TagSense: A Smartphone-based Approach to Automatic Image Tagging

Chuan Qin[†§*] qinc@cse.sc.edu
Xuan Bao[§] xuan.bao@duke.edu
Romit Roy Choudhury[§] romit.rc@duke.edu
Srihari Nelakuditi[†] srihari@cse.sc.edu

[†]University of South Carolina
Columbia, SC, USA

[§]Duke University
Durham, NC, USA

ABSTRACT

Mobile phones are becoming the convergent platform for personal sensing, computing, and communication. This paper attempts to exploit this convergence towards the problem of automatic image tagging. We envision *TagSense*, a mobile phone based collaborative system that senses the people, activity, and context in a picture, and merges them carefully to create tags on-the-fly. The main challenge pertains to discriminating phone users that are in the picture from those that are not. We deploy a prototype of TagSense on 8 Android phones, and demonstrate its effectiveness through 200 pictures, taken in various social settings. While research in face recognition continues to improve image tagging, TagSense is an attempt to embrace additional dimensions of sensing towards this end goal. Performance comparison with Apple iPhoto and Google Picasa shows that such an out-of-band approach is valuable, especially with increasing device density and greater sophistication in sensing/learning algorithms.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software; C.2.4 [**Computer-Communication Networks**]: Distributed Systems; H.5.5 [**Information Interfaces and Presentations**]: Sound and Music Computing

General Terms

Design, Experimentation, Performance

Keywords

Image Tagging, Face Recognition, Sensing, Smartphone, Context-awareness, Activity Recognition

*This research was conducted while Chuan Qin was visiting systems and networking research group at Duke University.

MobiSys'11, June 28–July 1, 2011, Bethesda, Maryland, USA.

1. INTRODUCTION

Automatic image tagging has been a long standing problem. While the fields of image processing and face recognition have made significant progress, it remains difficult to automatically label a given picture. However, digital pictures and videos are undergoing an explosion, especially with the proliferation of high quality digital cameras embedded in mobile devices. As these pictures get stored in online content warehouses, the need to search and browse them is becoming crucial [1]. Furthermore, the growing sophistication in textual search is raising the expectations from image retrieval – users are expecting to search for pictures as they do for textual content. Efforts to engage humans for labeling pictures (with crowdsourcing or online gaming [2–5]) may be a stop-gap solution, but is not likely to scale in the longer run. The volume of content is growing at dramatic speeds, and its dependence on a pair of human eyes is likely to become the bottleneck.

This paper breaks away from established approaches to image tagging, and explores an alternative architecture rooted in multi-dimensional, out-of-band sensing. The core idea is simple. Consider a scenario in which Bob is taking a picture of his friends. Bob's phone establishes a short-lived wireless connection with all his friends' phones, and instructs them to activate their sensors for a short time-window around the photo-click. The sensors sense the "moment", summarize the measured data, and communicate them back to Bob's phone. Bob's phone processes this data to identify which of the individuals are in the picture, their activities, and other contextual tags about the occasion. These tags are systematically organized into a "`when-where-who-what`" format, ultimately creating an automatic description for the picture.

If done well, automatic image tagging can enable a variety of applications. One may imagine improved image search in the Internet, or even within one's own computer – Bob may query his personal photo collection for all pictures of Alice and Eve playing together in the snow. Another application may tag videos with important event/activity markers; a user of this application may be able to move the video slider to the exact time-point where President Obama actually walks up to the podium, or starts speaking. Today, such functionalities may be available in select images and videos, where some humans have painstakingly tagged them [2]. TagSense aims to automate this process via sensor-assisted tagging.

A natural question is *why should sensor-assisted tagging be any easier or better than image processing/face recognition?* We believe this may be true because the different sensors are likely to capture the "moments" across multiple sensing dimensions. Laughing may be more naturally detectable via the microphone; dancing may exhibit an accelerometer signature; light sensors may easily discern between indoor and outdoor environments [6–8]. Further, people in the picture may have the direction of their smartphone's compass opposite to that of the camera's compass; those that pose for the picture may exhibit a motion signature through their accelerometers. Recognizing all these attributes through image processing alone may be difficult. The diversity offered by multiple sensing dimensions may allow TagSense to "cast a wider net" – the chances of capturing the individuals/actions, over at least one of these dimensions, is likely to be higher.

Of course, it is necessary to clarify the notion of "tags" here. We do not fix the meaning of tags to be those that can only be obtained visually, by looking at an image or a video. Instead, we define them to be keywords that describe the ongoing scenario/event/occasion during which the picture was taken. For instance, "`noisy cafe`" may be a valid tag, even though nothing in the picture visually suggests that the place is noisy. Similarly, Eve may not be fully visible in a picture because she is hiding behind Alice, yet, tagging the picture with Eve is valid as per our interpretation.

Translating the overall vision into a deployable system entails a number of challenges. (1) TagSense needs to identify the individuals in the picture – since Bob's phone gathers sensed information from all phones within wireless range, it is unclear which of the phones were in the picture. (2) Sensor readings gathered from different phones need to be mined to identify activities and contextual information. (3) The energy budget for sensing, communicating, and computing needs to be optimized to facilitate wide-scale adoption.

The goal of this paper is to address these challenges, and consolidate them into an extensible system framework (with provisions to plug-in image processing and crowd-sourcing modules). We focus primarily on the first problem – identifying the people that are in the picture – and draw on multi-dimensional sensing to address it. For instance, we find that when people pose for a picture, their stillness during the pose presents a unique signature on the accelerometer. We also observe a relationship between the camera's compass direction and the compass directions of the subjects. Finally, for pictures with moving subjects, we find that it is possible to infer their motion through a sequence of camera snapshots and correlate them against accelerometer/compass measurements. Harnessing these opportunities, while coping with practical challenges like variable phone orientation, forms the crux of TagSense. Towards a complete system, we also generate activity/contextual tags by drawing on established activity recognition and acoustic signal processing algorithms.

Our main contributions in this paper may be summarized as:

- **Envisioning an alternative, out-of-band opportunity towards automatic image tagging.** We believe that this opportunity will be catalyzed by the growing proliferation of camera-equipped smartphones, and concurrent advances in personal sensing.
- **Designing TagSense, an architecture for coordinating the mobile phone sensors, and processing the sensed information to tag images.** The diversity in multi-dimensional sensing helps overcome problems that are otherwise difficult on a single dimension.
- **Implementing and evaluating TagSense on Android NexusOne phones.** Compared to face recognition capabilities in Apple iPhoto and Google Picasa, TagSense exhibits a fairly good *precision*, and a significantly higher *recall*. Moreover, activity tags with TagSense are far more relevant than Google Goggles [9], giving us confidence to pursue TagSense as a long-term research project.

The rest of this paper expands on each of these contributions, beginning with the problem space, and followed by the system overview, design, and evaluation.

2. PROBLEM SPACE

This section introduces TagSense with an example, and uses it to describe the problem landscape.

Figure 1 shows three pictures labeled by TagSense. The left and right were taken while our research group got together in the Duke University's Wilson Gym, and later again at the Nasher Museum. The middle picture was taken outside the Hudson Hall while snowing. Seven of the students had a phone in their pockets, running the TagSense application. For each picture, the sensor information from all the phones were assimilated and processed offline. TagSense generated the following tags automatically:

Picture 1: `November 21st afternoon, Nasher Museum, indoor, Romit, Sushma, Naveen, Souvik, Justin, Vijay, Xuan, standing, talking.`

Picture 2: `December 4th afternoon, Hudson Hall, outdoor, Xuan, standing, snowing.`

Picture 3: `November 21st noon, Duke Wilson Gym, indoor, Chuan, Romit, playing, music.`

With out-of-band sensing over multiple dimensions, tagging can be relatively easier, compared to image processing/face recognition. Tags like "`Nasher Museum`" and "`Wilson Gym`" are extracted from logical location services, or by reverse looking up geographic databases. "`Indoor/outdoor`" is extracted from light-sensor readings; the names of each individual from the phones; "`standing, playing`" from accelerometers; and "`talking, music`" from sound. Perhaps more importantly, the tags do not include the names of the people who were left out of these pictures, even though these people were within wireless vicinity. Thus, although TagSense-generated tags are not highly sophisticated, we believe they improve the state of the art. Google Goggles was not able to produce any tags for the same pictures in Figure 1.

We also asked an arbitrary person (who was not present at the scene and does not know the people in the pictures) to assign tags. This represents what one might expect from a crowd-sourcing solution, such as from Mechanical Turk [2]. The resulting tags were as follows:

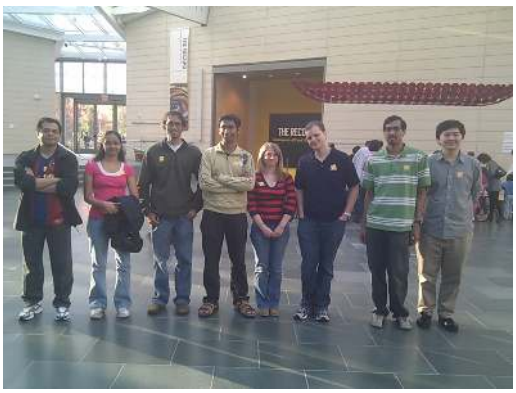

Figure 1: Three example pictures. TagSense tags each picture with the time, location, individual-name, and basic activity. Face recognition by iPhoto and Picasa can tag people in the left picture, less so in the middle, and not so in the right picture. An arbitrary human (emulating crowd-sourcing) is able to label with semantically rich tags, however cannot name individuals. Google Goggles, relying on image processing, offers poor tags.

Picture 1: `many people, smiling, standing`
Picture 2: `one person, standing, snowing`
Picture 3: `two guys, playing, ping pong`

We observed that human assigned tags were not a strict superset of TagSense. In fact, they were somewhat complementary in terms of semantic richness. While humans easily recognized the semantics of a picture, such as "`smiling`", electronic sensing extracted low-level attributes such as names of people, simple actions, location, ambience, etc. Even though these sensed attributes may only add up to a small "tag vocabulary" today, recent research has made significant advances in enriching this vocabulary. Activity recognition [10], ambience sensing [7], human behavior learning [11], and location sensors [12] are being actively mined, resulting in the extraction of rich contextual information. TagSense identifies this developing opportunity and proposes to harness it through a broad architectural framework. We instantiate the framework with an automatic image tagging application.

Scope of TagSense

TagSense is a first step and certainly not a complete solution for image tagging. Images of objects (e.g., bicycles, furniture, paintings), of animals, or of people without phones, cannot be recognized. Put differently, TagSense requires the content in the pictures to have an electronic footprint that can be captured over at least one of the sensing dimensions. If the objects do not present this footprint, one has to rely on the visual domain alone. Arguably, one may envision RFIDs on bicycles, furniture, paintings, and even pet animals in the future. If future cameras come equipped with RFID readers, TagSense will immediately be able to tag each picture based on the objects in it. However, without RFID readers on today's phones, TagSense narrows down the focus to identifying the individuals in a picture, and their basic activities.

Basis for Comparison

It is natural to contrast the person-identification capabilities of TagSense against face recognition algorithms. While we will indeed compare with iPhoto and Picasa (both of which allow face tagging via some human-assistance), we observe that the visual and sensor-based approaches can be complementary. Face recognition may work well under good lighting conditions, when a person's face is clearly visible; TagSense may be as good even under bad lighting conditions. For instance, unlike TagSense, both iPhoto and Picasa did not recognize the people in the middle and right pictures in Figure 1. TagSense does not depend on the physical features of a person's face (whether he is wearing a dark glass, or sporting a new beard), whereas face-recognition applies well to kids who do not carry phones. Finally, face recognition falsely detected faces in a wall-painting, and got confused between twins in a picture; TagSense avoided both these pitfalls. In summary, a hurdle to visual recognition may not hinder recognition on other sensing dimensions, and vice versa. Therefore, we believe that TagSense in conjunction with face recognition algorithms could make people-identification even more robust.

3. SYSTEM OVERVIEW

Figure 2 shows a high level overview of the TagSense system. We consider an example scenario where Bob is taking a picture of Alice and Eve, while John is in the vicinity (but not in the picture). We describe the operations step-by-step.

When TagSense is activated at the beginning of an event (e.g., party, dinner, vacation, picnic), the application prompts the user for a session password. People participating in that event, and willing to run TagSense, decide on a common password and enter it in their respective phones. This password acts as a shared session key, ensuring that sensed information is assimilated only from group members. Thus, when Bob takes a picture of Alice in a crowded place, the picture does not get tagged with names of all other people in the crowd. Privacy remains preserved.

Once Bob is ready to take the picture, he activates the camera on the phone. Bob's phone immediately broadcasts an `activate-sensor` beacon, encrypted with the shared key. Phones in the group activate their respective sensors. Once Bob clicks the picture, Bob's camera sends a beacon with its local timestamp and the phones record it. Phone to phone communication is performed using the WiFi *ad hoc mode*. After a threshold time from the click, the phones deactivate their sensors,

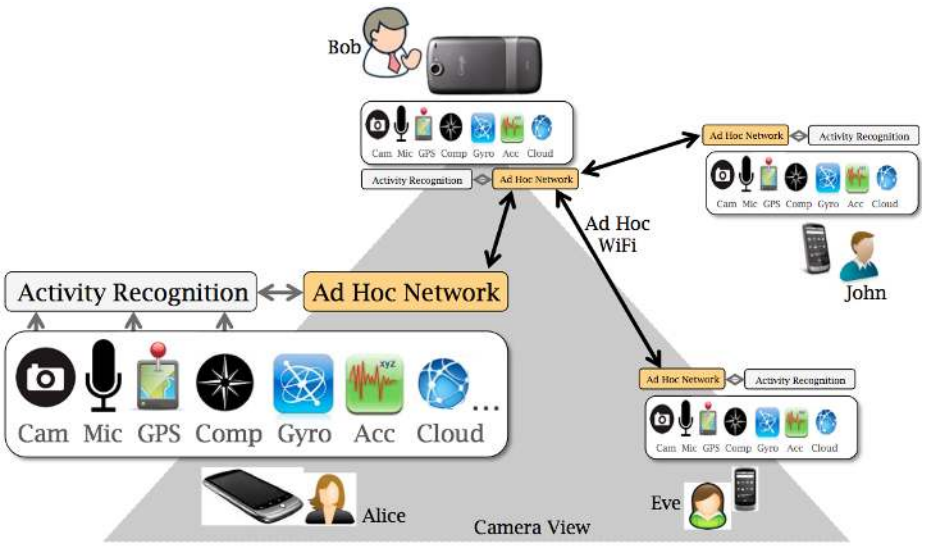

Figure 2: TagSense architecture – the camera phone triggers sensing in participating mobile phones and gathers the sensed information. It then determines who is in the picture and tags the picture with the people and the context.

perform basic activity recognition on the sensed information, and send them back to Bob's phone. Bob's phone assimilates these per-person activities, and also infers some contextual information from its own sensors, including location, ambient sound, light, etc.

The per-person activities are received from each phone in the group, not necessarily those who were in the picture. Thus, Bob's phone must tell which phone-owners were in the picture, and tag it accordingly. Briefly, TagSense adopts three mechanisms. (1) When people explicitly pose for the picture, TagSense extracts a pause signature from the accelerometer readings. This pause signature correlates well with the timing of the photo-click, and is found to be mostly absent in people who are not posing for the picture. (2) People in the picture are often faced towards the camera. TagSense leverages the phones' and camera's compass directions to infer a "mutually facing" relationship; this heuristic improves the confidence of the posing signatures. As will be evident later, unknown and time-varying phone orientations make the problem difficult. (3) For pictures in which the subjects do not pose explicitly, the TagSense camera takes multiple snapshots. The motion vectors for the subjects are computed from the sequence of snapshots, and then correlated to the motion derived from the phones' accelerometer/compass readings. Phones that exhibit a good correlation (between the visual and acceleration dimensions) are used for tagging. The next section visits the design of these techniques in detail.

Knowing which phones/people are in the picture, TagSense extracts the context only from these phones. In some cases, the context information is adequate for direct tagging (e.g., sitting, walking, indoor, outdoor, etc.). However, some measurements require CPU-intensive processing (e.g., laughter recognition), and others rely on external databases (e.g., GPS-to-address). In these cases, TagSense exports the measurements to a cloud and retrieves additional tags. These tags are then ordered in a *when-where-who-what* format as follows,

```
<time, logical location,
name1 <activities for name1>,
name2 <activities for name2>, ...>
```

and uploaded into a specified repository for image-search and other applications.

4. DESIGN AND IMPLEMENTATION

This section zooms into the design and implementation of the individual components in TagSense. We address the primary challenge first, namely *who* are in the picture. We then describe modules that handle *what* they are doing, and *when* and *where* the picture was taken.

4.1 WHO are in the picture

Alice and John may be spatially close by, but Bob may choose to take Alice's picture alone. Since Bob's phone will communicate to all phones through WiFi, and because phones use omnidirectional antennas, it is hard to tell which phones are part of the picture. TagSense explores combinations of multiple sensing dimensions, along with observations about human behavior, to identify who are in the picture. We present 3 main opportunities: (1) accelerometer based motion signatures, (2) complementary compass directions, and (3) motion correlation across visual and accelerometer/compass.

(1) Accelerometer based motion signatures

We imagined that when subjects pose for a picture, their phones are likely to exhibit a motion signature that is different from those not posing. The intuition is that the subjects of the picture often *move into a specific posture in preparation for the picture, stay still during the picture-click, and then move again to resume normal behavior.* TagSense expects to find such a signature around the time of the picture-click, but only in the accelerometers of subjects posing for the picture. For those

not posing, the expectation is that their motions would not be in "lock-step" with the subjects, and hence, their accelerometers will not reveal such a signature.

To verify the existence of posing signatures, we distributed NexusOne phones to 4 students, and took 20 pictures at different times and locations. Different subsets of students posed in each of the pictures, and all their accelerometer readings were recorded. The measurements were processed offline, and the variance visualized over time. Figure 3(a) shows the results from a random subset of people that were in the picture, while Figure 3(b) shows the same for people outside the pictures. The black vertical line indicates the time at which the pictures were taken. The posing signature appears to be distinct – the accelerometer variance subsides for a few seconds around the picture-click, and grows back again. Section 5 presents additional results from ≈ 70 posing pictures, reaffirming that the presence of the signature is a reliable discriminator.

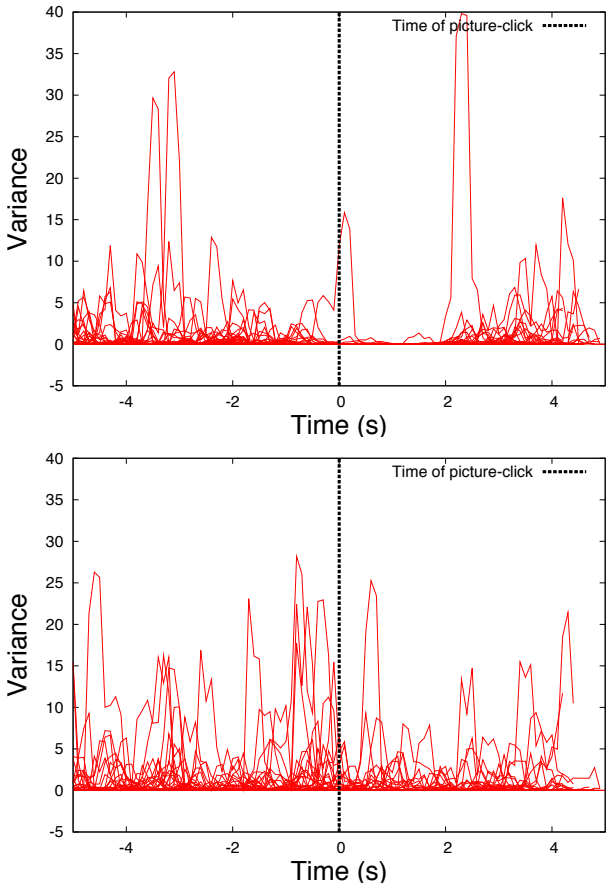

Figure 3: The variance of accelerometer readings from phones of (a) those in the picture and (b) those outside the picture. Posing signature is evident in (a) and absent in (b).

(2) Complementary Compass Directions

While the posing signature may be a sufficient condition to detect a person, it is obviously not necessary. A person may behave naturally when the picture is being taken (e.g., eating, playing a guitar, making a speech) – the posing signature will be absent. Even if a person poses, it may not be distinct on the accelerometer. Figure 4(a) shows a picture where the subject was seated on his chair and only looked up when the picture was taken – the "looking up" did not reflect on the accelerometer. TagSense makes an assumption to solve this problem. The assumption is that people in the picture roughly face the direction of the camera, and hence, the direction of their compasses will be roughly complementary to the camera's facing direction. Thus, by analyzing people's compass directions, TagSense expects to tell who are in the picture.

The challenge, however, is that the user and her phone may not be facing the same direction. This is because the phone may be in the pant's side-pockets, in a women's purse, back-pockets, etc. Let *personal compass offset* (PCO) denote this angle between the user's facing direction and her compass (see Figure 4(b)). The TagSense camera computes PCO as:

$$UserFacing = (CameraAngle + 180)\ mod\ 360$$

$$PCO = ((UserFacing + 360) - CompassAngle)\ mod\ 360$$

Figure 4(c) shows the distribution of PCO, derived from 50 pictures in which people were facing the camera. Evidently, a reasonable fraction of the phones are not oriented in the opposite direction of the camera even though its user is actually facing the camera. Therefore, blindly using the compass direction to detect the subjects of a picture can be erroneous.

TagSense mitigates the compass problem by periodically re-calibrating the PCO. The idea is to find pictures in which subjects can be reliably identified using other methods, and use these pictures for recalibration. Specifically, if TagSense identifies Alice in a picture due to her posing signature, her PCO can be computed immediately. In subsequent pictures, even if Alice is not posing, her PCO can still reveal her facing direction, which in turn identifies whether she is in the picture. This can continue so long as Alice does not change the orientation of her phone. However, if she moves the phone and changes its PCO (say at time t_i), then all pictures taken after t_i may get erroneously tagged. This is because TagSense will make decisions based on a stale value of Alice's PCO, leading to false positives or false negatives.

We believe that TagSense can be made robust to such changes in PCO. Let us assume that TagSense takes a picture at time t_j, $t_j > t_i$, where Alice was again identified reliably through a posing signature. TagSense recalibrates Alice's PCO at t_j, and revisits pictures that were taken between t_i and t_j. All these pictures are re-analyzed with Alice's new PCO – if the new PCO indicates that Alice was actually facing the camera in a prior picture, the correction is made. In general, as long as an individual's PCO gets periodically re-calibrated, her presence in other pictures may be reliably identified. Since posing pictures are common, we believe such recalibration is viable.

We note that the time of changing the PCO, t_i, may not be always known. If the phone's display is turned on at some point, TagSense notes that as a time at which the PCO may have changed. However, if the user does not activate the phone and only moves it from one orientation to another, t_i will remain undetectable. In such cases, TagSense orders all the pictures in time, and identifies the ones in which Alice was detected reliably – call these Alice's *anchor pictures*. For

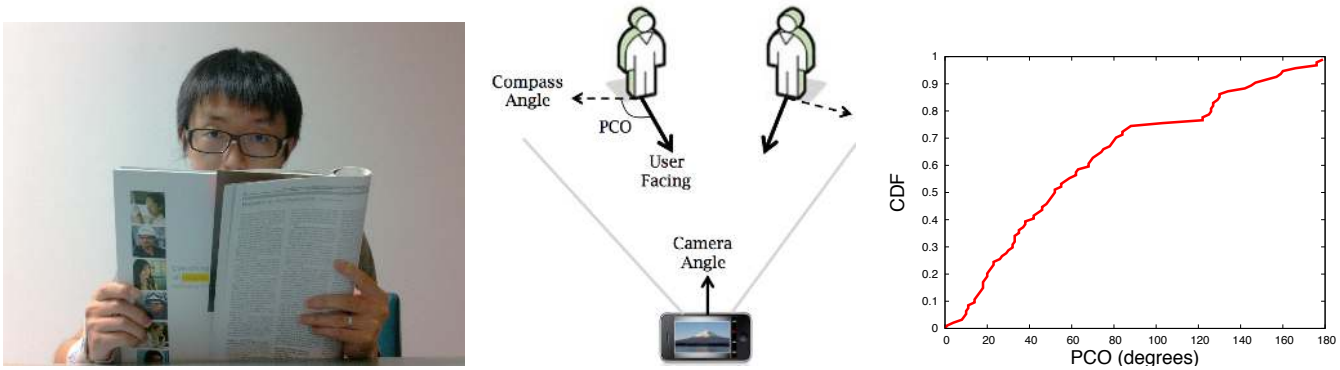

Figure 4: (a) Posing signature absent in the subject. (b) Personal Compass Offset (PCO) (c) PCO distribution from 50 pictures where subjects are facing the camera. PCO calibration is necessary to detect people in a picture using compass.

every other picture, P, TagSense identifies an anchor picture that is closest in time to P. Alice's PCO in this anchor picture is now used to update Alice's PCO in picture P; then a decision is made about Alice's presence in P. This offline process repeats for all users in the group, ultimately leading to improved detection accuracy. Of course, false positives will still occur since some people outside the picture may have their compasses parallel to those in the picture. TagSense is unable to avoid these at present; however, WiFi power control and/or phone-to-phone communication are promising approaches.

(3) Moving Subjects

Some pictures may have subjects moving in them – playing ping-pong, finish line in a race, people dancing, etc. Posing signatures will clearly be absent in these; even compass orientation is unreliable because the moving subjects' compass reading may continuously change over time. TagSense relies on a multi-dimensional sensing heuristic to identify the moving subjects. The essential idea is to take multiple snapshots from the camera, derive the subject's motion vector from these snapshots, and correlate it to the accelerometer measurements recorded by different phones. The accelerometer motion that matches best with the optically-derived motion is deemed to be in the picture. We elaborate on this next.

When a user clicks for a picture, the TagSense camera takes several snapshots[1] following the click. These time-sequence of snapshots are then analyzed to identify the motion vectors of the subjects. To achieve this, the component techniques are drawn from the literature. Figure 5 illustrates the intermediate steps for calculating the motion vectors between two snapshots. First, the velocity of each pixel is computed by performing a spatial correlation across two snapshots — this operation is well known as *Optical Flow*. TagSense adopts a recent `Matlab` implementation for Optical Flow [13], and the outcome is shown in Figure 5(c). Second, the average velocity for the four corner pixels are computed, and subtracted from the object's velocity – this compensates for the jitter from the cameraman's hand. Third, the color of each pixel is redefined based on its velocity. Neighboring pixels with different velocities are assigned different colors, producing clear boundaries for moving objects (Figure 5(d)). Fourth, by leveraging the outcome of the third step, an edge finding algorithm identifies the objects in the picture, as shown in Figure 5(e). Now, a bounding box is created around each object. Finally, the average velocity of one-third of the pixels, located in the center of each object, is computed and returned as the motion vectors of the people in the picture. Figure 5(f) shows the result.

[1]A snapshot is at a much lower resolution compared to actual photos. In our prototype, they are just screenshots of the camera preview, taken at time intervals of $0.3s$.

Once the optical motion vectors are in place, TagSense assimilates the accelerometer readings from different phones and computes their individual velocities. Of course, standard noise suppression and smoothing techniques are first applied on the acceleration [10]. TagSense now matches the optical velocity with each of the phone's accelerometer readings. Direct matching is cumbersome – the magnitude and direction of velocity will not directly relate to any instantaneous acceleration reading. Instead, we match the coarse-grained properties of the two motion vectors. A person walking is likely to exhibit a uniform change across the different snapshots, different from those biking, or moving back and forth while playing. The accelerometer readings are also classified into these coarse buckets, and the person is identified based on such a match. The process repeats for every object, and every match tags a new person in the picture.

Combining the Opportunities

After taking a picture, TagSense attempts to leverage the above opportunities for tagging it with the names of people. TagSense first searches for the posing signature in the accelerometer readings of every phone, and also computes that user's facing direction (assuming that it already knows her PCO). If the posing signature is present, the person is immediately deemed to be in the picture, and her PCO is recalibrated. In the absence of the posing signature, TagSense checks if the person is reasonably static. If she is static, and her facing direction makes less than $\pm\ 45^\circ$ angle with the camera's direction, then her name is added to the tags. Finally, if the person is not static, TagSense computes the picture's optical motion vectors and correlates with the person's accelerometer/compass readings. The person is included upon a high-confidence match.

Points of Discussion

(1) We are aware that *TagSense cannot pinpoint people in a picture*. It can say that Alice is in the picture but may not be able to point out which of the three people in the picture is Alice. Nevertheless, we believe that tagging the picture with

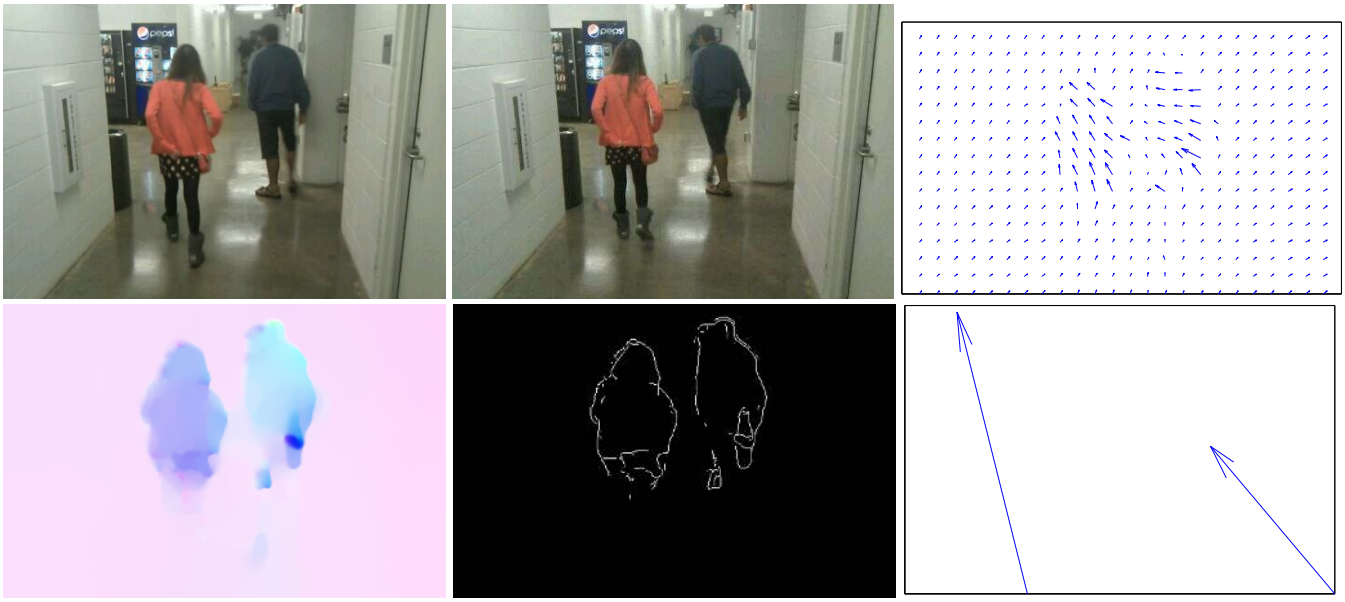

Figure 5: Extracting motion vectors of people from two successive snapshots in (a) and (b): (c) The optical flow field showing the velocity of each pixel; (d) The corresponding color graph; (e) The result of edge detection; (f) The motion vectors for the two detected moving objects.

only the names is still valuable for a variety of applications. For instance, TagSense-style tagging can relieve users of the cumbersome task of manually organizing their personal photos into albums. Instead, users will have the option of creating a new album by simply associating it with a tag string. Relevant pictures can automagically become part of that album (akin to the notion of labels in Gmail). Such an application does not require pinpointing a person inside a picture.

In another example, TagSense may allow an image-search application that uses the presence of a person as a context of the query. For instance, "show all Birthday party pictures where Eve was present". Again, such a useful application does not require knowing which one is Eve – just knowing that she is in the picture is adequate. Finally, tagging pictures with names of people can be useful for sharing pictures instantaneously among the people present in that picture. Even social networking applications such as Facebook may benefit – instead of defining groups that have access to these pictures, the tags of the pictures can be used as a self-defined group. Those present in the picture can automatically be given the right to view the picture, making content management intuitive and user-friendly.

(2) *TagSense cannot identify kids as they are not likely to have phones*. This is a major limitation, however, even at a young age, kids are beginning to listen to music and play games on mobile devices like iPods/PSPs. TagSense works with any such device that has a wireless footprint and basic sensors. Of course, tagging pictures of babies will still be hard, and babies may be a reason for taking many pictures.

(3) *TagSense's compass based method assumes people are facing the camera*. Though there is some leeway in the facing direction, this assumption is invalid when someone is turned sideways or around in that picture, and not posing or moving. We currently do not have a remedy for this case, and leave it for future investigation.

4.2 WHAT are they doing

Activity recognition with the aid of mobile phones has been an active area of research lately [7, 14]. TagSense can avail the schemes resulting from that research to identify activities while tagging pictures. Therefore, the focus of this paper is not on devising new activity recognition schemes. Instead, we implement a few schemes to provide a sample set of activity tags for the sake of completeness in developing the TagSense prototype. We start with a limited vocabulary of tags to represent a basic set of activities. This vocabulary can be later enlarged to incorporate the further advances in activity recognition in the future. In the following, we discuss a few of the supported activities in our vocabulary.

Accelerometer: Standing, Sitting, Walking, Jumping, Biking, Playing. Most of the activity recognition schemes rely on accelerometer sensor. It has been observed that many of the physical activities produce distinct motion patterns. There is a clear signature from accelerometer readings to determine whether someone is sitting or standing. Similarly, using statistics of accelerometer readings (e.g. variance, 4th moment, dynamic range, and zero-crossing rate) as well as location information, it is possible to differentiate between walking, jumping, biking, or playing. There are many other activities that can be easily recognized with the aid of accelerometer. But we have not done this exhaustively since our aim is only to show a representative set of activity tags to indicate what is possible with TagSense approach.

Acoustic: Talking, Music, Silence. This sensor provides information that is quite distinct from what could be gleaned from the visual picture. From a picture of a person in front of a microphone, it is hard to say whether that person is talking

or singing. On the other hand, with the audio samples from the acoustic sensor, it becomes easier to differentiate between these two cases. In our TagSense prototype, we provide basic information regarding ambient sound when the picture is taken. The classification is done by feeding Mel-frequency Cepstral coefficients into SVM. A few audio samples around the picture click would suffice for this purpose.

We evaluate the accuracy of our prototype in recognizing activities in Section 5.2 and show encouraging results. We believe more sophisticated techniques can further improve the vocabulary and accuracy of activity tags.

4.3 WHERE is the picture taken?

The location of a picture conveys semantic information about the picture – a photo taken inside a restaurant conveys a sense of food and fun. It also enables location based photo search, such as *all pictures from the Disneyland food court*. Importantly, GPS based location coordinates are unsuitable for these purposes. In many cases, its important to distill out a semantic form of location, such as the name of a place (gym, airport, cafe), indoor or outdoors, or even descriptions of nearby landmarks (e.g., near Eiffel Tower). Tagging the background of the picture (e.g., Atlantic Ocean in the background) may be even more attractive. TagSense leverages mobile phone sensors and cloud services to approach these goals.

The "place" is derived by performing a reverse lookup on the GPS coordinates. We assume that such databases will emerge over time, or SurroundSense [12] like services will become prevalent. Now, to infer whether the picture was taken indoors or outdoors, TagSense utilizes the light sensor on the camera phone. We find that in most cases, the intensity of outdoor environments are either far above or far below the light intensity in indoor environments. Figure 6 shows the variation of light intensity measured at 400 different times, across days and nights in outdoor and indoor environments. Evidently, it is feasible to compute light intensity thresholds (one for daytimes and another for nights), using which indoor environments can be reliably discriminated from outdoors. TagSense uses the light intensity measurement (from the camera) during the picture-click, and uses it to tag the picture as "indoors" or "outdoors".

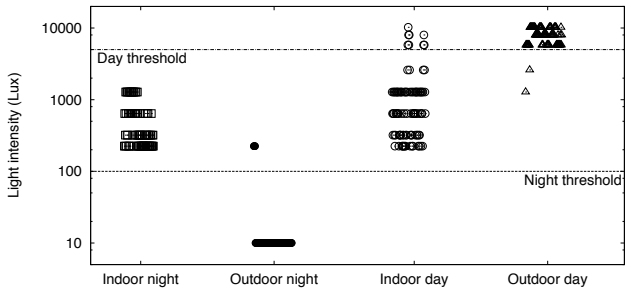

Figure 6: Indoor/outdoor light intensities

TagSense employs the combination of location and phone compasses to tag picture backgrounds. For example, knowing that the picture is at the California beach and the camera is facing westward, one can infer the ocean in the background. Further, Enkin and similar services [15] have developed a location/orientation database for frequently visited locations. Given a <location, orientation> tuple, the database returns names of visible objects. One can create similar mini databases for their personal spaces, such as homes, office, gardens, or university campus. Google Streetview is also expected to expose such an API in the near future. TagSense exploits these capabilities for background tagging.

4.4 WHEN is the picture taken

Tagging the picture with current time is a standard feature in today's cameras, and TagSense trivially inherits it. However, TagSense adds to this by contacting an Internet weather service and fetching the weather conditions. If the picture happens to be taken outdoors, and if the whether suggests snowing or raining, TagSense associates that tag with the photo. Finally, if the picture is taken after sunset (determined by sending the current time to the weather service), TagSense tags the picture as "`at night`". Together, the "`when-where-who-what`" tags offer a reasonable description of the picture. The following section evaluates the overall efficacy.

5. PERFORMANCE EVALUATION

To evaluate TagSense, we have conducted real-life experiments with 8 Google Nexus One phones. One phone is used as a camera while the others are carried by 7 participants naturally in their pockets. When a picture is taken, the camera triggers other phones and gathers sensor readings through WiFi ad-hoc mode. The sensing data is later processed to generate tags. To collect a diverse set of pictures, we visited four different settings: (1) Duke University's Wilson Gym, (2) Nasher Museum of Art, (3) a research lab in Hudson Hall, and (4) a Thanksgiving party at a faculty's house. For brevity, we refer to these scenarios as *gym*, *museum*, *lab*, and *house*.

Our evaluation aims to answer the following questions: (1) How well does TagSense tag people compared to approaches based on face recognition; (2) How does human behavior in different scenarios affect the individual tagging methods (posing, compass, motion) employed by TagSense. (3) How well can TagSense recognize activities and context. We begin with tagging people in the picture, and later evaluate activity and context-tagging. We end with a toy image search tool using our collection of 200 pictures tagged by TagSense.

5.1 Tagging People

We compare TagSense with Apple's iPhoto and Google's Picasa, two popular products that employ face recognition. Once a person's face is manually tagged with a name, iPhoto and Picasa attempt to tag similar faces in other pictures. In our evaluation, we tagged each participant's face once and let iPhoto and Picasa tag other pictures in the set. One may argue that iPhoto and Picasa perform better with more training. However, we believe that our evaluation setting is fair as it ensures similar amount of human assistance in all these schemes.

Figure 7 illustrates how well TagSense tags people in a picture: Figure 7(a) shows how accurately people were included in the picture, while Figure 7(b) shows how accurately they were excluded[2]. For example, in the last picture in Figure 7(a), TagSense correctly identifies 2 of the 3 people in the picture.

[2]In an attempt to preserve natural behavior, we allowed people to freely move in and out of the camera's communication

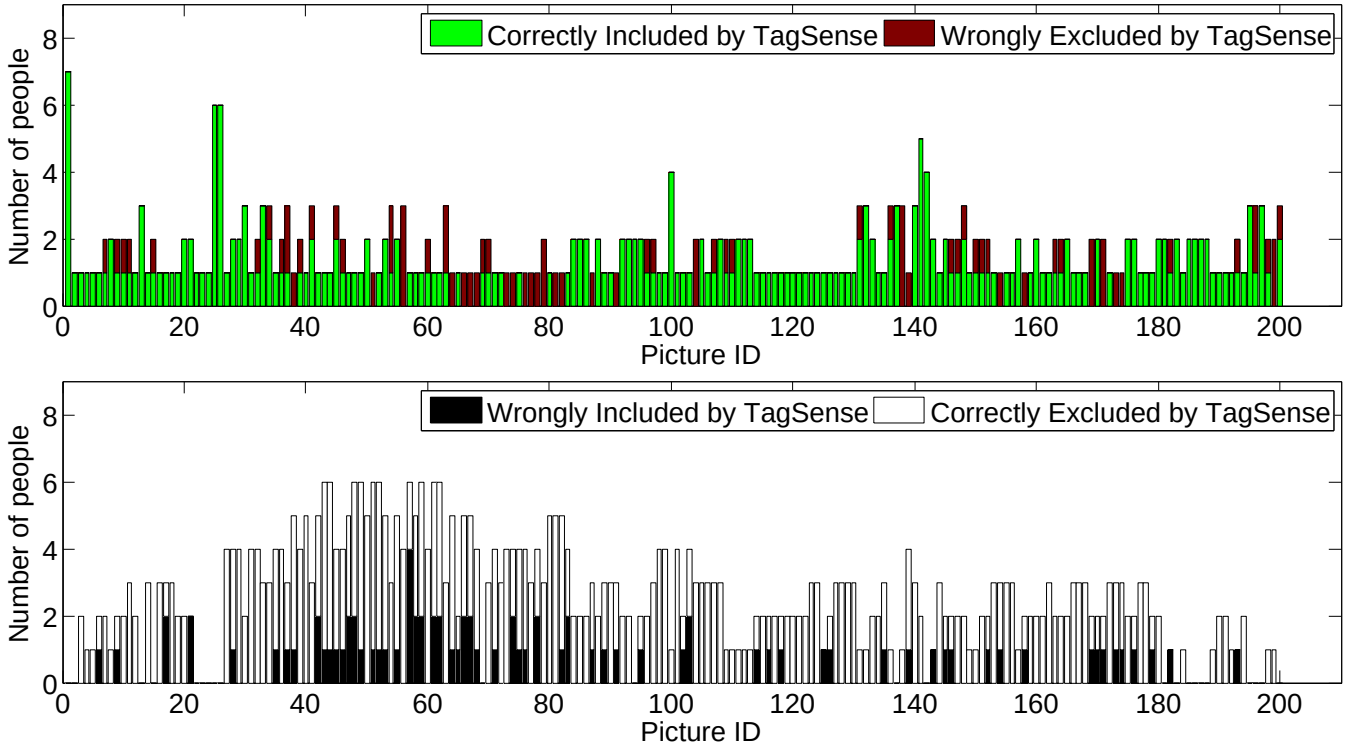

Figure 7: Performance of TagSense: (a) Top and (b) bottom graphs show people inside and outside each picture. Wrongly excluded/included ones are shown in red/black. Overall, TagSense does well in tagging people.

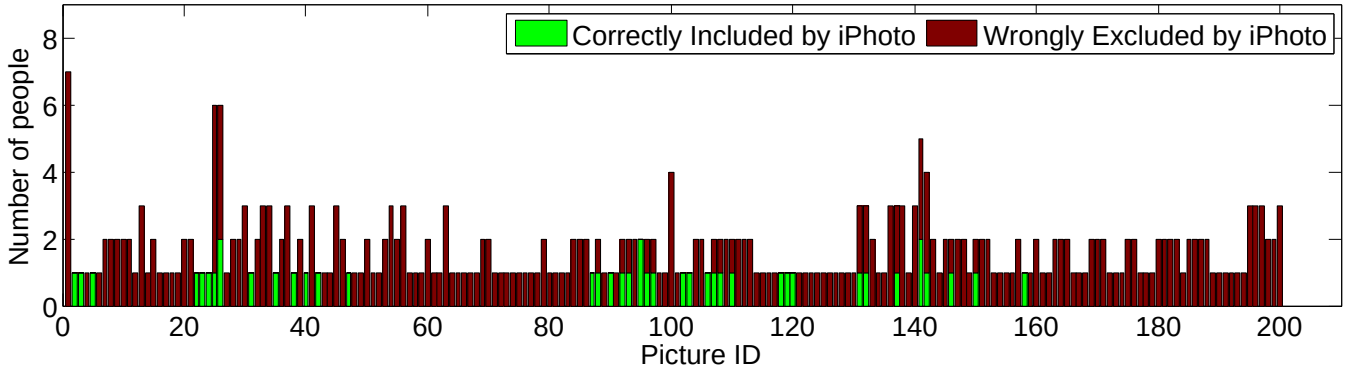

Figure 8: iPhoto wrongly excludes quite a few people. But only a few are wrongly included (graph not shown).

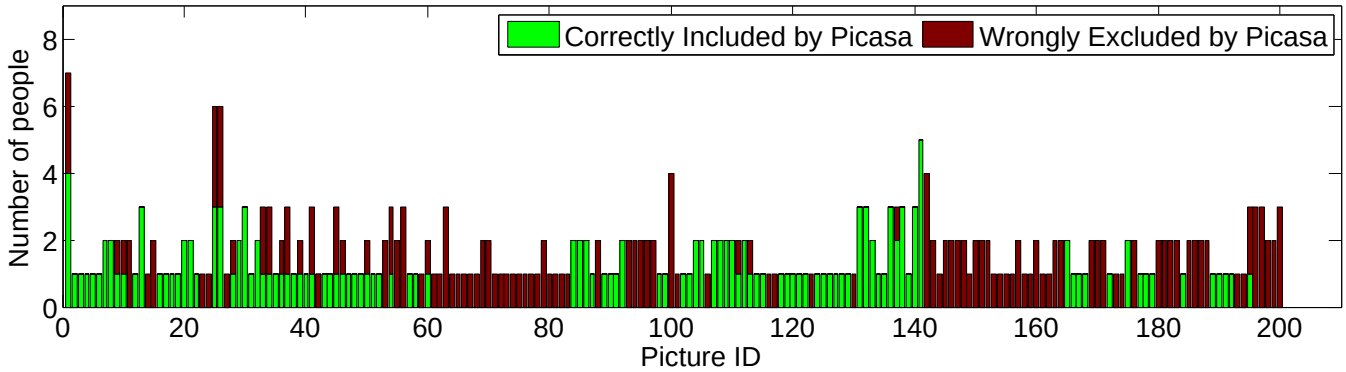

Figure 9: Picasa too wrongly excludes many people. But just one is wrongly included (graph not shown).

range. Hence, in some of the pictures, the sum of people inside and outside the picture adds up to less than seven.

Overall, TagSense performs reasonably well in separating the people outside from those inside a picture. In contrast, Figure 8 shows that iPhoto has very high false negatives though only a few false positives, i.e., iPhoto is accurate when it detects a face, but it fails to detect a large fraction of faces. Picasa, as shown in Figure 9, performs better than iPhoto on our picture set, but it too does not recognize many faces. To formally evaluate TagSense and compare it with iPhoto and Picasa, we employ the metrics commonly used for information retrieval – *precision*, *recall* and *fall-out*.

Metrics

We know the people in each picture, so the ground truth is known. Therefore, the precision, recall, and fall-out of TagSense can be defined as follows.

$$\text{precision} = \frac{|\text{People Inside} \cap \text{Tagged by TagSense}|}{|\text{Tagged by TagSense}|}$$

$$\text{recall} = \frac{|\text{People Inside} \cap \text{Tagged by TagSense}|}{|\text{People Inside}|}$$

$$\text{fall-out} = \frac{|\text{People Outside} \cap \text{Tagged by TagSense}|}{|\text{People Outside}|}$$

Similarly, we can compute the precision, recall, and fall-out for iPhoto and Picasa. The goal of a tagging scheme is to achieve high precision, high recall, and low fall-out.

Overall Performance

Figure 10 compares the performance of TagSense with iPhoto and Picasa using these metrics. The precision, recall, and fall-out are computed over the entire set of pictures. While the precisions of iPhoto and Picasa are better than TagSense, their recalls are much lower. Importantly, recall is a key metric for search-like applications. A low recall implies that when a user searches for a picture, the results are unlikely to include the one she is looking for. On the other hand, a scheme with high recall (albeit with low precision) is more likely to return the sought picture, along with several less relevant ones. TagSense is suitable for the latter type of service, which perhaps is more desirable in image-search applications.

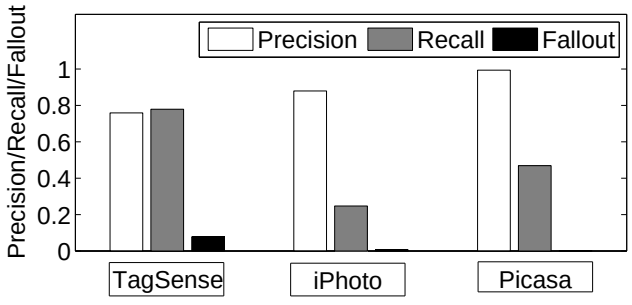

Figure 10: The overall *precision* of TagSense is not as high as iPhoto and Picasa, but its recall is much better, while their fall-out is comparable.

Method-wise and Scenario-wise Performance

Figure 11 shows how the 3 different people-tagging methods (posing, compass, and motion) perform in different scenarios. Evidently, posing signatures work reliably in the majority of museum and lab pictures (Figure 11(a)), where participants explicitly posed for the camera. On the contrary, people were mostly sitting/eating/talking in the house, and this did not present a distinct posing signature on the accelerometer. Thus, the compass-based identification proved beneficial in this scenario (Figure 11(c)). Similarly, motion-based methods were suitable for gym pictures, where people were engaged in playing racquetball, ping-pong, or running (Figure 11(d)). The performance of iPhoto and Picasa also varies; both precision and recall are relatively better in the museum and lab, where pictures are close-ups, and people mostly face the camera. The recall degrades significantly in the gym and house, where people may not be always facing the camera, and behave more naturally. These results convey a sense of complementary behavior between TagSense and iPhoto/Picasa, and we believe that their merger can be powerful.

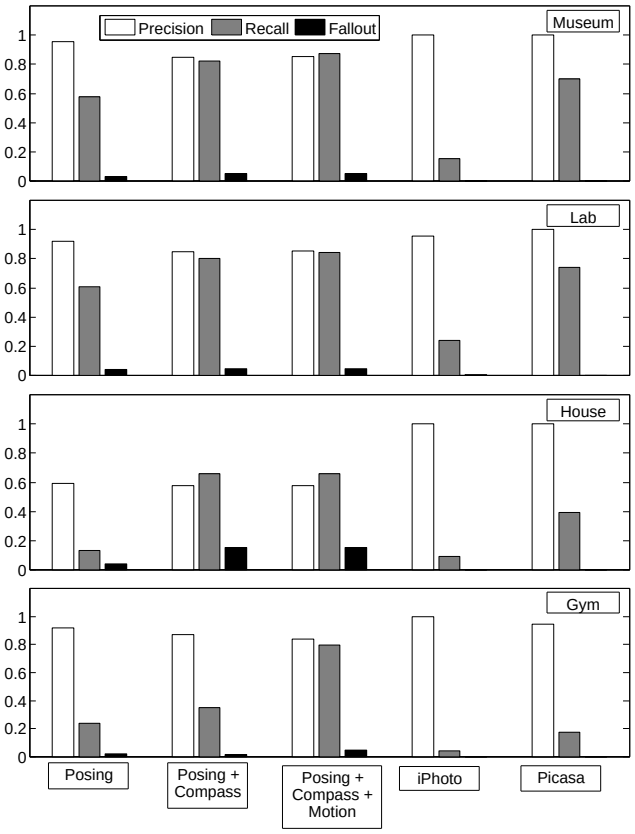

Figure 11: The performance of different TagSense methods under different scenarios (from top to bottom): (a) museum, (b) lab, (c) house, and (d) gym.

Searching Images by Name

An obvious application of tagging people is name-based image search. The user types one or more names and the system should retrieve all the images containing them. Figure 12 shows TagSense's performance in comparison to iPhoto and Picasa. It shows the results for 9 individual searches and for 4 pairs of names (e.g., Alice and Eve). The pairs are chosen such that there are several pictures with both the individuals in them. The results demonstrate once again that TagSense offers reasonable precision and better recall than iPhoto and

Picasa. The lower recall of iPhoto and Picasa gets amplified when searching for pictures of a pair of people, as in pair ID 12. Also, note that both iPhoto and Picasa recognize some individuals better than others, whereas TagSense provides similar performance across all people.

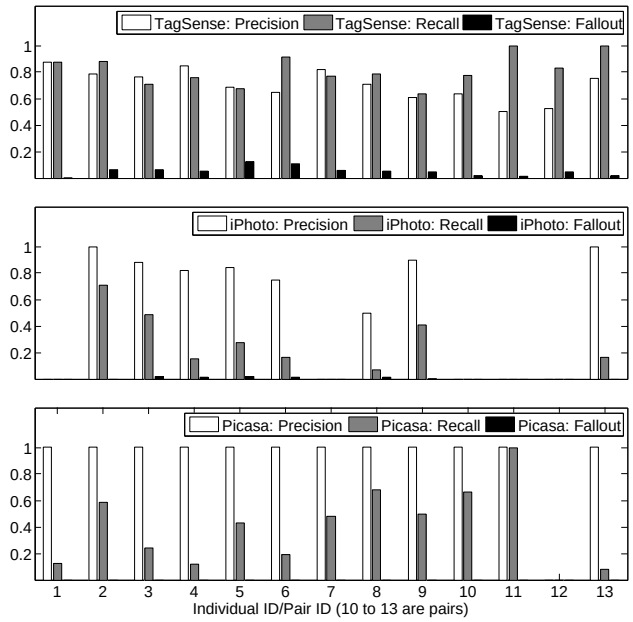

Figure 12: Performance comparison of TagSense with iPhoto and Picasa for name based image search.

We hypothesize that Picasa and iPhoto may be able to match TagSense's performance by tuning certain parameters. However, TagSense is a nascent system (compared to years of research behind Picasa/iPhoto), and we believe that further investigation can improve TagSense's capabilities as well. Perhaps more importantly, the integration of both the approaches can be superior to either of them. In our future work, we plan to perform such an integration, reinforcing sensing-based approaches with face recognition, and vice versa.

5.2 Tagging Activities and Context

TagSense processes the sensor data from phones to create activity/contextual tags for the picture. As an example, the figure below shows a cloud of tags assigned to pictures in the gym. Observe that the location, participants, and simple activities (such as playing and jumping) appear in the cloud (the size of each tag scaled by its frequency of occurrence).

The assessment of an activity-tagging scheme is not straightforward since the ground truth is rather subjective. Therefore, we rely on humans to judge the relevance and completeness of tags generated by TagSense. Since completeness is a function of the "vocabulary" of tags, we limited the experiment to only TagSense's vocabulary. We asked humans to pick a set of tags for the picture from the given vocabulary. We can then define *precision* and *recall* as follows.

$$\text{precision} = \frac{|\text{Tags by Humans} \cap \text{Tags by TagSense}|}{|\text{Tags by TagSense}|}$$

$$\text{recall} = \frac{|\text{Tags by Humans} \cap \text{Tags by TagSense}|}{|\text{Tags by Humans}|}$$

We do not define a metric like fall-out here since we can not meaningfully bound the total set of tags that are considered irrelevant to the picture, and then find the fraction of those that are incorrectly tagged by TagSense.

Figure 13 shows the results with 5 human volunteers assessing TagSense's activity tags[3]. As evident, most of the tags produced by TagSense are relevant and also somewhat describe the context of the people in the pictures. Of course, with a small tag vocabulary (of around 30 activity tags), this should not be viewed as a concrete result. Rather, this only suggests that the TagSense framework can be useful, if its vocabulary grows over time by borrowing from activity recognition, acoustic sensing, and signal processing communities.

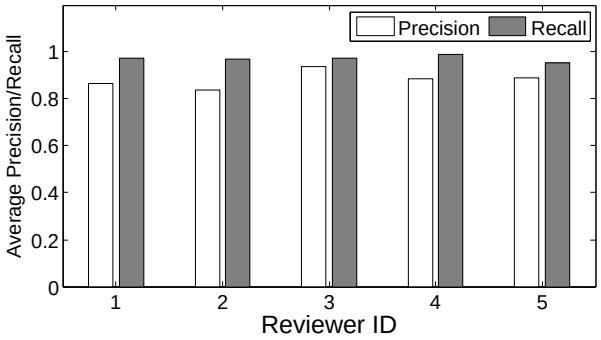

Figure 13: Assessment of tags given by TagSense.

5.3 Tag Based Image Search

To understand user-centric performance of TagSense, we implemented a toy image-search engine running on our set of 200 tagged pictures. We recruited 5 volunteers and showed them 20 pictures each (randomly picked from the 200). They were then asked to compose query strings from our tag vocabulary and retrieve each of the pictures – they could also use names of people in their queries. For the given search string, our system returned a set of pictures. The volunteer marked the number of relevant pictures and also whether the target picture is one of them (we call this a *hit*). Table 1 shows the per-volunteer performance results.

[3]The volunteers were recruited from our research group members and their friends. Though there is a potential for bias here, the volunteers have been explicitly asked to provide objective assessment. Furthermore, we note that this assessment by a small group of people is meant to be illustrative, and a broader user study is essential to validate our results.

Table 1: Performance of tag based image search

Name	Avg. Relevant	Avg. Irrelevant	Hit rate
User 1	2.75	4.85	0.85
User 2	5.6	1.8	0.65
User 3	4.05	2	0.5
User 4	4.05	2.35	0.7
User 5	2.55	1.6	0.55

Figure 14 zooms into the results of user 4 with medium search satisfaction. For each search string, it shows the number of relevant and irrelevant pictures. The search string is marked with a '*' if it is a hit, a 'x' otherwise. While TagSense does not precisely return the target pictures in all cases and returns irrelevant ones in some cases, we believe that the overall results are encouraging for continued research in this direction.

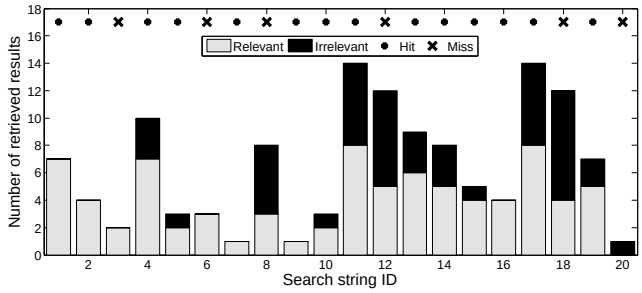

Figure 14: The measure of satisfaction of a user of TagSense based image search system.

6. LIMITATIONS OF TAGSENSE

In this section, we present some potential concerns with TagSense, and discuss how they can be alleviated.

TagSense vocabulary of tags is quite limited. TagSense is necessarily limited by the activities that can be recognized with phone sensors. But the tags assigned by humans are much richer than any sophisticated auto tagging approach. However, from the perspective of image search and retrieval, the limited vocabulary may suffice.

TagSense does not generate captions. Though a natural next step, captioning a picture is hard for TagSense given its limited vocabulary. However, if it generates tags like, "Chuan, Xuan, playing, beach", natural language processing may be able to string them together to make a caption like "Chaun and Xuan playing at the beach".

TagSense cannot tag pictures taken in the past. It requires active recording of sensor readings while taking the picture. In that sense, image processing approaches are broadly applicable to a variety of images. However, looking forward, TagSense can greatly help in tagging photos involving people, which form a bulk of image searches.

TagSense requires users to input a group password at the beginning of a photo session. One could argue that users may a feel a little inconvenienced by this requirement. However, this can be alleviated since most pictures frequently involve family members, close friends, or regular contacts. Therefore, it is conceivable to have preset passwords for these groups and default to one based on the picture context/location. Then, TagSense users need to explicitly specify a password only on a few occasions.

TagSense methods for tagging people are complex. A simpler manual alternative to TagSense, one might consider, is to display the list of people near the camera and let the camera-user select from that small list. We note that determining the list of people in the vicinity (using an ad hoc WiFi network) is part of the TagSense architecture. Moreover, when taking the picture, people are keen on capturing the moment and may easily get annoyed with any distractions. While it is possible to record some meta information to aid the users offline, they may not bother to tag them later. We believe that users prefer seamless tagging and TagSense methods can be redesigned and refined to make tagging more efficient and accurate.

7. FUTURE OF TAGSENSE

TagSense is an alternative way to address a longstanding problem of automatic image tagging. While our current prototype has some limitations, we argue that future technological trends are well aligned with the TagSense architecture. We list some of these trends below and discuss their implications.

Smartphones are becoming context-aware with personal sensing [14]. Activity recognition research is beginning to determine where a person is and what she is doing using her smartphone's sensors [16]. **TagSense can be broadly viewed as the process of applying a spatio-temporal filter on personal sensing**. It amounts to identifying the devices that fall within the space covered by the camera angle, and gathering their activity and context at the time of click. Viewed this way, all the research on activity recognition feeds into and acts as a force multiplier for TagSense.

Smartphones may have directional antennas [17]. Some researchers advocate equipping smartphones with directional antennas to achieve high throughput [17]. Given the insatiable demand for bandwidth, this may be a viable option in the near future. When the camera has a directional antenna, upon a click, it can send a directional broadcast limiting the area of reception to that covered by the camera angle. Then camera will receive responses only from those that are in the picture. This makes tagging people both simple and accurate.

The granularity of localization will approach a foot [18]. Given the drive towards localization, it will not be long before a person can be localized to a spot within a feet. Then, it is straightforward to identify people that are located within the area covered by the camera angle. Even without absolute coordinates, localization relative to camera would suffice. The ability to measure distance between two nearby smartphones would also aid TagSense [19].

Smartphones are replacing point and shoot cameras [20, 21]. The recent trend has been to equip smartphones with sophisticated camera features diminishing the need for traditional cameras. Moreover, in many instances people forget to bring their cameras and instead take pictures with their phones (which they typically carry everywhere). Therefore, the fraction of pictures taken with phones is already large, which will only grow further, amplifying the utility of TagSense.

8. RELATED WORK

Image tagging has been a topic of extensive research given its applications to image retrieval [22–24]. Because of the challenging nature of this problem, out-of-band solutions have attracted attention of late. An approach in [25] computes event and location groupings of photos based on their time and location. As the user tags some people in their collection, patterns of re-occurrence and co-occurrence of different people in different locations and events are expected to emerge. The system uses these patterns to assist users in tagging pictures by suggesting a short list of candidates. While this eases the process of manual tagging, TagSense aims to automate it. Mobile Media Metadata (MMM) [26] gathers contextual metadata such as location at the time of capture. It also generates more metadata for the captured image based on the metadata of "similar" images at the server. The similarity is determined based on image location and pixel data. When the system is not certain about the similarity, the user is prompted for confirmation. The MMM approach is more suitable for landmark pictures whereas TagSense targets people-centric pictures.

Among all the earlier works on tagging, ContextCam [27] is the most relevant to our work. Both TagSense and ContextCam have similar objectives but their solutions are quite different. ContextCam annotates videos at the point of capture with people in and around a scene. It achieves this by placing ultrasound receivers on a horizontal plane in front of the camera. Moreover, each individual has to wear a device that periodically chirps an ultrasound sequence. These ultrasound receivers and emitters are used to determine the relative distance between an individual and the camera, and whether a person is in the view of the camera. In contrast, TagSense offers a more practical solution. TagSense, to the best of our knowledge, is the first image tagging system that leverages the sensing ability of the current smartphones.

The ideas in TagSense build on three threads of research: mobile sensing, activity recognition and image/audio processing. While each thread is composed of rich bodies of research, we survey only a few of them in the interest of space.

Mobile sensing. Smartphones are becoming a convergent platform for people-centric sensing [28, 29]. Various applications have been developed to exploit sensing across multiple dimensions. SenseCam [30] proposes an innovative idea of using sensors on the camera to decide when pictures should be taken. SoundSense [7] taps into the sound domain to identify events in a person's life, and building a audio journal. SurroundSense [12] shows the possibility of ambience-sensing in determining a user's location. Censeme and Nericell [14, 31] detect user and traffic status and share the information through online networks. All these applications share the vision of leveraging mobile phones as a "point of human attachment" to gain insights into human behavior and activity. TagSense is in an opportune position to fully benefit from this growing body of research.

Activity recognition. Activity recognition has recently taken prominence with advancements in data mining and machine learning algorithms. Researchers are looking into activities through various information sources [10, 32], temporal activity correlations [11] and in different environment settings [33, 34]. We foresee some of these algorithms coming to mobile phones, and becoming amenable to TagSense.

Image/audio processing. Image and audio processing are integral to TagSense [35]. We already use *optical flow* [13, 36], but there are several other opportunities to use image processing for labeling. The literature is vast, but can be well summarized by how Google Goggles is approaching the problem. Any advancements with Google Goggles and similar softwares will be amplified through the multi-dimensional sensing based approach in TagSense.

9. CONCLUSION

Mobile phones are becoming inseparable from humans and are replacing traditional cameras [37]. TagSense leverages this trend to automatically tag pictures with people and their activities. We developed three different methods based on posing, compass, and movement, to identify the people in a picture. We implemented a prototype of TagSense using Google Nexus One phones and evaluated its performance on around 200 pictures. Our experiments show that TagSense has somewhat lower precision and comparable fall-out but significantly higher recall than iPhoto/Picasa. TagSense and iPhoto/Picasa employ complementary approaches and can be amalgamated yielding a robust scheme for tagging images.

10. ACKNOWLEDGEMENT

We sincerely thank our shepherd Jason Hong, as well as the anonymous reviewers, for their valuable feedback on this paper. We thank the students of the SyNRG research group at Duke University for agreeing to participate in the picture-sessions, and providing thoughts and ideas for the betterment of TagSense. We are also grateful to NSF for partially funding this research through the following grants – CNS-0448272, CNS-0917020, CNS-0916995, and CNS-0747206.

11. REFERENCES

[1] Tingxin Yan, Deepak Ganesan, and R. Manmatha, "Distributed image search in camera sensor networks," *ACM SenSys*, pp. 155–168, Nov 2008.

[2] Amazon, "Amazon Mechanical Turk," *https://www.mturk.com/mturk/welcome*.

[3] Google Image Labeler, "http://images.google.com/imagelabeler/," .

[4] L. Von Ahn and L. Dabbish, "Labeling images with a computer game," in *ACM SIGCHI*, 2004.

[5] Tingxin Yan, Vikas Kumar, and Deepak Ganesan, "Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones," in *ACM MobiSys*, 2010.

[6] T. Nakakura, Y. Sumi, and T. Nishida, "Neary: conversation field detection based on similarity of auditory situation," *ACM HotMobile*, 2009.

[7] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell, "SoundSense: scalable sound sensing for people-centric applications on mobile phones," in *ACM MobiSys*, 2009.

[8] A. Engstrom, M. Esbjornsson, and O. Juhlin, "Mobile collaborative live video mixing," *Mobile Multimedia Workshop (with MobileHCI)*, Sep 2008.

[9] Google Goggles, "http://www.google.com/mobile/goggles/," .

[10] L. Bao and S.S. Intille, "Activity recognition from user-annotated acceleration data," *Pervasive Computing*, 2004.

[11] D.H. Hu, S.J. Pan, V.W. Zheng, N.N. Liu, and Q. Yang, "Real world activity recognition with multiple goals," in *ACM UbiComp*, 2008.

[12] M. Azizyan, I. Constandache, and R. Roy Choudhury, "SurroundSense: mobile phone localization via ambience fingerprinting," in *ACM MobiCom*, 2009.

[13] C. Liu, "Beyond Pixels: Exploring New Representations and Applications for Motion Analysis," in *Doctoral Thesis MIT*, 2009.

[14] E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell, "Sensing Meets Mobile Social Networks: The Design, Implementation and Evaluation of CenceMe Application," in *ACM Sensys*, 2008.

[15] M. Braun and R. Spring, "Enkin," *http://enkinblog.blogspot.com/*.

[16] E. Aronson, N. Blaney, C. Stephan, J. Sikes, and M. Snapp, "The jigsaw classroom," *Improving Academic Achievement: Impact of Psychological Factors on Education*, 2002.

[17] A.A. Sani, L. Zhong, and A. Sabharwal, "Directional Antenna Diversity for Mobile Devices: Characterizations and Solutions," in *ACM MobiCom*, 2010.

[18] K. Chintalapudi, A. Padmanabha Iyer, and V.N. Padmanabhan, "Indoor localization without the pain," in *ACM Mobicom*, 2010.

[19] C. Peng, G. Shen, Z. Han, Y. Zhang, Y. Li, and K. Tan, "A beepbeep ranging system on mobile phones," in *ACM SenSys*, 2007.

[20] Nokia Siemens Networks, "Unite: Trends and insights 2009," 2009.

[21] Sam Grobart, "In Smartphone Era, Point-and-Shoots Stay Home," *New York Times*, Dec 2010.

[22] R. Datta, D. Joshi, J. Li, and J.Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM CSUR*, 2008.

[23] Gustavo Carneiro, Antoni B. Chan, Pedro J. Moreno, and Nuno Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 2007, 2007.

[24] Alipr, "Automatic Photo Tagging and Visual Image Search ," *http://alipr.com/*.

[25] Mor Naaman, Ron B. Yeh, Hector Garcia-Molina, and Andreas Paepcke, "Leveraging context to resolve identity in photo albums," in *Proc. of the 5th ACM/IEEE-CS joint conference on Digital libraries*, 2005, JCDL '05.

[26] Risto Sarvas, Erick Herrarte, Anita Wilhelm, and Marc Davis, "Metadata creation system for mobile images," in *ACM MobiSys*, 2004.

[27] Shwetak N. Patel and Gregory D. Abowd, "The contextcam: Automated point of capture video annotation," in *Proc. of the 6th International Conference on Ubiquitous Computing*, 2004.

[28] R. Want, "When cell phones become computers," *IEEE Pervasive Computing, IEEE*, 2009.

[29] R.K. Balan, D. Gergle, M. Satyanarayanan, and J. Herbsleb, "Simplifying cyber foraging for mobile devices," in *ACM MobiSys*, 2007.

[30] D.H. Nguyen, G. Marcu, G.R. Hayes, K.N. Truong, J. Scott, M. Langheinrich, and C. Roduner, "Encountering SenseCam: personal recording technologies in everyday life," in *ACM Ubiquitous computing*, 2009.

[31] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: Rich monitoring of road and traffic conditions using mobile smartphones," in *ACM SenSys*, 2008.

[32] J. Lester, B. Hannaford, and G. Borriello, "ÒAre You with Me?Ó-Using Accelerometers to Determine If Two Devices Are Carried by the Same Person," *Pervasive Computing*, 2004.

[33] T. van Kasteren, A. Noulas, G. Englebienne, and B. Krose, "Accurate activity recognition in a home setting," in *ACM Ubicomp*, 2008.

[34] M. Leo, T. D'Orazio, I. Gnoni, P. Spagnolo, and A. Distante, "Complex human activity recognition for monitoring wide outdoor environments," in *IEEE ICPR*, 2004.

[35] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *ISMIR*, 2000.

[36] S. Baker, D. Scharstein, JP Lewis, S. Roth, M.J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," in *IEEE ICCV*, 2007.

[37] Joshua J. Romero, "Smartphones: The Pocketable PC," *IEEE Spectrum*, Jan 2011.