Semi-Supervised Face Image Retrieval using Sparse Coding with Identity Constraint

Bor-Chun Chen, Yin-Hsi Kuo, Yan-Ying Chen, Kuan-Yu Chu, and Winston Hsu
National Taiwan University, Taipei, Taiwan
{siruishpa, kuonini, yanying}@gmail.com, b96902119@ntu.edu.tw, winston@csie.ntu.edu.tw

ABSTRACT
We aim to develop a scalable face image retrieval system which can integrate with partial identity information to improve the retrieval result. To achieve this goal, we first apply sparse coding on local features extracted from face images combining with inverted indexing to construct an efficient and scalable face retrieval system. We then propose a novel coding scheme that refines the representation of the original sparse coding by using identity information. Using the proposed coding scheme, face images with large intra-class variances will still be quantized into similar visual words if they share the same identity. Experimental results show that our system can achieve salient retrieval results on LFW dataset (13K faces) and outperform linear search methods using well known face recognition feature descriptors.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms
Algorithms, Performance

Keywords
Face, retrieval, Sparse coding, Semi-supervised, Identity

1. INTRODUCTION
The goal of face image retrieval is to find the ranking result from most to least similar face images in a face image database given a query face image. Such work has many applications in different areas. For instance, when it applies on personal multimedia, it can enable automatic face tagging and face image clustering; while in forensics, it can help with crime investigation.

Face retrieval task is closely related to face recognition task, a problem has been investigated for decades. Pentland and Turk introduced the concept of eigenfaces [13] which has been widely used; Ahonen et al. proposed using local texture descriptors: local binary pattern (LBP) [10] for face recognition [2]. The descriptor proposed is efficient and has good performance. The difference between face recognition and face retrieval is that face recognition task requires completely labeled data in the training set, and it uses learning based approach to find classification result while neither training set nor learning process is needed in face retrieval task, and it provides a ranking result. Although descriptors used in face recognition can be directly applied on retrieval task, it is not trivial to apply it on an indexing system due to its high dimensionality. One might consider using traditional image retrieval system such as bag-of-words (BoW) model [12].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Figure 1. (a) Because face images have high intra-class variance, patches from two bush images are quantized into different sparse representations ($v_0$, $v_1$). (b) Augment face sparse coding with identity constraint. Even though the original sparse representations are different, the proposed coding scheme will propagate identity-related visual words between images of the same identity and construct more semantic sparse representations ($v_2$, $v_3$). In this example, number of overlapped visual words increase from zero to three.

constructed by local descriptors like SIFT features [7] for this problem. But when applying these methods on face images the performance is unsatisfactory. It is because face images have higher intra-class variance and these methods neglect the important spatial information in face images.

Wu et al. proposed a face image retrieval framework using identity based quantization and multi-reference re-ranking [16] to solve this problem. Motivated by their work, we propose a framework using more general coding scheme with the help of partial identity information to form a semi-supervised face image retrieval system. Similar idea was discussed in [15] for semantic hashing. It is easy to see the work has its necessity because in most of application for face retrieval, we might have some side information about our database images, including identity information, gender information, social network, etc. It is important if we can properly use such (noisy and partially available) information in the retrieval system to improve the results. To the best of our knowledge, little work aims to deal with this problem.

Recently, sparse coding has shown promising results on many different applications such as image denoising and classification. Raina et al. proposed a machine learning framework using unlabeled data to learn basis of sparse coding for classification task [11]. Yang et al. [17] further improved it by applying on SIFT descriptors along with spatial pyramid matching [6]. Gao et al. proposed a variant of sparse coding called laplacian sparse coding [4] by adding a locality constraint in the objective function. All of them show that sparse representations are well adapted to natural signals. Taking advantage of sparse coding, we introduce a...
face retrieval system using inverted index built from sparse representation of face images.

Although using sparse coding combined with inverted indexing results in an efficient retrieval framework, it does not take advantage of using identity information. To address this point, we propose a novel coding scheme to refine the result of sparse coding based on identity information. One might think our method is similar to [4], but they are essentially different. Firstly, no template features and large laplacian matrix is needed in our method. Secondly, since we quantize each person separately, we can scale up to as many people as we want. Last but not least, we integrate (partially available) identity information to automatically derive semantic-related visual words.

Figure 1 illustrates the proposed method. Using sparse coding with identity constraint, a query will be able to retrieve most images with the same identity as long as it is similar to one of them. To sum up, our contributions include identifying the problem of semi-supervised face image retrieval, introducing a general coding scheme for face retrieval problem, and further improving the performance by using identity information in the coding scheme. Extensive experiments show that the proposed method significantly outperforms linear search methods (cf. Table 1).

2. SYSTEM FRAMEWORK

For every face image in database, we first employ a frontal face cascade detector [14] to find the location of the face, and then active shape model [9] is applied to locating five facial components, including two eyes, two mouth corners and nose tip. The face image is then aligned using the locations of the eyes. After alignment, we define 5x7 grids around each facial component to get a total of 175 grids using the same procedure in [16]. From each grid, we extract a 59 dimension uniform local binary pattern feature descriptor. All 175 descriptors then are first quantized into visual words separately by using sparse coding (Nonzero entries in the sparse representation are considered as visual words). Images with identity information are further refined using coding scheme described in section 3.1 to embed the inverted information. Inverted index is built using the final sparse representation. Note that each descriptor from different grid locations will be quantized separately, hence two visual words will match only if it is extracted from the same grid location. Through this way, we can encode important geometric information of face into visual words.

Given a query image, we apply the same detection, alignment, and feature extraction methods described above. The extracted features are then quantized into visual words by sparse coding and used for retrieving inverted index. In the retrieval stage, we simply use histogram intersection to compute similarity score. The system overview is shown in Figure 2.

3. SPARSE CODING WITH IDENTITY CONSTRAINT

3.1 Sparse Coding (SC)

Traditional BoW model use k-means clustering algorithm to learn dictionary for quantization. The algorithm solves the following optimization problem:

$$\min_{D, v} \sum_{i=1}^{n} ||x_i - Dv_i||_2$$

subject to $Card(v_i) = 1, \forall i$, $v_i \geq 0, \forall i$ where $D = [d_1, ..., d_K]$ is a dictionary matrix with the size of $59xK$, each column represents a centroid (totally $K$); $v = [v_1, ..., v_n]$ is the centroid indicator matrix, each $v_i$ indicates that the original feature $x_i$ belongs to a centroid in $D$. The constraint $Card(v_i) = 1$ means each feature can only be assigned to one centroid. This constraint is considered to be too strict because some features might be at boundary of two or more centroids, therefore many people suggest relaxing the constraint and putting L1 regularization term on $v_i$ instead. This then turn into another optimization problem known as sparse coding:

$$\min_{D, v} \sum_{i=1}^{n} ||x_i - Dv_i||_2 + \lambda ||v_i||_1$$

subject to $||v||_1 \leq 1, \forall i$

This optimization problem can be efficiently solved by an online optimization algorithm [8]. We adopt this method to learn dictionary $D$ for later use. Because we quantize feature descriptors from each grid separately, we have to learn different dictionary from them. Consequently, there are total 175 dictionaries learned.

Once the dictionary is learned, we can fix $D$ in the above formulation and minimize the objective function along with $v_i$ separately to find the sparse representation of each feature. When $D$ is fixed, the optimization problem turns into a least square problem with L1 regularization. Since the L1 regularization term makes the objective function non-differentiable when $v_i$ contains zero elements, we cannot solve it by standard unconstraint optimization method. This challenge has led to many literatures using different approaches to solve this problem. Here we use LARS algorithm [3] to solve this problem. Because of the L1 regularization term, each sparse representation $v_i$ will only contain several nonzero elements out of $K$ dimension. These nonzero entries are then considered as the visual words from the descriptor $x_i$. Since there are 175 grid locations, total dictionary size is $175xK$. 

![Figure 2. The proposed system framework. 59 dimension uniform local binary pattern feature descriptors are extracted from 175 grids in an aligned face image. These features then are quantized into visual words using sparse coding. Images with identity information are further refined via the proposed method. Finally, inverted index is built using the final sparse representation.](Image)
3.2 Sparse Representation Refinement using Identity Constraint (SC+I)

Sparse representations computed from section 3.1 are ready for inverted indexing, but it does not embed the identity information. Hence, the retrieval results suffer from low recall rate due to high intra-class variance. Therefore, we introduce a method to reduce the intra-class variance by using identity information. Intuitively, the sparse representation extracted from the images of the same identity should be alike. To achieve this, we propose to add an identity constraint into the original sparse coding problem to refine each sparse representation $v_i$ separately:

$$\min \beta \sum_i \|v_i - D_i\|^2 + (1 - \beta) \sum_i \|v_i - \hat{v}_i\|^2 + \gamma \|v_i\|_1$$

where $V = [v_1, ..., v_m]$ are $m$ sparse representations computed from section 3.1 which share the same identity with $v_i$, $\beta$ is a parameter for adjusting weight between identity information and visual feature, and $\gamma$ is another parameter for adjusting the sparsity of the result. Since there are two different parameters and each term in the objective function is in different scales, it is hard to set the parameters. Hence we propose a simply way to normalize the objective, the results are as follow:

$$\min \beta \sum_i \frac{\|v_i - D_i\|^2}{\|D_i\|^2} + (1 - \beta) \sum_i \frac{\|v_i - \hat{v}_i\|^2}{\|\hat{v}_i\|^2} + \gamma \frac{\|v_i\|_1}{\|v_i\|_1}$$

Although the objective function is non-differentiable, the objective function without L1 regularization term is twice differentiable and convex. Therefore we can still solve this problem by using generalized algorithms for L1 constraint optimization problem. Specifically, we use a projected scaled sub-gradient algorithm [1] proposed by Schmidt.

After refinement, sparse representations from the same identity will propagate visual words to each other, and query will be able to retrieve most images from the same identity as long as it is similar to at least one of them (cf. Figure 1 (b)).

4. EXPERIMENTS

4.1 Experimental Setting

4.1.1 Data Set

We use all the images from LFW dataset [5] for our experiments. There are total 13,233 face images among 5,749 people in this dataset, while 1,680 people with two or more images and 12 people with 50 or more images. We take ten images each from these 12 people, total 120 images, as our query set. Figure 3 shows some example images from the dataset. Throughout the experiments we use mean average precision (MAP) and precision at one (p@1) as our performance measurement.

4.1.2 Face Retrieval Baselines

We use two kinds of face descriptors with linear search as baselines to compare with our system. The first one is based on the face recognition descriptor proposed in [2]. We divided the face into 7 by 7 fixed grids, extract 59 dimension uniform LBP from each grid, and then concatenate the features from each grid to form a global descriptor with 2,891 dimension, we call it baseline – fixed grids (BF) in the following experiments. The second one is the simple concatenation of local descriptors extracted in our proposed method, it has 175 grids with 59 dimension feature, total 10,325 dimension, and we call it baseline – component based (BC) in the following experiments.

For fair comparison, in the experiments with one hundred percent identity information, after retrieving images using linear search in baselines, we perform re-ranking by using the identity information to retrieve images from top 10 ranked identities and only use these images as ranking results. (BF+R, BC+R)

4.2 Sparse Coding Retrieval Performance

In sparse coding, we need to decide the size of dictionary $K$ and sparsity parameter $\lambda$. In our experiments, we have tried different dictionary size from 100 to 400 with $\lambda$ range from $10^{-3}$ to $10^{-1}$. Results are shown in Figure 4. We find out that if $\lambda$ is properly set (around $10^{-5}$ to $10^{-3}$) the performance affects little by the size $K$. When $\lambda$ is big (around $10^{-3}$), it penalizes the nonzero entry so much that all the entries in sparse representation become zero, therefore, the MAP drops to zero. When $\lambda$ is small (around $10^{-5}$), the performance drops faster with smaller dictionary size, this is because the sparse representation are so dense that each visual word lost its discriminative power.

When $\lambda$ is chosen properly, the performance of sparse coding exceeds baselines because the baselines with high dimensional features suffer from the curse of dimensionality. Note that in Figure 4 the component based baseline (BC) outperforms the fixed grids baseline (BF), this is because that grids from component based baseline have many overlaps in the central face area where contains more information, this overlapping acts like the weight in the fixed grid method, and proper weight increasing the performance agrees with the finding in [2]. In the following experiments, we use $K=100$, $\lambda=0.1$ as our default parameters.

4.3 Sparse Coding with Identity Constraint Retrieval Performance

In sparse coding with identity constraint, we need to decide $\gamma$, the weight between visual features and identity information, and...
sparsity parameter $\gamma$. In our experiments, we have tried different $\gamma$ range from 0.1–0.9 with $\gamma$ range from 0.1–0.5 and find out it performs best when $\beta$ is 0.1 and $\gamma$ is 0.15 in the case of a hundred percent images with identity information. The best $\beta$ is small indicating that we weight toward the identity information. This is probably because the identity information used in our experiments is reliable. The results are shown in Table 1. We can see that sparse coding with identity constraint achieve 70% relative improvement in MAP over the baselines using the same feature descriptors (BC+R). It suggests that our method effectively utilizes the identity information for the retrieval task while maintaining an efficient and scalable index structure.

In the case of partial information, the performance decreases when the information is less, but as long as there is some identity information, using sparse coding with identity constraint always outperforms using sparse coding only. Figure 5 shows the performance of sparse coding with identity constraint under different percentage of identity information compares with sparse coding without identity constraint. Note that the query images do not carry any identity information.

Intuitively, adding the identity constraint will increase the number of visual words in sparse representation and increase the retrieval time. Figure 6 shows the relationship between average number of visual words and performance by fixing $\beta=0.1$ and adjusting $\gamma$. We find out that under the same sparsity with sparse coding, sparse coding with identity constraint can achieve much higher performance.

Table 1 summarizes the performance of the proposed method and baselines under the best setting. Note that the performance of proposed method outperforms the linear search method in both cases while using an efficient index structure.

Table 1. Performance of the proposed method. Sparse coding outperforms linear search no matter with or without identity information. The left three columns are without any identity information while the right are with 100% identity information. Note that we did not compare the result that using partial information because the baselines cannot deal with partial identity labels.

<table>
<thead>
<tr>
<th>Method</th>
<th>BF</th>
<th>BC</th>
<th>SC</th>
<th>BF+R</th>
<th>BC+R</th>
<th>SC+I</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.10</td>
<td>0.12</td>
<td>0.16</td>
<td>0.33</td>
<td>0.42</td>
<td>0.72</td>
</tr>
<tr>
<td>P@1</td>
<td>0.61</td>
<td>0.71</td>
<td>0.80</td>
<td>0.60</td>
<td>0.71</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Figure 6. Relationship between sparsity and performance in sparse coding and sparse coding with identity information. Using identity constraint can achieve much higher performance while maintaining roughly the same sparsity.

5. CONCLUSIONS

We propose a face retrieval method based on inverted index structure with sparse coding and achieve better performance than linear search based on well-known face recognition descriptors. We then show how to extend the proposed method into semi-supervised case. In the experiments, we show that using the identity information, we can achieve higher performance while maintain the same sparsity. In the future, we will test the scalability of our system and combine more context information into the proposed framework to further boost up the performance.

6. REFERENCES