Egocentric Field-of-View Localization Using First-Person Point-of-View Devices

Vinay Bettadapura[1,2]
vinay@gatech.edu

Irfan Essa[1,2]
irfan@cc.gatech.edu

Caroline Pantofaru[1]
cpantofaru@google.com

[1]Google Inc., Mountain View, CA, USA
[2]Georgia Institute of Technology, Atlanta, GA, USA
http://www.vbettadapura.com/egocentric/localization

Abstract

We present a technique that uses images, videos and sensor data taken from first-person point-of-view devices to perform egocentric field-of-view (FOV) localization. We define egocentric FOV localization as capturing the visual information from a person's field-of-view in a given environment and transferring this information onto a reference corpus of images and videos of the same space, hence determining what a person is attending to. Our method matches images and video taken from the first-person perspective with the reference corpus and refines the results using the first-person's head orientation information obtained using the device sensors. We demonstrate single and multi-user egocentric FOV localization in different indoor and outdoor environments with applications in augmented reality, event understanding and studying social interactions.

1. Introduction

A key requirement in the development of interactive computer vision systems is modeling the user, and one very important question is *"What is the user looking at right now?"* From augmented reality to human-robot interaction, from behavior analysis to healthcare, determining the user's egocentric field-of-view (FOV) accurately and efficiently can enable exciting new applications. Localizing a person in an environment has come a long way through the use of GPS, IMUs and other signals. But such localization is only the first step in understanding the person's FOV.

The new generation of devices are small, cheap and pervasive. Given that these devices contain cameras and sensors such as gyros, accelerometers and magnetometers, and are Internet-enabled, it is now possible to obtain large amounts of first-person point-of-view (POV) data unintrusively. Cell phones, small POV cameras such as GoPros, and wearable technology like Google Glass all have a suite of similar useful capabilities. We propose to use data from

these first person POV devices to derive an understanding of the user's egocentric perspective. In this paper we show results from data obtained with Google Glass, but any other device could be used in its place.

Automatically analyzing the POV data (images, videos and sensor data) to estimate egocentric perspectives and shifts in the FOV remains challenging. Due to the unconstrained nature of the data, no general FOV localization approach is applicable for all outdoor and indoor environments. *Our insight is to make such localization tractable by introducing a reference data-set*, i.e., a visual model of the environment, which is either pre-built or concurrently captured, annotated and stored permanently. All the captured POV data from one or more devices can be matched and correlated against this reference data-set allowing for transfer of information from the user's reference frame to a global reference frame of the environment. The problem is now reduced from an open-ended data-analysis problem to a more practical data-matching problem. Such reference data-sets already exist; *e.g.,* Google Street View imagery exists for most outdoor locations and recently for many indoor locations. Additionally, there are already cameras installed in many venues providing pre-captured or concurrently captured visual information, with an ever increasing number of spaces being mapped and photographed. Hence there are many sources of visual models of the world which we can use in our approach.

Contributions: We present a method for egocentric FOV localization that directly matches images and videos captured from a POV device with the images and videos from a reference data-set to understand the person's FOV. We also show how sensor data from the POV device's IMU can be used to make the matching more efficient and minimize false matches. We demonstrate the effectiveness of our approach across 4 different application domains: (1) egocentric FOV localization in outdoor environments: 250 POV images from different locations in 2 major metropolitan cities matched against the street view panoramas from those
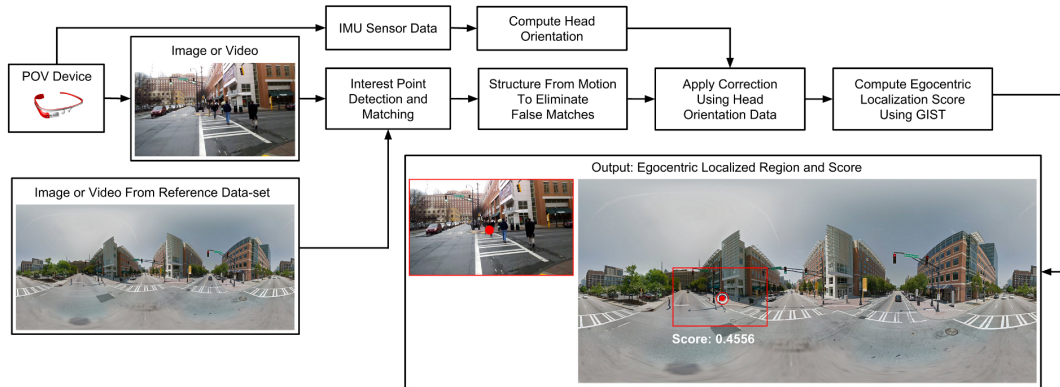
Figure 1. An overview of our egocentric FOV localization system. Given images (or videos) and sensor data from a POV device, and a pre-existing corpus of canonical images of the given location (such as Google street view data), our system localizes the egocentric perspective of the person and determines the person's region-of-focus.

locations; (2) egocentric FOV localization in indoor spaces: a 30 minute POV video in an indoor presentation matched against 2 fixed videos cameras in the venue; (3) egocentric video tours at museums: 250 POV images of paintings taken within 2 museums in New York City (Metropolitan Museum of Art and Museum of Modern Art) matched against indoor street view panoramas from these museums (available publicly as part of the Google Art Project [1]); and (4) joint egocentric FOV localization from multiple POV videos: 60 minutes of POV videos captured concurrently from 4 people wearing POV devices at the Computer History Museum in California, matched against each other and against indoor street view panoramas from the museum.

2. Related Work

Localization: Accurate indoor localization has been an area of active research [12]. Indoor localization can leverage GSM [24], active badges [31], 802.11b wireless ethernet [16], bluetooth and WAP [2], listeners and beacons [25], radiofrequency [3] technologies and SLAM [18].

Outdoor localization from images or video has also been explored, including methods to match new images to streetside images [27, 32, 28]. Other techniques include urban navigation using a camera mobile phone [26], image geotagging based on travel priors [15] and the IM2GPS system [11].

Our approach leverages these methods for visual and sensor data matching with first-person POV systems to determine where the user is attending to.

Egocentric Vision and Attention: Detecting and understanding the salient regions in images and videos has been an active area of research for over three decades. Seminal efforts in the 80s and 90s focused on understanding saliency and attention from a neuroscience and cognitive psychology perspective [30]. In the late 90s, Illti *et al.* [14] built a vi-

sual attention model using a bottom-up model of the human visual system. Other approaches used graph based techniques [9], information theoretical methods [4], frequency domain analysis [13] and the use of higher level cues like face-detection [5] to build attention maps and detect objects and regions-of-interests in images and video.

In the last few years, focus has shifted to applications which incorporate attention and egocentric vision. These include gaze prediction [19], image quality assessment [22], action localization and recognition [29, 7], understanding social interactions [6] and video summarization [17]. Our goal in this work is to leverage image and sensor matching between the reference set and POV sensors to extract and localize the egocentric FOV.

3. Egocentric FOV Localization

The proposed methodology for egocentric FOV localization consists of five components: (i) POV data consisting of images, videos and head-orientation information, (ii) a pre-captured or concurrently captured reference dataset, (iii) robust matching pipeline, (iv) match correction using sensor data, and (v) global matching and score computation. An overview of our approach is shown in Figure 1. Each step of the methodology is explained in detail below.

Data collection: POV images and videos along with the IMU sensor data are collected using one or more POV devices to construct a "pov-dataset". For our experiments, we used a Google Glass. It comes equipped with a 720p camera and sensors such as accelerometer, gyroscope and compass that lets us effectively capture images, videos and sensor data from a POV perspective. Other devices such as cellphones, which come equipped with cameras and sensors, can also be used.

Reference dataset: A "reference-dataset" provides a visual model of the environment. It can either be pre-captured

(and possibly annotated) or concurrently captured (i.e. captured while the person with the POV device is in the environment). Examples of such reference datasets are Google Street View images and pre-recorded videos and live video streams from cameras in indoor and outdoor venues.

Matching: Given the person's general location, the corresponding reference image is fetched from the reference-dataset using location information (such as GPS) and is matched against all the POV images taken by the person at that location. Since the camera is egocentric, the captured image provides an approximation of the person's FOV. The POV image and the reference image are typically taken from different viewpoints and under different environmental conditions which include changes in scale, illuminations, camera intrinsics, occlusion and affine and perspective distortions. Given the "in-the-wild" nature of our applications and our data, our matching pipeline is designed to be robust to these changes.

In the first step of the matching pipeline, reliable interest points are detected both in the POV image, I_{pov}, and the reference image, I_{ref} using maximally stable extremal regions (MSER). The MSER approach was originally proposed by [20], by considering the set of all possible thresholdings of an image, I, to a binary image, I_B, where $I_B(x)=1$ if $I(x) \geq t$ and 0 otherwise. The area of each connected component in I_B is monitored as the threshold is changed. Regions whose rates of change of area with respect to the threshold are minimal are defined as maximally stable and are returned as detected regions. The set of all such connected components is the set of all extremal regions. The word extremal refers to the property that all pixels inside the MSER have either higher (bright extremal regions) or lower (dark extremal regions) intensity than all the pixels on its outer boundary. The resulting extremal regions are invariant to both affine and photometric transformations. A comparison of MSER to other interest point detectors has shown that MSER outperforms the others when there is a large change in viewpoint [21]. This is a highly desirable property since I_{pov} and I_{ref} are typically taken from very different viewpoints. Once the MSERs are detected, standard SIFT descriptors are computed and the correspondences between the interest points are found by matching them using a KD tree, which supports fast indexing and querying.

The interest point detection and matching process may give us false correspondences that are geometrically inconsistent. We use random sample consensus (RANSAC) [8] to refine the matches and in turn eliminate outlier correspondences that do not fit the estimated model. In the final step, the egocentric focus-of-attention is transferred from I_{pov} to I_{ref}. Using three of the reliable match points obtained after RANSAC, the affine transformation matrix, $\mathbf{A}$, between I_{pov} and I_{ref} is computed. The egocentric focus-of-attention $\mathbf{f_{pov}}$ is chosen as the center of I_{pov} (the red dot

in Figure 1). This is a reasonable assumption in the absence of eye-tracking data. The focus-of-attention, $\mathbf{f_{ref}}$, in I_{ref}, is given by $\mathbf{f_{ref}} = \mathbf{A} \mathbf{f_{pov}}$.

Correction using sensor data: The POV sensor data that we have allows us to add an additional layer of correction to further refine the matches. Modern cellphones and POV devices like Glass come with a host of sensors like accelerometers, gyroscopes and compasses and they internally perform sensor fusion to provide more stable information. Using sensor fusion, these devices report their absolute orientation in the world coordinate frame as a 3×3 rotation matrix R. By decomposing R, Euler angles ψ (yaw), θ (pitch), ϕ (roll) can be obtained. Since Glass is capturing sensor data from a POV perspective, the Euler angles give us the head orientation information, which can be used to further refine the matches. For example, consider a scenario where the user is looking at a high-rise building that has repetitive patterns (such as rectangular windows), all the way from bottom to the top. The vision-based matching gives us a match at the bottom of the building, but the head orientation information suggests that the person is looking up. In such a scenario, a correction can be applied to the match region to make it compatible with the sensor data.

Projecting the head orientation information onto I_{ref}, gives us the egocentric focus-of-attention, $\mathbf{f_s}$, as predicted by the sensor data. The final egocentric FOV localization is computed as: $\mathbf{f} = \alpha \mathbf{f_s} + (1 - \alpha) \mathbf{f_{ref}}$, where α is a value between 0 and 1 and is based on the confidence placed on the sensor data. Sensor reliability information is available in most of the modern sensor devices. If the device sensors are unreliable then α is set to a small value. Relying solely on either vision based matching or on sensor data is not a good idea. Vision techniques fail when the images are drastically different or have fewer features and sensors tend to be noisy and the readings drift over time. We found that first doing the vision based matching and then applying a α-weighted correction based on the sensor data gives us the best of both worlds.

Global Matching and Score Computation: We now have a match window that is based on reliable MSER interest point detection followed by SIFT matching and RANSAC based outlier rejection and sensor based correction. Although this match window is reliable, it is still based only on local features without any global context of the scene. There are several scenarios in the real world (like urban environments), where we have repetitive and commonly occurring patterns and local features that may result in an inaccurate match window. In this final step, we do a global comparison and compute the egocentric localization score.

Global comparison is done by comparing the match window, W_{ref} located around $\mathbf{f_s}$ in I_{ref}, with I_{pov} (*i.e.*, the red match windows of the bottom image in Figure 1). This

Figure 2. Egocentric FOV localization in outdoor environments. The images on the left are the POV images taken from Glass. The red dot shows the focus-of-attention. The panorama on the right shows the localization (target symbols) and the shifts in the FOV over time (red arrows). Note the change in season and pedestrian traffic between the POV images and the reference image.

Figure 3. Egocentric FOV localization in indoor environments. The images on the first column show the room layout. The presenter is shown in Green and the person wearing Glass is shown in Blue with his egocentric view shown by the blue arrow. The second column shows the POV video frames from Glass. The red dot shows the focus-of-attention. The third and fourth column show the presenter cam and the audience cam respectively. The localization is shown by the target symbol and the selected camera is shown by the red bounding box. The person wearing Glass is highlighted by the blue circle in the presenter camera views.

comparison is done using global GIST descriptors [23]. A GIST descriptor gives a global description of the image based on the image's spectral signatures and tells us how visually similar the two images are. GIST descriptors q_{pov} and q_{ref} are computed for I_{pov} and W_{ref} respectively and final egocentric FOV localization score is computed as the $L2$-distance between the GIST descriptors: $\| q_{pov} - q_{ref} \| = \sqrt{(q_{pov} - q_{ref}) . (q_{pov} - q_{ref})}$. Scoring quantifies the confidence in our matches and by thresholding on the score, we can filter out incorrect matches.

4. Applications and Results

To evaluate our approach and showcase different applications, we built 4 diverse datasets that include both images and videos in both indoor and outdoor environments. All the POV data was captured with a Google Glass.

4.1. Outdoor Urban Environments

Egocentric FOV localization in outdoor environments has applications in areas such as tourism, assistive tech-

nology and advertising. To evaluate our system, 250 POV images (of dimension 2528x1856) along with sensor data (roll, pitch and yaw of the head) was captured at different outdoor locations in two major metropolitan cities. The reference dataset consists of the 250 street view panoramas (of dimension 3584x1536) from those locations. Based on the user's GPS location, the appropriate street view panorama was fetched and used for matching. Ground truth was provided by the user who documented his point of attention in each of the 250 POV images. However we have to take into account the fact that we are only tracking the head orientation using sensors and not tracking the eye movement. Humans may or may not rotate their heads completely to look at something; instead they may rotate their head partially and just move their eyes. We found that this behavior (of keeping the head fixed while moving the eyes) causes a circle of uncertainty of radius R around the true point-of-attention in the reference image. To calculate its average value, we conducted a user-study with 5 participants. The participants were instructed to keep their heads still and use only their eyes to see as far to the left and to the right as
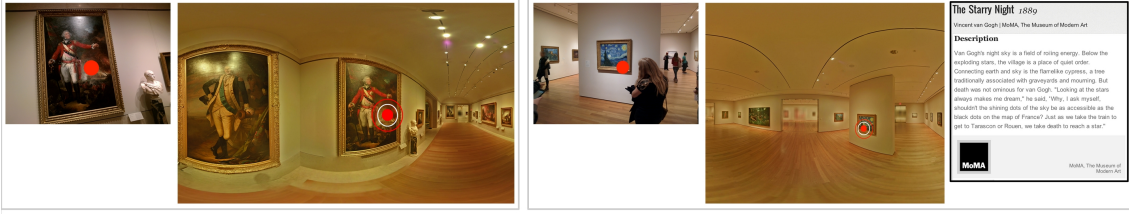
Figure 4. Egocentric FOV localization in indoor art installations. The images on the left are the POV images taken from Glass. The red dot shows the focus-of-attention. The images to their right are panoramas from indoor streetview that correctly shows the localization result (target symbol). When available, the details of the painting are shown. This information is automatically fetched, using the egocentric FOV location as the cue. For the painting on the right (Van Gogh's "The Starry Night"), an information card shows up and provides information about the painting.

they could without the urge to turn their heads. This mean radius of their natural eye movement was measured to be 330 pixels for outdoor urban environments. Hence for our evaluation we consider the egocentric FOV localization to be successful if the estimated point-of-attention falls within a circle of radius $R = 330$ pixels around the ground truth point-of-attention.

Experimental results show that without using sensor data, egocentric FOV localization was accurate in 191/250 images for a total accuracy of 76.4%. But when sensor data was included, the accuracy rose to 92.4%. Figure 2 shows the egocentric FOV localization results and the shifts in FOV over time. Discriminative objects such as landmarks, street signs, graffiti, logos and shop names helped in the getting good matches. Repetitive and commonly occurring patterns like windows and vegetation caused initial failures but most of them were fixed when the sensor correction was applied.

4.2. Presentations in Indoor Spaces

There are scenarios where a pre-built reference dataset (like street view) is not available for a given location. This is especially true for indoor environments that have not been as thoroughly mapped as outdoor environments. In such scenarios, egocentric FOV localization is possible with a reference dataset that is concurrently captured along with the POV data. To demonstrate this, a 30 minute POV video along with sensor data was captured during an indoor presentation. The person wearing Glass was seated in the audience in the first row. The POV video is 720p at 30 fps. The reference dataset consists of videos from two fixed cameras at the presentation venue. One camera was capturing the presenter while the other camera was pointed at the audience. The reference videos are 1080p at 30 fps. Ground truth annotations for every second of the video were provided by the user who wore Glass and captured the POV video. So, we have 60*30 = 1800 ground truth annotations. As with the previous dataset, we empirically estimated R to be 240 pixels. Experimental results show that egocen-
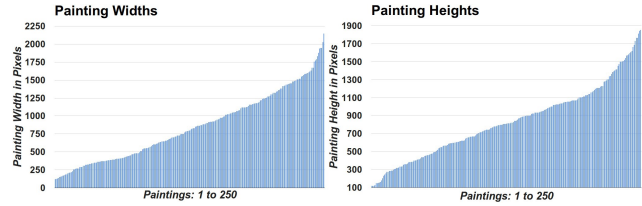

Figure 5. The widths and heights of the 250 paintings, sorted in ascending order based on their value. We can see that our dataset has a good representation of paintings of varying widths and heights.

tric FOV localization and camera selection was accurate in 1722/1800 cases for a total accuracy of 95.67%. Figure 3 shows the FOV localization and camera selection results.

4.3. Egocentric Video Tours in Museums

Public spaces like museums are ideal environments for an egocentric FOV localization system. Museums have exhibits that people explicitly pay attention to and want to learn more about. Similar to audio-tours that are available in museums, we demonstrate a system for attention-driven egocentric video tours. Unlike in an audio tour where a person has to enter the exhibit number to hear details about it, our video tour system recognizes the exhibit when the person looks at it and brings up a cue card on the wearable device giving more information about the exhibit.

For our evaluation, we captured 250 POV images of paintings at 2 museums in New York City - The Metropolitan Museum of Art and The Museum of Modern Art. The reference dataset consists of indoor street view panoramas from these museums, made available as part of the Google Art Project [1]. Since this dataset consists of paintings, which have a fixed structure (a frame enclosing the artwork), we have a clear definition of correctness: egocentric FOV localization is deemed to be correct if the estimated focus-of-attention is within the frame of the painting in the reference image. Experimental results show that the localization was accurate in 227/250 images for a total accuracy of 90.8%. Figure 5 shows the distribution of the widths and heights of the paintings in out dataset. We can see that paintings of all widths and heights are well represented.

The Google Art panoramas are annotated with information about the individual paintings. On successful FOV localization, we fetch the information on the painting that the person is viewing and display it on Glass or as an overlay. Figure 4 shows the FOV localization results and the painting information that was automatically fetched and shown on Glass.

4.4. Joint Egocentric FOV Localization

When we have a group of people wearing POV devices within the same event space, egocentric FOV localization becomes much more interesting. We can study joint FOV localization (i.e. when two or more people are simultaneously attending to the same object), understand the social dynamics within the group and gather information about the event space itself.

Joint FOV localization can be performed by matching the videos taken from one POV device with the videos taken from another POV device. If there are n people in the group, $P = \{p_i | i \in [1, n]\}$, then we have n POV videos: $V = \{v_i | i \in [1, n]\}$. In the first step, all the videos in V are synchronized by time-stamp. In the second step, k videos (where $k \leq n$) are chosen from V and matched against each other, which results in a total of $\binom{n}{k}$ matches. Matching is done frame-by-frame, by treating frame from one video as I_{pov} and the frames from the other videos as I_{ref}. By thresholding the egocentric FOV localization scores, we can discover regions in time when the k people were jointly paying attention to the same object. Finally, in the third step, the videos can be matched against the reference imagery from the event space to find out *what* they were jointly paying attention to.

We conducted our experiments with $n = 4$ participants. The 4 participants wore Glass and visited the Computer History Museum in California. They were instructed to behave naturally, as they would on a group outing. They walked around in the museum looking at the exhibits and talking with each other. A total of 60 minutes of POV videos and the corresponding head-orientation information were captured from their 4 Glass devices. The videos are 720p at 30fps. The reference dataset consists of indoor street view panoramas from the museum. Next, joint egocentric FOV localization was performed by matching pairs of videos against each other, i.e. $k = 2$, for a total of 6 pairs of matches. Figure 7 shows the results for 25,000 frames of video for all the 6 match pairs. The plot shows the instances in time when groups of people were paying attention to the same exhibit. Furthermore, we get an insight into the social dynamics of the group. For example, we can see that P2 and P3 were moving together but towards the end P3 left P2 and started moving around with P1. Also, there are instances in time when all the pairs of videos match which indicates that the group came together as a whole. One such instance is
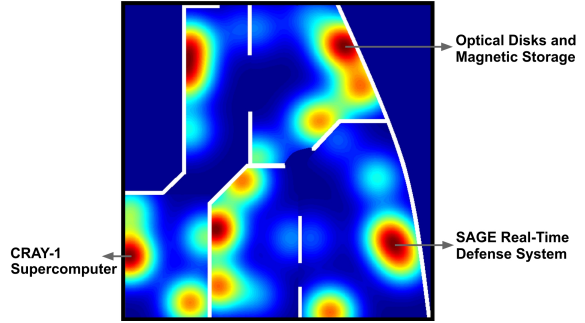

Figure 6. A heatmap overlaid on a section of the Computer History Museum's floorplan. Hotter regions in the map represent exhibits which had joint egocentric attention from more people. Three of the hottest regions are labeled to show the underlying exhibits that brought people together and probably led to further discussions among them.

highlighted in Figure 7 by the orange vertical line. There are also instances when the 4 people split into two groups. This is shown by the green vertical line in Figure 7.

Joint egocentric FOV localization also helps us get a deeper understanding of the event space. Interesting exhibits tend to bring people together for a discussion and result in higher joint egocentric attention. It is possible to infer this from the data by matching the videos with the reference images and labeling each exhibit with the number of people who jointly viewed it. By overlaying the exhibits on the floorplan, we can generate a heat map of the exhibits where hotter regions indicate more interesting exhibits that received higher joint attention. This is shown in Figure 6. Getting such an insight has practical applications in indoor space planning and the arrangement and display of exhibits in museums and other similar spaces.

5. Discussion

One of the assumptions in the paper is the availability of reference images in indoor and outdoor spaces. This may not be true for all situations. Also, it may not be possible to capture reference data concurrently (as in the indoor presentation dataset) due to restrictions by the event managers and/or privacy concerns. However, our assumption does holds true for a large number of indoor and outdoor spaces which makes the proposed approach practical and useful.

There are situations where the proposed approach may fail. While our matching pipeline is robust to a wide variation of changes in the images, it may still fail if the reference image is drastically different from the POV image (for example, a POV picture taken in summer matched against a reference image taken on a white snowy winter). Another reason for failure could be when the reference dataset is outdated. In such scenarios, the POV imagery will not match well with the reference imagery. However these drawbacks are only temporary. With the proliferation of cameras and

the push to map and record indoor and outdoor spaces, reference data for our approach will only become more stable and reliable.
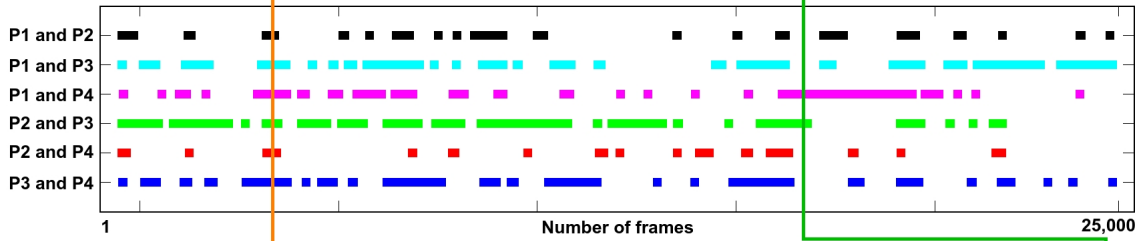
Our reference images are 2D models of the scene (for example, Street View panoramas). Moving to 3D reference models could provide a more comprehensive view of the event space and result in better FOV localization. But this would require a computationally intensive matching pipeline which involves 2D to 3D alignment and pose estimation.

6. Conclusion

We have demonstrated a working system that can effectively localize egocentric FOVs, determine the person's point-of-interest, map the shifts in FOV and determine joint attention in both indoor and outdoor environments from one or more POV devices. Several practical applications were presented on "in-the-wild" real-world datasets.

References

[1] Google Art Project. http://www.google.com/culturalinstitute/project/art-project. 2, 5

[2] L. Aalto, N. Göthlin, J. Korhonen, and T. Ojala. Bluetooth and WAP push based location-aware mobile advertising system. In *Int. Conf. Mobile systems, applications, and services*, pages 49–58. ACM, 2004. 2

[3] P. Bahl and V. N. Padmanabhan. Radar: An in-building RF-based user location and tracking system. In *INFOCOM*, volume 2, pages 775–784. IEEE, 2000. 2

[4] N. Bruce and J. Tsotsos. Saliency based on information maximization. *NIPS*, 18:155, 2006. 2

[5] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In *NIPS*, 2007. 2

[6] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *CVPR*, pages 1226–1233, 2012. 2

[7] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *ECCV*, pages 314–327. 2012. 2

[8] M. A. Fischler and R. C. Bolles. Random sample consensus. *Communications of the ACM*, 24(6):381–395, 1981. 3

[9] J. Harel, C. Koch, P. Perona, et al. Graph-based visual saliency. *NIPS*, 19:545, 2007. 2

[10] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[11] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *CVPR*, pages 1–8, 2008. 2

[12] J. Hightower and G. Borriello. A survey and taxonomy of location systems for ubiquitous computing. *IEEE computer*, 34(8):57–66, 2001. 2

[13] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *Trans. PAMI*, 34(1):194–201, 2012. 2

[14] L. Itti, C. Koch, E. Niebur, et al. A model of saliency-based visual attention for rapid scene analysis. *Trans. PAMI*, 20(11):1254–1259, 1998. 2

[15] E. Kalogerakis, O. Vesselova, J. Hays, A. A. Efros, and A. Hertzmann. Image sequence geolocation with human travel priors. In *ICCV*, pages 253–260, 2009. 2

[16] A. M. Ladd, K. E. Bekris, A. Rudys, L. E. Kavraki, and D. S. Wallach. Robotics-based location sensing using wireless ethernet. *Wireless Networks*, 11(1-2):189–204, 2005. 2

[17] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, pages 3–2, 2012. 2

[18] J. J. Leonard and H. F. Durrant-Whyte. Mobile robot localization by tracking geometric beacons. *Trans. Robotics and Automation*, 7(3):376–382, 1991. 2

[19] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *ICCV*, 2013. 2

[20] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004. 3

[21] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International journal of computer vision*, 65(1-2):43–72, 2005. 3

[22] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barbba. Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric. In *ICIP*, volume 2, pages II–169, 2007. 2

[23] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001. 3

[24] V. Otsason, A. Varshavsky, A. LaMarca, and E. De Lara. Accurate GSM indoor localization. In *UbiComp*, pages 141–158, 2005. 2

[25] N. B. Priyantha, A. Chakraborty, and H. Balakrishnan. The cricket location-support system. In *Int. Conf. Mobile computing and networking*, pages 32–43. ACM, 2000. 2

[26] D. P. Robertson and R. Cipolla. An image-based system for urban navigation. In *BMVC*, pages 1–10, 2004. 2

[27] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *CVPR*, pages 1–7, 2007. 2

[28] G. Schroth, R. Huitl, D. Chen, M. Abu-Alqumsan, A. Al-Nuaimi, and E. Steinbach. Mobile visual location recognition. *Signal Processing Magazine, IEEE*, 28(4):77–89, 2011. 2

[29] N. Shapovalova, M. Raptis, L. Sigal, and G. Mori. Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization. In *NIPS*, pages 2409–2417, 2013. 2

[30] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. 2

[31] R. Want, A. Hopper, V. Falcao, and J. Gibbons. The active badge location system. *ACM Trans. on Information Systems*, 10(1):91–102, 1992. 2

[32] A. R. Zamir and M. Shah. Accurate image localization based on Google maps street view. In *ECCV*, pages 255–268, 2010. 2

Figure 7. The plot on the top shows the joint egocentric attention between groups of people. The x-axis shows the progression of time, from frame 1 to frame 25,000. Each row shows the result of joint egocentric FOV localization, i.e. the instances in time when pairs of people were jointly paying attention to the same exhibit in the museum. The orange vertical line indicates an instance in time when all the people (P1, P2, P3 and P4) were paying attention to the same exhibit. The green vertical line indicates an instance in time when P1 and P4 were jointly paying attention to an exhibit while P2 and P3 were jointly paying attention to a different exhibit. The corresponding frames from their Glass videos is shown. When matched to the reference street view images, we can discover the exhibits that the groups of people were viewing together and were probably having a discussion about. Details of the exhibit was automatically fetched from the reference dataset's annotation.