InSight: Recognizing Humans without Face Recognition

He Wang
Duke University

Xuan Bao
Duke University

Romit Roy Choudhury
Duke University

Srihari Nelakuditi
University of South Carolina

ABSTRACT

Wearable cameras and displays, such as the Google Glass, are around the corner. This paper explores techniques that jointly leverage camera-enabled glasses and smartphones to recognize individuals in the visual surrounding. While face recognition would be one approach to this problem, we believe that it may not be always possible to see a person's face. Our technique is complementary to face recognition, and exploits the intuition that colors of clothes, decorations, and even human motion patterns, can together make up a "fingerprint". When leveraged systematically, it may be feasible to recognize individuals with reasonable consistency. This paper reports on our attempts, with early results from a prototype built on Android Galaxy phones and PivotHead's camera-enabled glasses. We call our system *InSight*.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software; C.2.4 [**Computer-Comunication Networks**]: Distributed Systems

General Terms

Design, Experimentation, Performance

Keywords

Wearable camera, visual fingerprinting, smartphones, augmented reality, matching, recursion, distributed cameras

1. INTRODUCTION

Imagine a near future where humans are carrying smartphones and wearing camera-embedded glasses, such as the Google Glass. This paper intends to recognize a human by looking at him or her from any angle, even when her face is not visible. For instance, Alice may look at people around her in a social gathering and see the names of each individual – like a virtual badge – suitably overlaid on her Google Glass display. Where revealing names is undesirable, only a tweet message could be shared. People at the airport could tweet "looking to share a cab ", and Alice could view each individual's tweet above their heads. In general, we intend to extend augmented reality [1, 2] to humans and the key challenge pertains to differentiating individuals. We explore options outside face recognition [3, 4].

Our core technique exploits the intuition that faces are not necessarily the only "visual fingerprint" of an individual. Features combined from clothing colors, body structure, and motion patterns can potentially be fingerprints for many practical scenarios. There is evidence of this opportunity given that humans can often recognize another human without looking at her face. This paper asks: *can sensor-enabled smartphones and wearable glasses together achieve the same?*

Our main idea is simple and illustrated through Figure 1. Whenever a user Bob uses his phone (such as while checking emails), the phone's camera opportunistically "takes a peek" at Bob. Through image segmentation and processing [5, 6] the phone extracts a visual fingerprint – a feature vector that includes clothing color and their spatial organization. The spatial organization captures the relative locations of each color in 2D space, hence a red over blue shirt is different from blue over red. This spatio-chromatic information – called Bob's *self-fingerprint* – is announced in the vicinity. Nearby smartphones receive a tuple:

$$S_{Bob} = \ <\textit{Bob, Bob's self-fingerprint}>.$$

Now consider Alice (wearing a Google Glass and carrying a smartphone) looking at a group of people that includes Bob. A picture from the glass is processed on Alice's phone (or in the cloud), and through image segmentation and analysis, the phone computes each individual's spatio-chromatic fingerprint, F_i. Since Alice has separately received Bob's self-fingerprint, S_{Bob}, a matching algorithm computes the similarity between F_i and S_{Bob}. If one of the fingerprints, F_j matches strongly with S_{Bob}, then Alice's phone can recognize Bob against the group of people. An arrow labeled "Bob" can now be overlaid on the image segment that generated F_j; Alice can view this either on her Google Glass display or on her smartphone screen (as shown in Figure 2).

Realizing the above idea presents a number of challenges. Bob's self-fingerprint is likely to capture only some parts of his clothing, and may not be adequately discriminating (particularly when Alice views Bob from the back, or when Bob is partially visible in the crowd). Even if front and back fingerprints are somehow available, and Bob is fully visible, ambiguity can arise when people are wearing similar dresses, say in a birthday party

ACM HotMobile'13, February 26–27, 2013, Jekyll Island, Georgia, USA.

Figure 1: Sketch of *InSight*: Bob's phone announces his own name and fingerprint to the vicinity; Alice's phone computes fingerprints from her glass, matches them against the ones received from the vicinity, and recognizes Bob.

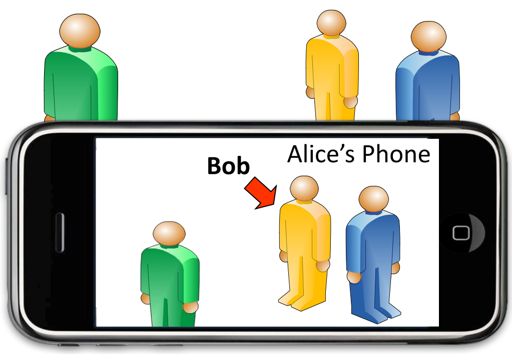

Figure 2: An arrow labeled "Bob" overlaid on Bob in the smartphone's screen.

with a dress theme. Finally, even when all is well, different lighting conditions, wrinkles on clothes, and human mobility can inject errors into the system. Coping with these challenges are indeed non-trivial, however, we believe that certain opportunities can help, as described next.

(1) Even if Bob's self-fingerprint is not highly discriminating, Charlie may luckily identify Bob when Bob happens to be alone or around a few people wearing sharply contrasting clothes. Since Charlie now sees Bob's full clothing, he could enrich Bob's self-fingerprint with more information and upload it to the cloud. The enriched self-fingerprint helps others discriminate Bob better, which in turn enables more frequent opportunities to enrich his fingerprint – *a recursive process*. We find that the system converges in our limited set of experiments, enriching almost everyone's self-fingerprints.

(2) If the system still contains visual ambiguity, we observe that short term motion can be exploited. When Bob moves at a certain pace in a certain direction, Bob's accelerometer and compass could compute his motion vector and include it in his self-fingerprint. Alice could compute a similar motion vector from a short Google Glass video (by comparing a few consecutive video frames [7]), and match against all the motion vectors embedded in received self-fingerprints. If the best match corresponds to Bob, then Alice can disambiguate Bob despite visual similarity.

This paper is targeted to harness these opportunities. Our early prototype, built on Android phones and PivotHead camera-equipped glasses, implements basic self fingerprinting and matching. Offline experiments with 15 people in a clique yield promising results – 93% of correct recognition when viewed from the front. When viewed from the back, the accuracy degrades sharply. However, when different views are used to recursively enrich fingerprints (implemented via *monte carlo* simulations), the system converges to 96% accuracy even when viewed from the back; the front-side accuracy is perfect, and the convergence time is not long. Overall, *InSight* is an autonomous, self-correcting scheme, much different from a crude color matching idea proposed in our earlier work [8]. If successful, *InSight* could perhaps trigger new thinking in human-centric augmented reality applications [9].

2. SYSTEM SETTING

InSight assumes that Bob uses his smartphone to check emails or browse the Internet. When the phone is held in a specific orientation and the display senses finger taps – partly ensuring that the front-facing camera is facing Bob's upper body – the phone takes a few opportunistic pictures[1]. The pictures are analyzed, visual fingerprints extracted, and concatenated with the name "Bob", or any content/tweets that Bob intends to make visible. This self-fingerprint is either announced to the vicinity via Bluetooth Low Energy (BLE) or transmitted to the cloud along with the Bob's rough location. With BLE beacons, nearby phones directly receive the fingerprint. For cloud-based access, all phones update the cloud with their locations; the cloud matches the fingerprints and pushes back the recognition results. While both approaches present tradeoffs, we use the cloud based approach. As we will see later, the cloud based approach allows a central repository of fingerprints that can be distributedly updated by different people over time.

A viewer Alice looks at different people, and when in need to recognize a person, presses a button on her camera-enabled glass. A short video – of around 3 seconds – is recorded and transferred to her smartphone via WiFi or BLE, whichever is available on the glass. In the default case, the phone processes one of the frames in this video, separates different individuals in this image, and extracts visual fingerprints corresponding to each of them. For each computed fingerprint (sent to the cloud), the cloud computes a "similarity" with Bob's self-fingerprint. When the similarity is greater than a high confidence threshold,

[1] We discuss privacy issues in Section 5.

the cloud identifies Bob – Alice's phone superimposes an arrow on her phone screen or the glass display. When the similarity is sub-threshold, InSight explores motion patterns (speed and walking direction) for better recognition.

A natural question might be: *why not utilize the user's location as a form of identification?* While this is indeed a possible approach, we believe that such precise location in indoor spaces is unavailable today. Moreover, it may not be easy to extract depth information from the video, i.e., if Alice is looking down a corridor, its unclear what "location" she is looking at. Finally, locations need to be combined with compasses to compute the line of sight of the viewer; given compasses have a reasonably large error, especially indoors, pure location-based solutions may not suffice. However, location and compasses can be used to narrow down the search space for visual recognition. While we don't leverage this opportunity (to understand the limits of our techniques), we certainly intend to optimize InSight with location, compass, and face recognition in the future.

3. SYSTEM DESIGN

We sketch the basic design decisions underlying *InSight*. Several deliberations underpinning these decisions are curtailed in the interest of space.

3.1 Extracting Self-Fingerprints

Figure 3(a) shows an example photo taken opportunistically by InSight[2]. The key task here is to extract a visual self-fingerprint that is robust to lighting conditions, viewing angle, and viewing distance. Put differently, even if different people look at the same person from different positions (Figure 3(b)), the fingerprint from all these views should reasonably match the self-fingerprint. As a first step, InSight automatically crops out a rectangular region from picture – the part below the face/neck. It then applies two well known techniques on the cropped image, namely (1) spatiograms, and (2) wavelets.

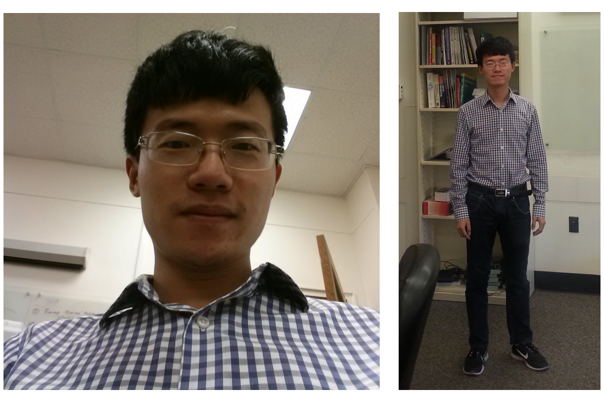

Figure 3: (a) Upper body view when user browsing on his smartphone (b) View from user wearing a glass.

(1) Spatiograms: Spatiograms are essentially color histograms with spatial distributions encoded in its structure. Put differently, while basic color histograms only capture the relative frequency of each color, spatiograms capture how these colors are distributed in 2D space. The second order of spatiogram can be represented as [10]:

$$h_I(b) = < n_b, \mu_b, \sigma_b >, \quad b = 1,2,3,\cdots,B,$$

where B is the number of color bins, n_b is the number of pixels whose value falls in the b^{th} bin, and μ_b and σ_b are the mean vector and covariance matrices of the coordinates of those pixels respectively. Through such a representation, a white over red stripe can be distinguished from a red over white stripe, even if the number of red and white pixels are identical in both. Also, to cope with various viewing distances, we normalize the spatial information with respect to the *shoulder width* so that all the spatial representation is relative to the captured body size in each photo. Finally, to decouple lighting conditions from the colors, we convert the pixels from RGB to HSV, and quantize them into $B = 10x4x4$ bins.

(2) Wavelets: Apparels are often fashioned with patterns that run horizontally, vertically, or along diagonals. InSight captures them by computing the energy distribution over wavelet sub-bands [11, 12] along the vertical (E_v), horizontal (E_h) and diagonal (E_d) dimensions. As a result, different organizations of edges exhibit distinct feature vectors in our representation (Figure 4). We also use the ratio between E_v and E_h to improve robustness against different viewing distances. This is because viewing from afar usually leads to loss in resolution, which implies fewer detected edges. However, since this lossy behavior affects vertical and horizontal stripes equally, the ratio between E_v and E_h remains almost unchanged.

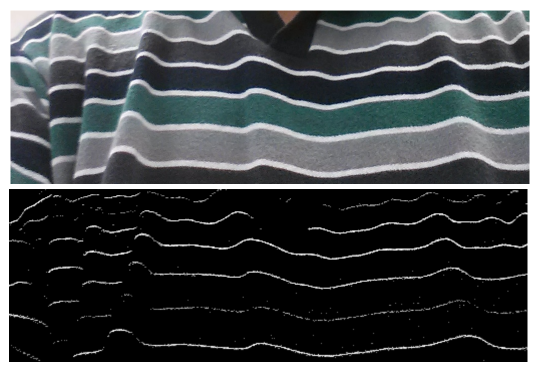

Figure 4: (a) Image for self-fingerprint (b) corresponding energy over wavelet sub-band along horizontal axis.

3.2 Extracting Fingerprints from Glass View

Bob's self-fingerprint is a combination of the spatiogram and wavelet representations. Later, when Alice views Bob through her glass – either from the front or from the back – InSight again crops out a rectangular region around Bob's upper body (below the face/neck), and applies the same fingerprinting operations on this image. These fingerprints – one from Bob and another from Alice – are now ready for matching.

3.3 Fingerprint Matching

Our matching algorithm first computes the spatiogram similarity between each person in Alice's view with the given self-fingerprint (from Bob). Denote the spatiograms to be compared as $S = \{n, \mu, \sigma\}$ and $S' = \{n', \mu', \sigma'\}$, both having B color bins. We define the similarity measure as [13]:

$$\rho = \sum_{b=1}^{B} \sqrt{n_b n_b'}\, 8\pi |\Sigma_b \Sigma_b'|^{1/4} \mathcal{N}(\mu_b; \mu_b', 2(\Sigma_b + \Sigma_b'))$$

[2] Recall that this occurs when the accelerometer senses that the phone is at an appropriate angle, and the user is typing.

Essentially, the similarity decreases (following a Gaussian function) with increasing difference between the colors and their spatial locations.

Following this, we dynamically train a model using the wavelet features of the same two fingerprints. The classifier in use is a bagged decision tree (BDT). The BDT selects random samples from the training data, builds multiple decision trees (each with a subset of samples), and eventually chooses a weighted majority voting result as the final output. The classification results are accompanied by confidence scores that quantify the uncertainty. In the end, the algorithm combines the similarity values from spatiograms with the confidence-scores from wavelet classifiers, and selects a candidate whose confidence exceeds a high threshold. When the confidence is below this threshold, our current system declares “unsure”, an attempt to minimize incorrect recognition.

3.4 Refining the Self-Fingerprint

Bob’s self-fingerprint is derived from a sliver of his dress, and may not be adequately discriminating against a background of many individuals. Moreover, Alice may view Bob from his back, and this “back fingerprint” may not match well with Bob’s self-fingerprint (derived from his front view). This could be due to differing patterns at the back of Bob’s shirt; differing wrinkles; and/or unequal lighting conditions. We identify opportunities to consolidate front and back fingerprints, which in turn can improve the robustness of recognizing Bob. Our core intuition exploits natural human mobility as described next.

Consider a social gathering where humans are naturally walking around, to the extent that from any camera view, people in Bob’s background changes over time. This *diversity* of backgrounds is likely to become contrasting to Bob at some point. In other words, even if Bob’s self-fingerprint is not highly discriminating, in certain favorable situations, the fingerprint may suffice because people in Bob’s background are wearing different colored clothes. At this point in time, since Charlie may be able to recognize Bob and actually see his full attire (through her glasses), he can enrich Bob’s fingerprint. Enriching would entail informing the cloud that Bob’s fingerprint should be updated with spatiogram and wavelet features derived from his trousers, center of the shirt, etc. Later, if Julie happens to view Bob from the back, this enriched fingerprint may now help recognize Bob (perhaps because the trouser colors are discriminating). This can in turn lead to further enrichment – Bob’s fingerprint can now be updated with the visual features of his back.

Over time, one may envision everyone’s fingerprint getting enriched, which improves recognition, which in turn facilitates enrichment. This recursive process may eventually converge, resulting in fairly unique fingerprints for almost all. Our controlled experiments in the next section will endorse this intuition and indicate room for improvement.

4. EVALUATION

We implement a prototype of InSight using PivotHead camera enabled glasses (Figure 5) and Samsung Galaxy phones running Android. We conduct experiments with 15 users dressed in their regular attires. We explicitly asked these participants to actively use their smartphones. Each phone opportunistically takes “profile” pictures of the user. In this process, InSight selects the most suitable pictures via angle detection using accelerometer readings, face detection, and blur estimation. The automatically chosen pictures are then used to form the “self-fingerprint” for the user. The PivotHead glass was worn by a single user who captured all the other users from the front and from the back. In our preliminary experiments, the users captured in the glass view do not overlap with each other – we controlled the experiment in this manner for the purpose of simplicity.

Our main findings may be summarized as follows: (1) We confirm that color spatiograms and pattern wavelets capture complementary features of a person’s dress – together, they are effective in discriminating an individual from the rest 14. (2) When people are facing the glass, they can be accurately recognized using their self-fingerprints. (3) Through *monte carlo* simulations on real fingerprints, we demonstrate how recursive fingerprint refinement can help recognize a person, even when she is facing away from the glass.

Figure 5: PivotHead camera glasses used for InSight.

4.1 Combining Colors and Patterns

To assess the discriminative abilities of color and pattern features, we first evaluate them separately when 15 people are facing the glass – we extract their features from the glass view. Figure 6(a) shows the confusion matrix represented using a heat map. Element ij of the matrix reflects the similarity score when InSight compares the spatiogram corresponding to person i in the glass-view with that of self-fingerprint of person j. A lighter color indicates higher similarity, and vice versa. If diagonal elements are much lighter than the rest, then spatiograms alone may be considered discriminative. With Figure 6(a), this is true for 80% of the cases.

Figure 6(b) reflects the confidence scores in the confusion matrix, when wavelets are used to extract features from clothing patterns – the recognition accuracy is 73%. While this is not high, we find that spatiograms and wavelets exhibit complementary behavior (compare failure cases in Figure 6(a) and Figure 6(b)). When color spatiograms fail to differentiate two people, pattern wavelets are able to distinguish them. Therefore, we combine these two approaches by computing the product of their similarity and confidence scores. Figure 6(c) presents the result of this hybrid approach. The accuracy improves distinctly; 14 out of 15 people are recognized without ambiguity. The rest of the evaluation employs this hybrid approach.

4.2 Performance with Self-Fingerprints

Since self-fingerprints capture the front view of a person from a close range, its important to characterize whether they are effective when others view the person from a distance, and

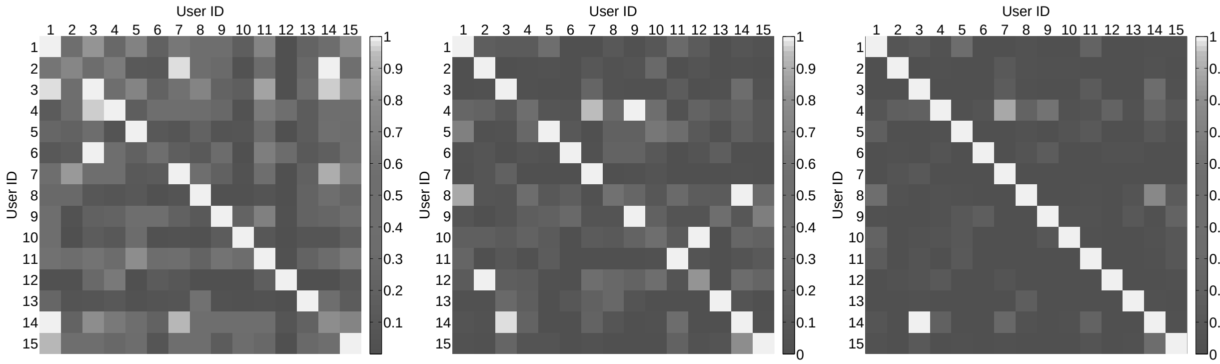

Figure 6: All users facing the glass: (a) Similarity scores from spatiograms; (b) Confidence scores from classification using wavelet features; (c) The effect of combining color with patterns.

sometimes from the back. To address this question, we evaluate the discriminative power of InSight by varying the number of users facing towards and away from the glass. We conduct experiments with *all possible user combinations*, ranging from only 1 user to all 15 users.

Figure 7 evaluates scenarios when all users are facing the glass. For scenarios with n users (on the x axis), the graph shows the average percentage of users correctly recognized, falsely recognized, and unrecognized. The average is computed over all the possible combinations, e.g., 105 combinations in case of 2 users in the view. Evidently, the accuracy drops slightly (from 100% to 93%) from the 1-user to the 15-user scenario. None of the users are recognized incorrectly as someone else. This suggests that when Bob is facing Alice, self-fingerprints may be adequate for reliable human recognition.

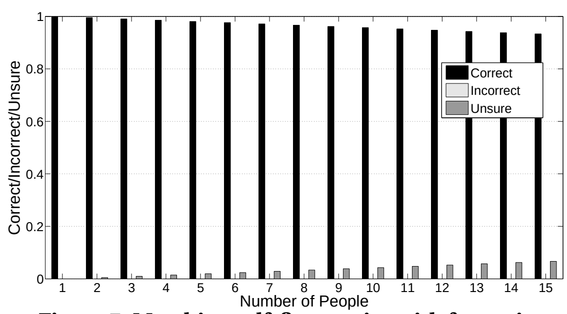

Figure 7: Matching self-fingerprint with front view

Now consider the case where users are facing away from the glass, but their self-fingerprints used to recognize them. Figure 8 shows the results. In most cases, InSight is unable to recognize individuals, particularly when there are many users. This is not too surprising since the glass only captures users' back views and needs to compare them with the self-fingerprints taken from the front. However, a positive outcome is that very few users are incorrectly recognized. Moreover, when there are only few people around, some of them can be recognized from their back view. Next, we describe how these few instances can be leveraged to bootstrap "fingerprint refinement", such that even back-views can discriminate people in a crowded situation.

4.3 Performance with Refined Fingerprints

Consider those few lucky instances when Alice recognizes Bob even though Bob has his back facing Alice (note, these instances are more probable when few people are around). Knowing that

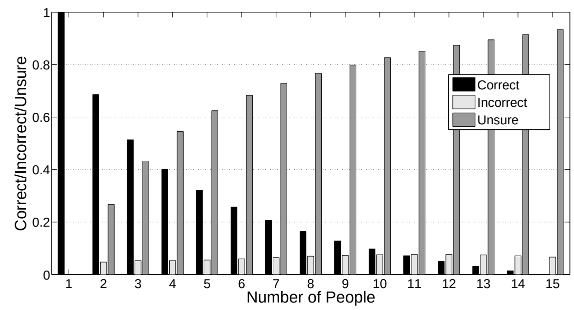

Figure 8: Matching self-fingerprint with back view

this is Bob's back view, InSight can refine Bob's fingerprint (i.e., the cloud updates Bob's self-fingerprint to also contain features from the back-side of his dress). This refinement is feasible only because InSight is rarely wrong in recognizing people (when unsure, InSight refrains from making a recognition). Thus, if Bob is recognized once, Bob can be recognized quite accurately thereafter even in a crowded place. This is regardless of whether his front or back is facing the glass.

To validate the fingerprint refinement approach, we conduct the following monte carlo simulation. We randomly choose 4 people with their backs facing the glass. We compare each of their back views with their self-fingerprints. When there is a strong match, the corresponding <ID, back view> is added to the InSight system. This step is repeated 200 times, and over time more such back views get added to the system. Once the same ID gathers 5 or more back views, we pick the most dominant one as that ID's back fingerprint. Gradually, the accuracy of recognizing a person with a back view improves since it would be compared to back fingerprints when available. Even when Bob's back fingerprint is not in the system, back fingerprints of others help narrow down the search space considerably, enhancing the chances of recognizing Bob using his front-side fingerprint. We perform 500 runs of this simulation and show the average results in Figure 9. The system converges, and increasing number of users get recognized correctly over time, even when all of them have their back facing the glass. Some errors indeed occur, but they do not propagate in the system due to the overwhelming number of correct recognitions.

5. DISCUSSION

Many more challenges need to be addressed, several opportunities need to be exploited. We discuss a few here.

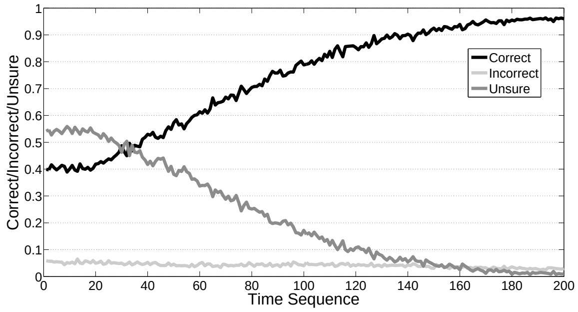

Figure 9: Matching back view after refining fingerprints.

(1) Incremental Deployment. Non-participants of this system – those not running InSight – are likely to be mis-recognized (even though ideally they should be labeled "unknown"). While this is indeed a problem, work arounds may be feasible. If Alice views Charlie but finds that none of the announced fingerprints match with him, then Alice can suspect that Charlie is not a part of the system. Over time, this suspicion could grow as more people are unable to recognize Charlie, eventually tagging Charlie as "unknown". Of course, if Charlie is wearing a dress similar to Bob, then the situation is harder. InSight has to wait for adequate opportunities to find Charlie separate from Bob, so that the suspicion value rises above a threshold – a key challenge for our ongoing work.

(2) Privacy of Opportunistic Pictures. Taking opportunistic pictures, while creating the self-fingerprint, may raise privacy issues. Although the camera takes pictures only at specific orientations, we believe that a concerned user can choose to manually create the self-fingerprint. For instance, if InSight is used occasionally – only at certain conferences or events, or when the user needs to broadcast a message – it may not be burdensome to take a self-picture only at those times. However, regular use may call for additional privacy precautions; one simple way could be to show the automatically-taken picture to the user before using it for fingerprinting.

(3) Overlapping Users in View. When one views a crowded gathering, it may not be easy to crop out each individual from the image. People may be overlapping in their views, and only a small part of their dresses may be visible. InSight will need to evaluate the impact of such complicated views of individuals, especially in crowded settings.

(4) Application scenarios. InSight enables use-cases in which a user Bob intends to convey some information to anyone who visually looks at him. One may view this as a form data broadcast using a visual "source address"; the recipients are all users whose line of sight intersects with the transmitter. Specific instances in which such visual broadcasts are applicable include virtual badges in a conference, students tweeting about their areas of interest in a job fair, etc. Also, several use-cases do not require revealing the user's identity – a person at a basketball stadium can simply tweet "selling an extra ticket for tonight's game". One may even view InSight as a way of social messaging, similar to how people where T-shirts with interesting captions on them.

6. CONCLUSION

This paper pursues a hypothesis that colors and patterns on clothes may pose as a human fingerprint, adequate to discriminate one individual from another in low/moderate density situations. If successful, such a fingerprint could be effectively used towards human recognition or content announcement in the visual vicinity, and more broadly towards enabling human-centric augmented reality. Pivoted on this vision, we develop a proof of concept – *InSight* – using which users create a visual fingerprint of an individual and compare it with that individual's self-created fingerprint. Preliminary evaluation with 15 people wearing natural clothes, suggest promise – we find that clothes indeed exhibit good entropy, and can be automatically fingerprinted/matched with reasonable accuracy. Our ongoing work is focussed on coping with the issue of incremental deployment, as well as exploring motion patterns when visual fingerprints are not unique. We believe there is promise, and are committed to building a fuller, real-time, system.

7. ACKNOWLEDGMENT

We sincerely thank our shepherd, Dr. Stefan Saroiu, as well as the anonymous reviewers for their valuable comments. This work was supported in part by the NSF grant IIS:0910846.

8. REFERENCES

[1] D. Wagner, G. Reitmayr, A. Mulloni, T. Drummond, and D. Schmalstieg, "Real-time detection and tracking for augmented reality on mobile phones," *IEEE Visualization and Computer Graphics*, 2010.

[2] P. Mistry and P. Maes, "Sixthsense: a wearable gestural interface," in *ACM SIGGRAPH ASIA 2009 Sketches*, 2009, p. 11.

[3] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys (CSUR)*, 2003.

[4] M.A. Turk and A.P. Pentland, "Face recognition using eigenfaces," in *IEEE CVPR*, 1991.

[5] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE TPAMI*, 2002.

[6] Andrew C. Gallagher and Tsuhan Chen, "Clothing cosegmentation for recognizing people," in *IEEE CVPR*, 2008.

[7] C. Qin, X. Bao, R. Roy Choudhury, and S. Nelakuditi, "Tagsense: a smartphone-based approach to automatic image tagging," in *ACM MobiSys*, 2011.

[8] Ionut et. al. Constandache, "Did you see bob?: human localization using mobile phones," in *ACM*, 2010, MobiCom '10.

[9] A.T. Campbell, S.B. Eisenman, N.D. Lane, E. Miluzzo, R.A. Peterson, H. Lu, X. Zheng, M. Musolesi, K. Fodor, and G.S. Ahn, "The rise of people-centric sensing," *IEEE Internet Computing*, 2008.

[10] S.T. Birchfield and S. Rangarajan, "Spatiograms versus histograms for region-based tracking," in *IEEE CVPR*, 2005.

[11] B.S. Manjunath and W.Y. Ma, "Texture features for browsing and retrieval of image data," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1996.

[12] Y. Tian and S. Yuan, "Clothes matching for blind and color blind people," *Computers Helping People with Special Needs*, 2010.

[13] CO Conaire, N.E. O Connor, and A.F. Smeaton, "An improved spatiogram similarity measure for robust object localisation," in *IEEE ICASSP*, 2007.