




Data-Driven Lightweight Interest Point Selection for Large-Scale Visual Search

Feng Gao, Xinfeng Zhang , *Member, IEEE*, Yicheng Huang, Yong Luo ,
Xiaoming Li, *Senior Member, IEEE*, and Ling-Yu Duan , *Member, IEEE*

I. INTRODUCTION

Abstract—With the explosive increase of images and videos, visual analysis has become an essential technique in dealing with the big visual data, which utilizes the visual feature descriptors to search or recognize the images or frames with target objects or events. Subject to the constraints of resources (e.g., memory, bandwidth, storage, etc.), interest point selection is crucial to generate robust compact descriptors for high-efficiency visual analysis by selecting and aggregating the most discriminative local feature descriptors, which has been demonstrated in the state-of-the-art low bit rate visual search works. In this paper, we propose a data-driven lightweight interest point selection approach to significantly improve the performance of visual search, while ameliorating the efficiency of extracting feature descriptors. Comprehensive experimental results over benchmarks have shown that the proposed interest point selection algorithm has significantly improved image matching and retrieval performance in the completed MPEG Compact Descriptors for Visual Search (CDVS) standard as well as the emerging MPEG Compact Descriptors for Video Analytics (CDVA) standard, say 20% mAP gain by data-driven selection against random selection of interest points. In particular, the presented data-driven interest point selection has been adopted by MPEG-CDVS and MPEG-CDVA as a normative technique to improve the aggregation of handcrafted features, which has contributed to the combination of handcrafted features and deep learning (CNN) features as well.

Index Terms—Visual search, interest point selection, compact descriptors, regression, classification, feature selection, MPEG-CDVS, MPEG-CDVA.

Manuscript received December 4, 2017; revised February 19, 2018; accepted March 6, 2018. Date of publication March 21, 2018; date of current version September 18, 2018. This work was supported in part by the National Natural Science Foundation of China under Grants 61661146005, U1611461, and 61390515; in part by the National Key Research and Development Program of China under Grant 2016YFB1001501; and in part by the PKU-NTU Joint Research Institute (JRI) sponsored by a donation from the Ng Teng Fong Charitable Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jian Zhang. (*Corresponding author: Ling-Yu Duan.*)

F. Gao, Y. Huang, and L.-Y. Duan are with the School of Electronics Engineering and Computer Science, Institute of Digital Media, Peking University, Beijing 100871, China (e-mail: gaof@pku.edu.cn; anorange0409@pku.edu.cn; lingyu@pku.edu.cn).

X. Zhang is with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90007 USA (e-mail: xinfengz@usc.edu).

Y. Luo is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: yluo180@gmail.com).

X. Li is with the School of Electronics Engineering and Computer Science, Institute of Network Computing and Information Systems, Peking University, Beijing 100871, China (e-mail: lxm@pku.edu.cn).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMM.2018.2818012

WITH the explosive increase of images and videos, the intelligent visual signal analysis plays a more and more important role [1]–[5]. Herein, the visual feature descriptor extraction, compression, and learning is a fundamental approach in numerous computer visual tasks. Take visual search for example, the reference image/video databases are usually hosted at the remote servers, e.g., cloud, while the query images are captured at the frontal end, e.g., mobile phones or surveillance cameras, as shown in Fig. 1. However, confined to limited bandwidth, memory and computational resource, the computational lightweight visual descriptors with compact representation directly obtained from frontal devices are urgently demanded in large-scale visual analytic applications.

A considerable number of visual feature descriptors are proposed for visual signal analysis to achieve a good balance between the accuracy and descriptor compactness requirements, e.g., in [6]–[9]. The Scale-Invariant Feature Transform descriptors (SIFT) [6] is one of the classical visual features, which achieves good performance in many visual analysis applications, especially for visual search. However, SIFT is a cumbersome local descriptor due to its high computational complexity and bandwidth requirement.

More recently, the Moving Picture Experts Group (MPEG) has released the standard, Compact Descriptor for Visual Search (CDVS) [10], which provides the state-of-the-art solution for visual search with low bandwidth cost and high memory efficiency. The MPEG-CDVS consists of two kinds of feature descriptors, i.e., global descriptors (Scalable Compressed Fisher Vector (SCFV) [11]) and local descriptors (compressed SIFT-like descriptors), and defines 6 descriptor lengths, 512 bytes, 1 KB, 2 KB, 4 KB, 8 KB, and 16 KB, to adapt to different bandwidth scenarios, among which the descriptors have very good interoperability. To implement the compact descriptor representation, an important technique, interest point selection, is adopted in MPEG-CDVS, which selects a subset of most effective features for image representation. It also significantly reduces the computation complexity for subsequent local feature description, compression and aggregation. In general, the MPEG-CDVS only selects about 300 local features instead of thousands of features to represent an image, which achieves about 50% running time saving and also significantly reduces the memory and bandwidth requirement for visual search.

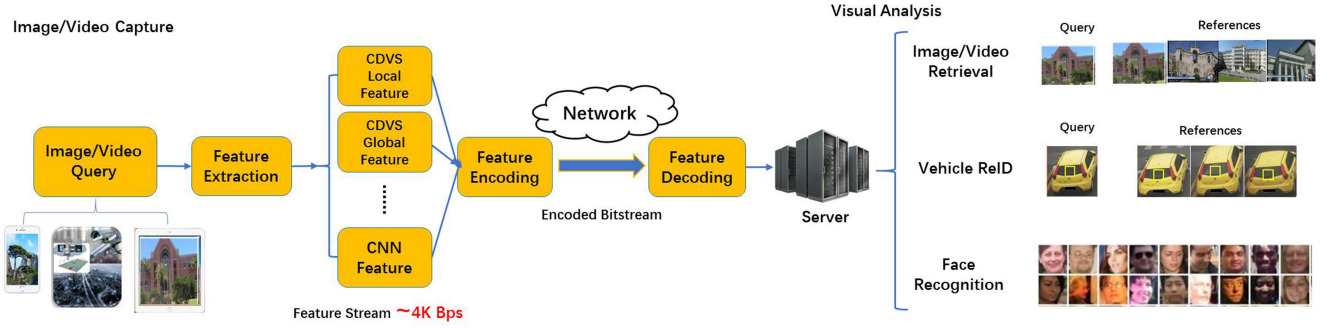


Fig. 1. The framework of the large-scale image/video analysis with feature extraction and compression from frontal end devices.

With the explosive increase of videos analysis demands, directly applying the MPEG-CDVS descriptors to frame-by-frame video analysis tasks will lead to excessive consumption of storage and bandwidth, and cannot fulfil high-performance large-scale analysis, especially under distributed camera networks. For example, as suggested by CDVS, the descriptor length for each frame is 4 KB, and for a typical 30 fps video, the bit rate of feature stream is approximately 1 Mbps, which still poses high burden for state-of-the-art wireless bandwidth. To further reduce the feature bit rates, MPEG initiated a new standard for video analysis, Compact Descriptor for Video Analysis (CDVA) [12], in 2015, and it not only inherits the spatial interest point selection strategy from CDVS, but also adopts color histogram based temporal sampling strategy to reduce the number of selected interest points for video analysis tasks. In addition, the deep learning based feature descriptors have exhibited substantial performance improvement in various computer vision tasks, and achieved very promising performance in image retrieval with more compact representation, e.g., RMAC [13], NIP [14] and Gordo [15]. The emerging MPEG-CDVA has adopted the combination of handcrafted CDVS and deep learning based feature descriptors due to their proven complementarity. To maximize the advantages of combining handcrafted and deep learning features, the interest point selection consistently plays a crucial role in removing the redundancy amongst the handcrafted local feature descriptors in spatial-temporal domain, and improving the matching and retrieval performance by combining the handcrafted feature descriptors and the deep learning feature descriptors.

As illustrated in Fig. 1, the compact descriptors directly extracted from frontal camera devices may significantly reduce the bandwidth cost, which is crucial for large-scale visual analysis. However, most of the feature extraction and compression for query images/videos are subject to the resource constraints of the frontal end devices, which are power and memory limited, and cannot afford heavy computing. Although the interest point selection plays an important role in compact image representation, there are few systemic study, especially from the perspective of improving search performance as well as developing lightweight algorithms to fit into smart frontal end devices with constrained computational capacity. In [6], Lowe utilized the response of the DoG (Difference of Gaussian) filter to detect interest points, but did not further leverage data-driven learning approach to filter in a compact set of discriminative

interest points for local feature descriptors. In [7], the Speeded-Up Robust Features (SURF) [7] was proposed by utilizing a Hessian matrix-based measure as the detector instead of the DoG in SIFT, which significantly reduces the descriptor extraction time. However, no explicit pruning of salient interest points is applied.

In this paper, we give in-depth investigation on the interest point selection for visual search, which is formulated as the problem of learning an optimal predictor via classification or regression in data-driven manner. To realize an extremely low complexity predictor, we propose to leverage the detector results (i.e., scale space filtering statistics) rather than the characteristics of local descriptors. Promising performance over benchmarks have been obtained by combing the improved compact handcrafted features (consisting of selected local descriptors and aggregated descriptors) derived from optimizing interest point selection, and the state-of-the-art deep learning features. The main contributions of this paper are three-fold:

- 1) We study the role of selecting local feature descriptors in aggregating the discriminative global representation, and formulate the problem of selecting those salient local patches for effective and efficient visual search mathematically. The proposed formulation serves as a guidance for optimizing the interest point selection via learning an optimal predictor of valid patches for successful matching to improve image search performance, and especially set up an open framework for improving the discriminative power of aggregated descriptor through selecting effective patches by utilizing various image statistical characteristics or other useful priors.
- 2) We have proposed a fast yet effective solution for the interest point selection problem. The optimal solution is learned from the inlier and outlier pairs, which are generated from a collection of image matching/non-matching pairs by geometric consistent check (GCC). In addition, a useful preprocessing stage is introduced to further reduce the influence of the inlier and outlier pairs with low confidence.
- 3) Furthermore, we have studied the significant impact of the interest point selection on both the accuracy and efficiency of image matching and retrieval, and also extensively analyzed the behavior of different prediction algorithms. Two lightweight predictors derived from regression functions have reported encouraging visual search performance in

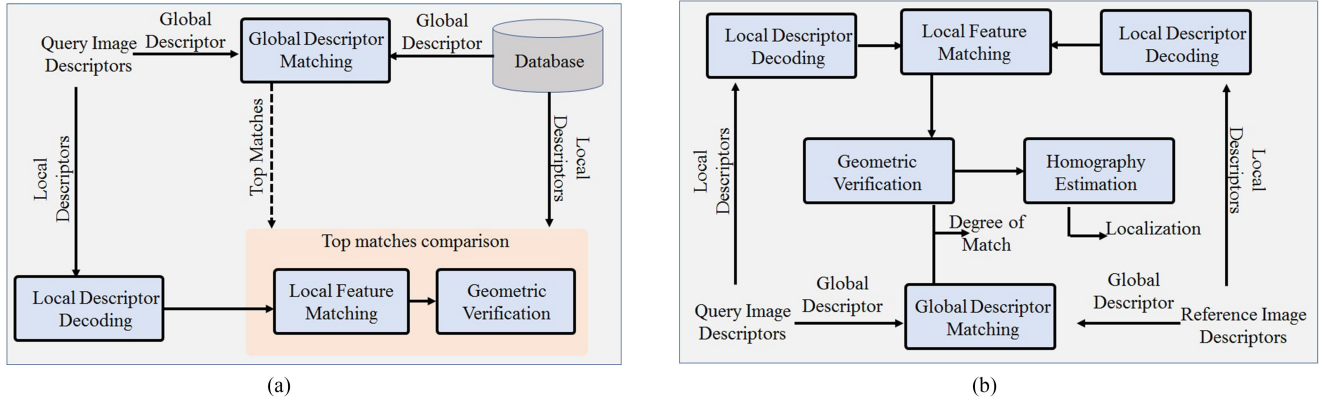


Fig. 2. State-of-the-art visual search framework, (a) image retrieval pipeline, (b) image pairwise matching pipeline.

terms of mean average precision (mAP), which outperform classification based solutions as well as the traditional LoG detector. In particular, the MPEG-CDVS standard and the emerging MPEG-CDVA standard have adopted a regression function to select interest points for generating compact handcrafted features, which has significantly improved search performance at reduced computational cost, together with deep learning based feature descriptors.

The remainder of this paper is organized as follows. Section II reviews the related works of interest point selection. Section III presents the interest point selection problem and its significance in low-bandwidth visual search. Section IV introduces the attributes or features used in interest point selection, and formulates the interest point selection as an optimization problem from the perspectives of classification and regression respectively. Section V introduces our implementation for the formulated optimal problem to improve the interest point selection performance. Extensive experimental results and discussions are reported in Section VI, and some explorations for future work on interest point selection are discussed. Finally we conclude this paper in Section VII.

II. RELATED WORK

A. Introduction of Visual Search

Visual search is a basic computer vision task that employs visual feature descriptors to find the images sharing the same or similar scenes or objects with that in query images, which is also denoted as image matching/retrieval. For more accurate visual search, we not only need to retrieve the images, but also need to localize the corresponding objects by pairwise matching. Fig. 2 shows the typical pipelines for pairwise matching and image retrieval. The state-of-the-art visual search systems work on both global descriptor (e.g., aggregated handcrafted descriptors, CNN deep descriptors) and compressed local feature descriptors, where the binary global descriptor can significantly speed up visual search process while the local descriptors are for further verification to improve the accuracy.

As shown in Fig. 2(a), the Hamming distance of global feature descriptors between query image and database images are first calculated to generate a short list of candidates. Then, the

geometric consistency check (GCC) is optionally performed by local feature matching and RANSAC [16], or other re-rank strategies. The global feature plays a crucial role, which directly determines the recall rate of candidates, and impact the retrieval performance. The local features contribute to pairwise matching, in which the matched local features can be applied to localize the objects by estimating the homography matrix. However, limited by the bandwidth, memory and computational resources, practical visual search system may not afford to represent an image using all the local feature descriptors. The matchability or efficiency of selected local feature descriptors is useful for improving the performance of aggregated descriptor by removing the impact of noisy local descriptors, as well as maintaining the localization accuracy via a limited number of local descriptors.

B. The Interest Point Selection for Visual Search

Many research efforts have been devoted to the selection of local features, and these work can be divided into three categories according to their derivation process.

1) *The Interest Point Selection based on Heuristic Rules:* During the development of MPEG-CDVS, Samsung Electronic proposed to select the interest points in visual attention regions [17] based on the assumption that more relevant descriptors are located in salient regions for human visual system. But it is difficult to measure the saliency of each interest point. Simone Buoncompagni *et al.* [18], [19] proposed to rank the interest points by measuring the distinctiveness, repeatability and detectability heuristically. The distinctiveness quantifies the difference between a given interest point descriptor and other interest point descriptors of the same object. The repeatability quantifies the difference between a given interest point descriptor and its corresponding descriptors of projected interest point on transformed images. The detectability quantifies the aptitude of a given interest point to be detected under various viewpoint and lighting changes. However, the distinctiveness and detectability rely on the local feature descriptors, e.g., BRIEF [20] in their work, which requires us to extract all the redundant local feature descriptors and calculate their distance, which thus significantly increases the computational costs for feature extraction. In [21], Mukherjee *et al.* further extended these heuristic rules to AKZE [22] and SIFT descriptors, and introduced Gabor

filters for constructing a texture map to enhance the salient map. Li and Calway [23] proposed a point selection process, which provides an even distribution of the “the most matchable” points across the scene in terms of pairwise 3D distance consistency, normal-normal and normal-position consistency, and selects the points by a non-maximum suppression algorithm. However, this kind of intuitive methods may lead to performance loss because they did not consider the variance of the distribution and the matching points.

2) *The Interest Point Selection Based on Unsupervised Learning*: In [24], Jan Knopp *et al.* proposed an interest point selection method by removing the points on the “confusing objects” in place recognition such as trees or road markings, which frequently occur in the database and cause significant confusion between different places. They learned the local confusion score distribution by GPS tags to automatically detect the “confusing objects” such as trees and window blinds, and excluded the candidate points with their scores greater than a certain threshold. Instead of the GPS information, Turcot *et al.* [25] utilized each image in the reference database as a query, and defined the *useful* features which are robust enough to be matched with a corresponding feature in the same object, stable enough to exist in multiple viewpoints, and distinctive enough that the corresponding features are assigned to the same visual word. Thus, an unsupervised preprocessing is applied to identify correctly matching features between the training images, and database images, and then the *useful* feature descriptors are selected based on an adjacency graph among database images. But this method relies on the given database and is difficult to extend to individual query images or other databases. In [26], Tolias *et al.* proposed to keep the frequent features in the query extension stage as useful features, and refine local descriptors by comparing the distance between two local descriptors assigned to the same visual word. If the distance is above a predefined threshold, the descriptors are regarded as non-matching ones. However, the unsupervised learning methods using a single image’s statistical attributes cannot well predict the matchable of interest points.

3) *The Interest Point Selection based on Supervised Learning*: To further improve the prediction accuracy of matchable features, Hartmann *et al.* utilized the random forest to classify the local descriptors as matchable and non-matchable ones. The training samples are generated by matching 13 preceding and 13 subsequent images in each image sequences which are captured at different localtions and varying lighting conditions. In [27], Dorkó and Schmid presented a more accurate feature classifier by utilizing the manually labeled positive and negative descriptors. In their method, two step classification is applied, and the first step is a unsupervised local descriptor clustering, and the second step is a supervised learning process using two types of classifiers: Support Vector Machines (SVMs) and Caussian mixture model (GMM) based classification. In [28], Demirci and Kacka used a classifier that predicts which features are salient among neighboring views of the same objects, and constructed the training samples from 10 objects with 72 rotated views with a fixed illumination angle and color. Similar approaches are also developed in [29], [30] for different descrip-

tors, e.g., ORB and VLAD. In [31] Dymczyk *et al.* exploited the Convolutional Neural Networks (CNN) as the classifier for descriptors using both raw image and depth information, which obviously increase the computational cost for interest point selection. Moreover, these classifiers heavily depend on the specific features and are difficult to extend to general features.

In addition, the emerging object instance search needs to efficiently find out all the object proposals in an image with minimum redundancy among them [32]. To reduce the visual search complexity, they utilize the *k* – *means* method to cluster all the proposal features and generate a small set of fuzzy objects which are treated as the atoms of an image specific dictionary, and then generate image compact feature representation by sparse coding according to the dictionary. Thereafter the features of all object proposals generated from images are encoded into a set of sparse locality-constrained linear codes [33], [34]. Although these works aim to construct compact visual features which are different from interest point selection, they also reduce the visual feature redundancy in image representation. However, they are feature-dependent methods, which incur heavy extraction complexity.

III. PROBLEM DESCRIPTION

Descriptor extraction and matching for a large amount of interest points are time-consuming for smart front-end devices. When transmitting the descriptors to the cloud server for remote matching, the bandwidth is also constrained. It is desirable to select a subset of effective (or valid) interest points prior to the subsequent descriptor extraction, transmission and matching. As illustrated in Fig. 3, we aim to find out those interest points with high matchability so that the selection process has little influence on the final matching.

We formulate the problem of determining interest point validity in the local and global matching to guide the interest point selection. Let $\mathcal{P} = \{p_n, n = 1, 2, \dots, N\}$ denotes the set of N interest points detected from an image. In the local matching, if an interest point p in an image is matched with an interest point in another image and passes geometric consistency check, then p is called an inlier. For an inlier, $\psi(p) = 1$ and otherwise $\psi(p) = 0$. Then the local matching score of a pair of images is given by

$$s_L = \sum_{n=1}^N \psi(p_n). \quad (1)$$

If the network bandwidth is adequate, we may include all interest points in the image matching. But when the bandwidth is limited, we need to select a subset of interest points for matching. We use $\phi_L(p_n) = 1$ to denote that p_n is selected and call it a valid interest point in the local matching, otherwise $\phi_L(p_n) = 0$. After the interest points selection, the local matching score becomes

$$s'_L = \sum_{n=1}^N \phi_L(p_n) \psi(p_n). \quad (2)$$

Intuitively, if a valid interest point is not an inlier ($\psi(p_n) = 0$), it will not contribute to the final score. That is, if an interest

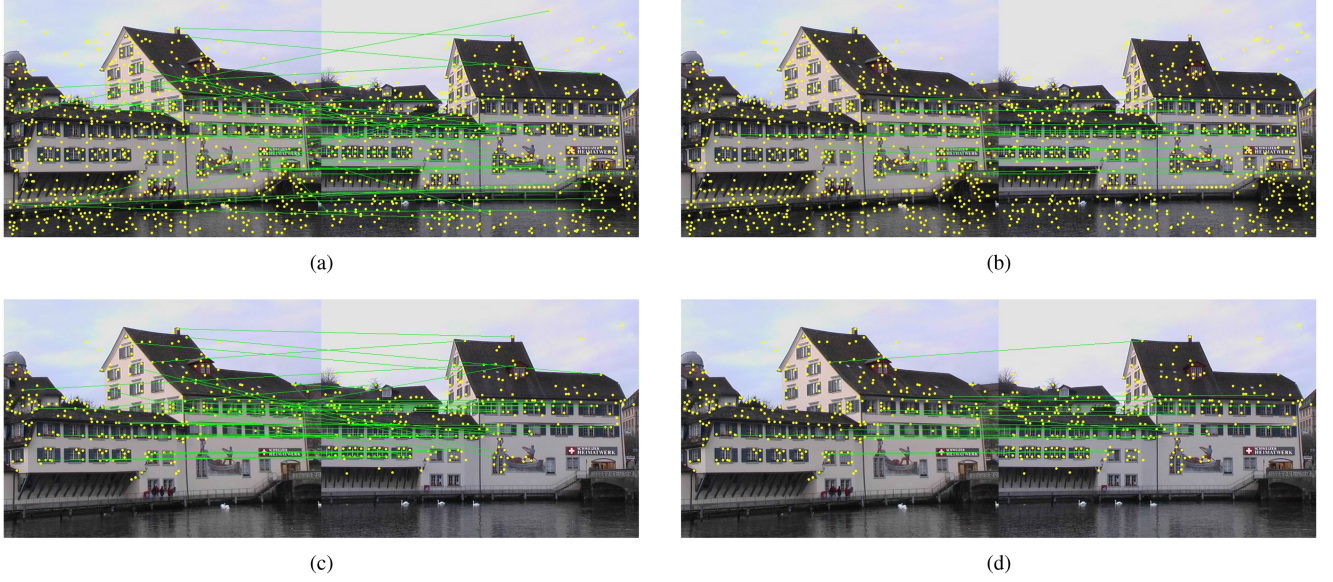


Fig. 3. Matching between a pair of images. (a) Without interest points selection, a large amount of candidate matches can be found. (b) With geometric consistency check, only a few good matches left. (c) By adopting the interest points selection, the number of interest points are reduced significantly and so are the candidate matches. (d) But there is little impact on the good matches since effective interest points (with high matchability) have been selected.

point is an inlier, this point is supposed to be selected as a valid interest point with high probability.

In addition to the local matching, we also aggregate the interest points to generate a global descriptor for improving the matching accuracy. In particular, the descriptor of each interest point p is assigned to a certain visual word in the codebook to form a statistical vector $\mathbf{r}(p)$ (such as the residual between the descriptor of p and its assigned visual word). Let \mathcal{P}_k be the subset of interest points that are assigned to the k -th visual word, then the k -th component of the global feature vector is

$$\mathbf{g}_k = \sum_{p \in \mathcal{P}_k} \mathbf{r}(p), \quad (3)$$

which is obtained by aggregating the descriptor statistics of all interest points that are projected on the k -th visual word. The global feature is a concatenation of all components, i.e., $\mathbf{g} \propto [\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_K]$. If many irrelevant or redundant interest points are involved in the aggregation process, the discriminative ability of the global feature would decrease. Therefore, it is also critical to select valid interest points for aggregation. We use $\phi_G(p) = 1$ to indicate that p is selected as a valid point for global aggregation, and $\phi_G(p) = 0$ otherwise. Then the k -th component of the global feature is

$$\mathbf{g}'_k = \sum_{p \in \mathcal{P}_k} \phi_G(p) \mathbf{r}(p), \quad (4)$$

Hence the global matching score is

$$s_G = \sum_{k=1}^K w_k \left\langle \gamma(\mathcal{P}) \sum_{p \in \mathcal{P}_k} \phi_G(p) \mathbf{r}(p), \cdot \right\rangle, \quad (5)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, w_k is a scalar that indicates the importance of the k -th visual word and $\gamma(\mathcal{P})$ is a normalization factor. If p is an inlier, then the statistical vectors

between an interest point p and its corresponding interest point in another image are with large similarity value. Herein, the other term in inner product of (5) has been omitted since it shares the same expression with the first term but for the other image. This results in a large inner product and contributes to the final matching score. Therefore, we should select as many inliers as possible.

In this paper, we define a prediction function to denote the validity of an interest point, and our goal is to learn this function so that the output value is large for inlier, and small otherwise. However, learning such a validity function is nontrivial due to the following major challenges:

- First, this is an ill-posed problem. We need to predict the validity of interest points from a single query image. Without image matching pairs, we cannot exactly tell whether an interest point is an inlier or not.
- Second, the interest point selection is required to be carried out within the detector phase, definitely before descriptor extraction, to save more computation. That is, we cannot leverage detailed descriptor content.
- Third, the bandwidth and computational resources are limited. The interest point selection algorithm must be efficient and has low memory consumption. Besides, as the number of resulting points is bounded by the network bandwidth, the algorithm should be able to select a varied number of points by ranking validity values.

IV. INTEREST POINT SELECTION FOR VISUAL SEARCH

For effective and efficient interest point selection, we first extract lightweight features that can indicate the importance of an interest point in the matching. Then we generate abundant labeled interest points by performing geometric consistent check on matching image pairs. Finally, given the features of the

interest points and their associated labels, we learn a model to predict validity scores of the interest points in a single image and perform selection according to the scores.

A. Features Extraction and Training Data Generation

Directly utilizing the feature descriptor (say, 128-dimensional SIFT) for validity prediction has several flaws: 1) extracting descriptions for all interest points is time-consuming; 2) the high dimension of the description may lead to over-fitting when the labeled training data is limited; 3) the description heavily depends on the visual content, which would lead to more domain-specific selection method.

Therefore, we propose to use the low-level features derived from interest point detection for validity prediction. Below we list the representative features.

- Peak (extreme value) of the LoG response. It indicates the contrast around an interest point, and a high contrast usually corresponds to high distinctiveness and stability.
- Scale. It indicates the patch size of details. Those small-scale or large-scale patches are hard to be matched correctly since its size may be outside of the range or it is too large to be detected in the other image.
- Orientation. Empirical finding shows that the interest points in horizontal and vertical directions are more likely to be matched than diagonal ones.
- Distance from center. An interest point closer to the center is more likely to be an inlier since the target/interest objects usually locate in the center of an image.
- Other statistical moments. For example, curvature ratio is calculated as $(r + 1)^2/r$, where r is the ratio of the two eigenvalues of the Hessian matrix at the interest point, the $\log_hessian_eign1$ and $\log_hessian_eign2$ mean the first and the second eigenvalues for the LoG Hessian matrix respectively, the $g_hessian_eign1$ and $g_hessian_eign2$ represent the first and second eigenvalues of the Gaussian Hessian matrix, the $\log_hessian_det$ denotes the determinant value of the LoG Hessian matrix, $\log_hessian_trace$ is the trace of the LoG Hessian matrix, and the $g_hessian_det$ and $g_hessian_trace$ mean the determinant and trace of the Gaussian Hessian matrix.

For training data, we first collect good number of image pairs that contain the same visual object. The interest points passing the geometric check are labeled as positive samples (inliers), otherwise as negative samples (outliers).

B. Interest Point Selection Method

Given the features and associated labels of abundant interest points, we propose a lightweight data-driven approach to effective and efficient interest point selection from a single image. As shown in Fig. 4, this approach consists of three main procedures: feature selection, preprocessing and validity prediction. Below, we describe these procedures in details.

1) *Feature Selection*: There are different types of features that may be useful for predicting the validity of interest points. However, to achieve efficient interest point selection, it is desirable to use as fewer features as possible. Moreover, there exist

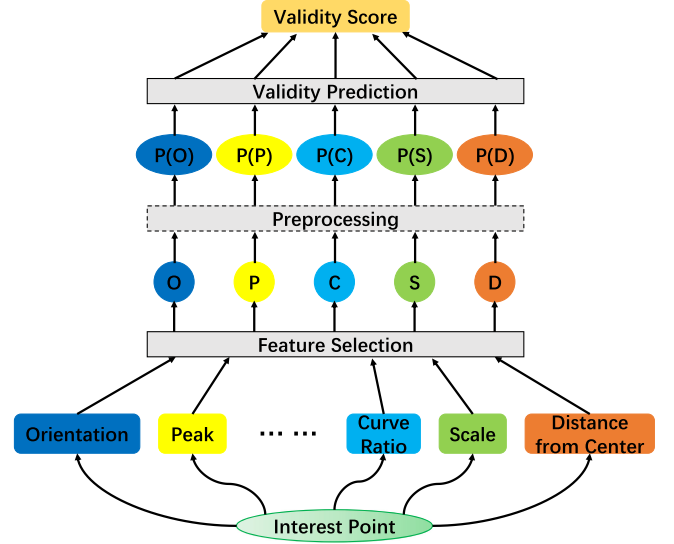


Fig. 4. The proposed lightweight interest point selection approach. Given the extracted different features (such as peak of LoG response, distance from center, etc.), we first perform feature selection, and then preprocess different features. Finally, a prediction model is learned to predict the validity score of an interest point.

redundancy. Therefore, we propose to perform feature selection prior to the validity prediction.

There are a variety of feature selection approaches in the literatures. For example, we can evaluate the importance of each feature or feature subset according to some pre-defined statistical measures, such as Information Gain and Mutual Information [35]. To select features for a certain prediction model, we can divide all features into different subsets and the importance of each subset is evaluated by testing their combination performance in visual search application. The subset which leads to the best performance is selected. However, the computational complexity of this strategy is quite high. Therefore, recent feature selection methods focus on learning the feature importance together with the prediction model. The importance of each feature is determined by learning a weight for it, and some sparsity constraints [36] are usually enforced to select a small number of features and sometimes exploit the feature correlations [37]. After feature selection, we obtain a feature vector $\mathbf{x} = [x_1, x_2, \dots, x_D]^T$ for interest point p .

2) *Preprocessing*: In interest point selection, the values of original features may vary within a wide range. This would deteriorate the predictor performance. Therefore, we propose to preprocess the original data by imposing the range constraints. Rather than simply normalizing the original feature vector to a unit range, we propose to utilize the label information for more effective normalization. Specifically, an individual predictor $g(\cdot)$ is learned for each feature and the output score $g(x_d)$ is the normalized value for each input feature $x_d, d = 1, 2, \dots, D$. This preprocessing strategy not only makes the normalized features comparable, but also increases discriminative ability of each feature since the label information is utilized. The performance impact from this preprocessing will be reported in our experiments.

Algorithm 1: Interest Point Selection for Visual Search**Input:** The training matching image pairs $\{(I_t^1, I_t^2)\}_{t=1}^T$;**Output:** The selected interest point subset for each query image I_q , i.e., $\{p_{q,n}\}_{n=1}^{N_S}$, N_S is the number of selected features;

- 1: Detect the interest points in training images, extract the features $\{\mathbf{x}_{t,n}\}_{n=1}^{N_{I_t}}$ for each interest point $p_{t,n}$;
- 2: Perform pair-wise interest point matching in training images, label the matched interest points as positive sample pair and the non-matched ones as negative pair;
- 3: Preprocess the features by learning an individual predictor $g(\cdot)$ for each feature, and learn an additional prediction model $f(\cdot)$ on the preprocessed features;
- 4: **For a query image** I_q
- 5: Detect the interest points $\{p_{q,n}\}_{n=1}^{N_{I_q}}$, extract the features $\mathbf{x}_{q,n}$ for each interest point $p_{q,n}$;
- 6: Preprocess the features and predict the validity score $s_{q,n}$ for each interest point using the learned $f(\cdot)$;
- 7: Sort the interest points according to their validity scores, and select the top-ranking ones $\{p_{q,n}\}_{n=1}^{N_S}$ for subsequent description extraction and image matching.
- 8: **End for**

3) *Validity Prediction:* To predict the validity of any interest point, we learn a prediction model using the training samples $\{g(\mathbf{x}_n), y_n\}$ after feature selection and preprocessing, where each $g(\mathbf{x}_n) = [g(x_{n1}), g(x_{n2}), \dots, g(x_{nD})]^T$ and y_n is the corresponding label. The presented prediction model together with the preprocessing is closely related to multi-view learning [38], [39], which aims to combine information of multiple representations (such as features and modalities). In our method, each dimension of the selected features can be regarded as a view. It is not appropriate to directly concatenate different views since they have quite different statistical properties. In this work, we preprocess different views so that the output results have similar statistical properties. The final validity score is an integration of the preprocessed outputs, i.e.,

$$f(\mathbf{x}) = f(g(x_1), g(x_2), \dots, g(x_D)). \quad (6)$$

We propose to combine the outputs of different views (features) by training an additional prediction (via classification or regression) model. We do not adopt more sophisticated multi-view learning strategies such as multi-view subspace learning [40], because we have only one scalar element in each view. This is well aligned with our target of lightweight feature selection method. In addition, we allow the predictors adopted in preprocessing and validity prediction to be different. For example, naive Bayesian can be used for feature preprocessing (normalization) and linear regression can be employed for view (feature) combination. Hence, the proposed method is able to take advantages of different predictors and thus achieves satisfactory performance as shown in our experiments. The pseudo-code of the proposed interest point selection algorithm is summarized in Algorithm 1.

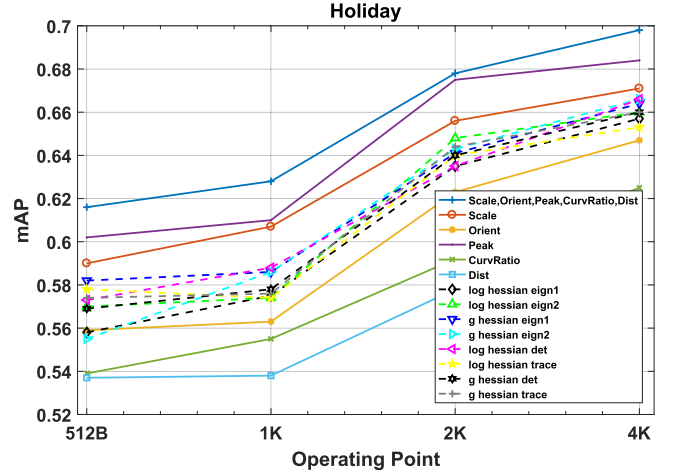


Fig. 5. The performance comparison by using different features or combination.

V. OUR IMPLEMENTATION

In this work, we select a subset of effective features by mutual information and heuristic rules. A comparison of different features is shown in Fig. 5, to reflect their individual performance and complementary nature. We finally choose five features for interest point selection, i.e., scale, main orientation, the peak response value of LoG, distance from center, and the ratio of the squared trace to the determinant of the Hessian. From Fig. 5, we may give several reasons for the selection as follows: 1) “peak” and “scale” achieve higher performance than other features and are complementary to each other; 2) although “Orient”, “CurveRatio” and “Dist” perform worse than the different hessian statistic features, we obtain larger improvements when combining them with “peak” and “scale”. The main reason is that the three features are more complementary to “peak” and “scale” than the hessian features. All the selected features can be extracted quite efficiently in the interest point detection stage and hence significantly accelerate the whole process of generating compact descriptors including local and global descriptors.

For preprocessing, we first quantize the continuous features by clustering all features in the training set into different groups. Then given the quantized feature values $[X_1, X_2, \dots, X_M]^T$ for a certain interest point, we replace the m -th value X_m with the following conditional probability

$$P(y = 1 | x_m = X_m) = \frac{P(y = 1 \cap x_m = X_m)}{P(x_m = X_m)}. \quad (7)$$

The probability $P(x_m = X_m)$ is the probability of the m -th feature assigned with a specific value X_m . This is calculated as the ratio of interest points that contains the feature value X_m in the training set. The probability $P(y = 1 \cap x_m = X_m)$ is the ratio of inliers that contains the feature value X_m . Therefore, the conditional probability $P(y = 1 | x_m = X_m)$ is the possibility of an interest point to be an inlier if it contains the feature value X_m . This preprocess normalizes all feature values into the range $[0, 1]$. An interest point that contains larger normalized feature values are more likely to be an inlier, and hence should be selected.

Let $\hat{\mathbf{x}}_n$ be the preprocessed feature vector of \mathbf{x}_n , and suppose the training set is given by $\{\hat{\mathbf{x}}_n, y_n\}$, where $y_n \in \{1, 0\}$ indicates that p_n is an inlier or not. Our ultimate goal is to learn a prediction function $f(\hat{\mathbf{x}}_n)$, where the output is large for inlier and small otherwise, i.e.,

$$f(\hat{\mathbf{x}}_n) > f(\hat{\mathbf{x}}_m), \forall (n, m) : y_n = 1 \ \& \ y_m = 0. \quad (8)$$

Since the given label is binary, it is natural to regard the validity prediction as a binary classification problem. From the perspective of classification, we compare and analyze the behavior of two representative classification models, i.e., SVM [41] and random forest [42]. For SVM, the validity score for a test interest point is given by its distance to the separating hyperplane, i.e., $f(\hat{\mathbf{x}}) = \mathbf{w}^T \hat{\mathbf{x}} + b$, where \mathbf{w} and b are the learned weight vector and bias. For random forest, the validity score is the ratio of leaves with positive labels.

However, the objective of a classifier essentially is to separate different classes, not to predict a (continuous) score. Besides, the computational cost is often high for a large amount of training samples. For example, the complexity of support vector machines (SVM) is more than quadratic w.r.t. the number of training samples. On the contrary, regression approaches are more appropriate than classification models in estimating a continuous function, and they are usually more efficient. For example, stochastic gradient descent can be directly employed for optimizing linear and logistic regression. Moreover, the memory complexity of regression approaches is extremely low or nearly zero in the online prediction process. By directly minimizing the difference between the output score and groundtruth label (0/1), we can learn a continuous function that tends to assign large value (close to 1) for inliers and small values (close to 0) for other points. The major issue of using regression for our validity prediction application is that only binary scores are provided in the training. Therefore, the predicted score does not exactly indicate the importance of an inlier, and hence we cannot guarantee that a regression model can achieve higher accuracy consistently than classification. This can be observed in our experiments.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

A. Databases and Evaluation Criteria

To analyze the performance of different interest point selection methods, we perform the image retrieval tasks on the popular databases, including MPEG-CDVS benchmark database [10], [43], Holiday [44], Oxford5k [45], Paris6k [46] and multi-view database [47]. Herein, MPEG-CDVS benchmark database consists of 5 classes: graphics, paintings, video frames, landmarks and common objects with a total of 8314 query images, 18840 reference images and a distractor set of 1 million images from Flickr [48] for the retrieval experiment. The Holiday database includes 500 query images and 1491 reference images, which covers a very large variety of scene types (natural, man-made, water and fire effects, etc). The Oxford5k (55 query images, 5062 reference images) and Paris6k (55 query images, 6412 reference images) mainly includes the building images.

Specially, the multiview database is utilized to verify the role of the depth information in interest point selection, and it contains the depth information for each query and reference images. To construct this database, we capture the indoors object images with a mobile camera (Samsung S6). For each target object, 5 ~ 6 videos with average duration of 10 s are taken in different cluttered scenes, and the key frames are selected every 50 frames to form the query image set. For each query image, the depth information is estimated based on the subsequent selected frame. In total, there are 1009 query images, and 108 reference images, where there are 3 clean reference images are captured for each target object. We merge the reference images with about 1 million distractor images in our experiments.

For the image retrieval, the Mean Average Precision (mAP) is widely utilized to measure its performance. The mAP for a set of queries is calculated as the mean of the average precision scores for each query, which is defined as follows,

$$\text{mAP} = \frac{\sum_{q=1}^Q AP(q)}{Q}, \quad AP = \int_0^1 p(r) dr, \quad (9)$$

where Q is the number of queries, and AP is the average precision, and $p(r)$ is the precision function at recall r .

B. Performance Analysis on Interest Point Selection

To analyze the performance of different interest point selection strategies sufficiently, we first carry out the visual search on MPEG-CDVS database. Two baseline interest point selection methods, i.e., varying the LoG threshold and random selection, are utilized in our experiments to verify the effectiveness of the derived selection strategies. Herein, the LoG serves as the state-of-the-art interest point detector in SIFT features, and shows higher values near regular structures with meaningful information. Thus, we can adapt the local feature number per image by properly thresholding the LoG response to select the points with the most abundant information. The random selection serves as a sanity check which any reasonable prediction should beat.

In our performance testing, we compare the classification solutions using Support Vector Machine (SVM), Random Forest (RF) and the regression solutions using the linear and logistic solution. Herein, the the learned parameters for linear and logistic regression are [0.7044, 0.0262, 0.3680, 0.4653, 0.6822] and [7.0848, 0.0652, 4.0804, 4.8527, 6.9627], which correspond to scale, main orientation, the peak response value of LoG, the ratio of the squared trace to the determinant of the Hessian and distance from center, respectively. The linear kernel function is utilized in SVM.

Firstly, we analyze the impact of the preprocessing using the Probability Distribution Function (PDF) in interest point selection stage. Given an image, we first detect the interesting point candidates using the ALP method [49] in CDVS, and then apply the regression and classification based interest point selection method w/o preprocessing, respectively. For each image, we select the number of local features from 50 to 800 and aggregate them into Scalable Compressed Fisher Vector (SCFV), which is adopted in MPEG-CDVS standard. Fig. 6 shows the comparison results for image retrieval application on different

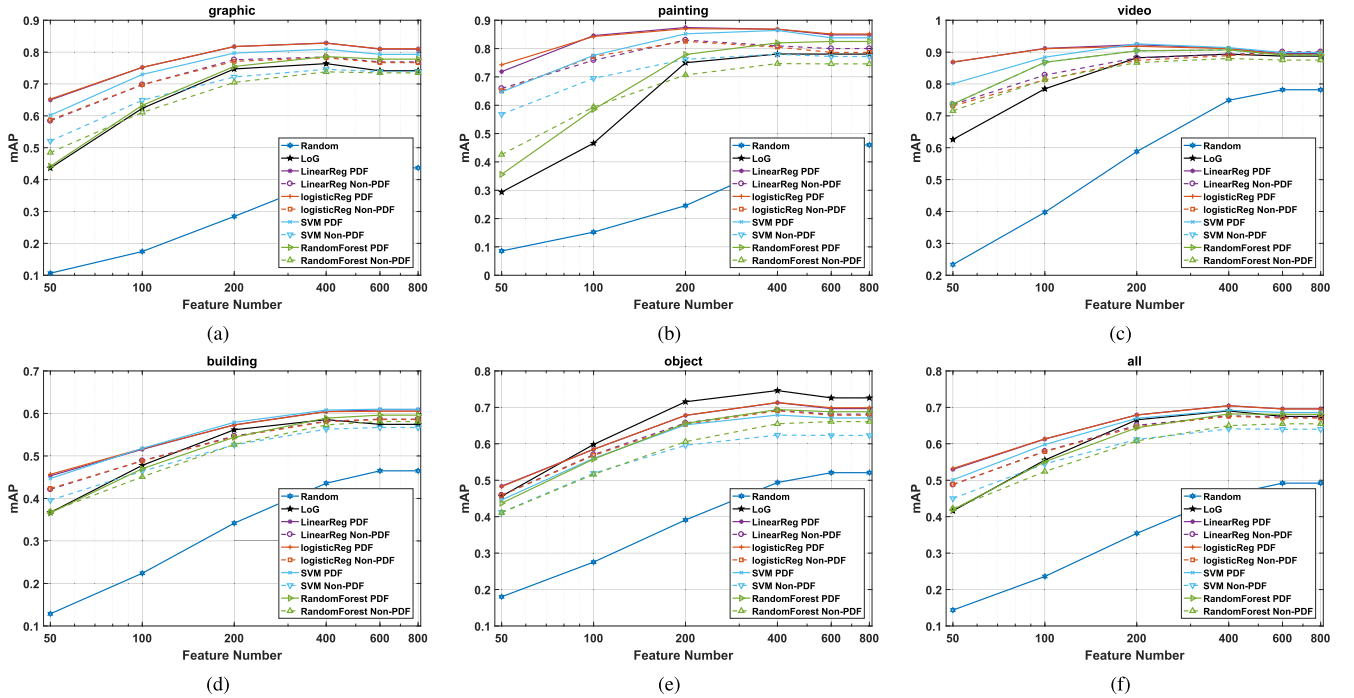


Fig. 6. Performance comparisons on MPEG-CDVS datasets. (a) Graphics, (b) Painting, (c) Video, (d) Building, (e) Objects, (f) Average results on all the datasets.

sub-databases of CDVS. From the results, we can see that the preprocessing is very useful in improving retrieval performance, achieving about 5%~10% improvement on mAP compared with that without preprocessing. In addition, when the number of the local features is beyond 400, the image retrieval performance is almost stable with marginal performance fluctuation. This phenomenon shows that it is enough to represent an image with meaningful or salient objects using only 300~400 local features by a SCFV aggregated descriptor. The more local features with trivial information would serve as noise and drop the expressiveness ability of SCFV to capture the salient information in an image, which also shows the importance of interest point selection in global feature construction. In general, the regression based methods perform better than that of classification methods, especially with much fewer local features. The classification based method, SVM, achieves better results compared with that of regression method for the building database, where there are much more regular structures.

To further evaluate the effectiveness of interest point selection, we carry out the image retrieval and pairwise matching experiments on CDVS database at different descriptor lengths using both the global and local features. Table I lists the numerical results for the six interest point selection strategies, where the regression and classification based interest point selection methods have applied the PDF preprocessing. Compared with the random selection method, the others all achieve very obvious performance improvement at all the descriptor length scenarios, which further proves that selecting local features have significant influence in visual search. In addition, compared with the LoG selector, the regression and classification based methods are more effective in low bit-rate scenario. While at high bit-rate

cases, although the LoG detector, regression and classification based method achieve approximate performance, the reasonable interest point selection still have significant influence on the retrieval and matching performance. This conclusion can be derived from the comparison with random selection. Even at high bit-rate scenario, the LoG, regression and classification based selection methods significantly outperforms the random selection. Moreover, the regression based methods performs much better than that of classification based methods, and this may be attributed to that the regression methods attempt to project the local feature importance into a continuous space which is more suitable for feature ranking than that of binary classification. Furthermore, in the pairwise matching experiments, we only use the local feature matching to show the influence of interest point selection on the local features. From the matching results in Table I, the regression based methods, especially the logistic regression, achieve more accurate results, which shows that the selected interest points using regression methods, usually falling into the texture region with consistent structure, are suitable for geometric consistent check.

By comparison between proposed logistical regression method and random method, we can see that the MPEG-CDVS features with 512 bytes generated using the proposed interest point selection method can outperform the MPEG-CDVS with 16 KB length generated from random selected interest points. In other words, it means that the proposed interest point selection achieves more than 96% bandwidth saving compared with random selection strategy for visual search.

To show the selection results intuitively, we visualize the interest point candidates and the final determined points by different interest point selectors in Figs. 7 and 8, where the yellow

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT INTEREST POINT SELECTION METHODS USING THE PREPROCESSING
AT DIFFERENT DESCRIPTOR LENGTH ON MPEG-CDVS DATASET

Methods/Descriptor Length	Image retrieval (mAP)						Pairwise Matching, TPR@FPR = 0.01					
	512 B	1 KB	2 KB	4 KB	8 KB	16 KB	512 B	1 KB	2 KB	4 KB	8 KB	16 KB
Random	0.4210	0.4619	0.5200	0.5747	0.5771	0.5712	0.5509	0.6410	0.7578	0.8268	0.8907	0.9110
LoG	0.6380	0.6894	0.7226	0.7515	0.7533	0.7504	0.6605	0.8241	0.8701	0.8965	0.9177	0.9255
Linear regression (PDF)	0.6721	0.7060	0.7334	0.7573	0.7564	0.7535	0.7322	0.8481	0.8881	0.9067	0.9242	0.9308
Logistic regression (PDF)	0.6717	0.7062	0.7330	0.7584	0.7571	0.7546	0.7344	0.8470	0.8892	0.9070	0.9246	0.9306
SVM (PDF)	0.6580	0.7010	0.7300	0.7530	0.7520	0.7490	0.6979	0.8486	0.8873	0.9065	0.9249	0.9322
Random Forest (PDF)	0.6440	0.6890	0.7240	0.7500	0.7490	0.7440	0.6190	0.8169	0.8749	0.9017	0.9226	0.9283

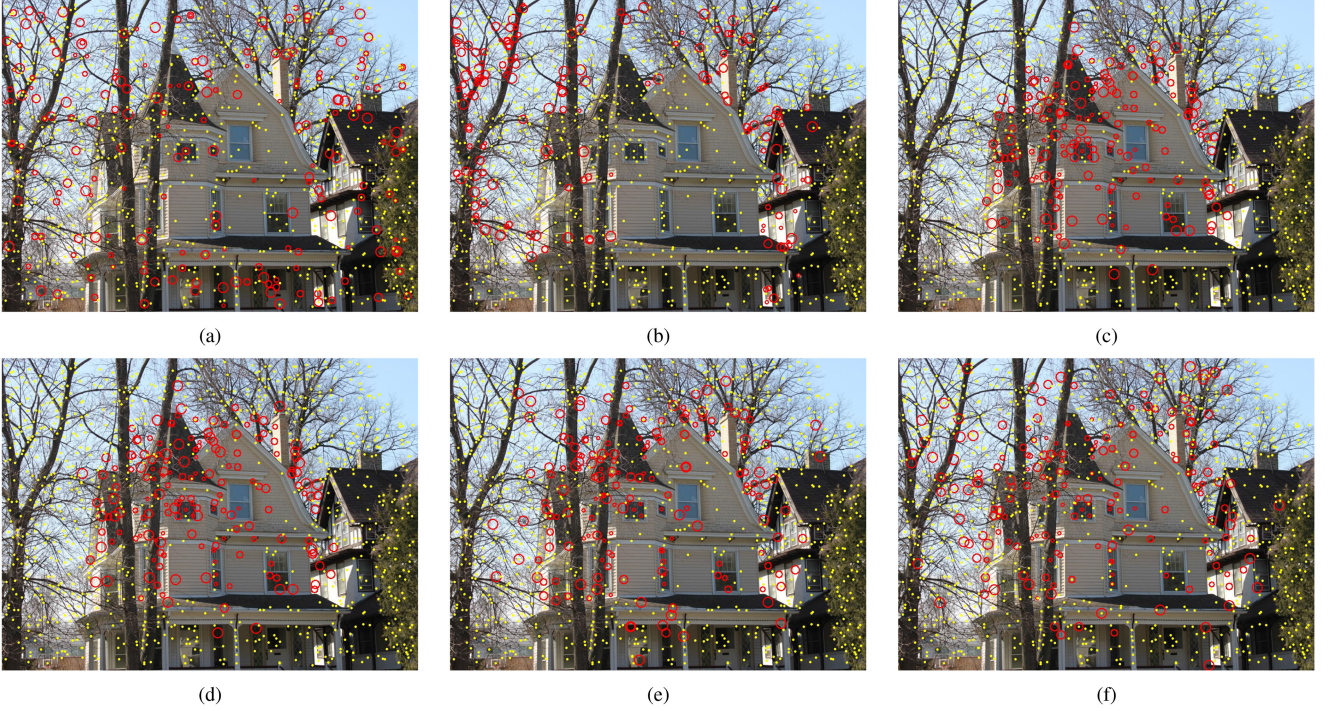


Fig. 7. Illustration of selected features by using different interest point selection methods. (a) Random selector, (b) LoG selector, (c) Linear regression selector, (d) Logistic regression selector, (e) SVM selector, (f) Random Forest selector.

dots represent the interest point candidates obtained from the local extremes in LoG scale space, and the red circles represent the final determined points (selecting 200 local features). We can see that although there are many extremes in LoG scale space, most of them are meaningless for image retrieval, because for most of the users the building are the most concerned content, which are the most likely to be queried information, while the trees or the background sky are useless for users in visual search which even may mislead the retrieval results when too many useless local features are selected. The LoG selector focuses on the points with large variations in a neighborhood, e.g., edges or textures, which does not consider the salient content prior distribution. Thus, most of selected local features are distributed among the tree trunks. By combining more attributes in the regression and classification based selectors, most of the selected interest points are more concentrated on the buildings. Especially, as a practically useful prior the attribute of “Distance from Center” is an effective complementary to LoG,

which assumes that most of the meaningful objects tend to be located in the center of images. So we can get a more reasonable results for interest point selection by jointly utilizing these attributes.

C. Computation Complexity Analysis on Feature Extraction

Another merit of interest point selection is to directly reduce the local feature description complexity since the interest point selection method does not rely on the descriptor. As shown in Fig. 9, we compare the percentage change of running time for different modules of CDVS, and the results are tested on a DELL PC with 3.40 GHz Intel Core(TM) i7@6700 CPU and 16 GB RAM in a mode of single core and single thread by averaging the CDVS extraction time for 1000 images with VGA resolution. When applying the interest point selection after the local feature description, the computational bottleneck is local feature description, but if we apply the interest point selection before feature description, the running time percentage

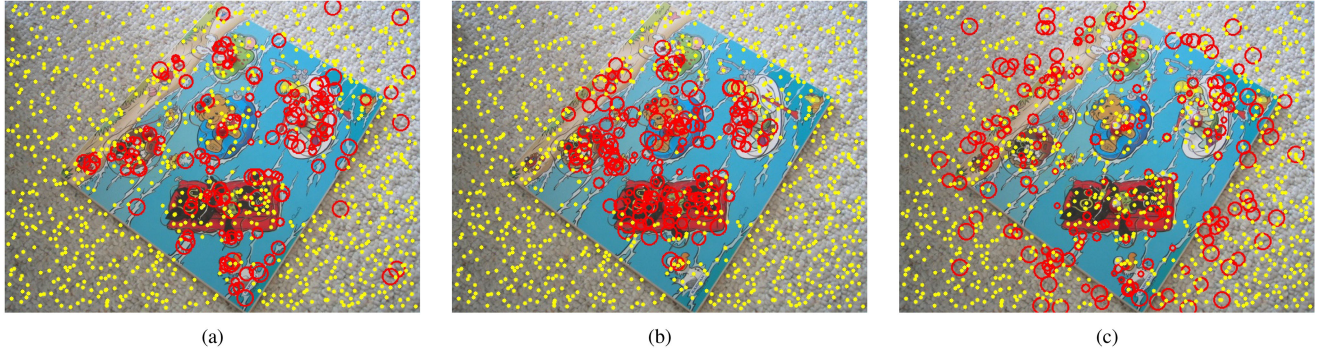


Fig. 8. Illustration of selected features by using different interest point selection methods. (a) LoG selector, (b) Logistic regression selector, (c) SVM selector.

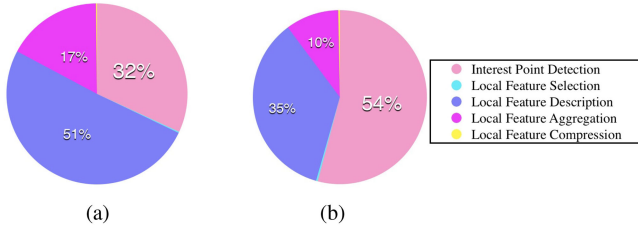


Fig. 9. The time cost comparison of MPEG-CDVS core modules by incorporating interest point selection, (a) extracting local feature descriptors before interest point selection, (b) extracting local feature descriptors after interest point selection.

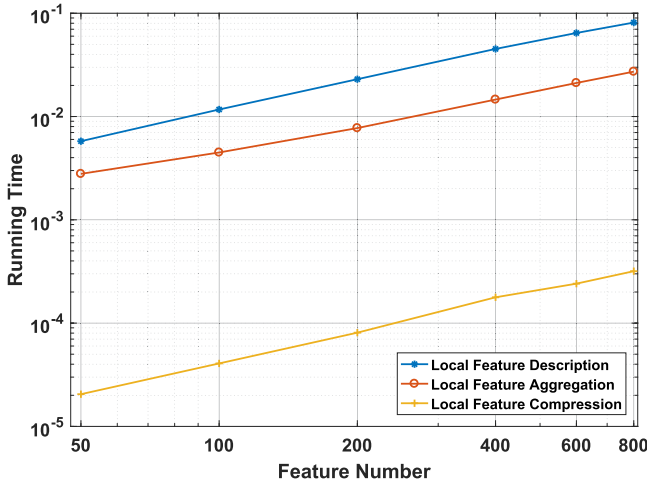


Fig. 10. The time consumption for local feature description, compression and aggregation at different number of local features.

of local feature description decrease from 51% to 35%. This is why MPEG-CDVS and MPEG-CDVA adopt the interest point selection before description without using the descriptors as an attribute for interest point selection.

From the above results, we also note that the running time consumption of interest point selection is marginal, smaller than 1%. That's what we claim as a merit of lightweight interest point selection method. To show the effects of the interest point selection on its subsequent modules, we further show the running time variations of the three major modules, i.e., local feature description, compression and aggregation, along with the number of local features, as shown in Fig. 10. We can see that the

running time of the three modules almost increase linearly along with the number of local features, indicating that interest point selection is very useful to reduce the computation complexity (especially at smart frontal-end devices) for large-scale visual search with real-time requirements.

D. Outlook for Interest Point Selection

Although we employ five attributes derived from scale space analysis to solve the formulated problem of interest point selection by classification or regression, there are more other types of image attributes like those shown in Fig. 5. What are the optimal image attributes or useful priors and what are the global optimal combination for them remain as open issues for future work, which is expected to be explored under the formulated regression and classification frameworks. For example, as a meaningful extension, we further explore another interesting attribute, *Relative Depth Characteristic* (RDC) [47], which can be derived from neighboring frames when applying visual search on videos. In the attribute combination, we utilize the nonlinear regression function to combine the above five attributes and the RDC to predict the function of selecting interest points (denoted as *RDC_regression*). Table II shows the image retrieval results on the multi-view database at different descriptor lengths. We can see that RDC is also a very good attribute for visual search, which achieves the best performance on the multi-view database compared with the other attributes individually. However, when we combine it with other attributes, the performance of image retrieval is further improved, which shows that the significance of interest point is also associated with the depth information. The objects with salient information are usually with moderate depth near the focal length of the camera.

Recently, very promising results have been achieved by combining the CNN feature descriptors and handcrafted feature descriptors [14], [51]. To explore the performance impact of interest point selection on this combination, we compare the visual search performance of state-of-the-art CNN feature descriptors combined with handcrafted SCFV descriptors (CDVS global feature descriptor) generated from different interest point selection strategies, as shown in Table III, where three state-of-the-art CNN feature descriptors, RMAC [13], NIP [14], [52] and Gordo [15], are utilized. In our experiments, the dimensions of NIP and RMAC descriptors are 512, and Gordo's CNN feature is 2048

TABLE II
PERFORMANCE COMPARISON FOR DIFFERENT INTEREST POINT SELECTION METHODS AT DIFFERENT
DESCRIPTOR LENGTH ON MPEG-CDVS DATASET MEASURED BY MAP

Methods/Descriptor Length	512 B	1 KB	2 KB	4 KB	8 KB	16 KB
RDC	0.8213	0.8900	0.9041	0.9426	0.9366	0.9356
Scale	0.7193	0.8409	0.8848	0.9271	0.9352	0.9355
Orientation	0.6855	0.6477	0.7811	0.8914	0.9244	0.9199
LoG	0.7770	0.8927	0.9151	0.9464	0.9476	0.9481
CurveRatio	0.7506	0.8798	0.9072	0.9351	0.9401	0.9390
DistCenter	0.8065	0.8885	0.9102	0.9430	0.9483	0.9470
Method in [50]	0.8450	0.9118	0.9229	0.9495	0.9510	0.9505
RDC_regression	0.8859	0.9189	0.9259	0.9497	0.9500	0.9496

TABLE III
PERFORMANCE COMPARISON OF CNNs AND ITS COMBINATION WITH MPEG-CDVS GLOBAL FEATURES GENERATED FROM SELECTED LOCAL FEATURE
DESCRIPTORS FROM DIFFERENT SELECTION STRATEGIES ON MPEG-CDVS DATASET MEASURED BY MAP

Descriptors/Database	Holiday	Oxford5k	Paris6k	Graphic	Paintings	Video	Buildings	Objects	Descriptor Length
SCFV [11]	0.7122	0.4291	0.4535	0.9254	0.9457	0.9775	0.6760	0.7904	4 KB
RMAC [13]	0.8623	0.6213	0.7917	0.4207	0.3930	0.5190	0.5450	0.9320	2 KB
Random + RMAC	0.8715	0.6275	0.7982	0.5485	0.6189	0.8393	0.6910	0.9486	4 KB
LoG + RMAC	0.8868	0.6357	0.8043	0.7254	0.7269	0.8838	0.7018	0.9646	4 KB
Regression + RMAC	0.8887	0.6422	0.8107	0.7817	0.7360	0.8890	0.7020	0.9660	4 KB
NIP [14]	0.8980	0.6323	0.7695	0.5057	0.4410	0.7280	0.7070	0.9630	2 KB
Random + NIP	0.8991	0.6414	0.7775	0.6276	0.5134	0.8649	0.7395	0.9752	4 KB
LoG + NIP	0.9147	0.6498	0.7840	0.7965	0.7595	0.9444	0.7712	0.9793	4 KB
Regression + NIP	0.9180	0.6518	0.7909	0.8259	0.8020	0.9440	0.7840	0.9790	4 KB
Gordo [15]	0.9147	0.8289	0.9279	0.5250	0.6917	0.8106	0.7065	0.9500	8 KB
Random + Gordo	0.9187	0.7666	0.9062	0.5970	0.7859	0.8866	0.7544	0.9688	10 KB
LoG + Gordo	0.9376	0.7762	0.9082	0.7641	0.8840	0.9482	0.8062	0.9775	10 KB
Regression + Gordo	0.9378	0.8346	0.9264	0.8025	0.8921	0.9387	0.8076	0.9735	10 KB

TABLE IV
IMAGE RETRIEVAL AND PAIRWISE MATCHING PERFORMANCE COMPARISON WHEN APPLYING DIFFERENT INTEREST POINT
SELECTION METHOD IN THE COMBINATION OF MPEG-CDVS AND NIP ON MPEG-CDVA DATASET

Descriptors/Database	Image retrieval (mAP)			Pairwise Matching, TPR@FPR = 0.01		
	large objects	small objects	scenes	large objects	small objects	scenes
Random	0.480	0.716	0.380	0.602	0.820	0.443
LoG	0.576	0.903	0.602	0.767	0.946	0.650
Regression	0.598	0.917	0.594	0.784	0.951	0.645
NIP [14]	0.719	0.817	0.843	0.832	0.928	0.953
Random + NIP	0.734	0.849	0.846	0.822	0.947	0.946
LoG + NIP	0.739	0.914	0.865	0.833	0.951	0.951
Regression + NIP	0.749	0.923	0.861	0.830	0.952	0.951

dimensional. 4 bytes are used for each dimension, leading to 2 KB representation for NIP and RMAC, 8 KB for Gordo’s CNN feature. From the results, we can see that although both CNN feature descriptors achieve good results, their overall performances are further improved by incorporating handcrafted SCFV descriptors. In Table III, the methods with “Regression + X” mean that they utilized the proposed interest point selection method to generate the MPEG-CDVS features, while the methods with “Random + X” and “LoG + X” mean that they utilized the random strategy and LoG detector to select interest points. The improvements of NIP with MPEG-CDVS global descriptors generated from the proposed interest point selection method (i.e., the Regression + NIP) are up to and

0.36 compared with NIP in terms of mAP. For Paris6k and Oxford5k, the Gordo’s CNN descriptors achieve much better performance since its model has been fine-tuned on the images with buildings, but the incorporation of MPEG-CDVS global descriptors may further improve its performance on Holidays and Oxford5k, as illustrated in the Regression+Gordo. By using different interest point selection methods we can see varying performance when combining the MPEG-CDVS global descriptors with CNN. And the regression based interest point selection outperforms the random and LoG selectors in most of the databases except for the *Objects*, where the LoG local feature selector achieves better performance. Therefore, the globally optimal interest point selector is still an open research issue,

and the most efficient lightweight image attribute and their optimal combination should be further explored in future.

To further verify the effectiveness of interest point selection, we perform the large scale visual search and pairwise matching tasks on more challenging CDVA dataset, a >1000 hours video benchmark for the emerging MPEG-CDVA standard. Table IV shows the retrieval and pair matching results using SCFV descriptors with different interest point selection strategies, i.e., random, LoG and regression. The same conclusion is reached that the multiple image statistical attributes based interest point selection method achieves better performance on both the retrieval and pairwise matching tasks. This also proves that interest point selection consistently plays an important role in compact descriptor construction for video analysis, especially for the emerging MPEG-CDVA standard.

VII. CONCLUSION

In this paper, we have explored the interest point selection problem in visual search, and formulated it as an optimization problem. The problem can be solved using regression and classification via data-driven methods. We suggested an effective and efficient approach to learning a low-complexity predictor based on matching and non-matching points. A light-weight regression approach is proposed to resolve the problem of interest point selection in MPEG-CDVS and MPEG-CDVA. Moreover, we also explored the potential performance for interest point selection with the extension to combine other types of attribute like depth cue. Finally, the combination of the CNN feature with handcrafted global feature descriptors generated from different interest point selectors is also explored, which demonstrates the performance improvement room for the state-of-the-art interest point selector. How to construct the global optimal solution with more image statistical attributes or priors will be further explored in our future work.

REFERENCES

- [1] Y.-G. Jiang, Q. Dai, T. Mei, Y. Rui, and S.-F. Chang, "Super fast event recognition in internet videos," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1174–1186, Aug. 2015.
- [2] L.-Y. Duan, J. Lin, Z. Wang, T. Huang, and W. Gao, "Weighted component hashing of binary aggregated descriptors for fast visual search," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 828–842, Jun. 2015.
- [3] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Deep video hashing," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1209–1219, Jun. 2017.
- [4] X. Zhang *et al.*, "Rate-distortion optimized sparse coding with ordered dictionary for image set compression," *IEEE Trans. Circuits Syst. Video Technol.*, Sep. 1, 2017, doi: 10.1109/TCSVT.2017.2748382.
- [5] B. Girod *et al.*, "Mobile visual search," *IEEE Signal Process. Magazine*, vol. 28, no. 4, pp. 61–76, 2011.
- [6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
- [7] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.
- [8] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.
- [9] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2548–2555.
- [10] L.-Y. Duan *et al.*, "Overview of the MPEG-CDVS standard," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 179–194, Jan. 2016.
- [11] J. Lin *et al.*, "Rate-adaptive compact fisher codes for mobile visual search," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 195–198, Feb. 2014.
- [12] L.-Y. Duan *et al.*, "Compact descriptors for video analysis: The emerging MPEG standard," arXiv:1704.08141, to be published, 2017.
- [13] G. Toliás, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," presented at the Int. Conf. Learn. Represent., San Juan, Puerto Rico, May 2–4, 2016.
- [14] Y. Lou *et al.*, "Compact deep invariant descriptors for video retrieval," in *Proc. Data Compression Conf.*, 2017, pp. 420–429.
- [15] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *Int. J. Comput. Vis.*, vol. 124, no. 2, pp. 237–254, 2017.
- [16] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [17] L.-T. Cheok, J. Song, and K. Park, *CDVS: Telecom Italia's Response to CE1 ? Interest Point Detection* ISO/IEC JTC1/SC29/WG11/M23822, Feb. 2012.
- [18] S. Buoncompagni, D. Maio, D. Maltoni, and S. Papi, "Saliency-based keypoint selection for fast object detection and matching," *Pattern Recognit. Lett.*, vol. 62, pp. 32–40, 2015.
- [19] S. Buoncompagni, D. Maio, D. Maltoni, and S. Papi, "Saliency-based keypoint reduction for augmented-reality applications in smart cities," in *Proc. Int. Conf. Image Anal. Process.*, Berlin, Germany: Springer, 2015, pp. 209–217.
- [20] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 778–792.
- [21] P. Mukherjee, S. Srivastava, and B. Lall, "Salient keypoint selection for object representation," in *Proc. 22nd Nat. Conf. Commun.*, 2016, pp. 1–6.
- [22] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "Kaze features," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Germany: Springer, 2012, pp. 214–227.
- [23] S. Li and A. Calway, "RGBD localisation using pairwise geometry and concise key point sets," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 6374–6379.
- [24] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *Proc. Comput. Vis.*, 2010, pp. 748–761.
- [25] P. Turcot and D. G. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops*, 2009, pp. 2109–2116.
- [26] G. Toliás and H. Jégou, "Visual query expansion with or without geometry: Refining local descriptors by feature aggregation," *Pattern Recognit.*, vol. 47, no. 10, pp. 3466–3476, 2014.
- [27] G. Dorkó and C. Schmid, "Selection of scale-invariant parts for object class recognition," in *Proc. IEEE 9th Int. Conf. Comput. Vis.*, 2003, vol. 1, pp. 634–640.
- [28] M. F. Demirci and S. Kacka, "Object recognition by distortion-free graph embedding and random forest," in *Proc. IEEE 10th Int. Conf. Semantic Comput.*, 2016, pp. 17–23.
- [29] W. Zhou, H. Fu, and X. An, "A classification-based visual odometry approach," in *Proc. IEEE 8th Int. Conf. Intell. Human-Mach. Syst. Cybern.*, 2016, vol. 2, pp. 85–89.
- [30] H. Jin Kim, E. Dunn, and J.-M. Frahm, "Predicting good features for image GEO-localization using per-bundle VLAD," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1170–1178.
- [31] M. Dymczyk, E. Stumm, J. Nieto, R. Siegwart, and I. Gilitschenski, "Will it last? Learning stable features for long-term visual localization," in *Proc. IEEE 4th Int. 3D Vis.*, 2016, pp. 572–581.
- [32] S. D. Bhattacharjee, J. Yuan, Y.-P. Tan, and L.-Y. Duan, "Query-adaptive small object search using object proposals and shape-aware descriptors," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 726–737, Apr. 2016.
- [33] T. Yu, Y. Wu, S. D. Bhattacharjee, and J. Yuan, "Efficient object instance search using fuzzy objects matching," in *Proc. 31st Conf. Assoc. Advancement Artif. Intell.*, 2017, pp. 4320–4326.
- [34] T. Yu, Y. Wu, and J. Yuan, "Hope: Hierarchical object prototype encoding for efficient object instance search in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3195–3204.
- [35] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, no. Mar, pp. 1157–1182, 2003.
- [36] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
- [37] Y. Luo, Y. Wen, D. Tao, J. Gui, and C. Xu, "Large margin multi-modal multi-task feature extraction for image classification," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 414–427, Jan. 2016.

- [38] Y. Luo *et al.*, “Multiview vector-valued manifold regularization for multilabel image classification,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 709–722, May 2013.
- [39] Y. Luo, T. Liu, D. Tao, and C. Xu, “Multiview matrix completion for multilabel image classification,” *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2355–2368, Aug. 2015.
- [40] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen, “Tensor canonical correlation analysis for multi-view dimension reduction,” *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 3111–3124, Nov. 2015.
- [41] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
- [42] A. Liaw *et al.*, “Classification and regression by randomforest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002.
- [43] *Evaluation Framework for Compact Descriptors for Visual Search* ISO/IEC JTC1/SC29/WG11/N12202, Jul. 2011.
- [44] H. Jegou, M. Douze, and C. Schmid, “Hamming embedding and weak geometric consistency for large scale image search,” in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 304–317.
- [45] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
- [46] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Lost in quantization: Improving particular object retrieval in large scale image databases,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
- [47] Z. Liu, L.-Y. Duan, J. Chen, and T. Huang, “Depth-based local feature selection for mobile visual search,” in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 276–280.
- [48] M. J. Huiskes, B. Thomee, and M. S. Lew, “New trends and ideas in visual concept detection: The MIR flickr retrieval evaluation initiative,” in *Proc. ACM Int. Conf. Multimedia Inf. Retrieval*, 2010, pp. 527–536.
- [49] G. Francini, S. Lepsoy, and M. Balestri, *CDVS: Telecom Italia’s Response to CEI Interest Point Detection* ISO/IEC JTC1/SC29/WG11/M30256, Jul. 2013.
- [50] W. Hartmann, M. Havlena, and K. Schindler, “Predicting matchability,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 9–16.
- [51] L. Duan *et al.*, “Fast MPEG-CDVS encoder with GPU-CPU hybrid computing,” *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2201–2216, 2018.
- [52] J. Lin *et al.*, “Hnlp: Compact deep invariant representations for video matching, localization, and retrieval,” *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 1968–1983, Sep. 2017.



Feng Gao received the B.S. degree in computer science from University College London, London, U.K., in 2007, and the Ph.D. degree in computer science from Peking University, Beijing, China, in 2018. He is currently a Postdoctoral Research Fellow at the Future Laboratory, Tsinghua University, Beijing, China. His research interests include the intersection of computer science and art, including but not limited to artificial intelligence and painting art, deep learning, painting robot, etc.



Xinfeng Zhang (M’16) received the B.S. degree in computer science from Hebei University of Technology, Tianjin, China, in 2007, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2014. From July 2014 to October 2017, he was a Research Fellow with Nanyang Technological University, Singapore. He is currently a Postdoctoral Fellow with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA. His research interests include image and video processing, and image and video compression.



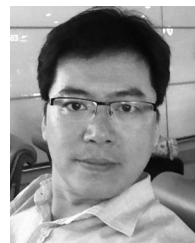
Yicheng Huang received the Bachelor’s degree in computer science and technology from Peking University, Beijing, China, in 2015. He is currently working toward the M.S. degree at the School of Electrical Engineering and Computer Science, Peking University, Beijing, China. His current research interests include large-scale image retrieval and fast nearest neighbor search.



Yong Luo received the B.E. degree in computer science from the Northwestern Polytechnical University, Xi’an, China, in 2009, and the D.Sc. degree from the School of Electronics Engineering and Computer Science, Peking University, Beijing, China, in 2014. He is currently a Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He was a visiting student with the School of Computer Engineering, Nanyang Technological University, and the Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW, Australia. He has authored several scientific articles at top venues including IEEE T-NNLS, IEEE T-IP, IEEE T-KDE, IJCAI, and AAAI. His research interests include machine learning and data mining with applications to visual information understanding and analysis. Dr. Luo was the recipient of the IEEE Globecom 2016 Best Paper Award, and was nominated for the IJCAI 2017 Distinguished Best Paper Award.



Xiaoming Li (SM’03) is currently a Professor of computer science and technology and the Director with the Institute of Network Computing and Information Systems, Peking University, Beijing, China. He has been leading the effort of developing a Chinese search engine (Tianwang) since 1999. He is also the Founder of the Chinese Web Archive (Web InfoMall). His current research interests include search engine and web mining.



Ling-Yu Duan (M’06) is a Full Professor with the National Engineering Laboratory of Video Technology, Peking University, Beijing, China, and was the Associate Director of the Rapid-Rich Object Search Laboratory (ROSE), a joint lab between Nanyang Technological University (NTU), Singapore, and Peking University (PKU), China, since 2012. His research interests include multimedia indexing, search, and retrieval, mobile visual search, visual feature coding, and video analytics, etc. Prof. Duan was a Co-Editor of MPEG Compact Descriptor for Visual Search Standard (ISO/IEC 15938-13), and is a Co-Chair of MPEG Compact Descriptor for Video Analytics (CDVA). He is currently an Associate Editor of *ACM Transactions on Intelligent Systems and Technology* and *ACM Transactions on Multimedia Computing, Communications, and Applications*. He was the recipient of the EURASIP Journal on Image and Video Processing Best Paper Award in 2015, the Ministry of Education Technology Invention Award (First Prize) in 2016, the National Technology Invention Award (Second Prize) in 2017, the China Patent Award for Excellence (2017), and the National Information Technology Standardization Technical Committee “Standardization Work Outstanding Person” Award in 2015.