Improving Content-Based Image Retrieval by Identifying Least and Most Correlated Visual Words

Leszek Kaliciak[1], Dawei Song[2,3], Nirmalie Wiratunga[1], Jeff Pan[4]
[1]The Robert Gordon University, Aberdeen, UK
[2]Tianjin University, Tianjin, China
[3]The Open University, Milton Keynes, UK
[4]Aberdeen University, Aberdeen, UK
{l.kaliciak,n.wiratunga}@rgu.ac.uk; Dawei.Song@open.ac.uk; jeff.z.pan@abdn.ac.uk

Abstract. In this paper, we propose a model for direct incorporation of image content into a (short-term) user profile based on correlations between visual words and adaptation of the similarity measure. The relationships between visual words at different contextual levels are explored. We introduce and compare various notions of correlation, which in general we will refer to as image-level and proximity-based. The information about the most and the least correlated visual words can be exploited in order to adapt the similarity measure. The evaluation, preceding an experiment involving real users (future work), is performed within the Pseudo Relevance Feedback framework. We test our new method on three large data collections, namely MIRFlickr, ImageCLEF, and a collection from British National Geological Survey (BGS). The proposed model is computationally cheap and scalable to large image collections.

Keywords: content-based image retrieval and representation, local features, correlation, pseudo relevance feedback, similarity measure

1 Introduction

Recently, content-based image retrieval (CBIR) based on local feature extraction has attracted a lot of attention. One of the widely used approaches is based on so-called "bag of visual words" or "bag of features" (BOF) [1]. This model was inspired by the "bag of words" (BOW) framework from text information retrieval. The BOW represents documents as orderless "bag" of terms containing some words from the dictionary. In CBIR, terms from the text retrieval correspond to groups of local image patches (called visual words). A BOF representation of an image is a histogram of the visual words' counts in the image. The BOF approach is a mid-level representation that helps to reduce semantic gap between human perception and machine representation of images.

The local features based on BOF disregard the information about correlations between visual words. However, when the vocabulary size (the number of clusters) is small, the BOF's coefficients tend to be highly correlated. Such correlations can be exploited in order to improve the BOF performance.

[1] Terms "bag of visual words" and "bag of features" will be used interchangeably in this paper.

Proximity-based correlations are often utilized to capture the spatial relative information between instances of visual words and enhance the visual representations. Here, we will utilize both proximity-based and image-level correlations to adapt the similarity measure and re-rank the top images returned in the first round retrieval. To the best of our knowledge, no systematic comparison has been conducted between image-level and proximity based notions of correlation in the context of query expansion in image retrieval.

Existing approaches (query expansion like frameworks) often modify the current query, which leads to the normalization of histograms. This may not be desirable, since the (mid-level) semantic meaning of bins may be lost and the representations may become less discriminative due to the varied complexity of images. Moreover, many researchers incorporate the $tf \cdot idf$ weighting scheme from text retrieval although some experiments suggest that even the most frequent visual words are important to the retrieval (see [1]). However, others ([13]) report performance improvement for $tf \cdot idf$ and thus the results are not conclusive. We believe that this may be domain specific. $tf \cdot idf$ may work better in case when the precise object matching is important. In this paper we are concerned with generic image retrieval only and our model avoids the re-normalization by modifying the similarity measure.

Current approaches are also data storage and computationally expensive which makes them less suitable for real user oriented applications, for example, to incorporate content into a user profile.

To tackle the aforementioned problems, we propose a novel approach to exploit the inter-relationships between the visual words. We introduce and test a few notions of correlation. First, we generate a matrix of correlations between visual words for each top image returned in the first round retrieval. Second, we aggregate the matrices and identify the dominant and least correlated coefficients. Thus obtained information, along with the visual words' frequencies from the current query, is then utilized to weight the similarity measure. Certain coefficients in the similarity measure corresponding to highly correlated terms are then increased, while the coefficients related to least correlated visual words are deemphasized. The images returned in the first round retrieval are then re-ranked according to the modified similarity measure. The improved performance, observed on three different data collections, is in our opinion a promising indicator for the real user evaluation. The proposed approach should let us directly incorporate image content into user profiles, where each profile would be represented in a form of a matrix of correlations between visual words obtained from the query history. Thus obtained user profile, which would store user visual preferences, could be utilized to adjust the similarity measure with respect to each individual user.

2 Related Work

Readers interested in local features and the "bag of visual words" approach are referred to [1,2,3] for the detailed description and application of aforementioned methods.

An interactive image retrieval model with adaptive similarity measure is introduced in [14]. The weights for adjusting the similarity measure (with respect to the image content representation - global features) are calculated according to the consistency of the vectors' components representing images collected from user relevance feedback. First, the representations of images deemed relevant by the user are stacked to form a matrix. Next, if a column contains elements with similar values then this particular dimension is considered to be a good indicator of the user's information need and the weight is calculated as an inverse of standard deviation across this dimension.

Liu et. al. ([6]) exploit co-occurrence information in spatial domain. Authors make an assumption that the related visual words would appear in a certain neighbourhood. They utilize the equivalent of $tf \cdot idf$ weighting scheme from text retrieval. Having obtained the information about the relationships, they use it to update the current query by weighting all the coefficients in the histogram. This leads to the normalization process which may hamper the performance [1].

Another approach [7] tries to capture the spatial relationships between pairs of visual words by building a visual word tree. The tree is generated by clustering interest points that co-occurred within some spatial distance. Latent Semantic Analysis (LSA) is then applied to compute the importance of each visual word to the given query, and the most important ones become so-called topic words. The $tf \cdot idf$ weighting scheme and the topic words are then utilized to re-rank the images. This approach, although quite efficient in comparison with others, is not applicable to real user evaluation because of the computational cost (high dimensional Scale Invariant Feature Transform descriptor, costly LSA).

Model proposed in [8] utilizes data mining techniques to discover spatially co-occurrent patterns of visual words. Authors report limitations of standard codebook generation techniques (related to synonymy and polysemy of visual words) and propose a novel approach, which constructs a higher-level visual phrase lexicon consisting of groups of co-located visual words.

Spatial correlations are also exploited in [9] where they are represented by correlograms. Experimental results show, that the joint models (B.O.F and correlogram) outperform standard appearance-only models. However, models based only on correlograms perform worse than standard B.O.F approach.

Jamieson et. al. [10] propose to group features that exist within a local neighbourhood, claiming that arrangements or structures of local features are more discriminative. Such groups of visual words are then associated with annotation words.

Trigram model is proposed in [11] to help in image classification. The method captures spatial correlations between image patches. Comparison between unigram and trigram models shows that the latter one improves the classification accuracy.

Another model [12] defines visual phrase-based image similarity. First, they count occurrences of each visual word. Then the occurrences of adjacent patch pairs formed by frequent visual words are counted and finally, the visual phrases are generated by selecting the adjacent patch pairs whose occurrences are higher than the threshold. The similarity between two images is measured by cosine metric with $tf \cdot idf$ weighting scheme adapted from text retrieval.

In general, methods that utilize information about correlations between visual words try to group semantically similar visual words' together. They usually consider co-occurrences at one contextual level and are computationally expensive and not scalable. Our method, in contrast, is computationally and data storage cheap, utilizes co-occurrences at various contextual levels, and avoids the normalization of histograms of visual words' counts. These properties make it suitable for a real user evaluation, where a user profile would represent user visual preferences. Such type of user profile would be utilized to put a query into the right context.

3 Notions of Correlation Between Visual Words

Here, we introduce image-level and proximity-based notions of correlation. In text retrieval (see [5]) document-level correlations seem to be stronger. A document may contain correlated terms not because of their proximity, but because they refer to the same topic.

Because our histograms of visual words' counts can be classified as a mid-level representation (the BOF reduces the semantic gap), we can introduce the correlations in a relatively intuitive way. Let us first focus on the correlations at the image level.

Correlation 1 can be regarded as the number of all pairs between the instances of different visual words (see Figure 1). Here, for instance, the squares denoted as A represent different instances of the same visual word (image patches) that appears within an individual image. When dealing with a set of images, we would aggregate the correlation matrices generated for each image. In case of Correlation 1, this would be equivalent to putting histograms of visual words counts as rows in a matrix and multplying the transposition of this matrix by itself. This is an analogy to document-level correlation in text IR. Correlation 2 is a normalized version of Correlation 1, where the denominator is a total number of all possible pairs between occurrences of visual words (Figure 2).

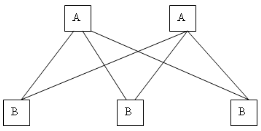


Fig. 1. Interpretation of Correlation 1. This is the common document/image level correlation. Here, squares denote instances of visual words (image patches) and the links the relationships between them.

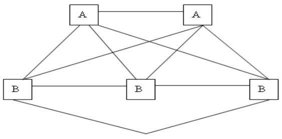


Fig. 2. Normalization factor in Correlation 2. Here, squares denote instances of visual words (image patches) and the links the relationships between them.

Correlation 3 (Figure 3) can also be regarded as the number of pairs between the occurrences of different visual words, but this time the correspondence is as follows (see Figure 3).

1. $corr(vt_i, vt_j) = vt_if \cdot vt_jf$

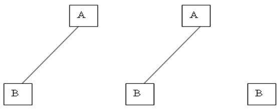


Fig. 3. Interpretation of Correlation 3. Here, squares denote instances of visual words (image patches) and the links the relationships between them.

2. $corr(vt_i, vt_j) = \frac{2 \cdot vt_i f \cdot vt_j f}{(vt_i f + vt_j f) \cdot (vt_i f + vt_j f - 1)} = \frac{vt_i f \cdot vt_j f}{\binom{vt_i f + vt_j f}{2}}$
3. $corr(vt_i, vt_j) = \min(vt_i f, vt_j f)$
4. $corr(vt_i, vt_j) = \frac{vt_i f \cdot vt_j f}{\binom{vt_i f + vt_j f}{2}} + \min(vt_i f, vt_j f)$

where vt_i, vt_j denote the ith and jth visual term respectively, and $vt_i f$, $vt_j f$ denote the frequencies (number of occurrences) of the terms. By calculating the correlations between all visual words in a particular image, we will obtain a matrix of correlations:

$$\begin{pmatrix} corr(vt_1, vt_1) & corr(vt_1, vt_2) & \dots & corr(vt_1, vt_n) \\ corr(vt_2, vt_1) & corr(vt_2, vt_2) & \dots & corr(vt_2, vt_n) \\ \vdots & \vdots & \dots & \vdots \\ corr(vt_n, vt_1) & corr(vt_n, vt_2) & \dots & corr(vt_n, vt_n) \end{pmatrix}$$

The matrix corresponding to the first notion of correlation can also be obtained by calculating the inner product of a transposed vector image representation and itself $h^t \cdot h$.

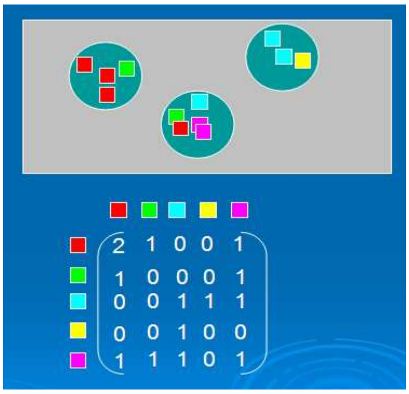
At first, there does not seem to be much difference between these three relationships. A closer look will show us the contradictions with our intuition of correlation.

Let us focus on Correlation 1. If the frequencies of two pairs of visual words are $\{5, 10\}$ and $\{5, 100\}$ then the latter will be assigned higher correlation value. We would, however, expect the former pair to be at least equally correlated.

Normalization (Correlation 2) helps to overcome the above issue. However, if the frequencies are proportional, for example $\{10, 20\}$ and $\{40, 80\}$ then the former will score higher. But, intuitively, the latter is more correlated.

Correlation 3 seems to be intuitively right, but will ignore the additional information from the frequencies (see example for Correlation 1). Normalization of correlation 3 will produce similar side effects to Correlation 2. Therefore, we introduce the Correlation 4, which does not seem to contradict our intuition. Experimental results confirm the superiority of this notion of correlation in the user simulation.

Above notions of correlation consider two instances of visual words to co-occur if they appear somewhere within an image (visual context - the whole image). Let us now introduce, by analogy to text retrieval, what we will refer to as proximity-based correlation. Two instances of visual words will be considered correlated if they appear together within a certain neighbourhood (visual context - "sliding window"). In case of dense sampling this is rather straightforward. When dealing with sparse sampling, however, we need to shift the window (square, circular) from one instance of visual word to another. Figure 4 shows an example of proximity-based correlation. Here, the squares denote instances of various visual words. Now we can show how to incorporate the information about correlations into the Pseudo Relevance Feedback (PRF). PRF assumes that the top documents from the first round retrieval are all relevant to the query. Then, the additional information from the top documents is usually utilized to expand the query.



	🟥	🟩	🟦	🟨	🟪
🟥	2	1	0	0	1
🟩	1	0	0	0	1
🟦	0	0	1	1	1
🟨	0	0	1	0	0
🟪	1	1	1	0	1

Fig. 4. Proximity-based correlation. For the clarity of presentation, the matrix corresponds to only three instances of visual words (circles' centres)

Initially, the first round retrieval is performed. Then, for each image from the top returned images, the matrix of correlations will be created. We aggregate all the matrices in order to obtain the final matrix from which the most and least dominant correlations will be identified (in terms of values). Notice, that in case of Correlation 1, this approach would be equivalent to constructing a matrix with rows corresponding to each image representation (from the top returned images)

$$M = \begin{pmatrix} vt_1^1 f & vt_2^1 f & \dots & vt_n^1 f \\ vt_1^2 f & vt_2^2 f & \dots & vt_n^2 f \\ \vdots & \vdots & \dots & \vdots \\ vt_1^m f & vt_2^m f & \dots & vt_n^m f \end{pmatrix}$$

and multiplying $M^t * M$, where t denotes the transpose operation. The advantage of our method is that it does not restrict us to one notion of correlation and we can define it in a more intuitive way.

4 Adaptation of Similarity Measure Using Most and Least Correlated Visual Words

As aforementioned, we can identify a few most and least correlated visual words from the matrix of correlations. We can now utilize this information to modify the similarity measure. For this purpose, we are going to use Minkowski fractional similarity measure (the method may be used with any measure from the Minkowski's family of distances).

First, we must identify a certain number (see Experimental Setup section for details) of most and least correlated visual terms by looking at the correlation matrix's elements' values above or below the diagonal (symmetrical matrix). Let's assume that we have identified the dominant correlation $\{v_k, v_l\}$ and the least correlated pair $\{v_n, v_m\}$. Let us now look at the query and extract the frequencies of visual terms corresponding to v_k, v_l, v_n, v_m. We can assume that $v_k f \geq v_l f$ and $v_n f \geq v_m f$, where vf denotes the frequency of a visual word taken from the query.

We can weight the similarity measure as follows

$$
\begin{aligned}
d(Q,I) = &\left(\sum_{i=1}^{N} \sqrt{|v_{Q_i} f - v_{I_i} f|}\right)^2 = \\
&= \left(\sqrt{|v_{Q_1} f - v_{I_1} f|}\right)^2 + \left(\sqrt{|v_{Q_2} f - v_{I_2} f|}\right)^2 \\
&+ \ldots + \left(\sqrt{\frac{v_{Q_k} f}{v_{Q_l} f \cdot \log_b c\,(v_k, v_l)}\, |v_{Q_l} f - v_{I_l} f|}\right)^2 \\
&+ \ldots + \left(\sqrt{\frac{v_{Q_m} f}{v_{Q_n} f \cdot \log_b c\,(v_n, v_m)}\, |v_{Q_n} f - v_{I_n} f|}\right)^2 \\
&+ \ldots + \left(\sqrt{|v_{Q_N} f - v_{I_N} f|}\right)^2
\end{aligned}
$$

where Q denotes the query representation, I is an image representation from the data collection, and $c\,(v_k, v_l)$ and $c\,(v_n, v_m)$ are the correlation values taken from the correlation matrix..

Thus, we increase the elements corresponding to the visual word in the query with lower frequency value (dominant correlations), and decrease the elements corresponding to the visual word in the query with higher frequency value (least correlated pairs). Having done the similarity measure weighting, we re-rank the top images by calculating the new distance between the query and the images returned in the first round retrieval.

5 Experiments and Discussion

We evaluate the proposed method on three large data collections: ImageCLEFphoto 2007 (20000 images), MIRFlickr 25000 (25000 images) and a collection from British Geological Survey (BGS, 7432 images). The collections differ significantly in size and content. For each of the 100 query topics (60 for ImageCLEF) we retrieve 16 images and calculate Mean Average Precision (MAP). To test the influence of the correlations on the retrieval performance, we generate the correlation matrix from these 16 images, weight the similarity measure, and re-rank the top images. Next, we compute the Mean Average Precision and compare it with the baseline (which does not take the correlations into account).

Some images belong to a few categories. The evaluation on MIRFlickr and BGS collections was the "lenient" one. We assumed that an image is relevant if it shares at least one category with the query image (based on the ground truth data provided).

5.1 Experimental Setup

The implemented local features utilize the random sampling technique. We set the number of sample points to 900. A large number of sample points (in random sampling) is expected to give better results than other sampling methods (see [3]). For each sample point of an image, a 10×10 square patch around it was characterized as multidimensional vector by applying a local descriptor. Each image patch has 9 dimensions (3 for each colour channel), and the codebook size is 40. The visual features, despite using low dimensional vectors and small vocabulary, are comparable with more sophisticated approaches (ImageCLEF2010 Wikipedia Retrieval Task). For a detailed description of the local features used, the reader is referred to [15].

When exploiting the correlations between visual words, we identify 5 dominant and 1 least correlated pair. The p and c parameters' values in the similarity measure are set to 0.5 and 1.31 for all three data collections and were determined experimentally. In case of proximity-based correlation, we will consider two instances of visual words to be correlated if they both appear within a circle of radius 14.15. This is approximately the sum of two diagonals of square sub-images (image patches may overlap).

5.2 Experimental Results and Discussion

Tables 1, 2, and 3 show the experimental results. They present the results for the case when no Pseudo Relevance Feedback was incorporated (NP), when only the dominant correlations were taken into account (D), and the MAPs for both dominant and least correlated pairs (DL). The performance of five notions of correlation is also depicted in the tables. Labels C1, C2, C3 and C4 correspond to correlations 1, 2, 3 and 4 accordingly (see Correlation Between Visual Words section, image-level) whereas C0 denotes proximity-based correlation. The computation of correlation 1 and then addition of matrices is equivalent to commonly used multiplication of the transpose of an image representation matrix by itself. It is one of the standard ways for capturing correlation. Therefore, correlation 1 can also be considered as another baseline. Results presented in bold font are significantly different (two-tailed t-test, 0.05) from the baseline.

It can be seen that C4 and C3 correlations obtained the best results on all three data collections. The addition of information about the least correlated visual words often further improves the performance. Moreover, image level correlations outperformed proximity based one. This may be due to the notion that an image may contain correlated visual words not because of their proximity but because they refer to the same topic.

Table 1. ImageCLEF2007 results (MAP)

	C4	C3	C2	C1	C0
NP	0.0204	0.0204	0.0204	0.0204	0.0204
D	0.0211	0.0211	0.0210	0.0208	0.0206
DL	0.0213	0.0213	0.0211	0.0209	0.0207

Table 2. MIRFlickr results (MAP)

	C4	C3	C2	C1	C0
NP	0.6794	0.6794	0.6794	0.6794	0.6794
D	**0.6938**	**0.6936**	0.6859	0.6869	0.6802
DL	**0.6951**	**0.6936**	0.6854	0.6871	0.6807

Table 3. BGS results (MAP)

	C4	C3	C2	C1	C0
NP	0.3158	0.3158	0.3158	0.3158	0.3158
D	**0.3286**	**0.3286**	0.3187	0.3199	0.3172
DL	0.3268	0.3265	0.3194	0.3193	0.3176

We should be aware, however, that the assumption in PRF framework that all the top documents are relevant to the query may produce a number of false correlations. The process will therefore depend on the adequacy (the ability to capture relevant properties) of the image representation and the retrieval performance of the implemented methods. The real user evaluation should, however, be able to overcome these limitations because all the queries will be selected by the user.

6 Conclusions and Future Work

In this paper we propose a new approach for identifying and utilizing the information about correlations between visual words. We implement and test various notions of correlation at different contextual levels (we refer to them as image-level and proximity based). To the best of our knowledge, this is the first time these two were compared within this type of framework in image retrieval.

Experimental results show the superiority of two notions of correlation, C4 and C3, which are image level correlations. For these two correlations, we report significant improvement in terms of Mean Average Precision on two data collections within PRF evaluation framework. Moreover, the addition of information about the least correlated visual words often further improves the performance. Proximity based notion of correlation does not show a significant improvement in the context of this model.

The proposed method is computationally and data storage cheap, utilizes correlation at different contextual levels, and avoids the normalization of histograms. We believe that the our approach can be successfully incorporated into the experiment involving real users. Thus, a user profile (correlation matrix generated from the query history) could be stored for each individual user, and the information from the profile would be utilized to put the query in the right visual context.

We are planning to extend our evaluation to other various weighting schemes and similarity measures. The ultimate goal, however, would be the aforementioned real user evaluation. The proposed method was developed for this purpose. We will try to take into account the ranking of the top retrieved images and the order of the queries in the query history, as the current query should be given more importance than others. When it comes to the automated methods, like Pseudo Relevance Feedback for example, the assumption that all the top documents are relevant to the query may produce a number of false correlations. The process will therefore depend on the adequacy (the ability to capture relevant properties) of the image representation and the retrieval performance of the implemented methods. The real user evaluation should, however, be able to overcome these limitations and the promising results encourage us to pursue the proposed approach.

References

1. J. Yang, Y.G. Jiang, A.G. Hauptmann, C.W. Ngo. Evaluating bag-of-visual-words representations in scene classification. *In Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, 206, 2007.
2. E. Nowak, F. Jurie. Learning visual similarity measures for comparing never seen objects. *IEEE Conference on Computer Vision and Pattern Recognition*, 1–8, 2007.
3. E. Nowak, F. Jurie, B. Triggs. Sampling strategies for bag-of-features image classification. *Lecture Notes in Computer Science*, 3954(490), 2006.
4. M. Grubinger, P. Clough, A. Hanbury, H. Muller. Overview of the ImageCLEFphoto 2007 photographic retrieval task. *Advances in Multilingual and Multimodal Information Retrieval*, 433–444, 2008.
5. C. Biancalana, A. Lapolla, A. Micarelli. Personalized web search using correlation matrix for query expansion. *Web Information Systems and Technologies*, 186-198, 2009.
6. T. Liu, J. Liu, Q. Liu, H. Lu. Expanded bag of words representation for object classification. *16th IEEE International Conference on Image Processing (ICIP)*, 297–300, 2010.
7. S. Zhang, Q. Huang, Y. Lu, G. Wen, Q. Tian. Building pair-wise visual word tree for efficient image re-ranking. *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 794–797, 2010.
8. J. Yuan, Y. Wu, M. Yang. Discovery of collocation patterns: from visual words to visual phrases. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR'07*, 1–8, 2007.
9. S. Savarese, J. Winn, A. Criminisi. Discriminative object class models of appearance and shape by correlatons. *IEEE Computer Society*, 1063-6919, 2006.
10. M. Jamieson, S. Dickinson, S. Stevenson, S. Wachsmuth. Using language to drive the perceptual grouping of local image features. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:2102–2109, 2006.
11. L. Wu, M. Li, Z. Li, W.Y. Ma, N. Yu. Visual language modeling for image classification. *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, 115–124, 2007.
12. Q.F. Zheng, W.Q. Wang, W. Gao. Effective and efficient object-based image retrieval using visual phrases. *Proceedings of the 14th Annual ACM International Conference on Multimedia*, 77–80, 2006.
13. J. Sivic, A. Zisserman. Video Google: Video Google: A Text Retrieval Approach to Object Matching in Videos. *Proceedings of the 9th IEEE International Conference on Computer Vision*, 2:1470 - 1477, 2003.
14. Y. Rui, T. S. Huang, M. Ortega, S. Mehrotra. Relevance Feedback: A Power Tool in Interactive Content-Based Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8:644 - 655, 1998.
15. L. Kaliciak, D. Song, N. Wiratunga, J. Pan. Novel local features with hybrid sampling technique for image retrieval. *Proceedings of Conference on Information and Knowledge Management (CIKM)*,1557–1560 , 2010.