A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2018,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2018 1

Facial Attribute Guided Deep Cross-Modal Hashing for Face Image Retrieval

Fariborz Taherkhani[1], Veeru Talreja[2], Hadi Kazemi[3], Nasser Nasrabadi[4]

Abstract: Hashing-based image retrieval approaches have attracted much attention due to their fast query speed and low storage cost. In this paper, we propose an Attribute-based Deep Cross Modal Hashing (ADCMH) network which takes facial attribute modality as a query to retrieve relevant face images. The ADCMH network can efficiently generate compact binary codes to preserve similarity between two modalities (i.e., facial attribute and image modalities) in the Hamming space. Our ADCMH is an end to end deep cross-modal hashing network, which jointly learns similarity preserving features and also compensates for the quantization error due to the hashing of the continuous representation of modalities to binary codes. Experimental results on two standard datasets with facial attributes-image modalities indicate that our ADCMH face image retrieval model outperforms most of the current attribute-guided face image retrieval approaches, which are based on hand crafted features.

Keywords: Facial Attributes, Face Image Retrieval, Deep Hashing Network.

1 Introduction

Semantic attributes have been significantly exploited by the computer vision society to improve performance of object recognition, face verification and image search. Facial attributes are invariant and semantic visual properties (i.e., visual properties that have names) which can be used to express contents of a face image in practice. For example, the content of a face image can be described by its facial attributes such as "a bald old man wearing glasses". Facial attributes have been used in a variety of computer vision applications such as face search engine and face image retrieval [KBN08, Ku09, TND18, Ka18, TVN17]. Cross-modal retrieval is a key type of image retrieval method which provides similarity retrieval across different modalities. In this paper, we address the problem of cross-modal retrieval of relevant face images in response to facial attributes queries by utilizing a deep cross-modal hashing framework.

A fast and an advantageous solution for an approximate binary nearest neighbors (ANN) search for image retrieval has been *hashing*. Hashing methods transform the high-dimensional

[1] West Virginia University, Lane Department of Computer Science and Electrical Engineering, Morgantown, WV, USA, fariborztaherkhani@gmail.com
[2] West Virginia University, Lane Department of Computer Science and Electrical Engineering, Morgantown, WV, USA, vtalreja@mix.wvu.edu
[3] West Virginia University, Lane Department of Computer Science and Electrical Engineering, Morgantown, WV, USA, hakazemi@mix.wvu.edu
[4] West Virginia University,Lane Department of Computer Science and Electrical Engineering, Morgantown, WV, USA, nasser.nasrabadi@mail.wvu.edu

media data into similarity-preserving binary codes for efficient image search. Using this methodology, ANN search can be done extremely fast by just calculating the hamming distance between the binary vectors. Furthermore, using binary hash codes to represent the original data also dramatically reduces the storage cost.

Learning-based hash approaches [AI06, Gi99, Go12] have become popular as they leverage the semantic similarity in the training samples for code-construction. However, many of these methods use hand-crafted features, which do not give satisfactory performance because feature extraction method for these hand crafted features is completely independent of the hash code learning procedure. To counter this issue it is very important to combine the feature extraction method and the hash code learning procedure in an end-to-end framework.

Recently, application of deep learning to hashing methods [Li16, Ca17] have shown that end-to-end learning of feature extraction and hash coding using deep neural networks is more efficient than using the hand-crafted features. Particularly, it proves crucial to jointly learn similarity preserving features and also control the quantization error of hashing continuous representation to binary codes.

In many applications, the data may have an image content and text content as well such as information tags from Flicker images. This kind of data is known as multi-modal data. There has been a surge in the development of multi-modal hashing (MH) techniques used for ANN search (retrieval) on multi-modal datasets. One very extensively used MH technique is cross-modal hashing (CMH). CMH returns relevant results of one modality in response to query of another modality, where respective hash codes in the same latent hamming space are generated for each individual modality. Most of the CMH techniques tackle the problem of text-based image retrieval (TBIR) and image-based text retrieval (IBTR). We are utilizing the cross-modal hashing framework for face image retrieval based on semantic attributes. In our framework, a user can simply query on the statement such as ”smiling old man with wavy hair” to retrieve relevant face images from a large dataset.

As already mentioned, the application deep learning to hashing methods give improved performance when compared to other hashing techniques. There also exist some methods [Ya17, JL17] which adopt deep learning for cross-modal hashing (CMH) and give improved performance over other CMH techniques which use handcrafted features [ZL14, ZY12].

We are looking to exploit the deep learning framework for cross modal hashing and retrieval of facial images in response to a facial attribute query. Searching for facial images of people in response to a facial attribute query has been investigated in the past [KBN08, Va09, SFD11]. Vaquero *et al*. [Va09] argued that face recognition could be challenging in surveillance scenarios and hence proposed to search for people in surveillance systems based on a parsing of human parts and their attributes, including facial hair, eyeglasses, clothing color, etc. Kumar *et al*. [KBN08] used a combination of Support Vector Machines and Adaboost to built an image search engine FaceTracer, which allows users to retrieve face images based on queries involving multiple visual attributes. However, these methods did not consider the correlation between attributes. Siddiquie *et al*.[SFD11]

proposed a ranking and image retrieval system for faces based on multi-attribute queries, which explicitly modeled the correlations that are present between the attributes.

However, all of these methods use hand-crafted features to perform a cross-modal retrieval. We present a novel CMH framework called ADCMH for attribute guided deep cross-modal hashing for face-image retrieval from large datasets. Main contributions of this paper include: (1) **Attribute guided deep cross modal hashing (ADCMH)** : We utilize deep cross modal hashing for face image retrieval in response to an attribute query which has not been done previously. (2) **Scalable cross-modal hash**: ADCMH performs facial image retrieval using point wise data, and thereby requires neither pairs nor triplets of training inputs. This characteristic makes it scalable to large scale datasets.

2 Facial Attribute Guided Cross Modal Hashing

The block diagram of the proposed framework ADCMH is given in Fig. 1. In this framework, we provide an algorithm which generates hash codes to retrieve relevant images from a database based on given facial attributes. The proposed algorithm contains three main components: 1) Learns a coupled neural network (one of which is used to represent image modality features while the other one is used to represent facial attribute modality features) via a distance-based logistic loss to preserve the cross-modal similarity. 2) Minimizes quantization loss between the original real-valued neural network output features for each modality and the learned hash codes to preserve high retrieval performance. 3) Maximizes the entropy corresponding to each bit to obtain the maximum information provided by the hash codes.

Assume that n is the number of training samples, each sample has two modalities, image and attribute features. We use $\mathbf{X} = \{\mathbf{x}_\mathrm{i}\}_{i=1}^{n}$ to represent the image modality, in which $\mathbf{x}_\mathrm{i}$ is the raw image i in a training set of size n. In addition, we use $\mathbf{Y} = \{\mathbf{y}_\mathrm{i}\}_{i=1}^{n}$ to represent the attribute modality, in which $\mathbf{y}_\mathrm{i}$ is the annotated facial attributes vector related to image i. Furthermore, we are given a cross-modal similarity matrix $\mathbf{S}$ in which $S_{ij} = 1$ if image $\mathbf{x}_\mathrm{i}$ contains a y_j facial attribute, and $S_{ij} = 0$ otherwise. Based on the given training information (i.e., $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{S}$), the main goal of ADCMH is to learn two hashing functions: $h^{(x)}(\mathbf{x}) \in \{-1,+1\}^c$ for image modality and $h^{(y)}(\mathbf{y}) \in \{-1,+1\}^c$ for attribute modality where c is the number of the bits used in the hash codes. The hash codes need to be learned such that the cross-modal similarity in $\mathbf{S}$ is preserved in the Hamming space, which implies that if $S_{ij} = 1$, the Hamming distance between the binary codes $\mathbf{c}_i^{(x)} = h^{(x)}(\mathbf{x}_\mathrm{i})$ and $\mathbf{c}_j^{(y)} = h^{(y)}(\mathbf{y}_\mathrm{j})$ should be small and if $S_{ij} = 0$, the corresponding Hamming distance should be large enough. Assume that $f(\mathbf{w}_x, \mathbf{x}_\mathrm{i}) \in \mathbb{R}^d$ represents the learned CNN features for sample $\mathbf{x}_\mathrm{i}$ corresponding to image modality, and $g(\mathbf{w}_y, \mathbf{y}_\mathrm{j})$ denotes the learned MLP features for sample $\mathbf{y}_\mathrm{j}$ corresponding to attribute modality. Here, $\mathbf{w}_x$ are the CNN network weights for image modality, and $\mathbf{w}_y$ are the MLP network weights for facial attribute modality as shown in Fig. 1. We define the total objective function for ADCMH as follows:

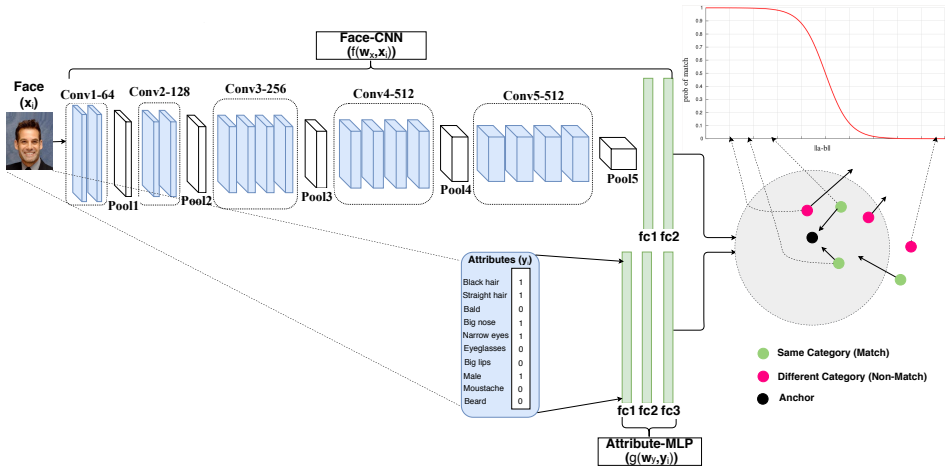


Fig. 1: Block Diagram of the ADCMH.

$$\min_{\mathbf{C}_{x,y},\mathbf{w}_x,\mathbf{w}_y} \mathscr{J} = \sum_{i=1}^{n}\sum_{j=1}^{n} \underbrace{\ell_c(p(\mathbf{F}_{*i},\mathbf{G}_{*j}),S_{ij})}_{\text{distance- based logistic loss}} + \alpha \underbrace{(||\mathbf{F}-\mathbf{C}_x||_F^2 + ||\mathbf{G}-\mathbf{C}_y||_F^2)}_{\text{quantization loss}} + \beta \underbrace{(||\mathbf{F1}||_F^2 + ||\mathbf{G1}||_F^2)}_{\text{entropy maximization}} \quad \text{s.t.} \quad \mathbf{C}_{x,y} \in \{+1,-1\}^{c\times n}, \tag{1}$$

where $\mathbf{F} \in \mathbb{R}^{c\times n}$ is the image feature matrix constructed by placing CNN features of training samples column-wise and $\mathbf{F}_{*i} = f(\mathbf{w}_x, \mathbf{x}_i)$ is the CNN feature corresponding to sample $\mathbf{x}_i$. Likewise, $\mathbf{G} \in \mathbb{R}^{c\times n}$ is the facial attribute feature matrix constructed by placing MLP features of training samples column-wise; $\mathbf{C}_x$ is the binary hash code matrix for image modality and $\mathbf{C}_y$ is the binary hash code matrix for attribute modality. Notation $\mathbf{1}$ represents a vector that all its elements are set to 1.

$p(\mathbf{F}_{*i},\mathbf{G}_{*j}) = \frac{1+exp(-m)}{1+exp(||\mathbf{F}_{*i}-\mathbf{G}_{*j}||-m)}$ is distance-based logistic probability; this function returns a value between 0 and 1 which represents the probability of the match between two feature vectors $\mathbf{F}_{*i}$ and $\mathbf{G}_{*j}$ from image and attribute modalities respectively, given their squared distance. Then we can use the cross entropy loss similar to the classification case for optimization: $\ell_c(p,s) = -slog(p) + (s-1)log(1-p)$. This loss function tries to bring features of the two modalities referring to the same sample close to each other while push them away if they refer to two different samples. The m is the margin parameter that matched/non-matched samples are pushed away from it in the inward/outward direction.

The second term is added to the objective function to generate the hash codes by setting $\mathbf{C}_x = \text{sign}(\mathbf{F})$ and $\mathbf{C}_y = \text{sign}(\mathbf{G})$. Therefore, we can consider that $\mathbf{F}$ and $\mathbf{G}$ are the continuous surrogates of $\mathbf{C}_x$ and $\mathbf{C}_y$, respectively. Because $\mathbf{F}$ and $\mathbf{G}$ preserve the cross-modal similarity in $\mathbf{S}$, this second term helps us to preserve the cross-modal similarity even in the binary domain using hash codes $\mathbf{C}_x$ and $\mathbf{C}_y$. The third term is added to the ADCMH loss

function to make each bit in the hash code to be balanced for all the training samples since minimizing $||\mathbf{F1}||_F^2$ and $||\mathbf{G1}||_F^2$ makes sum of all elements almost zero (i.e., number of $+1$ and -1 in each bit of the hash code on all the training be roughly same). Therefore, we can say that probability of appearing -1 is almost equal to that of $+1$ and this is equivalent to maximizing the entropy on the bits of the hash code.

During the experiments, we have noticed that we get better performance if we set $\mathbf{C}_x = \mathbf{C}_y = \mathbf{C}$ for our training points which means the binary codes from face modality and attribute modality are set to be same for the same training points. Note that this is true only for the training points. During testing, we will have to generate different hash codes for two different modalities for the same point if the point is a query point or a database point.

Parameters Learning: We used an alternating minimization (optimization) algorithm to learn network parameters $\mathbf{w}_x$, $\mathbf{w}_y$ and hash codes $\mathbf{C}$. In this algorithm, each time we optimize one parameter keeping other parameters fixed and when the algorithm converges, the converged result is returned as the solution.

Learning ($\mathbf{w}_x$) Parameter : In this step, we fix $\mathbf{w}_y$ and $\mathbf{C}$ and optimize the CNN parameters $\mathbf{w}_x$ for the image modality by using back propagation algorithm. We first compute loss function gradient with respect to output of image modality network as follows:

$$\frac{\partial \mathscr{J}}{\partial \mathbf{F}_{*i}} = \sum_{j=1}^{n} \frac{\partial \ell_c(p(\mathbf{F}_{*i}, \mathbf{G}_{*j}), S_{ij})}{\partial \mathbf{F}_{*i}} + 2\alpha(\mathbf{F}_{*i} - \mathbf{C}_{*i}) + 2\beta\mathbf{F1}. \tag{2}$$

The gradient of the first term in Eq. 2 is calculated as follows:

$$\frac{\partial \ell_c(p(\mathbf{F}_{*i}, \mathbf{G}_{*j}), S_{ij})}{\partial \mathbf{F}_{*i}} = \frac{-(1+exp(-m))}{(1+exp(||\mathbf{F}_{*i} - \mathbf{G}_{*j}|| - m))^2} \times (\frac{S_{ij}}{p(\mathbf{F}_{*i}, \mathbf{G}_{*j})} + \frac{1 - S_{ij}}{1 - p(\mathbf{F}_{*i}, \mathbf{G}_{*j})}).$$

In the next step, we compute $\frac{\partial \mathscr{J}}{\partial \mathbf{w}_x}$ with $\frac{\partial \mathscr{J}}{\partial \mathbf{F}_{*i}}$ by using the chain rule ($\frac{\partial \mathscr{J}}{\partial \mathbf{w}_x} = \frac{\partial \mathscr{J}}{\partial \mathbf{F}_{*i}} \times \frac{\partial \mathbf{F}_{*i}}{\partial \mathbf{w}_x}$), based on which back propagation is used to update the parameter $\mathbf{w}_x$.

Learning ($\mathbf{w}_y$) Parameter : Similar to the previous step, we fix $\mathbf{C}$ and $\mathbf{w}_x$ parameters and we optimize MLP network parameters $\mathbf{w}_y$ for the facial attribute modality by using the back propagation algorithm. We first compute the loss function gradient with respect to the output of the facial attribute network as follows:

$$\frac{\partial \mathscr{J}}{\partial \mathbf{G}_{*j}} = \sum_{i=1}^{n} \frac{\partial \ell_c(p(\mathbf{F}_{*i}, \mathbf{G}_{*j}), S_{ij})}{\partial \mathbf{G}_{*j}} + 2\alpha(\mathbf{G}_{*j} - \mathbf{C}_{*j}) + 2\beta\mathbf{G1}. \tag{3}$$

In the next step, we compute $\frac{\partial \mathscr{J}}{\partial \mathbf{w}_y}$ with $\frac{\partial \mathscr{J}}{\partial \mathbf{G}_{*j}}$ by using the chain rule ($\frac{\partial \mathscr{J}}{\partial \mathbf{w}_y} = \frac{\partial \mathscr{J}}{\partial \mathbf{G}_{*j}} \times \frac{\partial \mathbf{G}_{*j}}{\partial \mathbf{w}_y}$), based on which the back propagation algorithm is used to update the parameter $\mathbf{w}_y$. The gradient of the first term in Eq. 3 is similar to what we had for $\mathbf{F}_{*i}$ in (2), but with the negative sign.

Fig. 2: Qualitative results: Retrieved images using ADCMH for given facial attributes.

Learning Hash Code (C): Problem in Eq. 1 in can be reformulated as follows when $\mathbf{w}_x$ and $\mathbf{w}_y$ are fixed:

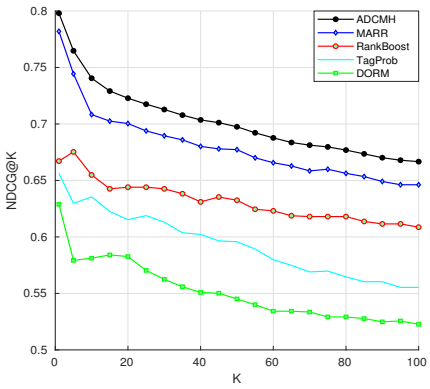
$$\max_{\mathbf{C}} \operatorname{tr}(\mathbf{C}^\top(\alpha(\mathbf{F}+\mathbf{G}))) = \operatorname{tr}(\mathbf{C}^\top\mathbf{D}) = \sum_{i,j} C_{ij}D_{ij} \quad \text{s.t. } \mathbf{C} \in \{+1,-1\}^{c\times n}, \tag{4}$$

where $\mathbf{D} = \alpha(\mathbf{F}+\mathbf{G})$. It can easily be shown that the hash code C_{ij} should preserve the same sign as D_{ij}; thus we can obtain $\mathbf{C}$ as follows: $\mathbf{C} = \mathsf{sign}(\mathbf{D}) = \mathsf{sign}(\alpha(\mathbf{F}+\mathbf{G}))$.

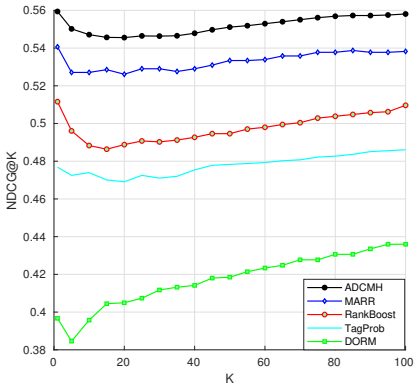
3 Experimental Results

Implementation: As shown in Fig. 1, the proposed framework of ADCMH network is composed of two networks : Convolutional Neural Network (CNN) and Multi-layer Perceptron (MLP). CNN is used to extract features for image modality while the MLP is used to extract features for facial attribute modality. For CNN network, we have used VGG19 network with the same filter size, convolutional layers, and pooling operation for learning the image modality features. However, the number of nodes in the last fully connected layer is the hash code length (the code length in all the experiment is 64 bits). We initialize our CNN parameters by a VGG19 pre-trained using the ImageNet dataset and then we fine tune it as a classifier by using the CASIA-Web Face dataset, which contains 10,575 subjects and 494,414 images. Fine-tuning using CASIA-Web Face is only performed for the image-modality as it helps the CNN to learn better and specialized set of facial features.

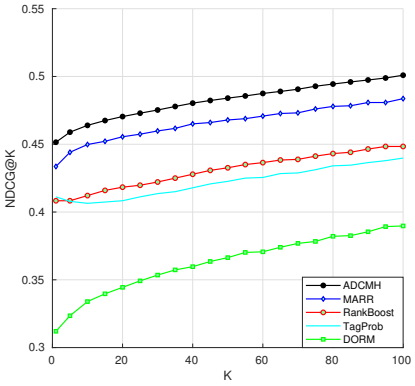
For MLP network, we have used three fully connected layers to learn features for facial attribute modality. To learn attribute features from this network, we first represent each input image with a vector formed by $1's$ and $0's$ which indicate the presence or absence of corresponding facial attribute, respectively. This facial attribute vector is used as input to the MLP network. The first and second layer in the MLP network contains 4096 nodes and the number of nodes in the last fully connected layer is the hash code length (64 bits for our experiments). The activation function for the first and second layers is ReLU, and



(a) Single Attribute Queries



(b) Double Attribute Queries



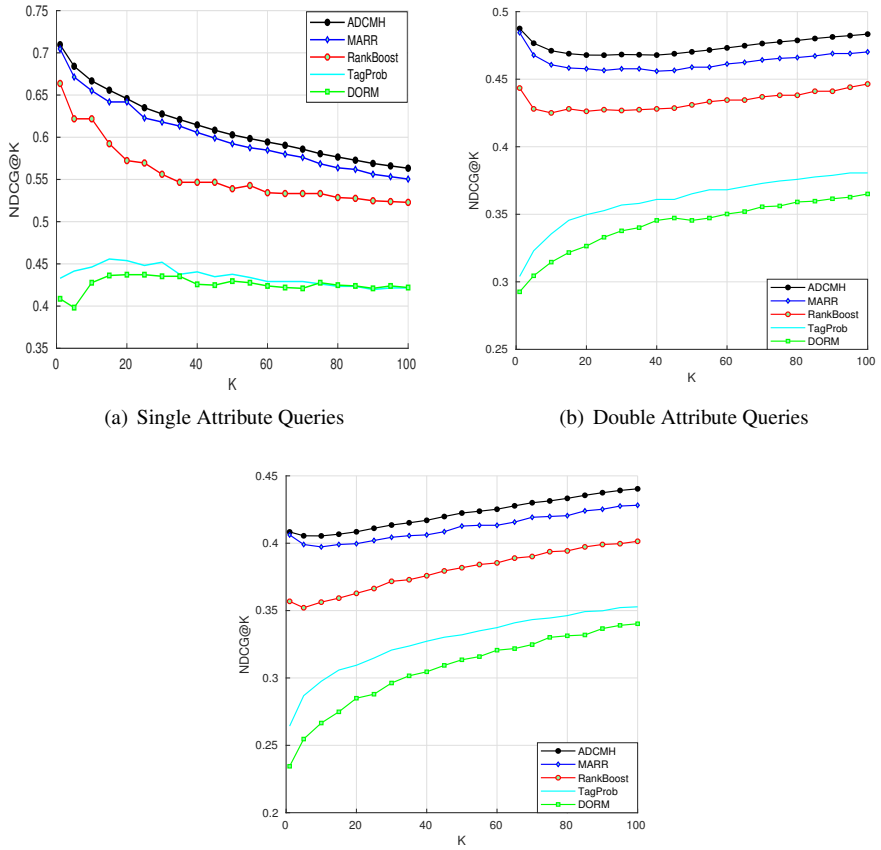
(c) Triple Attribute Queries

Fig. 3: Ranking performance on the LFW dataset.

for the third layer is the identity function. The weights of the MLP network are initialized by sampling randomly from $\mathcal{N}(0, 0.01)$ except for the bias parameters that are initialized with zeros. Note that we do not use CASIA-Web Face for fine-tuning the MLP because CASIA-Web Face does not provide annotated facial attributes.

We use the Adam optimizer [KB14] with the default hyper-parameter values ($\varepsilon = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) to train all the parameters using alternative minimization approach. The batch size in all the experiments is fixed to 128. Furthermore, during the experiments, we noticed that our ADCMH is not sensitive to hyper-parameters α and β when they are in the range $[1.5, 2.5]$. Our ADCMH is implemented in TensorFlow with python API and all the experiments are conducted on two GeForce GTX TITAN X 12GB GPUs.

Datasets: We evaluated ADCMH performance on two face datasets including the LFW [Hu07] and FaceTracer [KBN08] annotated by facial attributes. **LFW** is a well-known



(a) Single Attribute Queries

(b) Double Attribute Queries

(c) Triple Attribute Queries

Fig. 4: Ranking performance on the FaceTracer dataset.

dataset of more than 13,000 images of faces collected from the internet for face recognition as well as attribute classification. The **FaceTracer** database is a large collection of 15000 real-world face images, collected from the internet. For comparison purposes, we have been consistent with the train and test split of these datasets as used in MARR [SFD11].

Evaluation Results: We use NDCG (normalized discounted cumulative gain) to compare ADCMH performance with other methods. NDCG is a standard single-number measure of ranking quality that allows non-binary relevance judgments. It is defined as $NDCG@k = \frac{1}{Z}\sum_{i=1}^{k}\frac{2^{rel(i)}-1}{log(i+1)}$, where rel(i) is the relevance of the i^{th} ranked image and Z is a normalization constant to ensure that the correct ranking results in an NDCG score of 1. We have compared our retrieval and ranking results with some of the other state-of-the-art ranking approaches including Multi Attribute Retrieval and Ranking (MARR) [SFD11], rankBoost [Fr03], Direct Optimization of Ranking Measures (DORM) [LS07], TagProp [Gu09].

Fig. 2 indicates the qualitative result of ADCMH approach for the given facial attributes. Fig. 3 and Fig. 4 plots the NDCG scores, as a function of the ranking truncation level K, using different number of attribute queries for the LFW and FaceTracer dataset, respectively. From the Fig. 3 and Fig. 4, it is clear that our approach (ADCMH) is significantly better than the other methods for all three types of queries, at all values of K. For LFW dataset, at a truncation level of 20 (NDCG@20), for single, double and triple attribute queries, ADCMH is respectively, 2.5%, 2.5% and 0.5% better than MARR, the second best method. We can observe that the NDCG values for the FaceTracer dataset for all methods are relatively lower when compared to the LFW dataset. This is due to the difference in the distributions of the two datasets. For comparison with previous state-of-the-art ranking approaches, we have only used single, double and triple attribute queries for our ranking performance. However, we have used more than 3 queries to test our system as shown in the qualitative results in Fig. 2.

From the results, we can observe that our proposed deep-hashing based face image retrieval method ADCMH outperforms the other methods, which use hand-crafted features for face image retrieval.

4 Conclusion

In this paper, we proposed a facial attribute-based algorithm using deep hashing network to retrieve relevant images from the database.The method takes facial attributes as query and returns a list of images based on a Hamming distance similarity. The experimental results show that our method outperforms most of the current face image retrieval approaches.

References

[AI06] Andoni, A.; Indyk, P.: Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions. In: 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06). pp. 459–468, Oct 2006.

[Ca17] Cao, Zhangjie; Long, Mingsheng; Wang, Jianmin; Yu, Philip S: HashNet: Deep Learning to Hash by Continuation. arXiv preprint arXiv:1702.00758, 2017.

[Fr03] Freund, Yoav; Iyer, Raj; Schapire, Robert E; Singer, Yoram: An efficient boosting algorithm for combining preferences. Journal of machine learning research, 4(Nov):933–969, 2003.

[Gi99] Gionis, Aristides; Indyk, Piotr; Motwani, Rajeev et al.: Similarity search in high dimensions via hashing. In: VLDB. volume 99, pp. 518–529, 1999.

[Go12] Gong, Yunchao; Kumar, Sanjiv; Verma, Vishal; Lazebnik, Svetlana: Angular quantization-based binary codes for fast similarity search. In: Advances in neural information processing systems. pp. 1196–1204, 2012.

[Gu09] Guillaumin, M.; Mensink, T.; Verbeek, J.; Schmid, C.: TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 309–316, Sept 2009.

[Hu07] Huang, Gary B; Ramesh, Manu; Berg, Tamara; Learned-Miller, Erik: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[JL17] Jiang, Qing-Yuan; Li, Wu-Jun: Deep Cross-Modal Hashing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3232–3240, 2017.

[Ka18] Kazemi, Hadi; Soleymani, Sobhan; Dabouei, Ali; Iranmanesh, Mehdi; Nasrabadi, Nasser M: Attribute-Centered Loss for Soft-Biometrics Guided Face Sketch-Photo Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 499–507, 2018.

[KB14] Kingma, Diederik P; Ba, Jimmy: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[KBN08] Kumar, Neeraj; Belhumeur, Peter; Nayar, Shree: Facetracer: A search engine for large collections of images with faces. In: European conference on computer vision. Springer, pp. 340–353, 2008.

[Ku09] Kumar, Neeraj; Berg, Alexander C; Belhumeur, Peter N; Nayar, Shree K: Attribute and simile classifiers for face verification. In: Computer Vision, 2009 IEEE 12th International Conference on. IEEE, pp. 365–372, 2009.

[Li16] Liu, H.; Wang, R.; Shan, S.; Chen, X.: Deep Supervised Hashing for Fast Image Retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2064–2072, June 2016.

[LS07] Le, Quoc V.; Smola, Alexander J.: Direct Optimization of Ranking Measures. CoRR, abs/0704.3359, 2007.

[SFD11] Siddiquie, B.; Feris, R. S.; Davis, L. S.: Image ranking and retrieval based on multi-attribute queries. In: CVPR 2011. pp. 801–808, June 2011.

[TND18] Taherkhani, Fariborz; Nasrabadi, Nasser M; Dawson, Jeremy: A Deep Face Identification Network Enhanced by Facial Attributes Prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 553–560, 2018.

[TVN17] Talreja, V.; Valenti, M. C.; Nasrabadi, N. M.: Multibiometric secure system based on deep learning. In: Proc. IEEE Global Conference on Signal and Information Processing. pp. 298–302, Nov 2017.

[Va09] Vaquero, D. A.; Feris, R. S.; Tran, D.; Brown, L.; Hampapur, A.; Turk, M.: Attribute-based people search in surveillance environments. In: 2009 Workshop on Applications of Computer Vision (WACV). pp. 1–8, Dec 2009.

[Ya17] Yang, Erkun; Deng, Cheng; Liu, Wei; Liu, Xianglong; Tao, Dacheng; Gao, Xinbo: Pairwise Relationship Guided Deep Hashing for Cross-Modal Retrieval. In: AAAI. 2017.

[ZL14] Zhang, Dongqing; Li, Wu-Jun: Large-scale Supervised Multimodal Hashing with Semantic Correlation Maximization. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. AAAI'14. AAAI Press, pp. 2177–2183, 2014.

[ZY12] Zhen, Yi; Yeung, Dit-Yan: Co-Regularized Hashing for Multimodal Data. In (Pereira, F.; Burges, C. J. C.; Bottou, L.; Weinberger, K. Q., eds): Advances in Neural Information Processing Systems 25, pp. 1376–1384. Curran Associates, Inc., 2012.