the following way:

$$precision = \frac{|relevant\ images| \cap |retrieved\ images|}{|retrieved\ images|}$$

where *relevant images* are the images tagged with the target by users, and *retrieved images* being the images retrieved by the system.

The precision for tag recommendation tasks can be defined in the following way:

$$precision = \frac{|relevant\ tags| \cap |retrieved\ tags|}{|retrieved\ tags|}$$

with *relevant tags* being the ground truth tags labelled by users, and *retrieved tags* the tags retrieved by the system. We evaluate precision at a cut-off rank, such that only the top *n* retrieved results are used, defined as *precision@n*.

Construction and use of generalised attributes

Generalised attributes definition

We propose the concept of generalised attributes to bridge the semantic gap between image features and high-level tags. The generalised attributes exhibit the following characteristics:

1. Can represent any concept, concrete or abstract.
2. Defined computationally, an attribute can be defined in terms of image features, or in relation to cluster centroids. Attributes are defined for use of computation, and might not be directly interpretable to humans.

Visual score rank

We need to formalise the definition of abstract words: this can be achieved by calculating visualness scores, which measures the visualisability of words. Visualness of tags have been calculated using network inference methods (Xie and He 2013) and mixture models (Xu et al. 2013). Combining those results could produce a more reliable and robust measure for visualisability. We take the set of 3,018 tags considered in both studies. For each word, we rank the word according to its visualness score calculated in each method, then calculate the geometric mean of the ranks to get the combined score. We call this combined score Visual Score Rank (VSR). Note VSR is calculated using ranks instead of absolute visualness scores: the lower the VSR, the more visualisable a word is. Denoting a word by ω , and the rank of ω in Xie and He (2013) by r_1 and the rank in Xu et al. (2013) by r_2 , we calculate the VSR as follows:

$$VSR(\omega) = \sqrt{r_1(\omega) \times r_2(\omega)}$$

Target words and baseline

Target words were identified following the process below:

1. Illegal English words are filtered out using WordNet (Bird 2006).
2. The VSRs are calculated, words with high VSR are filtered out.
3. Parts of speech tags for words are obtained using WordNet: adjectives are considered as target candidates; for nouns we check its hypernym hierarchy; and nouns that represent physical entities are filtered out.
4. Words are ranked according to their occurrence frequency, the top 50 frequent terms are selected.

The prediction of such general high-level words has not been addressed in previous research, therefore we built a baseline system for comparison. For each target word, a linear support vector machine (SVM) is built using the Caffe image features and the associated tags. For each test image, the SVM output is transformed to probability of each target word using Platt scaling.

Clusters as generalised attributes

For abstract words, each word might have multiple meanings, for example the word ‘Asian’ might mean Asian food, Asian buildings, or Asian people. We cluster the training images, expecting each cluster specialises at one specific meaning of the target word. We first generate tag clusters: we represent images as tag vectors, and then cluster them using k-means clustering, thus varying numbers of clusters per target tag are tested. The Caffe features for images are then extracted and averaged to get the centroid (*k-means on tags*). We also generated clusters by performing k-means directly on Caffe image features (*k-means on image features*).

Anchor features

An anchor feature is a feature vector formed by the negated distances from an image to cluster centres. The distance is negated to represent the similarity between the image and each cluster centre. That is, for each image we build the anchor vector:

$$\overrightarrow{[-d(x_i, C_{jk})]_{j,k}}$$

where x_i is the centroid of cluster for the target word, iterates through all targets and iterates through the clusters for each target word.

Total distance feature selection

We use feature selection to select the clusters that are more relevant to each target word. We started feature selection using mutual information. We then proposed a novel feature selection scheme, which ranks the total distance from positive training images to each cluster centre for each target word. The clusters with minimum total distances are selected as features. Figure 2 illustrates this total distance feature selection method. Assume we want to select relevant clusters for ‘Asian’. Also consider a cluster from ‘traditional’ representing the concept ‘traditional clothing’. Those images should be visually similar with the images depicting Asian people wearing traditional clothing, leading to a smaller total distance comparing to, for example, the ‘snow mountain’ cluster from the word ‘cold’, which contains images that have very different visual features. So the cluster representing ‘traditional clothing’ will be selected, whereas the ‘snow mountain’ cluster will not.



Figure 2: Illustration of the total distance feature selection method.
Source: Author's analysis and NUS-WIDE Dataset.

Prediction

We train an SVM for each target word on those anchor features. For each testing image, anchor features are constructed and fed into the SVM for each target word. We then rank the images for each target word, or words for each target image, according to their SVM scores. We then output the ranked image list (for image retrieval tasks) or the ranked word list (for image annotation tasks).

Results and discussion

Image retrieval

Figure 3 shows the performance for image retrieval when the number of clusters varies. We observe that the feature selection using total distance outperforms feature selection using mutual information for all possible number of clusters. Furthermore, we observe that the performance is best when the number of clusters is around 35, which can be approximated by the square root number of clusters.

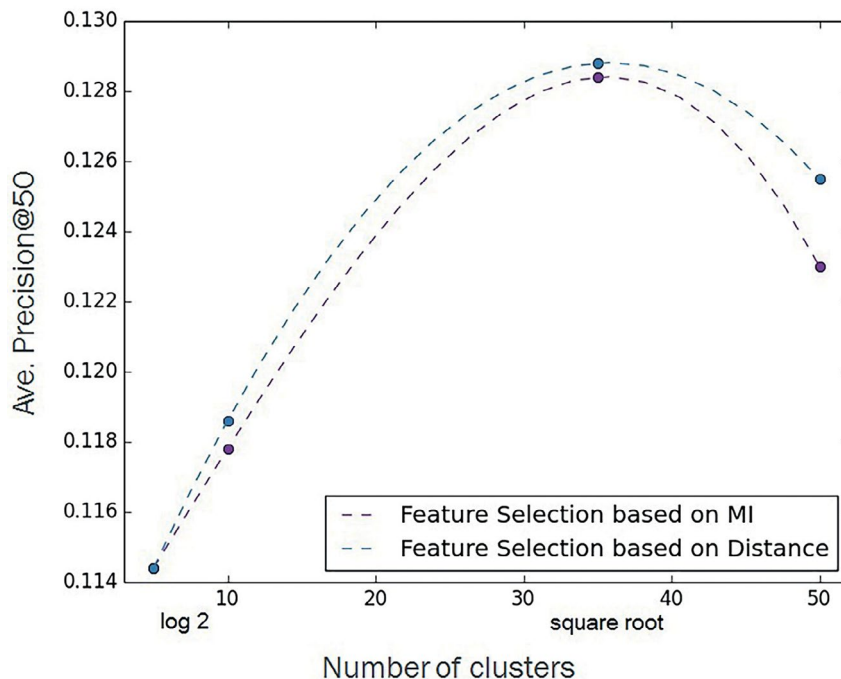


Figure 3: Image retrieval performance when varying the number of clusters, note the log 2 heuristic generates between 3 to 10 clusters, whereas the square root heuristic generates around 30 clusters.

Source: Author's analysis.

The performance statistics for image retrieval tasks using various methods are tabulated in Table 1. We observe that using total distance feature selection on image feature clusters yields the best performance.

Table 1: P@50 and uncertainty for the image retrieval task, averaged over all targets.

Clustering Method	k-means on tags	k-means on image features
Prediction		
Baseline	0.0358	0.0311
Anchor feature	0.1198	0.1288

Source: Author's analysis.

Figure 4 and Figure 5 present the best and worst image retrieval results using anchor features built on image feature clusters; features are selected using the total distance feature selection method. Figure 4 shows all images retrieved are relevant to the target tags, and the images retrieved illustrate different aspects of the tag. For example, the images retrieved for 'dawn' are from different scenes but all share a common visual characteristic. From Figure 5 we observe that even for tags that have a bad numerical performance, most of the retrieved images for the tags relate to the target. The bad numerical performance is likely caused by the imprecision and incompleteness of the ground truth.

Figure 6 compares the images retrieved for the same words 'perspective' and 'cold' using the baseline method and using our method: we again observe that most images retrieved by the baseline method are irrelevant to the targets, whereas the images retrieved using the anchor features are relevant yet cover a variety of scenarios.

The time taken to perform image retrieval for 50 target words from a pool of 5,000 images is less than 10 minutes. Out of the 10 minutes, approximately nine minutes are used to calculate the image features using the Caffe framework. This means that if the image features are extracted and stored, the image retrieval process for 50 tags takes less than one minute. This is 1.2 seconds per query, which suggests the possibility of using this system for live image retrieval.

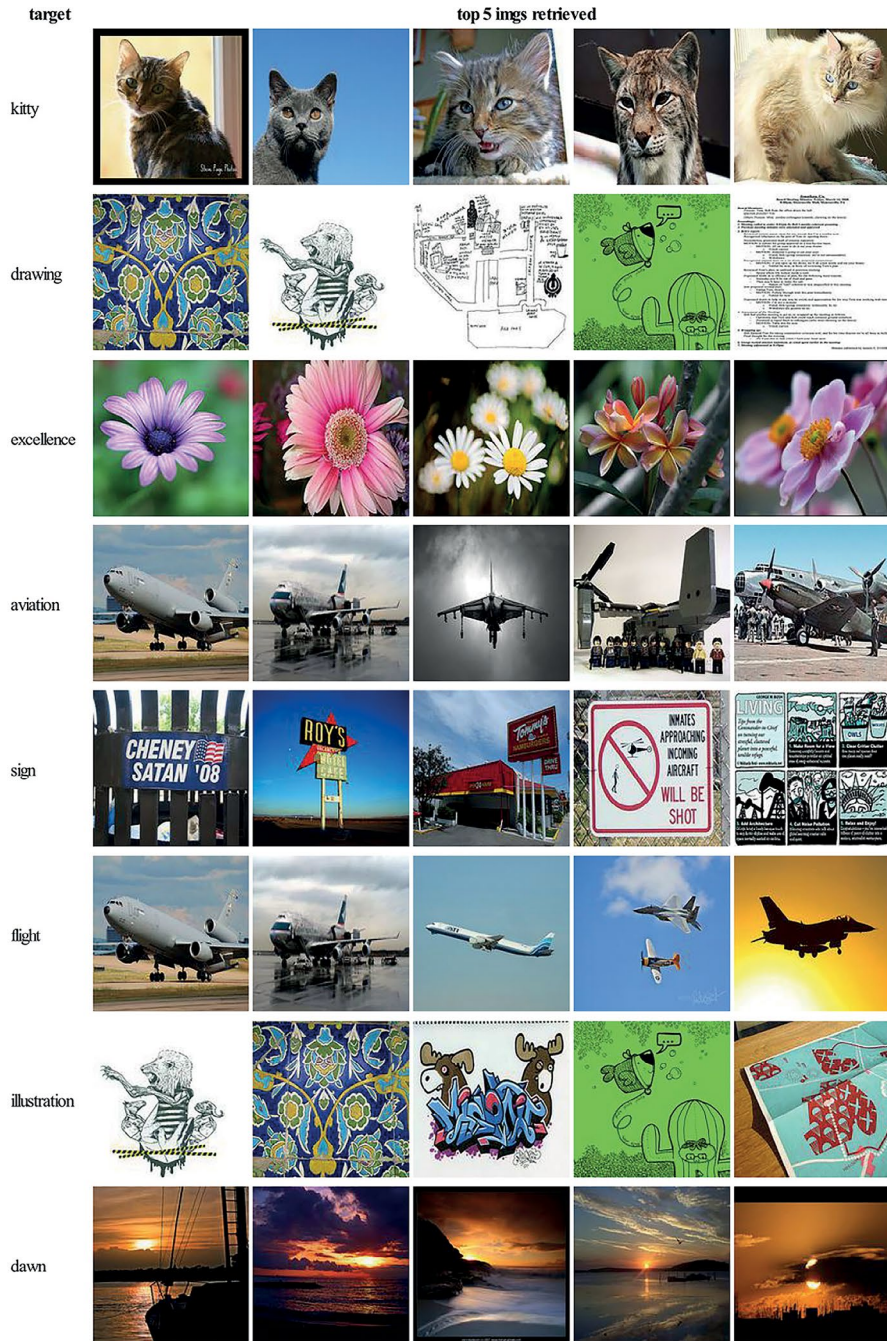


Figure 4: Top 8 tags and retrieved images using anchor features on square root number of image feature clusters with total distance feature selection.

Source: Author's analysis and NUS-WIDE Dataset.

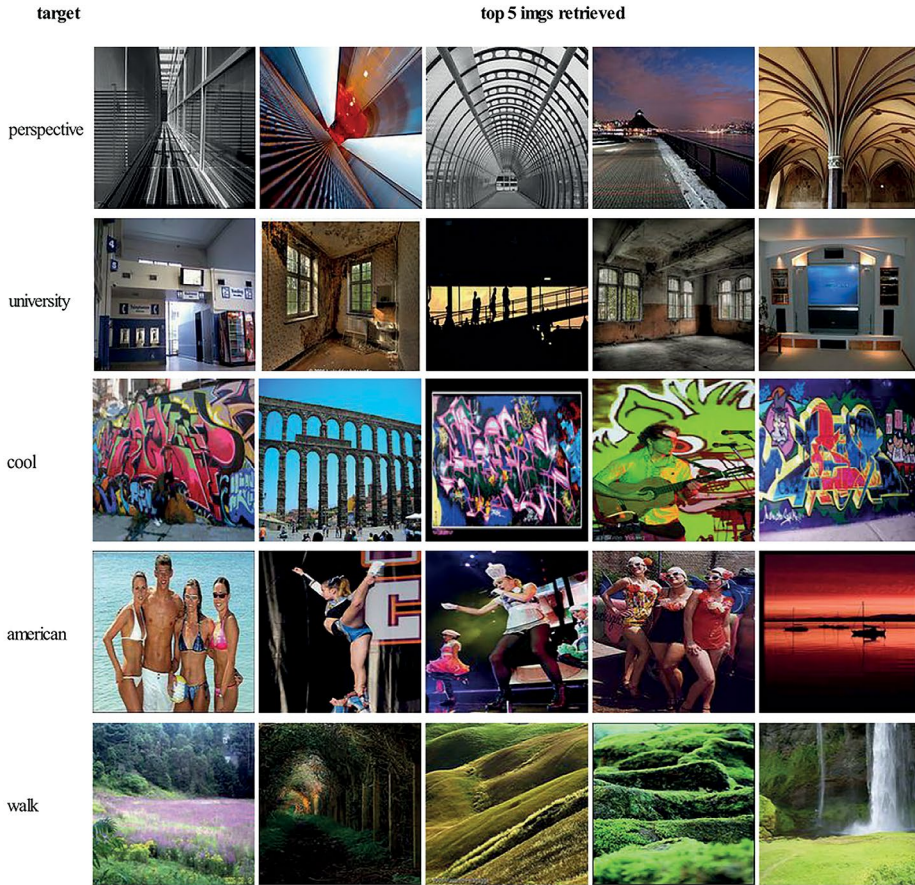


Figure 5: Bottom 5 tags and retrieved images using anchor features on square root number of image feature clusters with total distance feature selection.

Source: Author's analysis and NUS-WIDE Dataset.

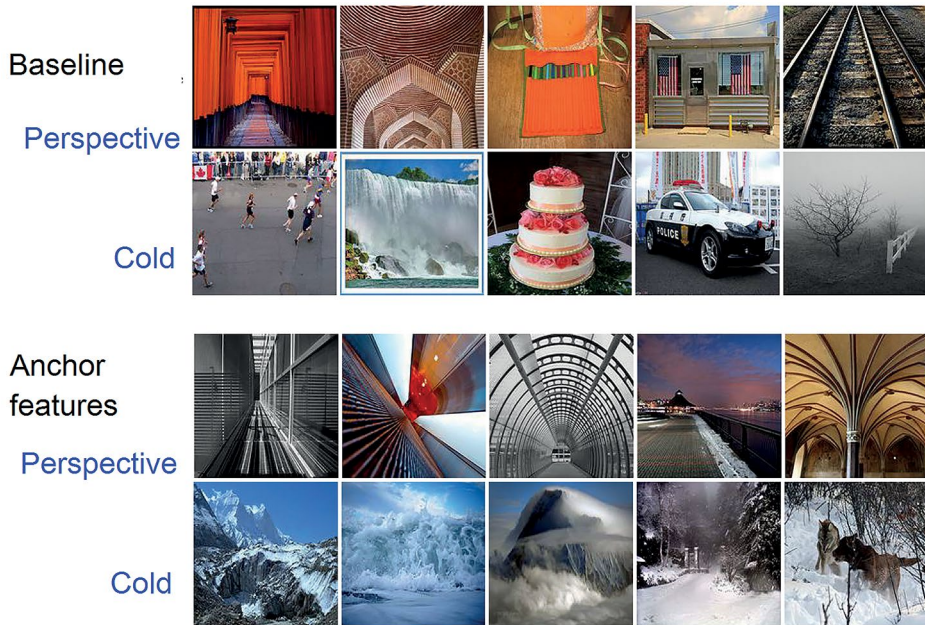


Figure 6: Performance comparison for the baseline method and the method using anchor features in the retrieving images for 'perspective' and 'cold'.

Source: Author's analysis and NUS-WIDE Dataset.

Tag recommendation

The TOP-n accuracy for tag prediction is tabulated in Table 2 and plotted in Figure 7.

Table 2: TOP-n accuracy and uncertainty for tag prediction using total distance feature selection, averaged over all targets.

Clustering Method	k-means on tags	k-means on image features
TOP-n		
1	0.168	0.177
3	0.111	0.116
5	0.090	0.092

Source: Author's analysis.

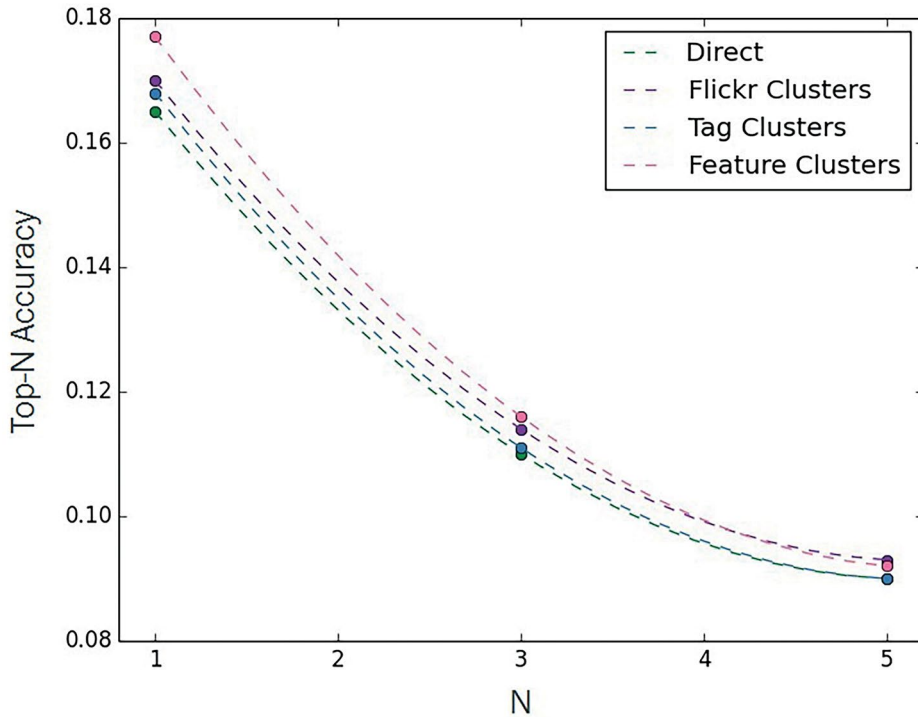


Figure 7: Performance comparison for tag prediction using various clustering methods, total distance feature selection on anchor features is used.

Source: Author's analysis.

We now predict tags from images and inspect the prediction results on 5,000 testing images. Figure 8 shows the top 20 images and predicted tags using a number of clusters with total distance feature selection. The predicted tags are mostly relevant to the images, especially the top-ranked tag for each image. This is very satisfactory performance as the number of target tags considered is only 50, so even a human labeller might not be able to label the image with five relevant tags from the target list. Figure 9 shows the bottom 20 images and predicted tags using a number of clusters. Again the predicted tags are mostly relevant to the images, which is much better than the direct approach.

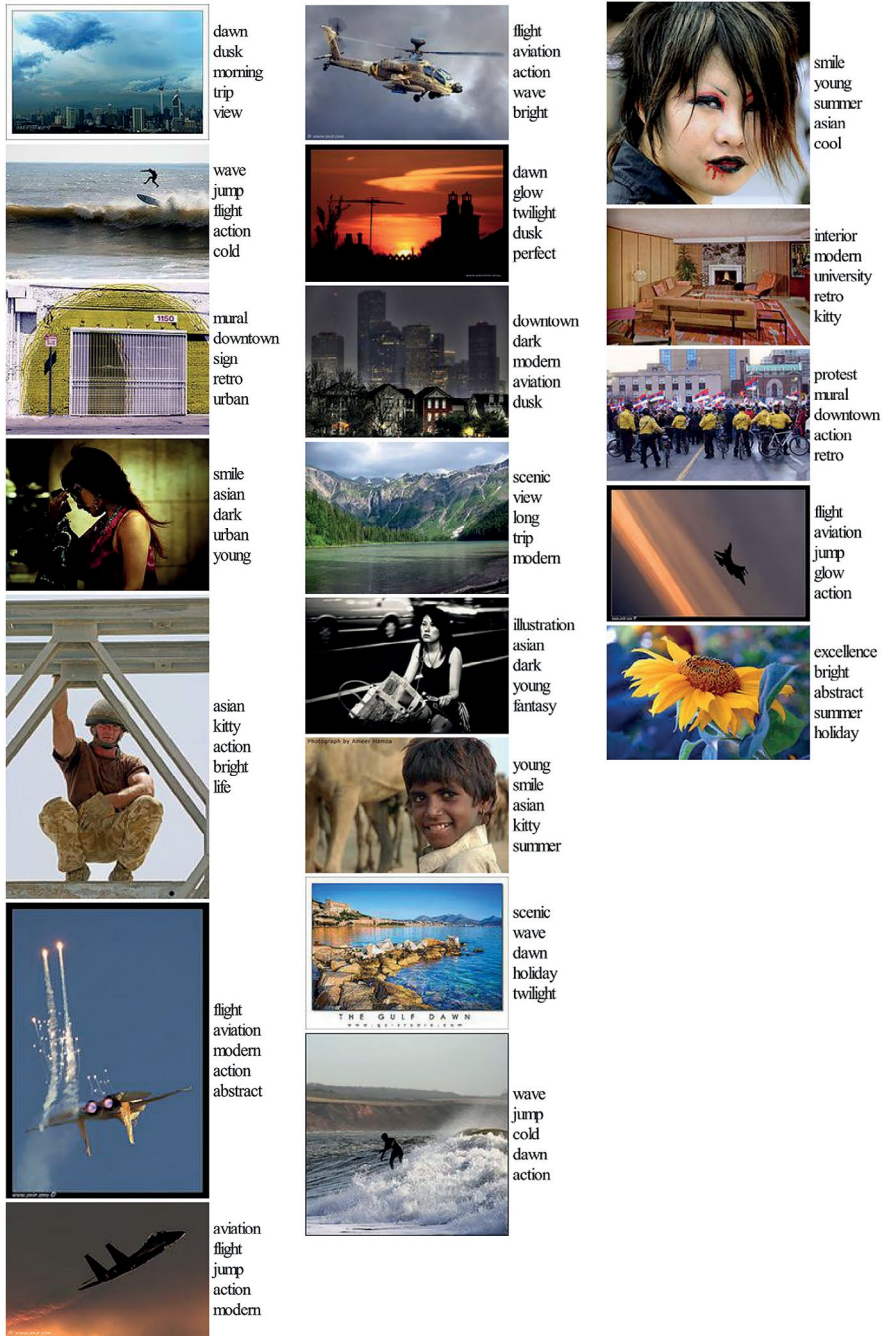


Figure 8: Top 20 images and predicted tags in the tag prediction task, prediction made using image feature clusters (log2 number of clusters) with total distance feature selection.

Source: Author's analysis and NUS-WIDE Dataset.

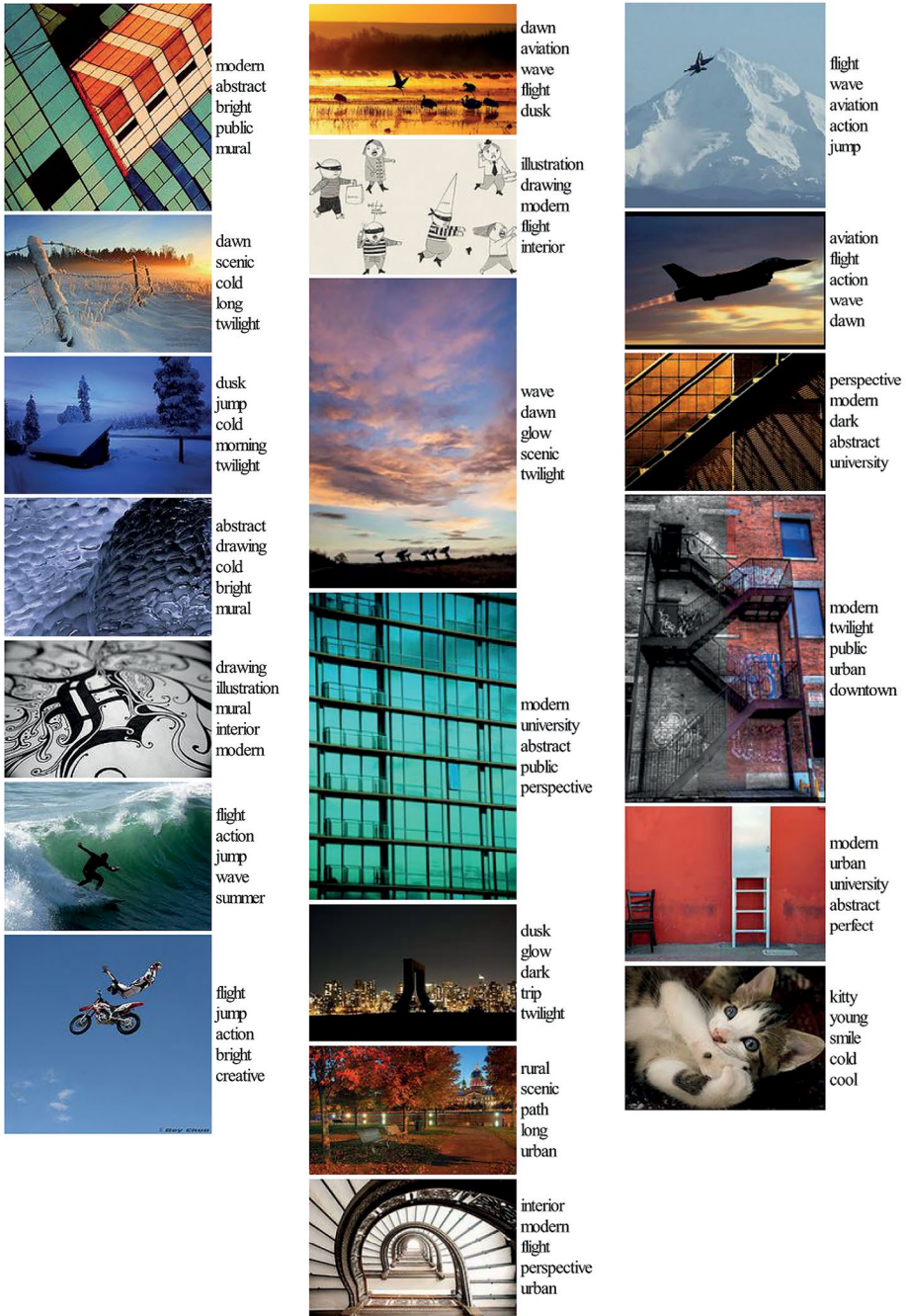


Figure 9: Bottom 20 images and predicted tags in the tag prediction task, prediction made using image feature clusters (log2 number of clusters) with total distance feature selection. Source: Author's analysis and NUS-WIDE Dataset.

Figure 10 presents three images with their predicted tags, predicted using the various methods discussed above. For the first image, there is only one ground truth tag; the tags predicted using the direct method do not seem to be very relevant to the image. The predictions made using our method generated four relevant tags. For the second image, there are five ground truth tags; however, apart from the tag ‘wave’, all other tags are not really relevant to the image. The baseline approach did not generate any relevant tags, whereas the predicted tags using anchor features on image feature clusters all seem applicable to the image. For the third image, there is no ground truth tag, and the direct prediction only predicts one relevant tag, ‘jump’. Our method predicts ‘jump’ and ranks it first. Moreover other predicted tags (‘action’, ‘young’) are also relevant.



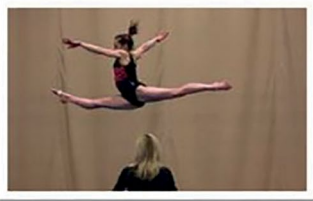
Ground truth	Predicted Tags		
	Direct	Flickr clusters	Image feature clusters
	rural	dawn	scenic
	path	twilight	dawn
	trip	wave	walk
	dusk	walk	twilight
	digital	scenic	wave
	twilight	aviation	dawn
	bright	cold	wave
	action	flight	cold
	dusk	abstract	flight
	wave	path	scenic
	sign	jump	jump
	jump	action	flight
	illustration	young	action
	flight	aviation	young
	downtown	bright	fantasy

Figure 10: Predicted tags for typical images using various methods. The incorrect tags are identified manually and highlighted in red.

Source: Author’s analysis and NUS-WIDE Dataset.

The time taken to recommend tags for 5,000 images using tags from the set of 50 target tags is less than 10 minutes. Approximately nine minutes were used to calculate the image features using the Caffe framework. If the image features

were extracted and stored, the image retrieval process for the 5,000 images would take less than one minute. This is 0.012 seconds per image, which suggests the possibility of using this system for live tag recommendation.

Conclusion

This thesis aimed to predict high-level words from images through the use of generalised attributes. Generalised attributes were proposed to bridge the semantic gap between image features and high-level words. This prediction of high-level words complements a wealth of previous research into the field of image tagging that has primarily focused on object identification. Such a system could enable accurate and practical context-based image retrieval and tag recommendation.

Words were identified as ‘high-level’ based on their visualness score and placement in the WordNet hierarchy. The 50 most frequent high-level words were kept as target tags. A baseline prediction system for these high-level words was constructed using linear classifiers trained on high-level image features, as no previous research had been conducted on such targets. Using the anchor features constructed on image clusters, together with the novel total distance feature selection method, we were able to improve the prediction performance for both image retrieval and tag recommendation tasks. In image retrieval tasks, poor accuracy was observed for certain tags. This is likely caused by incomplete ground truth in the form of non-comprehensive tagging rather than poor performance of the system.

These results demonstrate that it is possible to construct successful tag recommendation and image retrieval systems with high-level words. This meets real user needs for image querying and tag recommendations based on words representing more abstract concepts and ideas in addition to words which refer to objects.

Future work

Due to time constraints this thesis only selected 50 target words. However, with only 50 target words the number of relevant tags in the target list is too small to properly tag most images. Increasing the size of the target list would increase the usefulness and accuracy of the automatic tagging system.

Another possible future field of work is to employ new measurements for testing performance on testing sets with noisy and incomplete labelling. The current measures do not reflect the actual performance of the system in both image retrieval and tag prediction tasks.

Bibliography

- Agrawal, G., Chaudhary, R., and Singh, P. 2011. Relevancy tag ranking. Paper presented at 2nd International Conference on Computer and Communication Technology (ICCCCT), Allahabad, India., pp. 169–173. Institute of Electrical and Electronics Engineers (IEEE).
- Bird, S. 2006. Nltk: The natural language toolkit. *Proceedings of the COLING/ACL on Interactive presentation sessions*, pp. 69–72. Association for Computational Linguistics.
- Brahmi, D. and Ziou, D. 2004. Improving cbir systems by integrating semantic features. In *Proceedings of the first Canadian conference on Computer and Robot Vision*, 2004, pp. 233–240. IEEE.
- Chua, T., Tang, J., Hong, R., Li, H., Luo, Z., and Zheng, Y. 2009. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. Paper presented at ACM International Conference on Image and Video Retrieval. Greece. Jul. 8–10.
- Duygulu, P., Barnard, K., de Freitas, J. F., and Forsyth, D. A. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Computer Vision ECCV*, pp. 97–112. Springer.
- Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. 2009. Describing objects by their attributes. Paper presented at IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, Florida, pp. 1778–1785. IEEE.
- Ferrari, V. and Zisserman, A. 2007. *Learning visual attributes*. In *Advances in Neural Information Processing Systems*, pp. 433–440. MIT Press.
- Jeon, J., Lavrenko, V., and Manmatha, R. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 119–126. Association for Computing Machinery (ACM).
- Kennedy, L.S., Chang, S.-F., and Kozintsev, I.V. 2006. To search or to label?: Predicting the performance of search-based automatic image classifiers. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pp. 249–258. ACM.
- Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., and Berg, T.L. 2011. Baby talk: Understanding and generating simple image descriptions. Paper presented at IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011), Colorado Springs, pp. 1601–1608. IEEE.

- Lampert, C.H., Nickisch, H., and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. Paper presented at IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, Florida, pp. 951–958. IEEE.
- Liu, D., Hua, X.-S., Yang, L., Wang, M., and Zhang, H.-J. 2009. Tag ranking. In *Proceedings of the 18th international conference on World wide web*, pp. 351–360. ACM.
- Lu, Z., Ip, H.H., and He, Q. 2009. Context-based multi-label image annotation. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, p. 30. ACM.
- Smeulders, A.W., Worring, M., Santini, S., Gupta, A., and Jain, R. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380.
- Tong, H., He, J., Li, M., Ma, W.-Y., Zhang, H.-J., and Zhang, C. 2006. Manifold-ranking-based keyword propagation for image retrieval. *EURASIP Journal on Applied Signal Processing*, 1–10.
- Xie, L. and He, X. 2013. Picture tags and world knowledge: learning tag relations from visual semantic sources. In *Proceedings of the 21st ACM international conference on Multimedia*, pp. 967–976. ACM.
- Xu, Z., Wang, X.-J., and Chen, C.W. 2013. Mining visualness. Paper presented at IEEE International Conference on Multimedia and Expo (ICME), San Jose, California, pp. 1–6.
- Zhou, Z.-H. and Zhang, M.-L. 2006. Multi-instance multi-label learning with application to scene classification. In *Advances in Neural Information Processing Systems*, pp. 1609–1616. MIT Press.

This text is taken from *The ANU Undergraduate Research Journal*,
Volume Seven, 2015, edited by Daniel McKay, published 2016 by ANU eView,
The Australian National University, Canberra, Australia.