Shihan (David) Ma
University of Florida

Andrei P. Kirilenko
University of Florida

Automated Identification of Tourist Activities in Social Media Photographs: a Comparative Analysis Using Visual-based, Textual-based and Joint-based Methods

Abstract: The studies pertaining to image identification of tourist photographs are mainly dealing with objects/landscapes, while the *activities* of tourists interacting with these objects is not well researched. The eligible methods to identify in-depth activities are likewise greatly missing. In this paper, we first explore the feasibility of using different data approaches (visual and textual) to identify tourist activities in social media photos. We further develop a multimodal method combining both text-based and visual-based information. The performances of these methods are compared and validated by manual reviewing. The findings confirm that data fusing methodology is improving identification of micro-level activities.

Keywords: tourist activity, image identification, Instagram, social media photography, data fusion

Shihan (David) Ma, M.S.
Doctoral Candidate
Dept. of Tourism, Hospitality and Event Management
College of Health and Human Performance
University of Florida
206C Florida Gym,
P.O. Box 118208,
Gainesville, FL 32611-8208, USA
Email: david.ma@ufl.edu
Phone: +1 352 294 1679

Andrei P. Kirilenko, Ph.D.
Associate Professor
Dept. of Tourism, Hospitality and Event Management
College of Health and Human Performance
University of Florida
240B Florida Gym,
P.O. Box 118208
Gainesville, FL, 32611-8208, USA
Email: andrei.kirilenko@ufl.edu
Phone: +1 352 294 1648

Shihan (David) Ma is a Ph.D. candidate in the Department of Tourism, Hospitality and Event Management at University of Florida. His research concentration is social media data analysis in tourism, regarding tourist perception, behavior, mobility pattern and destination marketing opportunities through big data mining**.**

Dr. Andrei P. Kirilenko is Associate Professor in the Department of Tourism, Hospitality and Event Management at the University of Florida. His research interests include big data analysis, data mining, tourism analytics, climate change impacts, and sustainability issues.

Introduction

Tourist photographs are documentary evidence of the travel, and hence valuable data to analyze the tourist experience and the impression of the destinations (MacKay & Couldwell, 2004). With the exponentially growing popularity of photo sharing social networks among tourists, the access to tourist photography has vastly improved on terms of quantity and information richness. The latter comes with a new resource, photograph metadata, which was unavailable in the era of analog photography. Nowadays, the online digital photographs frequently come along with the information about the authors, keywords (tags), captions, timestamps, reactions ("likes"), and the geographical location of the photographed place. This auxiliary data allow tourist photography analysis from the temporal (when), spatial (where) content (what) or network (who is listening to whom) perspectives (Stepchenkova & Zhan, 2013).

The abovementioned increase in information richness of online shared photographic images has been largely unnoticed by the academic community, which concentrated on new possibilities coming from improvements in data quantity and accessibility. The photograph metadata resource meanwhile is rarely accounted for in photograph analysis. The following paper explore the possibilities of data fusion (that is, integrating multiple data sources) to improve tourist photography analysis. The specific problem tackled in the research is identification of tourist activities from the online photographs.

There have been numerous studies pertaining to identification and classification of tourist photographs based on systematic content analysis (e.g. Camprubi, Guia, and Comas 2014; Donaire 2011) that mainly dealt with identification of objects, such as landscape, buildings, people, etc. Meanwhile, another travel dimension, the *activities* of tourists interacting with these objects is not well researched (Vu, Li, Law, & Zhang, 2017). Admittedly, it is difficult to accurately identify the tourist activities from photographs, either manually, or using image recognition algorithms. For instance, when hiking or jogging, a social media user may post a picture of forests or natural landscape, while such activities cannot be discerned by the imagery depiction. The situation is more complicated if certain spots are shared by multiple activities. The posted photographs tend to resemble similar surroundings (e.g., forest) while the photographers participate in distinctly different activities (e.g., running and birdwatching).

The improved understanding of tourist activities may come from the descriptive text associated with the photos. As (Deng & Li, 2018) noticed, such descriptive information are rich and essential parts of the cognitive components of photographers, containing the knowledge about the place where the photos were taken, the objects pictured in the image, the people they were with, the sensation they were experiencing as well as the activity they were conducting.

The proposed methodology fuses two data sources, visual (photograph) and textual (photograph caption) in a single framework that allows automated identification of preferred tourist activities. The methodology was successfully tested on a large dataset of a popular travel destination, Lake Texoma, shared between Oklahoma and Texas, USA.

The following chapters first explore the feasibility of using different data approaches (visual and textual) to identify tourist activities in social media photos. We further develop a multimodal method combining both text-based and visual-based information. Finally, the performances of these methods are compared and validated by manual reviewing. The findings confirm that the data fusing methodology is improving identification of micro-level activities.

Related Works

Classification of destination photos has been a common practice in previous tourism photography researches, with the utilization of content analysis in numerous empirical studies (Camprubi et al., 2014; Choi, Lehto, & Morrison, 2007; Donaire, 2011; Donaire, Camprubí, & Galí, 2014; Galí Espelt & Donaire Benito, 2005; Govers & Go, 2005; Jenkins, 2003; Pritchard & Morgan, 2001; Pritchard, Morgan, & Morgan, 1995). Both textual and visual information have been scrutinized to classify destination pictures with the common image categories including culture and heritage, destination icons, landscape and nature, people, activity and services (Camprubi et al., 2014). The identification of imagery elements was further to segment tourists based on their common interests expressed in their photos (Donaire et al., 2014).

Recently, the advance of automatic image identification algorithms in computer science, and especially the machine learning made the automated analysis of large volumes of photos possible. Convolutional Neural Network (CNN) framework trained on the ImageNet dataset is the most popular method in such application (Zhou, Liu, Oliva, & Torralba, 2014). The emerging employment of fusion classifier approaches improved the accuracy of image identification and classification. For instance, Tang et al.(2015) demonstrated that significant improvement in identification of landscape features are possible when auxiliary geographical data associated with image geotags are used. Duong, Lebret, & Aberer (2017) proposed and successfully tested a multimodal classification approach with both textual and imagery information to identify the contextual emotions in social media photos.

Despite the demonstration of the advantages of image classification approaches in computer science, their application in tourism researches is still greatly missing. The limited tourist activity identification findings in existing studies are more or less the derivatives of other research questions. Sun and Fan (2014) used the spatial, temporal and semantic information in geotagged photograph to identify whether a photo is related to social events (e.g., festivals, parades, protests, sports, etc.) with binary logistic regression. Although the method resulted in good performance of 71% accuracy, the resultant classification categories were merely binary and specific to event identification. Oteros-Rozas et al. (2016) detected more detailed activity categorization during their evaluation on landscape images. These activities were actually the sub-categories of "people", including sunbathing, swimming, fishing, sailing/windsurf/motor, hiking/walking, biking, resting, skiing, and observing; and yet such classification was based on manual reviewing with small sample size.

Overall, automatic image identification and classification approaches for tourist activities are far from effective and adequate; the methods to identify in-depth activities are likewise greatly missing. In this study, we demonstrate how textual- and visual-based fusion improve in activity identification compared with either visual-based or textual-based identification alone.

Data and Methods

The object of study was Lake Texoma recreation area. Lake Texoma is an artificial reservoir shared between Oklahoma and Texas which is managed by the US Army Corps of Engineers. The lake area provides a wide range of settings and facilities for hiking, fishing, boating, camping and other outdoor activities and attracts 6 million visitors every year. The sheer variety of recreational opportunities and property rights (federal, state, county, and private) make direct observation of visitors' activities impossible. We collected 13,875 photographs taken within the lake's area and published on Instagram. The images along with metadata were obtained from a certified data aggregator Picodash. The photographs and accompanying hashtags were used in activity identification and geolocation information was used to separate visitors from locals and to identify activity areas (not covered in this short communication).

Image labeling i.e. recognizing the elements of an image was completed with Google Vision AI software (cloud.google.com/vision/), which uses a pre-trained machine learning model to automatically annotate the images. Thus, each photograph was converted into a sequence of labels (visual tokens). In total, we identified 3,018 unique tokens taking a long tail distribution. Following the standard content analysis procedure, we filtered out the least frequent and the most frequent tokens (those appearing in less than 0.5% and more than 50% of the images, accordingly), which resulted in a list of 314 tokens. The entire visual component of the dataset was then represented in a 13,875*314 image-visual token binary matrix.

The textual component of the dataset was created from the photographs' hashtags. The hashtags represent the self-reported cognitive components of the images. The data were first pre-processed by splitting the hashtags into natural language words, e.g., "#lakeday" was transformed into the words "lake day". Then, the stop-words were removed, the words were lemmatized with Gensim package (Řehůřek ; R. Řehůřek, Řehůřek, & Sojka, 2010), and the least and the most frequent words were removed as described above resulting in 244 words (tokens). The entire textual component of the dataset was then represented in a 12,348 * 244 image-textual token binary matrix. The differences in matrix dimension between the textual and visual components are due to missing textual component in some photographs.

Finally, data fusion was accomplished by creating the joint visual/textual token list for each image. Following the procedure described above, we found 538 joint visual-textual tokens describing the photographs. The fused photograph dataset was then represented in a 13,875*538 binary matrix.

The collected photographs were automatically assigned into different topics according to their content using the Rapidminer implementation (https://rapidminer.com) of the Latent Dirichlet Allocation (LDA) algorithm. LDA is a popular unsupervised classification approach to detect recurring patterns of words in a collection of documents by interpreting them as manifestations of hidden topics. While initially developed for textual data, LDA was also successfully applied in image classification (Elango & Jayaraman, 2005; Rasiwasia & Vasconcelos, 2013). The generated topics were then manually interpreted and named based on the semantic meanings of the words (tokens) included into each topic. Hence, the photographs were classified into multiple categories (topics) of natural sceneries, leisure activities, or a vacation selfie, together with the probabilities associated with each category.

Results

This section describes the outcomes of photograph classification using three ways of data representation: textual, visual, and fused.

LDA topic modeling applied to visual photograph representations resulted in 30 topics, with meaningful interpretation of 28 topics. Overall, those topics were grouped into 5 photograph types: nature, human, activity, structure, and others (Table 1). Importantly, in the activity type of photographs only five activities were recognized. These activities were commonly associated with a homogeneous background such as a body of water and with distinct objects such as a boat. The majority of tourist activities in the area such as *hiking*, *hunting*, or *birdwatching* were apparently missed by the topic modeling based on the visual data.

Table 1 Topics identified from visual-based representation

Type	Natural Scenery (7)	Human Presence (5)	Activity (5)	Objects (8)	Structure (3)
Topics	• water body • sky/sun • bank/shore • wood/jungle • grassland • flower • night/dark scene	• child & baby • part of face • bare chest • body part (hand/leg) • selfie/people • groups/community	• boating • watersport • fishing • camping	• dog • birds/beak • technical devices • auto/vehicle • food/dishes • swimwear • eyewear/glass • outerwear	• building interior • dock/marina • road sign

LDA topic modeling applied to the textual photograph representations resulted in 30 topics. Eight of the extracted topics expressed sentimental or similar information not relevant for identification of the objects of activities such as “look pretty” or “loving life”. The remaining 22 topics are identified as in 4 different types: activity, location, event and Object (Table 2). Importantly, unlike the topic modeling based on the visual data, classification of the textual data successfully recognized many tourist activities such as *camping*, *hunting*, and *retreating*. One of the contribution factors of this successful activity recognition was the presence of hashtags expressing business name such as Kent Outdoor (hunting service provider), location such as Cross Timbers Trailing (a popular hiking trail), or time such as the Fourth of July.

Table 2 Topics identified from text-based representation

Type	Activity (10)	Location (3)	Event (6)	Object (3)
Topics	• water sporting • fishing • hiking/trailing • retreating • golfing • camping • wedding • rock climbing • hunting • birdwatching	• Texoma • Eisenhower • Hagerman	• Thanksgiving • Labor Day • spring break • Forth July • Christmas • Mardi Gras	• dog • food • beautiful scenery

Finally, LDA topic modeling applied to the fused visual-textual photograph representations resulted in 50 different topics with 35 topics similar to those identified from either the visual or textual data as described above and 15 new topics (Table 3). Importantly, these newly identified topics included the activity types not previously recognized in either textual or visual topic model, namely *bicycling* and *yoga & fitness*. Other recognized activities became more solidly interpreted; for example, the *rock-climbing* topic, while recognized from the textual (but not visual) representation, in data fusing representation included additional tokens "geology formation", "cave", "boulder" and "soil", improving topic interpretation. Similarly, data fusion topic identification was able to recognize different types of water sports which were not present in textual or visual representations: *wakeboarding* and *jet-ski*.

Table 3 Topics identified from joint representation

Inherited from single-modal method	**Fusion and addition in multi-modal method**
Visual-based topics (22)	Newly identified topics (9)
• 7 natural scenery topics • 5 human presence topics • 3 structure topics • 4 object topics	• 7 non-activity topics (insect, reptile animals, cats, party event, music event, flag & symbol, art painting)

• 2 generic topics (meaningless) • 1 activity ***(boating)***	• 2 activity topics ***(bicycling, yoga & fitness)***
Textual-based topics (9)	Fusion activity topics (5)
• 2 location topics • 5 tag-related topics (meaningless) • 2 activity topics ***(retreat, birdwatching)***	• **golfing** (more environmental detail: venue, course, grass) • **rock climbing** (more environmental detail: geology formation, cave, soil) • **hunting** (more animal detail: bird, duck, geese) • **wakeboarding/boarding** (subcategory of watersport) • **jet-ski** (subcategory of watersport)
Common topics in both (5)	
• 2 object topics (dog, food) • 3 activity topics ***(hiking/run, fishing, camping)***	

Discussion

Overall, the data fusing topic identification outperformed both the visual and textual approaches in terms of interpretability of the identified topics of the photographs and the number of identified types of activities. This increase in performance comes from the synergistic combination of the visual representation's strength in recognizing the objects and the textual representation's strength in recognizing the action. Some examples of this synergy are provided in Table 4.

With topic identification using the visual or textual data alone, the textual-based activity identification outperforms the visual one. We hypothesize that the underlying reason is that while the visual data excels at describing the objects, the textual data is more informative in describing the actions performed with or between those objects. Three examples of this observation are shown in Table 4. The left columns example on Table 4 shows that the fusion classifier keeps consistence with textual classifier in prediction accuracy, while visual classifier may mislead. The second example illustrates how joint classifier perform when textual information is not adequate (lack of hashtags) and failing in predict activities. In the third example, joint classifier successfully identifies the yoga fitness activity, while textual classifier mis- classifies it into a generic tag group and visual classifier mis-classifies it in a technical device group.

Table 4 classification results with multiple classifiers. Manual classification is used for validation of the automated classifiers.

Image description		People sitting in a church hall	A man with a golf club	A women doing yoga streching by the lakeside
Textual data		bring the Word of God hard as we study Daniel 1 at the Bring Your Own Bible Retreat #truthsda #HappySabbath #Retreat #EvenSoCome #God #Jesus #Christian #JesusFollowe	Nice putt Willy.........miss you already!	This week on the mat I'm playing with flowing with the element of water. Extended triangle pose is a beautiful way to open the side bodies as you ground into the earth with the lower body. Open bodies make for open minds #yoga #yogi #yogateacher #yogainspiration #dallasyoga #triangle #pose #mindfulness
Visual data from image recognition software		community, event, architecture, building, crowd	golfer, golf, golf equipment, professional golfer, sport venue, golf club, putter, iron, pitch and putt, golf course	physical fitness, yoga, stretching, leg, balance, performance, wood, happy, sea, pilates
Classifi- cation	Visual	Community	Golfing	Technical devices
	Textual	Retreating	Generic tag topic	Generic tag topic
	Fusion	Retreating	Golfing	Yoga/fitness
	Manual	**Retreating**	**Golfing**	**Yoga**

Conclusion

Automated tourist activity recognition is possible with utilization of either visual or textual data extracted from the online photo sharing platforms. In comparison, the visual data seems to be better suitable in identifying objects and surrounding scenes, while using the textual data is more effective in identifying the activities. The data fusion approach which synergistically integrates the texts and images seems to keep the advantages of the text-only and visual-only classifiers, enabling recognition of a larger number of activity types with better reliability. Further improvement is possible with integrating additional auxiliary data such as geolocation of the photographs, which would account for geographical distribution of different activity types. Yet another improvement in activity recognition may be possible with switching from a pre-trained image classifier and unsupervised LDA topic modeling to image recognition algorithms specifically trained on tourism data and to a supervised LDA model.

REFERENCES

Camprubi, R., Guia, J., & Comas, J. (2014). Analyzing Image Fragmentation in Promotional Brochures. *Journal of Hospitality & Tourism Research*, *38*(2), 135–161. https://doi.org/10.1177/1096348012451451

Choi, S., Lehto, X. Y., & Morrison, A. M. (2007). Destination image representation on the web: Content analysis of Macau travel related websites. *Tourism Management*, *28*(1), 118–129. https://doi.org/10.1016/j.tourman.2006.03.002

Deng, N., & Li, X. (Robert). (2018). Feeling a destination through the “right” photos: A machine learning model for DMOs’ photo selection. *Tourism Management*, *65*, 267–278. https://doi.org/10.1016/j.tourman.2017.09.010

Donaire, J. A. (2011). Barcelona Tourism Image Within the Flickr Community. *Cuadernos de Turismo*, *27*, 1061–1062.

Donaire, J. A., Camprubí, R., & Galí, N. (2014). Tourist clusters from Flickr travel photography. *Tourism Management Perspectives*, *11*, 26–33. https://doi.org/10.1016/j.tmp.2014.02.003

Duong, C. T., Lebret, R., & Aberer, K. (2017). *Multimodal Classification for Analysing Social Media*. Retrieved from http://arxiv.org/abs/1708.02099

Elango, P. K., & Jayaraman, K. (2005). *Clustering Images Using the Latent Dirichlet Allocation Model*.

Galí Espelt, N., & Donaire Benito, J. A. (2005). The social construction of the image of Girona: A methodological approach. *Tourism Management*, *26*(5), 777–785. https://doi.org/10.1016/j.tourman.2004.04.004

Govers, R., & Go, F. M. (2005). Projected Destination Image Online: Website Content Analysis of Pictures and Text. *Information Technology & Tourism*, *7*(2), 73–89.

https://doi.org/10.3727/1098305054517327

Jenkins, O. H. (2003). Photography and travel brochures: The circle of representation. *Tourism Geographies*, *5*(3), 305–328. https://doi.org/10.1080/14616680309715

MacKay, K. J., & Couldwell, C. M. (2004). Using Visitor-Employed Photography to Investigate Destination Image. *Journal of Travel Research*, *42*(4), 390–396. https://doi.org/10.1177/0047287504263035

Oteros-Rozas, E., Martin-Lopez, B., Fagerholm, N., Bieling, C., & Plieninger, T. (2016). Using social media photos to explore the relation between cultural ecosystem services and landscape features across five European sites. *Ecological Indicators*. https://doi.org/10.1016/j.ecolind.2017.02.009

Pritchard, A., & Morgan, N. J. (2001). Culture, identity and tourism representation: marketing Cymru or Wales? *Tourism Management*, *22*(2), 167–179. https://doi.org/10.1016/S0261-5177(00)00047-9

Pritchard, A., Morgan, N., & Morgan, N. (1995). Evaluating vacation destination brochure images : the case of local authorities in Wales. *Journal of Vacation Marketing*, *2*(1), 23–38. https://doi.org/10.1177/135676679500200103

Rasiwasia, N., & Vasconcelos, N. (2013). Latent dirichlet allocation models for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(11), 2665–2679. https://doi.org/10.1109/TPAMI.2013.69

Řehůřek ; R. Řehůřek, R., Řehůřek, ; R, & Sojka, P. (2010). Fast and Faster: A Comparison of Two Streamed Matrix Decomposition Algorithms. In *Lecture Notes in Computer Science* (Vol. 6611). Retrieved from Springer website: http://www.eudml.eu

Stepchenkova, S., & Zhan, F. (2013). Visual destination images of Peru: Comparative content analysis of DMO and user-generated photography. *Tourism Management*, *36*, 590–601. https://doi.org/10.1016/j.tourman.2012.08.006

Sun, Y., & Fan, H. (2014). Event identification from georeferenced images. *Lecture Notes in Geoinformation and Cartography*, 73–88. https://doi.org/10.1007/978-3-319-03611-3_5

Tang, K., Paluri, M., Fei-Fei, L., Fergus, R., & Bourdev, L. (2015). Improving image classification with location context. *Proceedings of the IEEE International Conference on Computer Vision*, *2015 Inter*, 1008–1016. https://doi.org/10.1109/ICCV.2015.121

Vu, H. Q., Li, G., Law, R., & Zhang, Y. (2017). Tourist Activity Analysis by Leveraging Mobile Social Media Data. *Journal of Travel Research*, *57*(7), 1–16. https://doi.org/10.1177/0047287517722232

Zhou, B., Liu, L., Oliva, A., & Torralba, A. (2014). Recognizing city identity via attribute analysis of geo-tagged images. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8691 LNCS*(PART 3), 519–534. https://doi.org/10.1007/978-3-319-10578-9_34