LEARNING UNIFIED SPARSE REPRESENTATIONS FOR MULTI-MODAL DATA

Kaiye Wang, Wei Wang, Liang Wang[*]

Center for Research on Intelligent Perception and Computing,
National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China
{kaiye.wang, wangwei, wangliang}@nlpr.ia.ac.cn

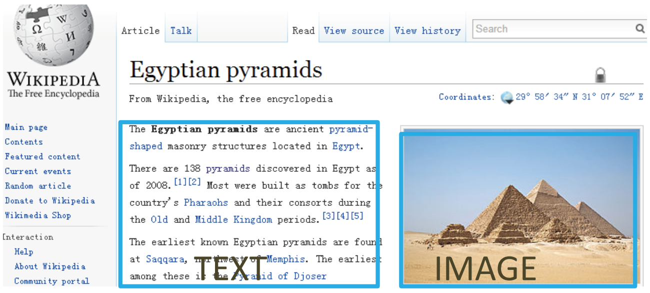
ABSTRACT

Cross-modal retrieval has become one of interesting and important research problem recently, where users can take one modality of data (e.g., text, image or video) as the query to retrieve relevant data of another modality. In this paper, we present a Multi-modal Unified Representation Learning (MURL) algorithm for cross-modal retrieval, which learns unified sparse representations for multi-modal data representing the same semantics via joint dictionary learning. The ℓ_1-norm is imposed on the unified representations to explicitly encourage sparsity, which makes our algorithm more robust. Furthermore, a constraint regularization term is imposed to force the representations to be similar if their corresponding multi-modal data have must-links or to be far apart if their corresponding multi-modal data have cannot-links. An iterative algorithm is also proposed to solve the objective function. The effectiveness of the proposed method is verified by extensive results on two real-world datasets.

***Index Terms*—** Cross-modal retrieval, unified representation learning, joint dictionary learning, multi-modal data

1. INTRODUCTION

In recent years, there has been a massive explosion of multimedia data on the web. Multimedia information comes through multiple input channels, and different types of data representing the same semantics usually exist together. For example, images or videos are often associated with text or tags in the webpage (as shown in Figure 1). We refer to such different types of data as *multi-modal data*. With the rapid growth of multimedia data, it is very desirable to support similarity search across the multi-modal data (called *cross-modal retrieval*), e.g., the retrieval of textual documents given a query image or vice versa. However, we cannot measure the similarity between different modalities of data directly due to *heterogeneity gap*. Hence, the key of the cross-modal retrieval is to reduce such heterogeneity gap.



Egyptian pyramids

From Wikipedia, the free encyclopedia

The **Egyptian pyramids** are ancient pyramid-shaped masonry structures located in Egypt.

There are 138 pyramids discovered in Egypt as of 2008.[1][2] Most were built as tombs for the country's Pharaohs and their consorts during the Old and Middle Kingdom periods.[3][4][5]

The earliest known Egyptian pyramids are found at Saqqara, northwest of Memphis. The earliest among these is the Pyramid of Djoser

Fig. 1. An example webpage including mult-modal data (text and image) from Wikipedia.

Recently, several subspace learning methods are proposed to reduce the heterogeneity gap, such as Canonical Correlation Analysis (CCA) [1, 2, 3] and Partial Least Squares (PLS) [4, 5]. In [6], Rastegari et al. apply CCA to model correlations between text and image, which learns two projections to map texts and images into a latent subspace. Sharma and Jacobs [5] use PLS to linearly map images of different modalities to a latent linear subspace in which they are highly correlated. Recently, Sharma et al. [7] propose a supervised version of CCA for cross-modal retrieval, referred to as Generalized Multiview Analysis (GMA). Lu et al. [8] and Wu et al. [9] formulate the cross-modal retrieval as a problem of learning to rank. However, supervised methods [7, 8, 9, 10, 11] generally need class information or rank lists, which are very expensive to obtain in the real-world applications.

As mentioned above, different modalities of data representing the same sematics usually exist together, which form a multi-modal document. A practical way for modeling cross-modal correlations is to explore the co-occurrence information and the links between multi-modal documents. Motivated by representation learning [12, 13], different modalities of data in a multi-modal document should have a unified representation. Furthermore, if multi-modal documents have must-links, their representations should be similar, and if they have cannot-links, their representations should be far away from each other.

[*] The corresponding author

978-1-4799-8339-1/15/$31.00 ©2015 IEEE

Based on the above considerations, this paper presents a joint dictionary learning method for cross-modal retrieval, named Multi-modal Unified Representation Learning (MURL), which learns unified representations for different modalities of data which represent the same semantics. The ℓ_1-norm and a constrained regularization term are further imposed on the unified representations, which makes our algorithm more robust and be able to capture more relationships. Different from previous works which usually map different modalities of data into different latent spaces of the same dimension, our method converts different modalities of data into a unified representation, which is more interpretable. The experimental results show the effectiveness of the proposed method when applied to the cross-modal retrieval task.

2. THE OVERVIEW OF OUR APPROACH

Suppose that we have a collection of data from M different modalities, i.e., $X^p = [\mathbf{x}_1^p, \mathbf{x}_2^p, ..., \mathbf{x}_N^p] \in \mathbb{R}^{d_p \times N}, p = 1, 2, ..., M$, where N is the number of samples, and d_p is the feature dimension. Let $x_i = \{\mathbf{x}_i^1, \mathbf{x}_i^2, ..., \mathbf{x}_i^M\}$ denote a multi-modal document, in which the data from M modalities represent the same underlying content or objects. Given a query from one modality, the goal of the cross-modal retrieval is to return the top k closest matches in another modality.

2.1. Formulation

The purpose of our approach is to learn unified representations $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N] \in R^{K \times N}$ for multi-modal data via joint dictionary learning. $\mathbf{D}^p = [\mathbf{d}_1^p, \mathbf{d}_2^p, ..., \mathbf{d}_K^p] \in R^{d_p \times K}$ denotes the learned dictionary from the p-th modality data and K is the size of the dictionary. The objective function for learning unified representations is defined as follows:

$$\begin{aligned} &\min_{\mathbf{D}^p,\mathbf{Z}} \sum_{p=1}^{M} \|\mathbf{X}^p - \mathbf{D}^p\mathbf{Z}\|_F^2 + \lambda_1\|\mathbf{Z}\|_1 \\ &+ \lambda_2\left(\sum_{(x_i,x_j)\in S}\|\mathbf{z}_i - \mathbf{z}_j\|^2 - \alpha\sum_{(x_i,x_j)\in D}\|\mathbf{z}_i - \mathbf{z}_j\|^2\right) \\ &s.t. \ \|\mathbf{d}_k^p\|_2^2 \le 1, \forall p = 1...M, k = 1...K \end{aligned} \tag{1}$$

where S and D denote the similarity constraints and dissimilarity constraints between different multi-modal documents, respectively. As shown in Eq.(1), the ℓ_1-norm is conducted as a penalty to explicitly encourage sparsity on the unified representations, and the third term is a constraint regularization term, which ensures that the representations are similar if their corresponding multi-modal documents belong to same class (i.e., have must-links) while the representations are far away from each other if their corresponding multi-modal documents are of different classes (i.e., have cannot-links).

The constraint regularization term of the objective function can be reformulated as:

$$\Omega(\mathbf{Z}) = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\|\mathbf{z}_i - \mathbf{z}_j\|^2 = tr(\mathbf{Z}\mathbf{L}\mathbf{Z}^T) \tag{2}$$

where

$$w_{ij} = \begin{cases} 1, & (x_i, x_j) \in S \\ -\alpha, & (x_i, x_j) \in D \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

and $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix. $\mathbf{D}$ is a diagonal matrix, $\mathbf{D}_{ii} = \sum_j w_{ij}$.

2.2. Optimization

As shown in Eq.(1), the objective function is non-convex with respect to $\mathbf{D}^p$ and $\mathbf{Z}$, but it is convex with respect to $\mathbf{D}^p$ while fixing $\mathbf{Z}$ and vice versa. Therefore, we present an iterative algorithm to solve the objective function.

Algorithm 1 The optimization of updating the unified representations.

```
Input: Initialized $\mathbf{Z}_0$, and the maximum iteration number $q$;
Output: $\mathbf{Z}$;
$\mathbf{Z}_1 = \mathbf{Z}_0, t_{-1} = 0, t_0 = 1$;
for $i = 1$ to $q$ do
    $\mu_i = (t_{i-2} - 1)/t_{i-1}, \mathbf{S}_i = \mathbf{Z}_i + \mu_i(\mathbf{Z}_i - \mathbf{Z}_{i-1})$;
    while true do
        Compute $\mathbf{Z}^* = \arg\min P(\mathbf{Z})$ where;
            $P(\mathbf{Z}) = f(\mathbf{S}) + \|\mathbf{Z}\|_1 + \langle \nabla f(\mathbf{S}), \mathbf{Z} - \mathbf{S} \rangle + \frac{\gamma_i}{2}\|\mathbf{Z} - \mathbf{S}\|_F^2$;
        if $f(\mathbf{Z}^*) \le P(\mathbf{Z}^*)$ then break the while loop
            else $\gamma_i = \gamma_i \times 2$;
        end if
    end while
    $\mathbf{Z}_{i+1} = \mathbf{Z}^*, \gamma_{i+1} = \gamma_i$
    if stopping criteria satisfied then break the for loop
    $t_i = \frac{1+\sqrt{1+4t_{i-1}^2}}{2}$
end for
$\mathbf{Z} = \mathbf{Z}_{i+1}$
```

The optimization of unified representations. Firstly, we optimize $\mathbf{Z}$ with fixed $\mathbf{D}^p$, the problem in (1) becomes

$$\min_{\mathbf{Z}} \sum_{p=1}^{M} \|\mathbf{X}^p - \mathbf{D}^p\mathbf{Z}\|_F^2 + \lambda_1\|\mathbf{Z}\|_1 + \lambda_2 tr(\mathbf{Z}\mathbf{L}\mathbf{Z}^T) \tag{4}$$

In (4), the ℓ_1-norm is not differentiable at zero and gradient doses not exist, so the gradient descent method is not available to solve the formulation. Here, we adopt the accelerated gradient method [14, 15] to solve the formulation. The key idea is to solve the proximal operator associated to ℓ_1-norm. Denote the smooth part in Eq.(4) as:

$$f(\mathbf{Z}) = \sum_{p=1}^{M} \|\mathbf{X}^p - \mathbf{D}^p\mathbf{Z}\|_F^2 + \lambda_2 tr(\mathbf{Z}\mathbf{L}\mathbf{Z}^T) \tag{5}$$

The optimization of updating unified representations is described in Algorithm 1 and we adopt the implementation in the MALSAR package [16].

The optimization of dictionaries. Fixing unified representations $\mathbf{Z}$, the problem in (1) becomes the least squares problem with quadratic constraints which can be solved using Lagrange dual [17]. We compute $\mathbf{D}^p$ alternately when other dictionaries are fixed.

$$\min \|\mathbf{X}^p - \mathbf{D}^p\mathbf{Z}\|_F^2 \quad s.t. \|\mathbf{d}_k^p\|_2^2 \leq 1, \forall k = 1...K \tag{6}$$

Consider the Lagrangian:

$$g(\mathbf{D}^p, \mu) = \|\mathbf{X}^p - \mathbf{D}^p\mathbf{Z}\|_F^2 + \sum_{i=1}^{N} \mu_i \left(\|\mathbf{d}_k^p\|_2^2 - 1\right) \tag{7}$$

where $\mu_i \geq 0$ is the Lagrange multipliers. Letting the derivative of (7) with respect to $\mathbf{D}^p$ equal to zero, the analytical solution of $\mathbf{D}^p$ can be computed as:

$$\mathbf{D}^p = \mathbf{X}^p\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \mathbf{\Theta})^{-1} \tag{8}$$

where $\mathbf{\Theta}$ is a diagonal matrix $\mathbf{\Theta} = diag(\mu)$ and it is obtained by optimizing the Lagrange dual problem as follows:

$$\min_{\mathbf{\Theta}} tr \left(\mathbf{X}^p\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \mathbf{\Theta})^{-1}(\mathbf{X}^p\mathbf{Z}^T)^T + \mathbf{\Theta}\right) \tag{9}$$

The Lagrange dual in (9) can be solved by using Newton's method or conjugate gradient.

The overall optimization of our approach is summarized in Algorithm 2.

Algorithm 2 The overall optimization

```
Input: $\mathbf{X}^p, p = 1,..,M$, initialized dictionaries $\mathbf{D}^p, p = 1,..,M$ via K-SVD [18];
Output: The learned dictionaries $\mathbf{D}^p, p = 1,..,M$;
Utilize the must-links and cannot-links to construct $\mathbf{L}$;
while not convergent do
    Fix dictionaries $\mathbf{D}^p, p = 1,..,M$, update the unified representations $\mathbf{Z}$ via Algorithm1;
    Fix $\mathbf{Z}$, compute the dictionary $\mathbf{D}^p$ one by one with other dictionaries fixed by solving (6).
end while
```

2.3. Extension to Out-of-Sample

For different modalities of data in the training dataset, we have obtained their corresponding unified representations. To compute the representations of new data, the learned dictionaries are exploited.

Assume that given a new data $\mathbf{x}_t^p$ from the p-th modality, using the learned dictionary $\mathbf{D}^p$, we can obtain the corresponding representation as follows:

$$\min_{\mathbf{z}_t} \|\mathbf{x}_t^p - \mathbf{D}^p\mathbf{z}_t\|_F^2 + \lambda_1\|\mathbf{z}_t\|_1 \tag{10}$$

We solve the above problem by using the SLEP (Sparse Learning with Efficient Projections) package [19].

Methods	Image query	Text query	Average
PCA	0.1443	0.1093	0.1258
PLS	0.1471	0.1329	0.1400
CCA	0.1580	0.1399	0.1489
MURL$_0$	0.1808	0.1516	0.1662
MURL	**0.1884**	**0.1595**	**0.1739**

Table 1. Comparison of MAP on the Wiki dataset.

Methods	Image query	Text query	Average
PCA	0.0903	0.0756	0.0829
PLS	0.1204	0.1101	0.1152
CCA	0.1177	0.1029	0.1103
MURL$_0$	0.1463	0.1268	0.1365
MURL	**0.1629**	**0.1357**	**0.1493**

Table 2. Comparison of MAP on the NUS-WIDE dataset.

3. EXPERIMENTAL RESULTS

3.1. Data Sets

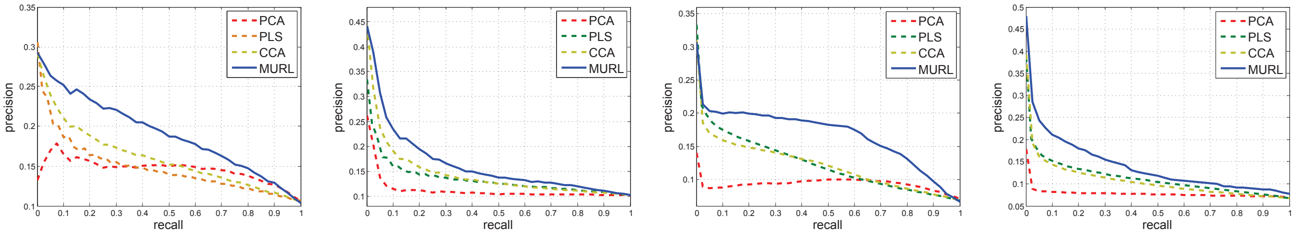
The *Wiki image-text* dataset [6] consists of 2866 image-text pairs. In each pair, the text is an article describing people, places or some events and the image is closely related to the content of the article. Each pair is annotated with a label from 10 semantic classes. Each image is represented by an 1000-dimensional bag-of-visual-words vector, and each text is represented with a 5000-dimensional vector by word frequency. We take 2146 image-text pairs as the training set and the remaining 720 image-text pairs as the testing set.

Another dataset used here is *NUS-WIDE* dataset [20]. Images are downloaded from Flickr and each image is associated with user tags. We select the 15 largest concepts with total 9000 images (600 images for each concept). The images are represented with 500-dimensional bag-of-visual-words vectors, and the text tags are represented with 1000-dimensional tag occurrence feature vectors. We take 50% of the data as the training set and the remaining 50% as the testing set.

3.2. Experimental Settings

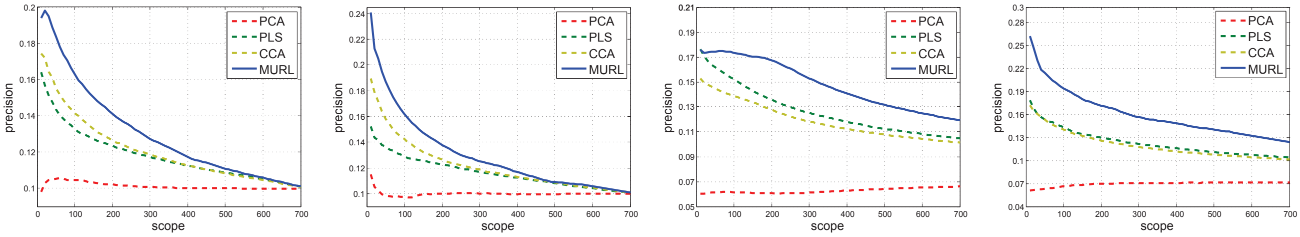
Experiments are performed with respect to two cross-modal retrieval tasks: (1) Image query vs. Text database, (2) Text query vs. Image database. We compare the proposed MURL algorithm with several related methods, namely, PCA, PLS [5] and CCA [3, 6]. Since MURL is semi-supervised, we don't compare with supervised methods. For our method, the must-links and the cannot-links are determined by class labels with only 1% randomly selected entries observed. The setting of λ_1, λ_2 and α is 0.1, 0.001 and 0.1, respectively.

The *mean average precision (MAP)* [6] is used to evaluate



(a) Wiki Image query (b) Wiki Text query (c) NUS Image query (d) NUS Text query

Fig. 2. Precision-recall curves on two datasets for two cross-modal retrieval tasks.



(a) Wiki Image query (b) Wiki Text query (c) NUS Image query (d) NUS Text query

Fig. 3. Precision-scope curves on two datasets for two cross-modal retrieval tasks.

the overall performance of the algorithms. Besides the MAP, we also use both *precision-scope curve* [21] and *precision-recall curve* [6] to evaluate the effectiveness of different methods. The scope is specified by the number (k=10 to 700) of top-ranked documents presented to the users.

3.3. Performance Comparisons

Table 1 and Table 2 show the performance in terms of MAP for the Wiki dataset and the NUS-WIDE dataset, respectively. To verify the role of the constraint regularization term, we add a light version of our MURL, namely MURL$_0$, by setting the parameter $\lambda_2 = 0$. Figure 2 and Figure 3 show the precision-recall curves and the precision-scope curves on the two datasets, respectively.

From the experimental results, we can make the following observations:

1) As shown in Table 1 and Table 2, The proposed MURL$_0$ without the constraint regularization term achieves better performance than CCA, PLS and PCA on both the Wiki dataset and the NUS-WIDE dataset. The reason is that different from CCA and PLS, the proposed MURL$_0$ converts the data from different modalities representing the same semantics into a unified representation, which is more effective for reducing the heterogeneity gap. In general, CCA and PLS performs better than PCA, This is because that both CCA and PLS exploit the co-occurrence information (or pairwise information) in the multi-modal data to model the cross-modal correlations, which naturally benefits cross-modal retrieval.

2) The MURL algorithm achieves superior performance over the MURL$_0$ algorithm in terms of MAP on both of the evaluated datasets. This observation reveals that the constraint regularization term is truly helpful for learning better unified representations, which further improves the performance by exploring the must-links and cannot-links between multi-modal documents.

3) Both precision-recall and precision-scope curves also validate the superiority of MURL. The proposed MURL algorithm gets higher precision than other methods for both image query and text query. It shows that MURL is more effective to precisely find the top k matches.

4. CONCLUSION

In this paper, we have proposed a cross-modal retrieval approach, named Multi-modal Unified Representation Learning, which learns unified sparse representations for multi-modal data via joint dictionary learning. We further present an iterative algorithm to solve the corresponding optimization problem. We have demonstrated the effectiveness of the proposed method on two benchmark datasets.

5. ACKNOWLEDGMENTS

This work is jointly supported by National Basic Research Program of China (2012CB316300), and National Natural Science Foundation of China (61175003, 61202328, 61420106015, U1435221).

6. REFERENCES

[1] D.R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: an overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[2] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *TPAMI*, vol. 29, no. 6, pp. 1005–1018, 2007.

[3] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *TPAMI*, vol. 36, no. 3, pp. 521–535, 2014.

[4] R. Rosipal and N. Kramer, "Overview and recent advances in partial least squares," *Subspace, Latent Structure and Feature Selection*, pp. 34–51, 2006.

[5] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," *In CVPR*, pp. 593–600, 2011.

[6] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," *In ACM MM*, pp. 251–260, 2010.

[7] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: a discriminative latent space," *In CVPR*, pp. 2160–2167, 2012.

[8] X. Y. Lu, F. Wu, S. L. Tang, Z. F. Zhang, X. F. He, and Y. T. Zhuang, "A low rank structural large margin method for cross-modal ranking," *In ACM SIGIR*, pp. 433–442, 2013.

[9] F. Wu, X. Y. Lu, Z. F. Zhang, S. C. Yan, Y. Rui, and Y. T. Zhuang, "Cross-media semantic representation via bi-directional learning to rank," *In ACM MM*, pp. 877–886, 2013.

[10] K. Y. Wang, R. He, W. Wang, L. Wang, and T. N. Tan, "Learning coupled feature spaces for cross-modal matching," *In ICCV*, pp. 2088–2095, 2013.

[11] Y. T. Zhuang, Y. F. Wang, F. Wu, Y. Zhang, and W. M. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *AAAI*, 2013, pp. 1070–1076.

[12] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011.

[13] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *NIPS*, 2012, pp. 2231–2239.

[14] A. Nemirovski, "Efficient methods in convex programming," *Lecture Notes*, 2005.

[15] Y. Nesterov, "Introductory lectures on convex optimization: A basic course," *Springer*, 2004.

[16] J. Zhou, J. Chen, and J. Ye, "MALSAR: Multi-task learning via structural regularization. arizona state university," *Arizona State University*, 2011, `http://www.public.asu.edu/~jye02/Software/MALSAR`.

[17] H. Lee, A. Battle, R. Raina, and Andrew Y. Ng, "Efficient sparse coding algorithms," in *NIPS*, 2006.

[18] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing over-complete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[19] J. Liu, S. Ji, and J. Ye, "SLEP: Sparse learning with efficient projections," *Arizona State University*, 2009, `http://www.public.asu.edu/~jye02/Software/SLEP`.

[20] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world web image database from national university of singapore," in *ACM International Conference on Image and Video Retrieval*, 2009.

[21] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos, "Bridging the gap: query by semantic example," *IEEE TMM*, vol. 9, no. 5, pp. 923–938, 2007.