Appearance, Context and Co-occurrence Ensembles for Identity Recognition in Personal Photo Collections

Archana Sapkota[1] Raghuraman Gopalan[2] Eric Zavesky[2] T.E.Boult[1]
[1] University of Colorado at Colorado Springs [2]AT&T Labs-Research, Middletown, NJ
[1] {asapkota,tboult}@vast.uccs.edu [2]{raghuram,ezavesky}@research.att.com

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-occurrence) information of samples in a dataset, to provide 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 comparable results with recently reported state-of-the-art results and 2.9% better classification accuracy than the previous best method, and (ii) experiment on the Gallagher personal photo collection where we demonstrate at least 17% relative performance gain using visual context and social co-occurrence 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 biometrics and vision research [32] 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.

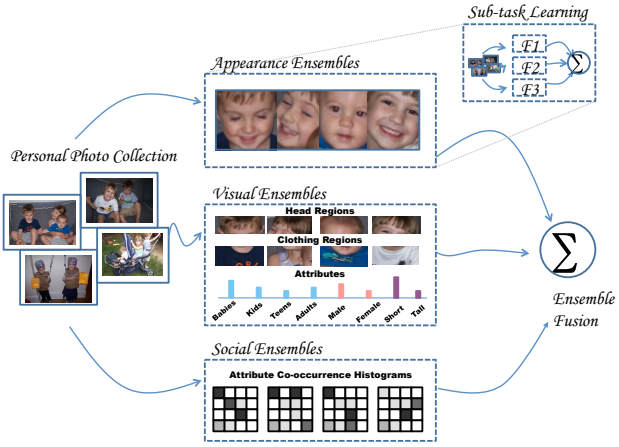


Figure 1. Our goal is to realize effective utilization of information available in personal photo collections. We propose an ensemble of appearance, visual context and social co-occurrence ensembles to boost the performance across several facial recognition tasks such as verification and identification.

In this work, we propose an ensemble-based approach for face recognition in personal photo collections. We consider three sources of information - appearance, visual context and social co-occurrences. 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 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

© 2013 IEEE. Pre-print of article that will appear in IEEE BTAS 2013
Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

from co-occurrence 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 [14]. 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 using inexpensive linear support vector machines(SVM).

To summarize, the main contributions of this work are:

1. Proposing appearance, visual, and social ensembles for a systematic utilization of diverse information in personal photo collections.
2. Leveraging image-level appearance features in an ensemble framework for a 2.9% the gain over the previous best verification results on the LFW dataset [13] with no external training data. [1]
3. Demonstrating the utility of social feature ensembles with more than 17.5% gains for face identification on the Gallagher personal photo collection [10].

2 Related Work

Face recognition tasks in personal photo collections have gained attention in the community due to the problem's complexity and its emerging applications.

Several works in this area propose ways to extract meaningful social context information from images to enable auto-tagging of images using faces. In [25], the appearance and contextual features are used to define the pairwise social relationships and identity in a probabilistic model and show that adding social-relationship as an attribute improves recognition accuracy. The utility of contextual information conveyed by a person's clothing has been investigated by [10] and [2]. Prior information about the people and their positions as an additional cue for resolving the identify information has been addressed by [9, 11]. Face and body information was combined in graphical models to improve recognition in [31]. Other recent works [23, 22] 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 [24] propose a face annotation system that uses an unsupervised approach to initialize the most evident clusters and allows user 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. Authors in [30] cast the face annotation as a problem of similar face search and annotation propagation. Finally, authors in [16] 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 [14, 26, 27, 19, 18, 29] 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. There are works in literature on learning from multiple sources and multiple weak classifiers [7, 6, 8, 20] . The work in [19] uses high dimensional Gabor features in Multi Kernel Learning framework. Ours is the first attempt in studying ensemble learning in personal photo collections. While it is not guaranteed that the ensemble of classifiers will always perform better than an individual *best* performing classifier [8], 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, such as 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,

[1] The current state-of-the-art was only published at the time of writing this paper.[17]

and social (or co-occurrence) information across samples in a dataset.

3.1 Ensemble Definition

We propose a linear ensemble-based approach to handle both pair-matching and identification problems. We first construct multiple base classifiers on multiple sources of information. For the pair matching (or verification) task, these classifiers learn whether a pair of images are from the same person or from different people. The classifier are constructed on Gabor [33] and Local Binary Pattern (LBP-Like) [21] features. Let $h_t, \{t = 1..T'\}$ represent these base classifiers (linear SVM's in this work) that are trained on multiple sources of information. Each h_t then provides a decision for every pair of test images. The SVM decision scores from these base classifiers are then normalized before constructing the final classifier. 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 and we do not have an information about the underlying correlation we use simple sum rule in combining these ensembles.

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

Training: $X = \{\{(x_i, y_i), z_i\}, ...\{(x_m, y_m), z_m\}\}$ where (x_i, y_i) represents a pair of training images and $z_i \in [1, -1]$ are labels. $z_i = 1$ if the images are from same person, otherwise $z_i = -1$
Testing image pair: (r, s)
Output: class label $c \in [1, -1]$. $(c = h_{final}(r, s))$
for $t = 1 \rightarrow T'$ **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
$\Phi(r, s) = 1/T' \sum_{t=1}^{T'} \phi_t(r, s)$
$h_{final}(r, s) = 1, if \Phi(r, s) > \tau, else -1$. τ is a threshold.

For the identification task, the base classifiers are constructed using appearance features, visual context features and co-occurrence features. Algorithm 2 shows the construction of final classifier. After learning these multiple base classifiers, the next step is to define their 'weights' (or relevance of their decisions) to construct a strong classifier. One of the problems here is that the performance of these statistically selected base classifiers are diverse. For example, in some identification scenarios, the appearance features can perform very well while visual features may not be as good. Constructing 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 [8]. 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 2 The ensemble for identification problem. The base classifiers are appearance, visual and social classifiers.

Training images: $X = \{(x_i, z_i), ...(x_m, z_m)\}$, x_i is an image, $z_i \in [1..C]$ is the class label.
Tesing image: r
Output: class label c from final classifier ($c = h_{final}(r)$)
for $i = 1 \rightarrow C$ **do**
for $t = 1 \rightarrow T'$ **do**
1. Let h_t be the base classifiers.
2. Let $\phi_{i,t}(r)$ be support of h_t for class i.

end for
$\Phi_i(r) = \max_{t=1}^{T'} \phi_{i,t}(r)$
end for
$h_{final}(r) = \text{argmax}_{i=1...C} \Phi_i(r)$

In fact, the base classifiers mentioned above are the ensembles by themselves. We build these base classifiers from different sources of information. The different sources of information here are the data partitions. We partition the available data (either from training data or outside the training data but not from testing data) into several reference partitions. Classifiers are trained on these reference partitions.

The training and testing images from the recognition problem at hand are tested on these classifiers and are represented with the output of the classifiers. The feature vectors thus obtained are used for the actual recognition problem, similar to stacked generalization [28]. Where, level-0 classifiers are constructed with different data folds and level-1 classifiers take the input from the output of level-0 classes. 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

Pair Matching: For the pair-matching problem, the reference partitions are defined and binary stage-0 classifiers are constructed as shown in Figure 2. In this work, we use LFW database and divide the training data into multiple partitions. Each classifier trained on each reference partition individually learns the match and non-match pairs. We use Gabor and LBP-Like features from image pairs and learn SVM classifiers. Unlike Tom vs Pete Classifiers [4], these binary stage-0 classifiers are constructed with samples of images from different individuals using the feature vector from a

Algorithm 3 Ensemble of ensembles: appearance, visual context and social co-occurrence ensembles for identification and only 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: $X = \{(x_i, z_i), ...(x_m, c_m)\}$, $z_i \in [1..C]$
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.
    $\Phi_t(r) = concat(\Phi_t(r), \phi_{i,t}(r))$
    $\Phi_t(x) = concat(\Phi_t(x), \phi_{i,t}(x))$
  end for
  $F(r) = concat(F(r), \phi_t(r))$ test sample r's feature vector
  $F(x) = concat(F(x), \phi_t(x)$ 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: $F(r)$
$h_{final}(r) = \text{argmax}_{k=1...C}\, \Phi_k(r)$
Verification:
Training: $(F(x_i, y_i), z_i), i \in \{1..m\}, z_i \in [1, -1]$
Testing vector: (r, s)
$h_{final}(r, s) = \Phi(r, s) > \tau ? 1 : -1$. τ is a threshold.

pair of images. Concatenation of the output of the stage-0 classifiers on the training pairs of images is used for training stage-1 binary classifiers.

Identification: For identification problem, the reference partitions contain the subjects outside the training and testing data. Each reference partition has multiple subjects. The stage-0 classifiers are the multi-class classifiers which learn the identities of these individuals in the reference partitions. We use Gabor and LBP-Like features from images to construct stage-0 classifiers. The training and testing images in hand are represented with the output from these classifiers. This means the training and testing images are represented in terms of their similarities with respect to the individuals in the reference partitions. Concatenation of the output of stage-0 classifiers on the training images is used for training stage-1 classifiers. Unlike attribute based classifiers [14], 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 is that humans tend to recognize new examples based on the previously learned categories. And learning from previous samples is popular in computer vision.

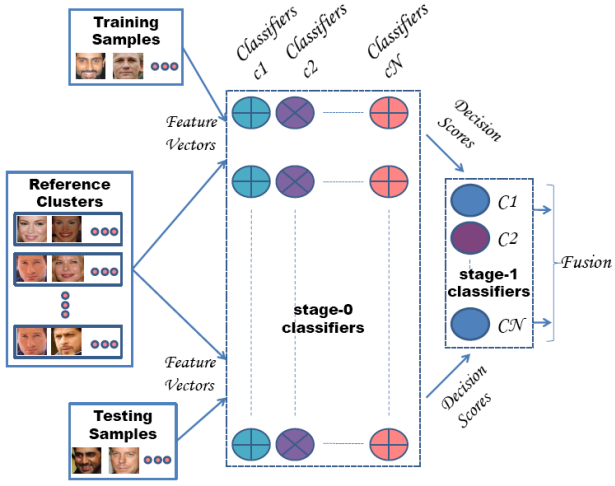


Figure 2. Framework for appearance, visual context and social co-occurrence ensembles. Face data is divided into different reference partitions and stage-0 classifiers are built on those partitions. Training and testing images are tested on the stage-0 classifiers and decision scores are obtained. The 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 Context Features and Ensembles

We define visual context features as the information from the facial images and the surrounding regions that define the attributes other than the actual identity. The visual context features in this paper are from two sources. They are either learned from the classifiers or are readily available from images [25]. 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 clusters are constructed using age, gender and expression database from AFD database [15] and FG-NET aging databases [1]. Within each cluster, we then construct multi-class SVM classifiers. A visual context feature vector for a face image is defined as a concatenation of the outputs of the classifiers from those classifiers.

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 [25]. The social information we incorporate in this paper consists of the following classifier-based social attributes: *Histogram of gender distribution:* This exploits the underling 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 [13] 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 database [12] 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.

Aglorithms	Performance
LBP-Like Features 1	0.7078 ± 0.0048
Ensemble LBP-Like Features 2	0.7456 ± 0.0046
LBP-Like Features & Ensemble 1,2	0.8081 ± 0.0044
Gabor Features 3	0.6505 ± 0.0055
Ensemble Gabor Features 4	0.7033 ± 0.0055
Gabor Features and Ensemble 3,4	0.7426 ± 0.0075
All combined 1,2,3,4	0.8168 ± 0.0055

Table 1. Mean ($\pm$ standard error) scores of different algorithms in this paper on strict LFW(no outside training), image restricted training benchmark. The ensemble-based approach performs better than the image level features such as LBP-like and Gabor. Combination of all gives significantly better results than the prior state-of-the-art results on this view of LFW database.

The LFW database was chosen primarily to evaluate our appearance features. We use Gabor [33] and LBP-like [21] features as they are widely chosen in face recognition works. We also apply our ensemble-based approach 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 such that there is a small correlation between partitions due to the presence of same subjects resulting in around 100 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 $sqrt|f(a) - f(b)|$, 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

Alorithms	Performance
Hybrid descriptor-based	0.7847 ± 0.0051
3x3 Multi-Region Histograms	0.7295 ± 0.0055
Pixels/MKL, funneled	0.6822 ± 0.0041
V1-like/MKL, funneled	0.7935 ± 0.0055
AEPM(fusion)	0.8408 ± 0.0120
This paper	0.8168 ± 0.0055

Table 2. Comparison of the results from this paper to prior state-of-the-art results on strict LFW database [13]. The method in this paper performs significantly better than prior methods.

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. We also used the fusion of classifiers that use LBP-Like or Gabor features only and the classifiers that use ensembles of these features. The performance of these classifiers is better than the individual classifiers because the two classifiers are constructed from different sources of information; from images and from the learnt classes in reference partitions. 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 and is comparable to recently reported results [17, 3].

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.* [10]. 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. We used the eye locations that come with the database and CSU normalization [5] to normalize the face images to a scale of 120x120 pixels and extracted LBP-like features from the different parts of the face regions in the same way as in [21]. 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

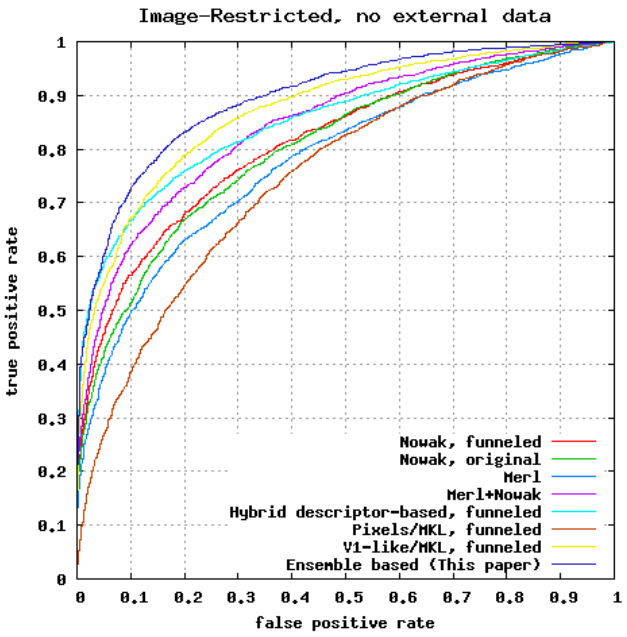
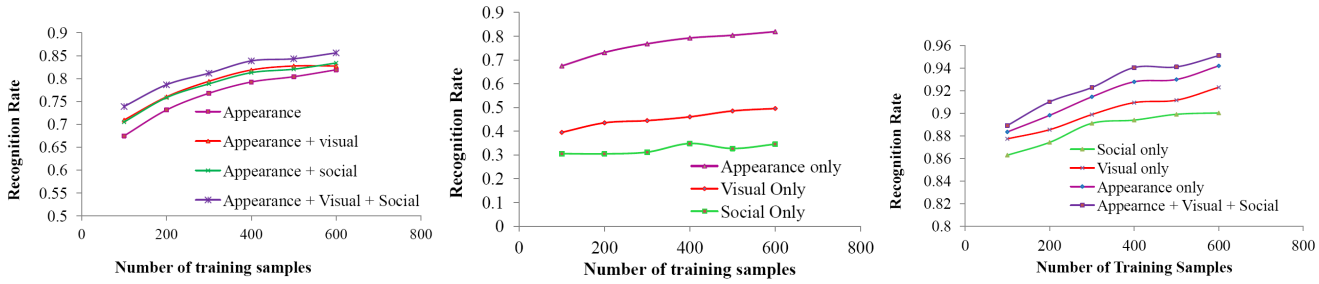


Figure 3. Roc curves for strict LFW (no outside data) image restricted training. Each point in the curve represents a 10-fold average of false positive and true positive rates for a fixed threshold.

classifiers using RAFD [15] 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 LFW610, [21] for stage-0 identity classifiers [13]. Three multi-class identity classifiers are constructed with these three sets of reference clusters as shown in Figure 2. The stage-0 SVM decision outputs represent the likeliness 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). The recognition rate without the features from external data and only with the LBP-Like and Gabor features from Gallagher database rang from 0.67 to 0.88 Rank 1 and 0.88 to 0.94 for Rank 9 for training examples ranging from 100 to 600. This itself is better than the results in [10]. The introduction of external data partitions is to enhance the performance and be consistent with our definition of ensembles.

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 [15] and FG-NET database [1], take the output from these classifiers, concatenate the



(a) (b) (c)

Figure 4. (a) Rank1 recognition on the Gallagher personal photo collection, set2. Results using appearance features and gains when using visual and social ensembles. (b) Rank1 recognition results using only appearance, visual and social ensembles. The performance of social and visual ensembles are not significant alone, but they complement each other when combined with appearance ensembles. (c) Rank9 recognition rates for comparison with the original results [10]. The published Rank9 recognition in [10] varies from 0.5 to 0.8, whereas the demonstrated rates are between 0.88 to 0.94 using our method.

output, and form a feature vectors for stage-1 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.

The social features are defined as mentioned in Section 3.4. Specifically, we divide the variations by age into multiple bins (e.g., baby, kids, teens, young adults and adults) and expressions into eight different expressions (e.g., happy, angry, sad, contemptuous, disgusted, neutral, fearful, and surprised) as defined by RAFD [15]. To construct a social co-occurrence feature for an individual, we concatenate that particular individual's values on these attributes with the histogram of these attributes with respect to the other people in the images. We also use features readily available from the images, such as relative height and face size of the person. Height is defined by the y-coordinate of the eyes and face width by the distance between eyes. If there is only one person in the image, we disregard this information as we do not have knowledge about the scale about the image. The rank1 results of social ensembles alone are as shown in Figure 4(b). The performance of social ensembles alone is also weak, but again the max confidence combination of social with appearance ensembles improves 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 [10]. Though our protocol is slightly different from the one presented in [10], we can see a lot of improvement in using our method compared to the results previously presented. [10] 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.

References

[1] FG-NET aging database. http://www.fgnet.rsunit.com. 5, 6

[2] D. Anguelov, K. chih Lee, S. Gokturk, and B. Sumengen. Contextual identity recognition in personal photo albums. In

Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, pages 1 –7, june 2007. 2

[3] L. Begeja, E. Zavesky, Z. Liu, D. Gibbon, R. Gopalan, and B. Shahraray. Vidcat: an image and video analysis service for personal media management. pages 86670F–86670F–11, 2013. 6

[4] T. Berg and P. N. Belhumeur. Tom-vs-pete classifiers and identity-preserving alignment for face verification. In *Proceedings of the 23rd British Machine Vision Conference (BMVC)*, September 2012. 3

[5] D. Bolme, R. Beveridge, M. Teixeira, and B. Draper. The CSU Face Identification Evaluation System: Its purpose, features and structure. In *ICVS*, pages 304–311, 2003. 6

[6] T. C. Faltemier, K. W. Bowyer, P. J. Flynn, and S. Member. A region ensemble for 3d face recognition. *IEEE Transactions on Information Forensics and Security*, pages 62–73, 2008. 2

[7] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm, 1996. 2

[8] G. Fumera and F. Roli. A theoretical and experimental analysis of linear combiners for multiple classifier systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(6):942 –956, june 2005. 2, 3

[9] A. Gallagher and T. Chen. Using group prior to identify people in consumer images. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8, june 2007. 2

[10] A. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, june 2008. 2, 6, 7

[11] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *Computer Vision and Pattern Recognition, 2009. CVPR '09. IEEE Conference on*, 2009. 2

[12] G. B. Huang and V. Jain. Unsupervised joint alignment of complex images. In *In ICCV*, 2007. 5

[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 2, 5, 6

[14] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *In IEEE International Conference on Computer Vision (ICCV*, 2009. 2, 4

[15] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg. Presentation and validation of the radboud faces database. *Cognition & Emotion*, 24(8):1377–1388, 2010. 5, 6, 7

[16] Y. J. Lee and K. Grauman. Face discovery with social context. In *Proceedings of the British Machine Vision Conference*, pages 36.1–36.11. BMVA Press, 2011. http://dx.doi.org/10.5244/C.25.36. 2

[17] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Computer Society Conference on*, 2013. 2, 6

[18] D. Parikh and K. Grauman. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503 –510, nov. 2011. 2

[19] N. Pinto, J. DiCarlo, and D. Cox. How far can you get with a modern face recognition test set using only simple features? In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2591 –2598, june 2009. 2

[20] A. Ross and A. Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24:2115–2125, 2003. 2

[21] A. Sapkota, B. Parks, W. Scheirer, and T. Boult. FACEGRAB: Face Recognition with General Region Assigned to Binary Operator. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 2010. 3, 5, 6

[22] Z. Stone, T. Zickler, and T. Darrell. Autotagging facebook: Social network context improves photo annotation. In *In Proceedings of CVPR Workshop on Internet Vision*, 2008. 2

[23] Z. Stone, T. Zickler, and T. Darrell. Toward large-scale face recognition using social network context. *Proceedings of the IEEE*, 98(8):1408 –1415, aug. 2010. 2

[24] Y. Tian, W. Liu, R. Xiao, F. Wen, and X. Tang. A face annotation framework with partial clustering and interactive labeling. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8, june 2007. 2

[25] G. Wang, A. Gallagher, J. Luo, and D. Forsyth. Seeing people in social context: recognizing people and social relationships. In *Proceedings of the 11th European conference on Computer vision: Part V*, ECCV'10, pages 169–182, Berlin, Heidelberg, 2010. Springer-Verlag. 2, 4, 5

[26] L. Wolf, T. Hassner, and Y. Taigman. Similarity scores based on background samples. In *Proceedings of the 9th Asian conference on Computer Vision - Volume Part II*, ACCV'09, pages 88–97, Berlin, Heidelberg, 2010. Springer-Verlag. 2

[27] L. Wolf, T. Hassner, and Y.Taigman. Descriptor based methods in the wild. In *Real-Life Images workshop at the European Conference on Computer Vision (ECCV)*, October 2008. 2

[28] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992. 3

[29] Q. Yin, X. Tang, and J. Sun. An associate-predict model for face recognition. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 497–504, Washington, DC, USA, 2011. IEEE Computer Society. 2

[30] L. Zhang, L. Chen, M. Li, and H. Zhang. Automated annotation of human faces in family albums. In *In MULTIMEDIA 03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 355–358. ACM Press, 2003. 2

[31] M. Zhao and S. Liu. Automatic person annotation of family photo album. In *Proc. International Conf. on Image and Video Retrieval*, pages 163–172, 2006. 2

[32] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, Dec. 2003. 1

[33] J. Zhu, M. Vai, and P. Mak. Anew enhanced nearest feature space $enfs$ classifier for gabor wavelets featuresbased face recognition. In D. Zhang and A. Jain, editors, *Biometric Authentication*, volume 3072 of *Lecture Notes in Computer Science*, pages 29–45. Springer Berlin / Heidelberg, 2004. 10.1007/978-3-540-25948-0_18. 3, 5