

DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY


Autotagging Facebook: Social Network Context Improves Photo Annotation

Citation
Stone, Zak, Todd Zickler, and Trevor Darrell. 2008. Autotagging Facebook: Social network context improves photo annotation. Paper presented at the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, AK. June 23-28.

Published version
https://doi.org/10.1109/CVPRW.2008.4562956

Link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:2920117

Terms of use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles (OAP), as set forth at
https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility
https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story
The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

Autotagging Facebook: Social Network Context Improves Photo Annotation

Zak Stone
Harvard University
zstone@fas.harvard.edu

Todd Zickler
Harvard University
zickler@seas.harvard.edu

Trevor Darrell
UC Berkeley EECS & ICSI
trevor@eecs.berkeley.edu

Abstract

Most personal photos that are shared online are embedded in some form of social network, and these social networks are a potent source of contextual information that can be leveraged for automatic image understanding. In this paper, we investigate the utility of social network context for the task of automatic face recognition in personal photographs. We combine face recognition scores with social context in a conditional random field (CRF) model and apply this model to label faces in photos from the popular online social network Facebook, which is now the top photo-sharing site on the Web with billions of photos in total. We demonstrate that our simple method of enhancing face recognition with social network context substantially increases recognition performance beyond that of a baseline face recognition system.

1. Introduction

An increasing number of personal photographs are uploaded to online social networks, and these photos do not exist in isolation. Each shared image likely arrives in a batch of related photos from a trip or event; these are then associated with their photographer and broadcast out to that photographer's hundreds of online friends, and they join a collection of billions of other photographs, some of which have been manually labeled with the people they contain and other information. Social networks are an important source of image annotations, and they also provide contextual information about the social interactions among individuals that can facilitate automatic image understanding.

Despite the large and growing number of images that are embedded in online social networks, surprisingly little is known about the utility of social context for automatically parsing images. In this paper, we take a first step in this direction; we focus on the specific problem of automatic face recognition in personal photographs. We show that social network context can help tremendously even when it is combined with image data in a simple manner.

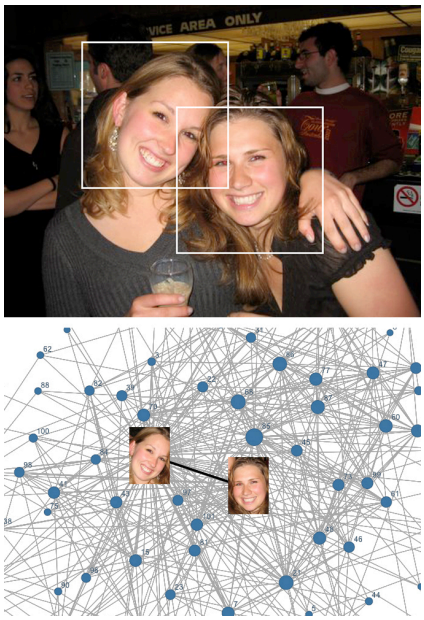


Figure 1. A typical hand-tagged photo from the Facebook social network with all tagged regions superimposed on the image. The visualization below the photo illustrates the social connections among these two individuals and their friends; for clarity, only a sample of the full graph is shown. These friendship links and other social network context can boost the accuracy of automatic face recognition in new photographs.

Our investigation uses photos and context drawn from the online social network Facebook, which is currently the most popular photo-sharing site on the Web. With over 70 million active users, Facebook already hosts a collection of billions of personal photographs, and more than 14 million new photos are posted every day [5, 6]. Many Facebook photos contain people, and these photos comprise an extremely challenging dataset for automatic face recognition. Individuals are observed with a wide variety of expressions and poses, lighting varies tremendously, and other complications such as sunglasses and heavy makeup abound.

To incorporate context from the social network, we define a conditional random field (CRF) model for each photograph that consists of a weighted combination of potential functions. One such potential comes from a baseline face recognition system, and the rest represent various aspects of social network context. The weights on these potentials are learned by maximizing the conditional log-likelihood over many training photos, and the model is then applied to label faces in novel photographs.

Drawing on a database of over one million downloaded photos, we show that incorporating social network context leads to a significant improvement in face recognition rates. We observe that much of this improvement is due to the fact that many individuals have already been manually tagged in hundreds of photographs, that they are typically photographed many times by the same photographer, and that they often appear together with a characteristic set of close friends. We conjecture that these and other phenomena are common to many photo collections in online social networks, and we argue that social network context should not be ignored by vision systems.

2. Related Work

The development of models for visual recognition of object categories has been intense in recent years, but at present, the ability to recognize large numbers of general categories in unconstrained environments remains out of reach. One way to improve object recognition performance is to consider the context in which particular objects or categories of objects are likely to be observed. An image descriptor can be computed that takes into account various properties of the scene, for example, and this scene context can then be applied to the recognition problem (*e.g.* [9, 22]).

More closely related to our work are methods that focus on the problem of person identification and use another source of context: the annotations and prior occurrances of people within a single individual's photo collection. In [13], temporal and geographic metadata was combined with co-occurrence statistics of pairs of people to provide a relevant short list of identities likely to appear in new photos. Face recognition was combined with prior co-occurrence context in [25]; the algorithm discussed therein also exploited temporal clustering and clothing matching. Other efforts have worked with ambiguously labeled data: [3], [4], and [7] resolved labels in photo collections using context from captions. Once ambiguous labels were resolved, [7] proceeded to apply a generative MRF model that took individual popularity in photos and co-occurrence statistics into account to classify faces in novel photos. Also related to our work are methods that have used hair and clothing features as context to match faces in photos taken at a single event [2, 19, 20, 24].

In contrast to the above efforts that worked with individual photo collections, we concentrate on applying the interactions and annotations of an entire community of people to improve recognition rates in the photo collections of everyone in the network. This can be viewed as a source of context that is complimentary to those previously explored.

Online social environments provide strong incentives for individuals to annotate photos, since annotation can be synonymous with sharing photos with friends. As the annotations are also shared, our system is likely to have access to many more labeled samples of each person than would be available in any individual's photo collection. The enormous and rapidly growing collection of human annotations in online social networks allows us to test algorithms in realistic conditions without exceptional data collection effort; ours is a form of "human computation" [17, 23] that exploits social behavior to gather the large-scale dataset we need to improve vision systems.

3. Facebook Faces

We conducted our study using a small portion of the Facebook social network. For this we relied on 53 volunteers, most of whom are college-age and active Facebook community members; these individuals agreed to contribute photos and metadata to our study through a web application. Using our web application, we retrieved all of the photos that had been posted by each volunteer, all photos that had been tagged with any of our volunteers' Facebook friends, all tags that were associated with any of these photos, and the network of friendships among our volunteers and their friends.

For the automatic labeling application discussed here, it is fortuitous that the tagging feature of the Facebook photo system is extremely popular. Though a sociological analysis is outside the scope of this work, tagging is at least partially driven by the fact that newly tagged photos are broadcast to the friends of the people who are tagged and to the friends of the photographer. Also, each user profile typically links to all tagged photos of that user. A more detailed analysis of user motivations behind the general practice of tagging is presented in [1]. However, it is important to note that even with the immense popularity of tagging, not all Facebook photos that contain people have been tagged; we estimate that roughly 70% of photos with people are associated with at least one tag. The proportion of faces that are tagged is difficult to estimate but is undoubtedly lower.

Our registered users and their friends number 15,752 individuals in total, and we retrieved 1.28 million tagged photos in all. From this collection, we automatically detected and aligned 438,489 face samples that could be associated with the identity labels manually entered by Facebook users. Of the users in our database, about 74% are tagged in a photo at least once, and 97% of those tagged present a computer-detectable frontal face at least once. In



Figure 2. Given the query face in the upper left, our baseline face recognition system selects the remaining faces as the most similar. The matches are presented in decreasing order of similarity from left to right and from top to bottom, and the ground-truth correct matches are highlighted with green squares. The baseline system is provided with a limited amount of context in order to be able to function at all; the search is restricted to the faces of the photographer and his or her Facebook friends. Due to the variability of our dataset, the correct matches are not highly ranked even in this restricted search space. (A small number of faces have been withheld by request from the results above, none of which were correct matches.)

this way, from only 53 community members, we compiled a labeled database of face samples that is significantly larger than most manually-collected databases reported in the literature (*e.g.* [14, 15, 16]). An alternative technique of automatically assembling large databases of labeled faces from the Web is presented in [3] and [4].

Facebook photos are taken under completely uncontrolled conditions, which makes our Facebook face dataset extremely challenging for standard automatic face recognition algorithms. Figure 2 illustrates the ranked results of a similarity search for the query face shown in the upper left using our baseline face recognition system, which was an implementation of the method reported in [10]. The results are displayed in order of decreasing similarity from left to right and then from top to bottom, and green squares indicate correct matches; they highlight the face samples of the individual who is actually shown in the query. To generate this figure, we provided the baseline face recognition system with a limited but essential amount of context: the baseline system only considers faces of friends of the photographer in whose photo the query face appeared.

We noticed several interesting features of the dataset that we collected that underscore the utility of context for automatic image labeling. First, well over 99% of the thousands of people tagged in our volunteers' albums are friends with their respective photographers, which makes it possible (as we did for Figure 2) to restrict the space of possible face labels from the millions of Facebook users to just several hundred Facebook friends without appreciable loss of performance. Second, we found that on average about 30% of the tagged faces in a photographer's albums belong to the photographer him or herself. Complete ground truth might reduce this number somewhat, since it seems reasonable that photographers would tag themselves and leave others untagged, but this statistic illustrates that the likelihood of particular individuals appearing in new photos is strongly nonuniform.

Finally, we found that people appear in photos with far fewer people than they count among their Facebook friends. In effect, photo co-occurrence defines a subgraph of an individual's friend graph that may be more relevant for predicting co-occurrence in new photos. We computed the percentages of our volunteers' Facebook friends with whom they had been tagged in a photo, and the percentages ranged from about 1% to 25% with an average of 9%. For example, a typical volunteer might have 700 Facebook friends but co-occur in photos with only 60 of them.

4. Social Network Context

In this section we develop a framework for incorporating social network context into an automatic image labeling system. For each photograph containing faces, we define a pairwise conditional random field (CRF) [8, 11, 21] with a node for each face and an edge connecting every pair of faces.

Let $\mathbf{x}$ represent the face data from all of the nodes in the photo of interest along with all known metadata, which might include the identity of the photographer, a list of the photographer's friends, and more. The goal is to infer a joint labeling $\mathbf{y} = \{y_i\}$ of face identities over all nodes i in the graph. We use the notation $y_i \in L = \{l_0 \dots l_N\}$ for the discrete label space of the nodes, which varies from photo to photo. An optimal joint labeling is found by maximizing the conditional density

$$\Pr(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} e^{E(\mathbf{y}|\mathbf{x})} \quad (1)$$

where the partition function $Z(\mathbf{x})$ is a data-dependent normalizing constant and the energy $E(\mathbf{y}|\mathbf{x})$ consists of sums of unary and pairwise potential functions corresponding to the nodes and edges of the graph:

$$E(\mathbf{y}|\mathbf{x}) = \sum_i \phi_i(y_i|\mathbf{x}) + \sum_{(i,j)} \phi_{ij}(y_i, y_j|\mathbf{x}). \quad (2)$$

In the CRF framework, the unary potential functions $\phi_i(y_i|\mathbf{x})$ capture information that is local to each node, and the pairwise potentials $\phi_{ij}(y_i, y_j|\mathbf{x})$ represent the compatibilities of possible label pairs across an edge; both types depend on all observed data.

In the case of faces, the label space $L = \{l_0, \ldots, l_N\}$ consists of a set of possible identities. As described above, this list might only include the photo's owner and his or her friends. We expand the potentials in Eq. 2 as a linear combination of univariate and bivariate functions:

$$\phi_i(y_i|\mathbf{x}) = \sum_k \alpha_k(\mathbf{x}) f_k(y_i, \mathbf{x}) \quad (3)$$

$$\phi_{ij}(y_i, y_j|\mathbf{x}) = \sum_k \beta_k(\mathbf{x}) g_k(y_i, y_j, \mathbf{x}). \quad (4)$$

The structure of this model is extremely flexible, and it can easily accommodate an arbitrary number of terms derived from the image evidence — the detected faces — and the available contextual information derived from the ambient social network. We list a few useful potentials below which follow from the discussion in the previous section.

4.1. Single-Node Potentials

Suppose a user inputs a single image that contains a set of faces and seeks a joint labeling of those faces with a known set of identities L. We will assume that identity l_0 corresponds to the user. For this inference task, we assume the availability of a face recognition system that accepts a face image and returns a distribution of likelihoods over the label space L ([26] provides a survey of possibilities). We denote this distribution of likelihoods at node i as follows:

$$f_1(y_i|\mathbf{x}) = \frac{1}{M_1}\text{FaceScore}(i, \mathbf{x}), \quad (5)$$

where M_1 is an optional normalization constant.

Unary potentials can also be used to encode contextual information from the photographic history of the user and that of his or her friends. Let P be the total set of photographs in the network that are visible to the user. If l_m represents the identity of a community member, let $F(l_m)$ be the set of identities of his or her immediate friends. Let the set of all photographs taken by community member l_m be $P(l_m)$. Finally, let $\delta(l_m, p)$ be an indicator function that is 1 when identity l_m is tagged in photograph p and 0 otherwise. Using these definitions, the following normalized distribution reflects the number of times that each person has been labeled in the user's existing personal photo collection:

$$f_2(y_i|\mathbf{x}) = \frac{1}{M_2}\left(\sum_{p \in P(l_0)} \delta(l_m, p)\right). \quad (6)$$

It is clear that we can generalize this to measure a related distribution summed over the corpus of images visible to the user (the set P) or the smaller set of images owned by the user and his or her friends (the set $P(F(l_0))$). These variations are essential, for example, when a new member joins the network and has little or no historical context available in his or her personal collection.

4.2. Pairwise Potentials

Several pairwise potentials representing interactions among pairs of individuals can be explored. Let $\text{Friend}(l_m, l_n)$ be an indicator function that is 1 when individuals l_m and l_n are 'friends' and zero otherwise; this may draw upon the explicit links of an online social network, or it may be estimated more indirectly from other interactions such as email exchanges. Note that by setting $\text{Friend}(l_m, l_m) = 0$, we can usefully bias the system away from assigning one person's identity to two or more faces in a single photo. We write a pairwise 'friendship' potential as follows:

$$g_1(y_i, y_j) = \frac{1}{N_1}\text{Friend}(l_m, l_n). \quad (7)$$

Due to the sparsity of the connections in many social networks, this alone is expected to be a useful quantity.

Next, we define potentials that measure the co-occurrence of pairs of people, either in the current photographer's previous photos or in the entire set of known photos. Let $\delta(l_m, l_n, p)$ be a pairwise indicator function that is 1 if the identities l_m and l_n have *both* been labeled in photo p and 0 otherwise; then we have the following potential:

$$g_2(y_i, y_j) = \frac{1}{N_2}\left(\sum_{p \in P} \delta(l_m, l_n, p)\right). \quad (8)$$

In addition to measuring the prior co-occurance of individuals in photographs, we may also be more broadly interested in their joint presence at previous events. Though we restrict our attention to single-photo models here, such models can easily be fused together by adding links among nodes from different photos; these links might join all faces in photos from the same event. Here, an *event* is considered to be a set of photos that are closely related and can be treated as a group. (Automated techniques for identifying events exist [12, 13].) Thus, by changing the summation in the equation above, we can also create potentials representing the

co-occurance of two individuals in events previously annotated by the user, events previously annotated by his or her friends, or events in the corpus.

Other unary, pairwise, and higher-order potentials could also be added to incorporate other forms of information, and these will be the subject of future work. For example, if we were to extract a global scene category from each photo, we could introduce potentials that represent the frequency with which individuals or groups of individuals have previously appeared in scenes of the same category as the photo being labeled. Additional nodes may also be added to the model to represent objects, clothing features, and other non-face data that could enhance identification accuracy.

4.3. Parameter Learning

In order to use our CRF model, it is necessary to set the weights α_k and β_k on each of the potentials defined above. There are many techniques for carrying out such parameter learning, and we employ the standard technique of maximizing the conditional log-likelihood of a set of training data by gradient ascent (*e.g.* [18]). In the experiments reported here, the CRF contains few enough nodes to permit exact computations, and we maximize the conditional log-likelihood with the L-BFGS-B routine [27].

5. Experiments

In this section, we present an experimental comparison between the performance of a baseline face recognition system and the performance of the same system when combined with social network context. When interpreting these results, it is important to keep in mind that although we assembled a set of over one million images, we were only able to test on several hundred faces from our volunteers' albums since we could only access full social network context from Facebook for our volunteers. A detailed discussion of this limitation and others is presented in the next section.

Our experiments measured the labeling accuracy that was achieved on held-out albums of our volunteers. In effect, we measured the performance that an automatic labeling system would achieve if each held-out album were the very last to be uploaded and could be labeled by drawing upon all of the remaining photos and annotations in the corpus. We performed this procedure both with our baseline face recognition system that ignores context and with the CRF model described above.

Before the CRF model could be applied, the weights on each potential needed to be set. Ideally, individual sets of weights would be learned for each user, but we chose to learn a single global set of parameters for the model in each experiment. Given an album to be held out for testing, we constructed the potentials defined above without using any image data or annotations from the held-out album and then optimized the potential weights as described in Section 4.3. We qualitatively observed that very similar parameters were learned no matter which album was held out, and we also saw that small variations in the weights had little effect on recognition performance later on. This led us to compute a single set of weights for a given set of potentials in the following manner: we held out a small number of randomly-selected albums one at a time, constructed potentials for each that did not include their data, found optimal weight vectors with each album held out, and then averaged these weight vectors together to produce a final set of weights for the given set of potentials.

Once a weight vector had been chosen for a set of potentials, we applied our model to each of the usable photos in our volunteers' albums in turn, always constructing our potentials to exclude both image and annotation data from the test photo's surrounding album. This means that both the gallery of labeled faces and the social network context changed for each test album. For each photo tested, we restricted the label space L defined above to contain the photographer and his or her Facebook friends. Without this restriction, our baseline face recognition performance would plummet even on our small dataset, and it is difficult to imagine successfully matching a face directly against Facebook's full database of millions of users. This illustrates one of the major advantages of gathering information from a social network instead of from isolated personal photo collections; for every photographer, a list of the people likely to show up in their photos has *already been collected* for other reasons, and without any further effort it can be applied to label those people in new photos automatically.

The CRF models applied in our experiments included various combinations of the potentials defined above. The baseline distribution of face scores in Eq. (5) was computed with an implementation of the method reported in [10]. The single-person photographic history potential in Eq. (6) was computed over all of the volunteer's photographs outside of the held-out album. Facebook friendships were used to compute the friendship potential from Eq. (7). The pairwise co-occurrence potential described by Eq. (8) was computed over *all* photographs outside the held-out album, whether they were taken by the photographer, by one of our volunteers, or by anyone else in our system. We found it useful to threshold the values of this potential, perhaps because global co-occurrence may not reliably predict co-occurrence in any individual's album, but it usefully restricts the space of friends who *might* appear together.

During inference, we computed the exact marginal probabilities for each face node. The putative label with the maximum marginal probability was assigned to each node and compared against ground truth to measure accuracy. The marginal probability estimates made it simple for us to compute a ranked list of labels for each face sample, and

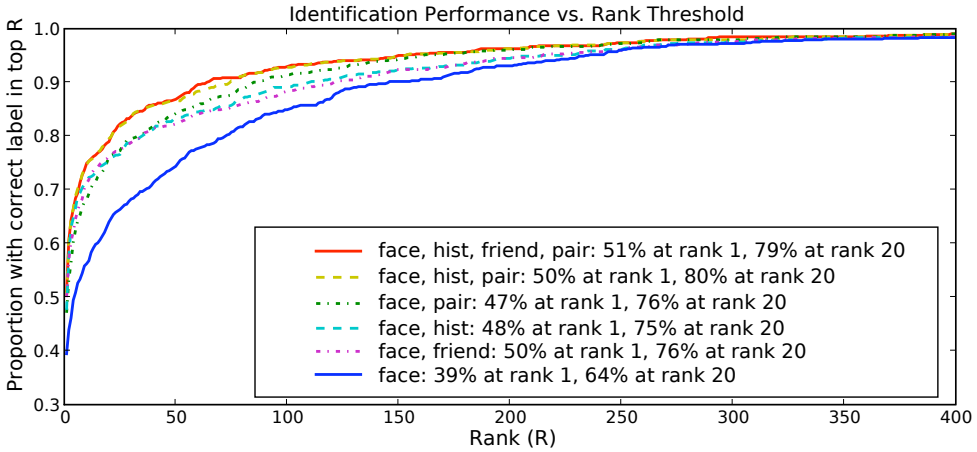


Figure 3. Identification performance as a function of rank threshold for a variety of CRF potential combinations. At each rank value R, this plot displays the proportion of all test samples for which the correct ground-truth label appeared in the top R predictions. While individual context potentials can boost performance on their own, combinations of context potentials improve performance further. The accuracy of our context-enhanced system starts higher and rises much faster than the accuracy of the baseline face recognition system, which suggests that it would be useful for the display of a short-list of likely names or corrections in a user interface.

we measured how often the correct label appeared in the top R ranks (*cf.* [15]). We then averaged these rank results across all faces in our test photos to compute average performance as a function of rank threshold, which we show in Figure 3. The curves in the figure show the average proportion of samples for which the correct label appears in the top R predictions as a function of R for a variety of potential combinations. The figure demonstrates that individual context potentials can improve substantially on the baseline face recognition output on their own but produce the best performance together. The curves clearly illustrate the advantage of including contextual information in our model; the context-enhanced curves climb much more steeply than the baseline curve in the useful range of small R. This implies that our system could provide an accurate short list that would be very helpful in any realistic annotation interface.

6. Limitations

It is important to recognize that our collection of over a million images only samples the online social universe of 53 people out of Facebook's 70 million users. Most of our volunteers attend the same college, and many of them know one another, so it is possible that their photographic habits do not adequately represent the habits of Facebook users as a whole. Some of our volunteers have posted over twenty albums of photos on Facebook and diligently tagged nearly every face they contain, while others have posted a single album with a handful of tags; in future work, we hope to incorporate an estimate of a user's activity level into our model to better accommodate these large variations. (In our experiments, we have simply used the same parameters α_k and β_k for all users.)

Our technique will be most effective when social network context is available about *all* of the people who are likely to appear in a photographer's photos, but this information may not be available from Facebook for many reasons. First, even in the populations where Facebook is most actively used, not everyone is a member and not all members have been tagged in photographs, so it will be impossible for the system to label some individuals in newly posted photos. Furthermore, Facebook provides its users with a wide variety of privacy settings that restrict the availability of personal information, and we are not able to find out how much of our volunteers' actual social network as represented on Facebook we have been allowed to access. We expect that our context-based labeling technique would perform far better with complete access to Facebook's data.

Since not all faces in existing Facebook photos are labeled and not all people in photos face the camera, we performed our experiments without access to complete ground truth. While eight people may appear in a photo, only two computer-detectable faces may have been tagged, and for this preliminary study, we treated such photos as if they contained only the detectable tagged individuals. In the future, we hope to provide our volunteers with a live interface that requests complete ground truth in exchange for mostly-automated tagging in order to improve the realism of our dataset. We expect that additional ground truth labels will only boost the performance of our algorithm by providing more face samples and richer contextual information.

Facebook photo tags are used for a variety of expressive purposes beyond accurately identifying people in photos,

so it is important to filter out tags that do not correspond to actual faces. We ran the OpenCV implementation of the Viola-Jones face detector on every photo to locate frontal faces, and we then matched detected faces with nearly concentric tags using a conservative threshold to assign labels to face samples. The OpenCV detector is relatively permissive, so a more detailed analysis of each matched face was then applied to discard low-quality samples. Enhancements at every step of this pipeline could increase the amount of training data available to our system.

To simplify the training and application of our model, we report results for photos containing exactly two faces with available ground-truth labels to highlight the effect of the pairwise context we have available. The majority of the photos we collected with more than one face contained only two, so this restriction allowed us to work with a large part of our available dataset while postponing the challenge of approximate parameter learning in larger graphs with variable numbers of nodes. In future work, as we have more photos with larger numbers of labeled faces available, we will apply our model to photos that contain arbitrary numbers of faces, including single-face photos for which only more limited context is available.

Our 53 volunteers had posted a total of 420 albums on Facebook at the time of this study, and under the constraints discussed above, these albums contained 722 labeled face samples suitable for testing. We drew upon the much larger set of 438,489 labeled face samples along with the Facebook friendships among the 15,752 corresponding individuals to perform our experiments.

7. Discussion

This study suggests a new paradigm for research on personal photographs: instead of working with small datasets that have been painstakingly collected and manually labeled by researchers, the computer vision community has the opportunity to gather large quantities of data from millions of volunteers — as long as we provide them with a genuinely useful service and guarantee that their privacy will be protected as long as their data is stored. By partially automating the entry of photo annotations that millions of people currently enter by hand, we can direct users' energy to provide the more difficult ground-truth labels that our automated systems cannot predict. With such a system, users will benefit by investing far less effort to achieve the same effect they do now, and the computer vision community will have access to human users who can be coaxed to answer any reasonable question about ground truth in any vision problem.

The amount of manual labeling available in Facebook is substantial: we were able to extract useful labeled face samples of nearly all of our 53 volunteers and of around three-quarters of their 15,699 Facebook friends. We believe that a fully automatic service based on our work would already be useful for the majority of Facebook users.

Among our volunteers, there are miniature social clusters, and we qualitatively observe that recognition performance appears higher for the photographers who are at the center of those clusters; it would be interesting to attempt to quantify this effect. As more information from the network becomes available, we expect that the accuracy results reported here will only improve, and it will become possible to incorporate other measures of interpersonal interactions from the network that are not currently available.

We chose to focus on data from Facebook for this study because it provides the photo tags and social network context we need on a significant scale. However, we emphasize that contextual information that is useful for improving person identification accuracy can be collected from an increasing variety of sources. For example, patterns of email exchange among individuals carry rich information about the strength of relationships, online calendars and PDA's record our scheduled commitments with increasing information about who will appear at each, and many people already own camera phones that can measure their own geographic coordinates.

In real life, context is time-dependent, and it would be valuable to allow time-dependent social context in our model. People move around the world, relationships begin and end, and the likelihood of appearing in a particular person's photos at a particular time depends strongly on such time-dependent behavior. It would also be desirable to introduce richer models of social groupings. People associate with each other in homes, schools, workplaces, clubs, societies, and a multitude of other organizations, and a more explicit representation of these would be interesting in its own right and might enhance labeling performance.

We have focused on small graphical models in this study for simplicity, but it would likely be very useful to aggregate single-photo graphical models into a much larger complete graph with a node for every person who appears in an event. The edge potentials from one photo to another would encourage same-label assignments, since people at an event are likely to be photographed more than once, and the edge potentials within photos would continue to suppress duplicate labelings. Additional unsupervised clustering techniques could be applied to group multiple appearances of unknown people (by faces, clothing, and other means) in an event into single nodes in such an event-wide CRF model, which would allow for more robust matching against the library of known faces.

8. Conclusions

Existing metadata from online social networks can dramatically improve automatic photo annotation. Personal photos are highly variable in apperance but are increasingly

shared online in social networks that contain a great deal of information about the photographers, the people who are photographed, and their various relationships. Our method combines image data with social network context in a conditional random field model to improve recognition performance.

We have applied our technique to a portion of the world's largest database of hand-labeled faces, the tagged faces in personal photographs posted on the popular social network Facebook. We demonstrate that social network context from Facebook provides a substantial increase in recognition performance over a baseline face recognition system on realistic photo collections drawn from 53 volunteers and their thousands of Facebook friends.

We believe that our current system could already provide a useful service to the millions of photographers who until now have tediously labeled billions of photographs by hand. As we scale up our system, the massive quantities of diverse online data we obtain and the test of real-world use will help us improve the system further.

9. Acknowledgments

We warmly thank the friends and volunteers who contributed photos and other data to this project. We also thank Mike Jones, Kuntal Sengupta, and Jay Thornton of MERL for helpful discussions regarding the implementation of the face recognition system used in this work. The first author gratefully acknowledges the support of a National Science Foundation Graduate Research Fellowship. The social graph visualization in Figure 1 was rendered at `http://many-eyes.com`.

References

[1] M. Ames and M. Naaman. Why We Tag: Motivations for Annotation in Mobile and Online Media. *SIGCHI*, pages 971–980, 2007.

[2] D. Anguelov, K. Lee, S. Gokturk, and B. Sumengen. Contextual Identity Recognition in Personal Photo Albums. *CVPR*, pages 1–7, 2007.

[3] T. Berg, A. Berg, J. Edwards, and D. Forsyth. Who's in the Picture? *NIPS*, 2004.

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

[5] The Facebook Blog: `http://blog.facebook.com/blog.php?post=2406207130`. (Accessed on 5/4/2008.).

[6] Facebook Statistics: `http://www.facebook.com/press/info.php?statistics`. (Accessed on 5/4/2008.).

[7] A. Gallagher and T. Chen. Using Group Prior to Identify People in Consumer Images. *CVPR*, 2007.

[8] X. He, R. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labeling. *CVPR*, 2004.

[9] D. Hoiem, A. Efros, and M. Hebert. Putting Objects in Perspective. *CVPR*, 2006.

[10] M. Jones and P. Viola. Face Recognition Using Boosted Local Features. *MERL Technical Report Number: TR2003-25*, 2003.

[11] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *ICML*, pages 282–289, 2001.

[12] M. Naaman, Y. Song, A. Paepcke, and H. Garcia-Molina. Automatic Organization for Digital Photographs with Geographic Coordinates. *International Conf. on Digital Libraries*, pages 53–62, 2004.

[13] M. Naaman, R. Yeh, H. Garcia-Molina, and A. Paepcke. Leveraging Context to Resolve Identity in Photo Albums. *International Conf. on Digital Libraries*, pages 178–187, 2005.

[14] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the Face Recognition Grand Challenge. *CVPR*, pages 947–954, 2005.

[15] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET Evaluation Methodology for Face-Recognition Algorithms. *PAMI*, 22(10):1090–1104, 2000.

[16] P. Phillips, W. Scruggs, A. OToole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 Large-Scale Results. Technical Report NISTIR 7408, NIST, 2007.

[17] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. Technical Report MIT-CSAIL-TR-2005-056, MIT, 2005.

[18] F. Sha and F. Pereira. Shallow Parsing with Conditional Random Fields. *Proc. of the Conf. of the Assoc. for Comp. Ling. on Human Lang. Tech.*, pages 134–141, 2003.

[19] J. Sivic, C. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. *British Machine Vision Conference*, 2006.

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

[21] C. Sutton and A. McCallum. An Introduction to Conditional Random Fields for Relational Learning. *Introduction to Statistical Relational Learning*, 2007.

[22] A. Torralba. Contextual Priming for Object Detection. *IJCV*, 53(2):169–191, 2003.

[23] L. von Ahn. Human Computation. *Proc. of the International Conf. on Knowledge Capture*, pages 5–6, 2007.

[24] L. Zhang, L. Chen, M. Li, and H. Zhang. Automated Annotation of Human Faces in Family Albums. *Proc. of the ACM International Conf. on Multimedia*, pages 355–358, 2003.

[25] M. Zhao, Y. Teo, S. Liu, T. Chua, and R. Jain. Automatic Person Annotation of Family Photo Album. *International Conf. on Image and Video Retrieval*, pages 163–172, 2006.

[26] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face Recognition: A Literature Survey. *ACM Computing Surveys*, 35(4):399–458, 2003.

[27] C. Zhu, R. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997.