Unifying Identification and Context Learning for Person Recognition

Qingqiu Huang, Yu Xiong, Dahua Lin
CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
{hq016, xy017, dhlin}@ie.cuhk.edu.hk

Abstract

Despite the great success of face recognition techniques, recognizing persons under unconstrained settings remains challenging. Issues like profile views, unfavorable lighting, and occlusions can cause substantial difficulties. Previous works have attempted to tackle this problem by exploiting the context, e.g. *clothes and social relations. While showing promising improvement, they are usually limited in two important aspects, relying on simple heuristics to combine different cues and separating the construction of context from people identities. In this work, we aim to move beyond such limitations and propose a new framework to leverage context for person recognition. In particular, we propose a Region Attention Network, which is learned to adaptively combine visual cues with instance-dependent weights. We also develop a unified formulation, where the social contexts are learned along with the reasoning of people identities. These models substantially improve the robustness when working with the complex contextual relations in unconstrained environments. On two large datasets, PIPA [27] and Cast In Movies (CIM), a new dataset proposed in this work, our method consistently achieves state-of-the-art performance under multiple evaluation policies.*

1. Introduction

Person recognition is a key task in computer vision and has been extensively studied over the past decades. Thanks to the advances in deep learning, recent years have witnessed remarkable progress in face recognition techniques [22, 19, 21, 17]. On LFW [10], a challenging public benchmark, the accuracy has been pushed to over 99.8% [17]. Nonetheless, the success on benchmarks does not mean that the problem has been well solved. Recent studies [27, 11, 13, 15] have shown that recognizing persons under an unconstrained setting remains very challenging. Substantial difficulties arise in unfavorable conditions, *e.g.* when the faces are in a non-frontal position, subject to extreme lighting, or too far away from the camera. Such conditions are very common in practice.

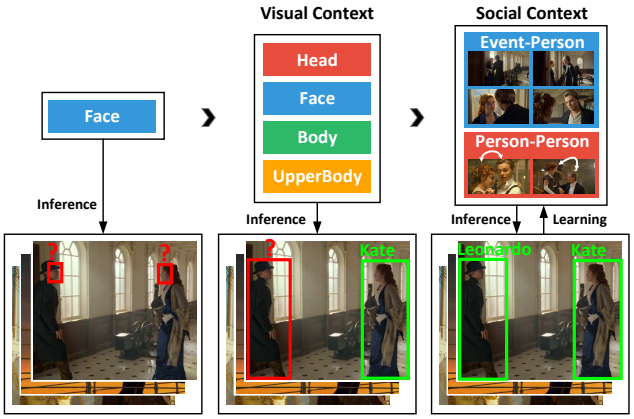

Figure 1: Person recognition under unconstrained settings remains a very challenging problem. Inference purely by face recognition techniques would fail in many cases. We propose a framework to tackle this problem, which combines *visual context* with adaptive weights and unifies person recognition with *social context* learning.

The difficulties above are essentially due to the fact that facial appearance is highly sensitive to environmental conditions. To tackle this problem, a natural idea is to leverage another important source of information, namely the *context*. It is our common experience that we can easily recognize a familiar person by looking at the wearing, the surrounding environment, or the people who are nearby. On the other hand, cognitive neuroscience studies [6, 1, 12] have shown that *context* plays a crucial role when we, as human beings, recognize a person or an object. A familiar context often allows much greater accuracy in recognition.

Exploiting context to help recognition is not a new story in computer vision. Previous efforts mainly follow two lines. The first line of research [2, 8, 27, 11] attempts to incorporate additional visual cues, *e.g.* clothes and hairstyles, as additional channels of features. The other line, instead, focuses on social relationships, *e.g.* group priors [7, 20] or people co-occurrence [4, 23]. There are also studies that try to integrate both visual cues and social relations [15, 13].

Whereas previous works have shown the utility of context in person recognition, especially in unconstrained en-

vironments, a key question remains open, that is, *how to discover and leverage contexts robustly*. Specifically, existing methods usually rely on simple heuristics, *e.g*. feature-based clustering, to establish contextual priors, and hand-crafted rules to combine contextual cues from different channels. Moreover, the construction of the context model is typically done *separately* and *before* person identification. The limitations of such approaches lie in two important aspects: (1) Heuristics designed manually are difficult to capture the diversity and complexity in unconstrained context modeling. (2) The identities of the people in a scene are also an important part of the context. Constructing the context separately would lose this significant connection.

In this work, we aim to explore more effective ways to leverage the context in person recognition (see Figure 1). Inspired by previous efforts, we consider two kinds of contexts, namely the visual context, *e.g*. additional visual cues, and the social context, *e.g*. the events that a person often attends. But, we move beyond the limitations of existing methods, by considering context learning and person identification as a unified process and solving both jointly. Driven by this idea, we propose novel methods for leveraging visual and social contexts respectively. Particularly, we develop a *Region Attention Network*, which is *learned end-to-end* to combine various visual cues adaptively with instance-dependent weights. We also develop a unified formulation, where the social context model is learned online *jointly* with the reasoning of people identities. As a by-product, the solution to this problem also comes with a set of *"events"* discovered from the given photo collection – an event not only share similar scenes but also a consistent set of attendants.

On PIPA [27], a large public benchmark for person recognition, our proposed method consistently outperform existing methods, under all evaluation policies. Particularly, in the most challenging *day split*, our method raised the state-of-the-art performance from 59.77% to 67.16%. To assess our method in more diverse settings and to promote future research on this topic, we construct another large dataset, *Cast In Movies (CIM)*, by annotating the characters in 192 movies. This dataset contains more than $150K$ person instances and 1218 labeled identities. Our approach also demonstrated its effectiveness on *CIM*.

Our contributions mainly lie in three aspects: (1) For visual context, we propose a *Region Attention Network*, which combines visual cues with instance-dependent weights. (2) For social context, We propose a *unified formulation* that couples context learning with people identification. It also discovers events from photo collections automatically. These two techniques together result in remarkable performance gains over the state-of-the-art. (3) We construct *Cast In Movies (CIM)*, a large and diverse dataset for person recognition research.

2. Related Work

Early efforts. The significance of context in person recognition has long been recognized by the vision community. Early methods mainly tried to use additional visual cues, such as clothing [2, 8], or additional metadata [7]. Yet, the improvement was limited. Later, more sophisticated frameworks [20, 2, 16] that integrate multiple cues (clothing, timestamps, scenes, etc) have been developed. Some of these works [2, 16] formulated the task as a joint inference process over a Markov random field and obtained further performance gains. Note that these MRF-based methods assume the same set of people and social relations in both training and testing, and thus the learned models are difficult to generalize to new collections.

Recent efforts. The rise of deep learning has led to new innovations on this topic. *Zhang et al.* [27] proposed a Pose Invariant Person Recognition method (PIPER), which combines three types of visual recognizers based on ConvNets, respectively on face, full body, and poselet-level cues. The PIPA dataset published in [27] has been widely adopted as a standard benchmark to evaluate person recognition methods. *Oh et al.* [11] evaluated the effectiveness of different body regions, and used a weighted combination of the scores obtained from different regions for recognition.

Recently, *Li et al.* [13] proposed a multi-level contextual model, which integrates person-level, photo-level and group-level contexts. Although this framework also considers the combination of visual cues and social context, it differs from ours essentially in two key aspects: (1) The visual cues are combined with a simple heuristic rule, instead of a learned network. (2) The groups are identified by spectral clustering of scene features *before* person recognition, as a *separate* step. Our framework, instead, formulates event discovery and people identification as a unified optimization problem and solves both jointly.

Another way of integrating both visual cues and social relations was proposed in [15]. This work formulates the recognition of multiple people into a sequence prediction problem and tries to capture the relational cues with a recurrent network. As there is no inherent order among the people in a scene, it is unclear how a sequential model can capture their relations. Note that we compared the proposed method with all of the four methods above in our experiments on PIPA. As we shall see in Section 5, our method consistently outperforms them under all evaluation policies.

Person Re-identification. Another relevant task is person re-identification [18, 28, 14], which is to match pedestrian images from different cameras, within a relatively short period. This task is essentially different, where visual cues are likely to remain consistent and social context is weak. General person recognition, instead, requires recognizing across events, where visual cues may vary significantly and thus the social context can be crucial.

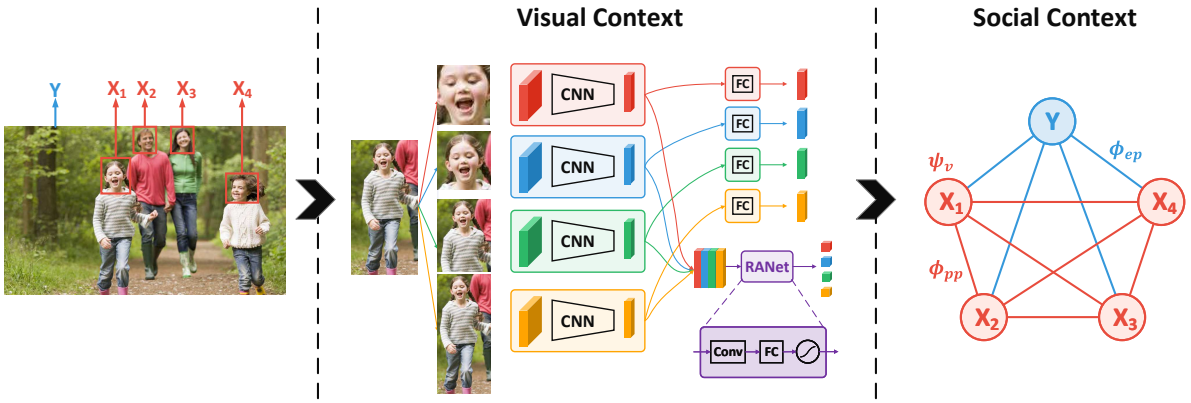

Figure 2: Our whole framework. We propose a Region Attention Network to get instance-dependent weights for visual context fusion and develop a unified formulation that join social context learning, including event-person relations and person-person relations, with person recognition.

3. Methodology

In general, the task of person recognition in a photo collection can be formalized as follows. Consider a collection of photos $I_1, \ldots, I_M$, where I_m contains N_m *person instances*. All person instances are divided into two disjoint subsets, the *gallery set*, in which the instances are all labeled (*i.e*. their identities are provided). and the *query set*, in which the instances are unlabeled. The task is to predict the identities for those instances in the query set.

As discussed, person recognition in an unconstrained setting is very challenging. In this work, we leverage two kinds of contexts, the *visual context* and the *social context*. Particularly, the *visual context* involves different regions of the person instances, including *face*, *head*, *upper body*, and *whole body*. These regions often convey complementary cues for visual matching. The *social context*, instead, captures the social behavior of people, *e.g*. the events they usually attend or the people whom they often stay with. It is worth noting that unlike visual cues, the social relations are reflected collectively by multiple photos and can not be reliably derived from a single photo in isolation.

3.1. Framework Overview

We devise a framework that incorporates both the visual context and the social context for person recognition. As shown in Figure 2, the framework recognizes the identities for all instances in the query set *jointly*, in two stages.

1. **Visual matching.** This stage computes a *matching score* for each pair of instances. For this, a *Region Attention Network* is learned to adaptively combine the visual cues from different regions, with instance-dependent weights.
2. **Joint optimization.** This is the key stage of our framework. In this stage, the *social context model*, which captures both *event-people* and *people-people* relations, will be jointly learned along with the identification of query instances, by solving a unified optimization problem.

3.2. Visual Matching

We combine the visual observations from different regions to compute the *matching score* between two instances. Particularly, we consider four regions: *face*, *head*, *upper body*, and *whole body*. These regions are often complementary to each other. This strategy has also been shown to be effective in previous work [27, 11, 13, 15].

However, existing methods mostly adopt uniform weighting schemes, where each region is assigned a fixed weight that is shared by all instances. Let $s(i, j)$ be the overall matching score between instances i and j, and $s^r(i, j)$ be the matching score based on the r-th region. Then, such a scheme can generally be expressed as

$$s(i, j) = \sum_{r=1}^{R} w^r s^r(i, j), \tag{1}$$

where R is the number of distinct regions. The weights $\{w^r\}$ are often decided by empirical rules [13] or optimized over a validation set [27, 11, 15].

The uniform schemes as described above are limited in two aspects, as illustrated in Figure 3. (1) Some regions may be invisible for an instance. The missing of such regions may be due to various reasons, *e.g*. limited scope of the camera and occlusion. With a uniform scheme, one would be forced to locate the missing parts with rigid rules and compute matching scores for them, which often leads to inaccurate results. (2) The contributions of different parts vary significantly across instances. For example, the facial features play a key role when the frontal face is visible. However, when we can only see one's back, we will have to resort to the clothing in the body region. A uniform scheme can not effectively handle such variations.

We propose to tackle this problem using *instance-dependent* weights, where the weight of a region is determined by whether it is visible and how much it contributes. Specifically, given an instance, we get the bounding boxes

(a) (b)

Figure 3: Here are examples to show the necessity of instance-dependent weights for recognition. (a) shows that some of the regions may be out of scope, like "body" of the first instance and "face" of the second instance. (b) shows that the contributions of parts vary across instances. The contribution of "face" for the first man is obviously more significant than that for the second one.

of the regions by either the annotation of the dataset or detectors, then resize each region to a standard size, and apply region-specific CNNs to extract their features.

To combine these features adaptively, we devise a *Region Attention Network (RANet)* as shown in Figure 2 to compute the fusion weights. Here, the RANet is a small neural network that takes the stacked features from all regions as input, feeds them through a convolution layer, a fully-connected layer, and a sigmoid layer, and finally yields four positive coefficients as the region weights. Then the combined *matching score* is given by

$$s(i,j) = \sum_{r=1}^{R} w_i^r w_j^r s^r(i,j). \tag{2}$$

Here, w_i^r and w_j^r are instance-dependent weights of the r-th region respectively for instances i and j; $s^r(i,j)$ is the cosine similarity between the corresponding features. We use the product $w_i^r w_j^r$ to weight a region score, which reflects the rationale that a region type should be active only when it is clearly visible in both instances. All region-specific CNNs together with the RANet are jointly trained in an *end-to-end manner* with the cross-entropy loss.

3.3. Unified Formulation with Social Context

In an unconstrained environment, certain instances are very difficult to recognize purely based on their appearance. For such cases, one can leverage the social context to help. Specifically, the *social context* refers to a set of social relations. We consider two types of social relations:

1. **Event-person relations.** Generally, an *event* can be conceptually understood as an activity that occurs at a certain place with a certain set of attendants [25]. Over a large photo collection, an event may involve just a small fraction of the people. Hence, an event can provide a strong prior for recognition if we can infer the event that a photo is capturing, as illustrated by the photo in Figure 4a.
2. **Person-person relations.** It is commonly observed that certain groups of people often stay together. For such groups, the presence of a person may indicate the presence of others in the group, as illustrated by the photo in Figure 4b. Note that *person-person relations* are complementary to the *event-person relations*, as such relations do not require the match of scene features.

(a) (b)

Figure 4: Some instances (with red boxes) are very difficult to recognize purely by their appearance. But we can leverage the social context to help. (a) If we know that the photo belongs to the event – "People cry for help after Titanic sunk", then the probability to recognize them as *Leonardo* and *Kate* will become higher. (b) Sometimes it's easy for us to get other instances' identities (with green boxes) in the same photo. So we can infer the red ones' identities by the person-person relation.

Unified Objective. Taking both the visual context and the social context into account, we can formulate a unified optimization problem where person identifications are coupled with event association and contextual relation learning. The objective function of this problem is given below:

$$J(\mathbf{X},\mathbf{Y};\widetilde{\mathbf{F}},\mathbf{P},\mathbf{Q} \mid \mathbf{S},\mathbf{F}) = \psi_v(\mathbf{X}|\mathbf{S}) + \alpha \cdot \phi_{ep}(\mathbf{Y},\mathbf{X};\widetilde{\mathbf{F}},\mathbf{P}|\mathbf{F}) + \beta \cdot \phi_{pp}(\mathbf{X};\mathbf{Q}). \tag{3}$$

The notations involved here are described below:

1. $\mathbf{X} \in \mathbb{R}^{L\times N}$ captures all people identities, where L is the number of distinct identities and N is the number of person instances. Each column of $\mathbf{X}$, denoted by $\mathbf{x}_j$, is the identity indicator vector for the j-th instance.
2. $\mathbf{Y} \in \mathbb{R}^{K\times M}$ is the association matrix between photos and events, where K is the number of events and M is the number of photos. The i-th column of $\mathbf{Y}$, denoted by $\mathbf{y}_i$, is the event indicator vector for the i-th photo. In particular, $y_i^k \triangleq \mathbf{Y}(i,k) \in \{0,1\}$ indicates whether the i-th photo is from the k-th event.
3. $\mathbf{S} \in \mathbb{R}^{N\times N}$ denotes the matrix of pairwise visual matching scores, derived by Eq.(2).
4. $\mathbf{F} \in \mathbb{R}^{D_f\times M}$ comprises the scene features of all photos, where the i-th column $\mathbf{f}_i$ is a D_f-dimensional feature for

the i-th photo. In this work, we obtained $\mathbf{f}_i$ for each photo with a CNN pretrained on Places [29].

5. $\widetilde{\mathbf{F}} \in \mathbb{R}^{D_f \times K}$ and $\mathbf{P} \in \mathbb{R}^{L \times K}$ are the parameters associated with the *events*. In particular, the k-th column of $\widetilde{\mathbf{F}}$, denoted by $\tilde{\mathbf{f}}_k$, is the prototype scene feature for the k-th event; and the k-th column of $\mathbf{P}$, denoted by $\mathbf{p}_k$, is a probability vector that captures the person identity distribution of the k-th event.
6. $\mathbf{Q} \in \mathbb{R}^{L \times L}$ is a matrix that captures the person-person relations. High value of $\mathbf{Q}(l, l')$ indicates that the identities l and l' are likely to co-occur in the same photo.

Among these quantities, the matching scores $\mathbf{S}$ and the scene features $\mathbf{F}$ are provided in the visual analysis stage, while others are jointly solved by optimizing this problem.

Potential Terms. The joint objective in Eq.(3) comprises three potential terms, which are introduced below.

1. **Visual consistency:** $\psi_v(\mathbf{X}|\mathbf{S})$ encourages the consistency between person identities and the visual matching scores, and is formulated as:

$$\psi_v(\mathbf{X}|\mathbf{S}) = \sum_{j=1}^{N} \mathbf{s}_j^T \mathbf{x}_j, \quad \text{with } \mathbf{s}_j(l) = \max_{j' \in \mathcal{G}_l} s(j, j'), \tag{4}$$

where $\mathbf{s}_j \in \mathbb{R}^L$, $\mathcal{G}_l$ refers to the set of gallery instances with label l, and thus $\mathbf{s}_j(l)$ is the maximum matching score of the l-th instance to those in $\mathcal{G}_l$.

2. **Event consistency:** $\phi_{ep}(\mathbf{Y}, \mathbf{X}; \widetilde{\mathbf{F}}, \mathbf{P}|\mathbf{F})$ concerns about the assignments of photos to events, and encourages them to be consistent in both scenes and attendants. This term is formulated as:

$$\phi_{ep}(\mathbf{Y}, \mathbf{X}; \widetilde{\mathbf{F}}, \mathbf{P}|\mathbf{F}) = \sum_{i=1}^{M} \sum_{k=1}^{K} a_i^k y_i^k$$
$$\text{with } a_i^k = \sum_{j \in \mathcal{I}_i} \log(\mathbf{p}_k)^T \mathbf{x}_j - \|\mathbf{f}_i - \tilde{\mathbf{f}}_k\|^2, \tag{5}$$

where, $\mathcal{I}_i$ is the set of instance indexes for the i-th photo. For each assignment of the k-th event to the i-th photo, this formula evaluates (a) whether the people in the photo are frequent attendants of the event (by $\mathbf{p}_k^T \mathbf{x}_j$) and (b) whether the photo's scene feature match the event's scene prototype (by $-\|\mathbf{f}_i - \tilde{\mathbf{f}}_k\|^2$).

3. **People cooccurrence:** $\phi_{pp}(\mathbf{X}; \mathbf{Q})$ takes into account the person-person relations, *i.e.* which identities tend to coexist in a photo. This term is formulated as:

$$\phi_{pp}(\mathbf{X}; \mathbf{Q}) = \sum_{i=1}^{M} \sum_{j \in \mathcal{I}_i} \sum_{j' \in \mathcal{I}_i : j' \neq j} \mathbf{x}_j^T \mathbf{Q} \mathbf{x}_{j'}. \tag{6}$$

This formula considers all pairs of distinct instances in each image, and sums up their person-person relation value. In particular, if $\mathbf{x}_j$ indicates label l and $\mathbf{x}_{j'}$ indicates label l', then $\mathbf{x}_j^T \mathbf{Q} \mathbf{x}_{j'} = Q(l, l')$.

To balance the contributions of these potential terms, we introduce two coefficients α and β, which are decided via cross validation.

3.4. Joint Estimation and Inference

We solve this problem using coordinate ascent. Specifically, our algorithm alternates between the updates of (1) people identities ($\mathbf{X}$), (2) assignments of events to photos ($\mathbf{Y}$), and (3) social relation parameters ($\widetilde{\mathbf{F}}$, $\mathbf{P}$, and $\mathbf{Q}$). These steps are presented below.

Person Identification. Given both the event assignments $\mathbf{X}$ and the social context parameters, the inference of people identities can be done separately for each photo, by maximizing the sub-objective as:

$$\sum_{j \in \mathcal{I}_i} \mathbf{s}_j^T \mathbf{x}_j + \alpha \sum_{j \in \mathcal{I}_i} \log(\mathbf{p}_{\hat{y}_i})^T \mathbf{x}_j + \beta \sum_{j \in \mathcal{I}_i} \sum_{j' \in \mathcal{I}_i : j' \neq j} \mathbf{x}_j^T \mathbf{Q} \mathbf{x}_j, \tag{7}$$

where $\hat{y}_i$ indicates the assigned event. Note that $\mathbf{x}_j$ here is constrained to be an indicator vector, *i.e.* only one of its entry is set to one, while others are zeros. When there is only one person instance, its identity can be readily derived as

$$\hat{x}_j = \operatorname*{argmax}_l \ \mathbf{s}_j(l) + \alpha \mathbf{p}_{\hat{y}_i}(l). \tag{8}$$

When there are two or more instances, we treat it as an MRF over their identities and solve them *jointly* using the max-product algorithm.

Event Assignment. We found that the granularity of the events has significant impact on the identification performance. If we group the photos into coarse-grained events such that each event may contain many people or scenes, then the event-person relations may not be able to provide a strong prior. However, for fine-grained events, it would be difficult to estimate their parameters reliably. Hence, it is advisable to seek a good balance.

In this work, we use two parameters ν_{min} and ν_{max} to control the granularity, and require that the number of photos assigned to an event fall in the range $[\nu_{min}, \nu_{max}]$. Then the problem of event assignment can be written as

$$\max \quad \sum_{i=1}^{M} \sum_{k=1}^{K} a_i^k y_i^k, \tag{9}$$

$$s.t. \quad \sum_{k=1}^{K} y_i^k \leq 1, \ \ \forall i = 1, \ldots, M, \tag{10}$$

$$\nu_{min} \leq \sum_{i=1}^{M} y_i^k \leq \nu_{max}, \ \ \forall k = 1, \ldots, K, . \tag{11}$$

Here, a_i^k Eq.(9) follows Eq.(5). Eq.(10) enforces the constraint that each photo is associated to at most one event; Eq.(11) enforces the granularity constraint above. This is a linear programming problem, and can be readily solved by an LP solver. Also, the optima is guaranteed to be integral.

Context Learning. As mentioned, the social context model, which is associated with three parameters $\widetilde{\mathbf{F}}$, $\mathbf{P}$, and $\mathbf{Q}$, are learned along the inference of people identities and event assignments. Given $\mathbf{X}$ and $\mathbf{Y}$, we can easily derive the optimal solution of the parameters listed above.

Specifically, for the scene prototypes in $\widetilde{\mathbf{F}}$, we have the optimal $\tilde{\mathbf{f}}_k$ (the k-th column of $\widetilde{\mathbf{F}}$) given by

$$\tilde{\mathbf{f}}_k = \operatorname*{argmin}_{\mathbf{f}} \sum_{i \in \mathcal{E}_k} \|\mathbf{f}_i - \mathbf{f}\|^2 = \frac{1}{|\mathcal{E}_k|} \sum_{i \in \mathcal{E}_k} \mathbf{f}_i, \tag{12}$$

where $\mathcal{E}_k = \{i \mid a_i^k = 1\}$ refers to the set of photos that are assigned to the k-th event. For the identity distributions $\mathbf{P}$, we have the optimal $\mathbf{p}_k$ given (the k-th column of $\mathbf{P}$) given by

$$\mathbf{p}_k = \operatorname*{argmax}_{\mathbf{p} \in \mathbb{S}^L} \sum_{i \in \mathcal{E}_k} \sum_{j \in \mathcal{I}_i} \log(\mathbf{p})^T \mathbf{x}_j, \tag{13}$$

where $\mathbb{S}$ is the L-dimensional probability simplex and $\mathbf{p} \in \mathbb{S}^L$ enforces that $\mathbf{p}$ be a probability vector. This is a maximum likelihood estimation over all people who attend the k-th event, and its optimal solution is

$$\mathbf{p}_k = \left(\sum_{i \in \mathcal{E}_k} |\mathcal{I}_i| \right)^{-1} \sum_{i \in \mathcal{E}_k} \sum_{j \in \mathcal{I}_i} \mathbf{x}_j. \tag{14}$$

With both $\tilde{\mathbf{f}}_k$ and $\mathbf{p}_k$, we can characterize an event with both scene features and attendants. For person-person relations, the optimal $\mathbf{Q}$ can be obtained by maximizing Eq.(6) with $\mathbf{X}$. Here, we enforce a constraint that $\mathbf{Q}$ is normalized, *i.e.* $\|\mathbf{Q}\|_F = 1$. Then, the optimal solution is

$$\mathbf{Q} = \mathbf{Q}' / \|\mathbf{Q}'\|_F, \text{ with } \mathbf{Q}' = \sum_{i=1}^{M} \sum_{j \in \mathcal{I}_i} \sum_{j' \in \mathcal{I}_i \setminus j} \mathbf{x}_j \mathbf{x}_{j'}^T. \tag{15}$$

It is worth emphasizing that all sub-tasks presented above are steps in the coordinate ascent procedure to optimize the unified objective in Eq.(3). We run these steps iteratively, and it usually takes about 5 iterations to converge.

4. New Dataset: Cast In Movies

In addition to photo albums, the proposed method can also be applied to other settings with strong contexts, *e.g.* recognizing actors in movies. To test our method in such settings, we constructed the *Cast In Movies (CIM)* dataset from 192 movies. We divide each movie into shots using an existing technique [3], sample one frame from each shot, and retain all those that contain persons. This procedure results in a dataset with $72,875$ photos.

Figure 5: Examples from *CIM*. Here are some instances of *Kate Winslet* and *Leonardo DiCaprio* from different movies in *CIM*.

Dataset	PIPA[27]	CIM (ours)
Images	37,107	72,875
Indentities	2,356	1,218
Instances	63,188	150,522
Instances (except "others")	51,751	77,598
Avg/identity	26.82	63.70

Table 1: Statistics of *CIM* compared with *PIPA*.

We manually annotated all person instances in these photos with bounding boxes for the body locations. We also annotated the identities of those instances that correspond to the 1218 main actors[1]. In this way, we obtained $77,598$ instances with known identities, while other instances are labeled as "others". Figure 5 shows some examples of our dataset. Table 1 shows the statistics of *CIM* in comparison with *PIPA* [27]. To our best knowledge, *CIM* is the first large-scale dataset for person recognition in movies.

5. Experiments

We tested our method on both *PIPA* [27], a dataset widely used for person recognition, and *CIM*, our new dataset presented above.

5.1. Experiment Setup

Evaluation protocols The *PIPA* dataset is partitioned into three disjoint sets: training, validation and test sets. The test set is further split into two subsets, one as the gallery set and the other as the query set. To evaluate a method's performance, we first use it to predict the identities of the instances in the query set and compute the prediction accuracy. Then, we switch the gallery and the query set and compute the accuracy in the same way. The average of both

[1]The main actors are chosen according to two criteria: 1) ranked top 10 in the cast list of IMDb for the corresponding movie, and 2) occur for more than 5 times in our sampled frames.

Split	Existing Methods on PIPA				Ours			
	PIPER [27]	Naeil [11]	RNN [15]	MLC [13]	Baseline	+RANet	+RANet+P	+RANet+P+E
original	83.05%	86.78%	84.93%	88.20%	82.79%	87.33%	88.06%	**89.73%**
album	-	78.72%	78.25%	83.02%	75.24%	82.59%	83.21%	**85.33%**
time	-	69.29%	66.43%	77.04%	66.55%	76.52%	77.64%	**80.42%**
day	-	46.61%	43.73%	59.77%	47.09%	65.49%	65.91%	**67.16%**

Table 2: Comparison of the accuracies of different methods on PIPA, under different splits of the query and gallery sets.

accuracies will be reported as the performance metric.

There are four different ways to split the test set, namely *original*, *album*, *time*, and *day*, for evaluating an algorithm under different application scenarios. In the *original* setup of *PIPA* [27], a query instance may have similar instances in the gallery. [11] defines the other three splits, which are more challenging. For example, *day* split requires that the query and the gallery instances of the same subject need to have notable differences in visual appearance. For *CIM*, we follow the rule in [27], dividing it into three disjoint subsets respectively for training, validation, and testing. Also, the test set is randomly split into a gallery set and a query set.

Implementation Details We use four regions of each instance: *face*, *head*, *upperbody*, and *body*. *PIPA* provides the head locations, while *CIM* provides the locations of whole body. Other regions are obtained by simple geometric rules based on the results from a face detector [26] and OpenPose [5]. Note that we only keep those bounding boxes that lie mostly within the photo. For those regions that are largely invisible, we simply use a black image to represent their appearance. We will see that our RANet can learn to assign such regions with negligible weights in our experiments. We adopt ResNet-50 [9] as our base model and train the feature extractor with OIM loss [24]. We chose design parameters empirically on the validation set. The coefficients α and β in Eq.(3) are set to 0.05 and 0.01. The number of events K is set to 300 for both *PIPA* and *CIM*.

5.2. Results on PIPA

We set up a *baseline* for comparison, which relies on a *uniformly* weighted combination of visual cues from all regions, where the weights are optimized by grid search. We tested three configurations of the proposed methods: **(1) +RANet:** This config combines region-specific scores following Eq.(2), using the instance-dependent weights from the Region Attention Network (see Sec. 3.2). **(2) +RANet+P:** In addition to the visual matching score RANet, it also uses the person-person relations in joint inference. **(3) +RANet+P+E:** This is the full configuration of our framework, which takes visual matching, person-person relations, and person-event relations into account. Moreover, we also compared with four previous methods: PIPER [27], Naeil [11], RNN [15], and MLC [13].

Table 2 shows the results under all the four splits, from which we can see that: (1) RANet, with adaptive weights, significantly outperforms the baseline with uniform weights. On the most challenging *day split*, it remarkably raises the accuracy from 47.09% to 65.49%. (2) With our proposed joint inference method, the use of social contexts leads to consistent improvement across all splits. (3) Our method also outperforms all previous works, including the state-of-the-art MLC [13], by a considerable margin on all splits. Particularly, the performance gain is especially remarkable on the most challenging *day split* (67.16% with ours vs. 59.77% with MLC).

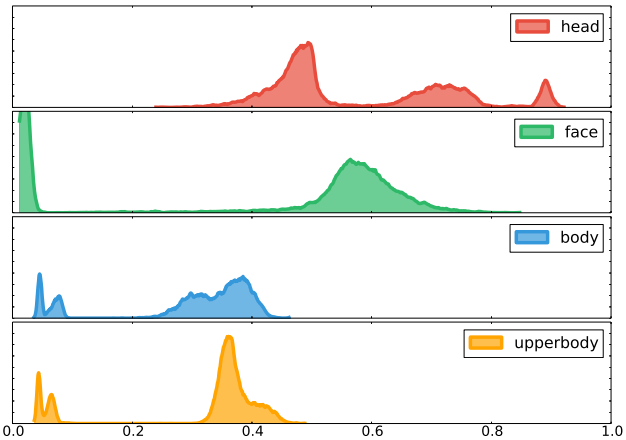

Figure 6: Weight distributions of different regions on *PIPA*.

Figure 7: Example events automatically discovered from *PIPA* by joint inference. Images in each row belong to the same "event".

Analysis on RANet Table 2 clearly shows the effectiveness of the Region Attention Network (RANet). To learn more about the RANet, we study the distributions of region-

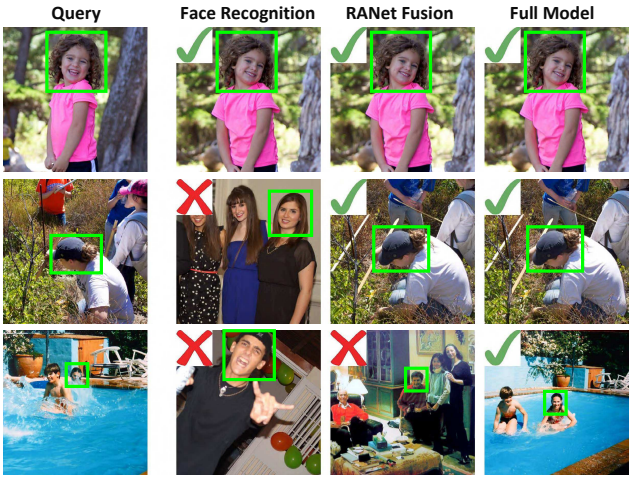

Figure 8: Example photos and recognition results. For each photo, the mark at the top left corner indicates whether the corresponding method predicts correctly for the highlighted instance.

specific weights on the test set of *PIPA*, and show them in Figure 6. This study reveals some interesting characteristics of RANet: (1) For each of the following region types, *face*, *body*, and *upper body*, there exist a fraction of instances with very low weights because the particular regions of them are out of scope. (2) The average weight of *faces* is the highest among all region types. This is not surprising, as *faces* are often the strongest indicators of identities when they are visible. (3) A small portion of instances have very high weights assigned to the *head* regions, because for such instances all other parts are largely invisible.

Analysis on Event Events are automatically discovered during joint inference and they play an important role in person recognition. Figure 7 shows some example events with their associated photos. We can see that our method can discover events in a reasonable way, and they can provide strong prior in a considerable portion of cases. More examples will be provided in the supplemental materials.

Case Study Figure 8 shows some photos and associated recognition results. We can see that 1) For an instance with frontal and clear face (1st row), all methods predict correctly. 2) When the face is not clearly visible (2nd row), our method with RANet can still correctly recognize the person with other visual cues, *e.g.* the head or upper body. 3) For the most challenging case where all visual cues fail (3rd row), our full model can still make a correct prediction by exploiting the social context.

5.3. Results on CIM

Table 3 shows the results on *CIM*, which again demonstrates the effectiveness of our approach. Only with RANet, it already outperforms the baseline (with uniform weighting) by over 4%. The whole framework, with social context taken into account, further improves the accuracy (6.3% higher than the baseline). Recognition results on example photos will be provided in the supplemental materials. It is also worth noting that the accuracies we obtained on *CIM* are generally lower than those on *PIPA*, implying that this is a more challenging dataset which can help to drive the progress on this task.

Baseline	+RANet	+RANet+P	+RANet+P+E
68.12%	71.93%	72.56%	**74.40%**

Table 3: Performance on *CIM*

5.4. Computational Cost Analysis

Our method obtains the improvement on recognition accuracy with substantially lower computing cost compared to some previous works. Note that PIPER [27] uses more than 100 deep CNNs and Naeil [11] uses 17 deep CNNs for feature extraction. While our model uses only 4 CNNs and a fusion module whose computing cost is negligible. Although MLC [13] uses just 3 deep CNNs for feature extraction, it additionally requires to train thousands of group-specific SVMs, which is also a costly procedure.

Compared with the CNN-based feature extraction components, the cost of the joint estimation and inference procedure is insignificant. Particularly, it takes about *30 minutes* to perform inference over the whole test set of *PIPA*, with one single 2.2 GHz CPU, while the feature extractors take over 40 hours to detect regions and compute CNN features for all test photos, with a Titan X GPU.

6. Conclusions

We presented a new framework for person recognition, which integrates a Region Attention Network to adaptively combine visual cues and a model that unifies person identification and context learning in joint inference. We conducted experiments on both PIPA and a new dataset CIM constructed from movies. On PIPA, our method consistently outperformed previous state-of-the-art methods by a notable margin, under all splits. On CIM, the new components developed in this work also demonstrated strong effectiveness in raising the recognition accuracy. Both quantitative and qualitative studies showed that adaptive combination of visual cues is important in a generic context and that the social context often conveys useful information especially when the visual appearance causes ambiguities.

7. Acknowledgement

This work is partially supported by the Big Data Collaboration Research grant from SenseTime Group (CUHK Agreement No. TS1610626), the General Research Fund (GRF) of Hong Kong (No. 14236516). We are grateful to Shuang Li and Hongsheng Li for helpful discussions.

References

[1] R. Adolphs. Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience*, 4(3):165–178, 2003. 1

[2] D. Anguelov, K.-c. Lee, S. B. 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. IEEE, 2007. 1, 2

[3] E. Apostolidis and V. Mezaris. Fast shot segmentation combining global and local visual descriptors. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 6583–6587. IEEE, 2014. 6

[4] M. Brenner and E. Izquierdo. Joint people recognition across photo collections using sparse markov random fields. In *International Conference on Multimedia Modeling*, pages 340–352. Springer, 2014. 1

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016. 7

[6] M. M. Chun and Y. Jiang. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive psychology*, 36(1):28–71, 1998. 1

[7] M. Davis, M. Smith, J. Canny, N. Good, S. King, and R. Janakiraman. Towards context-aware face recognition. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 483–486. ACM, 2005. 1, 2

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

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[10] 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, Technical Report 07-49, University of Massachusetts, Amherst, 2007. 1

[11] S. Joon Oh, R. Benenson, M. Fritz, and B. Schiele. Person recognition in personal photo collections. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3862–3870, 2015. 1, 2, 3, 7, 8

[12] K. S. LaBar and R. Cabeza. Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, 7(1):54–64, 2006. 1

[13] H. Li, J. Brandt, Z. Lin, X. Shen, and G. Hua. A multi-level contextual model for person recognition in photo albums. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1305, 2016. 1, 2, 3, 7, 8

[14] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 2

[15] Y. Li, G. Lin, B. Zhuang, L. Liu, C. Shen, and A. v. d. Hengel. Sequential person recognition in photo albums with a recurrent network. *arXiv preprint arXiv:1611.09967*, 2016. 1, 2, 3, 7

[16] D. Lin, A. Kapoor, G. Hua, and S. Baker. Joint people, event, and location recognition in personal photo collections using cross-domain context. *Computer Vision–ECCV 2010*, pages 243–256, 2010. 2

[17] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310*, 2015. 1

[18] B. J. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010. 2

[19] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 1

[20] Y. Song and T. Leung. Context-aided human recognition–clustering. *Computer Vision–ECCV 2006*, pages 382–395, 2006. 1, 2

[21] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015. 1

[22] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 1

[23] G. Wang, A. Gallagher, J. Luo, and D. Forsyth. Seeing people in social context: Recognizing people and social relationships. *Computer Vision–ECCV 2010*, pages 169–182, 2010. 1

[24] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *Proc. CVPR*, 2017. 7

[25] Y. Xiong, K. Zhu, D. Lin, and X. Tang. Recognize complex events from static images by fusing deep channels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1600–1609, 2015. 4

[26] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 7

[27] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4804–4813, 2015. 1, 2, 3, 6, 7, 8

[28] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013. 2

[29] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 5