DEEP REGIONAL FEATURE POOLING FOR VIDEO MATCHING

Yan Bai^{*,1,2}, *Jie Lin*^{*,3}, *Vijay Chandrasekhar*^{*,3,4},
Yihang Lou^{1,2}, *Shiqi Wang*^{5}, *Ling-Yu Duan*^{1}, *Tiejun Huang*^{1}, *Alex Kot*^{4}

[1]Institute of Digital Media, Peking University, Beijing, China
[2]SECE of Shenzhen Graduate School, Peking University, Shenzhen, China
[3]Institute for Infocomm Research, A*STAR, Singapore
[4]Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore
[5]Department of Computer Science, City University of Hong Kong

ABSTRACT

In this work, we study the problem of deep global descriptors for video matching with *regional* feature *pooling*. We aim to analyze the joint effect of ROI (Region of Interest) size and pooling moment on video matching performance. To this end, we propose to mathematically model the distribution of video matching function with a pooling function nested in. Matching performance can be estimated by the separability of these class-conditional distributions between matching and non-matching pairs. Empirical studies on the challenging MPEG CDVA dataset demonstrate that performance trends are consistent with the estimation and experimental results, though the theoretical model is largely simplified compared to video matching and retrieval in practice.

Index Terms— Convolutional Neural Networks, Pooling, Global Descriptor, Video Matching, Video Retrieval

1. INTRODUCTION

Recent years have witnessed a remarkable growth of interest in video retrieval, which refers to searching for videos representing the same object or scene as the one depicted in a query video. The main challenge is to develop compact and discriminative video feature representations towards highly efficient and effective video matching and retrieval. To this end, the Motion Picture Experts Group (MPEG) published the standardization of Compact Descriptors for Visual Search (CDVS) [1] in 2015, which came up with a normative bitstream of standardized descriptors for mobile visual search [2] and augmented reality applications [3]. State-of-the-art handcrafted descriptors (VLAD [4], Fisher vectors (FV) [5] and compact FV [6]) built on local invariant SIFTs have been adopted in CDVS as global descriptors. Very recently, MPEG has started a standardization effort titled Compact Descriptors for Video Analysis (CDVA) [7], to extend the CDVS standard to video analysis.

To deal with content redundancy along temporal dimension, the latest CDVA Experimental Model (CXM) [8] casts video retrieval into keyframe based image retrieval. Frame-level descriptors are extracted from a subset of detected keyframes in videos. Video matching is computed by aggregating matching results with keyframe-level descriptors. Besides, video-level descriptors aggregated from frame descriptors over time have also been explored for video retrieval [9,10]. In this work, we focus on keyframe based approaches.

Though handcrafted descriptors have achieved great success in the CDVS standard [1] and CDVA experimental model, many recent papers [11–16] have shown the advantage of deep global descriptors for image retrieval, which can be attributed to the remarkable success of Convolutional Neural Networks (CNN) [17, 18]. In particular, state-of-the-art deep global descriptors R-MAC [15] computes the max over a set of Region-of-Interest (ROI) cropped from each feature map output by intermediate convolutional layer, followed by the average of these regional max-pooled features. Results show that R-MAC offers remarkable improvements over other deep global descriptors like MAC [15] and SPoC [13], while maintaining the same dimensionality.

Previous work has focused on how to build high-quality regional pooled descriptors. Here, we aim to study the relationship between regional feature pooling and video matching. We make the following contributions,

- We propose a simple model to mathematically analyze how deep regional feature pooling affect video matching performance. In particular, we are interested in key parameters, i.e. ROI size and pooling moment, which jointly affect matching performance.
- Systematical and practical evaluations verify that model estimations on the performance order of various ROI sizes and pooling moments are largely consistent with empirical observations, for both pairwise video matching and video retrieval.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the keyframe based video matching framework using deep regional pooled global descriptors. Section 3 presents theoretical analysis of the matching framework. We systematically evaluate the theoretical analysis in Section 4, and summarize the paper in Section 5.

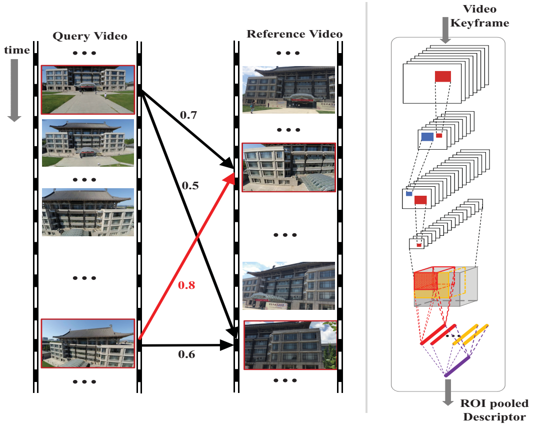
2. VIDEO MATCHING AND RETRIEVAL

2.1. Video Matching and Retrieval Pipeline

Fig. 1 (a) illustrates keyframe based video matching. First, color histogram comparison is applied to identify keyframes in both query and reference videos. Keyframe detection can reduces the temporal redundancy in videos, and reduces the matching complexity between video pairs. Subsequently, video matching is performed by comparing global descriptors extracted from keyframe pair <query keyframe, reference keyframe>. We denote query video $\boldsymbol{X} = \{X_1, ..., X_{N_x}\}$ and reference video $\boldsymbol{Y} = \{Y_1, ..., Y_{N_y}\}$,

$*$ Yan Bai, Jie Lin, Vijay Chandrasekhar contributed equally.
Ling-Yu Duan is the corresponding author.

978-1-5090-2175-8/17/$31.00 ©2017 IEEE



(a) Keyframe based Video Matching (b) Deep Descriptor Extraction

Fig. 1. (a) Keyframe based video matching. (b) Deep regional feature pooling over activation feature maps extracted from CNN architecture.

where X and Y denote keyframes. N_x and N_y are the numbers of detected keyframes in the query and reference videos, respectively. Each keyframe in $\boldsymbol{X}$ is compared to all keyframes in $\boldsymbol{Y}$. The video-level similarity $K(\boldsymbol{X}, \boldsymbol{Y})$ is defined as the largest matching score among all keyframe-level similarity scores.

$$K(\boldsymbol{X}, \boldsymbol{Y}) = \max_{X \in \boldsymbol{X}, Y \in \boldsymbol{Y}} k(f(X), f(Y)), \tag{1}$$

where $f(X)$ denotes a global descriptor extracted from keyframe X, and $k(\cdot, \cdot)$ represents a matching function.

For video retrieval, the video-level similarity between the query and each candidate database video is obtained following the same principle as pairwise video matching. The top ranked candidate database videos are returned for each query video.

2.2. Deep Regional Feature Pooling

Fig. 1 (b) introduces the global descriptor extraction pipeline for keyframes. Considering a keyframe image as input to CNN architecture, we describe it with the feature maps extracted from intermediate convolutional layer, denoted as $X = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_C\}$, where $\boldsymbol{x}_c$ represents feature map of size $W \times H$, C is the number of channels. ROIs can be sampled from feature maps with varied sizes and strides. For the c^{th} channel, regional feature pooling is first computed by α_s-norm pooling over ROIs sampled from the c^{th} feature map, followed by α_n-norm pooling,

$$f(\boldsymbol{x}_c) = f_{\alpha_n}^{N_{ROI}}(f_{\alpha_s}^{S_{ROI}}(\boldsymbol{x}_c)), \tag{2}$$

where S_{ROI} denotes ROI of size $w \times h$, with $w < W$ and $h < H$. N_{ROI} represents the number of sampled ROIs. $\alpha_s \in \{1, +\infty\}$ and $\alpha_n \in \{1, +\infty\}$ denote pooling moments. For instance, $\alpha = 1$ represents average pooling [13], while $\alpha \to +\infty$ denotes max pooling [15],

$$f_{\alpha}^{N}(\hat{\boldsymbol{x}}) = (\frac{1}{N}\sum_{i=1}^{N}(\hat{\boldsymbol{x}}_i)^{\alpha})^{\frac{1}{\alpha}}. \tag{3}$$

The C-dimensional global descriptor $f(X)$ is formed by concatenating $\{f(\boldsymbol{x}_1), ..., f(\boldsymbol{x}_C)\}$ for all channels.

Following [13], we consider a simple match kernel to compute the similarity between keyframes X and Y with their descriptors $f(X)$ and $f(Y)$,

$$k(f(X), f(Y)) = \beta(X)\beta(Y)\sum_{c=1}^{C} k(f(\boldsymbol{x}_c), f(\boldsymbol{y}_c)), \tag{4}$$

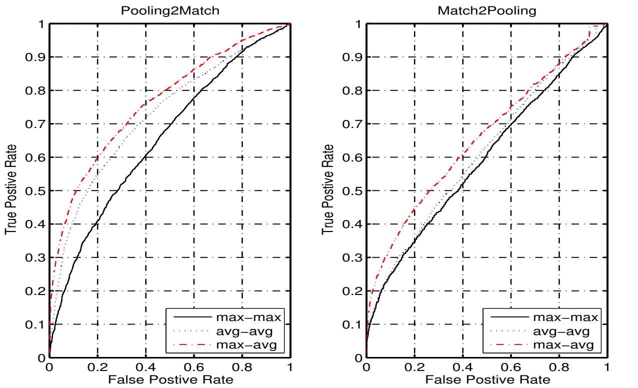


Fig. 2. Comparisons of pairwise image matching performance trends between Pooling2Match and Match2Pooling with different pooling strategies, in terms of ROC curve on the challenging Stanford Mobile Visual Search (SMVS) dataset.

where $k(f(\boldsymbol{x}_c), f(\boldsymbol{y}_c)) =< f(\boldsymbol{x}_c), f(\boldsymbol{y}_c) >$ is the scalar product of pooled descriptors, $\beta(.)$ is a normalization term computed as $\beta(X) = (\sum_{c=1}^{C} k(f(\boldsymbol{x}_c), f(\boldsymbol{x}_c)))^{-\frac{1}{2}}$. Eq. 4 refers to cosine similarity by accumulating the scalar product of normalized pooled features for each channel. Deep descriptors can be further improved by post-processing techniques such as PCA whitening [13, 15].

3. THEORETICAL ANALYSIS

Signal-to-Noise Ratio (SNR) measurement. We aim to theoretically analyze the effect of ROI pooling on keyframe-level matching performance. Inspired by Boureau et al. [19], keyframe-level matching can be regarded as a two-class classification problem (matching and non-matching pairs). As such, we first estimate the distribution (e.g. Binomial distribution with mean μ and variance σ^2) of matching function with ROI pooling function nested in. Then, we adopt Signal-to-Noise Ratio (SNR) to measure the separability of these two class-conditional distributions,

$$SNR = \frac{|\dot{\mu}_{k(.,.)} - \ddot{\mu}_{k(.,.)}|}{\dot{\sigma}_{k(.,.)} + \ddot{\sigma}_{k(.,.)}}, \tag{5}$$

where $\dot{\mu}_{k(.,.)}$ ($\ddot{\mu}_{k(.,.)}$) and $\dot{\sigma}_{k(.,.)}$ ($\ddot{\sigma}_{k(.,.)}$) denote mean and standard deviation of Eq. 4 for matching (non-matching) set, respectively. The larger the SNR, the better the separability between the two classes, implying better matching performance. In this work, we estimate video matching performance by simply analyzing the matching function $k(.,.)$ on per-channel basis (i.e., equal contribution for all channels).

Matching function. In Eq. 4, video matching is performed in two-stage: ROI pooling function (Eq. 2) over raw feature maps, followed by matching function $k(.,.)$ with pooled features. We term the two-stage pipeline as Pooling2Match. It is non-trivial to derive closed-form solution for estimating distribution of the composite function Pooling2Match. Thus, we propose a simplified alternative named Match2Pooling, which firstly performs matching between raw feature maps, followed by ROI pooling. In particular, given a matching/non-matching feature map $\boldsymbol{x}, \boldsymbol{y}$, a new feature map $\boldsymbol{g}$ is generated, a neuron/feature is 1 if and only if it is activated on both images $\boldsymbol{x}$ and $\boldsymbol{y}$, otherwise, 0. Finally, the pooled feature $f(\boldsymbol{g})$ reflects the similarity of a image pair (the larger, the more similar).

Recall that our objective is to study how video matching performance is affected by ROI pooling function. We expect matching performance order of various pooling moments is the same be-

Table 1. Parameters S_{ROI}, N_{ROI} for ROI pooling function with feature map of size $W \times H = 20 \times 15$.

$S_{ROI} = w \times h$ (w=h)	N_{ROI}	$S_{ROI} * N_{ROI}$
15 x 15	3	675
10 x 10	6	600
7 x 7	12	588
3 x 3	65	585

tween Pooling2Match and Match2Pooling. To this end, we design practical pairwise image matching performance to validate the hypothesis empirically. We construct matching and non-matching image pairs from the challenging Stanford Mobile Visual Search (S-MVS) dataset [20]. Both sets contain 3000 pairs. With input video keyframe size 640×480, feature maps of size 20×15 are extracted from the last pooling layer of VGG16 [18] pre-trained on ImageNet dataset. Feature maps are binarized (i.e. 1 if neuron activated, otherwise, 0), and ROIs sampling follows R-MAC [15]. Matching results of Pooling2Match and Match2Pooling are normalized to $[0, 1]$ for fair comparison.

Fig. 2 shows ROC curve comparing pairwise image matching between Pooling2Match and Match2Pooling, with different pooling strategies (Max-Max, Avg-Avg and Max-Avg). We observe that matching performance order of Match2Pooling is largely consistent with Pooling2Match, i.e. Max-Avg performs the best, while Max-Max is the worst.

ROI pooling function. Next, we introduce how to estimate the distribution of ROI pooling function Eq. 2 over new feature map $\boldsymbol{g}$, which depends on parameters S_{ROI}, N_{ROI} and pooling moments (α_s, α_n). We devote our analysis to these key parameters. Without loss of generality, we assume all binary spatial bins of $\boldsymbol{g}$ follow i.i.d. Bernoulli distribution $g_i \sim \text{Bern}(p)$, where p denotes the activation probability. In particular, p is specified as $\dot{p}$ and $\ddot{p}$ for matching and non-matching sets, respectively.

First, we consider the distribution estimation of Eq.3. For average pooling (α=1), the average of N i.i.d Bernoulli variables follows a scaled-down version of Binomial distribution with mean p and variance $\frac{p(1-p)}{N}$. For max pooling ($\alpha \to +\infty$), the maximum of N i.i.d Bernoulli variables still follows a Bernoulli distribution with mean $(1-(1-p)^N)$ and variance $(1-(1-p)^N)(1-p)^N$.

Then, we derive the distribution of ROI pooling function Eq. 2 in a similar way. We are interested in the distribution of Eq. 2 with $(\alpha_s, \alpha_n) \in \{(Max-Max), (Avg-Avg), (Max-Avg)\}$. Max-Max follows Bernoulli distribution with

$$\mu_{\boldsymbol{g}} = e(e(p, S_{ROI}), N_{ROI}), \quad \sigma_{\boldsymbol{g}}^2 = \mu_{\boldsymbol{g}}(1-\mu_{\boldsymbol{g}}) \tag{6}$$

Avg-Avg follows Binomial distribution with

$$\mu_{\boldsymbol{g}} = p, \quad \sigma_{\boldsymbol{g}}^2 = \frac{\mu_{\boldsymbol{g}}(1-\mu_{\boldsymbol{g}})}{S_{ROI} * N_{ROI}} \tag{7}$$

Max-Avg follows Binomial distribution with

$$\mu_{\boldsymbol{g}} = e(p, S_{ROI}), \quad \sigma_{\boldsymbol{g}}^2 = \frac{\mu_{\boldsymbol{g}}(1-\mu_{\boldsymbol{g}})}{N_{ROI}} \tag{8}$$

Where function $e(.,.)$ is defined as $e(p, m) = 1-(1-p)^m$.

Visualization and Estimations. In summary, we first estimate mean $\mu_{\boldsymbol{g}}$ and variance $\sigma_{\boldsymbol{g}}^2$ from the distribution of ROI pooling function by Eq. 6,7,8. Then, we compute SNR values following Eq. 5. Note that the former two steps are performed independently for matching and non-matching sets.

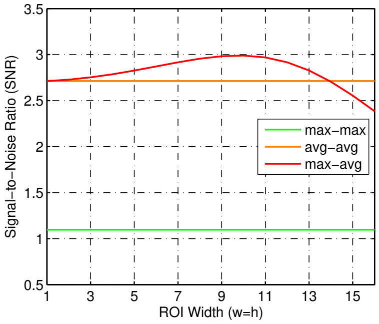


Fig. 3. Video matching performance estimation by measuring the Signal-to-Noise Ratio (SNR) as a function of ROI size, for Max-Max, Avg-Avg, and Max-Avg.

To visualize SNR values with varied S_{ROI}, N_{ROI}, we simplify our analysis by constraining that $S_{ROI} * N_{ROI} \sim 2 * W * H$ (i.e. each spatial bin is roughly scanned twice), in which $S_{ROI} = w \times h$ is with $w = h$ (i.e. square size ROI), following R-MAC [15]. Table 1 presents examples for S_{ROI}, N_{ROI} with feature map of size $W \times H = 20 \times 15$. Without loss of generality, we empirically set $\dot{p} = 0.02$ and $\ddot{p} = 0.001$, which are statistically computed from matching/non-matching sets. Correspondingly, we plot SNR as a function of ROI size ($w = h$) for Max-Max, Avg-Avg and Max-Avg, as shown in Fig. 3, which leads us to the main estimations:

- The overall performance order of ROI pooling moments is that Max-Max is worse than Avg-Avg, while Max-Avg is the best.
- The performance of Max-Max and Avg-Avg is independent on ROI sizes, while Max-Avg is sensitive to it.

Next, we evaluate the quality of our theoretical analysis with both pairwise video matching and video retrieval experiments on the challenging MPEG CDVA dataset.

4. EVALUATIONS

4.1. Datasets

We conduct evaluations on the MPEG CDVA datasets, which contain 9974 query and 5127 reference videos[1]. The videos have durations from 1 sec to 1+ min of 25∼30 fps. These videos depict 796 items of interest across three different categories, including 489 large **Landmarks** (5224 queries), 71 **Scenes** (915 queries) and 236 **Objects** (3835 queries, e.g. products). To evaluate video retrieval performance at large scale, we combine the reference videos with a set of user-generated and broadcast videos as distractors, which consist of content unrelated to the items of interest. There are 14537 distractor videos with more than 1000 hours data. To evaluate pairwise video matching, 4693 matching video pairs and 46911 non-matching video pairs are constructed from the query and reference videos.

We report pairwise matching results in terms of True Positive Rate (TPR), at 1% False Positive Rate (FPR). Video retrieval performance is evaluated by mean Average Precision (mAP).

4.2. Implementation Notes

In the CDVA evaluation framework, there are on average 1∼2 keyframes detected per second from raw videos (25∼30 fps). We

[1] The MPEG CDVA dataset and evaluation framework are available upon request at http://www.cldatlas.com/cdva/dataset.html

Table 2. Pairwise video matching with different combinations of ROI size (w=h) and pooling moment, on matching/non-matching video datasets. No PCA whitening is performed.

ROI size	Max-Max	Avg-Avg	Max-Avg
w = 15	71.9	76.8	73.7
w = 7			80.7
w = 3			77.6

Table 3. Small-scale video retrieval comparison (mAP) with different pooling moments. ROI size is fixed (w=h=10). No PCA whitening is performed.

Pool Op.	Landmarks	Scenes	Objects
Max-Max	62.8	79.6	70.5
Avg-Avg	65.3	82.3	69.0
Max-Avg	**67.9**	**83.4**	**73.8**

verify the theoretical analysis with the widely used VGG16 architecture [18], which is pre-trained on ImageNet ILSVRC classification data set. We resize all video keyframes to VGA size (640×480) images as the inputs of CNN, and extract feature maps of size 20×15×512 from the last pooling layer (i.e. pool5). Subsequently, a 512-dimensional deep descriptor is derived following Eq.2. For P-CA whitening, we randomly sample 40K frames from the distractor videos for learning the PCA projection.

4.3. Evaluation on Pairwise Video Matching

Table 2 reports pairwise matching performance in terms of True Positive Rate with False Positive Rate equals to 1%, for pooling moments Max-Max, Avg-Avg and Max-Avg, with ROI size $w \in \{15, 7, 3\}$. First, as shown in Table 2, matching performances of Max-Max and Avg-Avg do not vary with ROI size. Second, as ROI size reduced from 15 to 3, matching performance for Max-Avg rises first and drops later. These performance trends are consistent with the model estimations (Fig. 3). One may note that the model differs from practical matching experiments in that: (1) the former performs Pooling2Match, as opposed to Match2Pooling used in the latter. (2) The former accumulates similarities contributed by all keyframes and all channels, while the latter is on per-keyframe per-channel basis under i.i.d assumptions. (3) The former works on real-valued feature maps, rather than binary feature maps for the latter case.

4.4. Evaluation on Video Retreival

Pooling moment. We further design video retrieval experiments to evaluate the quality of the model. Table 3 studies the effect of pooling moments when ROI size is fixed, on all categories of CDVA dataset. We observe the performance order of pooling moments is in line with the estimations in Fig. 3, i.e. Avg-Avg is superior to Max-Max, and Max-Avg performs the best.

ROI size. Table 4 explores the effect of ROI size when pooling moment is fixed (Max-Avg). First, similar to the observations in Table 2, retrieval performance (mAP) increases then decreases as ROI size ranging from 15 to 3. Second, the best performing ROI size depends on the type of data category, i.e. w = 7 is the best for Landmarks and Scenes, while w = 10 for Objects. Also, there is a quick drop in mAP on Objects, i.e. from 71.4% (w = 7) to 58.6% (w = 3). This is probably due to the fact that larger ROI size is desirable for small objects to include contextual info, while it is not that vital for large landmarks and scenes.

Table 4. Small-scale video retrieval comparison (mAP) with different ROI sizes (w=h). Pooling moment is fixed (Max-Avg). No PCA whitening is performed.

ROI size	Landmarks	Scenes	Objects
w = 15	60.5	78.6	73.1
w = 10	67.9	83.4	**73.8**
w = 7	**69.2**	**84.3**	71.4
w = 3	64.1	81.3	58.6

Table 5. Large-scale video retrieval comparison of combination (average) of multi ROI sizes with state-of-the-art. The former (latter) number in each cell represents performance without (with) PCA whitening.

Method	Landmarks	Scenes	Objects
CXM [8]	61.4	63.0	92.6
MAC [15]	57.8/61.9	77.4/76.2	70.0/71.8
SPoC [13]	64.0/69.1	82.9/84.0	64.8/70.3
CroW [14]	62.3/63.9	79.2/78.4	71.9/72.0
Max-Avg-Multi	**69.4/74.6**	**84.4/87.3**	**73.8/78.2**

Combination of multi ROI sizes. As the best performing ROI size changes across categories, it motivates us to combine multi ROI sizes by averaging their pooled descriptors (termed as Max-Avg-Multi), similar to R-MAC [15]. Table 5 presents the comparison of Max-Avg-Multi against state-of-the-art deep and handcrafted descriptors. Handcrafted descriptors are compact Fisher vectors (FV) built upon SIFT for the initial search, followed by geometric reranking with compressed local SIFTs, which have been adopted by the ongoing CDVA standard (terms as CXM). For deep descriptors, we report retrieval performance without (the former number) and with (the latter number) PCA whitening.

We make the following observations: (1) Compared to numbers shown in Table 4, Max-Avg-Multi without PCA whitening achieves the best performance on all three categories, which verifies the effectiveness of combining multi ROI sizes. (2) PCA whitening improves the performance of deep descriptors in most cases. (3) Max-Avg-Multi performs consistently better than other deep descriptors. (4) Overall, deep descriptors outperform handcrafted descriptors by a large margin on Landmarks and Scenes, but underperform on Objects, This is reasonable as handcrafted descriptors based on SIFT are more robust to scale and rotation changes of rigid objects in the 2D plane, compared to CNN descriptors [16].

5. SUMMARY

In this work, we study the problem of deep regional feature pooling for video matching. In particular, we are interested in the joint effect of ROI size and pooling moment on video matching performance. We derive a mathematical model to measure matching performance as the separability of distributions of matching function respectively estimated from matching and non-matching sets. Experiments show that the model estimations are well-aligned with empirical results on both pairwise video matching and video retrieval. In future work, an in-depth study on the theoretical relationship between Pooling2Match and Match2Pooling is valuable to further clarify and enhance the model proposed in this paper.

6. ACKNOWLEDGMENTS

This work was partially supported by grants from National Natural Science Foundation of China (U1611461, 61661146005) and National Hightech R&D Program of China (2015AA016302).

7. REFERENCES

[1] Ling-Yu Duan, Vijay Chandrasekhar, Jie Chen, Jie Lin, Zhe Wang, Tiejun Huang, Bernd Girod, and Wen Gao, "Overview of the MPEG-CDVS standard," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 179–194, 2016.

[2] Bernd Girod, Vijay Chandrasekhar, David M Chen, Ngai-Man Cheung, Radek Grzeszczuk, Yuriy Reznik, Gabriel Takacs, Sam S Tsai, and Ramakrishna Vedantham, "Mobile visual search," *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 61–76, 2011.

[3] Mina Makar, Vijay Chandrasekhar, S Tsai, David Chen, and Bernd Girod, "Interframe coding of feature descriptors for mobile augmented reality," *IEEE Transactions on Image Processing*, 2014.

[4] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3304–3311.

[5] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier, "Large-scale image retrieval with compressed fisher vectors," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[6] Jie Lin, Ling-Yu Duan, Yaping Huang, Siwei Luo, Tiejun Huang, and Wen Gao, "Rate-adaptive compact fisher codes for mobile visual search," *IEEE Signal Processing Letters*, vol. 21, no. 2, pp. 195–198, 2014.

[7] "Call for Proposals for Compact Descriptors for Video Analysis (CDVA)-Search and Retrieval," *ISO/IEC JTC1/SC29/WG11/N15339*, 2015.

[8] Werner Bailer Massimo Balestri, Miroslaw Bober, "Cdva experimentation model (cxm) 0.2," *ISO/IEC JTC1/SC29/WG11/W16274*, 2015.

[9] André Araujo, Jason Chaves, Roland Angst, and Bernd Girod, "Temporal aggregation for large-scale query-by-image video retrieval," in *IEEE International Conference on Image Processing*. IEEE, 2015, pp. 1519–1522.

[10] Zhongwen Xu, Yi Yang, and Alex G Hauptmann, "A discriminative cnn video representation for event detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1798–1807.

[11] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky, "Neural codes for image retrieval," in *European Conference on Computer Vision*, 2014.

[12] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson, "From generic to specific deep representations for visual recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015.

[13] Artem Babenko and Victor Lempitsky, "Aggregating local deep features for image retrieval," in *IEEE International Conference on Computer Vision*, 2015.

[14] Yannis Kalantidis, Clayton Mellina, and Simon Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *arXiv:1512.04065*, 2015.

[15] Giorgos Tolias, Ronan Sicre, and Hervé Jégou, "Particular object retrieval with integral max-pooling of cnn activations," in *arXiv:1511.05879*, 2015.

[16] Vijay Chandrasekhar, Jie Lin, and Olivier Morère, "A practical guide to cnns and fisher vectors for image instance retrieval," *Signal Processing*, vol. 128, pp. 426–439, 2016.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012.

[18] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *arXiv:1409.1556*, 2014.

[19] Y-Lan Boureau, Jean Ponce, and Yann LeCun, "A theoretical analysis of feature pooling in vision algorithms," in *International Conference on Machine learning (ICML)*, 2010.

[20] V. Chandrasekhar, D.M.Chen, S.S.Tsai, N.M.Cheung, H.Chen, G.Takacs, Y.Reznik, R.Vedantham, R.Grzeszczuk, J.Back, and B.Girod, "Stanford Mobile Visual Search Data Set," in *ACM Multimedia Systems Conference (MMSys)*, 2011.