TagSense: A Smartphone-based Approach to Automatic Image Tagging

Chuan Qin, Xuan Bao, Romit Roy Choudhury, Srihari Nelakuditi
University of South Carolina, Duke University
Digital pictures are undergoing an explosion

Image retrieval becomes crucial
- Image searching
- Personal albums

Facebook: 2.5 billion/month
My PC: 10GB/vacation
Image tagging

- Image retrieval systems use tags
- Human tagging
  - Accurate, widely used
  - Slow and boring
- Image based auto-tagging
  - Still many constraints
Image tagging

- Image retrieval systems use tags

- Human tagging
  - Accurate, widely used
  - Slow and boring

- Image based auto-tagging
  - Still many constraints
Image tagging

• Image retrieval systems use tags

• Human tagging
  - Accurate, widely used
  - Slow and boring

• Image based auto-tagging
  - Still many constraints
Image tagging

- Image retrieval systems use tags
- Human tagging
  - Accurate, widely used
  - Slow and boring
- Image based auto-tagging
  - Still many constraints
Smartphone, the wild card

- Today’s smartphones have powerful built-in sensors
- People always carry their phones
Smartphone, the wild card

- Today’s smartphones have powerful built-in sensors
- People always carry their phones
TagSense

- A system for auto-tagging, with smartphone sensors
- Leverages multiple sensing domains
An example
An example

Smartphones
What happens at picturing time
What happens at picturing time
TagSense architecture

Figure 2: TagSense architecture – the camera phone triggers sensing in participating mobile phones and in ad hoc mode.

Knowing which phones/people are in the picture, TagSense analyzes these person activities, and also infers some contextual information from its own sensors, including the phones' accelerometer/compass readings. Phones take multiple snapshots, and then correlate them to the motion derived from the subjects. The context information is adequate for disambiguating the when, where, who, what, and other application-centric tagging, and also improves the confidence of the posing signature. As the motion vectors for the shots are often faced towards the camera, TagSense leverages these person activities, and also infers some contextual information from its own sensors, including the phones' accelerometer/compass readings. Phone to phone communication is performed using the WiFi, and retrieves additional tags. These tags are then organized and uploaded into a specified repository for image search.

In some cases, the context information is adequate for disambiguating the when, where, who, what, and other application-centric tagging, and also improves the confidence of the posing signature. Knowing which phones/people are in the picture, TagSense infers a "mutually facing" relationship; this heuristic rectifies this problem. For pictures in which the subjects do not pose explicitly, the TagSense shots and then correlated to the motion derived from the subjects are computed from the sequence of snaps.

However, some measurements require CPU-intensive processing, and others rely on external databases, like GPS, to address privacy and other application-specific needs. In these cases, TagSense exports the measurements to a cloud format as follows:

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

This section visits the design of these techniques in detail.
TagSense architecture
Figure 2: TagSense architecture

The camera phone triggers sensing in participating mobile phones and non-posing people in the picture and tags it accordingly. Briefly, when Bob is ready to take the picture, he activates the camera on his phone. Bob's phone immediately broadcasts a beacon encrypted with the shared key. Phones in the group activate their respective sensors and send the beacon with its local timestamp and the phones' acceleration dimensions to Bob's phone. Tags are derived from the visual and contextual information from its own sensors, including the phones' and camera's compass directions. Knowing which phones/people are in the picture, TagSense extracts a pause signature from the accelerometer readings. This improves the confidence of the posing signature. As a result, the subjects are computed from the sequence of snaps in which the subjects do not pose explicitly. The TagSense architecture is confined only to group members. Thus, Bob's phone must tell which phones/owners belong to him. Phone-to-phone communication is performed using the WiFi ad-hoc format as follows:

- WiFi Ad-hoc
- Acc
- Gyro
- Comp
- GPS
- Cam
- Mic
- Cloud

The next step is to process the data using the WiFi Ad-hoc format and store the results in a specified repository for image search. Date visits the design of these techniques in detail.
TagSense architecture

Figure 2: TagSense architecture – the camera phone triggers sensing in participating mobile phones and in-camera view are often faced towards the camera.

TagSense adopts three mechanisms. When people take a picture of Alice in a crowded place, the picture is often not taken towards the camera. The person activities are received from each phone and send them back to Bob's phone. Bob's phone assigns the activity to the owner of the phone and retrieves additional tags. These tags are then ors and other applications.

Knowing which phones/people are in the picture, TagSense exports the measurements to a cloud. The ability to recover the context information only from these phones when where who what.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be standing.

Once Bob is ready to take the picture, he activates the smartphone's sensors and sends a beacon with its local timestamp and the phones' accelerometer/compass readings. Phones in the group activate their respective sensors to retrieve additional tags. These tags are then ors and other applications.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be standing.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be standing.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be standing.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be standing.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be standing.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be standing.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be standing.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be standing.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be sitting.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be sitting.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be sitting.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be sitting.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be sitting.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be sitting.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.

In some cases, the context information is adequate for disambiguating the activity. For example, a person sitting in a chair is likely to be sitting, while a person standing in a chair is likely to be sitting.

However, some measurements require CPUs-intensive processes and cannot be effectively done in the smartphone's local memory. In these cases, TagSense exports the measurements to an external database. In these cases, TagSense exports the measurements to a cloud.
**TagSense architecture**

- **Bob**
- **Srihari**
- **John**
- **Chuan Cameraman**
- **Peter**

The diagram illustrates the TagSense architecture, where mobile phones are activated to sense data. The sensing data is collected and processed by Bob's phone, which acts as a shared key to ensure the security of the information. The data is then sent to Bob's phone for further processing. The diagram also shows the integration of WiFi Ad-hoc and processing for activities such as laughter recognition and others that rely on external databases and GPS.
TagSense architecture

This section zooms into the design and implementation of TagSense, focusing on the architecture. The diagram illustrates the TagSense system, where participants use their phone's camera to capture images of the scene. The captured images are then processed to identify and tag objects or people based on the sensed information.

Knowing which phones/people are in the picture is crucial for TagSense to function effectively. The system adopts three mechanisms to achieve this:

1. **Camera View**: Utilizes the phone's camera and sensors to detect and process visual information.
2. **WiFi Ad-hoc**: Leverages the WiFi ad-hoc network to communicate between devices.
3. **Sensing data**: Analyzes data from the phone's accelerometer and other sensors to infer user activities and orientation.

The architecture also includes a cloud-based repository where processed images and additional tags are uploaded for further analysis and retrieval.

Figure 2: TagSense architecture – the camera phone triggers sensing in participating mobile phones and ad hoc mode.

The system ensures privacy remains preserved, and the names of people are not tagged with the picture unless they explicitly pose for the photo. This approach improves the confidence of the posing signature as it leverages the phones' and camera's compass directions along with other contextual information.
TagSense architecture

The person activities are received from each phone location, ambient sound, light, etc., and contextual information from its own sensors, including the phones' accelerometer, compass, etc. Once Bob is ready to take the picture, he activates the camera and takes multiple snapshots. The motion vectors for the subjects are computed from the sequence of snaps, and then correlated to the motion derived from the phones' accelerometer, compass readings. Phones sends a beacon with its local timestamp and the phones' shared key – phones in the group activate their respective sensors. Privacy remains preserved.

When people explicitly pose for the picture, TagSense extracts a pause signature with a high correlation. The pause signature correlates well with the timing of the people's threat. As a result, the confidence of the posing signature improves. However, some measurements require CPU-intensive processing, and others rely on external databases, e.g., GPS, to get the addresses. In these cases, the context information is adequate for deriving tagging. For pictures, the context information is used for tagging; the next three sections zoom into the design and implementation of these techniques in detail.

In some cases, the context information is adequate for deriving tagging. This section zooms into the design and implementation of these techniques in detail. In the TagSense architecture, the camera phone triggers sensing in participating mobile phones and in-ad hoc mode. The sensing data are processed to generate tags. The WiFi Ad-hoc network and sensing data are used for tagging. The next three sections zoom into the design and implementation of these techniques in detail.
TagSense tag generation
TagSense tag generation

- When?
- Where?
- What?
- Who?
TagSense tag generation

- When?
- Where?
- What?
- Who?

May 4th afternoon

Clock + GPS + WiFi = May 4th afternoon
TagSense tag generation

- **When?**
  - Clock
  - GPS
  - WiFi
  - = May 4th afternoon

- **Where?**
  - GPS
  - WiFi
  - Comp
  - Light s.
  - = State house, outdoor

- **What?**

- **Who?**

---

*May 4th afternoon
State house, outdoor*
TagSense tag generation

- **When?**
  - Clock + GPS + WiFi = May 4th afternoon

- **Where?**
  - GPS + WiFi + Comp + Light s. = State house, outdoor

- **What?**
  - Acc + Gyro + Mic + Cloud = Standing, talking, sunny

- **Who?**
TagSense tag generation

- When?
  - Clock + GPS + WiFi
  = May 4th afternoon

- Where?
  - GPS + WiFi + Comp + Light s.
  = State house, outdoor

- What?
  - Acc + Gyro + Mic + Cloud
  = Standing, talking, sunny

- Who?
  = Bob, Sam, John
The challenge of ‘Who’

- Include only those in camera view
- Why not use localization?
- New opportunities enabled by multi-dimensional sensing
Possible opportunities

1. Accelerometer based motion signatures
2. Complementary compass directions
3. Correlating visual and acceleration
1. Accelerometer based motion signatures
1. Accelerometer based motion signatures

Move into posing
1. Accelerometer based motion signatures

- Move into posing
- Stable when posing

- Standing still
- Walking
- Standing still
- Standing still
- Standing still

Wednesday, August 3, 2011
1. Accelerometer based motion signatures

- Move into posing
- Stable when posing
- Move out of posing
1. Accelerometer based motion signatures
Is Bob a special case?

Acceleration plotting for 50+ pictures

People inside the picture:

People outside the picture:
Is Bob a special case?

Acceleration plotting for 50+ pictures

People inside the picture:

People outside the picture:

Picturing time

Variance

Time of picture-click

Time (s)

Variance

Time of picture-click

Time (s)
Is Bob a special case?

Acceleration plotting for 50+ pictures

People inside the picture:

People outside the picture:

Variance

Time (s)

Picturing time

Motion

Still

Motion

Time of picture-click

Time of picture-click

Wednesday, August 3, 2011
Is Bob a special case?

- People inside the picture:
  - Motion
  - Still
  - Motion

- People outside the picture:
  - Motion
  - Motion
  - Motion

Acceleration plotting for 50+ pictures.
Possible opportunities

1. Accelerometer based motion signatures
2. Complementary compass directions
3. Correlating visual and acceleration
2. Complementary compass directions

People in the picture are likely to face the camera
2. Complementary compass directions

People in the picture are likely to face the camera.
2. Complementary compass directions

People in the picture are likely to face the camera.
Is someone facing the camera?

- Compass reading != User’s orientation
- The diff: Personal Compass Offset (PCO)
- Need to calibrate PCO
  - Use posing picture for calibration
Possible opportunities

1. Accelerometer based motion signatures
2. Complementary compass directions
3. Correlating visual and acceleration
3. Correlating visual and acceleration motion vectors extracted from visual
3. Correlating visual and acceleration

Motion vectors extracted from visual
3. Correlating visual and acceleration motion vectors extracted from visual
Correlating visual and acceleration

- Taking several snapshots after shutter click
- Motion vector can be obtained by optical flow from adjacent snapshots
- Correlate with accelerometer readings to find who
Correlating visual and acceleration

- Taking several snapshots after shutter click
- Motion vector can be obtained by optical flow from adjacent snapshots
- Correlate with accelerometer readings to find who

Combining the Opportunities

Given a picture TagSense attempts to leverage the above opportunities for tagging it with the names of people. For this it 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 PC. If the posing signature is present the person is immediately deemed to be in the picture. In the absence of the posing signatures TagSense checks if the person is reasonably static. If the person is static and the person’s facing direction makes less than $\pm 45^\circ$ angle with the camera’s direction then the user is also included. Finally if the person is not static TagSense computes the picture’s optical motion vectors and correlates with the person’s accelerometer readings. The person is included upon a high match.

Points of Discussion

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 only the names is still valuable for applications such as image search and retrieval.

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

7
Taking several snapshots after shutter click

Motion vector can be obtained by optical flow from adjacent snapshots

Correlate with accelerometer readings to find who

Correlating visual and acceleration

Points of Discussion

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 only the names is still valuable for applications such as image search and retrieval.

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. TagSense works with any such device that has a wireless footprint and basic sensors. Of course, babies will still be hard, and babies may be a reason for taking many pictures.
Putting all together

Sensing data from phones
Putting all together

Sensing data from phones
Pose Signature
Putting all together

Sensing data from phones

Pose Signature

All Posing?
Putting all together

Sensing data from phones
Pose Signature
All Posing?

Y

END

Wednesday, August 3, 2011
Putting all together

Sensing data from phones

Pose Signature

All Posing?

Y

N

Motion?

END

Wednesday, August 3, 2011
Putting all together

Sensing data from phones

Pose Signature

All Posing?

N

Motion?

Y

Correlate Acc & Visual

END
Putting all together

Sensing data from phones

Pose Signature

All Posing?

N

Motion?

Y

Correlate Acc & Visual

All?

Y

END
Putting all together

Sensing data from phones

Pose Signature

All Posing?

N

Motion?

Y

Correlate Acc & Visual

All?

Y

END

Wednesday, August 3, 2011
Putting all together

Sensing data from phones
  Pose Signature
  All Posing? Y
  N
  Motion? Y
    Correlate Acc & Visual
    N
    Compass Direction
    All? Y
  N
  END
Putting all together

Sensing data from phones

Pose Signature

All Posing?

Motion?

Correlate Acc & Visual

Compass Direction

All?

END
TagSense evaluation

• A prototype on Android Nexus One phones
• Evaluated TagSense with 200+ pictures
• Compare people tagging results with Picasa & iPhoto
Evaluation for tagging people

Figure 7: Performance of TagSense: Top and 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).

Recognized
Missed

Picasa

Correctly Included by Picasa
Wrongly Excluded by Picasa

iPhoto

Correctly Included by iPhoto
Wrongly Excluded by iPhoto

TagSense

Correctly Included by TagSense
Wrongly Excluded by TagSense
Evaluation for tagging people

Precision/Recall/Fallout

\[
\text{Precision} = \frac{|\text{People Inside} \cap \text{Tagged by TagSense}|}{|\text{Tagged by TagSense}|}
\]

\[
\text{Recall} = \frac{|\text{People Inside} \cap \text{Tagged by TagSense}|}{|\text{People Inside}|}
\]

\[
\text{Fallout} = \frac{|\text{People Outside} \cap \text{Tagged by TagSense}|}{|\text{People Outside}|}
\]

TagSense achieves better Recall with some sacrifice on Precision
TagSense case 1

December 4th afternoon,
Hudson Hall, outdoor,
standing, snowing,
Xuan
TagSense case 2

November 21st afternoon,
Nasher Museum, indoor,
Romit, Sushma, Naveen,
Souvik, Justin, Vijay, Xuan,
standing, talking
TagSense case 3

November 21st noon,
Duke Wilson Gym, indoor,
playing, music,
Chuan, Romit
TagSense case 4 (failure case)

November 21st evening,
155 main street, indoor,
Talking,
(Not recognized)
TagSense case 4 (failure case)

November 21st evening,
155 Main Street, Indoor,
Talking,
(Not recognized)

Image processing and smartphone sensing could be complementary
Related works

• Image processing based auto-tagging

• Semi-auto tagging
  - Cloud based [Naaman JDCL2005] [Sarvas MobiSys2004]
  - Crowd-Sourcing based solution [CrowdSearch MobiSys2010]

• Special hardware
  - ContextCam[Shwetak UbiComp2004]
Limitations of TagSense

- TagSense cannot tag pictures taken in the past
- TagSense can not recognize people not carrying phones
- TagSense’s current vocabulary is quite limited
TagSense revision & future work

- TagSense -- emulating human recollection
  - Sensors are similar to human senses
  - Recorded sensing data is similar to human memory

- Future: tag generation is a work in progress
  - Short time scale, multiple sensing dimensions
  - Extensible: Face/Activity recognition techniques also fit in
    New sensors are being added to smartphones
  - A door to many possibilities: video-tagging, augmented reality, ...
TagSense revision & future work

Thanks & Questions?

- **TagSense** – emulating human recollection
  - Sensors are similar to human senses
  - Recorded sensing data is similar to human memory

- **Future**: tag generation is a work in progress
  - Short time scale, multiple sensing dimensions
  - Extensible: Face/Activity recognition techniques also fit in
    New sensors are being added to smartphones
  - A door to many possibilities: video-tagging, augmented reality, ...