KCCA FOR DIFFERENT LEVEL PRECISION IN CONTENT-BASED IMAGE RETRIEVAL

David R. Hardoon and John Shawe-Taylor

Computer Science Department
Royal Holloway University of London
Egham, Surrey, TW20 0EX, U.K.
{davidh, john}@cs.rhul.ac.uk

ABSTRACT

We use kernel Canonical Correlation Analysis to learn a semantic representation of web images and their associated text. In the application we look at two approaches of retrieving images based only on their content from a text query. The semantic space provides a common representation and enables a comparison between the text and image. We compare the approaches against a standard cross-representation retrieval technique known as the Generalised Vector Space Model.

1. INTRODUCTION

During recent years there has been a vast increase in the amount of multimedia content available both off-line and online. However we are unable to access or make use of this information unless it is organised in such a way as to allow efficient browsing, searching and retrieval. The authors of [7] review different techniques and approaches that have been applied in an attempt to solve this issue.

In [6] it is shown for the categorisation problem of multimedia content, the combination of the different types of data (text and images) is able to give a more accurate result then can each component separately. In previous work [8] we follow the motivation of [6] and apply the correlation approach to images and their associated text from web pages for mate retrieval. In Vinokourov et. al [9] a similar approach is applied for cross-lingual retrieval. We use the set of canonical correlation directions, to form a semantic representation of the text and image, which can be used not only to improve categorisation of the images [8] but also to retrieve them according to text queries. In this work we look at finding correlation between images and associated text and use that for retrieval methods. We suggest a general novel approach, which can be used both for content as well as mate based retrieval [8].

We give the description of the KCCA algorithm with a dimensionality reduction approach that uses incomplete Cholesky decomposition [1] and partial Gram-Schmidt orthogonalisation [3] in section 2. We explore image content and mate based retrieval using semantic projection of text queries and images components in section 3 where our experiments are presented, and in subsection 3.3 we motivate our selection of the KCCA regularisation parameter.

2. ALGORITHM DESCRIPTION

Using the KCCA algorithm [1, 9, 2] we try to obtain a standard eigenproblem for the kernel mapping of the text and image kernels. In [4] we observe that with full rank kernel matrices maximal correlation can be obtained, suggesting that learning is trivial. To force non-trivial learning we introduce a control on the flexibility of the projection mappings using Partial Least Squares (PLS) to penalise the norms of the associated weights. We convexly combine the PLS term with the KCCA term in the denominator (detailed description of CCA and KCCA can be found in [4]):

$$\rho = \max_{\alpha,\beta} \frac{\alpha' K_x K_y \beta}{\sqrt{\alpha' K_x^2 \alpha \cdot \beta' K_y^2 \beta}}$$

hence we obtain

$$\begin{aligned} \rho &= \max_{\alpha,\beta} \frac{\alpha' K_x K_y \beta}{\sqrt{(\alpha' K_x^2 \alpha + \kappa\|\alpha\|^2) \cdot (\beta' K_y^2 \beta + \kappa\|\beta\|^2)}} \\ &= \max_{\alpha,\beta} \frac{\alpha' K_x K_y \beta}{\sqrt{(\alpha' K_x^2 \alpha + \kappa\alpha' K_x \alpha) \cdot (\beta' K_y^2 \beta + \kappa\beta' K_y \beta)}} \end{aligned}$$

Observe that the added regularisation does not effect the scaling of α or β either together or independently. Hence, the optimisation problem is equivalent to maximising the numerator subject to

$$\begin{aligned} (\alpha' K_x^2 \alpha + \kappa\alpha' K_x \alpha) &= 1 \\ (\beta' K_y^2 \beta + \kappa\beta' K_y \beta) &= 1. \end{aligned}$$

The corresponding Lagrangain is

$$\begin{aligned} L(\lambda_\alpha, \lambda_\beta, \alpha, \beta) &= \alpha' K_x K_y \beta \\ &\quad -\frac{\lambda_\alpha}{2}(\alpha' K_x^2 \alpha + \kappa \alpha' K_x \alpha - 1) \\ &\quad -\frac{\lambda_\beta}{2}(\beta' K_y^2 \beta + \kappa \beta' K_y \beta - 1). \end{aligned}$$

Taking derivatives in respect to α and β we obtain

$$\frac{\partial f}{\partial \alpha} = K_x K_y \beta - \lambda_\alpha (K_x^2 \alpha + \kappa K_x \alpha) \quad (2.1)$$

$$\frac{\partial f}{\partial \beta} = K_y K_x \alpha - \lambda_\beta (K_y^2 \beta + \kappa K_y \beta). \quad (2.2)$$

Subtracting β' times the second equation from α' times the first we have

$$\begin{aligned} 0 = & \quad \alpha' K_x K_y \beta - \lambda_\alpha \alpha' (K_x^2 \alpha + \kappa K_x \alpha) \\ & -\beta' K_y K_x \alpha + \lambda_\beta \beta' (K_y^2 \beta + \kappa K_y \beta) \\ = & \quad \lambda_\beta \beta' (K_y^2 \beta + \kappa K_y \beta) - \lambda_\alpha \alpha' (K_x^2 \alpha + \kappa K_x \alpha) \end{aligned}$$

Which together with the constraints implies that $\lambda_\alpha - \lambda_\beta = 0$, let : $\lambda = \lambda_\alpha = \lambda_\beta$.

Considering the case that K_x and K_y are not full rank matrices, we explore the Gram-Schmidt orthogonolisation algorithm, which is described in [3], as it is equivalent to incomplete Cholesky decomposition. Complete decomposition of a kernel matrix is an expensive step and should be avoided with real world data. We slightly modify the Gram-Schmidt algorithm so it will use a precision parameter as a stopping criterion as shown in [1].

The Gram-Schmidt orthogonolisation works as follows; the projection is built up as the span of a subset of the projections of a set of m training examples. These are selected by performing a Gram-Schmidt orthogonalisation of the training vectors in the feature space. The algorithm used is summarised in Appendix A.

Setting via the Gram-Schmidt decomposition, where R is a lower triangular matrix, gives

$$\begin{aligned} K_x &\doteq R_x R_x' \\ K_y &\doteq R_y R_y' \end{aligned}$$

we can rewrite equation 2.1 and 2.2 as

$$\begin{aligned} R_x R_x' R_y R_y' \beta - \lambda (R_x R_x' R_x R_x' + \kappa R_x R_x') \alpha &= 0 \\ R_y R_y' R_x R_x' \alpha - \lambda (R_y R_y' R_y R_y' + \kappa R_y R_y') \beta &= 0 \end{aligned}$$

multiplying by R_x' with the first equation and similarly by R_y' with the second we obtain

$$R_x' R_x R_x' R_y R_y' \beta - \lambda R_x' (R_x R_x' R_x R_x' + \kappa R_x R_x') \alpha = 0 \quad (2.3)$$

$$R_y' R_y R_y' R_x R_x' \alpha - \lambda R_y' (R_y R_y' R_y R_y' + \kappa R_y R_y') \beta = 0. \quad (2.4)$$

Let Z be the new correlation matrix with the reduced dimensionality

$$\begin{aligned} Z_{xx} &= R_x' R_x \\ Z_{yy} &= R_y' R_y \\ Z_{xy} &= R_x' R_y \\ Z_{yx} &= Z_{xy}' \end{aligned}$$

and let $\tilde{\alpha}$ and $\tilde{\beta}$ be the new directions with reduced lengths

$$\begin{aligned} \tilde{\alpha} &= R_x' \alpha \\ \tilde{\beta} &= R_y' \beta. \end{aligned}$$

Substituting in equations 2.3 and 2.4 gives

$$Z_{xx} Z_{xy} \tilde{\beta} - \lambda Z_{xx} (Z_{xx} + \kappa I) \tilde{\alpha} = 0 \quad (2.5)$$

$$Z_{yy} Z_{yx} \tilde{\alpha} - \lambda Z_{yy} (Z_{yy} + \kappa I) \tilde{\beta} = 0 \quad (2.6)$$

Considering the Z_{xx} and Z_{yy} matrices as invertible we have

$$\begin{aligned} \tilde{\beta} &= \frac{(Z_{yy} + \kappa I)^{-1} Z_{yy}^{-1} Z_{yy} Z_{yx} \tilde{\alpha}}{\lambda} \\ &= \frac{(Z_{yy} + \kappa I)^{-1} Z_{yx} \tilde{\alpha}}{\lambda} \end{aligned}$$

and so substituting in equation 2.5 gives

$$Z_{xy} (Z_{yy} + \kappa I)^{-1} Z_{yx} \tilde{\alpha} = \lambda^2 (Z_{xx} + \kappa I) \tilde{\alpha} \quad (2.7)$$

Let S be the lower triangular matrix of the complete Cholesky decomposition of $Z_{xx} + \kappa I$ such that $(Z_{xx} + \kappa I) = SS'$ and let $\hat{\alpha} = S' \cdot \tilde{\alpha}$. So substituting in equation 2.7 gives

$$S^{-1} Z_{xy} (Z_{yy} + \kappa I)^{-1} Z_{yx} S^{-1'} \hat{\alpha} = \lambda^2 \hat{\alpha}$$

We are left with a generalised symmetric eigenproblem of the form $A\mathbf{x} = \lambda \mathbf{x}$.

3. EXPERIMENTS

In the following application the problem of learning semantics of multimedia content by combining image and text data is addressed. The synthesis is addressed by the kernel Canonical Correlation Analysis described in Section 2. We test the use of the derived semantic space in an image retrieval task that uses only image content. The aim is to allow retrieval of images from a text query but without reference to any labelling associated with the image. This can be viewed as a cross-modal retrieval task. We used the combined multimedia image-text web database, which was kindly provided by the authors of [6], where we are trying

to facilitate mate and content retrieval on a test set. The data was divided into three classes - Sport, Aviation and Paint-ball - 400 records each and consisted of jpeg images retrieved from the Internet with attached text. We randomly split each class into two halves, which were used as training and test data accordingly. The extracted features of the data were used the same as in [6] (detailed description of the features can be found in [6]): image HSV (Hue Saturation Values) colour, image Gabor texture and term frequencies in text.

We initially set the value of the regularisation parameter κ by running the kernel-cca with the association between image and text randomised. Let $\lambda(\kappa)$ be the spectrum without randomisation, the database with itself, and $\lambda_R(\kappa)$ be the spectrum with randomisation, the database with a randomised version of itself, (by spectrum it is meant that the vector whose entries are the eigenvalues). We would like to have the non-random spectrum as distant as possible from the randomised spectrum, as if the same correlation occurs for $\lambda(\kappa)$ and $\lambda_R\kappa$ then clearly over-fitting is taking place. Therefor we expect for $\kappa = 0$ (no regularisation) and let $\mathbf{j} = 1, \ldots, 1$ (the all ones vector) that we may have $\lambda(\kappa) = \lambda_R(\kappa) = \mathbf{j}$, since it is very possible that the examples are linearly independent. Though in practice only 50% of the examples are linearly independent but this does not effect the method of selection of κ . We choose κ so that the κ for which the difference between the spectrum of the randomised set is maximally different (in the two norm) from the true spectrum.

$$\kappa = argmax\|\lambda_R(\kappa) - \lambda(\kappa)\|$$

We find that $\kappa = 7$ and setting through a heuristic technique the Gram-Schmidt precision parameter to $\eta = 0.5$.

To perform the test image retrieval we compute the features of the images and text query using the Gram-Schmidt algorithm. Once we have obtained the features for the test query (text) and test images we project them into the semantic feature space using $\tilde{\beta}$ and $\tilde{\alpha}$ respectively. Now we can compare them using an inner product of the semantic feature vector. The higher the value of the inner product, the more similar the two objects are. Hence, we retrieve the images whose inner products with the test query are highest.

We compared the performance of our methods 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 text query and each training label or test image and each training image. For both methods we have used a Gaussian kernel, with $\sigma =$ max. distance$/20$ (we set the value of σ using a heuristic technique), for the image colour component and all experiments were an average of 10 runs, and the training and testing data was the same for both mate and content based approaches. For convenience we separate the mate-based and content-based approaches into the following subsections 3.1 and 3.2 respectively.

3.1. Mate-based retrieval

In the experiment we used the first 150 and 30 $\tilde{\alpha}$ eigenvectors and $\tilde{\beta}$ eigenvectors (corresponding to the largest eigenvalues). We computed the 10 and 30 images for which their semantic feature vector has the closest inner product with the semantic feature vector of the chosen text. A successful match is considered if the image that actually matched the chosen text is contained in this set. As shown in Table 1 we compare the success rate of the KCCA algorithm and GVSM over 10 and 30 image sets. In figure 1 we see the overall performance of the KCCA method against the GVSM for all possible image sets, this ad-hoc precision method is calculated by computing the overall average success rate for all possible queries in matching the exact image to text query. The success rate in Table 1 and Figure 1 is computed as follows

Image set	GVSM	KCCA (30)	KCCA (150)
10	8%	17.19%	59.5%
30	19%	32.32%	69%

Tab. 1. Success rate cross-results between kernel-cca & generalised vector space.

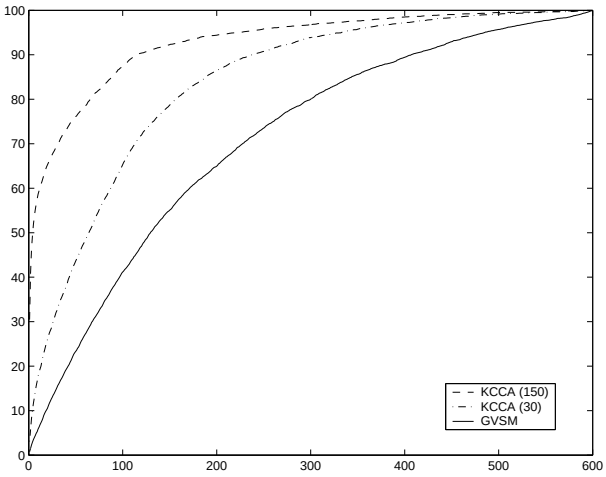

Fig. 1. Success plot for KCCA against GVSM (success (%) against image set size).

$$\text{success \% for image set } i = \frac{\sum_{j=1}^{600} count_j}{600} \times 100$$

where $count_j = 1$ if the exact matching image to the text query was present in the set, else $count_j = 0$.

We find that increasing the number of eigenvectors used will assist in locating the matching image to the query text. We speculate that this may be the result of added detail towards exact correlation in the semantic projection. Though we do not compute for all eigenvectors as this process would be expensive and the reminding eigenvectors would not necessarily add meaningful semantic information.

3.2. Content-based retrieval

In this experiment we used the first 30 and 5 $\tilde{\alpha}$ eigenvectors and $\tilde{\beta}$ eigenvectors (corresponding to the largest eigenvalues). We computed the 10 and 30 images for which their semantic feature vector has the closest inner product with the semantic feature vector of the chosen text. Success is considered if the images contained in the set are of the same label as the query text. In table 2 we compare the success

Image Set	GVSM	KCCA (30)	KCCA (5)
10	78.93%	85%	90.97%
30	76.82%	83.02%	90.69%

Tab. 2. Success rate cross-results between kernel-cca & generalised vector space.

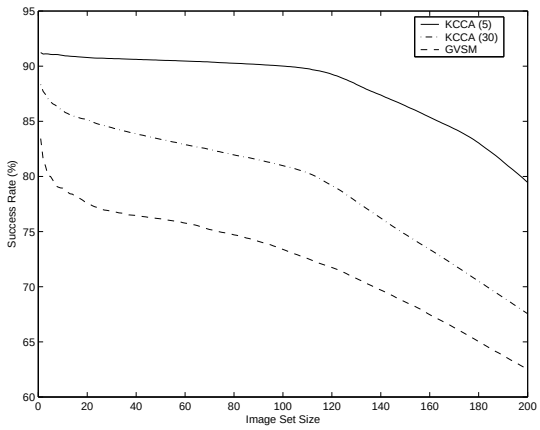

Fig. 2. Success plot for KCCA against GVSM

rate of the KCCA algorithm and GVSM over 10 and 30 image sets. In figure 2 we see the overall averaged success rate of the KCCA method against the GVSM per all queries for image sets $(1-200)$, as in the $200'th$ image set location the maximum of 200×600 of the same labelled images over all text queries can be retrieved (we only have 200 images per label). The success rate in Table 2 and Figure 2 is computed as follows

$$\text{success \% for image set } i = \frac{\sum_{j=1}^{600} \sum_{k=1}^{i} count_k^j}{i \times 600} \times 100$$

where $count_k^j = 1$ if the image k in the set is of the same label as the text query present in the set, else $count_k^j = 0$.

We observe that unlike the mate-based retrieval task (section 3.1) when we add eigenvectors to the semantic projection we will reduce the success of the content based retrieval. We speculate that this may be the result of unnecessary detail in the semantic projection and as the semantic information needed is contained in the first few eigenvectors. Hence a minimal selection of 5 eigenvectors is sufficient to obtain a high success rate.

Therefor we show that the KCCA can be adapted to two different types of problem, content and mate retrieval, by only changing the selection of eigenvectors used in the semantic projection. In both methods the success rate of the KCCA approach sharply outperforms the GVSM method.

3.3. Regularisation parameter

We next verify that the method of selecting the regularisation parameter κ a priori gives a value performed well. We randomly split each class into two halves which were used as training and test data accordingly, we keep this divided set for all runs. We set the value of the partial Gram-Schmidt orthogonolisation precision parameter via a heuristic method $\eta = 0.5$ and run over possible values κ where for each value we test its content-based and mate-based retrieval performance. Let $\hat{\kappa}$ be the previous optimal choice of the regularisation parameter $\hat{\kappa} = \kappa = 7$. As we define the new optimal value of κ by its performance on the testing set, we can say that this method is biased (loosely its cheating). Though we will show that despite this, the difference between the performance of the biased κ and our a priori $\hat{\kappa}$ is slight. In table 3 we compare the overall perfor-

κ	CB-KCCA (30)	CB-KCCA (5)
0	46.278%	43.8374%
$\hat{\kappa}$	83.5238%	91.7513%
90	88.4592%	**92.7936%**
230	**88.5548%**	92.5281%

Tab. 3. Overall success of Content-Based (CB) KCCA with respect to κ.

mance of the Content Based (CB) performance in respect to the different values of κ and in figures 3 and 4 we view the plotting of the comparison. We observe that the difference in performance between the a priori value $\hat{\kappa}$ and the

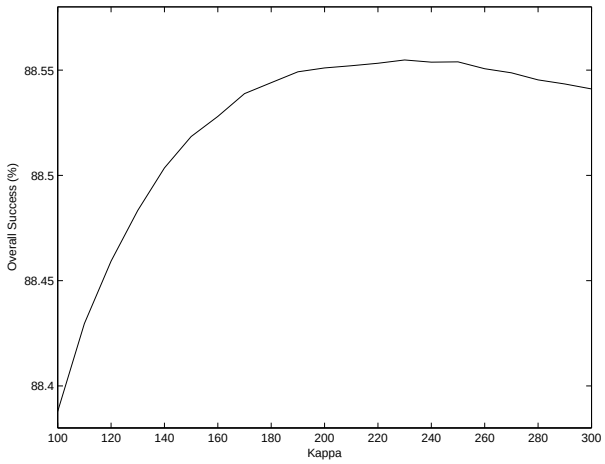

Fig. 3. Kapa selection over overall success for 30 eigenvectors.

new found optimal value κ for 5 eigenvectors is 1.0423% and for 30 eigenvectors is 5.031%. The more substantial increase in performance on the latter is due to the increase in the selection of the regularisation parameter, which compensates for the substantial decrease in performance (figure 1) of the content based retrieval, when high dimensional semantic feature space is used. In table 4 we compare the

κ	MB-KCCA (30)	MB-KCCA (150)
0	73.4756%	83.46%
$\hat{\kappa}$	84.75%	92.4%
170	$\mathbf{85.5086}\%$	92.9975%
240	$\mathbf{85.5086}\%$	93.0083%
430	85.4914%	$\mathbf{93.027}\%$

Tab. 4. Overall success of Mate-Based (MB) KCCA with respect to κ.

overall performance of the Mate-Based (MB) performance with respect to the different values of κ and in figures 5 and 6 we view s plot of the comparison. We observe that in this case the difference in performance between the a priori value $\hat{\kappa}$ and the new found optimal value κ is for 150 eigenvectors 0.627% and for 30 eigenvectors is 0.7586%.

Our observed results support our proposed method for selecting the regularisation parameter κ in an a priori fashion, since the difference between the actual optimal κ and the a priori $\hat{\kappa}$ is very slight.

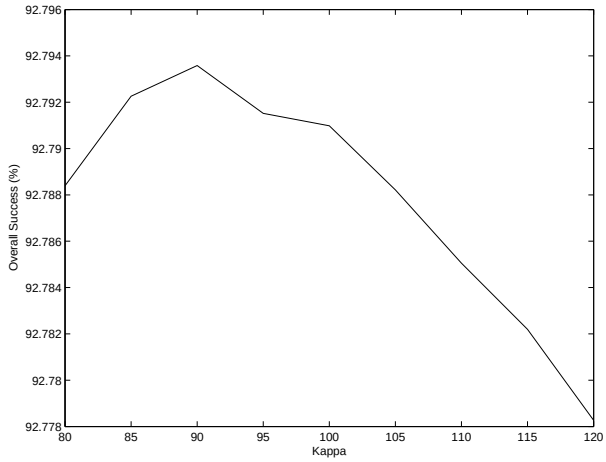

Fig. 4. Kapa selection over overall success for 5 eigenvectors.

4. CONCLUSIONS

Through this paper we have established a general approach to retrieving images based solely on their content. This is then applied to content-based and mate-based retrieval. Experiments show that image retrieval can be more accurate than with the Generalised Vector Space Model. We demonstrate that one can choose the regularisation parameter κ a priori that it gives a value that performs well in very different regimes. Hence we have come to the conclusion that kernel Canonical Correlation Analysis is a powerful tool for image retrieval via content. In the future we will extend our experiments to other data collections as well as to kernel canonical correlation analysis of multiple views of underlying semantic objects.

A. PARTIAL GRAM-SCHMIDT ORTHOGONOLISATION

Given a kernel K and precision parameter η:

Initializations:

$m =$ size of K
$size$ and $index$ are a vector with the same length as k
$feat$ a zeros matrix equal to the size of k
for $i = 1$ to m do
    norm2$[i] = K_{ii}$;

Algorithm:

$j = 1$;
while $\sum_{i=j}^{m} norm2[i] > \eta$ do
    $i_j = argmax_i$(norm2$[i]$);
    index$[j] = i_j$;

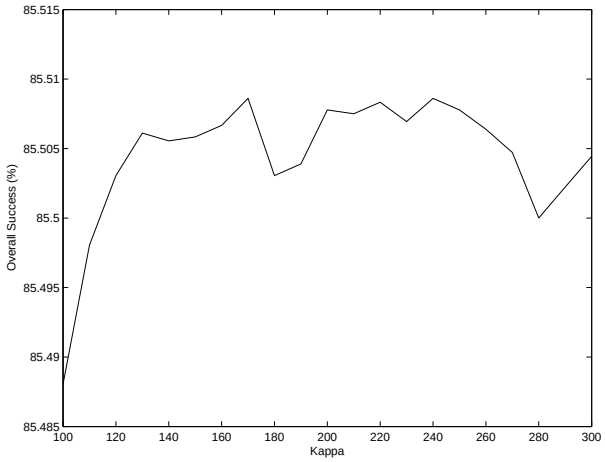

Fig. 5. Kapa selection over overall success for 30 eigenvectors.

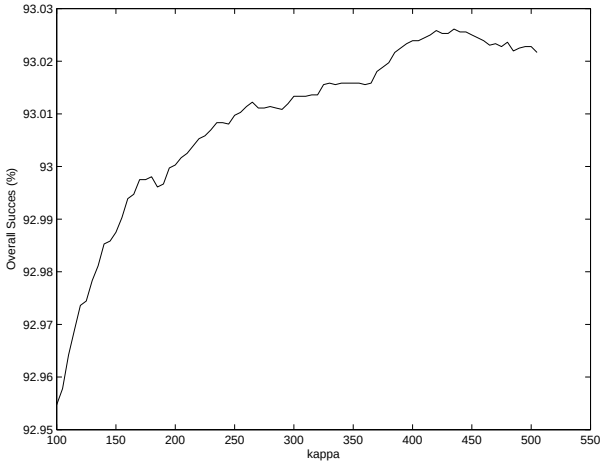

Fig. 6. Kapa selection over overall success for 150 eigenvectors.

size$[j] = \sqrt{norm2[i_j]}$;
for $i = 1$ to m do
feat$[i,j] = \frac{\left(K_{i,i_j} - \sum_{t=1}^{j-1} feat[i,t] \cdot feat[i_j,t]\right)}{size[j]}$;
norm2$[i]$ = norm2$[i]$− feat$(i,j)\cdot$ feat(i,j);
end;
$j = j + 1$;
end;
return feat$[i,j]$ as the $j-$th feature of input i;

Output:
Features satisfying $\|K - feat \cdot feat'\| \leq \eta$

To classify a new example:
for $j = 1$ to T
newfeat$[j]$ = $(K_{i,i_j} - \sum_{t=1}^{j-1}$ newfeat$[j,t] \cdot$ feat$[i_j,t])$
/size$[j]$;
end;

B. REFERENCES

[1] Francis Bach and Michael Jordan. Kernel independent component analysis. *Journal of Machine Leaning Research*, 3:1–48, 2002.

[2] Magnus Borga. Canonical correlation *a tutorial*, 1999.

[3] Nello Cristianini, John Shawe-Taylor, and Huma Lodhi. Latent semantic kernels. In Caria Brodley and Andrea Danyluk, editors, *Proceedings of ICML-01, 18th International Conference on Machine Learning*, pages 66–73. Morgan Kaufmann Publishers, San Francisco, US, 2001.

[4] 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 (to appear), 2003.

[5] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:312–377, 1936.

[6] 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.

[7] Yong Rui, Thomas S. Huang, and Shih-Fu Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communications and Image Representation*, 10:39–62, 1999.

[8] Alexei Vinokourov, David R. Hardoon, and John Shawe-Taylor. Learning the semantics of multimedia content with application to web image retrieval and classification. In *Proceedings of Fourth International Symposium on Independent Component Analysis and Blind Source Separation*, Nara, Japan, 2003.

[9] Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini. Inferring a semantic representation of text via cross-language correlation analysis. In *Advances of Neural Information Processing Systems 15 (to appear)*, 2002.