A CROSS-MODAL APPROACH FOR EXTRACTING SEMANTIC RELATIONSHIPS OF CONCEPTS FROM AN IMAGE DATABASE

Marie Katsurai, Takahiro Ogawa, and Miki Haseyama

Graduate School of Information Science and Technology, Hokkaido University
N-14, W-9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan
E-mail: {katsurai, ogawa}@lmd.ist.hokudai.ac.jp, miki@ist.hokudai.ac.jp

ABSTRACT

This paper presents a cross-modal approach for extracting semantic relationships of concepts from an image database. First, canonical correlation analysis (CCA) is used to capture the cross-modal correlations between visual features and tag features in the database. Then, in order to measure inter-concept relationships and estimate semantic levels, the proposed method focuses on the distributions of images under the probabilistic interpretation of CCA. Results of experiments conducted by using an image database showed the improvement of the proposed method over existing methods.

***Index Terms*—** semantic analysis, image databases, inter-concept relationships, canonical correlation analysis, cross-modal correlations

1. INTRODUCTION

Extracting semantic relationships of concepts from an image database has recently attracted much research attention [1-6]. The aim of this research has been to facilitate image annotation, retrieval, and tag recommendation. The following information is especially useful: (a) inter-concept relationships and (b) semantic levels of concepts. In order to extract these semantic relationships, concept co-occurrences have been used by regarding each image in the target database as a document containing tags [1, 2]. Recently, visual features have been used to describe visual correlations of concepts [3, 4]. These conventional methods use features in a single modality, i.e., tag features or visual features only. On the other hand, some methods use two features for characterizing the relationships [5, 6]. However, the conventional approaches analyze features in each modality separately and finally combine their results for describing the relationships. Each of these modalities has a specific structure and provides information that can be unique or common to other modalities, however, multi-modal information will provide new structures that cannot be found by a single modality [7]. If cross-modal information of these features can be used, performance improvement of the semantic relationship extraction can be expected.

In this paper, we present a cross-modal approach for extracting semantic relationships of concepts from an image database. First, we use canonical correlation analysis (CCA) [8] to capture the cross-modal correlations of visual features and tag features. Then, in order to realize the semantic relationship extraction, we focus on the distribution of images corresponding to a target concept in the subspace estimated by CCA. Specifically, we make two assumptions based on the latent variables that correlate the two modalities under probabilistic interpretation of CCA [9]. The main contributions of this paper are two-fold: (i) we consider the cross-modal correlations of the visual features and the tag features to extract semantic relationships of concepts from an image database; (ii) in order to measure the inter-concept relationships and estimate the semantic levels of concepts, we focus on the distributions of images corresponding to the target concept in the latent space.

2. CCA AND ITS PROBABILISTIC INTERPRETATION

This section describes canonical correlation analysis (CCA) and its probabilistic interpretation. Given a pair of features $\boldsymbol{x}$ and $\boldsymbol{y}$, CCA searches for the linear transformations $\mathbf{U}_x$ and $\mathbf{U}_y$ such that each dimension of $\mathbf{U}_x\boldsymbol{x}$ correlates maximally with the corresponding dimension of $\mathbf{U}_y\boldsymbol{y}$. If $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{pmatrix}$ is a sample covariance matrix of features pairs, then the projection matrices $\mathbf{U}_x$ and $\mathbf{U}_y$ can be computed as the solutions of the generalized eigenvalue problem:

$$\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\mathbf{U}_x = \boldsymbol{\Sigma}_{xx}\mathbf{U}_x\boldsymbol{\Lambda}^2, \tag{1}$$

$$\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\mathbf{U}_y = \boldsymbol{\Sigma}_{yy}\mathbf{U}_y\boldsymbol{\Lambda}^2, \tag{2}$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix of the first d canonical correlations.

In [9], Bach and Jordan propose a probabilistic interpretation of CCA. In this model, the features $\boldsymbol{x}$ and $\boldsymbol{y}$ are generated from the same latent variables $\boldsymbol{z}$ (Gaussian distribution with zero mean and unit variance) with unknown linear transformations by adding Gaussian noise. The probabilistic structure of CCA has been used in many applications since it can provide posterior expectations of the latent variables $\boldsymbol{z}$ that lie in the subspace found by CCA [10, 11]. Under this framework, we try to extract semantic relationships of concepts by exploiting the distribution of the latent variables that correlate two features.

3. CROSS-MODAL CORRELATION-BASED SEMANTIC RELATIONSHIP EXTRACTION

This section presents a cross-modal correlation-based semantic relationship extraction method. Given a tagged image I_i ($i = 1, 2, \cdots, N$, where N is the number of images), we first extract their D-dimensional visual features $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,D}]^T$ and K-dimentional tag features $\mathbf{y}_i = [y_{i,1}, y_{i,2}, \cdots, y_{i,K}]^T$. The set of training data is represented by $S = \{\boldsymbol{x}_i, \boldsymbol{y}_i | i = 1, 2, \cdots, N\}$. After applying CCA to the dataset S, a subspace in which two modalities are maximally correlated is obtained. We focus on the distribution of images corresponding to a target concept, based on the latent variables under probabilistic interpretation of CCA. By using the representation of *a concept centroid* in the latent space, we measure inter-concept relationships based on the cross-modal correlations (See 3.1). Furthermore, we derive a criterion for estimating semantic levels of concepts (See 3.2).

3.1. Measurement of inter-concept relationships

Let C_k ($k = 1, 2, \cdots, K$, where K is the number of concepts) be a concept in the database. The proposed method measures the relevance between concepts C_k and C_l ($l \neq k$) by making the following assumption:

Assumption 1. (Inter-concept relationship)
If two concepts are semantically correlated, their images distribute very closely in the latent space.

In order to evaluate the distribution of images, we use a probabilistic interpretation of CCA. Given visual features $\boldsymbol{x}$ and tag features $\boldsymbol{y}$, the posterior probability of the latent variables $\boldsymbol{z}$ follows a normal distribution whose mean and variance are:

$$E(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}) = \begin{pmatrix}\mathbf{M}_x\\ \mathbf{M}_y\end{pmatrix}^{\mathrm{T}} \begin{pmatrix}(\mathbf{I}-\mathbf{\Lambda}^2)^{-1} & -(\mathbf{I}-\mathbf{\Lambda}^2)^{-1}\mathbf{\Lambda}\\ -(\mathbf{I}-\mathbf{\Lambda}^2)^{-1}\mathbf{\Lambda} & (\mathbf{I}-\mathbf{\Lambda}^2)^{-1}\end{pmatrix} \begin{pmatrix}\mathbf{U}_x^{\mathrm{T}}(\boldsymbol{x}-\bar{\boldsymbol{x}})\\ \mathbf{U}_y^{\mathrm{T}}(\boldsymbol{y}-\bar{\boldsymbol{y}})\end{pmatrix}, \quad (3)$$

$$var(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}) = \mathbf{I} - \begin{pmatrix}\mathbf{M}_x\\ \mathbf{M}_y\end{pmatrix}^{\mathrm{T}} \begin{pmatrix}(\mathbf{I}-\mathbf{\Lambda}^2)^{-1} & -(\mathbf{I}-\mathbf{\Lambda}^2)^{-1}\mathbf{\Lambda}\\ -(\mathbf{I}-\mathbf{\Lambda}^2)^{-1}\mathbf{\Lambda} & (\mathbf{I}-\mathbf{\Lambda}^2)^{-1}\end{pmatrix} \begin{pmatrix}\mathbf{M}_x\\ \mathbf{M}_y\end{pmatrix}, \quad (4)$$

where $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ are the sample means of the visual features and tag features, respectively, and $\mathbf{M}_x$ and $\mathbf{M}_y$ are arbitrary matrices with spectral norms smaller than one, such that $\mathbf{M}_x\mathbf{M}_y = \mathbf{\Lambda}$. In our implementation, let $\mathbf{M}_x = \mathbf{M}_y = \mathbf{\Lambda}^{1/2}$. As shown in the Eq. (4), the variances are the same for all of the images. We use the means in Eq. (3) to measure the overlap of the two concepts in the latent space. An illustration of our approach is shown in Fig. 1. If $\boldsymbol{m}_i = E(\boldsymbol{z}|\boldsymbol{x}_i, \boldsymbol{y}_i)$, then we define *a concept centroid* by using its images in the latent space as follows:

$$Cent(C_k) = \frac{1}{|R_k|}\sum_{I_i \in R_k} \boldsymbol{m}_i, \quad (5)$$

where R_k denotes a set of images that are tagged with concept C_k, and $|R_k|$ is the total number of images in R_k. By using the above definition, we compute the semantic distance of concepts C_k and C_l as follows:

$$dist(C_k, C_l) = \|Cent(C_k) - Cent(C_l)\|^2 . \quad (6)$$

If $dist(C_k, C_l)$ is close to 0, it means concepts C_k and C_l are semantically correlated. The proposed approach can use the maximally correlated visual features and tag features, and the computed distance can effectively reflect both modalities. We verify whether making Assumption 1 provides effective results for measuring the inter-concept relationships in experiments (see Sec. 4).

3.2. Estimation of semantic levels for concepts

In our framework that uses distributions of images in the latent space, semantic levels of concepts can also be estimated. This is performed based on the following assumption:

Assumption 2. (Semantic level)
If a concept has a specific meaning, then its images should be close to each other in the latent space. On the other hand, if a concept is more abstract and higher-level, its images should distribute widely in the latent space.

Then we measure semantic level as the average distance between a concept centroid and its image as follows:

$$\Phi(C_k) = \frac{1}{|R_k|}\sum_{I_i \in R_k} \|\mathbf{m}_i - Cent(C_k)\|^2 . \quad (7)$$

A small $\Phi(C_k)$ means images of C_k are visually and contextually similar to each other. That is, concept C_k represents a specific meaning in the target image database. On the other hand, a large $\Phi(C_k)$ means the concept C_k is more abstract due to having several meanings. The above score is converted by concept frequency in the database in a similar way to [4], which is denoted by $\Phi_{norm}(C_k)$. By sorting the scores in descending order, we can estimate the semantic level of concepts. We also verify whether making Assumption 2 leads to an effective approach for estimating semantic levels of concepts in the next section.

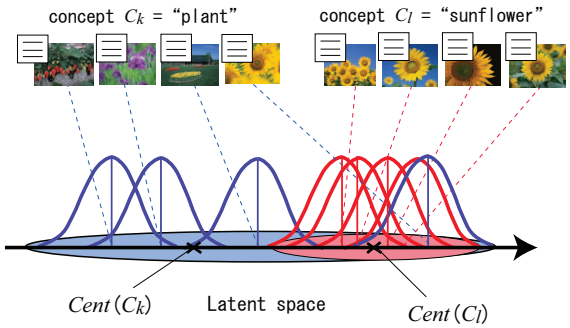


Fig. 1. Illustration of the proposed approach, where two concepts *"plant"* and *"sunflower"* are shown as examples.

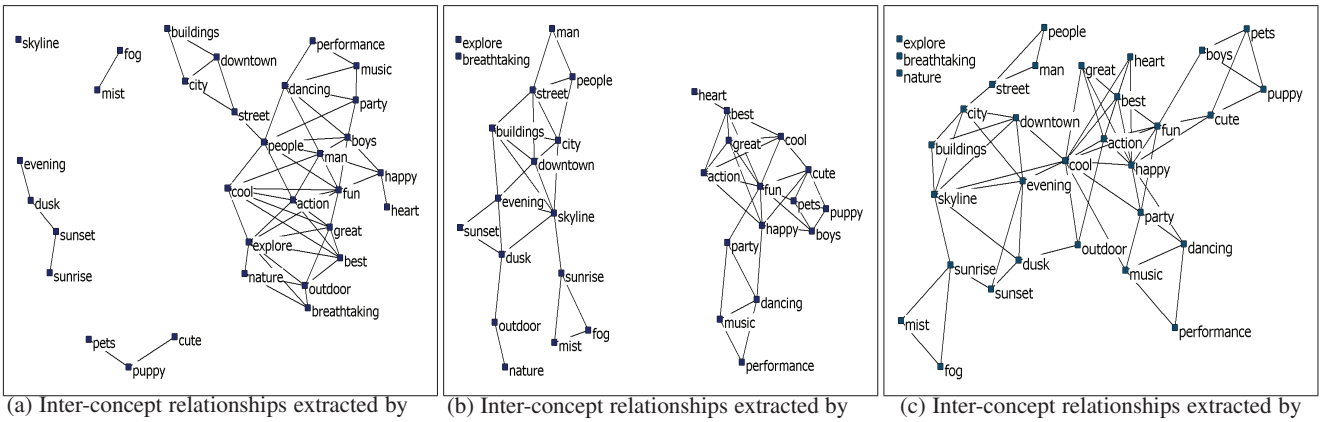
4. EXPERIMENTAL RESULTS

In this section, we show experimental results to verify the effectiveness of the proposed method. We used the NUS-WIDE dataset [12] containing 269,648 images from Flickr[1]. We divide the dataset into 168,650 images for training and the remainder for testing. Each image is manually tagged with a set of concepts, with the total number of concepts equal to 1815. In our experiments, we used 500-dimensional bag of features (BoF) with SIFT local descriptors for visual features. Furthermore, for tag features, we used 1815-dimensional binary features that represent whether the corresponding concept is related to the image. The results of inter-concept relationships based on Assumption 1 are shown in Sec. 4.1. We show the results of semantic level estimation based on Assumption 2 in Sec. 4.2.

4.1. Results of inter-concept relationship measurement

First, the extracted inter-concept relationships are depicted as networks by using NetDraw [13]. Since it is difficult to show relationships of all of the concepts in the database, we randomly selected 32 concepts from the database to draw relationships. Moreover, strong relationships are chosen by thresholding for easier viewing. For comparison, conventional methods [1, 5] were also applied to the same dataset. The drawn relationships are shown in Fig. 2, where each node represents a concept and each edge between concepts represents relevance. As shown in Fig. 2, although the three networks were constructed from the same database, their inter-concept relationships are different from each other. For example, the conventional method [1] extracts strong relationships such as *buildings - evening*, *best - heart*, and *sunrise - mist*. The conventional method [5] found the relationships such as *sunrise - sunset* by using visual features, however, other relationships such as *outdoor - action* are also extracted. On the other hand, our method effectively extracts relationships such as *heart - happy*, *boys - people*, and *nature - breathtaking*. We also find that our method cannot extract relationships involving the *skyline* concept. If the threshold is larger, then the relationship *skyline - evening* will appear.

[1] http://flickr.com/



(a) Inter-concept relationships extracted by the proposed method.

(b) Inter-concept relationships extracted by the conventional method [1].

(c) Inter-concept relationships extracted by the conventional method [5].

Fig. 2. Semantic inter-concept relationships extracted from NUS-WIDE dataset by the proposed method and the conventional method [1, 5]. Each node represents a concept and each edge between concepts represents a strong relationship.

Table 1. User assessments of inter-concept relationships extracted by the proposed method and the conventional methods [1].

Proposed method		Conventional method [1]	
Concept pair	Score	Concept pair	Score
landscape - scenery	4.82	roads - global	1.64
sunset - sunrise	4.55	football - soccer	4.82
sea - beach	4.73	underwater - scuba	4.34
portrait - face	4.18	jets - engines	4.55
plane - jet	4.73	rabbit - bunny	4.91
fun - hot	3.00	long - exposure	1.55
boy - male	4.34	secret - crime	2.64
child - kids	4.91	shore - coast	4.91
night - lights	3.64	pools - baths	3.55
explore - wow	3.00	adults - whites	1.91
ocean - coast	4.55	aeroplane - wing	4.10
color - photo	3.20	analog - health	1.37
Average	**4.13**	Average	3.01

Relationships extracted by thresholding were also manually evaluated by users. We presented 11 users a list of 60 concept pairs with the highest relevance. The users were required to give a score ranging from 1 to 5. A high score means the two concepts are correlated. Table. 1 shows the final score for some of the pairs, obtained by averaging the scores from all the users. The bottom row shows average scores for all 60 concept pairs. Some of the relationships rated as strong by the conventional method [1] received low user scores, e.g., "*roads - global*" and "*analog - health*." On the other hand, all relationships rated as strong by our method received moderate to high user scores. From these results, we can see that the proposed method extracts the semantic relationships better than the conventional methods by focusing on the cross-modal correlations.

Furthermore, we apply the extracted relationships to tag recommendation for quantitative evaluation. Given a set of initial tags for a testing image, relevant tags are recommended to assist manual annotator [14]. In our experiments, initial tags for each testing image are randomly chosen from its correct concepts. The sum of distances for the initial tags is computed for each concept, and then the most related concepts are recommended [15]. The average precision of the top 10 recommendations and the coverage over all correct recommendations are calculated to measure the performance of each method, which are shown in Table 2. From these results, we found that our approach with Assumption 1 extracts more semantic relationships due to the effective use of the cross-modal correlations between tag features and visual features.

Table 2. Tag recommendation results.

	Precision	Coverage
Proposed method	**0.535**	**0.517**
Conventional method [1]	0.485	0.494
Conventional method [5]	0.501	0.480

4.2. Results of semantic level estimation

We now show the results of semantic level estimation for each concept based on Assumption 2. For each concept, the scores representing semantic levels were calculated by using Eq. (7) and sorted in descending order. The results are shown in Table 3, where the top 12 and the bottom 12 concepts are shown. Concepts that have been used in image annotation benchmarks [16] are highlighted in bold for reference. If concept C approaches the top of the hierarchy, it means concept C is more abstract and higher-level in the database. From this table, we can find that the concepts in the bottom row actually represent specific objects or scenes. For the concepts in the top row, color-related concepts (e.g., red, yellow) and shape-related concepts (e.g., square) are placed as high-level concepts since our experiments do not use color and shape features. These results provide us with insight that several kind of features are necessary for comparisons.

Furthermore, for quantitative evaluation, all concepts were manually divided into objects, scenes, and others, which was performed by following the related work [4]. We investigate the effectiveness of the cross-modal correlations of tag features and visual features for estimating semantic level in experiments. Then, based on the results of semantic level estimation, we compute the precision $P@N$ as follows:

Table 3. Results of semantic level estimation for each concept, in which the top 12 and the bottom 12 concepts are shown. Concepts that were used in the image annotation benchmark [16] are highlighted in bold.

Rank	1	2	3	4	5	6	7	8	9	10	11	12
The 12 top concept C	macro	**square**	pattern	abstract	architecture	**flower**	yellow	red	blue	art	explore	patterns
$\Phi_{norm}(C)$	413.5	389.1	381.8	302.3	286.6	264.8	260.7	260.453	246.167	239.213	200.161	195.386
The 12 bottom concept C	**sunset**	**landscape**	**water**	**night**	**sea**	**beach**	**trees**	**car**	**clouds**	**mountains**	**boat**	**ocean**
$\Phi_{norm}(C)$	-606.6	-506.1	-475.3	-392.6	-370.1	-335.5	-278.9	-258.2	-230.9	-225.9	-221.2	-197.4

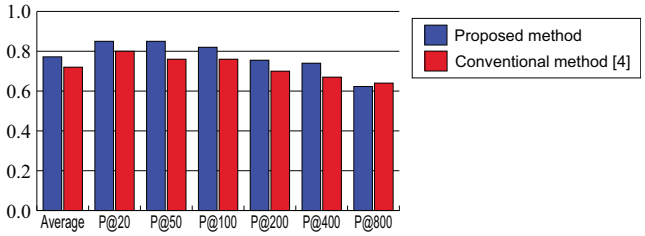


Fig. 3. Precisions at $N = 20, 50, 100, 200, 400$, and 800 for semantic level estimation.

$$\mathrm{P@}N = \frac{\#\text{concepts representing specific objects or scenes}}{\text{the bottom } N \text{ concepts in the sort of scores}}. \quad (8)$$

The results are shown in Fig. 3. As shown in this figure, the proposed method can place the concepts representing objects or scenes in the lower level more accurately than the conventional method [4]. Note that the method [4] aims at detecting visually representative tags, and they consider only visual aspects of each concept. However, the results show that their cross-modal correlations with tag features are effective to detect concepts representing specific objects or scenes. Our future works will include further evaluations and comparisons.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose a cross-modal approach for extracting semantic relationships of concepts from an image database. In the proposed method, we use canonical correlation analysis (CCA) to capture correlations between visual features and tag features in the database. Then, under the probabilistic interpretation of CCA, we make two assumptions based on distributions of images corresponding to a target concept in the latent space. Experimental results show that our cross-modal approach can extract more semantic relationships than the conventional method that use features in single modality. In future works, we should conduct more experiments and comparisons. In addition, we will investigate various kinds of visual features and tag features to improve the performance of the proposed approach.

6. ACKNOWLEDGEMENTS

This work was partly supported by Grant-in-Aid for Scientific Research (B) 21300030, Japan Society for the Promotion of Science (JSPS).

7. REFERENCES

[1] R. L. Cilibrasi and P. M. B. Vitanyi, "The google similarity distance," *IEEE Trans. on Knowl. and Data Eng.*, vol. 19, no. 3, pp. 370 –383, 2007.

[2] D. Liu, X. S. Hua, L. Yang, M. Wang, and H. J. Zhang, "Tag ranking," in *Proc. WWW*, pp. 351–360, 2009.

[3] L. Wu, X. S. Hua, N. Yu, W. Y. Ma, and S. Li, "Flickr distance," in *Proc. ACM MM*, pp. 31–40, 2008.

[4] A. Sun and S. S. Bhowmick, "Quantifying tag representativeness of visual content of social images," in *Proc. ACM MM*, pp. 471–480, 2010.

[5] Y. G. Jiang, Jun Wang, Shih-Fu Chang, and Chong-Wah Ngo, "Domain adaptive semantic diffusion for large scale context-based video annotation," in *Proc. ICCV*, pp. 1420 –1427, 2009.

[6] H. Kawakubo, Y. Akima, and K. Yanai, "Automatic construction of a folksonomy-based visual ontology," in *Proc. ISM*, pp. 330 –335, 2010.

[7] N. M. Correa, Y. O. Li, T. Adali, and V. D. Calhoun, "Fusion of fMRI, sMRI, and EEG data using canonical correlation analysis," in *Proc. ICASSP*, pp. 38–388, 2009.

[8] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, 1936.

[9] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Tech. Rep. 688, Department of Statistics, University of California, Berkeley, 2005.

[10] T. Harada, H. Nakayama, and Y. Kuniyoshi, "Image annotation and retrieval based on efficient learning of contextual latent space," in *Proc. ICME*, pp. 858 –861, 2009.

[11] T. Nakano, A. Kimura, H. Kameoka, S. Miyabe, S. Sagayama, N. Ono, K. Kashino, and T. Nishimoto, "Automatic video annotation via hierarchical topic trajectory model considering cross-modal correlations," in *Proc. ICASSP*, pp. 2380 –2383, 2011.

[12] T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: a real-world web image database from National University of Singapore," in *Proc. CIVR*, pp. 1–9, 2009.

[13] S. Borgatti, "Netdraw network visualization," `http://www.analytictech.com/netdraw/netdraw.htm`, Accessed: 18/01/2012.

[14] B. Sigurbjörnsson and Roelof van Zwol, "Flickr tag recommendation based on collective knowledge," in *Proc. WWW*, pp. 327–336, 2008.

[15] L. Wu and X. S. Hua, "Elements of visual concept analysis," in *Multimedia Analysis, Processing and Communications*, vol. 346 of *Studies in Computational Intelligence*, pp. 679–717. Springer Berlin / Heidelberg, 2011.

[16] M. Grubinger, P. Clough, H. Mller, and T. Deselaers, "The IAPR-TC12 benchmark: A new evaluation resource for visual information systems," in *Proc. of Int'l Conf. on Language Resources and Evaluation*, 2006.