Face Discovery with Social Context

Yong Jae Lee
https://webspace.utexas.edu/yl3663/~ylee/
University of Texas at Austin
Austin, TX, USA
Kristen Grauman
http://www.cs.utexas.edu/~grauman/

Abstract

We present an approach to discover novel faces in untagged photo collections by leveraging the "social context" of co-occurring people. Our idea exploits the social nature of consumer photos, in which people of the same clique (family, team, class, friends) often appear together. Initially, the system trains detectors for any individuals with tagged instances in the collection. Then, for each untagged image, it isolates any unfamiliar faces. Among those, it discovers novel face clusters by leveraging both their appearance, as well as descriptors encoding the (predicted) familiar faces with which the unfamiliar faces co-occur. The resulting discovered people can then be presented to a user for name-tagging, thereby efficiently propagating manually provided labels. Our experiments with real consumer photo collections demonstrate that the system outperforms baseline approaches that either lack any social context model, or rely solely on the appearance of co-occurring faces. Furthermore, we show it can successfully use the discovered models it forms to auto-tag unseen faces in a new collection.

1 Introduction

Photos are great for capturing monumental moments in life, such as birthdays, graduations, weddings; for capturing breathtaking sights; or for capturing artistic images of everyday life. A common theme of photos is that, in most cases, *people* are their main subjects. Photos can rekindle fond memories and even provide specific answers to questions such as: *What did my kindergarten teacher look like? Who was sitting next to me at my 16th birthday party?* Naturally, an automated method for organizing photos according to who is in them would be invaluable for the modern day digital-camera user who possesses large collections of photos.

Face detection algorithms can now provide highly accurate results in realistic images, and their use in conjunction with popular photo-sharing sites is to the point where "auto-tagging" functions are increasingly common in real commercial applications. Typically, the assumption is that a user will directly train the system about the faces of interest in his/her collection by providing tagged exemplars.

Vision researchers have explored a variety of innovative approaches to use tagged data to learn face models and perform recognition [6, 7, 16, 17, 24, 25]. The usual pipeline is as follows: (1) The user supplies name tags for a few images in the photo collection and trains classifiers that can recognize each labeled person; (2) The system detects faces in the remaining unlabeled images; and (3) The system applies the trained classifiers to tag those faces with candidate names. While face recognition methods perform quite well in more

BMVC 2011 http://dx.doi.org/10.5244/C.25.36

Figure 1: Main idea of our approach to unsupervised face discovery in personal photo collections. For any unfamiliar face not recognized by the system (in dotted green), we use the co-occurrence cues from familiar faces nearby (in solid yellow) to produce more reliable groups. In this example, an appearance-based grouping method that clusters the unfamiliar faces would likely fail to recognize the many instances of the boy, given their variability. In contrast, by also representing the *social context* of people appearing near each unfamiliar face, our approach computes more reliable clusters. Having discovered a novel face, the system would present the images to a user for name-tagging.

controlled environments, they become less reliable for natural consumer photo collections, where faces run the gamut in terms of pose (sitting, playing, dancing), occlusions (hats, sunglasses), and lighting variations (indoor, outdoor, night, day) [26]. Recent work shows that *context cues* such as clothing, timestamps, or nearby text data, are critical to overcoming such variations [1, 2, 6, 7, 12, 16, 17, 22, 25].

Nevertheless, a limitation of the above pipeline is that the user must teach the system about each face (person) of interest. While the system can automatically annotate new instances of *familiar* faces, it first requires a human to manually label samples of those faces. This means that the system's performance is bounded by the quality and scope of the labeled instances a user spends time providing, which is problematic once a photo collection grows to include new friends (a student goes to college), big events with many repeating new faces (a wedding on the in-laws' side), or when merging collections between users. For large photo collections with tens to hundreds of people, the user's role can become laborious.

We present an approach for face discovery that alleviates the costs of manual intervention, and allows users' collections and tagging functionality to evolve more fluidly. The goal is to perform unsupervised clustering on faces detected in the images, in order to come up with a batch of photos likely of the same individual, so that the user can efficiently tag or prune them with minimal effort. In contrast to previous face clustering algorithms (e.g., [2, 16, 19]), we propose to expand the representation of the detected faces to include not just their appearance, but also their *social context*. Specifically, the main idea is to use cues from co-occurring people in the same image in order to produce more reliable groups.

Why do co-occurrence cues help? New (yet unlearned) faces in a collection appear with some strong social context, as users' photos tend to dwell within different cliques of people: families, friends, co-workers, etc. This means the context of "familiar people" can both help disambiguate people with similar appearance, and help the system realize that instances of faces in different poses or expression are actually of the same person (see Figure 1).

We design a context descriptor to capture the predictions of previously trained face models, and show that this "face-level" cue is more reliable than simply using the appearance of nearby faces as context. A system using the proposed approach frees the user from manually identifying each new face. Instead, it discovers novel recurring faces—and, critically, discovers them more accurately by modeling the social context surrounding them. It can then present its discoveries (a cluster of photos) to the user, and he/she can confirm with tags (or reject). While related context cues have been explored to a limited extent for traditional supervised learning pipelines [6, 17, 22, 25], we are the first to consider unsupervised face discovery using social context. We demonstrate our approach mining for novel faces on a dataset drawn from multiple domains and two large personal photo collections that exhibit natural social context.

2 Related Work

Space does not permit a thorough review of face detection and recognition algorithms [4, 21, 26]; our contribution relates to managing photo collections of faces, and advances made in either of the above should only enhance our system's results.

Several face recognition systems intended for consumer photos have demonstrated the value of using co-occurrence statistics between people to improve predictions. However, in previous work the co-occurrence cues are learned from labeled examples and applied to help name familiar (trained) faces, e.g., [6, 17, 25], whereas we aim to discover new faces in the context of familiar ones. Since image-level tags for images with multiple faces are inherently ambiguous, researchers have explored ways to efficiently recover the correspondence between people's names and the face windows present [2, 6, 24]. Tracking and movie scripts also offer interesting ways to resolve ambiguities and collect face datasets [5, 15]. Context cues from familiar social relationships (e.g., mother-child, husband-wife) can improve face recognition accuracy in a weakly-supervised setting [22]. While the social relationships are manually provided in [22], our method automatically discovers the social context in an unsupervised manner.

Methods that tackle the face clustering problem have shown that clothing, timestamps [16], and captions [2, 12] are useful context, and that the most evident clusters can aid in interactive labeling [19]. We are the first to consider using the context of other faces to aid in discovering new faces in a photo collection.

In the object recognition community, much research has been done to exploit context between objects and the scenes that contain them (see [9] for a survey). Our approach has parallels with recent techniques that show how to discover useful context information in supervised or semi-supervised settings [10, 11, 14, 20]. The context cues for a set of specified generic objects (cats, trees, etc.) is learned directly from unseen test data in [10], and extracted iteratively for a fixed set of categories in [20]. Given a set of related scenes, one can also analyze spatial connections to discover semantically related objects [14].

Of the above work in object recognition, most closely related is our context-aware discovery method [11], which uses familiar objects surrounding a region of interest in an image to build a more reliable context descriptor. We design a social context descriptor that is directly inspired by the "object-graph", in that it records class posteriors rather than raw appearance. We follow a similar pipeline to our work in [11] for category discovery, but adapt it specifically for the face discovery setting, and show that it captures a very relevant form of social context that allows better unsupervised clustering in this domain. Given the central importance of face tagging for everyday consumer photo applications, this setting is particularly interesting to consider.

3 Approach

Our goal is to discover novel faces from untagged image collections by exploiting the social nature of consumer photographs. In particular, we aim to use the co-occurrence information from *familiar* people to better discover faces of new people.[1]

Given a pool of unlabeled photos, we first detect any faces in each image. We then identify novel faces that do not resemble any person for which we have trained models (Section 3.2). After isolating the unfamiliar faces, we form new people "categories" by grouping faces that have similar appearance *and* similar social networks (Section 3.3).

[1] We use "(un)familiar" and "(un)known", interchangeably.

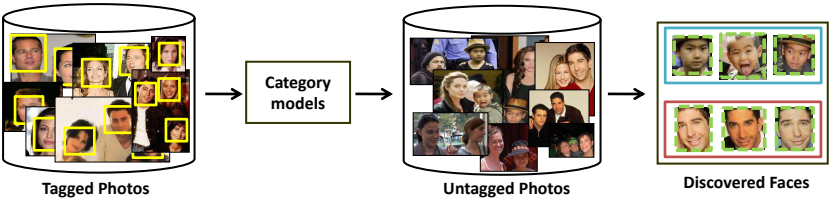


Figure 2: System overview. Given a photo collection with tagged faces, we train models for each person. Given a novel set of face images (that do not have any tags), we detect instances of familiar people in each image, and use their context to discover novel faces.

See Figure 2 for an overview of our system. In the following we describe the main steps.

3.1 Learning Models for Tagged Faces

For each face region r found with a face detector, we extract texture features to serve as the appearance descriptor $A(r)$. We use pyramid of HOG (pHOG) [3] or Local Ternary Patterns (LTP) [18]. We train SVM classifiers for N initial people, $\{c_1,\ldots,c_N\}$, for whom we have tagged face images. These classifiers will allow us to identify the instances of each initial familiar person in novel images. We will use those predictions to describe the social context for each *unfamiliar* face, as we describe in more detail in Section 3.3.

3.2 Identifying Unfamiliar Faces

For any unlabeled photo, we would like to detect the people in it, and determine whether any of them resembles a *familiar person*. Doing so will allow us to isolate the unknown faces, and to build social context descriptors that portray the co-occurring familiar people.

For all unlabeled images, we run a face detector [21] to extract candidate faces. To compute the known/unknown decision for a face region r, we apply the N trained classifiers from Section 3.1 to the face to obtain its class membership posteriors $P(c_i|r)$, for $i=1,\ldots,N$, where c_i denotes the i-th person class. Faces that resemble a known person c_i will produce a high value for $P(c_i|r)$, and low values for $P(c_j|r)$, $\forall j \neq i$. Faces that do not resemble any familiar person will have more evenly distributed posteriors.

Thus, to distinguish which faces should be considered to be unknown, we compute the entropy: $E(r) = -\sum_{i=1}^{N} P(c_i|r)\log P(c_i|r)$. Faces with low entropy values will likely belong to familiar people, while those with high values will likely be unfamiliar. We select a cutoff threshold t equal to one-quarter of the maximum possible entropy value, and treat faces with values above it as unknown. Our intentionally selective criterion allows us to compute accurate estimates on familiar people, and at the same time include as many unfamiliar faces as possible. We validate the impact of our conservative known/unknown decisions on discovery in Section 4.

3.3 Social Context Descriptors

For each unfamiliar face, we want to build a description that reflects that person's co-occurring familiar people, at least among those that we can already identify. Having such a description allows us to group faces that look similar (i.e., have similar appearance) and often appear among the same familiar people (i.e., have similar social context).

Suppose an image has T total faces: $r_1,\ldots,r_T$. We define the social context descriptor $S(r)$ as an N-dimensional vector that captures the distribution of familiar people that appear

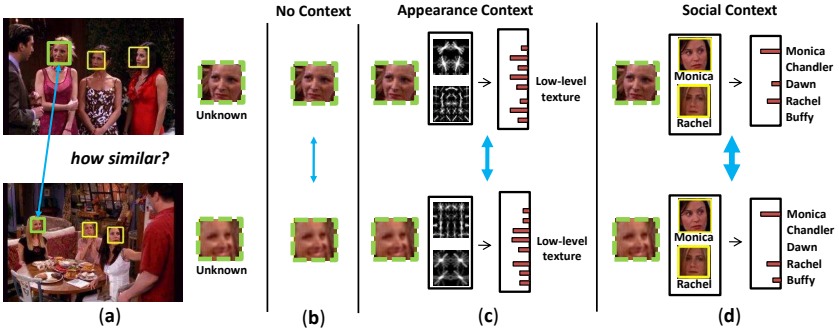


Figure 3: An example illustrating the impact of social context for discovery. The blue double-headed arrows indicate strength in affinity between the unknown regions. (a) Two images, where the unfamiliar faces are outlined in green. (b) Appearance information alone can be insufficient to deal with large pose or expression variations. (c) Modeling the context surrounding the face of interest can provide more reliable similarity estimates, but a context descriptor using raw *appearance* is limiting since it can only describe nearby faces with texture or color. (d) By modeling the *social* context using learned models of familiar people, we can obtain accurate matches between faces belonging to the same person.

in the same image:

$$S(r) = \left[\sum_{j=1}^{T} P(c_1|r_j), \ldots, \sum_{j=1}^{T} P(c_N|r_j) \right]. \quad (1)$$

If our class predictions were perfect, with posteriors equal to 1 or 0, this descriptor would be an indicator vector telling which other people appear in the image. When surrounding faces do belong to previously learned people, we will get a "peakier" vector with reliable context cues, whereas when they do not appear to be a previously learned person the classifier outputs will simply summarize the surrounding appearance.

Note that unlike existing discovery methods in object category recognition [10, 11, 20] that consider the spatial layout of the objects, we do not encode the spatial relationships between people. This is because we do not expect high regularity in how certain individuals arrange themselves (though this can be useful for broader traits like gender and age [8, 22]).

Alternatively, one can imagine forming a context description using the raw appearance of co-occurring faces—for example, by recording the pHOG or LTP descriptors of the other faces detected in the image. However, context in the form of low-level appearance information may be insufficient to provide reliable grouping cues, since the appearance variabilities of the same person (due to pose, expression changes, etc.) would not be accurately modeled (see Figure 3). By modeling social context using learned models of familiar people, we obtain more descriptive and compact representations. In Section 4, we directly evaluate the impact that the social context descriptor has on discovery over a baseline that utilizes low-level appearance features as context.

3.4 Discovering New Faces

Finally, we cluster all faces that were deemed to be unknown. We consider two clustering algorithms: (1) spectral clustering [13], and (2) complete-link agglomerative clustering. Spectral clustering provides flexibility in the choice of the affinity measure and is able to detect clusters of irregular shape. However, it requires the number of clusters as input, which is not always available for the discovery scenario. Agglomerative clustering offers more flexibility in this regard, since the size rather than the number of clusters can be targeted.

Each clustering method takes as input a matrix of the pairwise affinities between all current unknown faces.

We want the discovered groups to be influenced both by the appearance of the face regions themselves, as well as their surrounding context. Therefore, given two face regions r_m and r_n, we evaluate a kernel function K that combines their appearance similarity and context similarity:

$$K(r_m, r_n) = \alpha \cdot K_{\chi^2}(S(r_m), S(r_n)) + (1-\alpha) \cdot K_{\chi^2}(A(r_m), A(r_n)), \quad (2)$$

where α weights the contribution of social context versus appearance (recall $A(r)$ is a pHOG or LTP descriptor). Each K_{χ^2} is a χ^2 kernel function for histogram inputs x and y:

$$K_{\chi^2}(x,y) = \exp\left(-\frac{1}{2\Omega}\left(\sum_j \frac{(x_j - y_j)^2}{x_j + y_j}\right)\right), \quad (3)$$

where j indexes the histogram bins, and Ω is a data-dependent scaling factor, which we set as the average χ^2 distance between all face regions.

By considering both the appearance of the faces as well as their social context, we expect to be able to discover faces with occlusion (i.e., due to sunglasses or a hat) or large pose variations. For example, if the system knows what Monica and Chandler look like, it gets richer context descriptors to discover their pal Rachel, even in difficult cases such as when she is wearing sunglasses. Analyzing the facial appearance alone could have been inadequate to group the different instances of Rachel with and without sunglasses.

4 Results

In this section, we evaluate our method's face discovery performance.

Baselines We compare our method to two baselines: (1) a **no-context** baseline that simply clusters the face regions' appearance descriptors, and (2) an **appearance-context** discovery method that uses the appearance of surrounding faces as context (rather than the predicted categories). The second baseline substitutes the summed appearance descriptors of co-occurring faces for $S(r)$. These are important baselines to show that we would not be as well off simply looking at a model of appearance using image features, and to show the impact of social context analysis versus a low-level appearance context description for discovery.

Dataset We validate on three datasets. The first dataset (**Mixture**) is a compilation from three sources: The Gallagher Collection Person Dataset [7], an episode of *Buffy the Vampire Slayer* [5], and an episode of *Friends*. We chose these three since they contain natural cliques of people (family members, characters that appear in scenes together). There are a total of 12,542 images, 8,452 detected faces, and 23 unique people.

The second and third datasets are from [22], which are collected from real family photo albums from two different people. The second dataset (**Wang1**) has 1,125 images, 2,769 faces, and 47 people; the third dataset (**Wang2**) has 1,117 images, 3,282 faces, and 152 people. These datasets contain images encompassing real social relationships and thus are perfect testbeds for evaluating our method.[2] See [22] and the supplementary file at

[2] While the data from [22] is relevant to our task, their supervised labeling application is distinct from ours and so not relevant for comparison.

http://vision.cs.utexas.edu/projects/facediscovery/ for more details on the dataset statistics and example images.

We partition each dataset into two random subsets. The first is used to train N classifiers for the initial "knowns". These faces represent the set of people for which the system already has some tagged examples. On the second subset, we perform discovery using the N categories as context to obtain our set of discovered categories. To demonstrate that our method's improvements are robust with respect to N and which categories are chosen to be known, we test on four splits of the Mixture collection: two splits have 8 unknown people (489 and 540 face instances, respectively), the other two have 15 (1138 and 1044 face instances, respectively), all selected randomly. For the Wang1 and Wang2, we select as known the top 25% of the most frequently appearing people; the datasets have 16 and 104 unknown people (143 and 373 face instances), respectively. This reflects that the owner of the collection and his/her closest family members and friends would likely be labeled prior to those who appear less frequently.

Implementation details We use OpenCV for [21] and work only with true-positive detections. For the Mixture dataset, we use pHOG with two pyramid levels and eight bins to describe face appearance, and spectral clustering [13] to group the faces. For the Wang1 and Wang2 datasets, we use LTP with publicly available code by the authors [18] and default parameters to describe appearance, and agglomerative clustering for grouping. We worked with the pHOG descriptor in early experiments but later substituted it with the LTP descriptor due to it being more suitable for describing face patches. To compute class probabilities, we use one-vs-one SVM classifiers, and obtain posteriors using pairwise coupling [23]. We normalize the context descriptors to sum to 1. We set α to 0.5 for the Mixture dataset and 0.2 for Wang1 and Wang2 datasets. Due to the larger number of people and their varying frequencies in the Wang datasets, increasing the weight on appearance produces better clusters. In general, α could be determined interactively by observing qualitative examples of the clusters. Training the known classifiers, building the context descriptors, computing kernels, and clustering the unknowns takes 1-5 minutes with a Matlab implementation.

Evaluation metrics We use the *F-measure* to quantify discovery accuracy. The F-measure reflects the coherency (precision P) of the clusters, while taking into account the recall R of the same-category instances: $F = \frac{2 \cdot P \cdot R}{P+R}$. We set the number of clusters to discover to be equal to the number of true unfamiliar faces in the image collection, to meaningfully evaluate our method's discovery performance. To evaluate auto-tagging accuracy on novel images, we use standard multi-class recognition accuracy.

Face discovery Figure 4 shows discovery results. Our method significantly outperforms the baselines on all datasets, validating our claim that social context leads to better face discovery. In most cases, the appearance-context outperforms the no-context baseline, indicating that context can be useful even when described with low-level appearance features. However, our substantial improvement over the appearance-context baseline shows the importance of representing context with models of familiar people. The absolute performance on the more challenging Wang1 and Wang2 datasets is slightly lower than that of the Mixture dataset. Still, our method performs well, showing practical results for real personal photo collections. Furthermore, discovery succeeds just as well when the number of unknown people is increased (top to bottom in Figure 4 (b)).

We also explored taking the *least* frequent people to be known on the Wang datasets. In this case, our method attains similar clustering performance to the baselines. This is due to

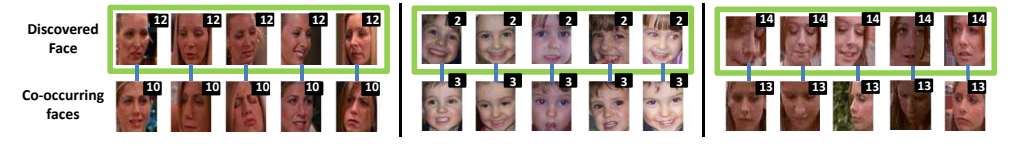
	# Unknowns	Ours	No-Context	App-Context
Mixture*	15	**0.30**	0.26	0.28
Wang1	16	**0.25**	0.20	0.21
Wang2	104	**0.24**	0.23	0.21

(a) Accuracy of discovery per dataset

	# Unknowns	Ours	No-Context	App-Context
split1	8	**0.34** (0.00)	0.24 (0.01)	0.26 (0.01)
split2	8	**0.32** (0.01)	0.23 (0.01)	0.29 (0.01)
split3	15	**0.30** (0.01)	0.26 (0.03)	0.28 (0.01)
split4	15	**0.33** (0.01)	0.28 (0.01)	0.30 (0.01)

(b) Impact of who is known ("splits")

Figure 4: Face discovery on the three datasets (a) and the different splits of the Mixture dataset (b) as judged by the F-measure. We compare our approach (Ours) with an appearance-context baseline (App-Context), and a baseline clustering only with the region descriptors (No-Context). Numbers in parentheses show range over 10 runs. Higher values are better. Our method outperforms both baselines in all cases, showing the impact of modeling the co-occurrence information of surrounding familiar people for discovery. **We take split3 to represent Mixture in (a), since it roughly corresponds to 25% of the people being known, parallel to the other datasets.*



(a) Discovery using social context

(b) Discovery using only appearance

Figure 5: Face discovery examples. (a) The first row shows representative faces of the dominant person for a discovered face, with their respective co-occurring faces below. The second row faces belong to a known person—their social context helps to group the diverse faces of the same person in the first row. The numbers indicate the ground-truth face ID. (b) Limitations of appearance-based grouping. The images show representative faces of the dominant person for a discovered face using only appearance features. Notice the limited variability in pose and expression of each grouped person, as compared to our discoveries in (a).

those people appearing in only one or two photos in the collection. Thus, meaningful models cannot be learned, which results in unreliable social context descriptors. Although this is a failure mode of our method, it is reasonable to assume that the most frequently appearing people, as opposed to those that seldom appear, would likely be tagged. In future work, we would like to consider how the system could even suggest which faces a user should tag as initially familiar, so as to maximize discovery performance.

Figure 5 (a) shows qualitative results. The representative faces of each discovered person exhibit a wide range of pose and/or illumination variations, and would not have been grouped if only facial appearance were considered. By leveraging the context from familiar people, we successfully group faces belonging to the same person. In contrast, when forming groups using only appearance cues, the discovered faces exhibit limited variability in pose or expression (see Figure 5 (b)). We show the impact of these differences on predicting novel tags with the discovered face models at the end of this section.

Familiar/unfamiliar predictions We next evaluate how accurately we predict novel instances to be familiar or unfamiliar. For this, we compute precision-recall curves, treating the known instances as positive and the unknowns as negative. See Figure 6. Our choice of the known/unknown cutoff point (indicated by the red star) leads to accurate classification

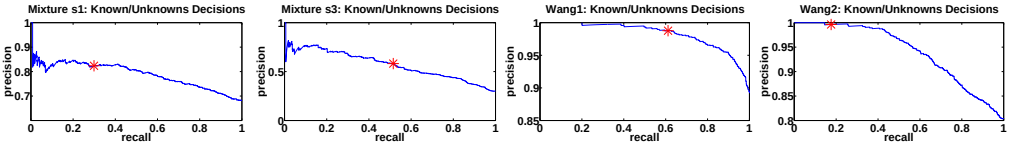


Figure 6: Precision-recall curves showing the known/unknown estimates.

	Mixture split1			Mixture split2		
	Ours	No-Context	App-Context	Ours	No-Context	App-Context
k=10	0.22	0.23	**0.29**	**0.28**	0.16	0.20
k=20	**0.30**	0.17	0.25	**0.21**	0.14	0.16
k=30	**0.27**	0.18	0.24	**0.19**	0.12	0.16

	Mixture split3			Mixture split4		
	Ours	No-Context	App-Context	Ours	No-Context	App-Context
k=10	**0.25**	0.18	0.22	**0.22**	0.18	0.21
k=20	**0.22**	0.17	0.20	**0.22**	0.14	0.19
k=30	**0.22**	0.14	0.17	**0.18**	0.10	0.13

Table 1: Face prediction on novel images with discovered faces on the Mixture dataset, as measured by classification accuracy. Note that the number of discovered clusters, *k*, is equivalent to the cost of human tagging effort required to map the discovered faces to predictive models. The models learned from faces discovered using social context generalize better than the baselines on novel face instances. The results show that our approach can serve to save human tagging effort.

for the true knowns (among the ones we determine to be known) at the cost of including some of them in the pool of unknowns. This result is especially relevant for the face tagging scenario, since the system should provide the user with a wide variety of unfamiliar (i.e., untagged) people to tag.

While we fix the selection criterion to make all known/unknown decisions in Figure 4, in order to further test our method's robustness to those predictions we measure discovery accuracy while varying the entropy cutoff value. When setting the maximum entropy value at which a face is unknown as $t = \{0.2, 0.3, \ldots, 0.6\}$, we observe consistent improvement (0.01 to 0.09 points) over the baselines.

Face recognition in novel images Finally, we evaluate how our discovered faces can be used to predict tags in novel photos. This experiment simulates an interactive face-tagging application, where the user is presented a cluster of faces that the system discovers, and the human tags it with the appropriate name. The system can then automatically tag other instances of that person given new images (for example, when the user uploads new batches of photos to her online photo collection). For this task, we use the Mixture dataset since it has a more balanced distribution in frequency counts of people in the data, providing a better testbed to evaluate prediction accuracy. The Wang datasets have heavy-tailed distributions in which a handful of people occur very frequently while the remaining people appear in only a few photos.

We classify the unknown instances in a third subset of the image data that is disjoint from both the subset on which we learned the initial familiar people models and the subset on which we performed discovery. There are 510, 600, 1152, and 1043 test instances for each split (1-4), respectively.

We train one-vs-one SVM classifiers for the discovered faces using the appearance descriptors. We label each discovered face cluster with its majority instance ground-truth tag. For this experiment, we vary the number of face clusters *k* that the system discovers in order to analyze the tradeoff between manual tagging effort and recognition accuracy.

Table 1 shows the result. For almost all *k* on each split, we consistently classify novel instances of discovered people much better than either baseline (the App-Context baseline

performs the best on split1, $k = 10$). This result shows that the models learned from faces discovered using social context generalize better on novel face instances than those learned from faces discovered using appearance alone, and is evidence that our approach can indeed serve to save human tagging effort.

5 Conclusions and Future Work

We introduced the idea of social context based discovery for faces, and demonstrated the clear advantages of replacing a traditional appearance-based framework with a learner that uses the context of familiar faces.

In future work, we will consider how to best add human supervision. The method could present a summary of each discovery (e.g., the most confident instances) to the human, who would then label it for the system to learn a model for automatic prediction in novel images. Finally, we want to consider ways in which the groupings can be revised incrementally as more data is seen.

Acknowledgements Many thanks to Andrew Gallagher, Gang Wang, and Mark Everingham for sharing their datasets. This research is supported in part by NSF CAREER IIS-0747356 and CSSG N11AP20004.

References

[1] D. Anguelov, K. Lee, S. Gokturk, and B. Sumengen. Contextual Identity Recognition in Personal Photo Albums. In *CVPR*, 2007.

[2] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller, and D. Forsyth. Names and Faces in the News. In *CVPR*, 2004.

[3] A. Bosch, A. Zisserman, and X. Munoz. Representing Shape with a Spatial Pyramid Kernel. In *CIVR*, 2007.

[4] S. Chopra, R. Hadsell, and Y. LeCun. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In *CVPR*, San Diego, CA, June 2005.

[5] M. Everingham, J. Sivic, and A. Zisserman. Hello! My Name is... Buffy - Automatic Naming of Characters in TV Video. In *BMVC*, 2006.

[6] A. Gallagher and T. Chen. Using group prior to identify people in consumer images. In *CVPR*, 2007.

[7] A. Gallagher and T. Chen. Clothing Cosegmentation for Recognizing People. In *CVPR*, 2008.

[8] A. Gallagher and T. Chen. Understanding Images of Groups of People. In *CVPR*, 2009.

[9] C. Galleguillos and S. Belongie. Context Based Object Categorization: A Critical Survey. Technical report, University of California at San Diego, 2008.

[10] S. Lazebnik and M. Raginsky. An Empirical Bayes Approach to Contextual Region Classification. In *CVPR*, 2009.

[11] Y. J. Lee and K. Grauman. Object-Graphs for Context-Aware Category Discovery. In *CVPR*, 2010.

[12] T. Mensink and J. Verbeek. Improving People Search Using Query Expansions: How Friends Help to Find People. In *ECCV*, 2008.

[13] A. Ng, M. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an Algorithm. In *NIPS*, 2001.

[14] D. Parikh and T. Chen. Hierarchical Semantics of Objects (hSOs). In *ICCV*, 2007.

[15] D. Ramanan, S. Baker, and S. Kakade. Leveraging Archival Video for Building Face Datasets. In *ICCV*, 2007.

[16] Y. Song and T. Leung. Context-Aided Human Recognition Clustering. In *ECCV*, 2006.

[17] Z. Stone, T. Zickler, and T. Darrell. Autotagging Facebook: Social Network Context Improves Photo Annotation. In *First IEEE Workshop on Internet Vision*, 2008.

[18] X. Tan and B. Triggs. Enhanced Local Texture Feature Sets for Face Recognition under Difficult Lighting Conditions. *IEEE Transactions on Image Processing*, 19(6): 1635–1650, June 2010.

[19] Y. Tian, W. Liu, R. Xiao, F. Wen, and X. Tang. A Face Annotation Framework with Partial Clustering and Interactive Labeling. In *CVPR*, 2007.

[20] Z. Tu. Auto-context and Application to High-level Vision Tasks. In *CVPR*, 2008.

[21] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *CVPR*, 2001.

[22] G. Wang, A. Gallagher, J. Luo, and D. Forsyth. Seeing People in Social Context: Recognizing People and Social Relationships. In *ECCV*, 2010.

[23] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability Estimates for Multi-Class Classification by Pairwise Coupling. *JMLR*, 5:975–1005, August 2004.

[24] L. Zhang, Y. Hu, M. Li, W. Ma, and H. Zhang. Efficient Propagation for Face Annotation in Family Albums. In *ACM MM*, 2004.

[25] M. Zhao, Y. Teo, S. Liu, T.-S. Chua, and R. Jain. Automatic Person Annotation of Family Photo Album. In *Conference on Image and Video Retrieval*, 2006.

[26] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips. Face Recognition: A Literature Survey. In *ACM Computing Surveys*, 2003.