An Interactive Tour Guide for a Heritage Site

Sahil Chelaramani, Vamsidhar Muthireddy, C.V. Jawahar
CVIT, IIIT Hyderabad
{sahil.chelaramani,vamsidhar.muthireddy}@research.iiit.ac.in, jawahar@iiit.ac.in

Abstract

Imagine taking a guided tour of a heritage site. Generally, tour guides have canned routes and stories about the monuments. As humans, we can inform the guide about topics which we are interested in, so as to ensure that we are presented stories which match our interests. Most digital storytelling approaches fail to take into account this aspect of human interaction when presenting stories to a user.
In this work, we take on the task of interactive story generation, for a casually captured video-clip of a heritage site tour. We leverage user interaction to improve the relevance of the stories presented to the user. The stories generated vary from user to user, with the stories progressively becoming more aligned with the captured interests, as the number of interactions increase. We condition the stories on visual features from the video, along with the interests of the user. We additionally present a mechanism to generate questions to be posed to a user to gain additional insights into their interests. The efficacy of this work is demonstrated at the Golconda fort, situated in Hyderabad, India.

1. Introduction

Seamless interaction represents the quintessential quality of being a human. There has been a long standing quest for a mechanism which allows machines to learn seamlessly from interactions with a user. In 1950, Alan Turing came up with the famous Turing test for an artificially intelligent system, requiring that a human being should be unable to distinguish the machine from another human being by using only the interactions with both. This requires machines to mimic the human capacity to learn on the fly and dynamically adjust its responses.

Machines learn differently than humans do. They require massive curated datasets which capture the minutiae of the aspects they are trying to learn. In the context of computer vision, the collection of a dataset requires a significant manual effort to map the relevant images to their corresponding labels. Every label needs to be tagged with multiple images from varied view-points and perspectives. Extensions of such kind to places of interest such as heritage sites, would thus require manual mapping of the whole site. We even need to have huge curated text corpus with site-specific information which has been sampled from a variety of topics. This makes it extremely taxing to attempt a problem of this kind at such a scale.

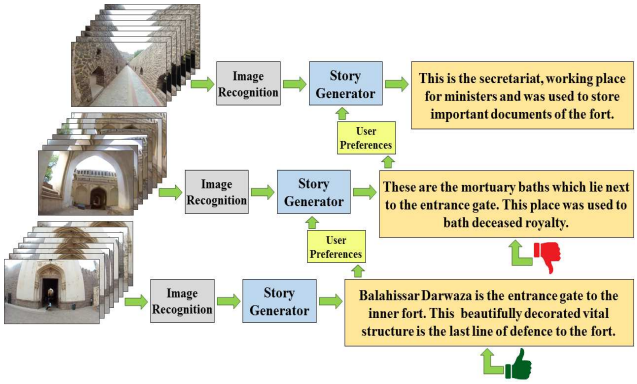

Figure 1. An overview of our approach. The user gives his feedback for stories related to the current segment of the video and we dynamically adjust the stories taking this interaction into account.

The suggested scheme should be precise enough to generate valid semantic associations which represent a user's interests. In this paper, we take on the task of interactive story generation for a casually captured video-clip of a heritage site tour. In the context of this paper, we define a story as a continuous piece of text which describes the content of the video in detail and is a cohesive entity. We leverage human interaction to improve the relevance of the stories presented to the user. The stories generated vary from user to user, with the stories becoming more and more aligned with the captured interests as the number of interactions increase.

We condition the stories on visual features from the video, along with the interests of the user. In the context of heritage sites, our features must be resistant to illumination variations, weather changes, pollution and other environmental degradations. Figure 1 illustrates our approach. We first identify monuments from the heritage site, which occur in the video. These recognitions define the structure

of our stories. We present an approach which optimizes content, coherence and relevance of the generated stories with the user's interests by formulating this problem as a binary integer program. We additionally present a mechanism to dynamically generate questions to be posed to a user to gain additional insights into their interests. We formulate questions like "Would you prefer to listen to stories related to architecture or culture?" to ascertain the interests of the user. These questions serve a dual purpose, first, it handles cases in which we are dealing with conservative users who do not explicitly interact with our system. Second, it helps garner an estimate for the user's interests quicker. We gradually decay the probability of asking questions to the user by the number of interactions.

The primary contributions of this paper are:-

- We dynamically adjust the stories presented to a user, for a video tour of a heritage site. These stories are adjusted so that the stories best represent the user's interests.
- We present a mechanism of generating a question, to gauge the topics that interests the user.
- We demonstrate the efficacy of our work at the Golconda fort, situated in Hyderabad, India. This fort has a total of 32 different monuments within its periphery of 7 km.

In summary, we address the problem of interactive storytelling for a video clip of a heritage site. In the following sections, we explain our approach in detail.

2. Related Works

Our work shares the high level goal of describing a video with natural language. The bridge between Computer Vision and Natural Language Processing(NLP) began with the task of associating words with images. Multiple papers [3, 36] studied the multimodal correspondence between words and images. Other tracks of research led to the task of describing images with sentences. A number of approaches pose the task as a retrieval problem, where the most compatible annotation in the training set is transferred to a test image [37, 9, 28, 41] or where training annotations are broken up and stitched together [21, 22].

There have been papers that have adapted these retrieval models to provide descriptions of images of a heritage site [30], which we adopt as a baseline for comparison. Many papers [2, 33, 46] propose using Convolutional Neural Networks (CNNs) in image retrieval due to their effectiveness in capturing image semantics. Although methods like [23, 11] show state of the art results in retrieval, they are slow during indexing and feature extraction. Other works like [26] use local features in the convolutional layer along with BoW encoding and take advantage of sparse representations for fast retrieval in large datasets. In [18], a generalized framework is presented for image search using cross-dimensional pooling on global features in a pre-trained network. We adopt this method as one of our retrieval methods, since it outperforms other frameworks.

Recently, several approaches also generate captions of images using neural networks [16, 19, 44, 48]. Videos are just a continuous sequence of images and hence there has been an increased interest in solving the problem of describing videos with text [29, 40]. These methods map videos to full sentences and can handle variable-length input videos. However, they do not scale to long tour videos that need to be described using multi-line text outputs.

Active learning methods have been gaining increased interest in the machine learning community [8, 4]. Active learning has found applications in a variety of subdisciplines of NLP, such as information extraction, text classification and natural language parsing [38, 24], to name a few. Chu and Ghahramani [7] present a Bayesian framework for preference learning, which takes advantages of Bayesian methods for model selection. Houlsby et al. [14] proposed a method for learning pairwise preferences expressed by multiple users. We formulate our interactive learning problem for storytelling, by understanding the numerous works dealing with active learning. Active learning, although initially developed within the classification framework, has been extended to handle a variety of multimedia applications [12, 15]. In the field of Computer Vision, active learning is adopted in image/video annotation [49, 47], image/video retrieval [42, 15] and image/video recognition [43, 17]. Even with this popularity, virtually none of the prior papers have applied active learning in the context of story-telling.

Stories have been a part of human culture and it is through storytelling that generations have been provided with a worldy compass. We are inspired by the works of Zhu et al. [50] which align movies and books to provide descriptive story-like explanations for visual content occuring in a movie. Majority of the work in the domain of technological storytelling is focused on creation of visual storyline graphs of images summarizing series of events that have chronological relationship [20, 45]. Existing systems such as [10] are built to dynamically adjust the stories being presented to a user depending solely on the participant's chosen tour path and approximated walking speed. Similar to [32], we focus on creating an interactive narrative system in which details of the heritage sites are unfolded steadily. While participants stroll around the site, they can inform the tour guide about the stories that they like, and the guide adapts the stories according to the participants interests.

Stories are powerful mechanisms which can instill a variety of emotions in the listener. A majority of these re-

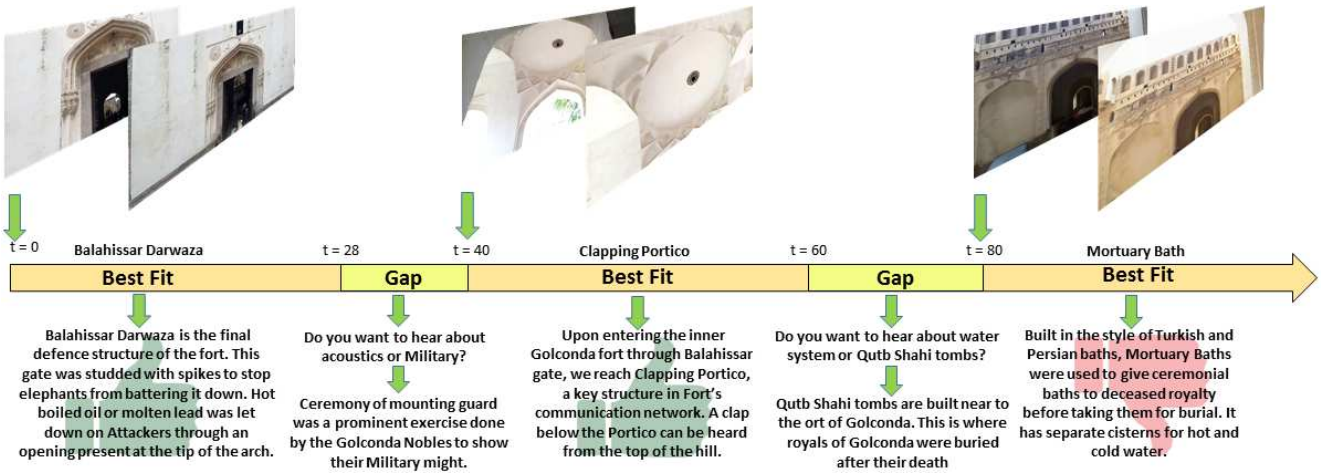

Figure 2. The working of our interactive story teller. Left to right indicates the timeline of the video. We interact with the users in the form of likes, dislikes and questions to gauge their interests. Gaps represent segments of the video where there is no relevant visual content to speak.

sponses are largely dependent on the listener. Most works described earlier, fail to capture this essential trait of a storyteller, which is to know their target audience. We propose a method which leverages interaction with a user to actively adjust stories related to a video tour at a heritage site, so that a user is presented with increasingly engaging content.

3. Method

Our approach for interactive storytelling begins with a video tour of a heritage site. Figure 2 depicts the pipeline for our proposed approach. The video is preprocessed by down-sampling and resizing each of the frames. Preprocessing helps improve the processing time for the videos. We proceed to identify the monuments occurring in the video by extracting features and using image retrieval to label each frame. We have a dataset of images of a variety of monuments from the Golconda fort, taken from different vintage points, along with their corresponding monument labels. We match the video frames to this set of images and transfer the corresponding labels to the frames. We encounter a class imbalance problem since most frames of the tour video correspond to the path rather than monuments. This results in the labels of the frames being noisy. We leverage the temporality of the video to smooth these noisy labels. Next, we generate stories about the monuments detected. Users can indicate whether they like or dislike the stories being narrated. Additionally, to help garner an estimate for the user's interests expeditiously, we present a mechanism where questions are formulated dynamically. This mechanism also addresses the case where we are dealing with a conservative user. We decay the probability of asking questions directly by the number of interactions with a user. The distribution of topics that a user may be interested in, are inferred from both forms of interactions and used to dynamically adjust the stories presented. This method is described in greater detail in the upcoming sections.

3.1. Monument Recognition

We begin with a video tour of a heritage site. A user records the video of the heritage site casually as he explores the site. This video cannot be processed as it is, since these videos tend to shake and can be blurry. We detect blurred frames by convolving the grayscale version of each frame with the laplacian operator and then compute the variance of the response [31]. We down-sample the number of frames and then resize the remaining, so as to obtain a reasonable processing time for a video.

We leverage existing image retrieval approaches to label each frame of the video. Our approach for labeling frames works by retrieving the closest matching image from our training set using an inverted index which stores the extracted image features. We transfer the label of the top retrieval to our query frame. We use the well known tf-idf score to compute the relevance between query frame and images from our training set [35]. We use a BoW model as a baseline, and compare it to a variety of other retrieval schemes [30, 10, 18]. We finally experiment with an ensemble of these schemes by combining the results from each method using Borda count for late fusion of retrieved results [1].

3.2. Label Smoothing

We take advantage of the temporality of videos and apply two different smoothing techniques to the labels obtained in the previous section. We note that the monuments which have been recognized should occur continuously in a video

segment, rather than interspersedly. Therefore, we apply a smoothing operation on the predicted labels in order to approximate this intuition. The first smoothing algorithm we employed is called fixed smoothing, in which we segment the n frames into continuous intervals of size w. This gives us a set of intervals of frames, denoted I_w.

$$I_w = \{[x_{1-w/2}, x_{1+w/2}], ..., [x_{n-w/2}, x_{n+w/2}]\} \quad (1)$$

For every interval I in I_w, we change the label of the middle frame in the interval to the mode label for that interval. We used 10-fold cross validation to discover effective values of w. The fixed smoothing technique is the simplest type of smoothing, and provides a useful comparison for the effectiveness of a smoothing operation. We have also employed a dynamic smoothing scheme, which allows for windows of varying width. We wish to find an ideal window size for an interval for a frame. For every possible window size w between a minimum and maximum window size W_{min} and W_{max}, we compute the mode label in the window and denote this mode by m_w. We then select the value of the window size w, which maximizes the number of labels that are equal to the mode, m_w, in an interval. This is described in equation 2.

$$w_I^{ideal} = \underset{w}{\text{argmax}} \frac{\|I(Y = m_w)\|}{w^2} \quad \forall w \in \{W_{min}, W_{max}\} \quad (2)$$

Here w_I^{ideal} denotes the ideal window size for a given interval I. Y is the label of the frames in the current interval. We select a value of w which maximizes the number of elements in the interval, which are equal to the mode. We have noticed that using larger window sizes has a negative influence on smoothed labels due to the presence of noisy frames. Hence, in the above optimization, we penalize the selection of large window sizes, by dividing the objective by a factor of w^2.

3.3. Selection of Stories

We associate a set of M stories of different lengths $S = \{s_i^m\}$ to each of the N monuments, indexed by i at the heritage site. We solve an optimization problem to obtain stories about the monuments for our video [10]. We also have generic stories, stories related to kings and facts about the heritage site, which are used to stitch stories about the monuments together. Since these stories have been sampled from a large variety of topics, finding the optimal set of stories is NP-hard [5]. To address this, we leverage user interactions. This gives us a two-fold advantage, first, it eases finding an appropriate story structure, second, the users are presented only with content relevant to them. Our problem is reduced to the selection of a sequence of generic stories with stories about kings and facts to fit in between the stories of each of the monuments, such that it forms a cohesive narrative which is relevant to the user. We formulate a binary integer program for this purpose. The formulation optimizes a score α_i^m for each of the stories, which is defined in equation 3.

$$\alpha_i^m = \sum_{a \in A} S(s_a^m, s_i^m) + \sum_{u \in U} S(s_u^m, s_i^m) + \sum_{d \in D} (1 - S(s_d^m, s_i^m)) \quad (3)$$

The first term of α_i is a coherence term between the set of stories of the adjacent monuments A detected in the video and the current story s_i^m. The second term is a measure of similarity to the set of stories U, that a user liked. Similarly, the third term represents the dissimilarity from the set of disliked stories D. All values are scaled and centered to have the same units. To compare the stories for similarity, we train a Latent Dirichlet Allocation(LDA) model [6]. The function S represents the similarity between the stories. This is computed as a function of the hellinger distance between the topical representations of the stories represented by the variable s_i^m. This function is given in equation 4.

$$S(p, q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{k=1}^{T} (\sqrt{p_k} - \sqrt{q_k})^2} \quad (4)$$

Here, T represents the number of topics we use to represent our stories. We solve the following binary integer program to select the appropriate stories.

$$\begin{aligned} &\max_X \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_i^j \; L(s_i^j) \; X_i^j \\ &\text{subject to} \\ &\sum_{j=1}^{M} X_i^j = 1 \qquad \forall i \\ &\sum_{i=1}^{N} X_i^j = 1 \qquad \forall j \\ &L(s_i^j) \; X_i^j <= l_i^* \qquad \forall i \\ &X \in \{0, 1\}. \end{aligned} \quad (5)$$

We define a function L which computes the time to narrate a story. It 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. The variable X_i^j is a binary variable which indicates the selection of the appropriate story. The first two constraints indicate that each story

should be selected only once. The optimal story time l_i^* depends on the number of frames which have been labeled as a monument. We estimate this time by normalizing the number of frames by the frame-rate of the video. To summarize, in the above linear program we attempt to maximize the coherence between the stories of the monuments and the similarity between user's preferred stories.

3.4. Question Generation

To ensure that the approximated distribution of topics that a user is interested in, converges to the true distribution faster, we present an additional mechanism of interaction to help gauge the interests of a user. This mechanism directly poses a question to users. It also addresses the problem of conservative users which are users who do not explicitly indicate their preferences for stories. To build questions, we first process the captured video. Then we build topic models from the stories about the monuments occurring in the video and compute the semantic similarity on pairwise topic coherence [27]. We select the two stories which have the most topical entropy [13]. The topical entropy is given in equation 6.

$$H(z|s) = -\sum_{k=1}^{T} P(z_k|s) \ \ logP(z_k|s) \tag{6}$$

The function H represents the conditional topical entropy, while P represents the conditional probability of a topic z given the story s. From these selected stories, we formulate a question of the form "Would you like to listen to a story related to X or Y?". We leverage the works of [25] to automatically generate topic labels from the selected stories that can possibly fill the values of X and Y. The story which corresponds to the user's answer is added to the list of liked stories, and the earlier pipeline continues.

4. Experiments and Results

The proposed approach identifies details imbibed in tour videos related to monuments and generates a text based story along with corresponding questions to pose to a user. The dynamic stories are simultaneously optimized over content, length, and relevance thus creating a 'digital tour guide' like experiences. This tour guide like experience is further enhanced by presenting a user with a mechanism to interact with the system to generate even more interesting stories. A variety of experiments have been documented below.

4.1. Dataset

We describe the dataset used for the task of interactive storytelling below. In brief, our dataset comprises of two parts, tour videos for testing and training images of monuments along with their associated story text.

4.1.1 Visual Data

We used a dataset of 4400 images showing various monuments at the Golconda fort along with their corresponding monument labels. These images are used to train our retrieval algorithm. We compare frames from a query video to the images from the training set. For this purpose, we build a validation and testing set, which is a collection of about 20 tour videos of this site which were captured casually by 7 different users exploring this heritage site. We have a total of 285 minutes of video content with an average length of 14 minutes. We divide this set of videos using a 50% - 50% split for validation-test purposes and report the accuracy on the test sets.

4.1.2 Story Data

We also collected textual stories describing the variety of monuments, and a multitude of other topics such as architecture, relevant information regarding the kings who ruled the Qutb Shahi dynasty (The dynasty that built the Golconda fort), and various tidbits of additional information such as facts and adages pertaining to the monuments. This consists of approximately 5000 words which represent this content. These stories are hand-crafted summaries of the information collected from various open descriptions from available on-line sources and tour books. Multiple versions of different lengths are created for each story. The stories describe historical importance of a site, accompanied by anecdotes and other details. Stories 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.

4.2. Scene Identification

We study the effectiveness of various retrieval algorithms and present a comparison with our baseline which uses bag of visual words with a vocabulary size of 10k words, for retrieval [35]. In Table 1, we present the accuracies of our algorithm.

Method	mAP	$P@1$	$P@3$
SIFT-BoW [30]	0.46	0.59	0.53
RootSIFT-BoW [10]	0.62	0.77	0.70
CNN-Pooling [18]	0.57	0.73	0.66
Combined-Max-Vote	0.53	0.85	0.74
Combined-Unanimous	**0.75**	**0.96**	**0.93**

Table 1. mAP metric presented for various retrieval methods. We present P@N values, which denotes the percentage of the top N retrievals, which are correct. We specifically compute this metric for values of N equal to 1, 3.

These results are reported on the subset of frames which have retrieval scores above a certain threshold. For the

CNN-pooling algorithm, we use a CNN with the VGG-16 architecture, pre-trained on ImageNet as a feature extractor. We then proceed to use PCA on the extracted features to reduce the representation to 128 dimensions. We also explore the use of an ensemble of these retrieval algorithms, where we combine the retrieved lists from each method, using the Borda count [1]. These top retrievals are treated as votes in favor of a particular monument. Next, we consider two cases, first, all methods need to unanimously agree on a particular label for a frame, and second, a majority of them need to agree.

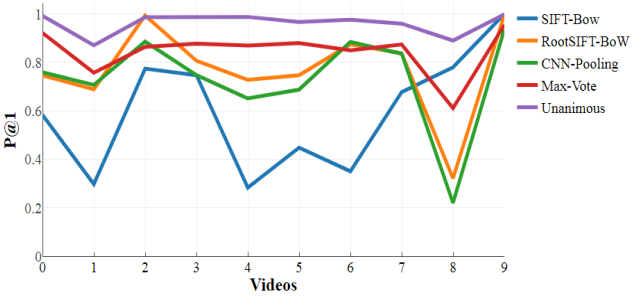

Figure 3. The graph depicts the variation of the $P@1$ for different methods across videos.

We present a comparison between the $P@1$ accuracies of the different methods in Figure 3. Our combined retrieval scheme outperforms all the other presented schemes. The reason why the combined scheme works better than the individual methods is evident from Figure 3. Considering only the graphs for SIFT and CNN pipelines, we can observe that they compliment each other well. SIFT features from our baseline model, are local features, while the CNN captures the global semantics of our images.

4.3. Label Smoothing

We leverage the temporality of videos to eliminate noisy retrievals. We experiment with two different smoothing techniques for our retrievals. The first smoothing algorithm employed is fixed smoothing. Using 10-fold cross validation we identify that a window size w of 30 gives us the best results on our validation set. This technique is intended as a baseline of comparison to our dynamic smoothing technique. We plot the precision and recall for the monuments identified as we vary the window sizes w of our smoothing algorithm in Figure 4. We note that for larger window sizes, even though the precision decreases slowly, the recall drops at a much faster rate.

Next, we consider our dynamic smoothing algorithm. We vary the size of window w between 5 and 100, and select the window size that minimizes the variance in that interval. We present the results comparing the $P@1$ accuracies obtained using fixed smoothing with the dynamic smoothing algorithm in Table 2.

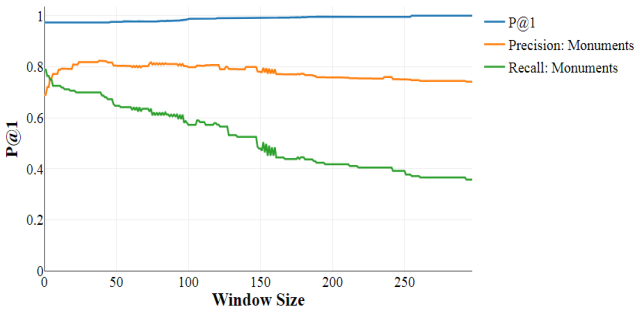

Figure 4. Variation of $P@1$ for the video frames, Precision and recall for monuments across different window sizes.

Method	$P_f@1$	$P_d@1$
SIFT-BoW [30]	0.75	0.80
RootSIFT-BoW [10]	0.91	0.95
CNN-Pooling [18]	0.79	0.82
Combined-Max-Vote	0.95	0.96
Combined-Unanimous	**0.97**	**0.98**

Table 2. $P_f@1$ denotes the precision obtained for fixed smoothing and $P_d@1$ denotes the precision for dynamic smoothing.

We note that the results obtained by the dynamic smoothing method are better than the results obtained using fixed smoothing.

4.4. Topic Modeling

To compare the stories for similarity, we train an LDA model [6]. For this, we first need to select the optimal number of topics. We use gap statistics [39] to get a reasonable estimate of the number of topics needed to represent our stories. We depict the sum of distances of samples from their closest cluster center, to the number of clusters in Figure 5.

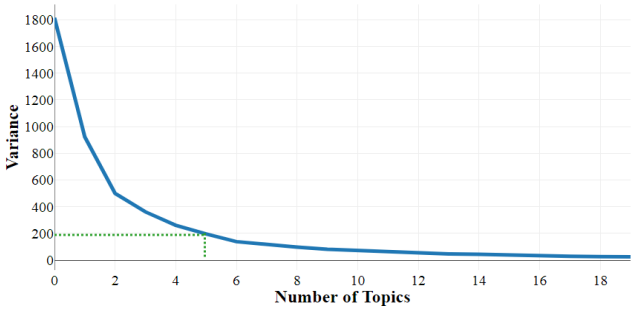

Figure 5. Graph to select the ideal number of topics for our story size.

Figure 5 shows that for our dataset, five topics would represent the data well. To further verify this, we present the visualization of the topic distribution learned on our dataset [34]. Figure 6 depicts the topics represented using multidimensional scaling, and the top salient terms corresponding to these topics.

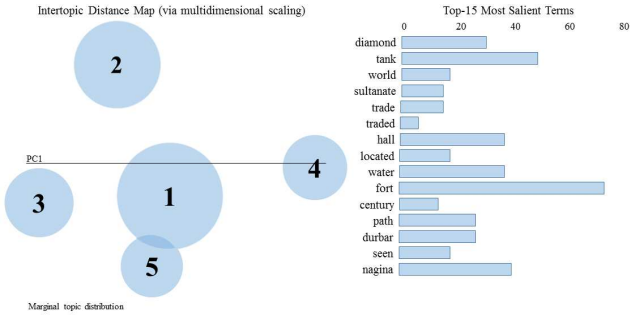

Figure 6. Visualization of the topic distributions learned by the LDA model with 5 topics.

As can easily be seen from the figure, the topics have very little overlap with each other and hence can be taken to be representative of a range of diverse topics. We proceed to model each of our stories in this LDA space trained with 5 topics, on the full set of stories, which consists of about 5000 words. We then project each of the stories into this space.

Additionally, we present sample topic labels generated for each of the topics using [25], in Table 3. We formulate questions with these topic labels. The topic labels are derived from stories which have the maximum entropy, as defined in Equation 6.

Topic	Topical words	Topic Labels
1	Fort Wall Gate	Defense Structure
2	Royal Sultan Darbar	England Crown
3	Kings Shahi Earthen	Vijayanagar Kingdom
4	Mughal Masjid Hindu	Ventilation System
5	Water Mosque Mahal	Engineering Prowess

Table 3. Sample topical words and their corresponding topic label.

4.5. Narratives Generation

Next, we present a quantitative mechanism for evaluation of the stories generated at the end of our pipeline. We collect a variety of user preferences from 40 users. These users are shown videos from our test set. The same video is shown to not more than five users. While we narrate the stories to them using a text to speech engine, they are allowed to interact with our system. We collect these interactions in the form of answers to questions, and likes and dislikes for different stories being narrated. Next, we compute the score for the average similarity between the stories corresponding to these user preferences and the stories generated. The similarity score is given by:-

$$\beta_i = \sum_{u \in U} S(s_u, s_i) + \sum_{d \in D} (1 - S(s_d, s_i)) \tag{7}$$

In Equation 7, β_i corresponds to the evaluation score of the i-th story. It represents the similarity of the selected stories to the user's preferences. This score has been normalized to a value between 0 and 1. A value of 1 indicates a perfect agreement between the selected stories and the user's preference, while 0 indicates a complete disagreement. The set U denotes the set of liked stories, and D denotes the disliked ones. We compare this value of β_i computed using our approach, against a greedy baseline, which fits the best stories based on a coherence score. The average evaluation score between the stories generated by the greedy algorithm and our method are plotted in the figure 7.

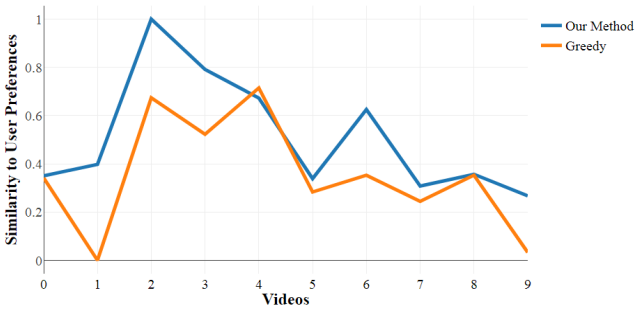

Figure 7. Comparison between the performance of a greedy baseline and our approach across videos. We demonstrate that for a variety of user preferences, the proposed approach outperforms the baseline.

The scores produced are a function of the user interactions and vary from video to video. As can be seen from Figure 7, even with the variations in similarity scores across videos, our method produces stories that are consistently superior to the baseline method. This shows that the stories generated indeed capture user interests.

4.6. Human Evaluation

Human perception can be highly non-linear and unstructured. Multiple facets of any human-computer interacting system needs 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 the relevance of the sequence of the text with current scene. Due to scarcity of any formal evaluation scheme to measure the effectiveness of our proposed approach, we contrive an evaluation procedure where we asked a group of 40 participants to evaluate the stories generated. Around half of the participants were unaware of the heritage site and had never been there. Figure 8 shows the results generated by our approach. The interactions have also been depicted in the same timeline. The participants were made to watch a video tour of the heritage site with the stories being spoken by a text-to-speech engine. They were asked to interact with the system, by liking and disliking the stories being presented to them. The system occasionally asked them questions about their interests.

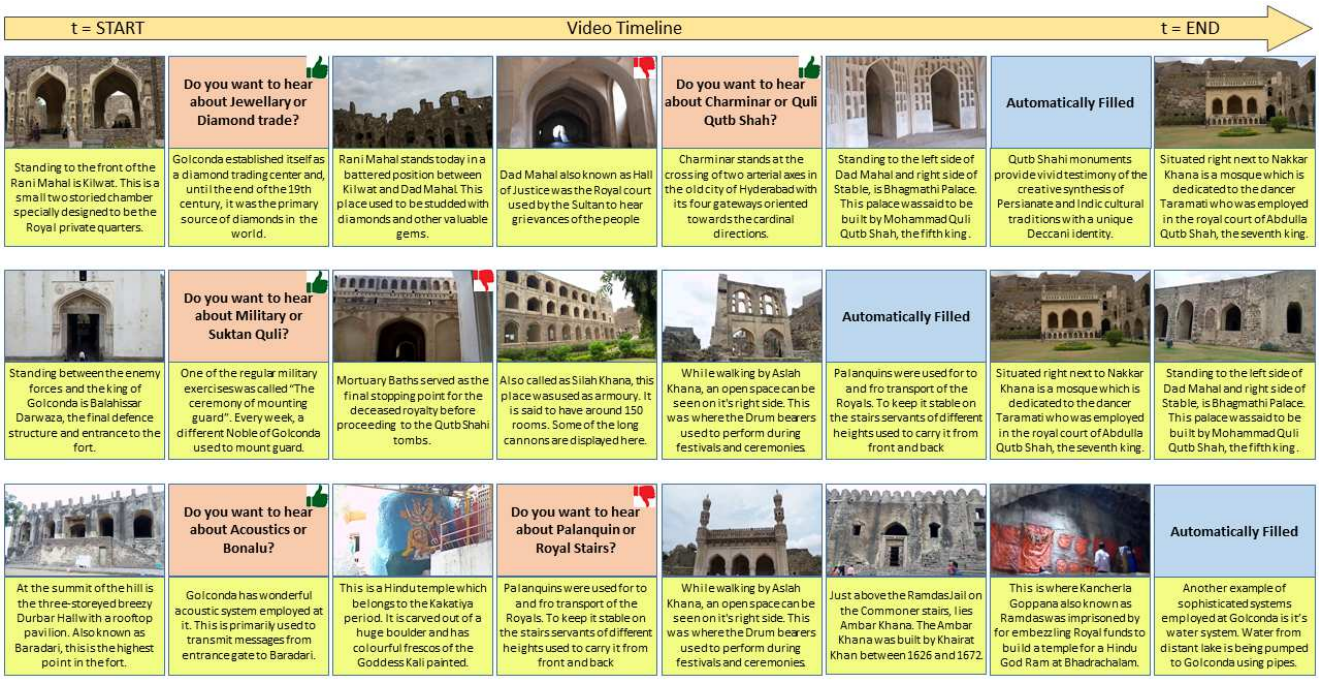

Figure 8. We illustrate stories generated by our approach for three different videos. When read left to right, we depict the video's timeline. A like or dislike on a particular time step indicates a user's interest in the story. Our approach prompts a user to answer a question, and progressively learns his interests.

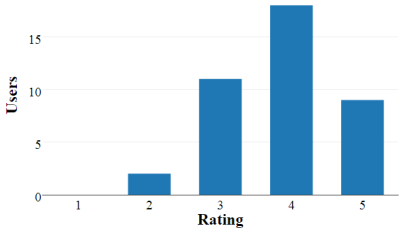

Figure 9. Coherence: We obtain an average rating of 3.85 out of 5.

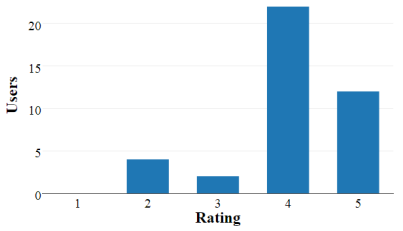

Figure 10. Relevance: We obtain an average rating of 4.05 out of 5.

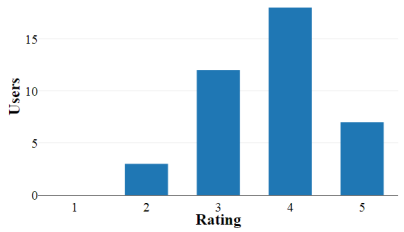

Figure 11. Alignment: We obtain an average rating of 3.73 out of 5.

The participants were then asked to rate the stories for the relevance to the video, cohesion of the overall structure and whether the stories align with their interests. They listened to the stories generated by the proposed approach, and were asked to rate the narrations based on the aforementioned metrics, 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 their knowledge about the site improved after listening to the stories and if the narration improved the overall experience of the tour. We plot the results for human evaluation in Figures 9, 10 and 11.

5. Conclusion and Future work

We suggest an approach to harness vision based technology, optimization theory and topic modeling methods to engage audiences in an unstructured environment. The proposed approach leverages user interaction to steer stories in the direction that a user chooses. The dynamic narratives generated on the fly are simultaneously optimized over content, length and relevance thus creating 'interactive tour guide' like experience, in which a user can "point" at different monuments and answer questions to improve stories. Museums, heritage parks, public places and other similar sites can reap the benefits of such an approach for both entertainment and educational purposes. This setup is ideal to enhance even virtual reality tours which have recently become common in many heritage sites. In future, we aim to apply interactive framework to a larger variety of problems and explore better modes of quantifying a user's interests.

References

[1] J. A. Aslam and M. Montague. Models for metasearch. In *SIGIR*, 2001.

[2] A. Babenko and V. Lempitsky. Aggregating local deep features for image retrieval. In *ICCV*, 2015.

[3] K. Barnard, P. Duygulu, D. Forsyth, N. d. Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *JMLR*, 2003.

[4] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *ICML*, 2009.

[5] M. Binshtok, R. I. Brafman, S. E. Shimony, A. Martin, and C. Boutilier. Computing optimal subsets. 2007.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 2003.

[7] W. Chu and Z. Ghahramani. Preference learning with gaussian processes. In *ICML*, 2005.

[8] B. Eric, N. D. Freitas, and A. Ghosh. Active preference learning with discrete choice data. In *NIPS*, 2008.

[9] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, 2010.

[10] A. Ghosh, Y. Patel, M. Sukhwani, and C.V. Jawahar. Dynamic narratives for heritage tour. In *ECCV-W*, 2016.

[11] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV*, 2014.

[12] P. H. Gosselin and M. Cord. Active learning methods for interactive image retrieval. *Transactions on Image Processing*, 2008.

[13] D. Hall, D. Jurafsky, and C. D. Manning. Studying the history of ideas using topic models. In *EMNLP*, 2008.

[14] N. Houlsby, F. Huszar, Z. Ghahramani, and J. M. Hernández-Lobato. Collaborative gaussian processes for preference learning. In *NIPS*, 2012.

[15] T. S. Huang, C. K. Dagli, S. Rajaram, E. Y. Chang, M. I. Mandel, G. E. Poliner, and D. P. Ellis. Active learning for interactive multimedia retrieval. *Proceedings of the IEEE*, 2008.

[16] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016.

[17] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, 2009.

[18] Y. Kalantidis, C. Mellina, and S. Osindero. Cross-dimensional weighting for aggregated deep convolutional features. In *ECCV*, 2016.

[19] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[20] G. Kim and E. P. Xing. Reconstructing storyline graphs for image recommendation from web community photos. In *CVPR*, 2014.

[21] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In *ACL*, 2012.

[22] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi. Treetalk: Composition and compression of trees for image descriptions. *TACL*, 2014.

[23] Y. Liu, Y. Guo, S. Wu, and M. S. Lew. Deepindex for accurate and efficient image retrieval. In *ICMR*, 2015.

[24] A. McCallum, K. Nigam, et al. Employing em and pool-based active learning for text classification. In *ICML*, 1998.

[25] Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In *SIGKDD*, 2007.

[26] E. Mohedano, K. McGuinness, N. E. O'Connor, A. Salvador, F. Marqués, and X. Giró-i Nieto. Bags of local convolutional features for scalable instance search. In *ICMR*, 2016.

[27] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *NAACL*, 2010.

[28] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

[29] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, 2016.

[30] J. Panda, S. Sharma, and C.V. Jawahar. Heritage app: annotating images on mobile phones. In *ICVGIP*, 2012.

[31] J. L. Pech-Pacheco, G. Cristóbal, J. Chamorro-Martínez, and J. Fernández-Valdivia. Diatom autofocusing in brightfield microscopy: a comparative study. In *ICPR*, 2000.

[32] M. O. Riedl and R. M. Young. From linear story generation to branching story graphs. *CG&A*, 2006.

[33] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPR-W*, 2014.

[34] C. Sievert and K. E. Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 2014.

[35] J. Sivic, A. Zisserman, et al. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[36] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*, 2010.

[37] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *ACL*, 2014.

[38] C. A. Thompson, M. E. Califf, and R. J. Mooney. Active learning for natural language parsing and information extraction. In *ICML*, 1999.

[39] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society*, 2001.

[40] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence-video to text. In *ICCV*, 2015.

[41] Y. Verma and C.V. Jawahar. Im2text and text2im: Associating images and texts for cross-modal retrieval. In *BMVC*, 2014.

[42] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. *IJCV*, 2014.

[43] S. Vijayanarasimhan, P. Jain, and K. Grauman. Far-sighted active learning on a budget for image and video recognition. In *CVPR*, 2010.

[44] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[45] D. Wang, T. Li, and M. Ogihara. Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs. In *AAAI*, 2012.

[46] L. Xie, R. Hong, B. Zhang, and Q. Tian. Image classification and retrieval are one. In *ICMR*, 2015.

[47] J. Yang et al. Automatically labeling video data using multi-class active learning. In *ICCV*, 2003.

[48] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *CVPR*, 2016.

[49] L. Zhang, Y. Tong, and Q. Ji. Active image labeling and its application to facial action labeling. *ECCV*, 2008.

[50] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, 2015.