A Multi-Level Contextual Model For Person Recognition in Photo Albums

Haoxiang Li[†], Jonathan Brandt[‡], Zhe Lin[‡], Xiaohui Shen[‡], Gang Hua[♮]

[†]Stevens Institute of Technology [‡]Adobe Research [♮]Microsoft Research
[†]hli18@stevens.edu [‡]{jbrandt, zlin, xshen}@adobe.com [♮]ganghua@microsoft.com

Abstract

In this work, we present a new framework for person recognition in photo albums that exploits contextual cues at multiple levels, spanning individual persons, individual photos, and photo groups. Through experiments, we show that the information available at each of these distinct contextual levels provides complementary cues as to person identities. At the person level, we leverage clothing and body appearance in addition to facial appearance, and to compensate for instances where the faces are not visible. At the photo level we leverage a learned prior on the joint distribution of identities on the same photo to guide the identity assignments. Going beyond a single photo, we are able to infer natural groupings of photos with shared context in an unsupervised manner. By exploiting this shared contextual information, we are able to reduce the identity search space and exploit higher intra-personal appearance consistency within photo groups. Our new framework enables efficient use of these complementary multi-level contextual cues to improve overall recognition rates on the photo album person recognition task, as demonstrated through state-of-the-art results on a challenging public dataset. Our results outperform competing methods by a significant margin, while being computationally efficient and practical in a real world application.

1. Introduction

After decades of research, the problem of face recognition as measured by standard benchmarks such as Labeled Faces in the Wild (LFW) [11, 12] is to the point of being nearly solved. For example, Schroff et al. [24] achieved 99.63% verification accuracy using a deep convolutional neural network (CNN). That said, this impressive result is misleading because such benchmarks are typically skewed towards images with clearly visible, high quality faces (see Figure 1). However, when applying face recognition for tagging faces in the real-world photo albums, faces are often not so clearly visible and present many challenges due to changes in body pose, illumination, heavy occlusion and so forth (see Figure 2). As an example, Zhang et al. [31] observed a dramatic drop in recognition accuracy when a top performing LFW algorithm was applied to a photo album test set.

Figure 1. Faces in the Labeled Faces in the Wild dataset: most faces are clear with good image quality.

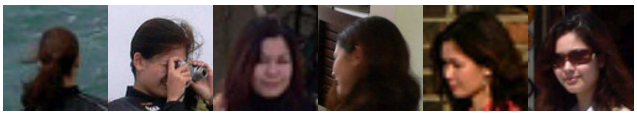
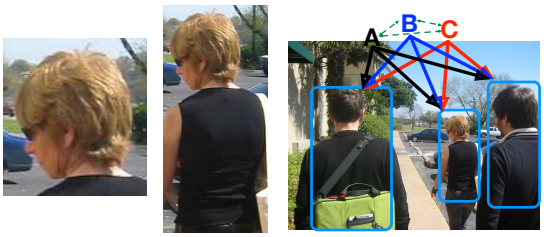


Figure 2. Faces in the People In Photo Albums dataset: the large visual appearance variations are very challenging for face recognition.

In addition to learning more discriminative and robust face features, we must look beyond the faces in order to approach human-level performance on thismore challenging task. In general, information beyond the faces can be viewed as context to guide recognition. Such "extra-face" context can naturally be divided into three levels: the person or body level, the photo level, and the photo group level, as depicted in Figure 3.

At the person level, perhaps the most obvious contextual cue is clothing, which has been shown to be an effective supplemental cue for face recognition in photo albums [1, 8, 25, 30]. However, it is still not well understood how best to fuse face and body appearance features as the relative importance of these features depends on higher level context. In this paper, we explore several alternatives for fusing person-level appearance features.

Contextual cues at the photo level include metadata (when present) such as geographical and temporal information [19], event labels, and social relationships [17, 26]. However, we cannot always rely on metadata to be present



(a) person-level context (b) photo-level context

(c) group-level context

Figure 3. The three levels of context that can be exploited for person recognition in photo albums: a) person-level context consists of face and body appearance features; b) photo-level context includes identity co-occurrences and mutual exclusion; c) group-level context presents higher intra-personal appearance consistency and reduced identity search space.

and accurate. Therefore, we seek a robust method that can take advantage of such metadata when available, but can still exploit photo level context when the metadata is absent. Aside from the metadata, we have found that we can leverage a rough prior on the co-occurrences of particular individuals within a photo, as well as a soft mutual exclusion constraint, to substantially improve recognition.

Going beyond a single photo, we observe that album photos frequently occur in groups that are closely related, such as being taken on the same day or same event or same setting. When such groupings are present and can be automatically determined, then it becomes possible to exploit mutual information across the photos in a group to improve recognition. In particular, we can adapt the person appearance classifiers to a given group. Metadata, when available, can be used for effective groupings. However, in order to be robust in the absence of metadata, we propose an unsupervised method to determine the effective photo groups based on photo appearance.

In summary, in order to go "beyond faces" for person recognition in photo albums, it is necessary to exploit contextual information. This paper presents a framework that is both effective and efficient, based on three levels of contextual information, namely person, photo, and photo group. Our contributions are as follows:

- a practical and efficient multi-level contextual model achieving state-of-the-art results on the People In Photo Albums (PIPA) person recognition benchmark;
- an iterative joint inference process to leverage the photo-level context cues;
- an unsupervised, metadata-free method to discover relevant photos that provide group-level context for improved person recognition;
- an effective confidence-aware method for fusing the person-level appearance cues.

2. Related Work

Researchers have been long interested in using context cues in recognition [6, 7, 8, 14, 17, 22, 25]. Clothing features [8], metadata [25], location and event labels [17] have been used as contexts to improve face tagging. Geographic contexts, spatial contexts, temporal contexts or even more high-level cultural contexts have been used for image classification, object detection and classification [4, 9, 10, 29].

However, most of these context cues are from metadata or require manual labels beyond the identity domain and are inapplicable in the absence of these information. In this paper, we propose a contextual model that does not rely on metadata but can benefit from it when available.

The most relevant literature to ours are Zhang et al. [31] and Oh et al. [21]. Zhang et al. [31] published the PIPA dataset to study person recognition. Both works observed that context cues beyond the faces help improve person recognition accuracy. Zhang et al. [31] used discriminative information from poselets [2]. Oh et al. [21] carefully evaluated the effectiveness of different body regions, the scene context and some long term attributes (e.g., age and gender).

However, the context cues in their works are mostly at the person-level. They do not exploit the joint identity assignment of multiple instances at the photo-level or the mutual information across multiple photos at the group-level. In our work, we have found that recognition accuracy improves significantly when we incorporate context cues at multiple levels.

Our work leverages Conditional Random Fields (CRF) to jointly infer the identities of instances in the same photo to exploit the identity co-occurrences. In this sense, the methods from Stone et al. [26] and Brenner et al. [3] are relevant to ours. However, the former relies on the social context to estimate the relationships between people and the latter jointly processes the entire set as a sparse Markov Random Fields, which can be computationally expensive in processing a large-scale photo collection.

To our knowledge, the framework presented here is the first to effectively leverage context cues across multiple levels for person recognition in photo albums. In addition to person-level appearance features and photo-level identity

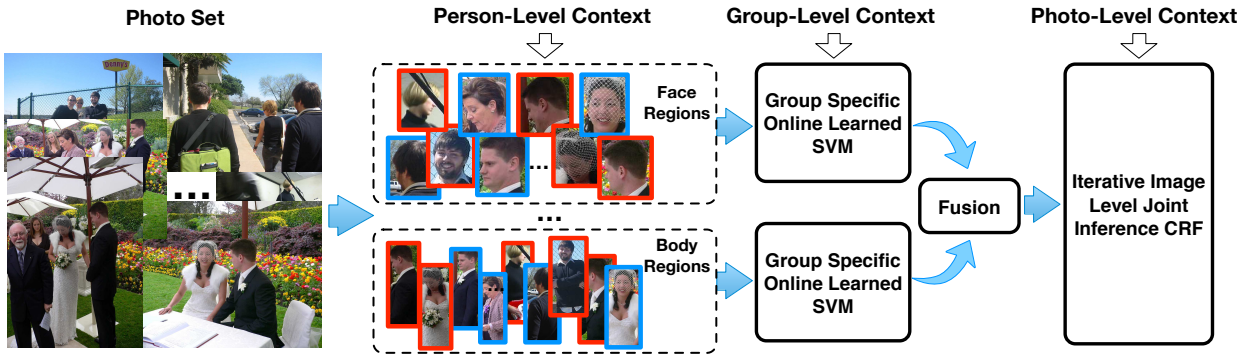


Figure 4. The proposed multi-level contextual model.

co-occurrences, we leverage groups of relevant photos specific to each testing photo as its group-level context.

3. The Multi-Level Contextual Model

This work addresses the identification setting for person recognition in a photo album. Specifically, we are given a set of photos containing a set of person instances that have been grouped into two disjoint sets: the gallery set where identity labels are assigned to each instance, and the probe set where the identity labels are unknown. Our task is to predict the identities of all unlabeled instances in the probe set. In this work, we assume that there is at least one labeled instance for each identity in the gallery set.

3.1. Framework

As shown in Figure 4, the proposed framework processes the face and body regions separately with the group specific online learned SVMs and then fuses the outputs. The outputs are then iteratively updated with the photo level joint inference CRF.

Multiple regions present complementary yet discriminative information in the person-level context. The proposed method uses face and body regions and can be naturally expanded to incorporate more regions. We discuss the options to fuse the predictions with different regions in Section 3.2.

In related photos, people present more consistent appearance. The proposed method discovers groups of related photos to learn group specific SVMs online for prediction. We describe the details in Section 3.3.

As shown in Figure 4, after leveraging the person-level and group-level context, the results are updated with an iterative image level joint inference CRF. In this step, we first estimate the identity co-occurrences based on the current predictions of the unlabeled instances and the given labeled instances. We then proceed to encode this prior knowledge to jointly infer identities of all unlabeled instances for each photo to update our predictions. After that, we update the prior knowledge based on the current predictions and repeat this process iteratively for several times. We describe this part in Section 3.4.

3.2. Person-Level Context

Identity information exists in the appearance of clothes, hairstyles and other regions. Oh et al. [21] carefully explored the effectiveness of different regions such as the head, body, upper body and scene. In this work, for simplicity, we only include the face and body regions shown in Figure 3. We use a face recognition system (detailed in the experiment section) to extract the face features from the face regions. We fine-tune a CNN pre-trained for image classification with the body regions to extract body features using the soft-max classification objective over identities. By fusing the information from different regions, person-level context is incorporated.

The proposed method can be expanded to incorporate more regions. Without loss of generality, we assume C regions are used in the proposed framework. Assume there are M people in the photo set. With the c-th region, we can apply an identity classifier (will be detailed in Section 3.3) to obtain the prediction $\mathbf{s}_c(x)$ as an M-dimensional vector. The y-th element in $\mathbf{s}_c(x)$ indicates the probability that the instance x is of identity label y. Given the instance x, we proceed with C regions and obtain C prediction score vectors: $\mathbf{s}_1(x), \ldots, \mathbf{s}_C(x)$. We fuse the scores from different regions as our final prediction.

Specifically, we explore the following options to fuse the C prediction vectors to obtain boosted recognition result.

3.2.1 Weighted Average Fusion

Zhang et al. [31] combined the predictions from the poselet classifiers with a linear Support Vector Machine (SVM), which is equivalent of taking the element-wise weighted sum of all prediction score vectors, i.e.,

$$\mathbf{s}(x) = \sum_{c=1}^{C} \omega_c \mathbf{s}_c(x), \tag{1}$$

in which ω_c is the weight for the c-th region. The weights can be learned with a binary linear SVM over the C-dimensional score vectors from the validation dataset [31].

3.2.2 Max Pooling Fusion

Without learning the weights, another straightforward option to fuse the predictions is taking the element-wise max operation, i.e.,

$$\mathbf{s}(x) = \max_{c \in [1,C]} \mathbf{s}_c(x) \tag{2}$$

3.2.3 Confidence-aware Fusion

Having fixed combination weights $\{\omega_c\}_{c=1}^{C}$ may not be optimal. Another option for fusion is to assign instance specific weights based on the prediction confidence scores.

By sorting the elements in $\mathbf{s}_c(x)$ in descending order: $[s_1, s_2, \ldots, s_M]$. , we define the weight $\omega_c(x)$,

$$\omega_c(x) = \frac{1}{Z}\frac{(s_1 - s_2)}{(s_1 - s_M)}, \quad \mathbf{s}(x) = \sum_{c=1}^{C} \omega_c(x)\mathbf{s}_c(x), \tag{3}$$

where Z is the normalization term to ensure $\sum_{c=1}^{C} \omega_c(x) = 1$.

We evaluate all these options in the experiments (see Section 4.4).

3.3. Group-Level Context

In the identification setting for person recognition, the problem naturally fits into a multi-class classification paradigm. With a specific region feature to represent the instance, we can learn a classifier from all the gallery instances. We name it the global multi-class classifier. In this work, we use linear SVM as the classifier. Linear SVM is efficient and it generalizes well to unseen data with limited amount of training samples. For multi-class classification, we follow the 1-versus-all paradigm to train the multi-class SVM using the LIBLINEAR [5] implementation.

When people change clothes, hairstyles or when the photos are taken from different viewpoints, the appearance of the same person can change dramatically. The large appearance variations lead to highly non-linear class boundary in the feature space and hence affect the robustness of the global classifier. However, we observe that when photos are properly grouped, the intra-personal variation is reduced, as shown in Figure 5, which allows us to leverage the mutual information inside the group to help person recognition. However, this extra information is not always available or is complete. Hence, we need a method to automatically discover the photo groups.

We first define the relevancy between two photos using the photo similarity and identity co-occurrence. The intuition is that when two photos are visually similar or contain

Figure 5. The instances of the same person in four photo groups: the intra-personal visual appearance variations become smaller within each group.

the same persons, they are more likely to be related. Given an instance x in a query photo, we determine a set of "neighbor" photos most relevant to the query photo, and use them to exploit the context cues at group-level.

We use the caffe [13] implementation of AlexNet [15] to extract holistic image feature F_I for photo I. Given two photos I_i and I_j, we denote the labeled instances in the two photos as $\{x_1, x_2, \ldots, x_{N_i}\}$ and $\{x'_1, x'_2, \ldots, x'_{N_j}\}$ respectively. The identity label for any instance x is denoted as $Y(x)$. We define the affinity $\Lambda_{i,j}$ between photos I_i and I_j as

$$\Lambda_{i,j} = cos(F_{I_i}, F_{I_j}) * (1 + \frac{N_{ij}}{N_i + N_j - N_{ij}})/2, \tag{4}$$

where

$$N_{ij} = \sum_{1 \le u \le N_i,\ 1 \le v \le N_j} [Y(x_u) = Y(x'_v)],$$

$$cos(F_{I_i}, F_{I_j}) = \frac{F_{I_i} \cdot F_{I_j}}{||F_{I_i}|| \cdot ||F_{I_j}||},$$

where $[P]$ is the Iverson bracket, i.e., $[P] = 1$ if P is true otherwise $[P] = 0$. When $N_i = 0$ or $N_j = 0$, we set $\Lambda_{i,j} = cos(F_{I_i}, F_{I_j})$.

N_{ij} measures the identity co-occurrences between two photos, while $cos(F_{I_i}, F_{I_j})$ indicates the photo similarity. The photo similarity part in $\Lambda_{i,j}$ is more important since the labeled instances are usually limited, which can lead to inaccurate estimation of the identity co-occurrences. We still choose to keep the identity co-occurrence part here because it slightly improves the recognition accuracy (around 0.3% on average) with an ignorable overhead.

We then use this affinity matrix for spectral embedding [20]. In this way, we embed all photos into the space in which the neighbors are of higher relevancy. Technically, we calculate the normalized Laplacian matrix L from Λ,

$$L = D^{-\frac{1}{2}} \Lambda D^{-\frac{1}{2}}, \quad D_{i,i} = \sum_j \Lambda_{i,j}. \tag{5}$$

Then we find the K largest eigen-vectors of L, i.e., $X_1, \ldots, X_K$. Assuming L is a $N \times N$ matrix, then each X_k is a N dimensional vector. The K vectors are stacked as rows in a $K \times N$ matrix. We represent the n-th photo by the n-th column in this matrix, which is a K dimensional vector. After embedding, the euclidean distance between two photos indicates their relevancy.

After the spectral embedding, given a probe instance in photo I, we choose the N_{contex} nearest neighbors of I as its group-level context. We then train an online, group-specific SVM classifier with the labeled instances in its group-level context for prediction. Note that when there is no labeled instance in the photo group, we use the global SVM for prediction. If there is only one identity in the photo group, we assign the identity label to the probe instance as its group specific output.

The group specific classifier may suffer from the insufficient training samples problem. For regularization, we combine the prediction from the globally trained SVM with the prediction from the group specific SVM. We observe that taking average of the two predictions works well. Intuitively, the global SVM is adapted to the group in this way.

3.3.1 Album Information

In many personal photo management softwares, the photos are usually organized as albums. If this information is available, given a photo, we simply use its photo album as the group-level context. An SVM is trained for each album as the group specific SVM.

3.4. Photo-Level Context

As shown in Figure 4, after fusing the recognition results from different regions, we further update the predictions with the photo-level context. At the photo-level, we jointly predict multiple instances in a single photo, in which we can encode the following prior knowledges.

First, intuitively, two people have a higher chance to appear in the same photo when they know each other. We can bias our predictions to assign co-occurred identities to instances in the same photo. Second, in general, two instances in the same photo are rarely the same person. We can largely reduce the possibilities of assigning duplicated labels in the same photo.

Given the current predictions, we incorporate these knowledges to update the predictions with the Conditional Random Fields (CRF). Technically, we define a label compatibility matrix ψ to encode the knowledge. Given two identity labels l_a and l_b, $\psi(l_a, l_b)$ indicates how likely the two identities are in the same photo.

$$\psi(l_a, l_b) = \begin{cases} \epsilon & l_a = l_b, \\ \alpha & l_a, l_b \text{ co-occurred}, \\ 1 & \text{otherwise}, \end{cases} \tag{6}$$

where $\alpha >= 1$ is a parameter for the strength of the identity co-occurrence assumption; a small $\epsilon = 0.01$ is used to enforce a soft mutual exclusion constraint; l_a and l_b are regarded as co-occurred, when two instances of them have been observed in the same photo.

We have explored other options for α. For example, instead of having α as a constant value, we tried setting it proportional to the number of times l_a, l_b co-occurred in the photo set. However, due to the limited number of samples, this statistic is not reliable. We observe that having a constant α provides more stable improvement.

In the CRF, we look for an identity assignment $Y = [y_1, \ldots, y_N], y_n \in [1, M]$ over the N instances $\{x_n\}, n \in [1, N]$ in the image I. Y is regarded as a random variable conditioned on the image I. We jointly predict the identities of the N instances to maximize the potential $E(Y)$,

$$E(Y) = \sum_{n=1}^{N} \phi(y_n) + \sum_{n,m} \psi(y_n, y_m). \tag{7}$$

The unary potential $\phi(y_n)$ is from the current prediction. i.e., the M-dimensional score vector $\mathbf{s}(x_n)$,

$$\phi(y_n) = y_n\text{-th element in } \mathbf{s}(x_n). \tag{8}$$

For any labeled instance of identity label l, we set $\phi(y_n) = [y_n = l]$, where $[P]$ is the Iverson bracket.

The pairwise potentials are from the label compatibility matrix ψ. The Loopy Belief Propagation [18] is applied for the CRF inference with the UGM implementation [23].

Lastly, since the label compatibility matrix is estimated from the current prediction, it may not be accurate enough. We propose to update the label compatibility matrix each time after we obtain the updated predictions from the joint inference. The updated label compatibility matrix is more accurate and helps the joint inference. This loop is iterated for T times. Typically after $T = 5$ iterations, it converges. We observe consistent improvement by having this iterative joint inference process.

4. Experiments on the PIPA Dataset

The proposed method is evaluated on the PIPA dataset and compared with the state-of-the-art methods. We noticed some incorrect labels in the test set of PIPA. To accurately evaluate the proposed method, we manually curated the test set to remove label noise and merge redundant labels. The curation does not make the problem easier. We show some examples of this in the supplemental material. For fairness, we use the original labels when compared with existing methods.

4.1. Setting

The PIPA dataset is composed of the training set, validation set and test set. The training set is for model training

Table 1. Evaluation of different strategies for multiple regions fusion.

Splits	Face	Body	Concatenated Features	Weighted Average	Max Pooling	Confidence-aware
original	67.89%	74.89%	71.15%	**84.38%**	82.70%	83.86%
album	64.87%	64.90%	67.81%	78.18%	77.41%	**78.23%**
time	58.86%	51.84%	61.76%	69.60%	68.76%	**70.29%**
day	54.23%	23.28%	54.46%	52.18%	54.95%	**56.40%**

Table 2. Evaluation of different components in the proposed framework with curated labels: a) baseline method: fusion of global classifiers from face and body regions. b) baseline method with photo-level context; c) baseline method with group-level context; d) multi-level contextual model; e) multi-level contextual model with album information.

stage	original	album	time	day
a)	83.86%	78.23%	70.29%	56.40%
b)	84.84%	79.13%	71.94%	57.37%
c)	85.10%	79.63%	72.02%	57.33%
d)	**88.20%**	**83.02%**	**77.04%**	**59.77%**
e)	**94.27%**	**83.79%**	**80.64%**	**61.77%**

and the test set is for evaluation. The test set has 7,868 photos in 357 albums. There are 12,886 labeled instances of 581 different individuals. The evaluation protocol on PIPA follows the person identification setting. The test set is split into the gallery set and probe set. Given the head bounding boxes of all instances in the test set and the identity labels for all instances in the gallery set, we train our person recognition system to predict the missing identity labels in the probe set. Then we switch the gallery set and probe set, repeat the experiment and report the average accuracy.

In the original setup, a probe instance may have near duplicate gallery instance. To study how well the person recognition system can address long-term appearance changes, Oh et al. [21] proposed three more challenging and realistic splits. We incorporate their evaluation protocols and report our results on all the four splits: original, album, time and day.

Generally, the four splits are in the order of increasing difficulty. In the album splits, the gallery and probe instances are split across different albums. In the time splits, instances are split into newest versus oldest based on the photo-taken time. They manually setup the day splits to enforce appearance changes. We refer the readers to Oh et al. [21] for details about the construction of these splits.

4.2. Face Region

We implemented a deep CNN based face recognition system following Sun et al. [27] for feature extraction from face regions. The number of parameters in the CNNs in the system is about one-fourth of the AlexNet [15]. On the face recognition benchmark LFW, it achieves an accuracy of 97.65% comparable to 97.35% from DeepFace [28].

On the PIPA dataset, with the given head bounding box, we apply a face detector we implemented following Li et al. [16] to detect the largest face inside the bounding box. Then we extract the face feature from the detection box.

4.3. Body Region

We follow Zhang et al. [31] to fine tune the AlexNet [15] using body regions on the training set. Given a head bounding box centered at (x, y) of size (w, h), we estimate the upper body region to be a box centered at $(x - 0.5l, y)$ of size $(2l, 3l)$, in which $l = min(w, h)$.

4.4. Score Fusion

To evaluate the different options for fusion, we train global multi-class classifiers with the face and body regions respectively for person recognition and fuse the results with all the options.

An alternative to score-level fusion is to concatenate the features from all regions to represent the instance and proceed with the other components for prediction. We observe that this fusion method performs worse than any of the score fusion methods, as shown in Table 1.

The weighted average fusion method works best for the original splits. However, when the setup becomes more challenging, the learned fixed weights may not help all probe instances. In the most challenging day splits, it fails to improve the face recognition performance. The max pooling fusion gives consistent improvements after fusion. However, since the prediction scores from different regions are not calibrated, it is not the most effective. The performance of the confidence-aware fusion method is stable across all setups. We use this method in our following experiments.

4.5. Framework Components

We introduce several comparison methods, with which we demonstrate how the components in the proposed framework contribute to the improvement in Table 2.

- *baseline method*: the confidence-aware fusion of the global face and body SVMs is our baseline method;
- *baseline method with photo-level context*: we directly update the results from the baseline method with the iterative image label joint inference CRF;

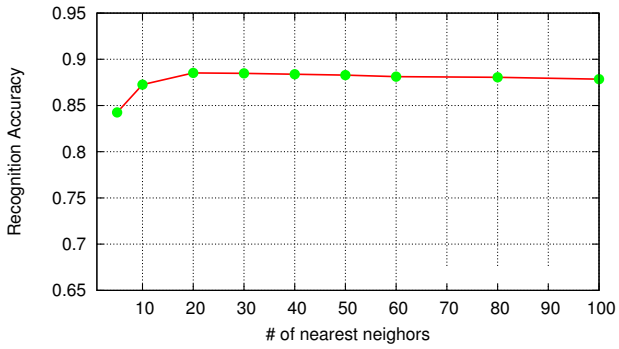


Figure 6. The recognition accuracy with respect to the number of nearest neighbors in group-level context on the original splits.

- *baseline method with group-level context*: we skip the iterative image label joint inference CRF in Figure 4;
- *multi-level contextual model*: the full proposed framework in Figure 4.

In Table 1, we demonstrate the effectiveness of exploiting person-level context. In Table 2, we observe that both the group-level context and photo-level context help improve the recognition accuracy independently. Moreover, the proposed framework leverages contexts at multiple levels and achieves further improvement.

4.6. Metadata

In Table 2, we also observe that the proposed framework can leverage extra information, when available, e.g., the album information, which is presented as grouping of photos. It helps because the same person usually has more consistent visual appearance in the same album and the total number of identities are limited in a photo album. As we can see, the album information consistently helps because it arbitrarily separate irrelevant photos.

4.7. Parameters

We set the spectral embedding dimension to be 400. We choose top $N_{context} = 50$ nearest neighbors after spectral embedding as the photo's group-level context. In Figure 7, we show examples of discovered group-level contexts. We observe the proposed method works with a range of values for the spectral embedding dimension. We evaluate the influence of $N_{context}$ to the recognition accuracy. As shown in Figure 6, the performance is stable with $N_{context}$ in the range of 10 to 100. A very small $N_{context} = 5$ degrades the accuracy due to limited training samples. With a very large $N_{context}$, the group specific SVMs degenerate to the global SVM.

In estimating the label compatibility matrix for the photo-level joint inference, we set $\alpha = 2$ for the strength of the identity co-occurrence prior and $T = 5$ for the number of iterations of the joint inference. The proposed method works with a range of reasonable values for these parameters.

Note that by setting $\alpha = 1$, we only encode the identity mutual exclusion knowledge in the label compatibility matrix, which leads to an accuracy drop compared with $\alpha = 2$. For example, in the original splits, the accuracy drops from 87.81% ($\alpha = 2$) to 85.74% ($\alpha = 1$).

Table 3. Recognition accuracy comparison with original labels: best results with and without extra information are highlighted.

methods	original	album	time	day
Oh et al. [21]	86.78%	78.72%	69.29%	46.61%
Zhang et al. [31]	83.05%	-	-	-
our method	**88.75%**	**83.33%**	**77.00%**	**59.35%**
ours with album info.	**93.91%**	**83.44%**	**80.23%**	**61.62%**

4.8. Results Comparison

We compare our results with the existing methods on PIPA in Table 3. The proposed method significantly outperforms all the existing methods. The improvement is more significant for more challenging splits, which demonstrate the effectiveness of exploiting the contexts at multiple levels. When album information is available, our method benefits from the extra information and achieves even dramatic improvements.

4.9. Computational Expense

Besides the improved recognition accuracy, we only use 3 deep CNNs in our method which is more efficient than previous state-of-the-art [21, 31]. Note that Zhang et al. [31] used more than 100 deep CNNs for feature extraction and Oh et al. [21] used 17 deep CNNs. Applying a deep CNN for feature extraction takes a large portion of time budget. Hence our method is more practical. The photo-level context based update is very efficient that it takes only 90 seconds to process 7,839 photos, in which half of the instances are unlabeled.

5. Experiments on the Gallagher Album

In this experiment, we test generalization of our method for another dataset. To this end, we pick the Gallagher Album [8] dataset, which is a small collection of family photos consists of 589 images with 931 annotated faces from 32 people. We show that our method also improves on a single photo album.

Following the same identification setting, for each identity, we randomly select half of the instances into the gallery

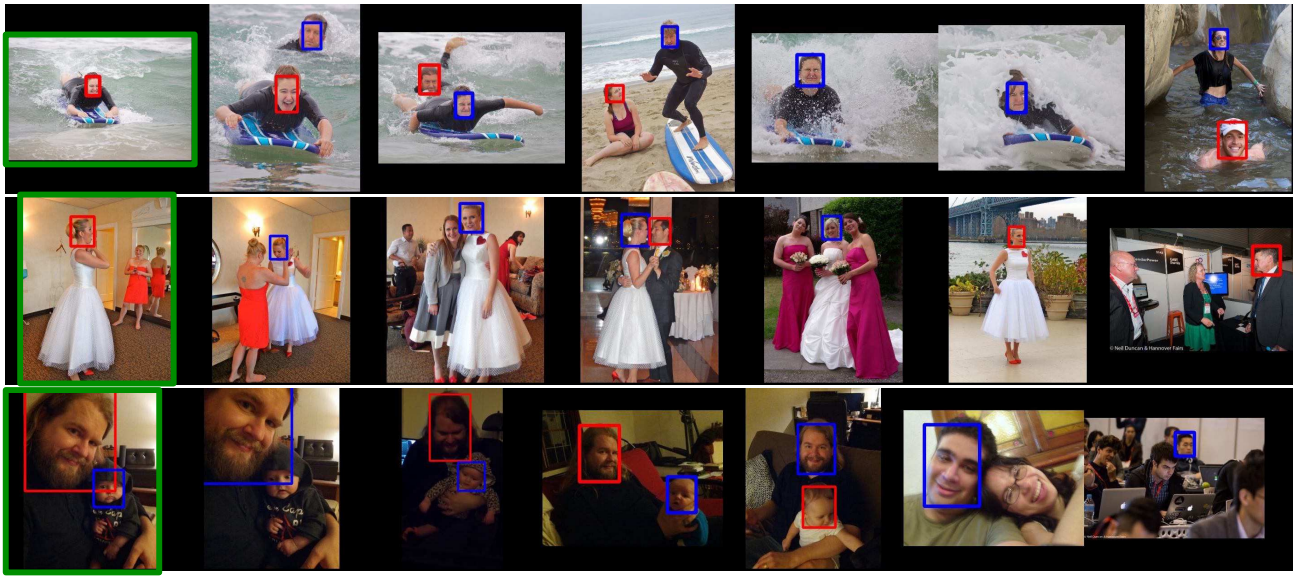


Figure 7. Examples of group-level contexts (showing top 7 nearest neighbors): red boxes indicate the unlabeled instances and blue boxes indicate the labeled instances; the first photo in each row (green bounding box) is the testing photo.

Table 4. Recognition accuracy on the Gallagher Album dataset.

components	Exp 1	Exp 2	Exp 3
face region global classifier	83.40%	83.19%	81.93%
body region global classifier	80.88%	78.78%	78.15%
baseline: confidence-aware fusion	86.34%	87.61%	85.50%
baseline with group-level context	85.50%	86.13%	83.19%
baseline with photo-level context	**89.08%**	**88.66%**	**87.61%**
full multi-level contextual model	88.87%	86.55%	86.76%

set and the rest into the probe set. We repeat the process to obtain 3 random splits. To test the generalization of our system, we use the same parameters in the evaluation on PIPA.

A small group of people dominate the identity distribution in this dataset. There are 5 people with totally 777 out of 931 instances. The rest of the people have a limited number of instances.

Because most identities are rare in the dataset, we need to select a very large group of photos as the group-level context to cover them, which is contrary to the motivation of exploiting the group-level context. Hence, it is better to regard the whole dataset as a single album and skip the group-level context component in the proposed framework. As shown in Table 4, the group-level context does not help because of the long-tailed identity distribution. However, we observe consistent improvements with the person-level and photo-level contexts.

6. Conclusion

In this paper, we propose a multi-level contextual model for person recognition. Our model exploits discriminative visual information from multiple regions. We propose to use a confidence-aware fusion method to integrate the discriminative information from all the regions. The proposed model uses photo group to reduce the identity search space and leverage the smaller intra-personal appearance changes within the same group. It naturally incorporates metadata such as the album information, when available, to help group photos, but does not rely on metadata. Additionally, the proposed method encodes an identity co-occurrence prior to jointly infer the identities of instances in the same photo. Multiple levels of context cues are leveraged with the proposed model. We demonstrate significant improvements over the state-of-the-art results on the challenging PIPA dataset while being more computationally efficient.

Acknowledgment

This work is partially done when the first author was an intern at Adobe Research. Research reported in this publication was partly supported by the National Institute Of Nursing Research of the National Institutes of Health under Award Number R01NR015371. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work is also partly supported by US National Science Foundation Grant IIS 1350763 and GH's start-up funds from Stevens Institute of Technology.

References

[1] D. Anguelov, K.-c. Lee, S. B. Gökturk, and B. Sumengen. Contextual identity recognition in personal photo albums. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 1

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *IEEE International Conference on Computer Vision*, 2009. 2

[3] M. Brenner and E. Izquierdo. Joint people recognition across photo collections using sparse markov random fields. In *MultiMedia Modeling*, pages 340–352. Springer, 2014. 2

[4] S. K. Divvala, D. Hoiem, J. H. Hays, A. Efros, M. Hebert, et al. An empirical study of context in object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[5] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 2008. 4

[6] A. Gallagher and T. Chen. Understanding images of groups of people. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[7] A. C. Gallagher and T. Chen. Using group prior to identify people in consumer images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 2

[8] A. C. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 1, 2, 7

[9] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 2

[10] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *European Conference on Computer Vision*, pages 30–43. 2008. 2

[11] G. B. Huang and E. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, 2014. 1

[12] 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, 2007. 1

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 4

[14] A. Kapoor, G. Hua, A. Akbarzadeh, and S. Baker. Which faces to tag: Adding prior constraints into active learning. In *IEEE International Conference on Computer Vision*, 2009. 2

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 1097–1105. 2012. 4, 6

[16] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[17] D. Lin, A. Kapoor, G. Hua, and S. Baker. Joint people, event, and location recognition in personal photo collections using cross-domain context. In *European Conference on Computer Vision*. 2010. 1, 2

[18] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 1999. 5

[19] M. Naaman, R. B. Yeh, H. Garcia-Molina, and A. Paepcke. Leveraging context to resolve identity in photo albums. In *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*, 2005. 1

[20] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2002. 4

[21] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele. Person recognition in personal photo collections. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 2, 3, 6, 7

[22] N. O'Hare and A. F. Smeaton. Context-aware person identification in personal photo collections. *Multimedia, IEEE Transactions on*, 2009. 2

[23] M. Schmidt. Ugm: A matlab toolbox for probabilistic undirected graphical models, 2010. 5

[24] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[25] Y. Song and T. Leung. Context-aided human recognition–clustering. In *European Conference on Computer Vision*. 2006. 1, 2

[26] Z. Stone, T. Zickler, and T. Darrell. Autotagging facebook: Social network context improves photo annotation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, (CVPRW)*, 2008. 1, 2

[27] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 6

[28] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 6

[29] K. Tang, M. Paluri, L. Fei-Fei, R. Fergus, and L. Bourdev. Improving image classification with location context. *arXiv preprint arXiv:1505.03873*, 2015. 2

[30] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. "knock! knock! who is it?" probabilistic person identification in tv-series. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1

[31] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 3, 4, 6, 7