



Learning a Semantic Space for Modeling Images, Tags and Feelings in Cross-Media Search

Sadaqat ur Rehman¹(✉), Yongfeng Huang¹, Shanshan Tu²,
and Basharat Ahmad¹

¹ Tsinghua National Laboratory for Information Science and Technology,
Tsinghua University, Beijing 100084, People's Republic of China
z-sun15@mails.tsinghua.edu.cn, yfhuang@mail.tsinghua.edu.cn

² Faculty of Information Technology, Beijing University of Technology,
Beijing 100124, China
sstu@bjut.edu.cn

Abstract. This paper contributes a new, real-world web image dataset for cross-media retrieval called FB5K. The proposed FB5K dataset contains the following attributes: (1) 5130 images crawled from Facebook; (2) images that are categorized according to users' feelings; (3) images independent of text and language rather than using feelings for search. Furthermore, we propose a novel approach through the use of Optical Character Recognition (OCR) and explicit incorporation of high-level semantic information. We comprehensively compute the performance of four different subspace-learning methods and three modified versions of the Correspondence Auto Encoder (Corr-AE), alongside numerous text features and similarity measurements comparing Wikipedia, Flickr30k and FB5K. To check the characteristics of FB5K, we propose a semantic-based cross-media retrieval method. To accomplish cross-media retrieval, we introduced a new similarity measurement in the embedded space, which significantly improved system performance compared with the conventional Euclidean distance. Our experimental results demonstrated the efficiency of the proposed retrieval method on three different public datasets.

Keywords: Cross-media search · Text-image-feeling embeddings · FB5K dataset

1 Introduction

The current era has seen rapid growth in Multimedia Information Retrieval (MIR). Despite constant hard work in the development and construction of new MIR techniques and datasets, the semantic gap between images and high-level concepts remains high. We need a promising model to focus on modeling high-level semantic concepts, either by image annotation or by object recognition to diminish this semantic gap. Numerous real-world methods [1, 2] have been introduced for this kind of concept-based multimedia search system. Among several

of these methodologies, the first step is dataset selection for high-level concepts and small semantic gaps, which are relatively easy for machine understanding and training.

This paper presents a novel resource evaluation dataset for cross-media searching, called FB5K along with a benchmark learning system. Existing cross-media or multi-modal retrieval datasets have some limitations. Firstly, some datasets lack context information i.e. link relations. Such context information is quite accurate, and can provide significant evidence to ameliorate cross-media retrieval system accuracy. Similarly, the Pascal VOC 2012 dataset¹ [3] consists of only 20 categories. However, cross-media retrieval implicates numerous domains under real-world internet conditions. Cross-media retrieval systems trained on scanty domain datasets have difficulties in handling queries from anonymous domains. Secondly, popular cross-media datasets are small in size, for example Xmedia [4], IAPR TC-12 [5], and Wikipedia² [6]. This deficiency in appropriate data makes it difficult for retrieval systems to learn and evaluate the robustness in real-world galleries. Thirdly, datasets such as, ALIPR [7], SML [8], either just used all the image annotation keywords associated with training images, or unenforced any constraint to the annotation vocabulary for example ESP [9], LabelMe [10], and AnnoSearch [11]. Therefore, these datasets essentially neglect the differences among keywords relating to semantic gaps.



Fig. 1. Some examples of FB5K dataset used for cross-media search.

¹ <http://host.robots.ox.ac.uk/pascal/VOC/>.

² <http://www.svcl.ucsd.edu/projects/crossmodal/>.

Considering the aforementioned problems, this paper makes three major contributions. The first is the collection of a new resource evaluation cross-media retrieval dataset, named FB5K. It contains 5130 image-feeling pairs collected from Facebook³, introduced for the first time in the cross-media retrieval research community. This dataset is differentiated from current datasets in three aspects: varied domains, high-level semantic information incorporation, and rich context information. Eventually, it should provide a more accurate standard for cross-media study. Therefore, we constructed a standard dataset, keeping in mind the research issues to focus researcher/developer efforts on cross-media retrieval algorithm development, instead of laboriously comparing methods and results. The second is that, to the best of our knowledge, this is the first effort to collect a dataset of high-level concepts with small semantic gaps based on users' semantic descriptions i.e. image-feeling relationships. Third, this approach aims to learn the cross-media embeddings of users' feelings, images, and tags/texts. We propose a novel method by using Optical Character Recognition (OCR), explicit incorporation of high-level semantic information, and a new similarity measurement in the embedded space, which significantly overcomes the conventional distance measurement methods and improve retrieval performance.

2 Proposed Dataset

This section describes a new dataset called FB5K, which comprises 5130 images collected from Facebook. The complete FB5K dataset will be made available via ABC⁴.

2.1 Dataset Collection

Each step in the dataset collection is briefly explained below.

Seed User Gathering. In order to obtain the genuine emotions of users associated with an image rather than image contents, we obtained seed users by sending queries to Facebook with numerous key words, for example *happy*, *hungry*, *love*, etc.

User Candidate Generation. To generate user candidates we implement a web spider to crawl the user accounts of individuals who were following the seed users. This step was repeated a number of times until we obtained a lengthy list of user candidates.

Feelings Collection. Another web spider collected feelings as text associated with the matching images by visiting the web pages of different users present in the candidate list. Our finding suggested that about 80% of the users' feelings accompanied the images.

³ [facebook.com](https://www.facebook.com).

⁴ <http://www.xyz.com/>.

Data Pruning. We refer to an image, tag, or feeling-text pair as a tweet. Data were pruned based on the following criteria (pruned out data were referred to as garbage data):

- Feelings without images;
- Tweets not associated with images or feelings;
- Repeated images with the same ID;
- Error images.

As a result, a total of 5130 image-tag pairs were obtained. Figure 1 presents some examples from this benchmark dataset.

2.2 Dataset Characteristics

The performance of cross-media retrieval methods is highly dependent on the nature of the dataset used for their evaluation. The FB5K dataset includes a set of images that are closely associated with user feelings. These images were crawled from Facebook along with the user-associated feelings. The FB5K dataset has the following attributes:

- First, since this dataset was collected from a social media website, it contains a broad variety of domains under single examples of feelings such as, *hungry, love, sad, thankful* etc.
- Second, the relationship between images and users’ feelings is often very strong. In the examples given in Fig. 1, the images have strong ties with the associated feelings. Such is the case in a realistic scenario.
- Third, FB5K is a large-scale dataset, containing 5130 image-text pairs, which helps to avoid overfitting during system training. In other words, it helps to test the cross-media retrieval method’s robustness via a wealth of data.
- Fourth, this dataset helps to reduce the semantic gap by providing more accessible visual content descriptors using high-level semantic concepts.

To our knowledge, this is the first cross-media dataset that consists of the above-mentioned characteristics. Also, we believe that FB5K is the first dataset collected from Facebook that comprises high-level concepts with minor semantic gaps between users’ semantic descriptions, and a ground-truth of 70 concepts for the whole dataset.

3 Proposed Retrieval Method for the FB5K Dataset

This section briefly explains the proposed cross-media retrieval algorithm for FB5K. Numerous features are used for image representation, for example SIFT [12], color features [13], GIST [14] and HOG [15,16]. These features are useful for extracting colors and shapes of images, but not for words represented by the images. In this regards, we first propose OCR then adopt explicit incorporation of high-level semantic information and finally develop a novel similarity

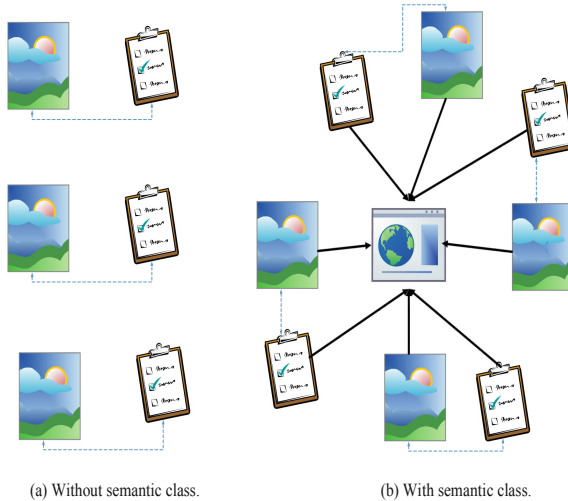


Fig. 2. Graphical representation of high-level semantic information incorporation: (a) without semantic class and (b) with semantic class.

measurement in the embedded space to improve the retrieval performance. A detailed explanation is provided as follows:

Text Extraction. First is the extraction of words on each image using tessart⁵.

Incorporation of High-Level Semantic Information. To facilitate OCR text extraction we incorporate high-level semantic information for learning a common space for image, text/tag, and semantic information (user feelings). Assume we have n training images having i_f -dimensional visual feature vectors and t_f -dimensional tag feature vectors. Where $I \in R^{n \times i_f}$ and $T \in R^{n \times t_f}$. Furthermore, we also associated each training image with a high-level semantic class, $C \in R^{n \times c}$, where c represents the number of categories. Individual images are labeled with one c class (only one specific class in each row of K is 1 and the remaining are 0).

Let i, j denote two points. We define similarity as:

$$K_x(i, j) = \psi_x(i)\psi_x(j)^T \quad (1)$$

where K_x is a kernel function and $\psi_x(\cdot)$ represent a function embedding the original feature vector into a nonlinear kernel space. The goal is to find matrices W_x that project the embedded vector $\psi_x(i)$ to minimize the distance between data items.

⁵ <https://github.com/tesseract-ocr/tesseract>.

The objective function can be mathematically expressed as:

$$\begin{aligned} \min_{w_1, w_2, w_3} &= \sum_{x, y=1}^3 \|\psi_1(I)W_1 - \psi_2(T)W_2\|_2 \\ &+ \|\psi_1(I)W_1 - \psi_3(C)W_3\|_2 + \|\psi_2(T)W_2 - \psi_3(C)W_3\|_2 \\ &\text{where } w_1 = w_2 = w_3 = 0 \end{aligned} \quad (2)$$

this equation tries to align corresponding images and tags [17], whereas, the remaining two terms try to align images with their semantic class. Figure 2 illustrates graphically the benefits of incorporating high-level semantic information.

Similarity Measure. We developed a novel similarity measurement that yielded better realistic results. Mathematically, this can be expressed as:

$$\text{sim}(x_i y_i) = \frac{(\psi_x(i)W_x)(\psi_y(j)W_y)^T}{\|(\psi_x(i)W_x)\|_2 \|(\psi_y(j)W_y)\|_2} \quad (3)$$

where x_i represents the training image and y_i represents the corresponding tweet. W_x projects the embedded vector $\psi_x(i)$ and W_y projects the embedded vector $\psi_y(j)$ to minimize the distance between image and text.

Distance in Common Subspace. In this paper, we represent the cosine distance between two different modalities in the common subspace as $\text{Cos}(Twt, Img)$, where Twt and Img represent the tweet and image. It was learned by retrieval methods such as, Correspondence Auto Encoder (Corr-AE) and subspace methods.

Ranking. Each candidate in the gallery was ranked, based on similarity distances between the queries and candidates.

4 Experimental Results and Discussion

4.1 Experimental Setup

All experiments were performed on four subspace learning methods, which were Canonical Correlation Analysis (CCA) [6], Bilinear Model (BLM) [18], Partial Least Square (PLS) [19], and Generalized Multi-view Marginal Fisher Analysis (GMMFA) [20] and three Corr-AE methods [21]: Corr-AE, cross Corr-AE and full Corr-AE.

In the case of subspace learning methods, we used the implementation from [20] to compute the linear projection matrix. For Corr-AE methods, we use the implementation of [21] to calculate the hidden vectors of the two different modalities. We employed a 1024-dimensional hidden layer. For Corr-AE, cross Corr-AE and full Corr-AE the weight factors for reconstruction errors and correlation distances were set to 0.8, 0.2 and 0.8, respectively.

Dataset Splitting. We used three datasets in each experiment: Wikipedia, Flickr30k and FB5K. We split each dataset into a training set, a testing set, and a validation set, as illustrated below:

1. *Wikipedia dataset.* In the case of subspace learning, we used 2173 and 500 image-text pairs for training and testing respectively, while for Corr-AE methods a further 193 pairs served as a validation set. We utilized all of the data in a test set as a query.
2. *Flickr30k dataset.* For subspace learning, we used 15000 image-text pairs for both training and testing while for Corr-AE methods an additional 1783 image-text pairs were added for validation. We randomly selected 2000 images and texts from the test set to function as a query.
3. *FB5K dataset.* We split the dataset into 80% and 20% image-text pairs for training and testing respectively. We used the same split for subspace learning, while for Corr-AE methods 250 additional image-text pairs served as a validation set.

Representation. All images were first resized to dimension of 224×224 . Then we extracted the last fully connected (fc7) Convolution Neural Network (CNN) features using VGG16 [22] with CAFFE [23] implementation. Text representation was based on Latent Dirichlet Allocation (LDA) [24]. An LDA model was learned from all texts and used to compute the probability of each text under 50 hidden topics. We used this probability vector for text representation. A Bag-of-Word (BoW) model was used for text representation in Corr-AE methods. Initially, texts were converted into lower case, with all stopping-words removed. A unigram model was adopted to form a dictionary of the most recurrent 5000 words. Based on this dictionary, for each text we generated a 5000-dimensional BoW model.

Evaluation Parameters. We assessed the retrieval performance using Cumulative Match Characteristic (CMC) curves and mean rank. CMC is a useful approach that is used as the evaluation metric in many applications such as face recognition [25–27] and biometric systems [28, 29]. For cross-media retrieval, CMC can be illustrated by a curve of average retrieval accuracy with respect to the average ranks of the correct matches for a series of queries, K , where rank is:

$$Rank = \frac{1}{|K|} \sum_{x=1}^K rank_x, \quad (4)$$

$rank_x$ refers to the rank position of the correct match for the x^{th} query.

4.2 Retrieval Methods Comparison Using Different Datasets

We tested different cross-media retrieval methods on Flickr30k, Wikipedia and FB5K datasets. Figure 3 shows the effectiveness of the different retrieval methods. We drew several conclusions from this.

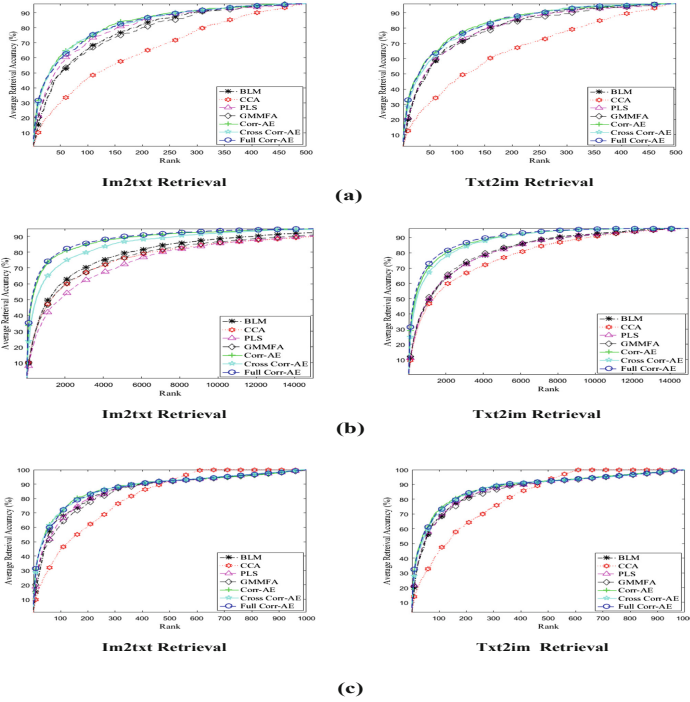


Fig. 3. CMC curves compared for different Corr-AE and subspace learning methods using different cross-media datasets: (a) Wikipedia, (b) Flickr30k, and (c) FB5K.

Corr-AE methods performed well compared to subspace learning methods with all three datasets. However, CCA showed a significant improvement in performance as the number of training samples increased using FB5K. The logic behind this is that correlation is ignored between different modalities in subspace learning when representation learning is performed. However, representation learning and correlation learning are merged into a single process in Corr-AE methods. Furthermore, Corr-AE is used to train a model by minimizing linear combinations of representation learning error and correlation learning error for individual modalities, and between hidden representations of two modalities [16]. This minimization of correlation learning error helps the model in learning hidden representations, while minimization of representation learning error makes better hidden representations to reconstruct the input of individual modalities.

The retrieval performance was highest for FB5K and lowest for Wikipedia. This shows that the tweets are highly correlated when using FB5K compared with Flickr30k and Wikipedia. The main reason that FB5K obtained the highest retrieval accuracy is twofold: first, it contained high-level concepts with small semantic gaps. Second, text and images were highly correlated in this dataset. We conclude that user descriptions on tweets are highly correlated to the scenarios.

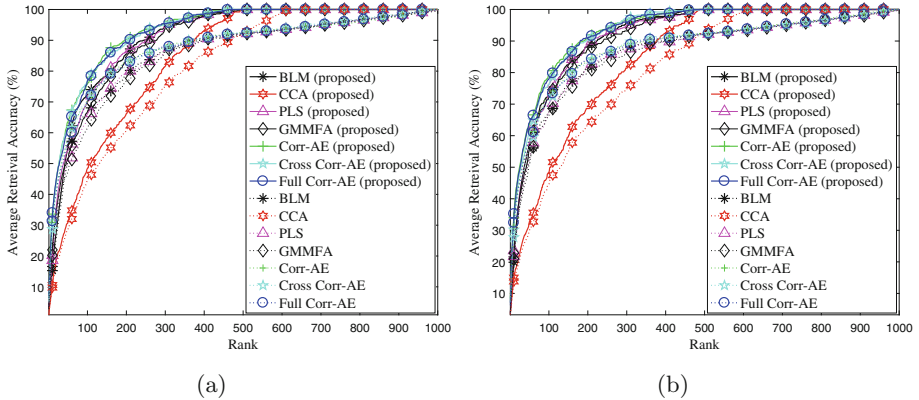


Fig. 4. CMC curves on FB5K with the proposed method and baselines. (a) Im2txt retrieval. (b) Txt2im retrieval.

4.3 Proposed Method Performance

In this section, we describe the proposed method for evaluation of FB5K. We compared its performance with the baseline methods.

Figure 4 clearly shows that using the proposed method in the baseline learning systems significantly improved their performance. In particular, using OCR, explicit incorporation of high-level semantic information, and a specially developed similarity measurement in the embedded space improved cross-media retrieval accuracy when similar retrieval methods were used. For example, in the case of Txt2im retrieval, CCA achieved 45% accuracy at rank = 110, whereas the BLM, PLS, and GMMFA methods achieved the same accuracy at ranks 20, 25 and 18, respectively. Incorporating the proposed method boosted the accuracy of CCA, BLM, PLS, and GMMFA to 6.5%, 4%, 5% and 7% respectively, at the same rank.

4.4 FB5K Retrieved Examples

This section describes the retrieval examples for FB5K using CCA and the proposed method.

Figure 5(a) shows the retrieval image results for different query tags. It shows that the proposed method was successful in learning colors, background and class information, e.g. in Fig. 5(a) we used the keyword *cold* to retrieve the images on right, which strongly indicate that the keyword information lay in the retrieved image. Moreover, incorporation of semantic class not only improved the retrieval accuracy, but also provided higher weights to more minor concepts during the formation of query tag vectors.

Figure 5(b) shows the tagging results retrieved by the proposed method on some test images. It is clear that using the proposed method with FB5K significantly outperformed the baseline methods, despite its diverse features.













Query tag	Retrieved images				
Cold					
Hungry					
(a)					
Query Image	Retrieved tags				
	Happy	Love	Laughing	Person	Shivering
	Cold	Shivering	Beautiful	Travel	Scenery
(b)					

Fig. 5. Retrieval examples for FB5K using CCA and the proposed method. The first two rows represent the query tag and its corresponding top five retrieved images, whereas the last two rows show query images and their corresponding top five retrieved tags. (a) tag/txt2img retrieval. (b) img2tag/txt retrieval.

Furthermore, FB5K provides information that is more realistic to the user. It incorporates high-level semantic information by providing the class probability for individual images. For example, in Fig. 5(b), with a query image of a baby, the proposed method retrieved *happy* and *love* as high frequency words in the retrieved text. This shows that despite the sentiment of an image being hidden under high-level concepts, opinion characteristics can have an impact on multi-modal retrieval.

5 Conclusion

This paper introduced a novel cross-media dataset called FB5K. We also presented a more realistic embedding approach for images, tags/texts, and their semantics. Specifically, in order to learn the cross-modal embeddings of user feelings, images and tags/texts, we developed a novel method by utilizing OCR, explicit incorporation of high-level semantic information, and a new similarity measurement in the embedded space, to improve the retrieval performance.

We believe that FB5K and the proposed cross-media retrieval method suffice as a reference guide for researchers and developers to facilitate the design and implementation of better evaluation protocols.

References

1. Lu, Y.-J., Nguyen, P.A., Zhang, H., Ngo, C.-W.: Concept-based interactive search system. In: Amsaleg, L., Guðmundsson, G.Þ., Gurrin, C., Jónsson, B.Þ., Satoh, S. (eds.) MMM 2017. LNCS, vol. 10133, pp. 463–468. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-51814-5_42
2. Kambau, R.A., Hasibuan, Z.A.: Concept-based multimedia information retrieval system using ontology search in cultural heritage. In: Second International Conference on Informatics and Computing (ICIC), pp. 1–6. IEEE (2017)
3. Hwang, S.J., Grauman, K.: Reading between the lines: object localization using implicit cues from image tags. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(6), 1145–1158 (2012)
4. Peng, Y., Huang, X., Zhao, Y.: An overview of cross-media retrieval: concepts, methodologies, benchmarks and challenges. *IEEE Trans. Circ. Syst. Video Technol.* **28**, 2372–2385 (2017)
5. Grubinger, M., Clough, P., Muller, H., Deselaers, T.: The IAPR TC-12 benchmark: a new evaluation resource for visual information systems. In: International Workshop onto Image, vol. 5, p. 10 (2006)
6. Rasiwasia, N., et al.: A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 251–260. ACM (2010)
7. Li, J., Wang, J.Z.: Real-time computerized annotation of pictures. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(6), 985–1002 (2008)
8. Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(3), 394–410 (2007)
9. Von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 319–326. ACM (2004)
10. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: LabelMe: a database and web-based tool for image annotation. *Int. J. Comput. Vis.* **77**(1–3), 157–173 (2008)
11. Wang, X.-J., Zhang, L., Jing, F., Ma, W.-Y.: Annosearch: Image auto-annotation by search. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1483–1490. IEEE (2006)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
13. Zheng, L., Wang, S., Liu, Z., Tian, Q.: Packing and padding: coupled multi-index for accurate image retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)
14. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **42**(3), 145–175 (2001)
15. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
16. Gong, Y., Ke, Q., Isard, M., Lazebnik, S.: A multi-view embedding space for modeling internet images, tags, and their semantics. *Int. J. Comput. Vis.* **106**(2), 210–233 (2014)
17. Haddoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.* **16**(12), 2639–2664 (2004)

18. Tenenbaum, J.B., Freeman, W.T.: Separating style and content. In: *Advances in Neural Information Processing Systems*, pp. 662–668 (1997)
19. Rosipal, R., Krämer, N.: Overview and recent advances in partial least squares. In: Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J. (eds.) *SLSFS 2005*. LNCS, vol. 3940, pp. 34–51. Springer, Heidelberg (2006). https://doi.org/10.1007/11752790_2
20. Sharma, A., Kumar, A., Daume, H., Jacobs, D.W.: Generalized multiview analysis: a discriminative latent space. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2160–2167. IEEE (2012)
21. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 7–16. ACM (2014)
22. Rasiwasia, N., Mahajan, D., Mahadevan, V., Aggarwal, G.: Cluster canonical correlation analysis. In: *Artificial Intelligence and Statistics*, pp. 823–831 (2014)
23. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 675–678. ACM (2014)
24. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**(Jan), 993–1022 (2003)
25. Rehman, S.U., Tu, S., Huang, Y., Liu, G.: CSFL: a novel unsupervised convolution neural network approach for visual pattern classification. *AI Commun.* **30**(5), 311–324 (2017)
26. Rehman, S.U., Tu, S., Huang, Y., Yang, Z.: Face recognition: a novel un-supervised convolutional neural network method. In: *IEEE International Conference of Online Analysis and Computing Science (ICOACS)*, pp. 139–144. IEEE (2016)
27. Rehman, S., et al.: Optimization of CNN through novel training strategy for visual classification problems. *Entropy* **20**(4), 290 (2018)
28. Damer, N., Opel, A., Nouak, A.: CMC curve properties and biometric source weighting in multi-biometric score-level fusion. In: *2014 17th International Conference on Information Fusion (FUSION)*, pp. 1–6. IEEE (2014)
29. Seha, S., Hatzinakos, D.: Human recognition using transient auditory evoked potentials: a preliminary study. *IET Biometrics*, *IET* **7**, 242–250 (2018)