https://doi.org/10.2352/ISSN.2470-1173.2020.10.IPAS-313

Multiscale Convolutional Descriptor Aggregation for Visual Place Recognition

Raffaele Imbriaco, Egor Bondarev and Peter H.N. de With;
Eindhoven University of Technology, Eindhoven, The Netherlands

Abstract

Visual place recognition using query and database images from different sources remains a challenging task in computer vision. Our method exploits global descriptors for efficient image matching and local descriptors for geometric verification. We present a novel, multi-scale aggregation method for local convolutional descriptors, using memory vector construction for efficient aggregation. The method enables to find preliminary set of image candidate matches and remove visually similar but erroneous candidates. We deploy the multi-scale aggregation for visual place recognition on 3 large-scale datasets. We obtain a Recall@10 larger than 94% for the Pittsburgh dataset, outperforming other popular convolutional descriptors used in image retrieval and place recognition. Additionally, we provide a comparison for these descriptors on a more challenging dataset containing query and database images obtained from different sources, achieving over 77% Recall@10.

Introduction

Accurate localization based on visual information is important for self-navigating devices and is a widely researched topic in computer vision and robotics. This problem is usually presented as an image retrieval task. Given a large collection of geotagged images and a query image, the geographically closest image, or a nearby image set should be retrieved. The recent development of convolutional neural networks (CNNs) has led to advances in the field of visual place recognition. Methods that traditionally used handcrafted features, have changed into techniques based on learned features extracted from CNNs [7, 8, 5].

The ideal place recognition system should be robust and invariant to viewpoint and appearance. Global descriptors generated by CNNs are useful for the image retrieval and place recognition tasks, since they are robust with respect to viewpoint and appearance. An advantage of convolutional descriptors is that these features can be tuned to any specific dataset (e.g. landmarks or cities) by retraining a network. However, global descriptors also include information about objects that are irrelevant to the scene. These objects are ubiquitous in urban scenes and provide no information about the location depicted in the image (e.g. pedestrians, vehicles).

Recent approaches extract regional [16] or attentive, deep local features (DELF) [10], to improve performance and reduce the impact of scene clutter. Some of these approaches generate a large number of multi-scale local descriptors. These descriptors can be used to increase the system's resilience to appearance and viewpoint changes. Furthermore, they may also enable geometric verification of the selected matches using techniques such as Hamming Embedding [17] or Random Sample Consensus [18]. The trade-off is that a single image can have hundreds or thousands of such descriptors, requiring descriptor aggregation to facilitate efficient searching in databases. In our work, we compare the performance of several different convolutional descriptors on three large-scale visual place-recognition datasets. Our contributions are threefold. First, we utilize a novel, memory vector-based aggregation algorithm to produce compact, multi-scale image representations. Second, we study how different parameters, such as descriptor dimensionality and aggregation modality impact the retrieval performance of the proposed system. Third, we compare the performance of the aggregated descriptor against other convolutional descriptors, which are commonly used for visual place recognition and image retrieval.

Until recent years, conventional image retrieval has depended on handengineered features and has been inspired by advances in other information fetching tasks, like document retrieval. Some methods and structures from document retrieval have been successfully adapted for image retrieval, such as, Bag-of-Words [4] and inverted file indices. However, recent advances in CNNs have resulted in new types of descriptors, generated from convolutional features.

Image retrieval with convolutional descriptors

Convolutional descriptors. Babenko *et al.* demonstrate in [19] that the high-level global features learned by CNNs for e.g. classification tasks do also apply to image retrieval. They show that the high-dimensional vectors generated by fully connected layers can be used to encode and retrieve images of landmarks based on visual similarity. Descriptors extracted from mid-level convolutional layers have better performance than those extracted from fully connected layers. However, these are high-dimensional tensors and require aggregation in order to produce compact descriptors. A common approach to vectorize the activation maps, is to perform a pooling operation on the extracted tensor. These operations can be max-pooling [21] or sum-pooling [1]. Pooling operations produce a single descriptor that encodes the visual content of an image. The similarity between a pair of images can then be computed as a simple metric, such as the Euclidean distance. However, geometric verification (e.g. using RANSAC) is not possible when using global convolutional descriptors. Recently, a new architecture for large-scale image retrieval has been presented. Deep Local Features (DELF) [10] is a CNN that uses visual attention to identify and extract the relevant local features of an image.

Datasets and testing. The performance of many of these descriptors is usually measured on relatively small datasets such as the Paris and Oxford Buildings datasets [22][23]. However, the above mentioned DELF is an exception to this. The previous two

datasets contain 6,412 and 5,062 images, which are generally extended with 100K unrelated images to complicate the image retrieval task. More recently, the Google-Landmarks dataset [10] has been made publicly available. This dataset contains 1M images of 13K different landmarks with 100K query images. Nevertheless, it is hard to estimate if these descriptors can help to address the challenges specific for large-scale place recognition. The problem is the difference between the visual content in common image retrieval datasets and those used for place recognition. Place recognition algorithms should also be able to recognize the location of areas with less informative features. This is further complicated by areas containing repetitive structures (such as residential areas), lacking uniquely identifying features (e.g. densely forested areas or highways); and by changes in appearance due to weather and scene clutter. Due to the unique characteristics of the visual place recognition datasets, it is difficult to ascertain whether good performance in image retrieval tasks translates to good performance in place recognition. We discuss some place recognition algorithms below.

Visual place recognition.

We propose to consider visual place recognition as a branch of image retrieval. Given an image depicting a real-world location, the system should identify visually similar locations in a large geotagged database of images. This task is complicated by variations in viewpoint and appearance that may exist between the database and query images. Sunderhauf *et al.* solve the visual place recognition problem in [5] by identifying landmarks in images and extracting convolutional descriptors. In their work, landmarks are salient objects detected automatically using the Edge Boxes [24] algorithm. Per detected object, a convolutional descriptor is generated. These descriptors are later used to match images of the same location. A drawback of this approach is the generation of multiple, possibly repeated, descriptor proposals per image. This requires a quadratic number of matching operations, thereby hindering its practical use for city-scale datasets. Arandjelovic *et al.* [13] learn the convolutional features and build the aggregated VLAD descriptors in an end-to-end fashion. This approach results in an excellent performance for place recognition. An alternative to NetVLAD descriptors is proposed in [8]. This work deploys memory vectors [9] to aggregate both the dataset and the query descriptors. Memory vectors reduce the number of necessary operations to find a match without requiring the computation of a codebook.

Our hybrid system approach. A common factor across most of the methods described above is that the geometric verification requires computation of additional local features. Having a single global descriptor is advantageous for performance-related reasons (fewer comparisons necessary per database image, more efficient use of memory). However, it prevents the use of methods like RANSAC, to remove similar but incorrect image matches from the retrieval sets. Our method combines the compactness of the global representations, while also retaining the ability to perform geometric verification. This becomes useful when the query and database images are acquired from different sources.

Methods

Image preprocessing

Street-level images are commonly available on the Internet as high-resolution panoramas. Working directly with these image types is computationally expensive because of their dimensions. Furthermore, common CNN architectures assume that their inputs are planar images. In this work, we decompose each panoramic image into a number of overlapping planar views. We remove the bottom area (20%) of the image, since this region usually depicts acquisition vehicles. Once the images have been preprocessed, we extract convolutional descriptors for each individual image.

Descriptor extraction

We extract either a single (global) or several (local) convolutional descriptors. We use global descriptors that are well known in image retrieval literature, such as MAC [26], R-MAC [16] and NetVLAD [13]. Local descriptors are extracted using the pretrained DELF [10] model. Each local descriptor also contains the attention score. This last metric represents the relevance measure that the network assigns to a particular descriptor. However, direct local descriptor matching would be computationally expensive. Thus, the local descriptors should preferably be aggregated into a single global representation of the image.

Descriptor aggregation

We use memory vectors for the aggregation process and aggregate descriptors based on their scale. Memory vectors [9] are an efficient descriptor aggregation technique. Given an image $\mathbf{I}$, local descriptors are extracted at various different scales. The matrix $\mathbf{X} = [\mathbf{x_1}, ...\mathbf{x_n}]$ is the feature-space representation of $\mathbf{I}$, with $\mathbf{x_n}$ the nth convolutional feature vector. The descriptor matrix $\mathbf{X}$ is of size $d \times n$, where n is the number of local descriptors extracted from $\mathbf{I}$, and d is the dimensionality of each descriptor. It is evident that for large values of n and d image matching is computationally costly, as it involves operations between large matrices. Furthermore, in this matrix representation the ordering of the feature vectors can influence the output of the similarity operation. To solve this problem, we generate a single representation per scale. Such a representation simultaneously addresses the problems of computational complexity and feature ordering. At each scale s, all descriptors belonging to that scale are aggregated using the *p-inv vector* formulation constructed as

$$\mathbf{m}(\mathbf{X_s}) = \mathbf{X_s}(\mathbf{X_s}^T\mathbf{X_s})^{-1}\mathbf{1_n}, \tag{1}$$

where $\mathbf{1_n}$ is an n-dimensional unity vector. The outcome of this operation is a single representation for all local descriptors at scale s. This procedure is repeated at each scale, and the resulting representations are concatenated into a single $s \times d$ matrix $\mathbf{X_{agg}}$. Another representation can be generated (*sum vector*). However, the *p-inv vector* provides better performance according to literature. An example is depicted in Figure 2.

Image matching

Finally, we compute the similarity metric for all descriptors as their inner product with the query descriptor. Each database image is ranked based on this metric. We then select the best N candidates as the matching set of the query. When using aggregated local descriptors, we first compute a similarity score per scale.

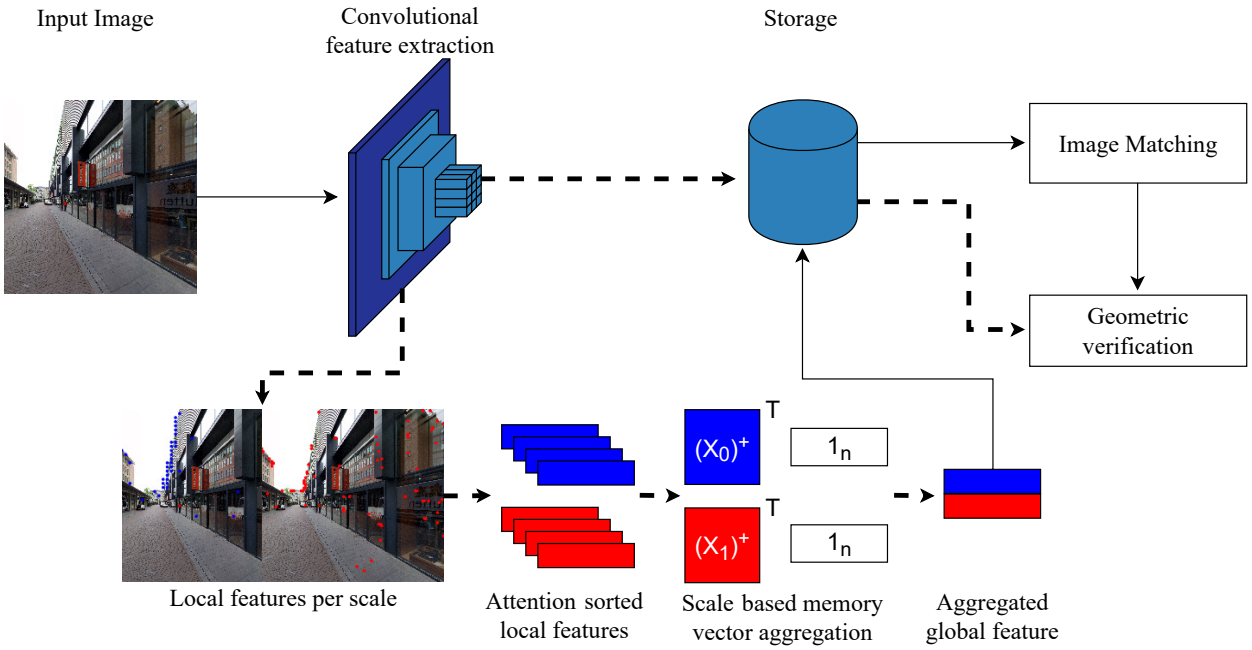


Figure 1: Visual place recognition pipeline. Solid arrows represent the flow of input data and global features. Dashed arrows represent the flow local features. Best viewed in color.

Figure 2: Local convolutional descriptors are generated at various scales. Each scale is aggregated into a single vector per scale forming a descriptor matrix. Scales are color-coded.

We perform a comparison either within the same scale value or we compare with adjacent scale values. For this, the similarity matrix is multiplied by either the identity matrix or a tri-diagonal matrix with non-zero elements equal to unity. This is done to prevent matching of features across large-scale variations. The scores are summed and provide an initial ranking of the database images. Finally, we perform geometric verification of the best 50 matches using the RANSAC algorithm. The candidates are then re-ranked based on one of two possible metrics, to evaluate their suitability. The metrics are (1) the number of inliers or (2) the total attention score of the inliers. Geometric verification allows us to remove erroneous candidates in the matching set.

Experiments

Data and metrics

We evaluate the performance of our visual place recognition system on three datasets. The first is the Pittsburgh dataset [27]. Each panoramic image of this dataset has been split into 24 overlapping planar images. We use the test set that is split into 83,952 database images and 8,280 query images. These images contain positioning information in the form of UTM coordinates. The second dataset has been provided by a Dutch company[1] and consists of 39,333 panoramic images of the city of Eindhoven and their corresponding coordinates in the Dutch *Rijksdriehoek* coordinate system. Each panoramic image is split into 8 overlapping planar images. Unlike the Pittsburgh dataset, images from the Eindhoven dataset are used exclusively as database images. The query images are obtained from a different source. These are street-level panoramas freely available on the Internet. The query dataset consists of 250 panoramic images (2000 planar query images), depicting locations available in the Eindhoven dataset. These images present significant viewpoint and appearance variations, since the images are acquired under different conditions for each source (e.g. day time, weather, road position of the acquisition vehicle) and provide a more challenging scenario than the one in the Pittsburgh dataset. The third is the Tokyo 24/7 dataset [12]. This dataset contains 75,984 planar database images and 1,125 query images acquired at 125 locations and at various times of the day. The images possess both GPS and UTM positioning information.

We evaluate the performance of the system using Recall@N. This is a common metric used in visual place recognition literature [13]. It considers a set of retrieved image candidates correct,

[1]Cyclomedia is a Dutch company that sells annually recorded pictures to government and civil engineering agencies.

Table 1: Average Recall@N for the preliminary aggregation modality and dimensionality experiments.

	Avg. recall@N for naive aggregation				
Dim.	**Top 1**	**Top 5**	**Top 10**	**Top 15**	**Top 20**
128	0.10	2.40	5.11	5.11	5.11
256	0.10	0.75	1.80	1.80	1.80
512	0.00	0.00	1.60	1.60	1.60
1024	0.00	0.00	0.80	0.80	0.80
	Avg. recall@N for scale aggregation				
128	35.09	46.70	50.80	53.65	55.71
256	43.84	57.16	61.71	64.72	66.42
512	47.30	60.21	65.57	68.02	69.82
1024	48.60	62.26	67.97	70.47	72.02

if at least one of the candidates is found within a certain distance of the query position. We consider a proposal a good match if it is in close proximity, i.e. within 25 meters from the actual query location.

Impact of aggregated descriptor dimensionality

Prior to comparing the proposed aggregated descriptor against those commonly found in image retrieval and place recognition literature, we first study possibilities on aggregation modalities and optimal descriptor dimensionality. We perform these preliminary experiments on a subset of 10,000 images from the Eindhoven dataset with their corresponding queries. We also study the retrieval performance of naive and scale-based descriptor aggregation. Naive aggregation produces a single memory vector from every local descriptor in the image. Meanwhile, scale-based aggregation generates a memory vector using all descriptors corresponding to a single scale. We use the same scales as in [10]. This experiment is repeated for various descriptor sizes. Furthermore, PCA is used to reduce the original, local descriptor size and repeat the retrieval experiments. Here, the largest descriptor size is 1024 and the smallest descriptor size 128. Geometric verification (GV) is performed on all experiments.

As can be observed from Table 1, naive aggregation produces poor retrieval results. This indicates that the aggregation of too many descriptors into a single memory vector significantly reduces its discriminatory strength. This is verified by the results obtained by our scale-based aggregation. By reducing the number of descriptors aggregated, we are able to significantly improve the retrieval performance. Even when using the smallest descriptor size, scale-based aggregation greatly outperforms naive aggregation. We observe that increasing the descriptor size yields gains in recall. This can be expected, since larger descriptors encode more information. However, this approach provides diminishing returns, as the difference between the two largest descriptor sizes is approx. 2%.

Another parameter that has an impact on the overall system performance is the metric used in the re-ranking step. We repeat the previous experiment using only scale-based aggregation and test three different metrics for re-ranking of candidate matches. The first metric is the similarity score, which is equivalent to the initial set of candidate matches, and is serving as the baseline for this experiment. In this case, no geometric verification is used. The second metric is the number of descriptor inliers found between the query and candidate images after a descriptor matching procedure (RANSAC). The third metric is the sum of the attention scores of the descriptor inliers. As mentioned previously, DELF descriptors are ranked based on their attention. This can be considered as a measure of the saliency of a particular descriptor. For this last metric, we compute the total attention score of the matched descriptors. The results of this experiment are presented in Table 2. From this, we observe that re-ranking provides a performance boost in every case. This is most noticeable for low values of N, where re-ranking provides gains of up to 20%. However, even for large values of N, re-ranking consistently outperforms the baseline. This is observed regardless of the descriptor dimensionality. With respect to the other two metrics, we have found that the descriptor inlier count metric slightly outperforms the sum of inlier attention metric. However, this gap rapidly vanishes as the value of N grows. This unexpected behavior is caused by matches with few, but very salient features, e.g. church steeples in the background of the images. These salient features are correctly matched in both query and database images, but may be detectable beyond the 25 meter radius used for determining the correctness of a candidate match. An example of this behavior is illustrated in Figure 3.

Table 2: Average Recall@N for different re-ranking metrics.

	Comparison of average Recall@N for three different re-ranking metrics				
Dim.	**Metric**	**Top 1**	**Top 5**	**Top 10**	**Top 15**
128	Baseline	15.72	30.78	38.74	43.24
	Inlier	37.23	46.70	51.60	54.00
	Attention	35.09	46.70	50.80	53.65
256	Baseline	27.13	44.14	52.70	56.71
	Inlier	46.65	58.21	62.41	64.91
	Attention	43.84	57.16	61.71	64.71
512	Baseline	32.43	49.90	57.66	61.91
	Inlier	48.95	61.66	65.77	68.32
	Attention	47.30	60.21	65.57	68.02
1024	Baseline	34.53	52.95	60.46	63.66
	Inlier	51.15	64.16	68.47	70.72
	Attention	48.60	62.26	67.97	70.47

Comparison of convolutional descriptors

In our experiments, we attempt to retrieve images neighbouring the location at which a query image was captured. We have generated convolutional descriptors using MAC, R-MAC, NetVLAD and DELF for the database and query images. Then, we identify potential matches by following the procedures described in the previous section and report the mean Recall@N on both datasets. When using DELF descriptors, they are aggregated by scale, and spatial verification is used for re-ranking. Re-ranking is done with the best 50 matches to reduce computational expense. The metric used for re-ranking of matches is the number of inliers.

Figure 4 depicts the recall curves of the different convolutional descriptors for the Pittsburgh, Eindhoven and Tokyo datasets, respectively. In the simpler case presented by the Pittsburgh dataset, we observe that the descriptors specialized in place recognition, DELF and NetVLAD, significantly outperform the more generic MAC and R-MAC descriptors. DELF, together with our scale-based aggregation, slightly outperforms NetVLAD. However, the NetVLAD descriptor has been trained on images of Pittsburgh, while DELF is trained only on land-

Figure 3: Example of retrieval failure due to salient features. The first image (from left to right) depicts the query location. The remaining five images show the highest-scoring matches. Out of these, only the first match is correct. The remaining matches depict locations further than 25 meters. All matches contain the same church steeple (circled in red) as a salient feature.

mark images of many locations around the world. This generalization ability becomes evident in the case of the Eindhoven dataset, where scale-aggregated DELF descriptors outperform all other convolutional descriptors. The Eindhoven dataset presents a more challenging case, since query and database images originate from different sources. This introduces appearance variations, due to weather or scene clutter and viewpoint variations due to the different position of the acquisition vehicle. In this more complex case, DELF descriptors are still capable of obtaining good performance with a Recall@20 over 80% and without requiring additional training. This robustness in performance is resulting from our hybrid representation, which retains the compact global representation for efficient image retrieval and the local descriptors for geometric verification. Nevertheless, in the last and most challenging dataset DELF outperforms NetVLAD in Recall@1. As N increases, the retrieval performance difference is severely reduced. This occurs because of the significant appearance variations of the query images during the evening and night.

As an additional experiment, we perform image retrieval with DELF descriptors and relax the matching criteria. We use a tri-diagonal matrix instead of a diagonal matrix for computation of the similarity score, thereby allowing neighbouring scales to be matched with each other. This produces a small increase in recall of roughly 1.3%, but increases the query time by 10% (from roughly 36 to 40 seconds per query). If no restrictions are placed during matching, the retrieved candidates are unusable. High similarity scores between disparate scales lead to increased incorrect retrieval. This is common in cases were vegetation is present.

Descriptors at very different scales (e.g. smallest and largest) may still be matched together producing matching sets with few correct candidates. Commonly, spatial verification would handle these cases and remove incorrect candidates. However, since we check only the best 50 matches, it is not possible to remove all erroneous candidates.

Conclusion

This work has presented a novel multi-scale representation to solve the visual place recognition problem. Our system uses state-of-the-art local convolutional descriptors and aggregates them into a compact, yet highly descriptive matrix. Using this representation, we obtain a Recall@10 of approximately 95% and 78% for the Pittsburgh and Eindhoven datasets, respectively. These results are obtained without tuning the descriptors to the datasets, thereby demonstrating the generalization capabilities of the aggregated descriptor. We have studied the different parameters that impact the performance of the system such as the aggregation method, the descriptor dimensionality and the re-ranking

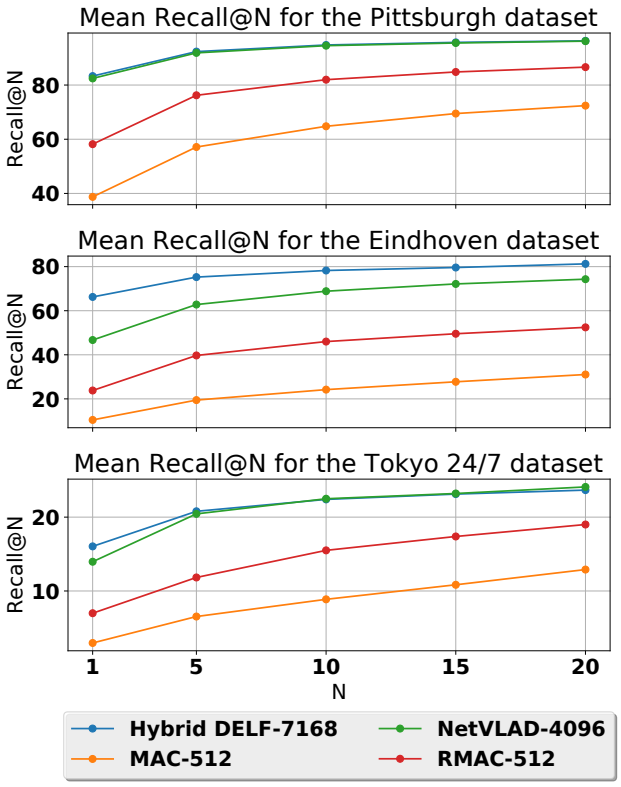


Figure 4: Mean Recall@N for various convolutional descriptors on three visual place recognition datasets.

metric. From these parameter studies we conclude that 1024-dimensional vectors using scale-based aggregation, provide the best retrieval results. Furthermore, we have observed that geometric verification improves recall by 16%. Both studied re-ranking metrics, inlier count and sum of inlier attention, behave comparably. When our system is compared against other convolutional descriptors used in image retrieval and place recognition, our aggregated DELF descriptor consistently outperforms all considered descriptors. In both datasets, the aggregated descriptor obtains a higher Recall@N.

References

[1] Babenko, A., & Lempitsky, V.S. (2015). Aggregating Local Deep Features for Image Retrieval. 2015 IEEE International Conference on Computer Vision (ICCV), 1269-1277.

[2] Chen, J., & Little, J.J. (2017). Where should cameras look at soccer games: Improving smoothness using the overlapped hidden Markov

model. Computer Vision and Image Understanding, 159, 59-73.
[3] Bay, Herbert, Tinne Tuytelaars and Luc Van Gool. SURF: Speeded Up Robust Features. ECCV (2006).
[4] Sivic, Josef and Andrew Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. ICCV (2003).
[5] Snderhauf, Niko, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft and Michael Milford. Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free. Robotics: Science and Systems (2015).
[6] Panphattarasap, Pilailuck, and Andrew Calway. "Visual place recognition using landmark distribution descriptors." In Asian Conference on Computer Vision, pp. 487-502. Springer, Cham, 2016.
[7] Hou, Yi, Hong Zhang and Shilin Zhou. Evaluation of Object Proposals and ConvNet Features for Landmark-based Visual Place Recognition. Journal of Intelligent & Robotic Systems 92 (2018): 505-520.
[8] Iscen, Ahmet, Giorgos Tolias, Yannis Avrithis, Teddy Furon, and Ondej Chum. "Panorama to panorama matching for location recognition." In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pp. 392-396. ACM, 2017.
[9] Iscen, Ahmet, Teddy Furon, Vincent Gripon, Michael Rabbat, and Herv Jgou. "Memory vectors for similarity search in high-dimensional spaces." IEEE Transactions on Big Data 4, no. 1 (2017): 65-77.
[10] Noh, Hyeonwoo, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. "Large-scale image retrieval with attentive deep local features." In Proceedings of the IEEE International Conference on Computer Vision, pp. 3456-3465. 2017.
[11] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
[12] Torii, Akihiko, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. "24/7 place recognition by view synthesis." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1808-1817. 2015.
[13] Arandjelovic, Relja, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. "NetVLAD: CNN architecture for weakly supervised place recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5297-5307. 2016.
[14] Rokach, Lior. "A survey of clustering algorithms." In Data mining and knowledge discovery handbook, pp. 269-298. Springer, Boston, MA, 2009.
[15] Piasco, Nathan, Dsir Sidib, Cdric Demonceaux, and Valrie Gouet-Brunet. "A survey on visual-based localization: On the benefit of heterogeneous data." Pattern Recognition 74 (2018): 90-109.
[16] Tolias, Giorgos, Ronan Sicre, and Herv Jgou. "Particular object retrieval with integral max-pooling of CNN activations." arXiv preprint arXiv:1511.05879 (2015).
[17] Jgou, Herv, Matthijs Douze, and Cordelia Schmid. "Improving bag-of-features for large scale image search." International journal of computer vision 87, no. 3 (2010): 316-336.
[18] Fischler, Martin A., and Robert C. Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography." Communications of the ACM 24, no. 6 (1981): 381-395.
[19] Babenko, Artem, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. "Neural codes for image retrieval." In European conference on computer vision, pp. 584-599. Springer, Cham, 2014.
[20] Sivic, Josef, and Andrew Zisserman. "Efficient visual search of videos cast as text retrieval." IEEE transactions on pattern analysis and machine intelligence 31, no. 4 (2008): 591-606.
[21] Azizpour, Hossein, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. "From generic to specific deep representations for visual recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 36-45. 2015.
[22] Philbin, James, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. "Object retrieval with large vocabularies and fast spatial matching." In 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8. IEEE, 2007.
[23] Philbin, James, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. "Lost in quantization: Improving particular object retrieval in large scale image databases." In 2008 IEEE conference on computer vision and pattern recognition, pp. 1-8. IEEE, 2008.
[24] Zitnick, C. Lawrence, and Piotr Dollr. "Edge boxes: Locating object proposals from edges." In European conference on computer vision, pp. 391-405. Springer, Cham, 2014.
[25] Jgou, Herv, Matthijs Douze, and Cordelia Schmid. "On the burstiness of visual elements." In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1169-1176. IEEE, 2009.
[26] Razavian, Ali S., Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. "Visual instance retrieval with deep convolutional networks." ITE Transactions on Media Technology and Applications 4, no. 3 (2016): 251-258.
[27] Torii, Akihiko, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. "Visual place recognition with repetitive structures." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 883-890. 2013.

Author Biography

Raffaele Imbriaco is a PhD at the Electrical Engineering faculty of Eindhoven University of Technology (TU/e, the Netherlands). He obtained his MSc from TU/e after concluding his research project on x-ray imaging at Philips Healthcare. His research interests include deep learning, image retrieval and visual place recognition. He is one of the researchers involved in the PS-CRIMSON project.

Egor Bondarev obtained his PhD degree in the Computer Science Department at TU/e, in research on performance predictions of real-time component-based systems on multiprocessor architectures. He is an Assistant Professor at the Video Coding and Architectures group, TU/e, focusing on sensor fusion, smart surveillance and 3D reconstruction. He has written and co-authored over 50 publications on real-time computer vision and image/3D processing algorithms. He is involved in large international surveillance projects like APPS and PS-CRIMSON.

Peter H.N. de With is Full Professor of the Video Coding and Architectures group in the Department of Electrical Engineering at Eindhoven University of Technology. He worked at various companies and was active as senior system architect, VP video technology, and business consultant. He is an IEEE Fellow, has (co-)authored over 400 papers on video coding, analysis, architectures, and 3D processing and has received multiple papers awards. He is a program committee member of the IEEE CES and ICIP and holds some 30 patents.

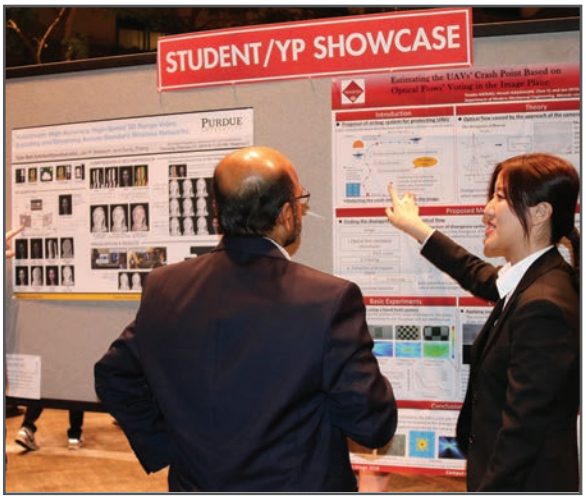
JOIN US AT THE NEXT EI!

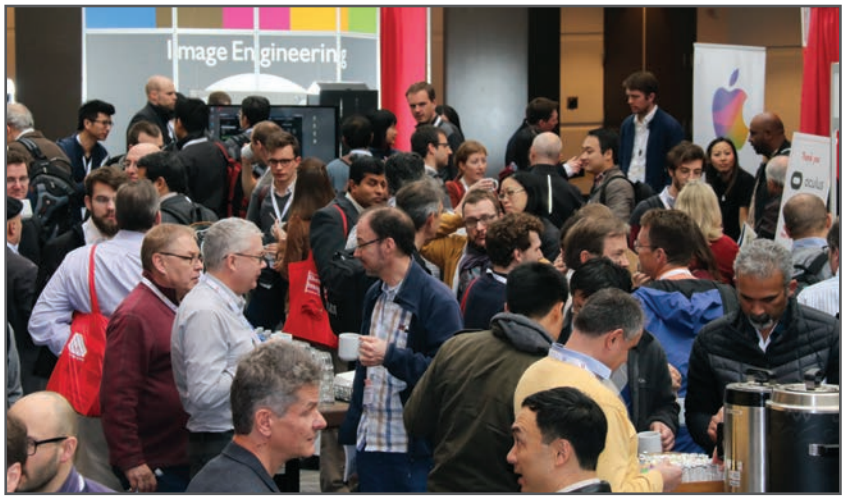
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org