Appearance, Visual and Social Ensembles for Face Recognition in Personal Photo Collections

Anonymous CVPR submission

Paper ID 2052

Abstract

While modern research in face recognition has focused on new feature representations, alternate learning methods for fusion of features, most have ignored the issue of unmodeled correlations in face data when combining diverse features such as similar visual regions, attributes, appearance frequency, etc. Conventional wisdom is that by using sufficient data and machine, one can learn the systematic correlations and use the data to form a more robust basis for core recognition tasks like verification, identification, and clustering. This however, takes large amounts of training data which is not really available for personal consumer photo collections. We address the fusion/correlation issue differently by proposing an ensemble-based approach that is built on different information sources such as facial appearance, visual context, and social (or co-appearance) information of samples in a dataset, to provide error diversity and higher classification accuracy for face recognition in consumer photo collections. To evaluate the utility of our ensembles and simultaneously generate stronger generic features, we perform two experiments - (i) a verification experiment on the standard unconstrained LFW (Labeled Faces in the Wild) dataset where by using an ensemble of appearance related features we report state-of-the-art results on protocol-1 that provides 2.9% better classification accuracy than the current best method, and (ii) an identification experiment on the Gal-lagher personal photo collection where we demonstrate at least 17% relative performance gain using visual and social ensembles.

1 Introduction

The availability and popularity of digital cameras has made capturing personal photographs extremely easy and inexpensive. While automated tools to organize them have not kept pace with this growth, leveraging the modern advances in a face recognition system is a natural choice to address this user need. Face recognition has a long history in vi-sion research [29] whose utility has been realized in security applications for decades, unlike forays into consumer photos which began more recently. Uncontrolled face variations such as lighting, expressions, pose, scale and age progression exist in both surveillance scenarios and personal photo collections, but an advantage with the latter is the availability of additional information from the surroundings, geo-tags, and co-occurrence statistics of people. However, efficient utilization and combination of this information to boost recognition accuracy still remains a challenge.

In this work, we propose an ensemble-based approach for face recognition in personal photo collections. We consider three sources of information - appearance, visual and social context. Appearance refers to the information pertaining to the facial region that could be used to discriminate one person’s face from another. Visual context is the information obtained from regions surrounding the face such as hair, clothes, accessories and some generic information about as face used in a non-discriminative sense. We also consider the visual attributes such as age and gender as visual context. Social context pertains to the information derived from co-
appearance statistics of people in the images. We do not consider information obtained from social network sites, because personal photo collections exist on cameras, and network-provided storage that are not restricted to the social media space.

The ensemble of features and classifiers trained on these three different information sources provides classification error diversity and, when leveraged in a ensemble fusion scheme, improves performance on facial recognition tasks on personal photo collections. It can be argued that direct concatenation of different features followed by efficient machine learning algorithms can result in strong performance, but variations in dimensionality across different features can lead to an imbalanced treatment. For example, the fusion of mid level attributes can produce better recognition accuracy than low level features, even when its resulting dimensionality is smaller than that of the low level features [11]. Also, from a computational viewpoint, training high dimensional statistical and kernel methods on these concatenated vectors is prohibitively expensive. Therefore, we adopt an ensemble-style approach for this problem by learning a multitude of data-driven ensembles corresponding to appearance, visual, and social features. We use linear support vector machines (SVM) as the base classifier since they are arguably good positions as an additional cue for resolving the identify in-formation has been addressed by [6, 8]. Face and body information was combined in graphical models to improve recognition in [28]. Other recent works [18, 17] consider recognizing faces in a photograph as a joint labeling problem and proposed a structure prediction model to solve it. They use pairwise Markov random fields (MRF) and define the relationship between image pairs using both visual information and social network context.

Other application-driven methods focus on interactive labeling aspects using face recognition. Authors of [19] propose a face annotation system that uses an unsupervised approach to initialize the most evident clusters and allows use interaction for labeling. A global uncertainty based approach suggests the labels and makes use of face features, context features such as clothes and time features. Another approach addresses the clustering problem using a graph based social structure to resolve identity ambiguity in face clustering and suggest a way to merge clusters [24]. Authors in [27] cast the face annotation as a problem of similar face search and annotation propagation. Finally, authors in [13] focus on discovering and presenting the novel face identities that do not have labels yet, but have social context with respect to previously labeled faces.

While there have been many other techniques [11, 21, 22, 15, 14, 25] that offered viable solutions to address unconstrained face recognition, the underlying problem of assessing importance of information conveyed by contextual sources (i.e., social relationships, clothing, age, and gender) with respect to the appearance information of the face has not been investigated thoroughly. In pursuit of this goal, we propose a joint ensemble approach focused on appearance, visual and social aspects. While it is not guaranteed that the ensemble of classifiers will always perform better than an individual best performing classifier [5], our intuition is that modeling diverse aspects of faces in the form of feature and classifier ensembles offers a more systematic solution than many previous approaches addressing the social context aspect of faces, and we supplement this by demonstrating competitive results on two widely used datasets. In the following sections we discuss the details of our approach.

3 Proposed Approach

Faces in personal photo collections have contexts and strong underlying correlations like similar visual regions, attributes, appearance frequency, etc. An increasing number of works explore such contexts, but efficient utilization of this information to boost recognition accuracy still remains a challenge. This section describes our ensemble-based approach, built on different contexts to provide error diversity and higher classification accuracy for face recognition in consumer photo collections. We use strategically defined ensembles from facial appearance information, visual context, and social (or co-appearance) information across samples in a dataset.
3.1 Ensemble Definition

We propose a linear ensemble-based approach to handle both pair-matching and identification settings. Assume that images are represented by \{d_i\}, where \(d_i\) refers to the dimensions in different features. For the pair matching (or verification) task, training images are defined by \(X = \{(x_i, y_i), z_i\}, \ldots (x_m, y_m), z_m\}\) where \((x_i, y_i)\) represents a pair of training images and \(z_i \in [1, -1]\). Let \(T = \{(r_1, s_1), \ldots (r_n, s_n)\}\) denote the test data, with \((r_i, s_i)\) representing a test image pair. Let \(h_t, \{t = 1..T\}\) represent the base classifiers, (linear SVM’s in this work) that are trained on different features \(\{d_i\}\). Each \(h_t\) then provides a decision for every pair of test images where if \(\Phi_t(r_i, s_i) > \tau, z_i = 1\), else \(z_i = -1\), where \(\Phi_t(.)\) is the marginal distance of the sample from the hyperplane obtained from a \(h_t\) and \(\tau = 0\) in binary SVMs. Algorithm 1 shows the ensemble of base classifiers using GABOR and Local Binary Pattern(LBP-Like) features for constructing the final pair-matching classifier. Since these classifiers have comparable performance, we use simple sum rule in combining these ensembles as it has been shown theoretically and empirically [5] that when the base classifiers exhibit similar classification accuracy, a simple weighting \(1/T’\) provides the optimal weights [5].

Algorithm 1 The ensemble for pair-matching. The base classifiers utilize only appearance features classifiers.

Training: \(X = \{(x_i, y_i), z_i\}, \ldots (x_m, y_m), z_m\}\) where \((x_i, y_i)\) represents a pair of training images and \(z_i \in [1, -1]\).

Testing image pair: \((r, s)\)

Output: class label \(c\) from final classifier \((c = h_{\text{final}}(r))\)

for \(i = 1 \rightarrow C\) do

1. Let \(h_t\) be the base binary classifiers.

2. Let \(\Phi_t(r, s)\) be the normalized support of \(h_t\).

end for

\[\Phi_t(r, s) = 1/T’ \sum_{t=1}^{T'} \Phi_t(r, s)\]

\[h_{\text{final}}(r, s) = \Phi(r, s) > \tau\]

The basic classifiers for the identification task in personal photo collection are appearance, visual and social classifiers, and the basic classifiers for pair matching task are appearance classifiers. These classifiers can actually be considered ensembles by themselves. Specifically, we consider different partitions of the data such that the classifiers built on those partitions learn different sub-tasks, which may or may not be related to the final classification task, to be ensembles. The training and testing images from the recognition problem at hand are now represented with the output of the classifiers. The feature vectors thus obtained are used for the actual recognition problem, similar to stacked generalization [23]. Where, level-0 classifiers are constructed with different data folds and level-1 classifiers take the input from the output of level-0 classes. These stage-0 classifiers are “independent” of the final task that is carried out in “stage-1”, which can be either pair-matching (verification) or identification. Algorithm 3 shows the general outline of these ensembles.

In the next few sections, we describe construction of these base classifiers (or ensembles) with different underlying features.

3.2 Appearance Features and Ensembles

We define facial features with respect to the prior learned classes as shown in Figure 2. The stage-0 and stage-1 classifiers are binary or multi-class depending upon the task. For conducting an ensemble which performs better than the best classifier is a challenging problem. However, uneven performances of individual classifiers can be counterbalanced by a weighted average rule and can, as a result, outperform the best classifier, even when classifiers with arbitrarily high error probability are added to a given ensemble [5]. However, since the correlation of the appearance, visual, and social classifiers is unknown and their expertise in recognition can be different, we use maximum rule. Algorithm 2 shows the ensemble of appearance, visual, and social features would be used to construct the final identification system in personal photo collection.
Algorithm 3 Ensemble of ensembles: appearance, visual and social ensembles for identification and appearance ensemble for verification.

STAGE-0

Reference Clusters: $R_t, t \in \{1..T\}$
Stage-0 Classifiers: $h_t, t \in \{1..T\}$
Number of classes in stage-0 classifiers: $C'_t$
Train images: $X = \{(x_i, z_i), ..., (x_m, z_m)\}, z_i \in \{1..C\}$
Test image: $r$

for $t = 1 \rightarrow T'$ do
  for $i = 1 \rightarrow C'_t$ do
    Let $\phi_{i,t}(r)$ be support of $h_t$ for class $i$.
    $\Phi_i(r) = \text{concat}(\Phi_i(r), \phi_{i,t}(r))$
    $\Phi_i(x) = \text{concat}(\Phi_i(x), \phi_{i,t}(x))$
  end for
  $F(r) = \text{concat}(F(r), \phi_i(r))$ test sample $r$’s feature vector
  $F(x) = \text{concat}(F(x), \phi_i(x))$ train sample $x$’s feature vector
end for

STAGE-1

Identification:
Training: $(F(x_i), z_i), i \in \{1..m\}, z_i \in \{1..C\}$
Testing vector: $F(r)$
$h_{f_{\text{final}}}(r) = \arg\max_{k=1..C} \Phi_k(r)$

Verification:
Training: $(F(x_i, y_i), z_i), i \in \{1..m\}, z_i \in \{1..1\}$
Testing vector: $(r, s)$
$h_{f_{\text{final}}}(r, s) = \Phi(r, s) > \tau$? $1 : -1$. $\tau$ is a threshold.

the pair-matching problem, we use an ensemble of binary classifiers which individually learn the match and non-match categories of face images.

For the identification problem, we use an ensemble of multi-class classifiers which learn the identities of multiple subjects appearing in each reference partitions. We use GABOR and LBP-like features to construct the image level feature vectors for stage-0 classifiers. Unlike Tom vs Pete classifiers [3], these binary stage-0 classifiers are constructed with samples of images from different individuals using the feature vector from a pair of images. And unlike attribute based classifiers [11], the output of these do not have any defined meaning such as ”long hair”, ”shiny cheeks” but have a meaning of undefined resemblance with respect to previously learned face categories. One motivation to this work in that human tend to recognize new examples based on the previously learned categories. The motivation behind learning from different linear classifiers is that SVMs are good at learning in the neighborhood of small number of examples from small number of classes [26]. We use an ensemble of liner classifiers instead of having to use expensive kernels. In order to construct the stage-0 classifiers, we use the face images, which are not in the test-

Figure 2. Framework for appearance, visual and social ensembles. Face data is divided into different reference partitions and stage-0 classifiers are built on those partitions. Training and testing images are tested on the stage-0 classifiers and decision scores are obtained. The feature vectors for the stage-1 classifiers are built with output decision scores from each of the stage-0 classifiers. We construct reference clusters using only training images for the verification task and using training and external data for the identification task.

3.3 Visual Features and Ensembles

We define visual features as the information from the facial images and the surrounding regions that define the attributes other than the actual identity. The visual features in this paper are either learned from the classifiers or are readily available from images [20]. We use the following classifier based visual attributes: Age: The age classifiers are trained based on the range of ages. We divide the age categories into multiple bins such as baby, kids, teen, young adult and older adult. Age is considered as a visual context in a smaller time duration because it provides a quick coarse level classification. People with huge differences in age may not be confused in identities considering a small time gap between images. Gender: We consider gender classifiers of kids as
well as for adults. Similar to age classifier, this visual information is useful in coarse level classification. Expression: Expression is considered as another visual feature. Most of the people tend to have certain expressions when they pose for pictures.

As mentioned in Section 3.1 and shown in Figure 2, we construct stage-0 classifiers on age, gender and expression partitions of the database. These partitions are from the outside database and images of these partitions are not in the training and testing images. Stage-0 classifiers take in LBP-like feature vectors from the face images. Input to the stage-1 identification classifier are the outputs from these stage-0 classifiers. These outputs are not decision labels based on age, gender or expression, but are the marginal scores or votes for all these categories from all the classifiers.

A set of $T'$ clusters are constructed using age, gender and expression database from RAFD database [12] and FG-NET aging databases [1]. Within each cluster $R_t$, $t \in \{1 .. T'\}$, we construct a multi-class SVM classifiers $h_t$ with $C'_t$ number of classes in the cluster. A visual feature vector for a face images $x$ is then defined as the concatenation of the outputs of the classifiers from the above clusters.

$$V(x) = [(\phi_{1_{h_1}}, \phi_{2_{h_1}}, ... \phi_{C'_{1_{h_1}}}],[\phi_{1_{h_2}}, \phi_{2_{h_2}}, ... \phi_{C'_{2_{h_2}}}],...$$

where, $\phi_{C'_{t_{h_t}}}$ is the decision value of the SVM for $C'_{t_{h_t}}$ class of $h_t$ classifier. Apart from the classifier-learned visual attributes, we also use face surrounding regions such as head region, shoulder and clothes regions as the visual context. We extract the histogram of RGB channels from these regions as it tends to provide a color information around those regions which can be helpful in identifying people in case the appearance information is ambiguous.

### 3.4 Social Features and Ensembles

In personal photo collections we will come across a lot of people but we are interested in tagging only a few people. However, we can use the information of other people appearing in the images and their visual context to identify the people we are interested in. We define social features based on the co-appearance of different individuals in the same picture [20]. The social information we incorporate in this paper consists of the following classifier-based social attributes: Histogram of gender distribution: This exploits the underlying clique in which the gender distribution that an individual is generally found. Histogram of age distribution: This also exploits the underlying clique and defines the age distribution of the people that an individual is co-appearing with. The visual attributes such as gender and age are obtained by visual ensembles. By analyzing these attributes with co-occurring individuals, we form the histograms of these attributes. Now the social feature vector is defined as the concatenation of all these histograms. In addition to the classifiers’ learned social attributes, we use relative height and relative face among the co-appearing people as social features. The stage-1 feature vector for social features is composed in the same way defined for visual features in Section 3.3.

### 4 Experiments and Results

We conduct experiments on datasets that demonstrate performance gains in two tasks: Labeled Faces in the Wild (LFW) as a large-scale face verification task and the Gallagher personal photo collection as a personal, socially annotated identification task. LFW also provides a setting to study the performance of our method when only appearance ensembles can be constructed.

#### 4.1 Large-Scale Verification Task

The Labeled Faces in the Wild database [10] provides face images collected from news articles on the web. This is an unconstrained setting similar to faces found in personal photo collections. It provides a protocol for face recognition where recognition task is defined as a pair matching (or verification) problem. The database consists of 3000 match pairs and 3000 non-match pairs with 10 fold cross-validation. Each validation set consists of 5400 training pairs, with 2700 match and non-match pairs each and 600 testing pairs with 300 match and non-match pairs. This is a binary classification problem where given a pair of images, decision is ‘match’ or ‘non-match’. We use funneled version of the dataset [9] and followed “Strict LFW” protocol with no external data used. The main reason for choosing this view of LFW is because we want to leverage the information from the data itself in an ensemble framework to enhance the performance and our goal is not to see how we can add additional information from the outside data to improve verification. The face images are cropped to the size of 150x150 from the center. This is to avoid the background information as much as possible while keeping the visual information related to an individual.

The LFW database was chosen primarily to evaluate our appearance features. We use GABOR [30] and LBP-like [16] features as they are widely chosen in face recognition models. We conduct experiments on datasets that demonstrate performance gains in two tasks: Labeled Faces in the Wild (LFW) as a large-scale face verification task and the Gallagher personal photo collection as a personal, socially annotated identification task. LFW also provides a setting to study the performance of our method when only appearance ensembles can be constructed.
Table 2. Comparison of the results from this paper to prior state-of-the-art results on strict LFW database.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid descriptor-based</td>
<td>0.7847 ± 0.0051</td>
</tr>
<tr>
<td>3x3 Multi-Region Histograms</td>
<td>0.7295 ± 0.0055</td>
</tr>
<tr>
<td>Pixels/MKL, funneled</td>
<td>0.6822 ± 0.0041</td>
</tr>
<tr>
<td>VI-like/MKL, funneled</td>
<td>0.7935 ± 0.0055</td>
</tr>
<tr>
<td>This paper</td>
<td>0.8168 ± 0.0055</td>
</tr>
</tbody>
</table>

To illustrate performance gains over using image level features alone, as mentioned in Section 3.2, we construct stage-0 classifiers using 5400 training samples for each validation set. These training samples are divided into multiple overlapping partitions and stage-0 binary match vs non-match classifiers are constructed. For this experiment, the feature vector for an image pair \( a - b \) consists of \( \sqrt{f(a)^2 + f(b)^2} \), where \( f(a) \) is the feature vector from image \( a \) and \( f(b) \) is the feature vector from image \( b \). All 5400 training pairs of images are tested using these stage-0 linear binary SVMs and the decision scores of these classifiers are concatenated to construct the feature vectors for the stage-1 classifiers. 600 testing image pairs are first tested using the stage-0 classifiers and the concatenation of decision scores from these linear binary SVMs are used to construct the feature vector for stage-1 classifier. We conduct our experiments on LBP-like features and GABOR features using this approach. We also conducted binary classification experiments with these features on LFW training and testing partitions and observed a significant improvement using this ensemble of classifiers. It should be noted that no external data was used for any of these approaches. Normalized fusion, based on a simple sum rule, combines decisions from ensemble classifiers and low level feature classifiers, achieving improvement over either of those methods alone.

Table 1 shows the results using GABOR and LBP-like features and performance improvements when using an ensemble. Using an ensemble, the accuracy from LBP-like features grew from 0.7048 to 0.7456 while the accuracy of GABOR features jumped from 0.6505 to 0.7033. Our results as shown on Table 2 and Figure 3 are significantly better than the prior state-of-the-art results on strict LFW with no outside training data. Combining the results from both the features and their corresponding ensembles, further boosts accuracy to 0.8168 which is significantly better than the prior published results.

4.2 Personal Photo Identification Task

To evaluate our ensemble-based approach on face recognition in personal photo collections, we use the dataset proposed by Gallagher et al. [7]. This database consists of 589 images with 931 face images. There are 220 images with multiple people. The images are captured on 50 different days over the period of 233 days and consists of 32 unique individuals. For extracting appearance based features, we used the eye locations that come with the database and CSU normalization [4] to normalize the face images to a scale of 12x120 pixels and extracted LBP-like features from the different parts of the face regions. These features perform well for pair-matching as shown in Section 4.1. We use the appearance features as described in Section 3.2. We construct the stage-0 classifiers using RAFD [12] and FG-NET database [1] as these images consist of people of varying ages, similar to most people in personal photo collections. We also use a subset from LFW database for stage-0 classifiers [10]. Three multi-class classifiers are constructed with these three sets of reference clusters as shown in Figure 2. The stage-0 SVM decision outputs represent the likeness of an individual being similar to each of the individuals from those three databases. This forms the input to the stage-1 classifiers. We combine the results from LBP-like features based appearance classifiers and stacked based ensemble classifiers. The results on appearance ensemble are shown in Figure 4(a), 4(b) and 4(c).

We extracted the visual features from the regions surrounding the face such as hair, clothes, and accessories, as well as the face region in non-discriminative sense. We divided the surrounding area in multiple patches and extracted RGB histograms from these patches. We also extracted the visual attributes using ensemble-based method as described in 3.3. We basically design an age, gender and expression stage-0 classifiers from RAFD [12] and FG-NET database.
The performance of social ensembles alone is also weak, but they complement each other when combined with appearance ensembles. (b) Rank1 recognition results using only appearance, visual and social ensembles. The performance of social alone is not very attractive; however, using maximum vote combination of visual and appearance ensembles improves the performance. (c) Rank9 recognition rates for comparison with the original results [7]. The published Rank9 recognition in [7] varies from 0.5 to 0.8, whereas the demonstrated rates are between 0.88 to 0.94 using our method.

[1], take the output from these classifiers, concatenate the output, and form a feature vectors for stage-I classifiers. The visual ensemble consists of the results from both image-based visual features and from visual ensemble. The rank1 results of visual ensemble only are shown in Figure 4(b). The performance of visual alone is not very attractive; however, using maximum vote combination of visual and appearance ensemble improves the performance.

Figure 4(a) and 4(c) show the results of our approach. Figure 4(a) shows the rank1 recognition rate. We assume that images are randomly labeled at take 100, 200, 300, 400, 500 and 600 labeled images as training each time and remaining images for testing and we repeat this 10 times. We could actually use the pair-matching technique we described in Section 4.1 for grouping and labeling the people and forming a training set, however we decided to stay close to the protocol set by [7]. Though our protocol is slightly different from the one presented in [7], we can see a lot of improvement in using our method compared to the results previously presented. [7] provides the accuracy between 0.7 to 0.85 as the number of training images increase from 100 to 600. For graphical comparison, Figure 4(c) shows the rank9 recognition results. The recognition accuracy from our method ranges between 0.88 to 0.94. Recognition accuracy does not reach 1 because it is an open set identification, meaning not all testing subjects are in training set. These results show that using the proposed ensembles based method, we can significantly improve face recognition accuracy.

5 Conclusions and Future Works

In this work, we demonstrated the power of an ensemble-based approach for face recognition in personal photo collections. The main motivation behind the ensemble was to utilize information from multiple sources in personal photos collections like facial regions, surrounding regions and information due to co-appearance of people in the same images, without the need for large amounts of data to “learn” the correlations between these features. We demonstrated that ensembles provide a more meaningful way of combining the diverse and complementary information within these features without modeling dependence. The ensemble-based image level feature fusion for image restricted strict LFW protocol pair-matching was statistically significantly better than the prior state of the art. This shows the potential of our approach to improve real-world recognition. Our ensemble-based approach for fusion of appearance, visual and social ensembles for the identification task on the Gallagher dataset also significantly advanced the state-of-the-art performance. Together these demonstrate that strategically selected ensembles offer potential for other researchers working in recognition problems. Future work will examine extending our pair-matching algorithm to group the similar faces from personal collection for easy annotation and to support automatic grouping of potential other attributes.
References


