CROSS MODAL METRIC LEARNING WITH MULTI-LEVEL SEMANTIC RELEVANCE

Yan Hua[⋆] *Shuhui Wang*[†] *Zhicheng Zhao*[⋆] *Qingming Huang*[‡†] *Anni Cai*[⋆]

[⋆]Beijing University of Posts and Telecommunications, Beijing, China
[†]Key Lab of Intell. Info. Process. (CAS), Inst. of Comput. Tech., CAS, Beijing, China
[‡]University of Chinese Academy of Sciences, Beijing, China

ABSTRACT

The *Mahalanobis* metric learning is an effective tool for constructing semantic consistent distance among data in single modal data analysis. However, distance metric learning is a more challenging issue for cross modal data, where less attention has been paid in previous studies. In this paper, we propose **C**ross m**O**dal **L**arge m**A**rgin metric lea**R**ning (COLAR) with multi-level semantic relevance. With large margin principle, we model different levels of the semantic relations across modalities, e.g., the one-to-one correspondence and intra-class relation, while traditional correlation learning approaches (such as CCA and its variants) can only handle the one-to-one correspondence or treat them indiscriminatively. As a result, the distances of multi-level relevance among cross modal data are optimized based on a regularized learning framework. Promising performance is achieved on cross modal retrieval, i.e., image-to-text retrieval and text-to-image retrieval.

Index Terms— cross modal metric learning, semantic relevance, large margin learning

1. INTRODUCTION

Co-occurred data from different modalities usually deliver the same semantic information, e.g., image and its surrounding text description on web, video frames and its accompanied voice messages. People can associate the different data types subconsciously with each other and understand their inherent semantics without much efforts. However, it is hard for computer to capture this coincident semantic information since data are represented in heterogeneous feature spaces.

Traditional content-based image retrieval methods with textual information train separate classifiers for each word [1] [2] and combine their outputs heuristically or by word semantic structure [3] for multi-word queries. Siddiquie et al. [4] propose a multi-attribute retrieval method which explicitly models the correlations that are present between the attributes, however attribute detectors are trained independently. With defined structured object queries, [5] develops a learning framework to jointly consider object classes and their relations, where the relations are modeled by latent variables. These separately trained models are not suited for describing the sophisticated semantic relation among multiple content modalities, which is an important requirement for further progress in cross modal retrieval [6].

(a) Sport (b) Sport (c) Art

Fig. 1. The images (a) and (b) belong to sport category, (a) is basketball star "Michael Jordan", and (b) is ice hockey star "Martin Brodeur". (c) belongs to art category. When searching with abundant textual information about Jordan, (a) is the perfectly matched item, (b) is somehow relevant, and (c) is totally irrelevant.

Canonical Correlation Analysis (CCA) and its variations, which conduct multi-view feature extraction and dimensionality reduction by cross modal correlation maximization, have been proved to be the workhorse in cross modal retrieval [6] [7]. The category information of single modality is further incorporated by Local discrimination CCA (LDCCA) [8]. Locality preserving CCA (LPCCA) [9] is proposed to incorporate local structure information of the single modality into CCA. Sun et al. [10] propose discriminant CCA utilizing category information of multi-views and perform discriminative feature extraction. More recently, Sharma et al. [7] propose a general multi-view feature extraction approach called Generalized Multi-view Analysis (GMA), which is also a supervised extension of CCA with category information. However,

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, National Natural Science Foundation of China (61025011, 61332016, 61303160, 61390511, 90920001, 61101212, 61372169), 863 program of China (2014AA015202, 2012AA012505, 2012AA012504), National Key Technology R&D Program (2012BAH63F00, 2012BAH41F03), and the Fundamental Research Funds for the Central Universities.

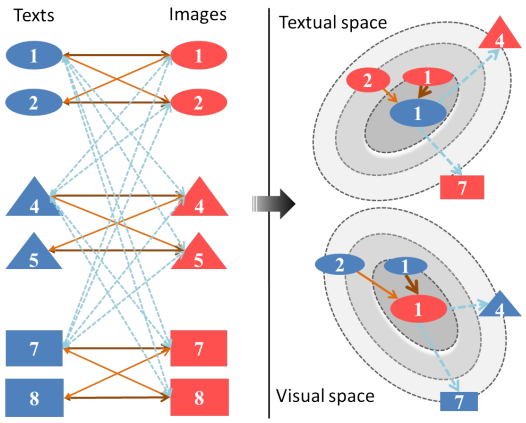


Fig. 2. Illustration of our approach. Data of different modalities and different categories are represented with nodes in different colors and shapes. The dark brown lines between nodes imply one-to-one correspondence (strongly relevant), light brown lines for intra-class relation (relevant), and blue lines for inter-class relation (irrelevant). The learned cross modal metrics aim to meet the semantic relation constraints, as shown in the right part of the figure.

these methods are not capable of modeling different levels of the semantic relations across modalities.

Intuitively, a good information retrieval system should present relevant documents high in the ranking, with less relevant documents following below [11]. Such different levels of semantic relations can also be observed in cross modal data involving images and texts. For example, see Fig.1, both images in Fig.1(a) and (b) belong to the "sports" category. However, their associated textual description cannot be the complementary description of each other. If queried with "Michael Jordan", the retrieved intra-class image (e.g., Fig.1(b)) is not as good as the true corresponding image (i.e., Fig.1(a)). The true one is considered most relevant to the query, thus to deliver the strongest semantic correlation, while the intra-class one only delivers a certain level of semantic correlation. Unfortunately, traditional approaches can not learn such kinds of multi-level relevance by explicitly modeling semantic relation difference among the cross modal data, therefore they cannot well meet the quality judgment for cross modal information retrieval [11].

In this paper, we propose COLAR, a novel cross modal metric learning method which models the multi-level semantic relation among cross modal data. Inspired by LMNN [12], we encode the one-to-one correspondence, the intra-class and inter-class relations across modalities with a unified large margin empirical loss minimization paradigm, as shown in Fig.2. By imposing penalties on the complexity of the learned metric, the objective function is a standard convex function which can be efficiently optimized with gradient descent. COLAR achieves promising performance on the benchmark dataset [6] and outperforms the state-of-the-art approaches.

2. EMPIRICAL LOSS

We are given $\mathbb{D} = \{x_i, y_i, c_i\}_{i=1}^n$, where $x_i \in R^{d_x}$ denotes the i-th training data from X, $y_i \in R^{d_y}$ from Y, and $c_i \in \{1, 2, \ldots, k\}$ denotes the category index of the i-th training pair. We want to learn two linear transformations parametrized by $\mathbf{M} \in \mathbb{R}^{d*d_x}$ and $\mathbf{N} \in \mathbb{R}^{d*d_y}$ to project X and Y into a unified d dimensional space where the learned cross modal distance $D(x_i, y_j) = (\mathbf{M}x_i - \mathbf{N}y_j)^T(\mathbf{M}x_i - \mathbf{N}y_j)$ better meets the multi-level semantic relation constraints.

Ideally, examples in X modality should be ranked according to the semantic relation to the query in Y modality with the learned transformations, and vise versa. We define three kinds of label sets, i.e., the one-to-one correspondence set $S_1 = \{s_{1i} = \{x_i, y_i\} \mid i = 1, 2, \ldots, n\}$, the intra-class relation set $S_2 = \{s_{2i} = \{x_i, y_j\} \mid c_j = c_i, j \neq i\}$ and the inter-class relation set $S_3 = \{s_{3i} = \{x_i, y_j\} \mid c_j \neq c_i\}$. The three label sets deliver multi-level semantic relevance information, where S_1 carries the strongest semantic relevance information, S_2 carries category level semantic relation and the cross modal data pairs in S_3 are totally irrelevant.

In mathematical terms, imposters are defined by a simple inequality. For an input x_i and its most relevant item y_i, an impostor is any input y_j in s_{2i} and s_{3i} such that $D(x_i, y_j) \leqslant D(x_i, y_i) + \Delta(x_i, y_j)$. In other words, an impostor y_j is any input of different semantic level that invades the perimeter plus certain margin defined by x_i and y_j. Of course, the margin of s_{3i} should be larger than that of s_{2i}. However, since all these constraints may not be perfectly satisfied for all the training data, we introduce slack variables on each term. The relaxed constraints are formulated as follows:

$$D(x_i, y_j) - D(x_i, y_i) \geq \Delta(x_i, y_j) - \varepsilon_{ij}, \varepsilon_{ij} \geq 0, \qquad (1)$$

where $\Delta(x_i, y_j)$ is determined by the category relation between examples x_i and y_j. For example, $\Delta(x_i, y_j) = 1$ if $c_i = c_j$ and $\Delta(x_i, y_j) = 2$ if $c_i \neq c_j$. It is similar with the cost of the predicted ranking label and the correct label for a sample in ranking SVM [11]. The linear transformations $\mathbf{M}$ and $\mathbf{N}$ are learned by minimizing the following loss function:

$$\begin{aligned} \mathbf{M}, \mathbf{N} = \underset{\mathbf{M},\mathbf{N}}{\operatorname{argmin}} L(\mathbf{M}, \mathbf{N}) = \underset{\mathbf{M},\mathbf{N}}{\operatorname{argmin}} c \sum_{S_1} D(x_i, y_i) \\ + \alpha \sum_i \sum_{s_{2i}} \varepsilon_{ij} + \beta \sum_i \sum_{s_{3i}} \varepsilon_{ij} \qquad (2) \\ \text{s.t. } D(x_i, y_j) - D(x_i, y_i) \geq \Delta(x_i, y_j) - \varepsilon_{ij}, \varepsilon_{ij} \geq 0. \end{aligned}$$

3. STRUCTURE RISK MINIMIZATION

By concatenating x_i and y_j, and the linear transformation metrics $\mathbf{M}$ and $\mathbf{N}$ in $D(x_i, y_j)$, we obtain the following representation $D(x_i, y_j) = ([\mathbf{M}, -\mathbf{N}][x_i; y_j])^T([\mathbf{M}, -\mathbf{N}][x_i; y_j])$. Denote $z_{ij} = [x_i; y_j]$ as a new vector concatenating x_i and y_j, $\mathbf{A} = [\mathbf{M}, -\mathbf{N}]$ as the row concatenation form of $\mathbf{M}$ and $\mathbf{N}$, and the covariance matrix $\mathbf{B} = \mathbf{A}^T\mathbf{A}$. The structured cross modal metric learning can be formulated as:

$$\mathbf{B} = \underset{\mathbf{B}}{\operatorname{argmin}} L(\mathbf{B}) = \underset{\mathbf{B}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{B}\|_F^2 + c\sum_{S_1} z_{ii}^T \mathbf{B} z_{ii} + \alpha \sum_i \sum_{s_{2i}} \varepsilon_{ij} + \beta \sum_i \sum_{s_{3i}} \varepsilon_{ij} \quad (3)$$
$$\text{s.t. } z_{ij}^T \mathbf{B} z_{ij} - z_{ii}^T \mathbf{B} z_{ii} \geq \Delta(x_i, y_j) - \varepsilon_{ij}, \varepsilon_{ij} \geq 0, \mathbf{B} \geq \mathbf{0},$$

where the squared Frobenius norm constraint $||\cdot||_F^2$ is imposed on $\mathbf{B}$ to avoid overly large elements, thus overfitting can be effectively alleviated. Then we substitute the inequality constraints with the following form:

$$\mathbf{B} = \underset{\mathbf{B}}{\operatorname{argmin}} L(\mathbf{B}) = \underset{\mathbf{B}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{B}\|_F^2 + c\sum_{S_1} z_{ii}^T \mathbf{B} z_{ii} + \alpha \sum_i \sum_{s_{2i}} [z_{ii}^T \mathbf{B} z_{ii} + \Delta(x_i, y_j) - z_{ij}^T \mathbf{B} z_{ij}]_+ + \beta \sum_i \sum_{s_{3i}} [z_{ii}^T \mathbf{B} z_{ii} + \Delta(x_i, y_j) - z_{ij}^T \mathbf{B} z_{ij}]_+. \quad (4)$$

The above formulation contains $\mathbf{B}$ with the form $z_{ij}^T \mathbf{B} z_{ij}$, and the gradient can be easily calculated as: $\frac{\partial z_{ij}^T \mathbf{B} z_{ij}}{\partial \mathbf{B}} = z_{ij} z_{ij}^T$. We adopt gradient descent during the optimization procedure. We denote the data in S_2 and S_3 where $\{z_{ii}, z_{ij} | z_{ii}^T \mathbf{B} z_{ii} + \Delta(x_i, y_j) - z_{ij}^T \mathbf{B} z_{ij} > 0\}$ with P_2 and P_3, respectively, which indicate the training subsets violating the inequility constraints in Eqn.3. The details are shown in Algorithm 1. We adjust the step size of gradient descent according to the loss value and ensure the objective function is descending at every step. In addition, the positive definiteness of $\mathbf{B}$ is not guaranteed. As discussed in [13], when considering the real world retrieval task, the relative distance or similarity among data is more crucial than the absolute values. Therefore, the positive definiteness of the learned metric is not neccessarily required for such cases. Meanwhile, it maximally preserves multi-level semantic relation among the cross modal data, thus achieves better model generality.

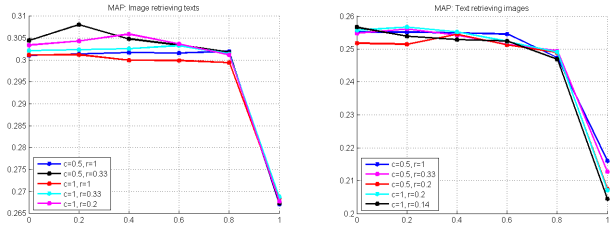
4. EXPERIMENTS

In this section, we conduct extensive experiments on Wikipedia dataset [6] to compare our approach COLAR with other approaches. The dataset contains 2866 image-text pairs of 10 categories. 2173 document pairs are randomly selected in [6] for training and the remaining 693 pairs are used for test. We use the features provided by [6] for model comparison.

The tasks in this paper are retrieving images for textual queries (text-to-image) and retrieving texts for image queries (image-to-text), therefore, the performance of both tasks in our experiments is reported. Mean average precision (MAP) is used as the evaluation criterion [6, 7] to compare different approaches. However, MAP cannot properly measure the performance on data with multi-level semantic relevance. To better evaluate the performance, we adopt the normalized discount cumulative gain (NDCG) [16] as a complementary measure. NDCG is defined as:

Algorithm 1 Model Optimization of COLAR

```
Initialize B with identity matrix, gradient descent step size s^t = 0.01, k = 1, c_1 = 1.2, c_2 = 0.8;
Find the sets P_2 and P_3, then compute the loss value L_old and gradient descent direction G(B);
while s^t > ε and k < K do
    Try B' = B − s^t · G(B);
    Find P_2 and P_3, then compute the loss value L_new;
    if L_old − L_new < (1/σ) L_old && L_old − L_new > 0 then
        B = B'; s^t = s^t · c_1;
        Compute the new gradient G(B); L_old = L_new;
    else if L_old − L_new ≥ (1/σ) L_old then
        B = B'; Compute new G(B); L_old = L_new;
    else if L_old − L_new ≤ 0 then
        s^t = s^t · c_2;
    end if
    k = k + 1;
end while
```



(a) Image-to-text (b) Text-to-image

Fig. 3. Parameter sensitivity test in MAP.

$NDCG@k = \frac{1}{N^k} \sum_{j=1}^{k} \frac{2^{rel(j)}-1}{log(1+j)}$, where N^k is a normalization constant to ensure that the optimal top k ranking of the query is 1. k is called a truncation or threshold level. In our evaluation, the relative gain (i.e., $rel(j)$) for the j-th cross modal document, which is the ground truth one-to-one corresponding and has identical category label with the query, is set respectively to be 3 and 1. Otherwise, the relative gain is 0.

Since there are several weight parameters (c, α, β) in our algorithm, we separate the training data into two parts for parameter tuning. We use 67% of the training data to train the model and use the remaining 33% of the training data to validate the performance of the parameter setting. For better understanding of the parameter validation process, we rewrite the loss function as follows:

$$\mathbf{B} = \underset{\mathbf{B}}{\operatorname{argmin}} L(\mathbf{B}) = \underset{\mathbf{B}}{\operatorname{argmin}} \frac{\|\mathbf{B}\|_F^2}{2} + c\sum_{S_1} z_{ii}^T \mathbf{B} z_{ii} + c \cdot r \cdot p \sum_i \sum_{s_{2i}} [z_{ii}^T \mathbf{B} z_{ii} + \Delta(x_i, y_j) - z_{ij}^T \mathbf{B} z_{ij}]_+ + c \cdot r \cdot (1-p) \sum_i \sum_{s_{3i}} [z_{ii}^T \mathbf{B} z_{ii} + \Delta(x_i, y_j) - z_{ij}^T \mathbf{B} z_{ij}]_+, \quad (5)$$

where c adjusts the relative importance between the empiri-

Table 1. MAP (%) for Wikipedia dataset

Query	Others							COLAR
	PLS[14]	BLM [15]	CCA[6]	SM[6]	SCM[6]	GMMFA[7]	GMLDA[7]	
Image-to-text	20.7	23.7	24.9	22.5	27.7	26.4	27.2	28.2
Text-to-image	19.2	14.4	19.6	22.3	22.6	23.1	23.2	22.4
Average	19.9	19.1	22.3	22.4	25.2	24.8	25.2	**25.3**

Table 2. NDCG@k for Wikipedia dataset

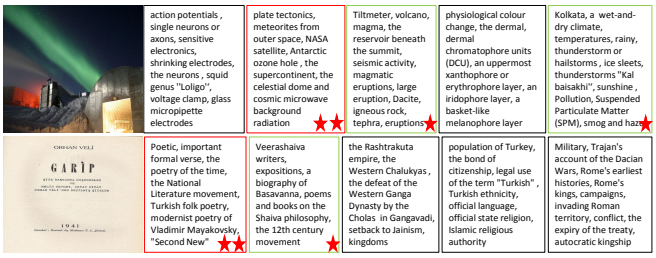
Query @k	CCA	GMLDA	COLAR
Image-to-text @10	0.1056	0.1139	**0.1276**
Image-to-text @20	0.1298	0.1395	**0.1563**
Image-to-text @50	0.1629	0.1719	**0.1899**
Image-to-text @100	0.2031	0.2108	**0.2284**
Image-to-text @693	0.5154	0.5189	**0.5296**
Text-to-image @10	0.1107	0.1410	**0.1672**
Text-to-image @20	0.1400	0.1649	**0.1928**
Text-to-image @50	0.1633	0.1931	**0.2159**
Text-to-image @100	0.1989	0.2235	**0.2400**
Text-to-image @693	0.5181	0.5293	**0.5416**

cal loss and the model complexity penalty. p $(0 < p < 1)$ measures the relative importance between the intra-class and inter-class terms. r measures the relative importance between the one-to-one correspondence and the intra-class / inter-class information. With p from 0 to 0.8, small changes are observed on the MAP curves, while there is a peak at $p = 0.2$, see Fig.3. Furthermore, we can obtain that ratio r should not be too large and setting $c = 0.5$ or $c = 1$ will not lead to drastic performance change. Therefore, we set $c = 0.5, r = 0.2$ and $p = 0.2$ for optimal performance.

The experimental results of cross modal retrieval in terms of MAP are shown in Table 1. We can see that COLAR slightly outperforms the state-of-the-art approaches, i.e., SCM [6] and GMLDA [7], where both approaches take advantage of the category information for correlation learning. Furthermore, we report the performance in terms of NDCG@k in Table 2 to evaluate how our method and other correlation learning based methods (CCA and GMLDA) perform on retrieving cross modal data with multi-level semantic relevance.

From Table 2, we can see that COLAR significantly outperforms CCA and GMLDA [7] with different setting of k, where 693 is the number of documents of the whole test set. Furthermore, COLAR performs much better than other methods on the top retrieved results (i.e., $k = \{10, 20, 50\}$). Specifically, for image-to-text retrieval, COLAR achieves 12% improvement than GMLDA on NDCG@10 and NDCG @20, and for text-to-image retrieval, COLAR achieves 18.6% and 16.9% improvement than GMLDA on NDCG@10 and NDCG@20. The results show that our method is more capable of encoding the multi-level semantic relevance, and the retrieved documents are appropriately ranked according to their relevance level with the queries.

We illustrate some top 5 retrieved examples using our COLAR method for cross modal retrieval in Fig.4. We further see from the examples that COLAR could better retrieve cross modal documents according to their relevance level. Specifically, the second example in Fig.4(a) and the first example in Fig.4(b) are "perfectly" ranked results that can be hardly obtained by other approaches. For visual documents with complex patterns, the retrieval performance seems to be influenced by the semantic consistency between the object and the background, as shown in the second example in Fig.4(b).



(a) Examples of image-to-text retrieval results

(b) Examples of text-to-image retrieval results

Fig. 4. Some examples of the top five results with COLAR for cross modal retrieval. Double red stars represent the ground truth one-to-one correspondence, and single red star represents the cross modal data with identical category labels as the queries.

5. CONCLUSION

We propose COLAR, which models different levels of the semantic relevance, e.g., the one-to-one correspondence and intra-class relation with a structure risk minimization for cross modal metric learning, while traditional correlation learning approaches can not effectively handle such information. Promising performance is achieved on cross modal retrieval. In future work, we will extend COLAR to nonlinear metric learning, and conduct extensive evaluations on other cross modal or multi-view data.

6. REFERENCES

[1] Cees GM Snoek, Marcel Worring, Jan C Van Gemert, Jan-Mark Geusebroek, and Arnold WM Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *Proceedings of the 14th annual ACM international conference on Multimedia*. ACM, 2006, pp. 421–430.

[2] Milind Naphade, John R Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis, "Large-scale concept ontology for multimedia," *Multimedia, IEEE*, vol. 13, no. 3, pp. 86–91, 2006.

[3] Cees GM Snoek, Bouke Huurnink, Laura Hollink, Maarten De Rijke, Guus Schreiber, and Marcel Worring, "Adding semantics to detectors for video retrieval," *Multimedia, IEEE Transactions on*, vol. 9, no. 5, pp. 975–986, 2007.

[4] Behjat Siddiquie, Rogério Schmidt Feris, and Larry S Davis, "Image ranking and retrieval based on multi-attribute queries," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 801–808.

[5] Tian Lan, Weilong Yang, Yang Wang, and Greg Mori, "Image retrieval with structured object queries using latent ranking svm," in *Computer Vision–ECCV 2012*, pp. 129–142. Springer, 2012.

[6] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 251–260.

[7] Abhishek Sharma, Abhishek Kumar, H Daume, and David W Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2160–2167.

[8] Yan Peng, Daoqiang Zhang, and Jianchun Zhang, "A new canonical correlation analysis algorithm with local discrimination," *Neural processing letters*, vol. 31, no. 1, pp. 1–15, 2010.

[9] Tingkai Sun and Songcan Chen, "Locality preserving cca with applications to data visualization and pose estimation," *Image and Vision Computing*, vol. 25, no. 5, pp. 531–543, 2007.

[10] Tingkai Sun, Songcan Chen, Jingyu Yang, and Pengfei Shi, "A novel method of combined feature extraction for recognition," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 1043–1048.

[11] Thorsten Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 133–142.

[12] Kilian Q Weinberger and Lawrence K Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[13] Gal Chechik, Uri Shalit, Varun Sharma, and Samy Bengio, "An online algorithm for large scale image similarity learning," in *Advances in Neural Information Processing Systems*, 2009, pp. 306–314.

[14] Abhishek Sharma and David W Jacobs, "Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 593–600.

[15] Joshua B Tenenbaum and William T Freeman, "Separating style and content with bilinear models," *Neural computation*, vol. 12, no. 6, pp. 1247–1283, 2000.

[16] Kalervo Järvelin and Jaana Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.