
Retrieving Keyword's to an Image Query using Kernel CCA

David R. Hardoon

School of Electronics & Computer Science
ISIS Group
University of Southampton
Southampton SO17 1BJ, U.K.
drh@ecs.soton.ac.uk

Sandor Szedmak

School of Electronics & Computer Science
ISIS Group
University of Southampton
Southampton SO17 1BJ, U.K.
ss03v@ecs.soton.ac.uk

John Shawe-Taylor

School of Electronics & Computer Science
ISIS Group
University of Southampton
Southampton SO17 1BJ, U.K.
jst@ecs.soton.ac.uk

Abstract

In this paper we propose an approach to automatically annotate query images with keywords. We use kernel Canonical Correlation Analysis to learn a semantic representation between images and their associated documents. The semantic space provides a common representation and enables a comparison between the documents and images. This representation is then used in the creation of new document, comprised from the keywords that best fit the image query. We compare our method against a standard cross-representation retrieval technique known as Generalised Vector Space Model.

1 Introduction

Due to the increasing rise of multimedia data available both on-line and off-line we are faced with the problematic issue of our ability to access or make use of this information, unless it is organised in such a way as to allow efficient browsing, searching and retrieval. One of the issues is image labelling or multi-labelling where we would like to annotate an image with several keywords which best describe it. A recent solution proposed by [1] is image segmentation which then has key words associated with the different segmented parts of the image. In this work we propose an approach which excludes any tampering with the image structure and the favouring of words to specific categories. In previous work [2] we presented an approach based on Kernel Canonical Correlation Analysis (KCCA) using the content of both views to retrieve images based on a text query by looking for the highest weighting between the query text and the test images in the feature space. We could reverse the system to look for the best matching document in the given test-set to an image query. Although we claim that a) you may not have documents to test

against, other than those in the training-set, and b) we would like to annotate an image query independently from the given set of words within a specific document. Hence we create a new document d^* which contains the keywords that best fit the query image. It is important to state that during the training stage we do not do any word annotation to the images other than using KCCA to find a common feature representation between the documents, and hence words, to the images.

The paper is divided as follows, in Section 2 we give a brief introduction to CCA for brevity we exclude the full kernelisation of CCA. In Section 3 we present a method of creating a new document d^* which best fits the query image while in Section 4 we present our experiments and results. Finally we present our conclusions in Section 5.

2 Canonical Correlation Analysis

Proposed by H. Hotelling in 1936 [4], Canonical correlation analysis can be seen as the problem of finding basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors are mutually maximised. Correlation analysis is dependent on the co-ordinate system in which the variables are described, so even if there is a very strong linear relationship between two sets of multidimensional variables, depending on the coordinate system used, this relationship might not be visible as a correlation. Canonical correlation analysis seeks a pair of linear transformations one for each of the sets of variables such that when the set of variables are transformed the corresponding coordinates are maximally correlated.

$\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product of the vectors \mathbf{x}, \mathbf{y} and is equal to $\mathbf{x}'\mathbf{y}$. Where A' to denote the transpose of a vector or matrix A .

Consider a multivariate random vector of the form (\mathbf{x}, \mathbf{y}) . Suppose we are given a sample of instances $S = ((\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_l, \mathbf{y}_l))$ of (\mathbf{x}, \mathbf{y}) , let S_x denote $(\mathbf{x}_1, \dots, \mathbf{x}_l)$ and similarly S_y denote $(\mathbf{y}_1, \dots, \mathbf{y}_l)$. We can consider defining a new co-ordinate for \mathbf{x} by choosing a direction \mathbf{w}_x and projecting \mathbf{x} onto that direction $\mathbf{x} \rightarrow \langle \mathbf{w}_x, \mathbf{x} \rangle$ if we do the same for \mathbf{y} by choosing a direction \mathbf{w}_y we obtain a sample of the new \mathbf{x} co-ordinate, let $S_{x, \mathbf{w}_x} = (\langle \mathbf{w}_x, \mathbf{x}_1 \rangle, \dots, \langle \mathbf{w}_x, \mathbf{x}_l \rangle)$ with the corresponding values of the new \mathbf{y} co-ordinate being $S_{y, \mathbf{w}_y} = (\langle \mathbf{w}_y, \mathbf{y}_1 \rangle, \dots, \langle \mathbf{w}_y, \mathbf{y}_l \rangle)$. The first stage of canonical correlation is to choose \mathbf{w}_x and \mathbf{w}_y to maximise the correlation between the two vectors. In other words the function to be maximised is

$$\max \rho = \text{corr}(S_{x, \mathbf{w}_x}, S_{y, \mathbf{w}_y}) = \frac{\langle S_{x, \mathbf{w}_x}, S_{y, \mathbf{w}_y} \rangle}{\|S_{x, \mathbf{w}_x}\| \|S_{y, \mathbf{w}_y}\|}$$

We use $\hat{\mathbb{E}}[f(\mathbf{x}, \mathbf{y})]$ to denote the empirical expectation of the function $f(\mathbf{x}, \mathbf{y})$, where $\hat{\mathbb{E}}[f(\mathbf{x}, \mathbf{y})] = \frac{1}{l} \sum_{i=1}^l f(\mathbf{x}_i, \mathbf{y}_i)$. We can rewrite the correlation expression as

$$\begin{aligned} \max \rho &= \frac{\hat{\mathbb{E}}[\langle \mathbf{w}_x, \mathbf{x} \rangle \langle \mathbf{w}_y, \mathbf{y} \rangle]}{\sqrt{\hat{\mathbb{E}}[\langle \mathbf{w}_x, \mathbf{x} \rangle^2] \hat{\mathbb{E}}[\langle \mathbf{w}_y, \mathbf{y} \rangle^2]}} = \frac{\hat{\mathbb{E}}[\mathbf{w}_x' \mathbf{x} \mathbf{y}' \mathbf{w}_y]}{\sqrt{\hat{\mathbb{E}}[\mathbf{w}_x' \mathbf{x} \mathbf{x}' \mathbf{w}_x] \hat{\mathbb{E}}[\mathbf{w}_y' \mathbf{y} \mathbf{y}' \mathbf{w}_y]}} \\ &= \frac{\mathbf{w}_x' C_{\mathbf{x}\mathbf{y}} \mathbf{w}_y}{\sqrt{\mathbf{w}_x' C_{\mathbf{x}\mathbf{x}} \mathbf{w}_x \mathbf{w}_y' C_{\mathbf{y}\mathbf{y}} \mathbf{w}_y}} \end{aligned}$$

subject to $\mathbf{w}_x' C_{\mathbf{x}\mathbf{x}} \mathbf{w}_x = 1$ and $\mathbf{w}_y' C_{\mathbf{y}\mathbf{y}} \mathbf{w}_y = 1$ as the quotient is not effected by rescaling of \mathbf{w}_x and \mathbf{w}_y . The total covariance matrix C is a block matrix where the within-sets covariance matrices are $C_{\mathbf{x}\mathbf{x}}$ and $C_{\mathbf{y}\mathbf{y}}$ and the between-sets covariance matrices are $C_{\mathbf{x}\mathbf{y}} = C_{\mathbf{y}\mathbf{x}}'$, although this is subjected to the data having a zero-mean.

The dual form of CCA can be formulated by expressing the image weights W_x and document weights W_y as a linear combination of the training examples $W_x = X'\alpha$ and $W_y = Y'\beta$ where X and Y are matrices with rows $\{\mathbf{x}_1, \dots, \mathbf{x}_1\}$ and $\{\mathbf{y}_1, \dots, \mathbf{y}_1\}$ respectively. Let $W_x = \{\mathbf{w}_x^1, \dots, \mathbf{w}_x^1\}$ and $W_y = \{\mathbf{w}_y^1, \dots, \mathbf{w}_y^1\}$ as we obtain a set of weight vectors for each sample. Respectively α and β are a sequence of feature vectors such that $\alpha = \{\alpha^1, \dots, \alpha^1\}$ and $\beta = \{\beta^1, \dots, \beta^1\}$.

Following [3, 7, for full derivation] the dual form of CCA with regularisation parameter τ will be given by solving

$$\max_{\alpha, \beta} \rho = \frac{\alpha' \mathbf{K}_x \mathbf{K}_y \beta}{\sqrt{((1-\tau)\alpha' \mathbf{K}_x^2 \alpha + \tau\alpha' \mathbf{K}_x \alpha)((1-\tau)\beta' \mathbf{K}_y^2 \beta + \tau\beta' \mathbf{K}_y \beta)}}$$

subject to $(1-\tau)\alpha' \mathbf{K}_x^2 \alpha + \tau\alpha' \mathbf{K}_x \alpha = 1$ and $(1-\tau)\beta' \mathbf{K}_y^2 \beta + \tau\beta' \mathbf{K}_y \beta = 1$. Where \mathbf{K}_x is the kernel matrix for the images and \mathbf{K}_y the respective kernel matrix for the documents.

3 Creating d^* Matching the Image Query

We confronted with the problem of creating a new document d^* which best matches our image query. Based on the idea of the CCA we are looking for a vector such that it has maximum correlation to the query image respect to the weight matrices α and β . Let $f = K_x^I \alpha$, where the vector K_x^I contains the kernelized inner products between the query image I and the images occurring in the training set. We have

$$\max_{d^*} \langle f, W_y d^* \rangle, \quad (1)$$

where $W_y = Y'\beta$. Assume that the new document is represented in the form $d^* = D\lambda$, where D a design matrix with size $n \times m$, where n is the number of known words in the training dataset and m is an arbitrary number depending on the expected structure of the document. The reader can find examples to the design matrix below. The λ is a vector variable giving a convex combination of the columns of the design matrix, thus it satisfies the constraints $\sum_{i=1}^m \lambda_i = 1$ and $\lambda_i \geq 0$, $i = 1, \dots, m$. The problem becomes

$$\max_{\lambda} f' W_y D \lambda \quad (2)$$

under the same constraints. Applying the notation $c' = f' W_y D$ we have

$$\max_{\lambda} c' \lambda, \quad (3)$$

therefore due to the constraints the components of the optimum solution λ^* equal to

$$(\lambda)_i^* = \begin{cases} 1 & i = \arg \max_j c_j, \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

If the design matrix D is the identity matrix then we receive a document comprising one word only. Let K be a given number, if every column of D contains 1 in K components and all other components are 0 and all possible but different vectors are included with this property in D then one can derive the best fitting document containing K words. It is easy to show the optimum solution for this design matrix gives the column of D corresponding to the K greatest values of the vector $f' W_y$, hence without building up a huge design matrix the K best words fitting to the query image can be found.

4 Experiements

In the following experiments the problem of learning the semantics of multimedia content by combining image and text data is addressed. The learnt semantics is then applied to

Table 1: Confusion matrix of the # of words in each subset.

	<i>A</i>	<i>B</i>	<i>C</i>
<i>A</i>	1269	447	324
<i>B</i>	447	850	86
<i>C</i>	324	86	327

Table 2: Example of words in categories:

Sports	aa, aaron, abdominal, abdul, abdur, abe, ability, abroad, absence, accomplish, accounted, achieve, achilles, acquired ...
Aviation	air, airborne, aircraft, airesearch, airfield, airflow, airframe, airline, airliner, airmotive, airpark, areonautical, alaska, alap ...
Paintball	warriors, watch, weather, web, wednesday, white, wildcards, wildfire, win, wing, winning, wizard, wordogz, work, world ...

the annotation of keywords to query images. The aim is to allow retrieval of keywords from an image query without reference to any labelling associated with the image. We use a combined multimedia image-text web database, which was kindly provided by the authors of [5]. The data was divided into three classes: Sport, Aviation and Paintball, 400 records each and consisted of jpeg images retrieved from the Internet with attached text. We randomly split each class into two halves. The extracted features of the data were used as in [5] (detailed description of the features can be found in [5]): image HSV (Hue Saturation Values) colour, image Gabor texture and term frequencies in the text. The representation of the words is a crucial one, as we wish to capture the information relating the words within the documents to the images, therefore we compare two approaches of word representation; The term frequency vector which is the number of occurrences of the word in the document, and Term Frequency Inverse Document Frequency (TFIDF) [6] which is

$$TFIDF(d_i, w_j) = (\#^1 \text{ of } w_j \text{ in } d_i) \cdot \log\left(\frac{N}{\# \text{ of documents that contain } w_j}\right) \quad (5)$$

where N is the number of documents, d_i is document i and w_j is word j .

We compare the performance of our method with a retrieval technique based on the Generalised Vector Space Model (GVSM). This uses as a semantic feature vector, the vector of inner products between either a query image and each training images or test documents and each training label. The first view was obtained by the combination of a Gaussian kernel (with σ as the minimum distance between the different images) with a linear kernel on the Gabor textures, and the second view by a linear kernel on the term frequencies or TFIDF features. We compute the KCCA regularisation parameter τ as described in [2].

Let \mathcal{W} be the set of all words in our database comprising of an overall of 3522 words and let $A, B, C \subseteq \mathcal{W}$ such that A is the Sports category with 2259 words, B the Aviation category with 1602 words and C the Paintball category with 956 words. The words in \mathcal{W} are associated to the categories by the following approach; for example, if word w is associated to an image that belongs to category A , w will belong to A as well. In table 2 an example of the different words in the categories are presented.

After finding the new document d^* as described in Section 3 we try and evaluate the

¹# - number

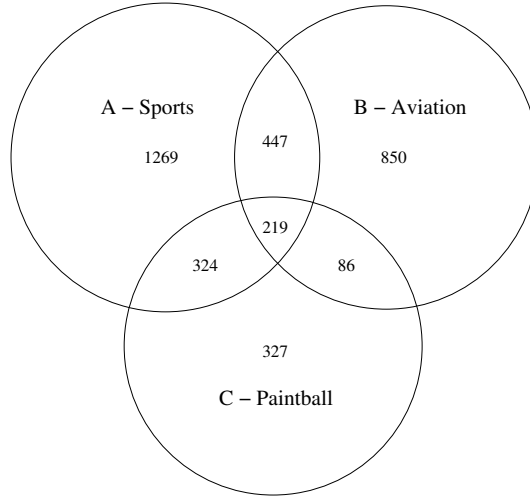


Figure 1: We separate the dictionary into its categories with overlapping regions.

relevance of the words found to the query image. As we attempt an automatic word retrieval where human intervention is at a minimum the definition of a “relevant” word is not trivial. We attempt this definition by denoting three levels of relevance, the first - level 1; If a word is in a category (including overlapping categories) and is the same label as the image, it is considered as a success with a weighting 1 otherwise the word will have a weighting of 0. In level 2 we follow the weightings of level 1 but start penalising the words that are in overlapping categories by -0.25 such that if a word belongs to either of the two overlapping subsets it will be given a weighting of 0.75 and a weighting of 0.5 if it belongs to all three. The last and final level, level 3, we further increase the penalty to -0.5 such that if a word belongs to either of the two overlapping categories it will have a weighting 0.5 and if the word belongs to all three categories it will be considered as a mistake (a weighting of 0). The choice of the penalty is arbitrary.

In tables 3 and 5 we present the KCCA comparison between the two word representation for a retrieval task of 1 and 10 words, the best success rate for the specific task is highlighted. The presented results are an average over all the query test images (600) repeated 10 times for the eigenvector selection giving the highest retrieval %. In each repeat we randomly split the testing and training samples. In tables 3 and 5 ‘eig’ are the corresponding eigenvectors used in the feature projection to obtain the result. As shown in [2] we do not need to test for all eigenvectors (600) as the success rate will generally saturate after half the number of eigenvectors. Therefore we suffice in testing for a selection of 1 – 300 eigenvectors and picking the best one for the specific test.

We expect that for some of the image queries the keywords would overlap, although as we would like to show that our approach does not return the trivial solution (i.e. the same keywords for all queries). We compute the variance² of the correctly retrieved keywords $variance = \frac{\# \text{ of different correct words found}}{\text{all correct words found}}$. In the case of retrieving 10 words we can not avoid overlapping keywords, therefore we normalise the variance by the maximum different keywords possible in that case ($\frac{3522}{600 \cdot 10}$).

As we expect when increasing the penalty in the three levels of relevance, the suc-

²In variance we mean the difference between words.

Table 3: KCCA: Success results using term frequency. (eig - eigenvectors)

	Retrieving 1 word			Retrieving 10 words		
	Retrieval	Variance	# eig	Retrieval	Variance	# eig
Level 1	96.27%	0.77%	3	89.51%	1.65%	5
Level 2	73.17%	31.17%	270	71.77%	21%	283
Level 3	51.19%	37.2%	273	45.45%	2.81%	6

Table 4: GSVM: Success results using Term Frequency

	Retrieving 1 word		Retrieving 10 words	
	Retrieval	Variance	Retrieval	Variance
Level 1	80.07%	0.41%	61.69%	0.58%
Level 2	52.43%	0.63%	44.55%	0.8%
Level 3	24.79%	0.68%	27.4%	1.03%

Table 5: KCCA: Success results using TFIDF features.(eig - eigenvectors)

	Retrieving 1 word			Retrieving 10 words		
	Retrieval	Variance	# eig	Retrieval	Variance	# eig
Level 1	88.14%	32.55%	264	86.22%	20.02%	299
Level 2	75.69%	38.93%	278	72.75%	23.72%	299
Level 3	63.34%	41.75%	278	59.24%	25.74%	296

Table 6: GSVM: Success results using TFIDF

	Retrieving 1 word		Retrieving 10 words	
	Retrieval	Variance	Retrieval	Variance
Level 1	35.62%	0.95%	54.33%	0.87%
Level 2	35.62%	0.95%	43.21%	1.1%
Level 3	35.62%	0.95%	32.09%	1.15%

cess of retrieving a “relevant” keyword will diminish. Although the TFIDF features need a larger number of eigenvectors for the feature projection it is able to produce a higher success rate than that of the term frequency vector, except in level 1 although we note the low variance that implies that the retrieved keywords are very similar. We see that as we increase the weight of the penalty TFIDF is able to extract words which are more singular to the topic and of a higher variance. In previous work [2] we have shown that increasing the number of eigenvectors for the feature selection will increase the success rates of the content-based image retrieval task, as visible in Tables 3 and 5 we can see a similar effect on the variance of the retrieved keywords. In tables 4 and 6 we present the baseline approach using term frequency and TFIDF. We are able to view that even with the high success rate of retrieval in level 1 using term frequency representation of the text the variance of the words are extremely low meaning that GSVM finds the same keywords for most of the queries. Clearly, KCCA is consistently better than the baseline method in both approaches.

In figures 2, 3, 4 and 6 we show examples of query images with their original text and the words retrieved using KCCA level 1 weighting scheme with TFIDF. The words that belong to the same category as the image are in bold while those which are a mistake are italicised. Figure 5 shows an example of an image with a more complicated text assigned to it.



Figure 2: Aviation: Original caption “is posed as if its nose gear has collapsed. executive decision is to the right of center. the 747-200 that appeared in the rookie appears at center. the convair 880 that was used in speed” best matching 10 words, from highest to lowest rank “convair, museum, cccp, eagle, voyage, cv, protivophozharny, tower, pima, roll”.



Figure 3: Aviation: Original caption “ec-121k, 141309, c/n 4433 . air force museum’s page about ec-121d, 53-0555 ec-121k, 141309 at the mcclellan afb museum on april 3, 1993. it was built as a navy wv-2, but it is displayed as air force ec-121d, 53-0552. its lockheed construction number is 4433.” best matching 10 words, from highest to lowest rank “museum, commando, goodyear, pima, page, takes, link, air, afb, castle”.



Figure 4: Paintball: Original caption “benini reffing” best matching 10 words, from highest to lowest rank “fate, darkside, kc, strange, team, wildcards, takeover, avljalde, hostile, check”.

Figure 5: Sports: Original caption “ap photo more photos february 18, 2002 toronto (ap) – sam cassell ’s injured toe didn’t hurt his effectiveness against the toronto raptors . but he might be selective about future games he plays in so he’s ready for the play-offs. “i’d rather take care of it now,” said cassell, who missed two games with a sprained left big toe before scoring 20 points in the milwaukee bucks ’ 91-86 victory over the toronto raptors on sunday. “this is not a joke,” he said. “this is probably the worst i’ve felt as a professional basketball player. it’s painful every step you take. coach says, ’you can take the pain!’ but not this kind of pain.” cassell, who scored eight points in the fourth quarter, said he would need 20 days to heal. “we have a game (monday) and there is a big possibility i might miss it,” he said. “this is the worst injury because there is nothing you can do for it. you can tape it, you can treat it. i never knew the big toe meant so much.” ray allen also had 20 points, michael redd 16 and tim thomas 15 for the bucks, who lost eight of their previous 10. alvin williams scored 24 points, but was 3-of-4 from the foul line late for the raptors, who have lost four straight since the all-star break. vince carter , hakeem olajuwon , jerome williams and dell curry missed the game with injuries. . . .” best matching 10 words, from highest to lowest rank “boyz, attitude, pt, hot, urban, quest ,rip, team,ap,matrix”.



Figure 6: Paintball: Original caption “all americans” best matching 10 words, from highest to lowest rank “ref, farside, american, team, trauma, stay, leader, flag, takeover, avljalde”.

5 Conclusions

The problem of retrieving information via content is still a non trivial one, although we present a relatively simple approach in annotating query images with keywords using kernel Canonical Correlation Analysis to find a semantic representation which is common for both views. We find that although the simplicity of the approach our results are promising and better than baseline method when using TFIDF with a sufficient number of eigenvectors for the feature projection. Though issues still remain such as, how can one define a better “relevance” test for the retrieved keywords? As we may also have image queries that do not have their associated keywords in the training corpus. It may also be relevant to devise a probabilistic scheme for the word penalty rather than using an arbitrary one. A further avenue that could be looked at, is the method of creating the new document d^* . Ideally to test the full potential of the system we would like to test it on a large scale image-text database system.

Acknowledgments

We would like to acknowledge the financial support of EU Projects LAVA, No. IST-2001-34405 and PASCAL network of excellence, No. IST-2002-506778

References

- [1] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Fretias, David M. Blei, and Michael I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.
- [2] David R. Hardoon and John Shawe-Taylor. KCCA for different level precision in content-based image retrieval. In *Proceedings of Third International Workshop on Content-Based Multimedia Indexing*, IRISA, Rennes, France, 2003.
- [3] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis; an overview with application to learning methods. Technical Report CSD-TR-03-02, Royal Holloway University of London, 2003.
- [4] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:312–377, 1936.
- [5] T. Kolenda, L. K. Hansen, J. Larsen, and O. Winther. Independent component analysis for understanding multimedia content. In H. Bourlard, T. Adali, S. Bengio, J. Larsen, and S. Douglas, editors, *Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII*, pages 757–766, Piscataway, New Jersey, 2002. IEEE Press. Martigny, Valais, Switzerland, Sept. 4-6, 2002.
- [6] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Berlin, 1983.
- [7] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.