Survey on Content Based Face Image Retrieval and Attribute Detection

Bharti S. Satpute[1], Archana A. Chaugule[2]

[1]ME Computer Student, DYPIET Pimpri, SavitriBai Phule Pune University, India

[2]Assistant Professor in Computer Department, DYPIET Pimpri, SavitriBai Phule Pune University, India

Abstract: ***Survey on content based face image retrieval and attribute detection***

Keywords: content-based image retrieval; sparse coding; content-based face image retrieval; attribute detection.

1. Introduction

Due to the popularity of digital devices such as digital cameras, image scanners, people can easily capture a photo and share it using the internet by various online tools like facebook, flickr, twitter, etc. Among vast digital images and photos shared on the internet, a big percentage of them are photos related to human faces .Because human faces are closely related to social activities of human beings. The rapid growth of facial images has created many research problems and opportunities for a variety of real-world applications.

Traditional CBIR systems mostly use low level visual features such as color, texture and shape features to represent images. Visual features of images are extracted automatically using image processing methods to represent the raw content of the image. Image retrieval based on color usually retrieves images with similar colors and image retrieval based on shape yields images that have clearly the same shape, etc. From that, such system that is used for the general purpose of image retrieval using the low level features is not effective with face images, especially when user's query is a kind of a verbal description, since it does not capture the semantic aspects of a face, while humans in their nature tend to use the verbal description of the semantic features (high level features) for describing what they looking for, and they encounter a difficulty to use language of low-level features. Human beings normally perceive facial images and compare their similarities using high-level features such as gender, race, age and the rating scale of the facial traits, and thus cannot relate these high-level semantic concepts directly to low-level features. Systems use the visual features, actually based on query by example strategy for adverse through the image database. If an example image is unavailable, it is not likely for such systems to perform the task of facial image retrieval efficiently. Generally, facial images are different from other images, because facial images are intricate, multidimensional and similar in overall configuration.

Facial images usually have high intra-class variances caused by expressions, poses and illuminance (lighting variations). Due to these intra-class variances in face images, content-based face image retrieval is very challenging problems. In content-based face image retrieval system, when input image is given, then it tries to find most similar images from a large image database. It is enabling technology for many applications including automatic face annotation, crime investigation etc.

This paper is closely related to several different research topics, including content-based image retrieval (CBIR), human attribute detection, and content-based face image retrieval. The rest of the paper is organized as follows. In Section 2, we briefly discussed various CBIR systems. Power of automatically detected human attributes is discussed in Section 3. In section 4, content-based face image retrieval frameworks discussed. Section 5 gives review on some sparse coding and dictionary learning techniques and section 6 conclude this paper.

2. Content-based Image Retrieval

Content-based image retrieval (CBIR) is a technique to automatically index images by extracting their (low-level) visual content, such as color, texture, and shape, and the retrieval of images is based solely upon the indexed image features. Mainly two kinds of indexing systems are used, to deal with large scale data. Many studies have move with inverted indexing or hash-based indexing combined with bag-of-word model (BoW) and local features like scale-invariant feature transform (SIFT), to achieve efficient similarity search. The bag-of-words model is a well-known and popular feature representation method for image categorization and annotation tasks. The key idea is to quantize high-dimensional local features into one of visual words, and then represent each image by a histogram of the visual words. For this purpose, a clustering algorithm (*e.g.*, K-means), is generally used to generate a codebook (or vocabulary) by converting the visual features to codewords or visual words. However, traditional BoW-like methods not succeed to address issues related to noisily quantize visual features and also problems related to variations in viewpoints, lighting conditions, etc., commonly observed in large-scale image datasets.These methods can achieve high precision on rigid object retrieval, but they suffer from low recall rate due to the semantic gap. In recent times, some researchers have focused on bridging the semantic gap by finding semantic image representations to improve the CBIR performance.

Wu et al. [1] propose a novel Semantic-Preserving Bag of Word model to learn optimized BoW models, by considering the distance between semantically identical features as a measurement of the semantic gap, and attempt to learn an optimized codebook by minimizing this gap, aiming to achieve the minimal loss of the semantics using effective distance metric learning. Kuo et al. [2] propose to enhance each database image with semantically related auxiliary visual words (AVWs).But these methods might require intensive human annotations (and tagging) to use extra information to construct semantic codewords. Figure 1.shows architecture of generalized content-based image retrieval systems.

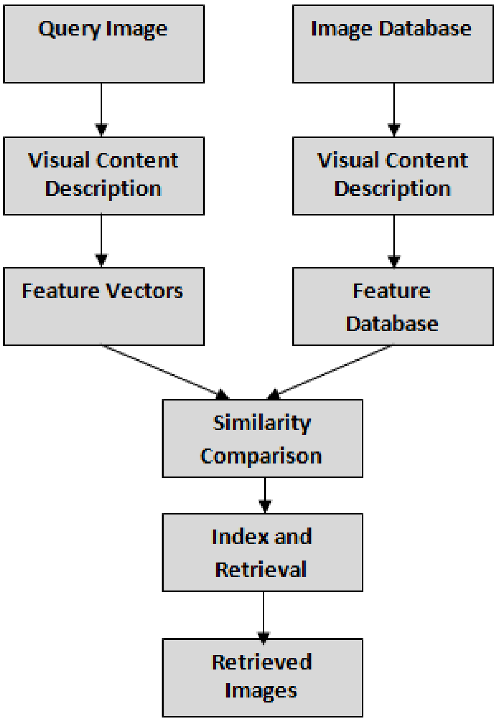

Figure 1: Generalized content-based systems.

3. Human Attribute Detection

Human attributes are high-level semantic descriptions about a person (e.g., gender, age, hair style, skin color). The recent work shows automatically detected human attributes have achieved promising results in different applications. The advantages of a describable human attributes are manifold: they are composable; they are generalizable, as from large image collections one can learn a set of attributes and then apply them to recognize new objects or categories without any further training; and attributes are also efficient. N. Kumar et al. [3] propose a learning framework to automatically detect describable aspects of visual appearance. In their approach, an extensive vocabulary of visual attributes is used to label a large data set of images, which is then used to train classifiers which measures the presence, absence, or degree to which an attribute is expressed in images and then these attribute classifiers can automatically label new images. First large number of images are collected from Internet using various online tools which having vast variations. Then, commercial face detector used to extract faces and fiducial points from downloaded images and stored in the Columbia Face Database. Facial images from Columbia Face Database are submitted to the Amazon Mechanical Turk (MTurk) service, for labelling images with attributes and identity. From these attribute and identity labels and face database, two publicly available face data Sets created, namely FaceTracer and PubFig data sets, respectively which have been publicly released for non-commercial use. A set of labeled positive and negative images for each attribute are requires for training attribute classifiers. For that purpose, all types of low-level features from the whole face are extracted and automatic, iterative selection procedure designed to select best features from a rich set of low-level feature options. The selected features are used to train the attribute or simile classifier. Using automatically detected human attributes with the help of attribute classifiers, they achieve excellent performance on face verification and keyword-based image search.

B. Siddiquie et al. [4] further extend the framework for ranking and retrieval of images based on multi-attribute queries. They propose a principled approach for multi-attribute keyword-based face image retrieval which explicitly models the correlations that are present between the different attributes which leading to improved ranking/retrieval performance. This recent works demonstrate the emerging opportunities for the human attributes but are not used to generate more semantic (scalable) codewords. Although these works achieve salient performance on keyword-based face image retrieval and face recognition, Chen et al.[8] further extend framework to exploit effective ways to combine low-level features and automatically detected facial attributes for scalable content based face image retrieval. To further improve quality of attributes, techniques based on the statistical Extreme Value Theory used by W. Scheirer et al. [16] to propose multi-attribute space to normalize the confidence scores from different attribute detectors for similar attribute search.

4. Content-based Face Image Retrieval

Now a days the rise of photo sharing/social network services (e.g. flickr, picasa, quickr etc.), needs for efficient large-scale content-based face image retrieval rises. Content-based face image retrieval task is closely related to face recognition task. The difference between face recognition and face retrieval is that face recognition requires completely labeled data in the training set, and it uses learning based approach to find classification result while in face retrieval task neither training set nor learning process is needed and it provides a ranking result. Face retrieval task focus on finding suitable feature representations for scalable indexing systems due to its high dimensionality. For now, facial images are more diverse and pose more visual variances (in poses, expressions, lighting effects etc.).

Conventional methods for face image retrieval usually use low-level features to represent face images which have lack of semantic meanings and face images generally have high intra-class variance (e.g., expression, posing), so the retrieval results are not good enough. To deal with this problem, Z .Wu et al. [7] proposes to use identity based quantization scheme and multi-reference re-ranking for scalable face image retrieval. Using bag-of-words representation and textual retrieval methods content-based image retrieval systems achieve scalability, but performance of such a system degrades quickly in the face image domain, mainly because they 1) produce visual words with low discriminative power for face images, and 2) ignore the special properties of the faces. This paper [7] develops a new scalable face representation using both local and global features. They exploit special properties of faces to design new component-based local features and then identity-based quantization scheme is use to quantize local features into discriminative visual words. In addition to local features, they also use a very small hamming signature (40 bytes) to encode the discriminative global feature for each face and then re-rank the top retrieved candidate images using hamming signature .It's improve the precision but without losing the scalability. But in this identity-based quantization scheme, construction of visual word vocabulary requires manual annotation (and tagging). Automation of this process is necessary to further improve the visual word vocabulary for face.

Wang et al. [6] investigated the retrieval-based face annotation problem and presented a promising framework to tackle this challenge by mining massive weakly labeled facial images freely available on Internet. A novel Weak Label Regularized Local Coordinate Coding (WLRLCC) algorithm was proposed to improve the annotation performance. This algorithm effectively exploits the principles of both local coordinate coding and graph-based weak label regularization.

B. C. Chen et al. [5] developed a scalable face image retrieval system using component-based local binary pattern (LBP) combined with sparse coding and which can integrate with partial identity information to improve the retrieval result. To achieve this goal, they first apply sparse coding on local features extracted from face images combining with inverted indexing to construct an efficient content-based face retrieval system. Then propose a novel coding scheme that refines the representation of the original sparse coding by using identity information. Their experimental results shows that the system can achieve salient retrieval results on LFW dataset (13K faces) and also achieve performance than linear search methods based on well-known face recognition feature descriptors. But disadvantage of this proposed coding scheme is that, face images with large intra-class variances will still be quantized into similar visual words if they share the same identity and identity information might need manual annotations. Figure 2 shows framework of face image retrieval using sparse coding with identity constraint method.

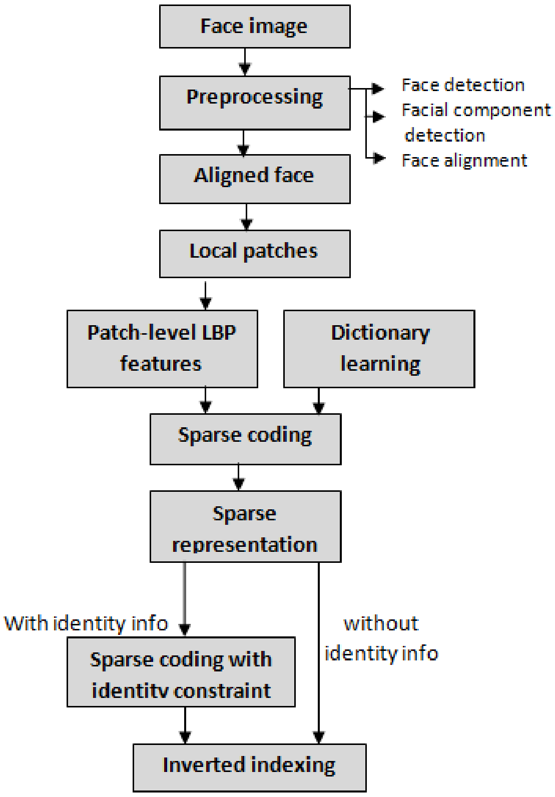

Figure 2 : Face image retrieval using sparse coding with identity constraint.

Face images of two different people might be very close in the traditional low-level feature space, because low-level features are lack of semantic meanings. Existing system only using low-level features and hence, retrieval result was not satisfactory. B. C. Chen et al. [8] provide a new perspective on content-based face image retrieval by combining low-level features with automatically detected high-level human attributes, they find better feature representations of face images and achieve better retrieval results. Although human attributes have been shown useful in different applications related to face images such as face recognition, keyword-based image retrieval and similar attribute search, it is non-trivial to apply it in content-based face image retrieval task because of following reasons. First, human attributes only contain limited dimensions. It loses discriminability, when there are too many people in the database because certain people might have similar attributes. Second, human attributes are represented as a vector of floating points. It does not work well with developing large-scale indexing methods, and therefore it suffers from slow response and scalability issue when the data size is huge. In this paper [8], two orthogonal methods are proposed, *attribute-enhanced sparse coding* and *attribute-embedded inverted indexing*. Attribute-enhanced sparse coding uses human attributes automatically detected by attribute detectors combined with low-level features to construct semantic codewords in the offline stage. Attribute-embedded inverted indexing further considers the local attribute signature of the query image in addition to sparse codewords computed from the facial appearance and gives efficient retrieval in the online stage. These methods treat all attributes as equal and further exploitation of contextual relationships between attributes needed.

5. Sparse Coding and Dictionary Learning

Image representation is a critical procedure for image processing and understanding, in computer vision. sparse coding is a powerful tool for capturing high-level semantics in visual data and represent images using only a few active coefficients. Hence, sparse representation on features provides efficient content-based image indexing and retrieval and has been successfully applied in many different applications such as image classification and face recognition. Sparse coding approximates to an expression of the input signal, y, by a sparse linear combination of items in which many of the coefficients are zero, from an over-complete dictionary D. The performance of sparse coding relies on the quality of learned dictionary D.

A well-known bag-of-features (BoF) approach ignores the spatial order of local descriptors, which rigorously limits the descriptive power of the image representation. To tackle this problem, spatial pyramid matching (SPM), one particular extension of the BoF model, has made a remarkable success on a range of image classification benchmarks like Caltech-101 and Caltech-256, and was the major component of the state-of-the-art systems. J. Yang et al. [9] propose an extension of the spatial pyramid matching approach. Their method computes a spatial-pyramid image representation based on sparse codes (SC) of SIFT features, instead of traditional K-means vector quantization to extract salient properties of appearance descriptors of local image patches. Further, this approach uses max spatial pooling that is more robust to local spatial translations, unlike the original non-linear SPM that performs spatial pooling by computing histograms. This new image representation works surprisingly well with simple linear classifiers which dramatically reduces the training complexity to O(n), and obtains a constant complexity in testing, while still improves classification accuracy in comparison with the traditional nonlinear SPM approach. This approach can be making faster with help of sparse coding by using a feed-forward network and accuracy could be further improved by learning the codebook in a supervised fashion.

Coding and classification, these are two phases of in sparse representation based classification. In first phase, the query image is collaboratively coded over a dictionary of atoms with some sparsity constraint, and then in second phase, classification is performed based on the coding coefficients and the learned dictionary. J. Wright et al. [10], propose a general classification algorithm for face recognition based on a sparse representation computed by l_1-minimization. Their framework addressing two crucial issues in face recognition namely feature extraction and robustness to occlusion and achieve state-of-the-art performance. Although their sparse representation based classification (SRC) scheme shows interesting face retrieval results, the dictionary used in it may not be effective enough to represent the query images because training samples of all classes directly used as the dictionary to code the query face image and original training images may contain uncertain and noisy information. Coding complexity also increases due to very large number of atoms of such a dictionary. In addition, using the original training samples as the dictionary could not fully exploit the discriminative information hidden in the training samples.

Various Dictionary Learning (DL) methods have been projected for image processing and classification. One representative DL method for image processing is the KSVD algorithm [11], which learns an over-complete dictionary from a training dataset of natural image patches. However, KSVD is not appropriate for classification tasks because it only requires that the learned dictionary could faithfully represent the training samples. To gain discrimination ability, Mairal et al. [12] added a discriminative reconstruction constraint in the DL model based on KSVD and used the learned dictionary for texture segmentation and scene analysis. However, this method does not explore the discrimination capability of sparse coding coefficients. Mairal et al. [13] further uses a trained classifier of the coding coefficients to proposed a discriminative DL method and verified their method for digit recognition and texture classification. Based on [11], Q. Zhang et al. [14] proposed an algorithm called discriminative KSVD (DKSVD) for face recognition. All the works in [13], [14] and [11] try to learn a common dictionary shared by all classes, as well as a classifier of coefficients for classification. However, the shared dictionary loses the correspondence between dictionary atoms and the class labels, and hence performing classification based on the reconstruction error associated with each class is not allowed. M. Yang et al. [15] present a Fisher Discrimination Dictionary Learning (FDDL) approach to improve the pattern classification performance. They have shown, a new classification scheme associated with the proposed Fisher discrimination DL (FDDL) method using both the discriminative information in the reconstruction error and sparse coding coefficients achieve state-of-art performance.

6. Conclusion

This paper provides survey on feature representation, work towards narrowing down the ‘semantic gap’, human attribute detection and face image retrieval performance. Sparse coding helps in discriminative image representation that outperforms state of art in content-based face image retrieval. Codewords generated by the attribute-enhanced sparse coding scheme can reduce the quantization error and achieve excellent result in face retrieval as compared to other approaches because scheme uses high level descriptions combining with low level facial features. This paper also gives review on sparse coding and dictionary learning.

7. Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments and also for everyone who directly and indirectly helps us in this work.

References

[1] L. Wu, S. C. H. Hoi, and N. Yu, “Semantics-preserving bag-of-words models and applications,” *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1908–1920, Jul. 2010.

[2] Y.-H. Kuo, H.-T. Lin, W.-H. Cheng, Y.-H. Yang, and W. H. Hsu, “Unsupervised auxiliary visual words discovery for large-scale image object retrieval,” *in Proc. IEEE Conf. Computer Vision and Pattern Recognit.*, 2011.

[3] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, “Describable visual attributes for face

verification and image search," *IEEE Trans. Pattern Anal. Mach. Intell., Special Issue on Real-World Face Recognition*,vol. 33, no. 10, pp. 1962–1977, Oct. 2011.

[4] B. Siddiquie, R. S. Feris, and L. S. Davis, "Image ranking and retrieval based onmulti-attribute queries," *in Proc. IEEE Conf. Computer Vision and Pattern Recognit.*, 2011.

[5] B.-C. Chen,Y.-H. Kuo, Y.-Y. Chen, K.-Y. Chu, and W. Hsu, "Semi-supervised face image retrieval using sparse coding with identity constraint," *in Proc. ACM Multimedia*, 2011.

[6] D. Wang, S. C. Hoi, Y. He, and J. Zhu, "Retrieval-based face annotation by weak label regularized local coordinate coding," *in Proc. ACM Multimedia*, 2011.

[7] Z. Wu, Q. Ke, J. Sun, and H.-Y. Shum, "Scalable face image retrieval with identity-based quantization and multi-reference re-ranking," *in Proc. IEEE Conf. Computer Vision and Pattern Recognit.*, 2010.

[8] B. C. Chen, Y. –Y. Chen, Y. –H Kuo, and Winston H. Hsu, "Scalable Face Image Retrieval Using Attribute-Enhanced Sparse Codewords", *IEEE Transactions On Multimedia, VOL. 15*, NO. 5 AUGUST 2013.

[9] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," *in Proc. IEEE Conf. Computer Vision and Pattern Recognit.*, 2009.

[10] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal.*
Mach. Intell., vol. 31, no. 2, pp. 210–227, Feb. 2009.

[11] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE TSP*, 54(11):4311–4322, 2006.

[12] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zissserman, "Learning discriminative dictionaries for local image analysis," *in CVPR*, 2008.

[13] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," *in NIPS*, 2009.

[14] Q. Zhang and B.X. Li, " Discriminative K-SVD for dictionary learning in face recognition," *in CVPR*, 2010.

[15] M. Yang, L. Zhang, X Feng and D.Zhang, "Fisher Discrimination Dictionary Learning for Sparse Representation", *in Proc. IEEE Conf. Computer Vision and Pattern Recognit.*, 2011.

W. Scheirer, N. Kumar, P. Belhumeur, and T. Boult, "Multi-attribute spaces: Calibration for attribute fusion and similarity search," *in Proc.IEEE Conf. Computer Vision and Pattern Recognit.*, 2012.

Author Profile

Bharti Satpute received the B.E. degree in Information Technology from K. K. Wagh College of Engineering, Nashik in 2009. During 2010-2011, she did lecturership in Mahavir Polytechnic College, Nashik. She now is pursuing Master degree in Computer Engineering from Padmashree Dr. D.Y. Patil Institute of Engineering & Technology, Pimpri, Pune.

Archana Chaugule working at Padmashree Dr. D.Y. Patil Institute of Engineering & Technology, Pimpri, Pune as an Asst. Professor in Department of Computer Engineering since Sept 2012. Published more than 10 Research papers in National /International Journals/National /International conferences.