Xplore-M-Ego: Contextual Media Retrieval Using Natural Language Queries

Sreyasi Nag Chowdhury Mateusz Malinowski Andreas Bulling Mario Fritz
Max Planck Institute for Informatics, Saarbrücken, Germany
sreyasi,mmalinowski,bulling,mfritz@mpi-inf.mpg.de

ABSTRACT

The widespread integration of cameras in hand-held and head-worn devices and the ability to share content online enables a large and diverse visual capture of the world that millions of users build up collectively every day. We envision these images as well as associated meta information, such as GPS coordinates and timestamps, to form a collective visual memory that can be queried while automatically taking the ever-changing context of mobile users into account. As a first step towards this vision, in this work we present *Xplore-M-Ego*: a novel media retrieval system that allows users to query a dynamic database of images using spatio-temporal natural language queries. We evaluate our system using a new dataset of real image queries as well as through a usability study. One key finding is that there is a considerable amount of inter-user variability in the resolution of spatial relations in natural language utterances. We show that our system can cope with this variability using personalisation through an online learning-based retrieval formulation.

1. INTRODUCTION

Due to the widespread deployment of visual sensors in consumer products and Internet sharing platforms, we have collectively achieved a detailed visual capture of the world in space and time over the years. In particular, mobile devices have changed the way we take pictures and new technology like life-logging devices will continue to do so. With efficient search engines, viewing images and videos of distant places is a few clicks away. But these search engines do not allow for complex natural language queries with spatio-temporal references. They also largely ignore the users' local context.

Similar to how mobile devices have changed the way we take pictures, we ask how media search should be transformed to make use of the rich context available at query time. What if we quickly want to know what is behind the building in front of us? What if we want to know what a particular cafe looks like to quickly locate it in a busy market area? What if we want to see what our new neighborhood looks like in winter? Our approach makes use of the user's ever-changing context to retrieve results of a spatio-temporal query on a mobile device. We have named our system *Xplore-M-Ego* (read *"Explore Amigo"*) – which stands for Exploration(Xplore) of Media(M) Egocentrically(Ego)".

Figure 1: Sample queries and retrieved images of our contextual media retrieval system Xplore-M-Ego.

2. RELATED WORK

Spatio-temporal Media Retrieval: Previous work [17, 22, 20, 19] allow for browsing media in their geographic context, optionally at different points of time. In contrast, our approach implements egocentrism by taking users' context (geographical location and viewing direction) into account.

Natural Language Query Processing: Answering natural language questions by machines is often realizable by a semantic parser that transforms the question into its formal representation. Modern approaches to training a semantic parser [2, 10, 1] using question-answer pairs have replaced traditional ones using manual annotations [23, 21]. The subjective-ness of question-answering tasks has been pointed out in [13, 15]. Unlike existing work, we target a dynamic and egocentric environment instead of static data.

Media Retrieval Using Natural Language Queries: Previous research have matched queries to manual descriptions of images [11, 7, 6, 4]. Our work does not use human annotations. We extract media based on meta data such as geographical location (GPS coordinates). [18, 8] have used

ICMR'16, June 06 - 09, 2016, New York, NY, USA

ISBN 978-1-4503-4359-6/16/06…$15.00

DOI: http://dx.doi.org/10.1145/2911996.2912044

spatial relations in restricted queries. We lift restrictions such as fixed sentence structure or limited vocabulary.

To the best of our knowledge, no previous work venture into media retrieval with the user's context. The introduction of egocentrism for browsing large media collections opens an unexplored dimension and also aids in human interaction with the computer.

3. CONTEXTUAL MEDIA RETRIEVAL

Our contextual media retrieval system allows users to explore a collective media collection in a spatio-temporal context through natural language questions like those in Figure 1. The formulation of our architecture is shown below.

3.1 Learning - based, Contextual Media Retrieval by Semantic Parsing

In this section we describe a semantic parser architecture and its extension towards a contextual media retrieval task. The probabilistic model of our architecture is shown in Figure 2. A question x is mapped to a latent logical form z, which is evaluated with respect to a world w (database of facts), producing an answer y. The world w consists of w_s (a static database of geographic facts) and w_d (a dynamic database which stores user and media metadata). The logical forms z are represented as labeled trees and are induced automatically from question-answer (x, y) pairs.

3.1.1 Question-Answering with Semantic Parsing

Our approach is built on an existing framework for semantic parsing [10] that is able to answer questions about static facts. In the framework (left part of Figure 2 labeled **Semantic Parsing** and **Interpretation**), 'parsing' translates a question into its logical form z, and 'interpretation' executes z on the dataset of facts w producing its denotation $[\![z]\!]_w$ - an answer. Parameter θ is estimated solely on the training question-answer pairs (x, y) with an EM algorithm maximizing the following posterior distribution:

$$\theta^* := \arg\max_{\theta} \sum_{(x,y)\sim\mathcal{D}} \sum_{z} \underbrace{\mathbf{1}\{y = [\![z]\!]_w\}}_{Interpretation} \quad \underbrace{p(z|x,\theta)}_{SemanticParsing} \tag{1}$$

where $\mathcal{D}$ denotes a training set, $\mathbf{1}\{a = b\}$ is 1 if a condition $a = b$ holds, and 0 otherwise. The posterior distribution marginalizes over a latent set of valid logical forms z. At test time, the answer is computed from the denotation $[\![z^*]\!]_w$ that maximizes the following posteriori:

$$z^* := \arg\max_{z} p(z|x, \theta^*) \tag{2}$$

The distribution over logical forms is modeled by a log linear distribution $p_\theta(z|x) \propto e^{\phi(x,z)^T\theta}$, where the feature vector ϕ measures compatibility between the question x and a logical form z. We perform a gradient descent scheme in order to optimize for parameters θ.

3.1.2 Static and Dynamic Worlds

Related existing work are based on a static environment [10, 1, 13]. In contrast, in our scenario an user (the source of the query - *user* in Figure 2) relocates herself in space and time in a continuously changing environment. The pool of media content (*Collective Visual Memory*) also grows as new media is added (crowd icon in Figure 2). Such an environment leads to decomposition of the world w into a static part w_s, which consists of geographical facts such as names of buildings and theirs GPS coordinates, and a dynamic part w_d. The dynamic world w_d breaks up into w_{d_m} that stores media metadata (timestamp, GPS coordinates) and is updated with continuously growing *Collective Visual Memory*, and w_{d_u} that models the user's context by storing her metadata (GPS coordinates, viewing direction). The latter is set anew for each query before it is fed into the semantic parser. Such representation renders the world $w = w_s + w_d$ static to the semantic parser although it is constantly changing.

3.1.3 Modelling User's Context

Understanding egocentric spatial relations in natural language forms a separate research area by itself [16, 8, 3, 14]. In our work, we approach ambiguity in the frame of reference [15] by defining predicates to resolve the spatial relations *front of, behind, left of, right of* based on the geomagnetic as well as the user-centric reference frames. Temporal references in questions (*what happened here five days ago?, how did this place look like in December?*) are modelled through predicates unifying the referenced time-stamp with those of the media files. The user's context is modelled by recording the GPS coordinates and viewing direction at query-time.

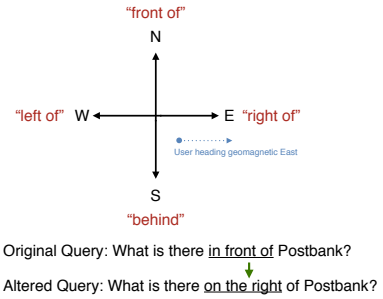

Figure 3: Modification of spatial reference in query for integrating *egocentrism* to media retrieval

However, it is hard to deduce the hidden intent for contextual questions since humans don't adhere to a consistent reference frame. They consider their own physical left or the *left side* of the geographic entity for *left of*. The prior requires an user-centric reference frame which is modelled by modifying the geomagnetic reference frame. Assuming that the user's viewing direction is the local north, the spatial reference in the query is modified in a pre-processing step. This is explained in Figure 3 – if the user faces east and queries for *What is there in front of postbank?*, the question is changed during pre-processing to *What is there on the right of postbank?*. The semantic parser predicts answer for the changed question. For simplicity we use only four basic heading directions - north, south, east and west.

3.1.4 Media Retrieval as Answers

In contrast to previous work on question answering [10, 13], we want to retrieve media instead of text as answers. This can be modeled by generating references to media files as denotations. For example, the question *What is there on the right of the campus center?* would predict the denotation (*image12, image58, ...*) which are references to relevant

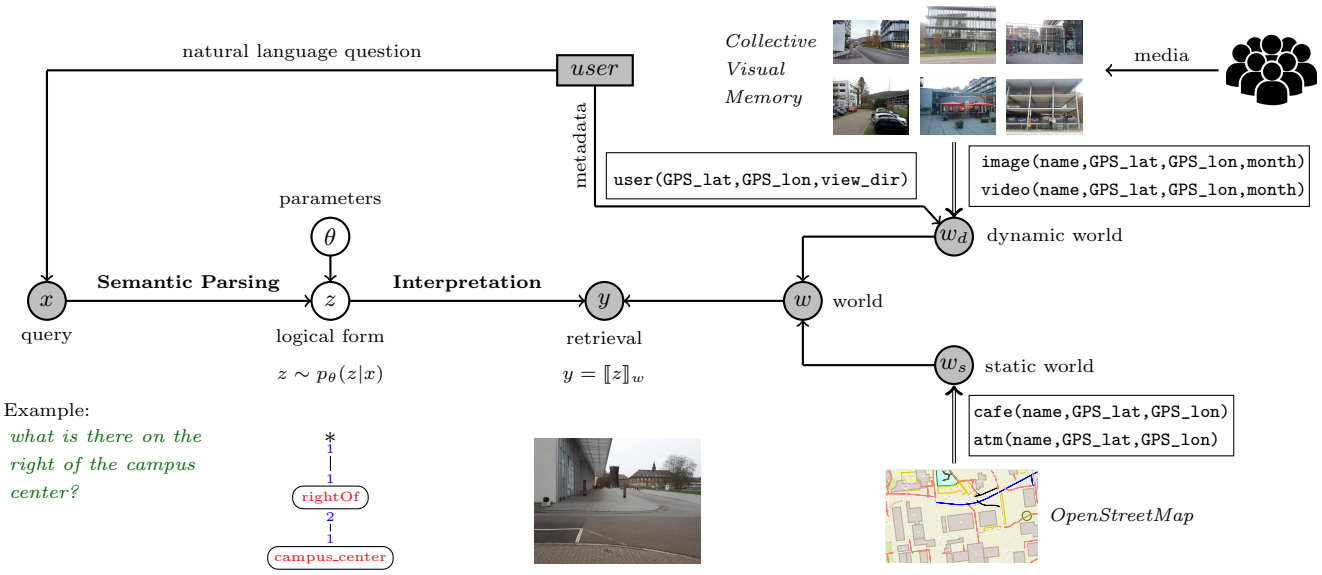

Figure 2: Our probabilistic graphical model: a question x in natural language is automatically mapped into a logical form z by the semantic parser. It is interpreted with respect to a world w to give retrievals y. The world w consists of a static part w_s (a database of geographical facts), and a dynamic part w_d (media content from *Collective Visual Memory* and the user's spatio-temporal context).

images. The actual images are then extracted from the *Collective Visual Memory* and returned to the user.

3.2 Data Collection

To enable spatio-temporal exploration of a geographic area we need a database which record physical features on the ground. To support media retrieval we need a database of images and videos rich with metadata. We also need natural language queries paired with corresponding media content as retrievals for training and testing our model. In the absence of a suitable benchmark, we record our own data set.

Geographical Facts: We extract the our geographic data from OpenStreetMap [5]. In our study, we restrict the spatial scope of our system to a university campus. We use information such as the type of the physical entity (*building, cafe, highway* etc.), their names, and their GPS coordinates.

Collective Visual Memory: Participants formulated questions and captured the photo(s)/video(s) that they expect as corresponding answers. 1000 questions-answer pairs with spatial references were collected. Question-answer pairs with temporal references were not collected because of the trivial infeasibility of capturing events from the past. The questions follow no particular template and contains a variety of spatial relations. The dataset was randomized and divided into 500 train and 500 test questions. To introduce sufficient amount of variations in natural language we chose participants from different cultural and linguistic background. Our instance of the *Collective Visual Memory* consists of 1025 images and 175 videos. The dataset is publicly available[1].

4. MODEL EVALUATION

Humans are inherently inconsistent in their perception of directions and idea of reference frames [9, 15]. The nature of speaking English questions also has variations based on a person's socio-cultural background. Hence, a system relying on fixed question templates and a particular set of rules to resolve spatial references does not guarantee high accuracy. To better understand these perceptual biases and efficiently analyze the system, a series of user studies were conducted.

4.1 Human Disagreements on Retrieved Results

The goal of this user study is to observe how accurate regular users find our system. Five users evaluated the retrieved results for 500 test questions as *relevant* or *irrelevant*. A canonical reference frame was used in this experiment to resolve spatial relations. According to this convention, *front of*=*north of*, *behind*=*south of*, *right of*=*east of* and *left of*=*west of*.

We observed that the opinions varied for each question. Based on this observation we divide the test questions into six groups – *(5,0)*, queries for which all five users agreed that the retrievals were relevant; *(4,1)*, queries for which four users found the retrievals relevant and one user found them irrelevant and likewise. Figure 4 depicts the result of this analysis. For 26.67% of the queries all five users deemed the retrievals relevant. However, if we consider the cases in which most of the users found the retrievals relevant, this number rises to 40%. The numbers in the middle region of the graph in Figure 4 point out the prominent difference in opinions among participants. This accounts for about 25% of all queries. We observed that the inter-user variability stems from the inherent inconsistencies with regards to reference frame resolution. This result also hints towards the difficulty of the problem at hand. From this observation we conjecture that the use of user-centric reference frames instead of geomagnetic reference frame could improve the performance of the system. In the deployment of the user-centric reference frame we mean to follow the user's physical egocentric directions – for example, her *right hand side* for *right of* etc.

[1]http://www.mpi-inf.mpg.de/xplore-m-ego

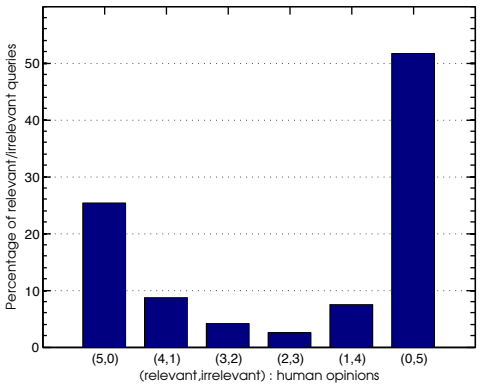

Figure 4: Inter-user variability in opinion

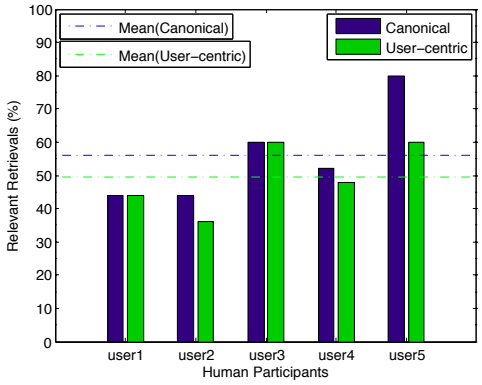

Figure 5: Difference in reference frame resolution among humans

4.2 Canonical and User-centric Reference Frame

The aim of this evaluation is to study the impact of using two different conventions of spatial relations resolution. Users were given two sets of retrieved results for each question – one set of media files retrieved according to the geomagnetic reference frame and the second set retrieved according to the user-centric reference frame. The experimental settings are similar to the previous user study.

Figure 5 shows the result of this user study. user1 and user3 remained neutral to the use of separate reference frames while the other users slightly preferred the canonical reference frame over the user-centric reference frame. This observation further highlights the subjectivity of the task.

4.3 Personalization of Xplore-M-Ego

Having observed this inter-person subjectivity, we hypothesize that personalization of our media retrieval system would increase its accuracy on a per user basis. By using an online relevance feedback mechanism, five users *(U1,U2,U3,U4,U5)* trained five different query-retrieval models *(M1,M2,M3,M4, M5)* with 500 questions. Every user was then asked to evaluate all five models keeping the identity of the models hidden.

The quantitative analysis of this study is shown in Figure 6. The diagonal shows the user-specific evaluation results and the rows depict inter-user evaluation results. It is clear from the figure that users deemed their own models more accurate than those trained by others. This observation leads us to believe that the query-retrieval model can be trained over time through relevance feedback to adapt to user-specific preferences of spatial relation resolution – hence, it should be personalized.

$$\begin{array}{c} \\ M1 \\ M2 \\ M3 \\ M4 \\ M5 \end{array} \begin{array}{c} \begin{array}{ccccc} U1 & U2 & U3 & U4 & U5 \end{array} \\ \begin{pmatrix} 25.36 & 8.84 & 14.84 & 12.21 & 15.79 \\ 20.38 & 35.9 & 33.9 & 24.06 & 25.25 \\ 15.66 & 13.25 & 37.33 & 15.06 & 28.59 \\ 14.49 & 24.32 & 33.38 & 36.86 & 28.06 \\ 8.58 & 11.34 & 23.53 & 17.48 & 37.84 \end{pmatrix} \end{array}$$

Figure 6: Quantitative analysis of personalization of Xplore-M-Ego: F1-score

USE Questionnaire	Mean	SD
It is useful.	6.2	0.63
It saves me time when I use it.	6.1	0.73
It is easy to use.	6.3	0.48
I can use it without written instructions.	5.8	1.22
Both occasional and regular users would like it.	5.4	1.42
I learned to use it quickly.	6.5	0.52
I quickly became skillful with it.	6.1	0.99
I am satisfied with it.	5.5	0.52
It is fun to use.	6.3	0.82
I would recommend it to a friend.	5.9	0.73

Table 1: User Experience Evaluation: Mean Rating and Standard Deviation. The grades are between 1 ('strongly disagree') and 7 ('strongly agree').

4.4 User Experience Evaluation

In this usability/desirability study, ten participants were given the Google Glass installed with our client-side application and asked to walk around in the university campus while making voice queries with spatio-temporal references. Afterward they were asked to fill in the USE Questionnaire [12]. This questionnaire has four groups of questions – Usefulness, Ease of Use, Ease of Learning and Satisfaction. Each question can be rated on a scale from 1 to 7, 1 meaning 'strongly disagree' and 7 meaning 'strongly agree'. Ten questions most representative of the entire questionnaire are chosen. The mean and standard deviation of the ratings of these questions are shown in Table 1.

The result of this evaluation shows that regular users find our application useful, easy to learn, and satisfying.

5. CONCLUSION

In this paper we propose *Xplore-M-Ego* – a novel system for media retrieval using spatio-temporal natural language queries in a dynamic setting. Our work brings forth a new direction to this paradigm by exploiting a user's current context. Our approach is based on a semantic parser that learns to infer interpretations of the natural language queries from question-answer pairs. We contribute several extensions which enable the user to dynamically refer to her context by spatial and temporal concepts. We analyze the system with various user studies that highlight the importance of our adaptive and personalized training approaches. For further details and discussions on our approach and observations we direct the readers to our technical report[2].

[2]http://arxiv.org/pdf/1602.04983.pdf

6. REFERENCES

[1] J. Berant and P. Liang. Semantic parsing via paraphrasing. In *Proceedings of ACL*, 2014.

[2] J. Clarke, D. Goldwasser, M.-W. Chang, and D. Roth. Driving semantic parsing from the world's response. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 18–27. Association for Computational Linguistics, 2010.

[3] S. Guadarrama, L. Riano, D. Golland, D. Gouhring, Y. Jia, D. Klein, P. Abbeel, and T. Darrell. Grounding spatial relations for human-robot interaction. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 1640–1647. IEEE, 2013.

[4] A. Hakeem, M. W. Lee, O. Javed, and N. Haering. Semantic video search using natural language queries. pages 605–608, 2009.

[5] M. Haklay and P. Weber. Openstreetmap: User-generated street maps. *Pervasive Computing, IEEE*, 7(4):12–18, 2008.

[6] M. Hwang, H. Kong, S. Baek, and P. Kim. A method for processing the natural language query in ontology-based image retrieval system. In *Adaptive Multimedia Retrieval: User, Context, and Feedback*, pages 1–11. Springer, 2007.

[7] O. Kucuktunc, U. Güdükbay, and Ö. Ulusoy. A natural language-based interface for querying a video database. *IEEE MultiMedia*, 14(1):83–89, 2007.

[8] T. Lan, W. Yang, Y. Wang, and G. Mori. Image retrieval with structured object queries using latent ranking svm. In *Computer Vision–ECCV 2012*, pages 129–142. Springer, 2012.

[9] S. C. Levinson. *Space in language and cognition: Explorations in cognitive diversity*, volume 5. Cambridge University Press, 2003.

[10] P. Liang, M. I. Jordan, and D. Klein. Learning dependency-based compositional semantics. *Computational Linguistics*, 39(2):389–446, 2013.

[11] V. Lum and K.-c. K. Kim. Intelligent natural language processing for media data query. In *Proc. 2nd Int. Golden West Conf. on Intelligent Systems*, 1992.

[12] A. M. Lund. Measuring usability with the use questionnaire. *Usability interface*, 8(2):3–6, 2001.

[13] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1682–1690, 2014.

[14] M. Malinowski and M. Fritz. A pooling approach to modelling spatial relations for image retrieval and annotation. *arXiv:1411.5190 [cs.CV]*, November 2014.

[15] M. Malinowski and M. Fritz. Towards a visual turing challenge. In *NIPS Workshop on Learning Semantics*, 2014.

[16] T. Regier and L. A. Carlson. Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2):273, 2001.

[17] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. *ACM transactions on graphics (TOG)*, 25(3):835–846, 2006.

[18] S. Tellex and D. Roy. Towards surveillance video search by natural language query. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, page 38. ACM, 2009.

[19] J. Tompkin, K. I. Kim, J. Kautz, and C. Theobalt. Videoscapes: exploring sparse, unstructured video collections. *ACM Transactions on Graphics (TOG)*, 31(4):68, 2012.

[20] J. Tompkin, F. Pece, R. Shah, S. Izadi, J. Kautz, and C. Theobalt. Video collections in panoramic contexts. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 131–140. ACM, 2013.

[21] Y. W. Wong and R. J. Mooney. Learning synchronous grammars for semantic parsing with lambda calculus. In *Annual Meeting-Association for computational Linguistics*, volume 45, page 960. Citeseer, 2007.

[22] F. Wu and M. Tory. Photoscope: visualizing spatiotemporal coverage of photos for construction management. pages 1103–1112, 2009.

[23] L. S. Zettlemoyer and M. Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *arXiv preprint arXiv:1207.1420*, 2012.