INVITED PAPER

Toward Large-Scale Face Recognition Using Social Network Context

The authors of this paper believe that social incentives can be used to obtain numerous facial images of faces and they propose a computational method for using these images.

By Zak Stone, *Student Member IEEE*, Todd Zickler, *Member IEEE*, and Trevor Darrell, *Member IEEE*

ABSTRACT | Personal photographs are being captured in digital form at an accelerating rate, and our computational tools for searching, browsing, and sharing these photos are struggling to keep pace. One promising approach is automatic face recognition, which would allow photos to be organized by the identities of the individuals they contain. However, achieving accurate recognition at the scale of the Web requires discriminating among hundreds of millions of individuals and would seem to be a daunting task. This paper argues that social network context may be the key for large-scale face recognition to succeed. Many personal photographs are shared on the Web through online social network sites, and we can leverage the resources and structure of such social networks to improve face recognition rates on the images shared. Drawing upon real photo collections from volunteers who are members of a popular online social network, we asses the availability of resources to improve face recognition and discuss techniques for applying these resources.

KEYWORDS | Face recognition; graphical models; social network context; structured prediction

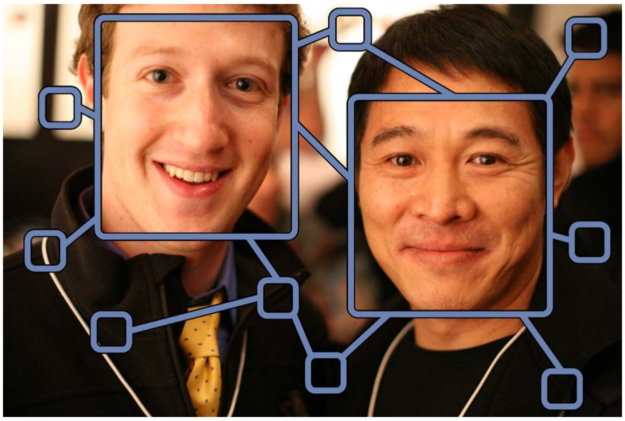

Fig. 1. *The billions of personal photographs shared in online social networks present a new opportunity to develop "socially aware" face recognition systems. By leveraging contextual information about the social relationships among photographers and their friends, these systems have the potential to achieve accurate recognition on Internet-scale photo collections that contain hundreds of millions of people. (Photo courtesy of Robert Scoble under a Creative Commons License [1].)*

I. INTRODUCTION

It has never been easier to capture a photograph, and trends indicate that personal digital photography will continue to become simpler, cheaper, and more available to people around the world. Unfortunately, our ability to automatically analyze and organize photos lags far behind our ability to create and store them. As petabytes of visual data are becoming commonplace, we require new tools to

Manuscript received November 3, 2009; revised February 6, 2010; accepted February 7, 2010. Date of publication May 24, 2010; date of current version July 21, 2010

Z. Stone and **T. Zickler** are with the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 USA (e-mail: zstone@post.harvard.edu; zickler@seas.harvard.edu).

T. Darrell is with the Department of Electrical Engineering and Computer Sciences, University of California Berkeley, Berkeley, CA 94720 USA (e-mail: trevor@eecs.berkeley.edu).

Digital Object Identifier: 10.1109/JPROC.2010.2044551

0018-9219/$26.00 ©2010 IEEE

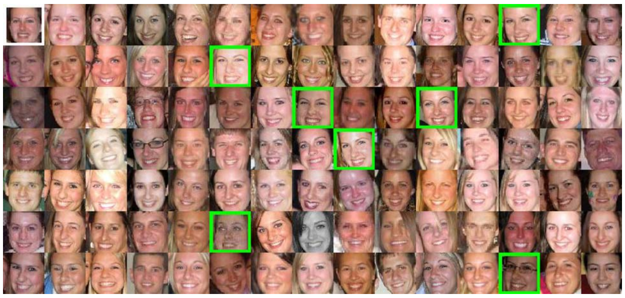

Fig. 2. ***Current face recognition performance is poor when faces are photographed "in the wild" with uncontrolled variations in pose, expression, and illumination. This figure shows the ranked results from a commercial face recognition system as it attempts to match the query face in the upper left (outlined in white) against a set of thousands of labeled face images of 731 individuals harvested from photos on Facebook, a popular online social network. The matches are presented in decreasing order of similarity from left to right and from top to bottom, and the correct identity matches are highlighted in green.***

automatically parse images so that they can be effectively indexed, browsed, searched, and shared.

One useful way to index photographs—especially personal photographs—is through the identities of the individuals they contain, and, in theory, this can be executed at scale using automatic face recognition. However, recognizing individuals from facial images is a hard problem, particularly when the images are like those in Figs. 1 and 2: collected "in the wild" with uncontrolled variations in pose, lighting, and expression. This difficulty is exacerbated in large online photo collections in which hundreds of millions of individuals might appear; the difference in appearance between individuals becomes very small relative to the appearance variation of any particular individual. Furthermore, even the preparation of training data (by manually labeling images, for example) to enroll people in an automatic recognition system becomes burdensome.

This paper argues that online social networks can provide the keys to successful face recognition in large photo collections on the Web. This argument is based on two observations. First, online communities induce social incentives for members to manually attach identity labels to facial images. The resulting practice of users voluntarily "tagging" themselves and their friends in photos can produce extraordinary quantities of labeled facial images, which reduces or eliminates the traditional enrollment burden. The second observation is that the social network graph of an online community, which is often available in machine-readable form, provides powerful contextual information that improves both performance and computational efficiency.

By drawing on photos embedded in the online social network Facebook, we assess the availability of labeled face data, and we build on our earlier study [2] to show how social network context can be leveraged to improve recognition. While these results are preliminary, they suggest that "socially aware" face recognition is a problem that deserves research attention.

II. FACE RECOGNITION IN PERSONAL PHOTOGRAPHS

Face recognition is a relatively mature topic in computer vision, and recognition rates on moderately large databases captured under controlled view and lighting conditions can be quite high (e.g., [3]). However, in personal photographs, the conditions are rather uncontrolled: as depicted in Fig. 2, faces exhibit a wide range of pose, expression, illumination, and makeup variation that is difficult for recognition systems to handle.

Recognition rates in such uncontrolled settings are improving thanks to ongoing developments in face detection and alignment [4], [5], feature extraction that is insensitive to changes in pose, expression, and illumination [6], [7], and face-specific metric learning and classification [8], [9]. This research is being spurred by the collection and dissemination of "standard" data sets containing hundreds or thousands of individuals [6], [10].

In parallel to these advances, there has been interest in understanding—as we seek to do in this paper—when and how contextual information of various forms can be used to improve recognition. For example, the performance of recognizing celebrities can be boosted by exploiting captions and scripts that accompany some video feeds [5], [10]–[13], or by exploiting the link structure and the text/image co-occurrence that exists on the Web [14], [15]. There is also significant contextual information available within an individual's personal photo collection, especially within subcollections corresponding to particular events. In this setting, recognition systems can exploit the fact that individuals have consistent clothing and hairstyles between photos, and that certain individuals and groups appear more frequently than others [16]–[20]. Contextual information can come from other places as well, such as apparent social norms for positioning in "group shots" [21], and census data that link names to visually salient attributes such as age and gender [22].

The use of social network resources, as described in this paper, can be seen as a source of contextual information that compliments the ones listed above. These resources come in two forms—labeled facial images and social network structure—and, as compared to other forms of contextual information, they are unique in terms of their scale and utility. As we discuss in later sections, these social network resources have several desirable properties.

1) The resources are "free" in that they accumulate as a natural by-product of human interaction online.

2) Data from social networks are already available in enormous quantities, and further growth seems likely.
3) Identity labels on photos tend to be highly accurate because there are social incentives for them to be correct.
4) With the increasing connectivity of the Web, these resources can potentially be exploited by a diverse set of recognition systems, some of which may be mobile and ubiquitous.
5) Trends suggest that additional resources, such as timestamps and geotags, will become available as quickly as technology permits.

III. PHOTO TAGGING IN ONLINE COMMUNITIES

One important source of information in online social networks—and Facebook in particular—is the vast quantity of facial images that have been manually labeled, or "tagged," by identity. The popularity of tagging is somewhat surprising, because tagging images by associating captions, annotations, or keywords is a tedious process—so tedious that very few people actually take the time to tag the images in their personal libraries [23]. This lack of personal tagging persists despite the fact that efficient tagging interfaces for personal photo collections have existed for almost a decade [24]–[26] and tags can significantly improve personal image organization and retrieval [27].

Interestingly, things seem to change when images are shared online. Online communities induce *social incentives* to tag, and, as evidenced by the density of tags in Facebook and other online communities, these incentives can be quite strong. Recent studies are beginning to explore this phenomenon [28]–[30], and they suggest that the social incentives for tagging can be quite diverse. On Facebook, tags typically correspond to the identities of individuals in an image, and these tags are used to ensure that an image will be seen by one's friends. When Avery tags Ben in a Facebook photo, Ben receives an e-mail message with a link to the image, and both Avery's friends and Ben's friends might find the tag mentioned in streaming "news feeds" on the site. In this way, Avery successfully shares the photo with Ben, and (perhaps) Ben's stature is enhanced among their combined group of friends.

Whatever the reasons for social tagging, the practice is a boon for recognition systems. At the time of this writing, Facebook has a rapidly growing population of more than 400 million users, and it hosts over 20 billion images, with more than 2.5 billion new photos being added every month [31], [32]. Many of these images have been manually tagged with individuals' identities, and, in this way, the members of this online community have inadvertently created an astoundingly large database of annotated facial images embedded in a social network structure that can be accessed (at least partially) in machine readable form.

Table 1 Results of an Empirical Study of the Social Network Resources Available to Aid Face Recognition Systems on Facebook. Data Were Accessed Using a Standard "Facebook Application" Authorized by 50 College-Age Volunteers

	Oct. 2007	July 2009
Friends per volunteer (avg.)	432	645
Volunteers and friends (total # individuals)	15752	22108
Individuals tagged at least once*	11183	14939
Photos	3.08M	7.7M
Tags	3.46M	8.1M
Photos with at least one tag	1.79M	3.98M

*These were individuals with at least one tag that could be associated with a machine-detectable frontal face. The raw numbers are slightly higher at 11645 and 15399.

A. An Empirical Study

We recently performed an empirical study to measure the availability of labeled face data. Our study was conducted using a very small portion of the Facebook social network associated with 50 college-age volunteers. We retrieved all photos that had been posted by the 50 volunteers and all photos taken by others that had been tagged with any of our volunteers or their friends. We also retrieved all of the identity tags and metadata associated with these photos, and we attempted to collect the network of "Facebook friendships" among our volunteers and their friends.[1]

The results of this empirical study are summarized in Table 1, along with the numbers we collected nearly two years earlier (see [2]) using almost the same volunteers. In this most recent study, the recovered network for the 50 registered volunteers and their friends includes 22 108 individuals in total, and the number of their photos that can be retrieved is more than seven million. There are more than eight million identity tags associated with these images, and nearly four million images have at least one attached tag.

The tagging interface in Facebook does not constrain the image location at which a user applies an identity tag. Fortunately, many users seem to apply these tags on or near individuals' faces, which makes associating identity tags with facial images reasonably accurate. In our data set, we used an open-source frontal face detector [33] to detect faces in the four million tagged images, and we found that 32% of the eight million manually attached tags could be very reliably associated with a machine-detectable frontal face.

The process of associating machine-detectable faces with identity tags ultimately produced a set of labeled facial images that includes 2.5 million samples of 385 624 individuals. The distribution of samples per individual in

[1] More precisely, our volunteers granted access to a Facebook Platform application that we developed, and the application acquired all network connections, photos, and tags that were accessible via the Facebook API. The privacy settings of many users prevented us from accessing complete information.

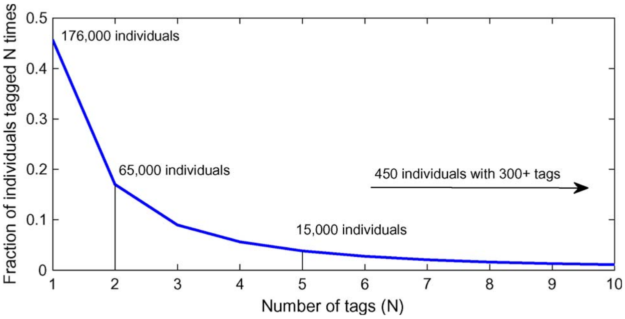

Fig. 3. ***Fractions of individuals in our study who are associated with N computer-detectable tagged face images. While most of the people referenced in our data set only appear in photos a handful of times, ten or more images are available for 44 000 people, and 300 or more images are available for 450 people. The number of human-assigned tags per individual is much higher; here we only count tags that could be assigned to computer-detected frontal faces with very high confidence.***

this set is shown in Fig. 3. For nearly half of the individuals, there is only one facial sample, but in many cases, there are more. For example, there are 15 000 individuals with exactly five facial samples, 450 individuals with 300 samples or more, and a handful of individuals with more than 1000 samples. If we view this set of images as a database for evaluating face recognition systems "in the wild," it is orders of magnitude larger than existing alternatives [6], [10].

We noticed several interesting features of this data set that are relevant to face recognition in the context of the social network. Of the 22 108 volunteers and friends in the network, 67% could be associated with at least one labeled facial sample. Our volunteers have 645 friends on average, which is substantially higher than Facebook's reported average of 130 friends per user on the entire site [31]. Since every photo is uploaded by a known user (whom we will call *the photographer*), it is possible for a recognition system to draw upon social context surrounding the photographer to reduce the set of possible identity labels that is considered for each detected face in each photo. Another interesting observation is that, on average, about 30% of the tagged faces in a photographer's albums belong to the photographer him or herself.

In this data set, people appear in photos with 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 average is only 13%.

IV. FACE RECOGNITION WITH SOCIAL CONTEXT

We consider the task of recognizing faces in a photograph as a joint labeling problem. As input, we are given an image and some associated metadata, which might include a timestamp, geotag, photographer identity, and one or more manually attached annotations. For simplicity, we will assume that the image has already been parsed into a discrete set of face regions via application of a face detection algorithm (e.g., [34]). We further assume that each detected face is associated with a discrete set of allowable identity labels. Our goal is to infer from these sets the correct label for each face.

An example input photo is shown in Fig. 4, where a face detector has located three faces to be identified. We seek a labeling that is supported by the image data (i.e., the appearance of each face) as well as the known social network structure (i.e., the relationships between individuals).

Formulated in this way, the recognition problem is one of *structured prediction* [36]–[40]. Given an input image-with-metadata $\mathbf{x}$, we seek to infer a joint labeling $\mathbf{y}$, and this is accomplished by learning a function $h : \mathcal{X} \longrightarrow \mathcal{Y}$ that maps inputs $\mathbf{x} \in \mathcal{X}$ to $\mathbf{y} \in \mathcal{Y}$. Using the notation of [39], this function is expressed as

$$h(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \tag{1}$$

where the function $f : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$ captures the essence of the problem and must be learned from training data. This learning process is made tractable by expressing the

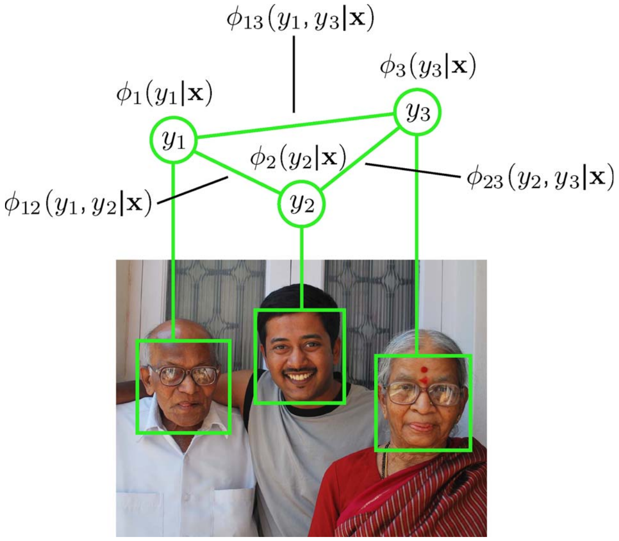
$\phi_{13}(y_1, y_3|\mathbf{x})$
$\phi_3(y_3|\mathbf{x})$
$\phi_1(y_1|\mathbf{x})$
y_3
y_1
$\phi_2(y_2|\mathbf{x})$
$\phi_{23}(y_2, y_3|\mathbf{x})$
$\phi_{12}(y_1, y_2|\mathbf{x})$
y_2

Fig. 4. ***A visualization of the input features ϕ and the output labels y_i in the structured prediction problem of jointly labeling faces in personal photographs. The graphical model above the photo contains a node for each detected face, and the nodes are connected in a complete graph; the goal is to infer the identity labels y_i. Inference is accomplished by drawing upon available "features" that correspond to each node and each edge. These features include both image data and context from the embedding social network. (Photo courtesy of Flickr user* mynameisharsha *under a Creative Commons License [35].)***

function f as a linear combination of fixed "feature functions" and then learning only the combining weights on these feature functions.

The example in Fig. 4 represents a specific structured prediction problem in which the function f is constructed from a pairwise Markov random field (MRF). The model contains a node for each detected face, and these nodes are connected with all possible edges to form a complete graph. There are feature functions associated with each node in the graph, such as recognition scores from a face recognition subsystem, and feature functions associated with each edge, such as the strength of the social tie between each pair of individuals. This model is represented as

$$f(\mathbf{x},\mathbf{y}) = \sum_i \phi_i(y_i|\mathbf{x}) + \sum_{(i,j),i\neq j} \phi_{ij}(y_i,y_j|\mathbf{x}) \qquad (2)$$

with

$$\phi_i(y_i|\mathbf{x}) = \sum_m \alpha_m(\mathbf{x}) f_m(y_i,\mathbf{x})$$
$$\phi_{ij}(y_i,y_j|\mathbf{x}) = \sum_n \beta_n(\mathbf{x}) g_n(y_i,y_j,\mathbf{x}).$$

Here, f_m and g_n represent different univariate and bivariate feature functions, and α_m and β_n are the learned weights that combine them.

The advantage of this approach is that arbitrary, possibly mutually dependent feature functions f_m and g_n can be proposed for f, and the combining weights can be trained discriminatively. Consider again the face example in Fig. 4. In addition to the image-based face recognition score mentioned above, the univariate functions might include measures of the social prominence of individuals [41], their likelihoods conditioned on a timestamp or geotag, or their likelihood of being photographed by this particular photographer at this particular time. Similarly, the bivariate functions might include a variety of measures for the social relations between pairs of individuals within the social network [42]–[45].

There are three main challenges to this approach to recognition, and all of them are surmountable. First, we require the means to perform inference by carrying out the argmax operation in (1). In many cases, this is intractable, but the problem has received intense interest during the past few years, especially for MRF-based structures like that in (2). A number of promising approximate inference schemes now exist, including those based on message passing (e.g., [46]) and graph-cuts (e.g., [47]). A second challenge is learning the weights α_m and β_n for a proposed set of feature functions. This requires a large number of pairs $\{\mathbf{x}_k,\mathbf{y}_k\}$—input images for which the true labels for all regions are known—and an efficient method for learning. The former can be obtained from the Internet as described in the previous section, and particularly useful for the latter are efficient large-margin techniques, such as those based on structural support vector machines [48], [49].

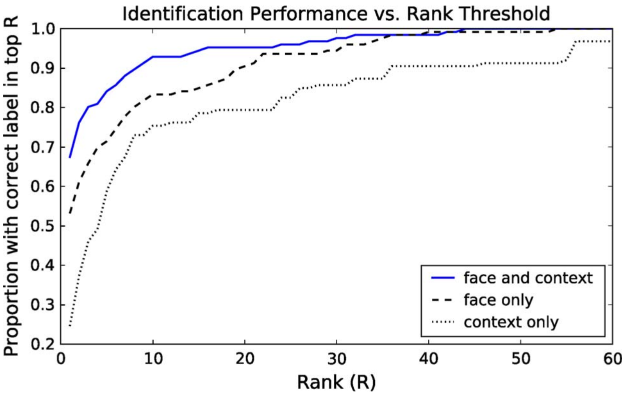

Fig. 5. ***Combining facial appearance and social network context for face recognition. The data set is split into training and testing sets (roughly 80%/20%) according to a time threshold*** t***. Training images are used to learn models of facial appearance and social (pairwise) relationships, and these models are used to recognize individuals in the test set. The figure displays identification performance as a function of rank threshold: at each rank value*** R***, it shows the proportion of all test samples for which the correct identity label appeared in the top*** R ***predictions. Results are shown for the facial appearance model*** (*face*)***, the social relationship model*** (*context*)***, and their combination.***

As a simple illustration of these techniques, Fig. 5 demonstrates that even basic contextual information can improve the recognition performance of a face recognition system. To generate this figure, we ran an improved version of the evaluation in [2] on our expanded data set. The data set was split into training and test sets based on a time threshold t. This imitates the realistic application of labeling people in a batch of new photographs uploaded after time t by drawing upon the network resources that were available before time t. For testing, we restricted our attention to photos that contain exactly two high-quality labeled faces to highlight the effect of context without the distractions of variable graph sizes and approximate inference, and the time threshold t was chosen such that the training set contained approximately 80% of the usable photos.

To set the label (identity) space for each test photo, one could use the list of Facebook friends of the photographer as described earlier, but timestamps for friend link creation are currently unavailable through the Facebook API. This prevents us from using friend links in a manner consistent with the time-based split defined above. Instead, we set the label space for each test photo to be the union of: 1) the users who have been tagged in the photographer's photos before time t; and 2) the individuals with whom the photographer has been jointly tagged in *any*

Facebook photo before time t. Finally, for the purposes of this evaluation, we considered the simplest possible feature functions: a single univariate function f_m that comes from a face recognition system and scores identities using appearance information, and a single bivariate function g_n that measures the thresholded pairwise photo cooccurences of the individuals in the training set. The weights on these face and context functions were learned by maximizing the conditional log likelihood of the training data (see [2]), and once these weights were learned, we performed inference [see (1)] and computed marginal probability distributions at each node to produce a ranked list of identity labels for each detected face.

The curves in Fig. 5 differ somewhat from the results of our previous study [2] for several reasons: we incorporated an improved face recognition subsystem based upon the work of Everingham *et al.* [5], the data set was approximately twice as large due to the passage of time (see Table 1), the label space per photo was different and somewhat smaller, and we only split the data set once based on a time threshold. However, the qualitative trend is the same—while face recognition beats guessing "with your eyes closed," face recognition and social context combined yield better recognition rates than either information source alone.

V. CONCLUSION

The ubiquity of identity tags in communities such as Facebook strongly suggests that social incentives can be leveraged to obtain significant quantities of labeled facial images of millions of individuals. To advance the state of the art in face recognition, the questions of how best to apply these data and how to build scalable recognition systems are worthy of attention. This paper argues that social network context is an important tool for assembling scalable recognition systems, and it provides an example of a simple computational architecture for utilizing this contextual information.

We have only begun to consider the wide variety of social signals that are readily available from Facebook and other online social networks to improve recognition, and additional sources of information will undoubtedly provide a far bigger boost in recognition accuracy than we observed in this small study. Photo timestamps, gender information, individuals' names [50] and positions within a photo [21], scene context [51]–[55], and various sources of within-album information [16]–[20], [56] are all immediate possibilities.

In order to put all of this information to use, it will likely be beneficial to move beyond the simple pairwise MRF structure described in this paper. For example, one might build graphs that span multiple photos to jointly recognize individuals over a short stretch of time or an entire event, and hierarchical models might capture group effects caused by shared affiliations having salient visual signatures (soccer teams, outdoor clubs, cultural societies, etc.).

In all of these cases, the increased complexity of the graphical model will make (approximate) inference and learning more difficult, and this provides an intriguing application for efficient techniques that have recently been proposed (e.g., [38], [40], and [57]–[61]). Also, since the size of the graphical model will often vary from one photo (or event) to the next, one must explore whether it is possible to learn a single set of parameters for variable-sized graphs or whether a separate set of parameters must be learned for each graph topology.

Ultimately, the growth of online social networks, the development of improved social tagging systems, and the increasing interconnectivity of the web have the potential to enhance our ability to achieve face recognition at scale. Exploring computational techniques that take advantage of these trends seems a worthwhile endeavor. ■

REFERENCES

[1] R. Scoble, Mark Zuckerberg, Founder of Facebook, and Jet Li, Famous Martial Arts Star. [Online]. Available: http://www.flickr.com/photos/scobleizer/3238530126/

[2] Z. Stone, T. Zickler, and T. Darrell, "Autotagging Facebook: Social network context improves photo annotation," in *Proc. 1st IEEE Workshop Internet Vis.*, 2008, DOI: 10.1109/CVPRW.2008.4562956.

[3] J. P. Phillips, T. W. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 large-scale results," Nat. Inst. Standards Technol., Gaithersburg, MD, Tech. Rep., Mar. 2007.

[4] G. Huang, V. Jain, and E. Learned-Miller, "Unsupervised joint alignment of complex images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, DOI: 10.1109/ICCV.2007.4408858.

[5] M. Everingham, J. Sivic, and A. Zisserman, "Hello! my name is . . . Buffy-automatic naming of characters in TV video," in *Proc. British Mach. Vis. Conf.*, 2006, pp. 899–908.

[6] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, DOI: 10.1109/ICCV.2009.5459250.

[7] L. Wolf, T. Hassner, and Y. Taigman, "Descriptor based methods in the wild," in *Proc. Faces in Real-Life Images Workshop*, Oct. 2008. [Online]. Available: http://hal.inria.fr/inria-00326729/en/

[8] Y. Taigman, L. Wolf, T. Hassner, and I. Tel-Aviv, "Multiple one-shots for utilizing class label information," in *Proc. British Mach. Vis. Conf.*, 2009. [Online]. Available: http://www.bmva.org/bmvc/2009/Papers/Paper391/Paper391.pdf

[9] M. Guillaumin, J. Verbeek, C. Schmid, I. Lear, and L. Kuntzmann, "Is that you? Metric learning approaches for face identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, DOI: 10.1109/ICCV.2009.5459197.

[10] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments Univ. Massachusetts, Amherst, MA, Tech. Rep. 07-49, Oct. 2007.

[11] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, vol. 2, pp. II-848–II-854.

[12] T. Berg, A. Berg, J. Edwards, and D. Forsyth, "Who's in the picture?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 137–144.

[13] S. Satoh and T. Kanade, "Name-it: Association of face and name in video," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 1997, pp. 368–373.

[14] J. Yagnik and A. Islam, "Learning people annotation from the web via consistency learning," in *Proc. Int. Workshop Multimedia Inf. Retrieval*, 2007, pp. 285–290.

[15] M. Zhao and J. Yagnik, "Large scale learning and recognition of faces in web videos," in *Proc. 8th Int. Conf. Autom. Face Gesture Recognit.*, 2008, DOI: 10.1109/AFGR.2008.4813381.

[16] M. Naaman, R. Yeh, H. Garcia-Molina, and A. Paepcke, "Leveraging context to resolve identity in photo albums," in *Proc. 5th ACM/IEEE-CS Joint Conf. Digit. Libraries*, 2005, pp. 178–187.

[17] L. Zhang, L. Chen, M. Li, and H. Zhang, "Automated annotation of human faces in family albums," in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 355–358.

[18] D. Anguelov, K. Lee, S. Gokturk, and B. Sumengen, "Contextual identity recognition in personal photo albums," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, DOI: 10.1109/CVPR.2007.383057.

[19] J. Sivic, C. Zitnick, and R. Szeliski, "Finding people in repeated shots of the same scene," in *Proc. British Mach. Vis. Conf.*, 2006, pp. 909–918.

[20] A. Gallagher, P. Pittsburgh, and T. Chen, "Using group prior to identify people in consumer images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, DOI: 10.1109/CVPR.2007.383492.

[21] A. Gallagher, P. Pittsburgh, and T. Chen, "Understanding images of groups of people," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 256–263.

[22] A. Gallagher and T. Chen, "Estimating age, gender, and identity using first name priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, DOI: 10.1109/CVPR.2008.4587609.

[23] K. Rodden and K. Wood, "How do people manage their digital photographs?" in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2003, pp. 409–416.

[24] L. Weiiyin, S. Dumais, Y. Sun, H. Zhang, M. Czerwinski, and B. Field, "Semi-automatic image annotation," in *Human-Computer Interaction-Interact '01.* Amsterdam, The Netherlands: IOS Press, 2001.

[25] B. Shneiderman and H. Kang, "Direct annotation: A drag and drop strategy for labeling photos," in *Proc. Int. Conf. Inf. Vis.*, 2000, pp. 88–95.

[26] A. Kuchinsky, C. Pering, M. Creech, D. Freeze, B. Serra, and J. Gwizdka, "FotoFile: A consumer multimedia organization and retrieval system," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 1999, pp. 496–503.

[27] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "HT06, tagging paper, taxonomy, Flickr, academic article, to read," in *Proc. Conf. Hypertext Hypermedia*, 2006, pp. 31–40.

[28] M. Ames and M. Naaman, "Why we tag: Motivations for annotation in mobile and online media," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2007, pp. 971–980.

[29] O. Nov, M. Naaman, and C. Ye, "What drives content tagging: The case of photos on Flickr," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2008, pp. 1097–1100.

[30] O. Nov, M. Naaman, and C. Ye, "Motivational, structural and tenure factors that impact online community photo sharing," in *Proc. AAAI Int. Conf. Weblogs Social Media*, 2009. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/206/426

[31] Facebook Statistics. [Online]. Available: http://www.facebook.com/press/info.php?statistics

[32] Engineering @ Facebook's Notes. [Online]. Available: http://www.facebook.com/note.php?note_id=76191543919

[33] Open Source Computer Vision Library (OpenCV). [Online]. Available: http://www.intel.com/research/mrl/research/opencv/

[34] P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. 511–518.

[35] K. R. Harsha, My Grandparents and Me! [Online]. Available: http://www.flickr.com/photos/mynameisharsha/3286864307/

[36] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 282–289.

[37] M. Collins, "Discriminative training methods for hidden Markov models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2002, DOI: 10.3115/1118693.1118694.

[38] B. Taskar, C. Guestrin, and D. Koller, "Maximum-margin Markov networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003. [Online]. Available: http://books.nips.cc/nips16.html

[39] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, vol. 6, pp. 1453–1484, Sep. 2005.

[40] G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, and S. Vishwanathan, *Predicting Structured Data.* Cambridge, MA: MIT Press, 2007.

[41] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications.* Cambridge, U.K.: Cambridge Univ. Press, 1994.

[42] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proc. 12th Int. Conf. Inf. Knowl. Manage.*, 2003, pp. 556–559.

[43] A. Clauset, M. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 70, no. 6, 2004, DOI: 10.1103/PhysRevE.70.066111.

[44] A. Mayer and S. Puller, "The old boy (and girl) network: Social network formation on university campuses," *J. Public Econom.*, vol. 92, no. 1–2, pp. 329–347, 2008.

[45] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis, "Tastes, ties, and time: A new social network dataset using Facebook.com," *Social Netw.*, vol. 30, no. 4, pp. 330–342, 2008.

[46] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1568–1583, Oct. 2006.

[47] P. Kohli, A. Shekhovtsov, C. Rother, V. Kolmogorov, and P. Torr, "On partial optimality in multi-label MRFs," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 480–487.

[48] T. Finley and T. Joachims, "Training structural SVMs when exact inference is intractable," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 304–311.

[49] SVMstruct. Support vector machine for complex outputs. [Online]. Available: http://svmlight.joachims.org/svm_struct.html

[50] A. Gallagher and T. Chen, "Estimating age, gender and identity using first name priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, DOI: 10.1109/CVPR.2008.4587609.

[51] A. Friedman, "Framing pictures: The role of knowledge in automatized encoding and memory for gist," *J. Exp. Psychol., Gen.*, vol. 108, pp. 316–355, 1979.

[52] A. Oliva and A. Torralba, "Modeling the shape of a scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[53] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 2169–2178.

[54] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 2, pp. 524–531.

[55] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in *Proc. Eur. Conf. Comput. Vis.*, 2006, vol. 4, pp. 517–530.

[56] Y. Song and T. Leung, "Context-aided human recognition-clustering," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 382–395.

[57] T. Joachims, T. Finley, and C.-N. Yu, "Cutting-plane training of structural SVMs," *Mach. Learn.*, vol. 77, pp. 27–59, 2009.

[58] C. Teo, A. Smola, S. Vishwanathan, and Q. Le, "A scalable modular convex solver for regularized risk minimization," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2007, pp. 727–736.

[59] D. Sontag and T. Jaakkola, "New outer bounds on the marginal polytope," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 21, pp. 1393–1400, 2007.

[60] N. Komodakis, N. Paragios, and G. Tziritas, "MRF optimization via dual decomposition: Message-passing revisited," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, DOI: 10.1109/ICCV.2007.4408890.

[61] N. Komodakis and N. Paragios, "Beyond loose LP-relaxations: Optimizing MRFs by repairing cycles," in *Proc. Eur. Conf. Comput. Vis. III*, 2008, pp. 806–820.

ABOUT THE AUTHORS

Zak Stone (Student Member, IEEE) received the A.B. degree in physics and mathematics and the S.M. degree in computer science from Harvard University, Cambridge, MA, in 2004 and 2009, respectively, where he is currently working towards the Ph.D. degree in computer science under the direction of Dr. T. Zickler.

His research interests include object recognition and scene classification in large-scale image collections, social network analysis, and scientific visualization. His research has been supported by a Harvard Peirce Fellowship and a National Science Foundation Graduate Research Fellowship.

Todd Zickler (Member, IEEE) received the B.Eng. degree in honors electrical engineering from McGill University, Montreal, QC, Canada, in 1996 and the Ph.D. degree in electrical engineering from Yale University, New Haven, CT, in 2004, under the direction of P. Belhumeur.

He joined the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, as an Assistant Professor in 2004 and was appointed a John L. Loeb Associate Professor of the Natural Sciences in 2008. He is the Director of the Harvard Computer Vision Laboratory, and his research is focused on modeling the interaction between light and materials and developing algorithms to extract scene information from visual data. His work is motivated by applications in face, object, and scene recognition; image-based rendering; content-based image retrieval; image and video compression; robotics; and human-computer interfaces.

Dr. Zickler is a recipient of the National Science Foundation Career Award and a Research Fellowship from the Alfred P. Sloan Foundation. His research is funded by the National Science Foundation, the Army Research Office, the Office of Naval Research, and the Sloan Foundation.

Trevor Darrell (Member, IEEE) received the B.S.E. degree in computer science from the University of Pennsylvania, Philadelphia, in 1988 and received the S.M. and Ph.D. degrees in media arts and sciences from the Massachusetts Institute of Technology (MIT) Media Lab, Cambridge, MA, in 1992 and 1996, respectively.

Having started his career in computer vision as an undergraduate researcher in Ruzena Bajcsy's GRASP lab, he was a member of the research staff at Interval Research Corporation from 1996 to 1999. He served on the faculty of the MIT Electrical Engineering and Computer Science Department from 19996 to 2008, where he directed the Vision Interface Group. He now leads the newly formed Computer Vision Group at the International Computer Science Institute and is on the faculty of the Computer Science Division at the University of California Berkeley. His group develops algorithms to enable multimodal conversation with robots as well as mobile devices and methods for object and activity recognition on such platforms. His interests include computer vision, machine learning, computer graphics, and perception-based human-computer interfaces.