

Utilizing Sensor-Social Cues to Localize Objects-of-Interest in Outdoor UGVs

Yingjie Xia¹, Luming Zhang²(✉), Liqiang Nie³, and Wenjing Geng³

¹ College of Computer Sciences, Zhejiang University, Hangzhou, China

² Department of CSIE, Hefei University of Technology, Hefei, China
zglum@gmail.com

³ School of Computing, National University of Singapore, Singapore, Singapore

Abstract. A huge number of outdoor user-generated videos (UGVs) are recorded daily due to the popularity of mobile intelligent devices. Managing these videos is a tough challenge in multimedia field. In this paper, we tackle this problem by performing object-of-interest (OOI) recognition in UGVs to identify semantically important regions. By leveraging geo-sensor and social data, we propose a novel framework for OOI recognition in outdoor UGVs. Firstly, the OOI acquisition is conducted to obtain an OOI frame set from UGVs. Simultaneously, the classified object set recommendation is performed to obtain a candidate category name set from social networks. Afterward, a spatial pyramid representation is deployed to describe social objects from images and OOIs from UGVs, respectively. Finally, OOIs with their annotated names are labeled in UGVs. Extensive experiments in outdoor UGVs from both Nanjing and Singapore demonstrated the competitiveness of our approach.

1 Introduction

Location-based services provided by social networks, such as Facebook and Twitter, remarkably enrich the quantity of multimedia content tagged by geo-sensor including latitude and longitude. Besides, the popularity of mobile devices with sensors makes capturing, uploading, and sharing of outdoor user-generated videos (UGVs) highly convenient. This motivates us to investigate effective techniques to manage these Internet-scale UGVs.

To handle the vast amount of UGVs on social networks, we focus on object-of-interest (OOI) recognition, *i.e.*, building an OOI recognition system by leveraging both visual features and sensor-social data. The major benefit is that not only it can localize OOIs, but also it is highly efficient and accurate by adopting sensor-social data. Such a recognition system would be of tremendous value and significance for a large body of multimedia applications. For example, Zheng *et al.* [1] proposed a web-scale landmark recognition engine by leveraging the vast amounts of multimedia data. However, most GPS-tagged recognition systems depend on a large collection of images to achieve accurate visual clusters. The existing techniques, however, are unsuitable for OOI recognition in UGVs because of two reasons: (1) OOI recognition in UGVs should be a lightweight

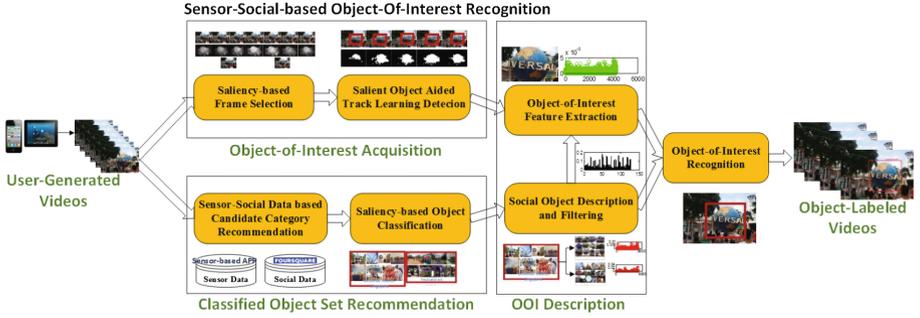


Fig. 1. The proposed OOI recognition pipeline using sensor-social data

application since UGVs are usually captured by mobile devices. Therefore, off-loading many recognition tasks onto cloud servers may increase latency and response time; (2) typical approaches which acquire a “complete” image dataset to handle object recognition may consume extra computation practically.

The explosive growth of UGVs leads to a significant challenge on how to efficiently organize large video repositories and make them searchable. Common approaches adopt content-based media analysis to extract visual features for similarity matching. However, due to the overwhelming amount of video materials, it is inappropriate to perform feature matching on a frame-by-frame level. In this work, we understand video content at object-level in a lightweight way. We propose to recognize OOIs in UGVs with user-intentionally captured objects. Similar to our work, Hao *et al.* [2] focused on point-of-interest detection in sensor-rich videos. It was achieved by analyzing a large number of sensor-rich videos automatically and comprehensively. This implies that the method is unsuitable for OOI recognition in a single UGV.

An overview of the proposed method is presented in Fig. 1. We focus on analyzing UGVs uploaded on social networks at object-level, by utilizing sensor-social data. Given a collection of UGVs, the OOI acquisition and the classified object set recommendation are conducted simultaneously. The former task can be formulated as salient objects extraction, where saliency indicates the informative/interesting regions within a scenery. To obtain the most representative frames, a saliency-guided selection algorithm is proposed to filter frames with similar saliency distributions. For the latter task, candidate categories are recommended by leveraging sensor-social data. Metadata including timestamps, GPS coordinates, accuracy, and visible distances are employed as sensor data. Afterward, salient objects are extracted in social images based on category classification. A spatial-pyramid architecture [3] is adopted to describe social objects and OOIs in UGVs for its robustness in scene modeling. And the Euclidean distance is employed to measure the similarity between the classified and labeled object sets. Finally, OOIs associated with their annotated names are labeled in UGVs frame-by-frame. Experiments on object-level video summarization and content-based video retrieval demonstrate the usefulness of our method.

2 Related Work

OOI detection is a widely used technique in a variety of domains, *e.g.*, video analysis and retrieval. Object/saliency detection and region of interest accumulation are typical approaches to localize OOIs. Most existing work on object detection depends on the sliding window approaches [4, 5]. They might be computationally intractable since windows detection with various scales are evaluated at many positions across the image. To accelerate computation, Harzallah *et al.* [8] and Vedaldi *et al.* [9] designed cascade-based methods respectively to discard windows at each stage, where richer features are adopted progressively. Cinbis *et al.* [10] developed an object detection system by employing the Fisher vector representation. State-of-the-art performance was achieved for image and video categorization. Kim *et al.* [11] proposed an OOI detection algorithm based on the assumption that OOIs are usually located near the image centroid. Zhang *et al.* [12] introduced a novel approach to extract primary object segments in videos from multiple object proposals. Although the above methods performs well on object detection, they are not lightweight algorithms. Thus, they cannot effectively handle OOI recognition toward mobile devices.

Many recent OOI detection algorithms are based on visual saliency prediction [13, 14]. It is generally accepted that OOIs are aroused by human perception and visual saliency can reflect the cognitive mechanism. Therefore, saliency prediction performance significantly influences these methods in detecting OOIs. Most of the existing saliency models are completely based on low-level visual features [15, 16]. However, some high-level semantic cues [17, 18] should also be integrated for saliency calculation [19]. Both biological and psychological studies [20] shown that, optimally fusing low-level and high-level visual features (including the location cue) can enhance saliency detection greatly. We employ the saliency detection by deploying the markov chain proposed by Jiang *et al.* [19]. One advantage of [19] is that both the appearance divergence and spatial distribution of foreground/background objects are integrated. It performs better on our multi-source location-aware dataset as compared with its competitors.

Many approaches have been proposed to predict where human perceives when viewing a scenery. The majority of the existing methods recognize OOIs based on the similarity of appearance features. Recently, the cheap availability of sensor-rich videos allows users to understand video semantics in a straightforward way [29]. For these different types of sensory metadata, we focus on the geo-attributes of sensor data throughout this paper. Associating GPS coordinates with digital photographs has becoming an active research domain over the last decade [30]. Toyama *et al.* [31] introduced a metadata-based image search algorithm and compiled a database which indexes photos using location and timestamp. Föckler *et al.* [32] developed a museum guidance system by utilizing camera-equipped mobile phones. Zheng *et al.* [1] constructed an efficient and effective landmark recognition engine, which organizes, models, and recognizes the landmarks on the world-scale. Gammeter *et al.* [33] introduced a fully functional and complete augmented reality system which can track both stationary

and mobile objects. By utilizing geo-sensor data, a number of object recognition tasks are implemented based on GPS coordinates.

3 Sensor-Social-Based OOI Recognition

Given an outdoor UGV, we detect its OOIs and annotate them by utilizing a variety of multimedia features. The key to recognize OOIs in outdoor UGVs is to fuse video content, sensor data, and social factors optimally. Thereafter, video sequences with annotated OOIs can be generated.

3.1 OOI Acquisition from UGVs

Saliency-Based Frames Selection. Obviously, semantics between sequential video frames are highly correlated. Existing summarization algorithms typically detect key frames to alleviate computational burden. These techniques are popularly used in video editing and compression. Notably, two factors should be emphasized in our method: the computational efficiency and representative OOI sequences. This means that the conventional key frames selection algorithms may not be able to preserve the diverse OOI sequences. In order to select representative frames at OOI-level, we propose a novel saliency-based frame selection. First, saliency map of each UGV frame is calculated based on Jiang *et al.*'s algorithm [19]. We employ [19] because it jointly describes the appearance divergence and spatial distribution of foreground/background objects. By adopting the Markov chain theory [21], the saliency detection is conducted rapidly. Let $Sal_{c,s}$ denote the calculated saliency map based on color and spatial distributions, e index the transient graph nodes, and y_w be the normalized weighted absorbed time vector, then the saliency map is simply obtained as:

$$Sal_{c,s}(e) = y_w(e), \quad i = 1, 2, \dots, t, \quad (1)$$

Afterward, region of OOI, denoted as $R_{bw}(\cdot)$, is binarized by an adaptive threshold τ_1 . The criterion of saliency-based frames selection is:

$$decision(i) = \begin{cases} 1 & \text{if } ||Th(i) - Th(i+1)|| > \tau_2 \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

where τ_2 denotes the divergence of saliency values between neighboring frames; $Th(i)$ is the salient area in frame i and $R(i)$ the salient object region; $R(i, i+1)$ denotes the salient area intersection between frames i and $i+1$.

Salient-Object-Assisted Track Learning Detection. To recognize OOIs, it is necessary to generate a number of OOI candidates extracted from UGVs. To balance the efficiency and accuracy, we employ the track learning detection framework proposed by Kalal *et al.* [22]. One advantage of [22] is that it can decompose a long-term tracking task into tracking, learning, and detection efficiently. Due to the complicated spatial context of a scenery, it is difficult to



Fig. 2. Tracking recognition for OOI in UGVs. The long box contains frames randomly selected from UGVs with a tracking box around OOI. The right column displays the extracted objects marked by the annotated names.

detect all the objects in UGVs accurately. To solve this problem, we propose a salient-object-assisted track learning detection. It combines object and saliency detection when processing each UGV frame. If the object detection fails, saliency detection will be conducted and assists the similarity measure between patches. Based on the assisting scheme, OOI acquisition is conducted for each frame. Thereafter, the new object modeling can be formulated as:

$$M = \{p_1^+, p_2^+, \dots, p_x^+, \dots, p_m^+, p_1^-, p_2^-, \dots, p_x^-, \dots, p_m^-\}, \tag{3}$$

where p^+ and p^- denote the foreground and background patches respectively; p_x^+ and p_x^- are the saliency patches of object and background respectively. Example OOI extracted from the UGVs are presented in Fig. 2. As can be seen, the proposed method not only detects those OOI accurately, but also tracks them within the UGV frames. The tracking is performed by localizing a bounding box centered around each detected OOI.

3.2 Classified Object Set Recommendation

Assisted by human interactions, social data has become an intellectual media conveying informative cues, *e.g.*, tagged images, video clips, and user comments. It is worth emphasizing that social data also contains lots of noises. Thus, effectively exploiting social data is a challenging task.

Sensor data is recorded by sensory modules embedded in mobile devices. In this work, we model sensor data of UGVs as a frame-related feature vector, which can be specified as:

$$S = \{(t_i, lat_i, long_i, accur_i, visD_i) | t_i \in T, (lat_i; long_i) \in G, accur_i \in A, visD_i \in V\}, \tag{4}$$

where T contains the capturing time of each frame; G is a set of GPS coordinates that describe the capturing location changes; A is a set of GPS location errors; V is a set of visible distances calculated by Arslan Ay *et al.* [28].

We constructed an image set containing candidate OOIs which are collected based on the category keywords. In particular, image retrieval is conducted by using different category names. Then, a collection of images are downloaded and classified from social networks. In order to compare at object-level, we calculate the saliency maps from these social images and then extract the salient objects as the OOIs adaptively. Saliency-based object classification minimizes the influence of noises resulted from the various backgrounds in social images. The classified OOIs from social images can be described as:

$$O_L = \{R_{bw}^1, R_{bw}^2, \dots, R_{bw}^n | n \in \mathcal{N}_L\}, \quad (5)$$

where O_L is the OOI set labeled by L ; and \mathcal{N} is the candidate category set.

3.3 OOI Description and Recognition

We adopt a spatial-pyramid-based [3] feature to represent an image, since it combines the advantages of standard feature extraction method. Spatial pyramid is a simple and efficient extension of an orderless bag-of-features image representation. It exhibits significantly improved performance on challenging scene categorization tasks. More specifically, local visual descriptors are quantized into a D -sized dictionary. Then, the spatial pyramid feature for the c -th class and n -th object is calculated as:

$$F_n^c = \{[f_1^1, f_2^1, \dots, f_t^1][f_1^2, f_2^2, \dots, f_s^2], \dots, [f_1^p, f_2^p, \dots, f_q^p]\}, \quad (6)$$

where p represents the pyramid level; t , s , and q denote the feature dimensionality of each pyramid level. Examples of the above spatial pyramid representation are presented on the left of Fig. 3. Noticeably, to maximally eliminate the negative effects caused by the complicated scenic backgrounds, we introduce a salient object based image filtering scheme, as elaborated in Fig. 4. We perform k-means clustering of two subsets from each classified object image set to constitute two class feature samples. Generally, people tend to capture images with similar salient objects for a category. Thus, we discriminate positive and negative samples using the intra-class variance. Additionally, we extract features with the same spatial pyramid architecture for OOI in UGVs, toward a consistent feature description. A few examples are shown on the right of Fig. 3.

As the last step, we recognize objects in UGVs using a similarity metric to compare objects extracted from an image set. The similarity is calculated between the mean features of OOIs extracted from UGVs and those of salient objects extracted from social images.

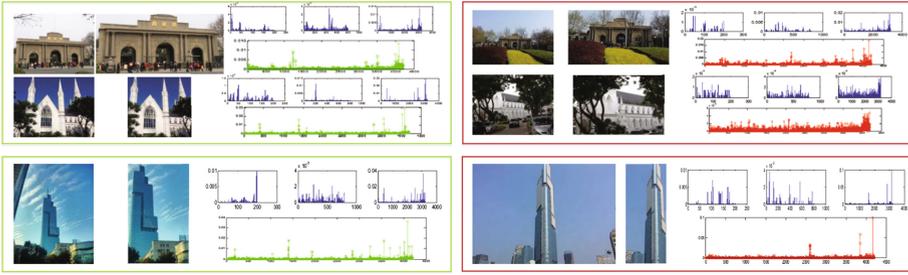


Fig. 3. Left: social objects and their three level spatial-pyramid features; right: OOIs of UGVs and their three level spatial-pyramid features

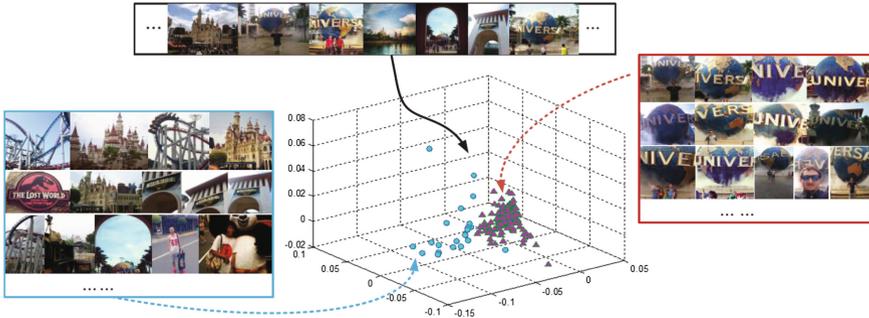


Fig. 4. Salient-object-based social images filtering

4 Experimental Results and Analysis

4.1 Dataset and Experimental Setup

The UGVs in our experiments consist of sensor-annotated videos captured from an Android/iOS device in Nanjing and Singapore. For the Nanjing dataset, five volunteers captured 676 UGVs using Sumsung Galaxy Note 3 and iPhone 6 respectively. Two resolutions 3840×2160 and 1920×1080 are employed. The Singapore dataset has 835 720×480 UGVs with complicated sceneries, *e.g.*, the Merlion, the Marina Bay, the Esplanade, and the Singapore Flyer.

Our approach is implemented on a desktop PC with an Intel i7-4770K CPU and 16 GB main memory. Java is adopted to parse the Json data collected from social servers. Matlab is used to implement the entire framework for its convenience in image/video processing. The location-based social network is implemented based on the Foursquare¹. The threshold τ_1 for salient region detection is adaptively calculated by OTSU. The frame selection threshold τ_2 is set to 0.2 and the spatial-pyramid level p is set to 3.

¹ <https://foursquare.com/>.

4.2 Experimental Results and Analysis

The experiments are designed to evaluate: (1) whether the proposed frames selection method is capable to preserve OOIs from UGV in order to accelerate computation, (2) users' satisfaction about the proposed tracking detection for OOIs in UGVs, and (3) the recognition accuracy.

Efficiency of Frame Selection. Figure 5 presents some results of the saliency-based frames selection. To better elaborate our proposed frames selection, we design a PSNR-loss histogram to measure the quality of the selected frames. The PSNR measure is popularly used to evaluate the reconstruction quality of the loss compression codec between images. In our experiment, we construct a PSNR-loss histogram $H = \{P_{12}, \dots, P_{ij}\}_L$ to calculate the PSNR difference between the i -th and j -th frames both in the original and the selected sequences. L is the frame number of the original UGVs. P denotes PSNR and is defined as:

$$P = 10 * \log_{10} \left(\frac{2^n - 1}{M_{SE}} \right), \quad (7)$$

where $M_{SE} = \sum_{x=1}^M \sum_{y=1}^M (f(x, y) - g(x, y)) / M * N$; n represents using n bits per sample, $f(x, y)$ and $g(x, y)$ are the grayscale of neighboring frames; $M \times N$ is the size of each frame.

The PSNR value of the selected sequences falls into the bin based on its frame number in the original videos. Therefore, the information loss can be compared at frame-to-frame level. Figure 6 presents the PSNR-loss histograms of one UGV, reflecting the information loss of the input UGV, and the selected sequences with $\tau_1 = 0.1$ and $\tau_2 = 0 : 2$. The red rectangle indicates that our method excludes the frames with very low information loss. It guarantees that the diversity changes of OOIs can be well preserved in the selected UGV frames.

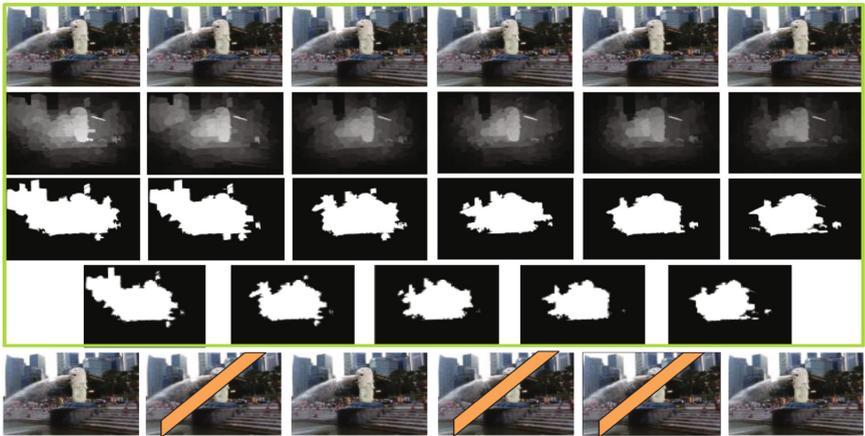


Fig. 5. Example frames of the saliency-based selection

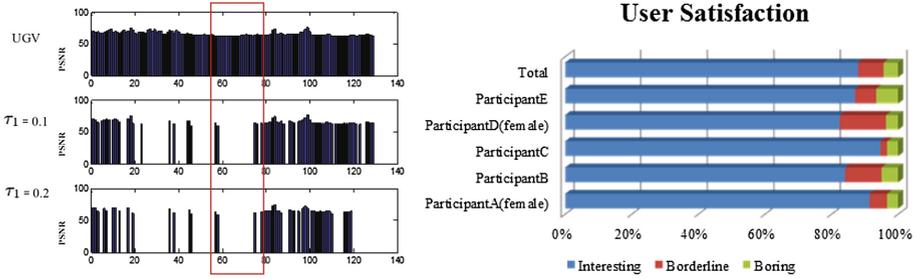


Fig. 6. Left: PSNR-loss histogram of the original UGV and the saliency-guided selected UGV frames; Right: user satisfaction with respect to the tracking detection

User Satisfaction. To evaluate the effectiveness of the proposed system, we invite five volunteers (two females and three males) whom are the photographers of the GeoVid² to participate our user study. As to the multi-source location-aware dataset in Singapore, we also invite them to rate the OOI tracking results generated by our system. Each volunteer rates the UGVs captured by himself/herself, and then randomly assigns one fifth part of the Singapore dataset. The participants are asked to choose from three feelings about the generated UGVs: “Interesting”, “Borderline”, and “Boring”, which reflect their opinions after viewing the UGVs with the OOI tracking box. Noticeably, the five volunteers label each video to determine whether the OOIs are recognized successfully or not. Afterward, we accumulate the feedbacks from the five volunteers, as shown on the right of Fig. 6. We also explore the reasons why they feel boring about some UGVs. We observe that the reason is that the wrong trackings occurred on several frames. The borderline opinion primarily due to the size of bounding boxes. Some of them cannot fully contain the OOIs.

Recognition Accuracy. Our multi-source location-aware dataset contains two cities: Nanjing and Singapore. We first calculate the recognition accuracies separately on the two cities. Afterward, we average them to obtain a final recognition accuracy of our designed system. All the experimental UGVs are captured by volunteers spanning a long time, and there is no ground truth presented. Therefore, all the UGVs are labeled by whether they can be correctly recognized during the user study. We employ the traditional method to label the dataset, “1” for the correct recognition while “0” for the mistaken one. In order to validate which distance measurement can achieve the best performance, we calculate 6 recognition accuracies. They are based on the Euclidean distance, the Seclidean distance, the Cosine distance, Histogram intersection, the Chebychev distance and the Hausdorff distance, respectively. The final recognition is calculated using the distance measure between feature vectors. All the accuracies on the two cities are presented in Fig. 7. Obviously, calculating the similarity by histogram intersection achieves the best

² www.geovid.org.

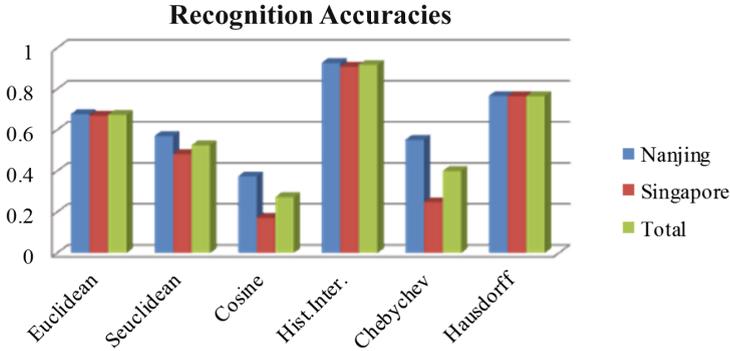


Fig. 7. OOI Recognition accuracies of UGVs extracted from Nanjing and Singapore

accuracy of 92.86 % on the Nanjing dataset, and 91.02 % on the Singapore dataset. Therefore, the average recognition accuracy of our system on the multi-sources dataset is 91.94 %.

5 Conclusions

OI recognition on UGVs is an important application in multimedia [24–27] and artificial intelligence [6, 7, 23, 34]. This paper proposes an automatic system to achieve OOI recognition on UGVs by leveraging sensor-social data. The key contributions of this paper can be summarized as follows. First, we propose a lightweight framework for recognizing OOIs in outdoor UGVs by leveraging geo-sensor data with the location-aware social networks. Second, we introduce a novel saliency-guided frame selection algorithm, which performs OOI recognition effectively and reduces the computational burden. Third, we compile a multi-source location-aware dataset containing two cities, Nanjing and Singapore, with three kinds of resolutions and two types of frame rates. Third, our system achieves an OOI recognition accuracy of 91.94 %, which demonstrated that it is useful in both mobile and desktop applications.

References

1. Zheng, Y.-T., Zhao, M., Song, Y., Adam, H.: Tour the world: building a web-scale landmark recognition engine. In: Proceedings of CVPR (2009)
2. Hao, J., Wang, G., Seo, B., Zimmermann, R.: Point of interest detection and visual distance estimation for sensor-rich video. *IEEE T-MM* **16**(7), 1929–1941 (2014)
3. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of CVPR (2006)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of CVPR (2005)

5. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D.: Object detection with discriminatively trained part-based models. *IEEE T-PAMI* **32**(9), 1627–1645 (2010)
6. Yang, K., Wang, M., Hua, X.-S., Yan, S., Zhang, H.-J.: Assemble new object detector with few examples. *IEEE T-IP* **20**(12), 3341–3349 (2011)
7. Wang, M., Hua, X.-S., Hong, R., Tang, J., Qi, G.-J., Song, Y.: Unified video annotation via multi-graph learning. *IEEE T-CSVT* **19**(5), 733–746 (2009)
8. Harzallah, H., Jurie, F., Schmid, C.: Combining efficient object localization and image classification. In: *Proceedings of ICCV* (2009)
9. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: *Proceedings of ICCV* (2009)
10. Cimbis, R.G., Verbeek, J.J., Schmid, C.: Segmentation driven object detection with fisher vectors. In: *Proceedings of ICCV* (2013)
11. Kim, S., Park, S., Kim, M.: Central object extraction for object-based image retrieval. In: Bakker, E.M., Lew, M., Huang, T.S., Sebe, N., Zhou, X.S. (eds.) *CIVR 2003*. LNCS, vol. 2728, pp. 39–49. Springer, Heidelberg (2003)
12. Zhang, D., Javed, O., Shah, M.: Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In: *Proceedings of CIVR* (2013)
13. Jiang, H., Wang, J., Yuan, Z., Liu, T., Zheng, N.: Automatic salient object segmentation based on context and shape prior. In: *Proceedings of BMVC* (2011)
14. Khuwuthyakorn, P., Robles-Kelly, A., Zhou, J.: Object of interest detection by saliency learning. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010, Part II*. LNCS, vol. 6312, pp. 636–649. Springer, Heidelberg (2010)
15. Margolin, R., Tal, A., Zelnik-Manor, L.: What makes a patch distinct? In: *Proceedings of CVPR* (2013)
16. Rosin, P.L.: A simple method for detecting salient regions. *Pattern Recogn.* **42**(11), 2363–2371 (2009)
17. Jia, Y., Han, M.: Category-independent object-level saliency detection. In: *Proceedings of ICCV* (2013)
18. Jiang, P., Ling, H., Yu, J., Peng, J.: Salient region detection by UFO: uniqueness, focusness and objectness. In: *Proceedings of ICCV* (2013)
19. Navalpakkam, V., Itti, L.: Modeling the influence of task on attention. *Vision. Res.* **45**(2), 205–231 (2005)
20. Borji, A.: Boosting bottom-up and top-down visual features for saliency estimation. In: *Proceedings of CVPR* (2012)
21. Bolch, G., Greiner, S., de Meer, H., Trivedi, K.S.: *Queueing Networks and Markov Chains*, 2nd edn. John Wiley, Hoboken (2006)
22. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. *IEEE T-PAMI* **34**(7), 1409–1422 (2012)
23. Zhang, L., Bian, W., Song, M., Tao, D., Liu, X.: Integrating local features into discriminative graphlets for scene classification. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) *ICONIP 2011, Part III*. LNCS, vol. 7064, pp. 657–666. Springer, Heidelberg (2011)
24. Zhang, L., Song, M., Sun, L., Liu, X., Wang, Y., Tao, D., Bu, J., Chen, C.: Spatial graphlet matching kernel for recognizing aerial image categories. In: *ICPR* (2012)
25. Zhang, L., Gao, Y., Zimmermann, R., Tian, Q., Li, X.: Fusion of multichannel local and global structural cues for photo aesthetics evaluation. *IEEE T-IP* **23**(3), 1419–1429 (2014)
26. Zhang, L., Wang, M., Nie, L., Hong, L., Rui, Y., Tian, Q.: Retargeting semantically-rich photos. *IEEE T-MM* **17**(9), 1538–1549 (2015)

27. Zhang, L., Gao, Y., Hong, R., Hu, Y., Ji, R., Dai, Q.: Probabilistic skimlets fusion for summarizing multiple consumer landmark videos. *IEEE T-MM* **17**(1), 40–49 (2015)
28. Ay, S.A., Zimmermann, R., Kim, S.H.: Viewable scene modeling for geospatial video search. In: *ACM Multimedia* (2008)
29. Zheng, Y.-T., Zha, Z.-J., Chua, T.-S.: Research and applications on georeferenced multimedia. *Multimedia Tools Appl.* **51**(1), 77–98 (2011)
30. Rodden, K., Wood, K.R.: How do people manage their digital photographs? In: *ACM SIGCHI* (2003)
31. Kentaro, T., Logan, R., Roseway, A., Anandan, P.: Geographic location tags on digital images. In: *ACM Multimedia* (2003)
32. Föckler, P., Zeidler, T., Brombach, B., Bruns, E., Bimber, O.: PhoneGuide: museum guidance supported by on-device object recognition on mobile phones. In: *Proceedings of Mobile and Ubiquitous Multimedia* (2005)
33. Gammeter, S., Gassmann, A., Bossard, L.: Server-side object recognition and client-side object tracking for mobile augmented reality. In: *Proceedings of CVPR* (2010)
34. Wang, M., Gao, Y., Ke, L., Rui, Y.: View-based discriminative probabilistic modeling for 3D object retrieval and recognition. *IEEE T-IP* **22**(4), 1395–1407 (2013)