Dynamic Narratives for Heritage Tour

Anurag Ghosh(✉), Yash Patel, Mohak Sukhwani, and C.V. Jawahar

CVIT, IIIT Hyderabad, Hyderabad, India
anurag.ghosh@research.iiit.ac.in

Abstract. We present a dynamic story generation approach for the egocentric videos from the heritage sites. Given a short video clip of a 'heritage-tour' our method selects a series of short descriptions from the collection of pre-curated text and create a larger narrative. Unlike in the past, these narratives are not merely monotonic static versions from simple retrievals. We propose a method to generate on the fly dynamic narratives of the tour. The series of the text messages selected are optimised over length, relevance, cohesion and information simultaneously. This results in 'tour guide' like narratives which are seasoned and adapted to the participants selection of the tour path. We simultaneously use visual and GPS cues for precision localization on the heritage site which is conceptually formulated as a graph. The efficacy of the approach is demonstrated on a heritage site, Golconda Fort, situated in Hyderabad, India. We validate our approach on two hours of data collected over multiple runs across the site for our experiments.

Keywords: Storytelling · Digital heritage · Egocentric perception

1 Introduction

Heritage sites are the places of interest commemorating people, places or events. These sites could be standalone structures or an agglomeration of multiple structures built across a large area. Spread across the site are the tales describing life and events over the centuries. Such tales are referred to as narratives in our present work. Digitalization and preservation attempts for heritage sites have ranged from methods dealing with restoration [1] to virtual reality based 3D re-constructions [2]. For the first time we attempt enhancing cultural diversity of heritage sites tour via a medium of text narration generation. We propose a method to create contextually aware, richer and multi-facet long descriptions instead of small 'tags'. Describing an area as "Built in twelfth century, this gate serves as an entrance to the main site. Accompanying eleven other gates, it is most remarkable of all and is located on the eastern..." is far more apt and relevant than a static annotation – 'Site entrance'.

In recent past, we have witnessed an increased interest in the use of computer vision and localization algorithms [3] to create digital representations for each

A. Ghosh and Y. Patel—Equal Contribution.

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part I, LNCS 9913, pp. 856–870, 2016.
DOI: 10.1007/978-3-319-46604-0_59

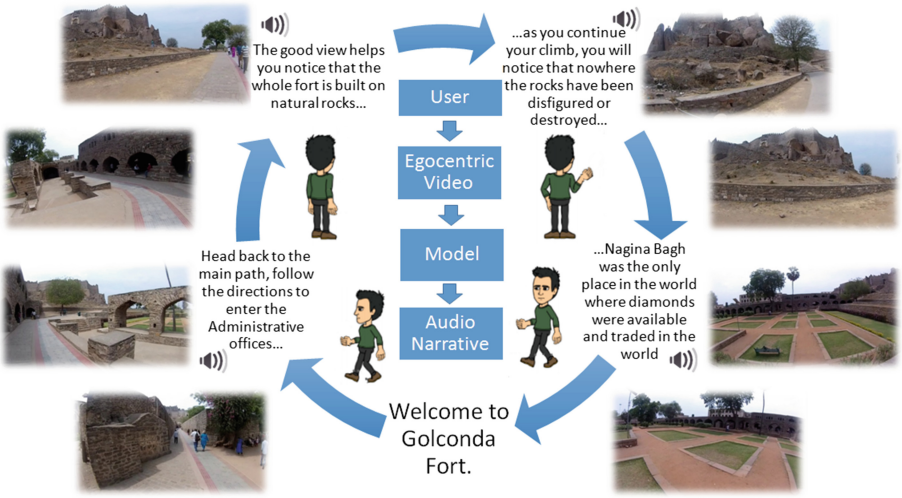

Fig. 1. Dynamic Narration generation for tour videos: Narrative generation using both vision cues and GPS tags while user walks across the site. Unlike past, the site specific audio narrations are detailed and optimized over the various parameters such as relevance and content.

of the aspects of cultural heritage sites [2,4]. Use of mobile based applications [5] to identify monuments on the fly by taking images have also gained significant traction. Story driven visual localization and summarization have also been attempted on egocentric videos [6]. We propose a method to generate text based summary for a given input egocentric tour video. Existing design selections for such vision based assistive technology are all centred around the utility for differentially abled people [7–11]. We propose a new use case which not (necessarily) caters to visually impaired users but in general to all. We suggest a new exemplar for vision based assistive technology with its use ranging from entertainment to educational purposes.

Accurrate localization is elusive in many situations. GPS based localization using mobile devices has proven unreliable in cases of short range distances [12]. Indoor environments where satellite signals find difficult to penetrate are also dead zones for GPS tags. People have used additional cues such as wi-fi signal strength [13], special purpose hardwares, blue-tooth [14] and probabilistic framework formulations over GPS to improve localization tags. Schroth et al. [15] use mobile recordings as a visual cue of the environment and match them to a geo-referenced database to provide pose information. In this work, we use combination of GPS and vision based localization to determine participants present location (Fig. 2). We use these location awareness cues to retrieve (from corpus of text describing the heritage area) the most relevant story of nearest heritage monument. Our stories comprise of details about the monument, its historical importance, events occurred and even includes the details of nearby structures.

Fig. 2. Need for precision localization: (best viewed in color) GPS signal in its raw form is noisy. Precise localization at GPS deadzones (and otherwise too) are achieved using vision cues. Such localization is achieved by matching current image with a set of GPS tagged images in the database. Inaccurate localization could lead to invalid semantic associations. (Color figure online)

Such a method of narrative generation imparts freedom to participants movement across the site and has potential to contribute to their learning.

The ability to create human like narratives to depict immersive and holistic heritage representations remains a challenging task. The proposed method needs to adapt to various external conditions simultaneously. The suggested localization scheme should be precise enough to generate valid semantic associations. Vision based multimedia retrieval methods must be real time and prone to illumination variations. Weather changes, pollution and other environmental degradations pose a direct threat to heritage monuments. The recognition module must be susceptible to ever changing (degrading) heritage structures. This requires a massive manual effort to map the full heritage area with relevant images (for localization) and text (for narration generation). Every monument needs to be tagged with multiple images from varied view-points and perspectives. We even need to have huge curated text corpus with site-specific information. Extensions of such kind to other monuments would thus require manual mapping of the whole site. This makes it extremely taxing to attempt a problem of this kind at such a scale.

Dynamic acquisition of narratives for heritage sites involves both content selection and ordering to create a seamless immersive experience for the participant. We use both vision cues and GPS measurements to design a method that generates narratives for heritage-site videos. The proposed solution generates narratives simultaneously optimized over content, coherence and relevance. This enables the generated narratives to be informative and dynamic in nature, i.e. it coexists with participants interaction and exploration of the space. We use the term 'relevant' to portray the applicability of the text with respect to the surrounding and the word 'informative' describes the 'comprehensive' nature of the selected narrative text. This paper is organized as follows: we describe the

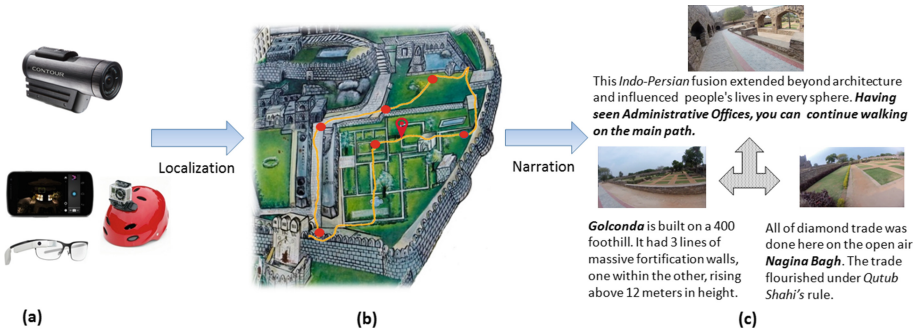

Fig. 3. Overview of the approach: (a) During training time, the site is mapped using images and high precision GPS sensors. At test instance, the participant can use any video capture device. (b) The heritage site is conceptualized as a graph and the user is localized on it using both vision and GPS cues. (c) The narrative generated are dynamic in nature and varied path selection lead to different narratives.

intricacies of narration generation task and discuss associated motivations in Sect. 2. User localization using visual cues are outlined in Sect. 3.1. The challenges associated with combining vision and natural language based processing are explicitly discussed in Sects. 3.2 and 3.3. The penultimate section describes the experiments on captured two hours of egocentric video of the tour.

2 Problem Statement

Storytelling using vision based approaches is still in the nascent stages. Majority of the work in this domain is focused on creation of visual storyline graphs of images summarizing series of events that have chronological relationship. Using methods described in [16,17] one can create such structural stories using collection of images, [16] uses friendship graphs to generate better stories. Similar to [18] we focus on creating interactive narrative systems in which details of heritage sites is unfolded steadily. While participants stroll around the site, the narrations adapt according to the participant's position. Dynamic story generation involves both content selection and ordering of the summaries depending participants present position and his overall tour path. The path selected and the speed of the trek determine the text selection. Thus, the variability is scripted at design time depending on user's preference. Figure 1 depicts one such use case – given an egocentric video of the stroll around the heritage site, we generate relevant narratives of the tour. Recent attempts for caption generation [19–21] do not bind variabilities at run time – participant's speed and pathway selection. Once trained these methods [19,20] resolve video details by learning temporal structure in the input sequence and associate these details to text. Deep learning based approaches [21] directly map videos to full sentences and can handle variable-length input videos but they do not scale to long tour videos described using mutliline text outputs.

Our problem can be stated as dynamic narration generation, i.e. given an egocentric video feed of the visitor we form a sequence of summaries aligned with the sequence of heritage sites that are being visited. Narrations in our case are the amalgamation of basic monument details – its historical importance and the details about nearby structures and form an integral part of cultural tourism of heritage sites. The ego-centric video frames are prone to camera motions which makes the inference of the heritage site or the location from the video frames an extremely challenging task. Factors like environmental and lighting conditions along with the crowd makes the problem even harder. GPS data in its raw form are error prone until explicit corrections are applied on measurements, as we can see in Fig. 2. We aim to generate an aligned set of relevant and cohesive narratives to maintain the engagement levels of participants high. The method is expected to operate at real time and within the computational and storage capacities of the device. Also, the method should be independent of GPS signal obtained from the visitors device as they are not very accurate in indoors and other GPS dead zones. The readers should note that the focus of the approach is to retrieve the best narrations linked to a particular GPS signal from the pre-curated text collection. We assume that sentences in the text collection are both grammatically and syntactically well formed.

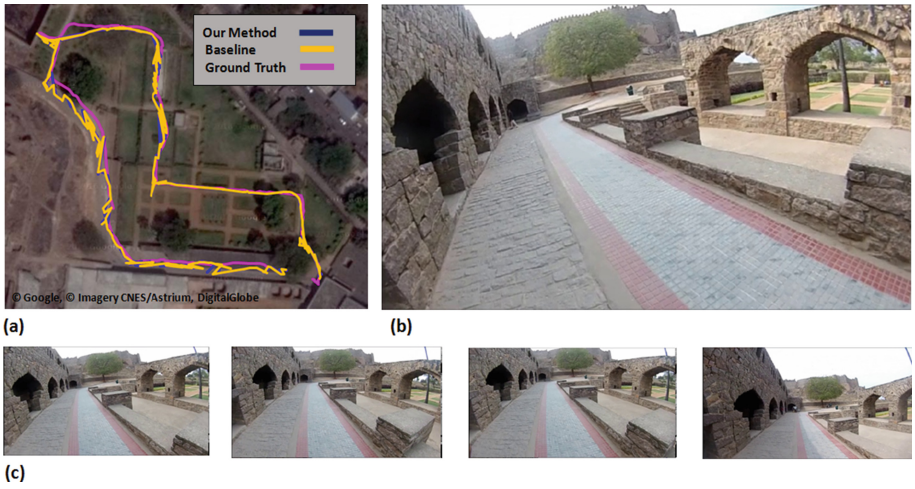

Fig. 4. Visual localization qualitative results: (best viewed in color) (a) The GPS estimate for a tour computed via the proposed method is shown in the 'blue' and the baseline in 'yellow' while the ground truth is overlayed in 'pink'. The estimates closely match the ground truth. (b) Frames from an egocentric viewpoint are taken as a query to find the top matches in the visual index. (c) From the query frame in sub-figure:b, the obtained top matches for the query are re-ranked by computing the most similar affine pair of the matches. The top matches are visually similar to the 'query frame'. (Color figure online)

3 Approach

Creating compelling narratives to highlight artistic and cultural significance of the heritage monument is formulated as an optimization problem. We propose a two-fold approach, we first identify participant's location using vision cues and later craft the most appropriate story linked to that location. The proposed solution primarily focuses on (1) use of vision cues over standalone GPS for accurate localization. (2) narrative generation for the localized structures as an optimization problem over a constructed graph. The overview of our approach is illustrated in the Fig. 3.

For a given heritage site, we first identify the major locations within a site ('interest points'). We conceptualize the heritage site as a graph where every interest point represents a vertex and an edge between any two interest points denotes a possible traversable trail between them. Every edge in a graph has following associations: (a) ground truth GPS tags, (b) relevant (hand curated) summaries linking both the interest points and (c) 'glue' sentences (which assist in connecting two selected summaries coherently).

Multiple egocentric tours are recorded along with loosely correlated GPS tags. The videos are partitioned as training and validation tours. The training set along with the embedded GPS tag are used to create frame level inverted index. For a test video of a tour, whose GPS tags are not known, the localization is achieved using a two-staged approach – (1) search similar frames using the inverted index (2) estimate GPS tags of a query frame by linearly interpolating location from the matched image pairs in the set of retrieved images. The computed GPS tags are associated to the (nearest) appropriate edge. The narrative generation is then formulated as an optimization problem over the constructed graph and other associations to obtain a sequence of summaries and 'glue' sentences. During training and inverted index creation, high precision GPS sensors are used. At test time we use vision based precise localization.

3.1 Localization

The images are represented by their corresponding bag-of-features representation, d. We use RootSIFT descriptors [22,23] and build a visual vocabulary using approximate nearest neighbor [24]. For fast image querying, we use inverted index data structure for our experiments. For every query point(q), the top-k visually similar images are retrieved, E_k. The retrieved image list is re-ranked using geometric verification of the estimated affine transformations using RANSAC [24]. To assure that image pairs share visual content and can be meaningfully used to estimate the GPS tag, the pairs are restricted by to *a unit graph distance* [3]. We instead employ the pairing of the top-k retrieved list of the images to achieve the same result.

The best matching pair (from E_k) is obtained by locating the most similar affine combination [3] in retrieved set of top-k similar images. The least affine similarity among the pairs is computed by minimizing the following:

$$s_1, s_2 = \underset{(i,j)\in E_k}{\operatorname{argmin}} \frac{(d_j - d_i)^T (q - d_i)}{\|d_j - d_i\|^2} \tag{1}$$

The location of q is expressed as an affine combination of the most similar affine pair s_1 and s_2. The relative similarity of q with d_{s_1} and d_{s_2} is computed as,

$$\beta = \frac{q^T d_{s_1}}{q^T d_{s_1} + q^T d_{s_2}} \tag{2}$$

We consider the following linear localization function to estimate the location (the location for a point p is represented as x_p):

$$x_q = x_{s_1} + (a_0 + a_1\beta)(x_{s_2} - x_{s_1}), \tag{3}$$

where, a_0 and a_1 are the regression parameters estimated from representative set, R, of randomly sampled GPS tags (details in Sect. 4.1). We thus obtain the estimated location from the linear combination of the most similar affines.

3.2 From Locations to Edges

A heritage area can be viewed as a culmination of prominent 'sites' and the 'pathways' joining these sites. In a conceptual framework, this can be represented by a graph $G = (V, E)$, with the sites being vertices V and the pathways being the edges E. Edges, e_i are represented as set of linear splines between the associated GPS points. Extended kalman filter [25,26] is used to smooth the input GPS signal, x. For each computed GPS coordinate, we assign the weights to every edge of the graph [27]. Each weight correspond to the shortest (euclidean) distance of the point from the edge. Starting from the first input coordinate to the last, each GPS tag is greedily assigned an edge label corresponding to the least weight, Algorithm 1. To suppress the errors in the labels assigned to GPS tags, which arise due to abrupt changes in the edge assignment, we smoothen the assigned labels and eliminate the infrequent ones. The smoothing is performed by considering sliding window of various sizes, say b. The effects of such varied window size selections are discussed in Sect. 4.3.

3.3 Forming Narratives

We associate text of various lengths to each edge of the graph G (Sect. 3.2), $S = [s_i^j \| \ e_i \in E]$ where j denotes the j^{th} summary (each summary is of a different length) and i represents the index of the edge. We also associate multiple glue sentences with an edge, $H = [h_{x,y}^k \| \ e_x, e_y \in E]$, where k is the k^{th} glue sentence and (x, y) represent indices of the two edges. These glue sentences are used to bind adjacent edge summaries together. We define a function L which computes the length of a summary. The optimal length of an edge, l_i^* depends on the participant's pace along the known length of an edge and is computed by

Algorithm 1. GPS to Edge classification: Each GPS tag is assigned an edge for narrative generation.

```
procedure LocationClassifier(x, E, b)
    x ← ExtendedKalmanFilter(x)                    ▷ x - GPS time signal
    y ← argmin_e EuclideanDistance(x, E)           ▷ E - edge set
    for i in 1..length(y) do                       ▷ y - classified points to edges
        if y[i] ≠ y[i + 1] then
            current, next ← y[i], y[i + 1]
            hit ← 0
            for j in i..i + b do                   ▷ b - window size
                if y[j] = next then
                    hit = hit + 1
                end if
            end for
            if hit ≠ b then
                y[i + 1] ← current
            end if
        end if
    end for
    return y
end procedure
```

counting number of GPS locations belonging to a particular edge. All values are scaled and centered to have the same units. The length (function, $L()$) of the summary is estimated by multiplying the number of words in a summary with average time taken to speak one word by the Text to Speech Engine.

Our problem is reduced to the selection of a sequence of summaries and glue sentences for a known path and speed such that it forms a narrative which is relevant, cohesive and informative. We maximize the following function,

$$R(l, l') = \sum_i \Big(\sum_j \alpha_i^j L(s_i^j) + \sum_k \beta_{i,i+1}^k L(h_{i,i+1}^k) \Big)$$

subject to the following constraints,

$$\sum_j \alpha_i^j = 1 \quad \forall\ i \tag{4}$$

$$\sum_k \beta_{i,i+1}^k = 1 \quad \forall\ i \tag{5}$$

$$\alpha_i^j L(s_i^j) + \beta_{i,i+1}^k L(h_{i,i+1}^k) \leq l_i^* \quad \forall \quad i, j, k \tag{6}$$

The objective function represents the selection of the most informative summaries and glue sentences. Equation 6 represents the constraints imposed to keep the summaries relevant to the localization. The glue sentence selection makes the narrative cohesive.

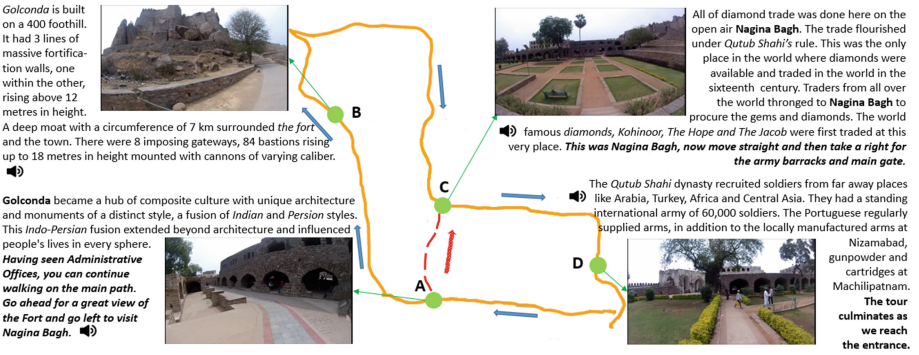

Fig. 5. Qualitative results: (best viewed in color) The narratives generated and spoken by a text-to-speech engine are not only site specific and encapsulate the details of the nearby sites. Although the participant does not walk via the dotted (red) path from site A, yet the narrative generated are comprehensive in nature and assimilate the details of site C and vice-versa. (Color figure online)

4 Experiments and Results

In this section, we present quantitative and qualitative results for each of the building blocks described in Sect. 3. We begin by describing the dataset that we formulate our model. Moving further we discuss the results of visual localization, edge classification and narrative generation modules.

4.1 Dataset

We demonstrate the application of the proposed framework on a dataset that is created by visiting a region inside a world heritage site. Built in distinctive style of Pathan, Hindu and Persian architecture the site is spread across an area of $20,000$ sq.m. We capture 2 h of 'tour' videos tagged with GPS values. We used a Contour Plus2 camera to capture multiple tour videos around 6 distinct sites in the same region. 8 tour videos are taped at 30 fps with GPS frequency of 1 Hz. 6 videos are used for training and 2 for validation and testing. The associated GPS tags act as a ground truth for localization. To bestow maximum lighting and occlusion variations in the video dataset, we record videos at varied times of day with varying crowd densities. While taking 'tour' videos we avoid following a standard and fixed pathways across the sites and take varied paths for every tour to capture the variabilities involved with participants path selection.

We sample GPS-annotated frames from the 'tour' videos at a *sampling rate* [3], $\triangle = 15$ and use them for training. In all we have 5952 images and corresponding GPS tags in our training set, M. To create validation set, T, we sample GPS annotated frames at $\triangle = 30$ to obtain 915 images. A representative set, R, of 124 GPS annotated images is randomly sampled from our training videos to estimate the regression parameters a_0 and a_1 (discussed in Sect. 3.1).

The narration corpus comprises of hand-crafted summaries of the site collected from various open descriptions available both on-line and off-line tour books. Multiple versions of different lengths are created for each site. The summaries describe historical importance of a site, accompanied by anecdotes and other details. Summaries of different length capture the essence of a site at various levels: the longest summary includes the maximum details while the shorter ones include only the most important details. In brief, our dataset is comprised of tour videos, associated GPS tags and associated narrations text.

4.2 Visual Localization

We evaluate the localization accuracy of the model on our validation set. We train a Bag of Visual Words model for a vocabulary size of $500K$ words using video training set. The localization performance is measured by calculating the percentage of query frames localized within d meters from their ground truth GPS location, loc_d [3]. The localization error is calculated using Harversine formula.

Baseline: We baseline the visual localization by comparing our results with the nearest neighbor (NN) [3] approach in Table 1

Table 1. Frame localization: The percentage of correctly localized frames sharply increases with the increase in the radius. The visual localization method (marginally) leads the baseline for different values of k. Here, k represents top-k retrieval

Measure	Our method			
	NN	$k = 2$	$k = 3$	$k = 5$
loc_5	48.1400	48.2495	49.1247	46.8271
loc_{10}	80.3063	81.9475	82.7133	83.1510
loc_{15}	90.4814	92.1225	93.3260	95.9519

Visual localization performs better than the nearest neighbour baseline (Table 1). The larger value of k improves the accuracy for loc_{15} as it considers more viewpoints which are sparsely represented but we lose out on the accuracy in loc_5 (where the location is densely represented). Thus we choose $k = 3$ as an optimal value our subsequent experiments. The qualitative results are depicted in Fig. 4 where we can see that an estimated tour closely matches the ground truth.

4.3 Edge Classification from Localization

We evaluate the edge classification accuracy on our validation set by manually labelling every GPS coordinate with an edge. The classification error is calculated as the percentage of coordinates which are misclassified by our algorithm while varying the window size.

Table 2. Edge classification: No smoothing or small window size indicates that aberrations cannot be removed without smoothing across an appropriate sized neighbourhood. Large window size indicates that a higher window size leads to loss of information instead of decreasing ambiguity.

Window size	0	5	10	15	20	25
Error percentage	7.6	5.4	4.3	4.3	4.3	10.41

The smoothing performed to eliminate aberrations in the GPS signal is tested over various window sizes and the best window size is selected. The results in Table 2 indicate that the window size of 15 yields the best results and we use this value for subsequent experiments. The error is due to the precision of the GPS signal and subsequent visual localization estimate which forms our input signal, thus coordinates close to two edges are classified ambiguously at times, resulting in abrupt aberrations. The suggested approach classifies each coordinate to an edge and then smoothens the resultant signal in the window.

4.4 Narrative Generation

The optimization problem is solved as an Binary Integer Programming Problem using PuLP python library [28]. The error for each edge, E_{e_i} is computed as the difference between the optimal time and spoken summary time.

$$E_{e_i} = l_i^* - (L(s_i^j) + L(h_{i,i+1}^k)) \tag{7}$$

where i denotes the index of the edge, j denotes the j^{th} summary selected for the edge from the set of all summaries associated with the edge and k denotes the k^{th} glue sentence selected from the set of all glue sentence associated for the two edges in order (e_i and e_{i+1}).

Table 3. Narrative generation: The overall tour percentage error of both baseline and the proposed approach are similar. The suggested approach being more robust performs marginally better than the baseline. Both the methods are comparable to the actual ground truth as can be seen for the tours in the validation set. The last two rows (Tour 3 and 4) correspond to videos in test set.

Tour	Our method	Nearest neighbour	Ground truth
Tour 1	7.85	7.14	5.50
Tour 2	5.39	6.19	5.37
Tour 3	6.12	6.32	–
Tour 4	10.20	17.63	–

The percentage error for a tour is calculated as the sum of the error components E_{e_i} divided by the sum of the optimal time for each edge. Say p is the walk followed, then percentage error is given by,

$$E_p = \frac{\sum_p E_{e_p}}{\sum_p l_p^*} \tag{8}$$

Baseline: We present our results for two video tours in the validation and test set. The best performing results, $k = 3$ (Sect. 4.2), of our method are compared with Nearest Neighbour (NN) baseline. The proposed method is marginally better than the NN baseline, Table 3. The errors in themselves are very marginal and it shows that our model is synchronous and relevant with respect to the entire tour.

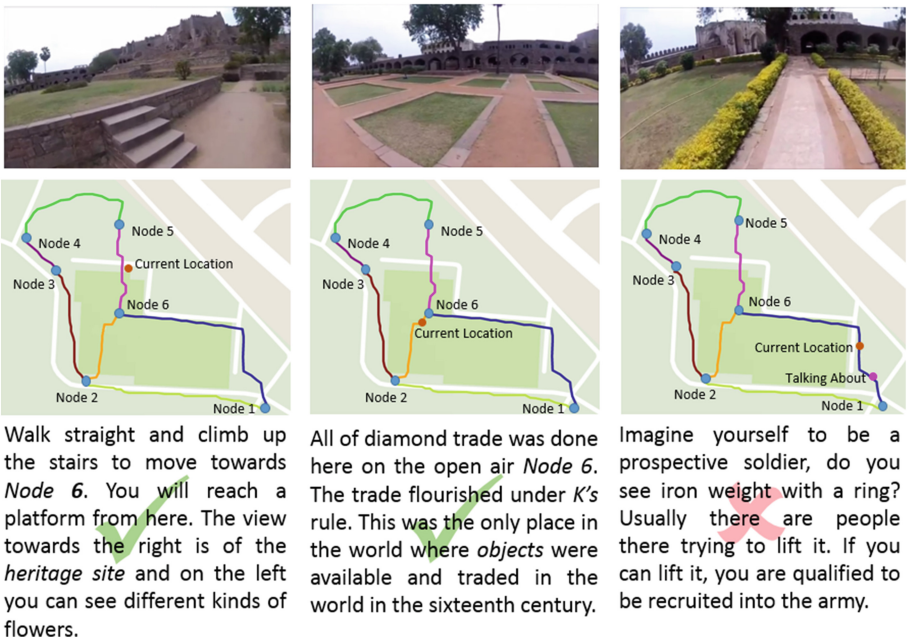

Fig. 6. Human evaluation test bench: The participants were shown narrations linked output videos and were asked to rate the experience. For every video snippet we show the 'current location' on the map and other nearby sites. The narrations generated should have information of the present location and other nearby sites. The last row exhibits the narration (spoken by a text-to-speech engine) generated by the system for the present location. The third instance shows a 'negative' case where the text generated does not imbibe the details of the present location and other nearby sites.

Fig. 7. Human evaluation: Qualitative assessment for 35 participants. The proposed approach garners an average score (shown by a straight line on the graph) of 3.97 for the relevance with respective to the surrounding, 3.4 for cohesion between narration sentences and 3.6 for participants overall experience.

4.5 Human Evaluation

Human perception can be highly non-linear and unstructured. Multiple facets of any human-computer interacting system need to be well thought out, profiled and tested to make them satisfactory to humans. In the context of the present scenario, we need to evaluate if the sequence of the text selected is relevant to the current location? Due to paucity of any formal evaluation scheme to measure the effectiveness of the proposed approach, we contrive an evaluation procedure where we asked the participants to watch a dynamic narrative video of the heritage site with the narratives being spoken by a text-to-speech engine. The participants were then asked to rate the relevance, cohesion and the overall experience, Fig. 6.

Around half of the participants were unaware of the heritage site and had never been there. We showed them the output of the proposed approach (videos, present location on map and the aligned narrations), and asked them if they thought that the narrations were relevant and cohesive on a five point scale – with 1 corresponding to strong disagreement and 5 corresponding to strong agreement. The same scale was used to ask if they were comfortable with the accent of the speaker and if the narration improved the overall experience of the tour.

A visualization of the human evaluation results can be seen in Fig. 7, with the vertical axis representing the five point scale and the horizontal axis representing the participants. Majority of the participants agree that the narrations were relevant while not as many believe that the narrations were cohesive in nature. The overall experience is even more varied. This can be attributed to the low scores (average score was 2.6) given for the accent of the narration as many participants found it slightly difficult to understand what was being said.

5 Conclusion

We suggest an approach to harness location based technology and text summarization methods to engage audiences in an unstructured environment. The proposed approach demystifies the details imbibed in the egocentric tour videos to generate a text based narration. We solve the adverse effects of inaccurate and

unreliable GPS signals using vision cues for robust localization. The dynamic narratives generated on fly are simultaneously optimised over content, length and relevance thus creating 'digital tour guide' like experience. Museums, heritage parks, public places and other similar sites can reap the benefits of such approach for both entertainment and educational purposes. Generating seamless denser stories and descriptions for diverse audiences over larger regions are some of the identified areas of future work.

References

1. Ikeuchi, K., Oishi, T., Takamatsu, J., Sagawa, R., Nakazawa, A., Kurazume, R., Nishino, K., Kamakura, M., Okamoto, Y.: The great buddha project: digitally archiving, restoring, and analyzing cultural heritage objects. In: IJCV (2007)
2. Adabala, N., Datha, N., Joy, J., Kulkarni, C., Manchepalli, A., Sankar, A., Walton, R.: An interactive multimedia framework for digital heritage narratives. In: ACMMM (2010)
3. Torii, A., Sivic, J., Pajdla, T.: Visual localization by linear combination of image descriptors. In: ICCV Workshop (2011)
4. Van Aart, C., Wielinga, B., Van Hage, W.R.: Mobile cultural heritage guide: location-aware semantic search. In: Knowledge Engineering and Management by the Masses (2010)
5. Panda, J., Brown, M.S., Jawahar, C.V.: Offline mobile instance retrieval with a small memory footprint. In: ICCV (2013)
6. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: CVPR (2013)
7. Chen, X., Yuille, A.L.: A time-efficient cascade for real-time object detection: with applications for the visually impaired. In: CVPR (2005)
8. Ezaki, N., Bulacu, M., Schomaker, L.: Text detection from natural scene images: towards a system for visually impaired persons. In: ICPR (2004)
9. Schwarze, T., Lauer, M., Schwaab, M., Romanovas, M., Bohm, S., Jurgensohn, T.: An intuitive mobility aid for visually impaired people based on stereo vision. In: ICCV Workshops (2015)
10. Rodríguez, A., Yebes, J.J., Alcantarilla, P.F., Bergasa, L.M., Almazán, J., Cela, A.: Assisting the visually impaired: obstacle detection and warning system by acoustic feedback. In: Sensors (2012)
11. Pradeep, V., Medioni, G., Weiland, J.: Robot vision for the visually impaired. In: CVPR (2010)
12. Lin, T.Y., Belongie, S., Hays, J.: Cross-view image geolocalization. In: CVPR (2013)
13. Martin, E., Vinyals, O., Friedland, G., Bajcsy, R.: Precise indoor localization using smart phones. In: ACMMM (2010)
14. Bay, H., Fasel, B., Gool, L.V.: Interactive museum guide. In: UBICOMP Workshop (2005)
15. Schroth, G., Huitl, R., Chen, D., Abu-Alqumsan, M., Al-Nuaimi, A., Steinbach, E.: Mobile visual location recognition. Signal Process. Mag. **28**(4), 77–89 (2011)
16. Kim, G., Xing, E.P.: Reconstructing storyline graphs for image recommendation from web community photos. In: CVPR (2014)
17. Wang, D., Li, T., Ogihara, M.: Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs. In: AAAI (2012)

18. Riedl, M.O., Young, R.M.: From linear story generation to branching story graphs. IEEE Comput. Graph. Appl. **26**(3), 23–31 (2006)
19. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: ICCV (2015)
20. Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., Schiele, B.: Translating video content to natural language descriptions. In: ICCV (2013)
21. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: ICCV (2015)
22. Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: CVPR (2012)
23. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. In: IJCV (2004)
24. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR (2007)
25. Hu, C., Chen, W., Chen, Y., Liu, D.: Adaptive Kalman filtering for vehicle navigation. Positioning **1**(04) (2009)
26. Tolman, B.W.: GPS precise absolute positioning via Kalman filtering. Ionosphere **2**(L1), L2 (2008)
27. Marchal, F., Hackney, J., Axhausen, K.: Efficient map-matching of large GPS data sets-tests on a speed monitoring experiment in Zurich. Arbeitsbericht Verkehrs-und Raumplanung (2004)
28. Mitchell, S., OSullivan, M., Dunning, I.: PuLP: a linear programming toolkit for python. The University of Auckland, New Zealand (2011)