# Improving Automatic Image Tagging using Temporal Tag Co-occurrence

Philip McParlane[1], Stewart Whiting, and Joemon Jose

The University of Glasgow,
Glasgow, G12 8QQ, UK
p.mcparlane.1@research.gla.ac.uk, s.whiting.1@research.gla.ac.uk,
Joemon.Jose@glasgow.ac.uk

**Abstract.** Existing automatic image annotation (AIA) systems that depend solely on low-level image features often produce poor results, particularly when annotating real-life collections. Tag co-occurrence has been shown to improve image annotation by identifying additional keywords associated with user-provided keywords. However, existing approaches have treated tag co-occurrence as a static measure over time, thereby ignoring the temporal trends of many tags. The temporal distribution of tags, however, caused by events, seasons and memes, etc, provides a strong source of evidence beyond keywords for AIA. In this paper we propose a temporal tag co-occurrence approach to improve AIA accuracy. By segmenting collection tags into multiple co-occurrence matrices, each covering an interval of time, we are able to give precedence to tags which not only co-occur each other, but also have temporal significance. We evaluate our approach on a real-life timestamped image collection from Flickr by performing experiments over a number of temporal interval sizes. Results show statistically significant improvements to annotation accuracy compared to a non-temporal co-occurrence baseline.

**Keywords:** Image Annotation, Tag Co-occurrence, Temporal

## 1 Introduction

With the amount of multimedia data rapidly increasing, it becomes important to organize this content effectively. To be able to facilitate efficient multimedia retrieval we must first categorize these objects with semantic features, such as keywords. However, unlike traditional text retrieval which can infer topics directly from the distributions of words in a document, multimedia objects provide little or no textual clues. Hence, content annotation with semantically related keywords is therefore necessary before indexing and retrieval can take place. The laborious nature of manual image annotation, however, combined with the need for effective large-scale image search has increased research in the field of automatic image annotation (AIA).

[1] This research was supported by the the European Community's FP7 Programme under grant agreements nr 288024 (LiMoSINe)

[2] For the remainder of this paper we refer to tags and keywords synonymously.

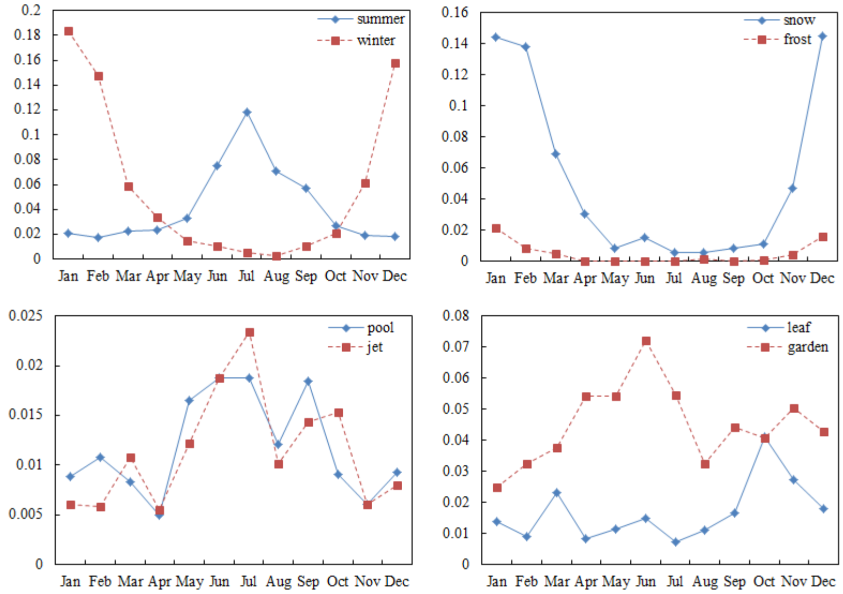
Current state-of-the-art AIA models, however, produce poor results, especially when tested on 'real-world' image collections [2]. Such collections are considered problematic often because of their noisiness, sparsity and diversity of image features. Bridging the semantic gap between low-level image features and high-level human concepts is still an unsolved research problem [23]. In any case, many fundamentally question if there even exists a correlation between these two levels [21]. Much research has focused on looking *beyond the pixel* to incorporate more robust evidence in the annotation process [20, 16]. We propose to explore beyond the visual contents of images in the annotation process by exploiting tag co-occurrence and temporality; by doing so we can avoid, to an extent, the problems associated with content-based image annotation.

Since the quality of AIA is very poor, a number of image sharing websites employ user tagging e.g. Flickr. However, the tagging process is either incomplete or often inaccurate. Automatic tagging techniques are often exploited to improve the quality of annotated tags. Tag co-occurrence has been used by existing tag recommendation [22] and AIA systems [16] to improve performance by discovering additional related tags. Tag co-occurrence for two keywords is defined as the number of documents in which both keywords co-exist; in the field of AIA, these documents are images. The motivation for exploiting tag co-occurrence is that keywords exist in a specific distribution which can be exploited. In the field of timestamped text analysis, a significant body of research has sought to exploit dynamic term distributions, most notably for Topic Detection & Tracking [1] and IR [26]. Analysis of user tags shows that tag co-occurrence is often linked with time. As such, two keywords which co-occur highly in June may not have the same relationship in December. Figure 1 shows example normalised tag distributions over time from a collated Flickr collection. Strong temporal distributions are seen for seasonal keywords such as `summer` and `winter`, which is expected. Further, tags related to weather cycles also observe a relationship with time. For example, `frost` and `snow` are most prominent during the winter months.

It may be argued, however, that only a restricted set of seasonal and weather related keywords will display such strong temporal distributions in image annotation but actually there are many tags with implicit temporality. For example, keywords such as `jet` and `pool` are seen to increase during the summer months i.e. typically when people go on vacation. Similarly, `garden` observes peaks during May through September which is expected due to the increase in outdoor activities in summer. By harnessing these temporal trends, we propose to improve tagging accuracy of an existing state-of-the-art model. Finally, research into tag co-occurrence has implications for a number of fields such as: tag recommendation systems as used on social bookmarking websites [5], query expansion [10], event detection [27] and personalised IR [3].

This paper is organised as follows. In Section 2 we present related work in the field of automatic image annotation and temporal IR. Section 3 describes the methodology behind our temporal co-occurrence based approach. In Section

**Fig. 1.** Tag distributions over time from our Flickr Collection



4 we discuss our experimental setup. Finally, Section 5 presents the results of our experiments and Section 6 concludes and discusses avenues for future work.

## 2 Related Work

The problem of image classification is often treated as a cross media modelling problem where we try to map low-level features in vector format to high-level textual concepts. Duygulu *et al.* [6] treated the problem of image annotation using a machine translation approach where images are segmented into small regions; keywords were then mapped based on a number of image features. In 2003, Joen *et al.* [9] adopted the cross lingual language model of Lavrenko *et al.* [13], Cross-Media Relevance Models (CMRM), to predict the probability of generating a word given blobs in an image in the training set. The model assumed regions in an image can be described by a small vocabulary of blobs, which were created from image features using clustering techniques. Lavrenko *et al.* [14] then proposed the Continuous-space Relevance Model (CRM) which generalised the previous CMRM to model highly dimensional continuous features without clustering and quantization. Bag of Visual Words (BOVW) has gained much interest in the field; Carneiro *et al.* [4] proposed a Gaussian mixture model using the bag of local features approach for class conditional dependencies.

More recently, Makadia *et al.* showed that all of the previously stated models could be outperformed by adopting a K-nearest neighbour approach trained on Gabor and HAAR image features [18]. In a similar experiment, Athanasakos *et al.* showed that these approaches were also out-performed by using an SVM approach trained on global features [2]. Further, they highlighted problems of the evaluation approaches of state-of-the-art annotation (SOTA) models, which are addressed in Section 4.2. We have chosen to implement the approach by Athanasakos *et al.* as a baseline, due to its simplicity and performance against other SOTAs.

Following research in text based IR [5, 10, 27, 3], tag co-occurrence has been used as a secondary source of evidence in tag recommendation systems [22] and image annotation models [17, 16]. Sigurbjornsson *et al.* proposed a tag recommendation strategy to support users annotating photos on Flickr [22]. The relationships between tags were exploited to suggest highly co-occurring tags. Sigurbjornsson *et al.* adopted two normalised measures for tag co-occurrence: the Jaccard (symmetric) and Asymmetric coefficients. Our approach follows this research by using these coefficients as a measure of keyword similarity. Llorente *et al.* incorporated tag co-occurrence in their annotation model which formulated the problem of image annotation as that of direct image retrieval [17]. Novelty was achieved by not only exploring the dependencies between words and their semantic context, but also between visual features and words.

Temporality has previously been studied and exploited in both information seeking and retrieval systems. Despite this, its implication on automatic image annotation has not yet been explored. Klieinberg *et al.* [12] developed a framework for modelling periodic bursts of keywords in a corpus with hierarchical structure using an infinite-state automaton. More recently, Leskovec *et al.* [15] performed a large-scale study of "memes" diffusing throughout news media as a result of temporal rhythms. As a result, a mathematical model was provided for analysing the temporal variation in the context of news. We propose to exploit these temporal trends of tags in a tag co-occurrence model.

## 3 Temporal Co-occurrence

In this section we present our temporal based co-occurrence approach for improving the effectiveness of tag suggestions made by an existing AIA model.

### 3.1 Problem Statement

Let $I = \{i_1, ..., i_m\}$ denote an image collection, where $m$ is the number of images in the image set. We denote $t$ as a tag and $T = \{t_1, ..., t_n\}$ our vocabulary, where $n$ is the number of keywords in our collection. We define $S(i_x, t_y)$ as a confidence score of matching tag $t_y$ to image $i_x$.

Every $i \in I$ has a time-stamp of when it was taken. We aim to cluster images based on time. We therefore define $\beta$ to be the number of time intervals in the year in which we wish to cluster images on. For example, $\beta = 3$ would group images into three, 122 ($\frac{366}{3}$) day, time intervals. We define $i_z \subset I$ where $i_z$ is a set of images taken between the start and end of time interval $z$, where $1 \leq z \leq \beta$.

Our approach improves image annotation by promoting the most highly co-occurring tags from our image classifier. For each subset of images taken within a given time interval, $i_z \in I$, we build a co-occurrence matrix $C_z$ mapping the number of images two tags co-occur in for the given time interval.

$$C_3 = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_n \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \dots & t_n \end{array} \\ \begin{bmatrix} 0 & 10 & \dots & 1 \\ 10 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix} \end{array} \tag{1}$$

where $C_3$ is the matrix constructed from images taken within the $3^{rd}$ interval *e.g.* tag $t_1$ occurs together with tag $t_2$ in 10 images. Tag co-occurrence measures, however, are actually normalised between 0 and 1, as explained in Section 3.3. We define $C_{overall}$ to be the co-occurrence matrix built from all images.

### 3.2 Content Based Annotation

Our proposed approach builds on top of a linear SVM based AIA approach. SVMs have been used for many years in text based information retrieval categorisation systems [24]. More recently, this methodology has been used in AIA systems and has been seen to outperform state-of-the-art annotation models [2]. Due to its performance against other baselines and simplicity in design, we will use this model as our baseline to improve upon.

We implement the $SVM_{light}$ model [11], which uses a linear kernel function, trained upon the MPEG-7 Global Edge Histogram (GEH) image feature [19]. This feature was seen to give greatest annotation accuracy in [2]. Our approach is to train $n$ classifiers in an one-versus-all scheme, where $n$ is the number of classes (tags). We use the normalised distance $-1 \leq d_{xy} \leq 1$ to the boundary plane as a measure of how trustworthy a tag $t_x$ is for a given image $i_y$. Therefore, we define $S(i_x, t_y) = d_{xy}$.

It could be argued that our approach could be improved by training $n(n-1)/2$ classifiers in the one-versus-one scheme where we train a SVM for every tag and every tag *combination*, thus retaining prior classification data. We argue, however, that this would quickly become computationally challenging as $n$ increases. For example, for our collection containing 270 tags, we would potentially have to train 36,315 SVMs. In a real-world collection where there are millions of keywords [22], this solution would become unscalable. Further, this method requires a heavily dense collection with each tag combination containing sufficient training data, which is not true in on-line collections [25].

### 3.3 Improving Annotation through Tag Co-occurrence

To improve annotations made by the SVM, we increase or decrease $S(i_x, t_y)$, using tag co-occurrence measures. $S(i_x, t_y)$ is therefore redefined as:

$$S(i_x, t_y) = \lambda \cdot P_{svm}(i_x, t_y) + (1 - \lambda) \cdot P_{cooc}(i_x, t_y) \tag{2}$$

where $P_{svm} = d_{xy}$. $P_{cooc}$ is the tag co-occurrence score for $t_y$ with the other SVM suggested tags. $\lambda$ is a parameter $(0 \leq \lambda \leq 1)$ which weights the amount of SVM and co-occurrence data we use for $S(i_x, t_y)$. $P_{cooc}(i_x, t_y)$ is as follows:

$$P_{cooc}(i_x, t_y) = \frac{\sum\limits_{t_w \in T_{svm} - t_y} C(t_w, t_y)}{|T_{svm}|} \tag{3}$$

where $T_{svm}$ is the set of tags suggested by the SVM (where $d \geq 0$), $|T_{svm}|$ is the number of tags suggested by SVM and $C(t_w, t_y)$ is the tag co-occurrence frequency between tag $t_w$ and $t_y$. Effectively, keywords in the SVM prediction set are promoted if they co-occur highly with the rest of the predictions, and demoted otherwise.

**Baseline** Our tag co-occurrence baseline takes normalised tag co-occurrence frequencies $C(t_w, t_y)$ from $C_{overall}$. Therefore, co-occurrence frequencies are static and taken from the entire collection, thus ignoring temporality.

**Temporal** Our temporal approach takes co-occurrence frequencies from the temporal interval in which the image was taken. e.g. if $\beta = 12$ (equivalent to 1 matrix per month) and an image is taken on the $15^{th}$ of March, co-occurrence scores are taken from $C_3$.

Using raw tag co-occurrence frequency is noisy, however, as the popularity of tags is not taken into account. This gives rise to weighting popular tags higher than less common keywords; we must first normalise these frequencies. We have decided to use two measures as chosen by previous work, namely the Jaccard and Asymmetric Measures: [22]:

$$J(t_i, t_j) = \frac{|t_i \cap t_j|}{|t_i \cup t_j|} \qquad\qquad P(t_j|t_i) = \frac{|t_i \cap t_j|}{|t_i|} \tag{4}$$

**Equation 4.** The Jaccard (left) and Asymmetric (right)

Both measures which are used to compute tag similarity and relatedness have an upper bound of 1 and a lower bound of 0. Previous work has stated that the Jaccard measure is more useful for identifying synonyms whereas the Asymmetric measure offers more diverse recommendations. We will compare the effectiveness of both measures in our work.

## 4 Experiments

Our experiments compare annotation accuracy made by three systems:

- **SVM (Contents)** The first system is the state-of-the-art (as defined in [2]) which annotates using SVM data only.
- **SVM$^{Cooc}$ (Contents + Co-occurrence)** Our baseline improves results from the SOTA by exploiting tag co-occurrence data.
- **SVM$^{TempCo}$ (Contents + Temporal Co-occurrence)** Our experimental approach improves on SVM$^{Cooc}$ by exploiting temporal information in the computation of tag co-occurrence measures.

### 4.1 Collection

We tested our approach on a collated real life image collection from Flickr[1]. Real life image collections pose problems for research as the tags are *inconsistent*, often *misspelt* and *sparse* (many tags are used in only one image). We therefore cleaned the collection to contain only tags which occurred in at least 40 images and where images contained at least 3 tags. We also filtered out tags which, when classified by WordNet [7], were not considered *nouns*. This would remove tags which were not suitable for AIA; for example, subjective (e.g. `so cute`, `nice`) and organisational tags (e.g. `me`, `avoid`). Once cleaned, the collection contained 12,985 images and 270 tags. Each image on average contained 4.07 tags.

### 4.2 Experimental Procedure and Settings

Our experiments are taken out in a two stage process. Initially, images are trained and tested on the keywords using a linear SVM based on the Global Edge Histogram feature, as described in Section 3.2. For each image, a list of keyword scores is returned, measuring the likelihood of a tag occurring in an image. Tag co-occurrence is then employed as a reweighing scheme by increasing or decreasing the given score for a tag, based on its co-occurrence with the other tags in the ground truth. After this reweighing stage, the tags with the highest scores are selected for annotation; the amount of tags selected is equal to that of the number of tags in the image's ground truth.

We introduce temporality by computing the tag co-occurrence measures in predefined intervals, whereas our baseline computes tag co-occurrence measures over the entire year i.e. 1 co-occurrence matrix. We varied our temporal interval size over a range of values from half a year to 2 days. Therefore, given a new image $i_x$, we select the co-occurrence measures from the co-occurrence matrix which is built upon images taken *in the same time interval* as image $i_x$. We compare results when using the following number of co-occurrence matrices: $\beta = 2, 6, 12, 18, 40, 52, 70,\ 90, 120$. For each interval size, we compare annotation accuracy between our 3 approaches using 10-fold cross validation over 10 iterations. For each iteration of the experiment we collate a *subset* of the overall collection which is *smoothed* and *normalised*. By *smoothed* and *normalised* we mean that most of the tags in the test collection contain approximately the same number of training images. Alternatively, popular keywords, such as `sky`, and unpopular keywords, such as `hammer`, are *not* selected for testing.

We have taken out this stage as using the whole collection would probably create an easier evaluation setting due to the following reasons. Firstly, by evaluating on the entire collection, popular keywords would more likely be selected for testing. Secondly, when annotating an image, the model would be more likely to select a more frequent tag. By normalising our collection we create a fairer evaluation test-bed where images are less likely to be annotated with tags based purely on their popularity. We therefore normalise our collection as is explained

[1] http://www.flickr.com/

in [2]. This stage is an important stage in our experiment, as it will reduce the perceived accuracy of our state-of-the-art, as the test collection is more "difficult" as the number of perceived "easy" keywords is reduced in the test collection.

For each iteration of the experiment, on average a subset of 1114 images were used for training and 124 used for testing. We tested using 100 keywords at each iteration, with each keyword containing at least 20 training images. In our experiments we compute *precision*, *recall* and the *number of words recalled.*

## 5 Results and Discussion

The following sections detail the results of our experiments showing the potential of temporal modelling in the annotation process and the conditions where accuracy is maximised.

### 5.1 Effects of Temporality

By exploiting temporality, we were able to achieve statistically significant improvements to annotation accuracy. Table 1 shows the results of our experiments comparing both normalization methods over all interval sizes. From Table 1 we can see that both coefficients give increases to recall, precision and number of words recalled when compared to our static baseline. Using the Jaccard coefficient produces marginally better results than when using the asymmetric co-occurrence measure. It may be noted that the measures appear somewhat low for a SOTA; this is a side effect of the collection normalisation as described in Section 4.2. In effect our model is annotating on a *difficult* subset of an already *difficult* real-life image collection, hence lower performance is expected.

Figure 2 illustrates the conditions where annotation accuracy is maximised. The scores in Figure 2 are taken as an average of recall, precision and number of words recalled over the baseline. The Jaccard coefficient produces best results when the interval window size is set to 6 days ($\beta = 70$); statistically significant improvements averaging 11.6% are observed. The Asymmetric coefficient produces best results using approximately the same interval size, i.e. 5 days ($\beta = 90$), achieving an 9.6% increase to annotation accuracy. By incorporating the temporal trends of tags in images, as seen in Figure 1, we are able to give precedence to temporally significant tags based on the time an image is taken, thus improving the annotation accuracy.

Using a large interval size, 183 days for example, has a slight detrimental effect on AIA accuracy however. This may be because temporal profiling barely exists at these levels. We believe it may have the opposite effect of adding noise to the co-occurrence measures. We therefore recommend that future temporal profiling of keywords should use a interval size of around 5 days.

Finally, $\lambda$ was trained giving a local maxima in annotation performance when $\lambda = 0.4$. Interestingly we achieve greatest accuracy when we use a higher weight of $P_{cooc}$ than $P_{svm}$ implying tag co-occurrence and temporality may be a more reliable source than image contents in the annotation process.

### 5.2 Tag Distributions

The following section gives real life examples of tags with high temporal relationships. Figure 3 shows the co-occurrence frequencies of temporarily significant keywords, `snow` and `winter`, with `tree` and `landscape` over 18 time intervals.

**Table 1.** Each column denotes the scores for each measure using the given number of intervals. On the left column, R, P and W stand for recall, precision and the number of words recalled respectively. Bolded columns denote the interval size which produced the largest average improvement over the baseline. Paired t-test statistical significance comparing our experimental approach against the baseline are denoted as * being $p < 0.05$, ** being $p < 0.01$ and *** being $p < 0.001$.

**Using the Jaccard Co-efficient**

| | **Intervals** | 2 | 6 | 12 | 18 | 40 | 52 | 70 | 90 | 120 |
|---|---|---|---|---|---|---|---|---|---|---|
| **R** | SVM | 0.0809 | 0.0782 | 0.0796 | 0.0769 | 0.0778 | 0.0732 | 0.0807 | 0.0726 | 0.0745 |
| | $\text{SVM}^{Cooc}$ | 0.0839 | 0.0789 | 0.0825 | 0.0824 | 0.0778 | 0.0755 | 0.0814 | 0.0757 | 0.0782 |
| | $\text{SVM}^{TempCo}$ | 0.0842 | 0.0816 | 0.0853 | 0.0836 | 0.0838** | 0.0813** | **0.0909*** ** | 0.0842** | 0.0855* |
| **P** | SVM | 0.0697 | 0.0619 | 0.0712 | 0.0598 | 0.0759 | 0.0596 | 0.0637 | 0.0691 | 0.0619 |
| | $\text{SVM}^{Cooc}$ | 0.0719 | 0.0588 | 0.0750 | 0.0637 | 0.0720 | 0.0610 | 0.0613 | 0.0709 | 0.0632 |
| | $\text{SVM}^{TempCo}$ | 0.0717 | 0.0625 | 0.0738 | 0.0640 | 0.0742 | 0.0649 | **0.0695*** ** | 0.0784* | 0.0718* |
| **W** | SVM | 20 | 19.5 | 20.3 | 18.9 | 20 | 18.4 | 18.1 | 19.3 | 18.5 |
| | $\text{SVM}^{Cooc}$ | 20.4 | 19.4 | 21 | 19.7 | 19.9 | 18.7 | 18.2 | 19.7 | 19.1 |
| | $\text{SVM}^{TempCo}$ | 20.2 | 19.8 | 21.3 | 20.3 | 21.1** | 20.3*** | **20.1*** ** | 21.5** | 20.7* |
| | +/- Over Baseline | -0.3% | +3.9% | +1.0% | +1.7% | +5.6% | +7.6% | **+11.9%** | +10.3% | +10.5% |

**Using the Asymmetric Co-efficient**

| | **Intervals** | 2 | 6 | 12 | 18 | 40 | 52 | 70 | 90 | 120 |
|---|---|---|---|---|---|---|---|---|---|---|
| **R** | SVM | 0.0809 | 0.0782 | 0.0796 | 0.0769 | 0.0778 | 0.0732 | 0.0807 | 0.0726 | 0.0745 |
| | $\text{SVM}^{Cooc}$ | 0.0849 | 0.0824 | 0.0855 | 0.0823 | 0.0786 | 0.0778 | 0.0862 | 0.0787 | 0.0795 |
| | $\text{SVM}^{TempCo}$ | 0.0837* | 0.0848 | 0.0886 | 0.0862 | 0.0834** | 0.0843** | 0.0909* | **0.0873** ** | 0.0851* |
| **P** | SVM | 0.0697 | 0.0619 | 0.0712 | 0.0598 | 0.0759 | 0.0596 | 0.0637 | 0.0691 | 0.0619 |
| | $\text{SVM}^{Cooc}$ | 0.0720 | 0.0628 | 0.0724 | 0.0630 | 0.0689 | 0.0589 | 0.0639 | 0.0722 | 0.0615 |
| | $\text{SVM}^{TempCo}$ | 0.0711* | 0.0628 | 0.0742 | 0.0662 | 0.0677 | 0.0644** | 0.0714** | **0.0786*** | 0.0676 |
| **W** | SVM | 20 | 19.5 | 20.3 | 18.9 | 20 | 18.4 | 18.1 | 19.3 | 18.5 |
| | $\text{SVM}^{Cooc}$ | 20.4 | 19.8 | 21.2 | 19.8 | 19.7 | 18.9 | 18.6 | 19.9 | 19.1 |
| | $\text{SVM}^{TempCo}$ | 20* | 19.9 | 21.5 | 20.4 | 20.7** | 20.5** | 20.5*** | **21.7** ** | 20.1 |
| | +/- Over Baseline | -1.6% | +1.1% | +2.5% | +4.3% | +3.1% | +8.7% | +9.1% | **+9.6%** | +7.4% |

**Table 2.** Jaccard and Asymmetric scores

| **Measure** | **Scores @ Interval** | | |
|---|---|---|---|
| *Time Interval* | *All* | *2* | *10* |
| J(tree, snow) | 0.09 | 0.10 | 0 |
| A(tree, snow) | 0.18 | 0.28 | 0 |
| J(landscape, winter) | 0.05 | 0.08 | 0 |
| A(landscape, winter) | 0.09 | 0.22 | 0 |

We can clearly observe the keywords' correlation with time. Both sets of keywords co-occur highly at the beginning and end of the year with almost no co-occurrence during time intervals 5 through 16. This produces different Jaccard and Asymmetric measures at different periods in the year. Table 2 compares these co-occurrence measures at different time intervals.

The temporal distribution shown in Figure 3 highlights that keyword co-occurrence measures should consider time in these calculations. In our example, `snow` only exists along side images of `trees` in images during the winter months.

Our baseline which ignores temporal profiling of tag co-occurrences can be represented by the *entire* column of Table 2. Columns *2* and *10* show the co-efficient scores for the given time intervals only. These intervals were chosen as

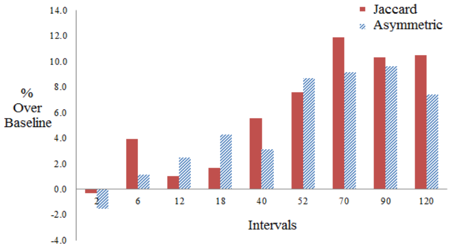


**Fig. 2.** Co-occurrence measures

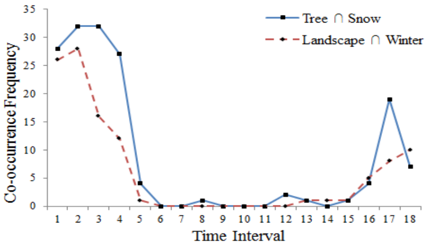


**Fig. 3.** Tag Co-occurrence distributions

they are the most divergent coefficient scores for the given keywords over the year. The coefficients in time *interval 2* are 73% higher, on average, than those taken over the whole year. We believe this is logical as the keywords, `tree` with `snow` and `landscape` with `winter`, co-occur most frequently in this time interval. Similarly, all the coefficients compute as 0 in *interval 10* as the keywords never co-occur in this time period. We believe this is sensible: if two keywords never co-occur in a given time period, they should produce a co-occurrence coefficient of 0 regardless of if they co-occur in other time intervals. By ignoring this noise and placing higher precedence to *temporally significant* keywords, we are able to achieve improvements to AIA accuracy.

## 6 Conclusion and Future Work

Accurate automatic image annotation is highly desired to be able to build effective multimedia retrieval systems. In this work we present a novel temporal based tag co-occurrence technique for the improvement of a state-of-the-art SVM based automatic image annotation model. Results from our experiments show that by exploiting temporal tag co-occurrences, we can produce statistically significant improvements to AIA accuracy.

This paper argues that static measures of normalised tag co-occurrence as used by previous methods are insufficient and that keywords co-occur in a non linear temporal distribution which can be exploited. We achieve this by constructing a number of co-occurrence matrices, one for each predefined interval, instead of building a co-occurrence matrix over the entire year. We further experiment by changing the interval sizes used to construct the co-occurrence matrices. We conclude that best results are achieved when the size of the temporal window is set to 5 or 6 days. Future work will look at extending our exploitation of temporal tag co-occurrence for AIA by incorporating more sophisticated techniques from temporal text-based IR systems.

## References

1. J. Allan. Topic detection and tracking. pages 1–16, Norwell, MA, USA, 2002. Kluwer Academic Publishers.
2. K. Athanasakos, V. Stathopoulos, and J. M. Jose. A framework for evaluating automatic image annotation algorithms. In *ECIR '10 Milton Keynes, UK*, 2010.

3. A. Byde and S. Cayzer. Personalized tag recommendations via tagging and content-based similarity metrics. *New York*, (2), 2007.
4. G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions*, 29, 2007.
5. C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic grounding of tag relatedness in social bookmarking systems. ISWC '08, pages 615–631, Berlin, 2008.
6. P. Duygulu, K. Barnard, J. De Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *ECCV '02*.
7. C. Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, illustrated edition edition, May 1998.
8. P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 1901.
9. J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. SIGIR '03, pages 119–126, NY, USA.
10. S. Jin, H. Lin, and S. Su. Query expansion based on folksonomy tag co-occurrence analysis. pages 300 –305, aug. 2009.
11. T. Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.
12. J. Kleinberg. Bursty and hierarchical structure in streams. KDD '02, pages 91–101, New York, NY, USA, 2002. ACM.
13. V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. SIGIR '02, pages 175–182, New York, NY, USA, 2002. ACM.
14. V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *IN NIPS*. MIT Press, 2003.
15. J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. KDD '09, pages 497–506, New York, NY, USA, 2009. ACM.
16. W. Li and M. Sun. Automatic image annotation based on wordnet and hierarchical ensembles. CICLing'06, pages 417–428, Berlin, Heidelberg, 2006. Springer-Verlag.
17. A. Llorente, R. Manmatha, and S. Rüger. Image retrieval using markov random fields and global image features. CIVR '10, pages 243–250, NY, USA, 2010. ACM.
18. A. Makadia, V. Pavlovic, and S. Kumar. Baselines for image annotation. *Int. J. Comput. Vision*, 90(1):88–105, Oct. 2010.
19. B. S. Manjunath. *Introduction to MPEG-7, Multimedia Content Description Interface*. John Wiley and Sons, Ltd., Jun 2002.
20. F. Monaghan and D. O'Sullivan. Leveraging ontologies, context and social networks to automate photo annotation. SAMT'07, pages 252–255. Springer-Verlag.
21. S. Santini, A. Gupta, and R. Jain. Emergent semantics through interaction in image databases. 13(3):337 –351, 6 2001.
22. B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. WWW '08, pages 327–336, New York, NY, USA, 2008. ACM.
23. A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. 22(12):1349 –1380, dec 2000.
24. V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.
25. K. Q. Weinberger, M. Slaney, and R. Van Zwol. Resolving tag ambiguity. MM '08, pages 111–120, New York, NY, USA, 2008. ACM.
26. S. Whiting, Y. Moshfeghi, and J. M. Jose. Exploring term temporality for pseudo-relevance feedback. SIGIR '11, pages 1245–1246, New York, NY, USA, 2011. ACM.
27. J. Yao, B. Cui, Y. Huang, and Y. Zhou. Data engineering (icde), 2010 ieee 26th international conference on. pages 780 –783, march 2010.