BURST-SURVIVE TEMPORAL MATCHING KERNEL WITH FIBONACCI PERIODS

Fan Yang, Shin'ichi Satoh

The University of Tokyo, National Institute of Informatics
{yang, satoh}@nii.ac.jp

ABSTRACT

In this paper we present a novel approach to improve temporal matching kernel (TMK) for video retrieval tasks. TMK has the ability to align videos during retrieval, but provides little to none retrieval performance improvement over baseline methods. We discovered that TMK cannot discriminate between a true match case in which two videos have long, consecutive segments of similar frames and a false match case in which two videos contain non-consecutive segments of randomly similar frames. Our proposed burst-survive temporal matching kernel adopts a novel shuffle strategy to rule out false match cases, with the assistance of multiple periods selected from Fibonacci series. As a result, we achieved significant performance improvement on the EVVE dataset.

Index Terms— Video retrieval, video alignment

1. INTRODUCTION

Video retrieval [1, 2] and alignment [3, 4, 5] are both important tools in the family of video analyses. Many works [6, 7] targeting the video retrieval or its subtask, video copy detection, attempted to create distinctive video descriptors, viewing videos as a bag of frames. Recently, Gao et al. [8] tried to exploit the temporal axis to compress redundant spatial information shared by adjacent frames, and achieved significant improvement on retrieval performance. However, these works focused on video retrieval without considering alignment. Douze et al. [3] adopted the Hough voting scheme to provide a precise temporal alignment approach for video copy detection, causing the computational complexity to be quadratic to the length of the video. Revaud et al. [4] proposed circulant temporal encoding (CTE), where fast Fourier transform (FFT) is used to create temporal encodings for videos. Despite the difficulty of handling complex numbers caused by FFT, CTE accelerated the video alignment significantly.

In recent works [5, 9, 10], temporal matching kernel (TMK) is developed to tackle video retrieval and alignment simultaneously. Instead of using FFT, the timestamp of each frame is encoded by Fourier series coefficients (real numbers) and embedded into the frame's visual representation. Video descriptors are then aggregated by their timestamp-embedded frame descriptors. With these video descriptors, TMK is able to compare video pairs by their frames with a simultaneous temporal consistency check. However, TMK performs worse than expected: 0.1% [5] higher and 1.3% [10] lower than the baseline, obtained by using the average of frame descriptors as video representations when using mAP as the evaluation. Baraldi et al. [10] equipped TMK with learned temporal layers by their proposed learning to align and match videos (LAMV) approach, and achieved 0.7% improvement over the baseline in terms of mAP. These results lead us to the question: what causes TMK to merely match the baseline level? From our observations, we found that the main problem comes from the way TMK computes video-wise similarity scores.

Following the assumption proposed in [4], many subsequent works [5, 11, 10] assume that the sum of similarities between the frame descriptors reflects the similarity of the videos. In particular, after shifting the videos to each possible offset, the similarity of two videos at each offset is computed by summing up the similarities of frames sharing the same timestamps. The final video-wise similarity score is the maximum value among the similarities at all possible offsets. However, the summation makes TMK lose the ability to discriminate if a high sum is resulted by a period of consecutive similar frames or non-consecutive randomly similar frames. For ease of explanation, we describe the pattern of frame-wise similarities caused by consecutive similar frames as the **burst**, and the pattern of noisy frame-wise similarities caused by non-consecutive yet randomly similar frames as **noise** in the following text. We define a true match case as a high similarity score resulting from a video pair mainly composed of bursts and a false match case as a false match case as a high similarity score resulting from noise. Since TMK cannot exclude false match cases from its results, its performance becomes hard to improve.

Inspired by the interleaving technique used for burst error correction, we propose a novel method to empower TMK with the ability of ruling out frame-wise similarities in noise from the video-wise similarity score, while keeping most of the similarities in bursts. To assist the TMK to precisely locate bursts, we also propose a strategy to select multiple periods from Fibonacci series, where the ratio of two adjacent numbers is asymptotically the golden ratio. Our proposed method shows significant improvement over baseline methods.

2. PRELIMINARIES

We show the processes of using temporal matching kernel (TMK) [5] to perform video retrieval in this section.

2.1. Video descriptors

Given a pair of videos denoted by $\mathbf{x} = [\boldsymbol{x}_0, \dots, \boldsymbol{x}_t, \dots]$ and $\mathbf{x}' = [\boldsymbol{x}'_0, \dots, \boldsymbol{x}'_{t'}, \dots]$, where $\boldsymbol{x}_t, \boldsymbol{x}'_{t'} \in \mathbb{R}^d$ are ℓ_2 normalized frame descriptors. To find the potential offset between videos for alignment, videos need to be shifted to have all possible relative offsets in units of frames. The similarity between $\mathbf{x}$ and $\mathbf{x}'$ at any possible offset δ is calculated by:

$$\begin{aligned}\mathcal{K}_\delta(\mathbf{x}, \mathbf{x}') &\propto \sum_{t=0}^{\infty} \boldsymbol{x}_t^\top \boldsymbol{x}'_{t+\delta} = \sum_{t=0}^{\infty}\sum_{t'=0}^{\infty} \boldsymbol{x}_t^\top \boldsymbol{x}'_{t'} k(t, t'+\delta) \\ &= \underbrace{\left(\sum_{t=0}^{\infty} \boldsymbol{x}_t \otimes \boldsymbol{\varphi}(t)\right)^\top}_{\boldsymbol{\psi}_0(\mathbf{x})} \underbrace{\left(\sum_{t'=0}^{\infty} \boldsymbol{x}'_{t'} \otimes \boldsymbol{\varphi}(t'+\delta)\right)}_{\boldsymbol{\psi}_\delta(\mathbf{x}')},\end{aligned} \tag{1}$$

where $\mathcal{K}_\delta$ denotes a similarity metric between a pair of videos when they are shifted to have an offset δ. In particular, $\mathcal{K}_\delta(\mathbf{x}, \mathbf{x}')$ is measured by summing up the frame-wise similarities between each pair of $\boldsymbol{x}_t$ and $\boldsymbol{x}'_{t+\delta}$. TMK decouples the integral comparison on features and timestamps to independent comparions: $\boldsymbol{x}_t^\top \boldsymbol{x}'_{t'}$ measures the similarity between any two feature vectors, and $k(t, t'+\delta)$ is a kernel shaped like Dirac's delta function, ensuring only the similarities between frame pairs having a δ offset can be counted into the video-wise similarity. Moreover, by introducing the operator of Kronecker product $\otimes$, the explicitly expanded kernel $k(t, t'+\delta) \approx \boldsymbol{\varphi}(t)^\top \boldsymbol{\varphi}(t'+\delta)$ is split into both video descriptors.

We dive a bit more into the explicit feature mapping used to expand the kernel $k(\cdot, \cdot)$. The expanded vector $\boldsymbol{\varphi}(t)$ can be obtained by:

$$\begin{aligned}\boldsymbol{\varphi}(t) = [&\sqrt{a_0}, \sqrt{a_1}\cos(\frac{2\pi}{T}t), \sqrt{a_1}\sin(\frac{2\pi}{T}t), \cdots, \\ &\sqrt{a_m}\cos(\frac{2\pi}{T}mt), \sqrt{a_m}\sin(\frac{2\pi}{T}mt)]^\top,\end{aligned} \tag{2}$$

according to [12]. Here, T is the period used to obtain the Fourier series coefficients $\{a_\cdot\}$ of the kernel function, and m is the number of frequencies. As m grows larger, the kernel function can be better precisely approximated. However, this adds additional memory and computational costs. In consistency with Eq. (2), the video descriptor of $\mathbf{x}$ can be written as

$$\boldsymbol{\psi}_0(\mathbf{x}) = [\mathbf{D}^\mathrm{T}, \mathbf{C}_1^\mathrm{T}, \mathbf{S}_1^\mathrm{T}, \dots, \mathbf{C}_m^\mathrm{T}, \mathbf{S}_m^\mathrm{T}]^\mathrm{T}, \tag{3}$$

where

$$\begin{aligned}\mathbf{D} \propto \sqrt{a_0}\sum_{t\in\mathcal{T}} \boldsymbol{x}_t,\ \mathbf{C}_i \propto \sqrt{a_i}\sum_{t\in\mathcal{T}} \boldsymbol{x}_t \cos\left(\frac{2\pi}{T}it\right), \\ \mathbf{S}_i \propto \sqrt{a_i}\sum_{t\in\mathcal{T}} \boldsymbol{x}_t \sin\left(\frac{2\pi}{T}it\right).\end{aligned} \tag{4}$$

Here, $\mathcal{T}$ is the set of timestamps of frames in video $\mathbf{x}$. We call $\mathbf{D}$ the direct current (DC) component, and the rest are alternating current (AC) components. Note that the DC component after ℓ_2 normalization is proportional to the average of feature vectors, which is usually used to obtain the baseline.

2.2. Retrieval and alignment

The key contribution of TMK is that it uses trigonometric transformation to perform retrieval and alignment at a low cost according to Eq. (5) below.

$$\begin{aligned}\mathcal{K}_\delta(\mathbf{x}, \mathbf{x}') \propto \mathbf{D}^\top \mathbf{D}' &+ \sum_{i=1}^{m} \cos\left(\frac{2\pi}{T}i\delta\right)\left(\mathbf{C}_i^\top \mathbf{C}'_i + \mathbf{S}_i^\top \mathbf{S}'_i\right) \\ &+ \sum_{i=1}^{m} \sin\left(\frac{2\pi}{T}i\delta\right)\left(\mathbf{C}_i^\top \mathbf{S}'_i - \mathbf{S}_i^\top \mathbf{C}'_i\right).\end{aligned} \tag{5}$$

That is to say, all the inner products such as $\mathbf{D}^\top \mathbf{D}'$, $\mathbf{C}_i^\top \mathbf{C}'_i$ and $\mathbf{C}_i^\top \mathbf{S}'_i$ etc. only need to be calculated once for any value of δ. The similarity score between a pair of videos is then computed by

$$\mathcal{S}(\mathbf{x}, \mathbf{x}') = \max_\delta \mathcal{K}_\delta(\mathbf{x}, \mathbf{x}'), \tag{6}$$

and the offset for alignment is obtained at the same time:

$$\Delta(\mathbf{x}, \mathbf{x}') = \arg\max_\delta \mathcal{K}_\delta(\mathbf{x}, \mathbf{x}'). \tag{7}$$

In addition to the entire TMK framework, Poullot et al. [5] introduced a multi-period strategy to use several periods shorter than the video length to create multiple video descriptors for one video. This requires more memory but provides better alignment accuracy, or localization accuracy in video copy detection tasks.

3. PROPOSED METHOD

The previous TMK framework cannot distinguish if a high video-wise similarity score is a result from a true match, where continuously similar contents are shared by a pair of videos, or a false match, where a lot of randomly similar yet irrelevant frames shows in both videos. We propose a solution to this problem neglected by previous works on temporal matching kernel.

The overview of our proposed method is presented in Fig. 1. We take two video pairs as an example, with both of them having high similarity scores measured by TMK. The left pair is from different events in which the street views and human beings are randomly similar in many frames, while the right pair is from the same event and the shared contents are highlighted by green background. To improve the retrieval performance, we equip TMK with the ability to discriminate the true and false match cases by using a shuffle strategy.

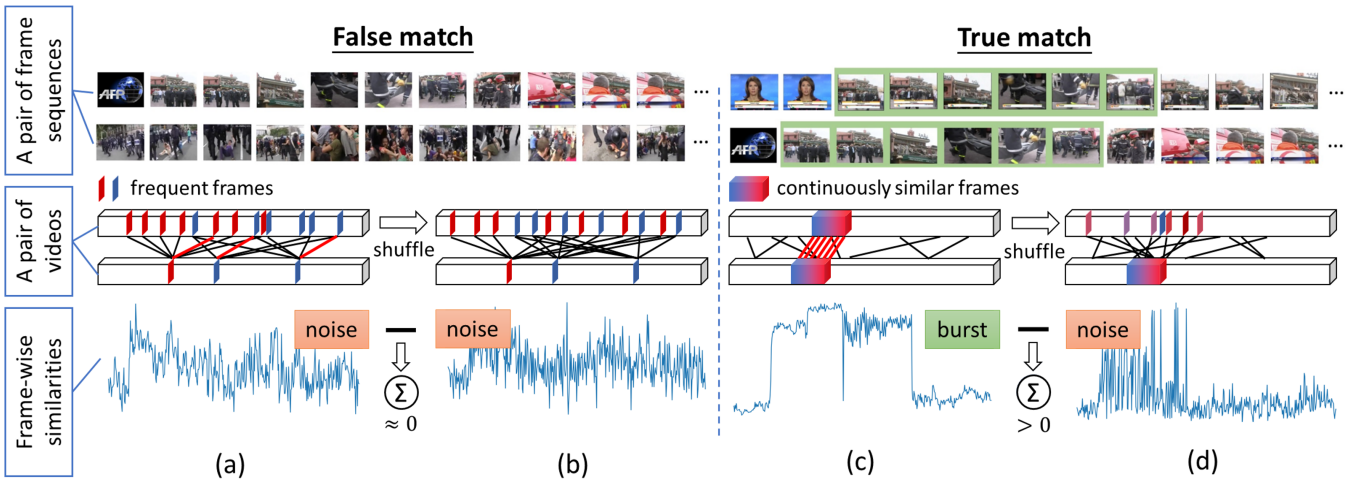

Fig. 1: Overview of the burst-survive temporal matching kernel process.

From our observations, the false match case contains noise before and after shuffling. When we calculate the differences between the before and after frame-wise similarities, the resulting sum is approximately zero. In contrast, since the shuffling collapses the temporal continuity in the true match case, it contains burst before and noise after shuffling. When we calculate the differences in this case, the resulting sum is much higher than zero. This is because the contribution of bursts is strong enough to "survive" the subtraction. The before and after frame-wise similarities along the axis of time measured by the inner product of feature vectors are shown in Fig. 1(a-d).

3.1. Burst-survive temporal matching kernel

We split the design of video descriptors of BSTMK into the DC and AC components as defined in Eq. (4). As mentioned in Section 2.1, the DC component itself can be used as the video descriptor. We keep the DC component unchanged and apply shuffling to the AC components. Similar to Eq. (1), we take the AC components in video descriptors defined as follows:

$$\mathcal{B}_\delta(\mathbf{x},\mathbf{x}') = \mathcal{K}_\delta(\mathbf{x}-\hat{\mathbf{x}},\mathbf{x}'-\hat{\mathbf{x}}') \propto \underbrace{\left(\sum_{t=0}^{\infty}(\boldsymbol{x}_t-\hat{\boldsymbol{x}}_t)\otimes\boldsymbol{\varphi}(t)\right)}_{\psi_0(\mathbf{x}-\hat{\mathbf{x}})}^{\top} \underbrace{\left(\sum_{t'=0}^{\infty}(\boldsymbol{x}'_{t'}-\hat{\boldsymbol{x}'}_{t'})\otimes\boldsymbol{\varphi}(t'+\delta)\right)}_{\psi_\delta(\mathbf{x}'-\hat{\mathbf{x}'})}, \quad (8)$$

where we use the notation $\hat{\mathbf{x}} = [\hat{\boldsymbol{x}}_0,\dots,\hat{\boldsymbol{x}}_t,\dots]$ for shuffled videos.

Note that we apply shuffling to both videos in the pair to create their descriptors. This provides simplicity in implementation, and also ensures that the longer video uses shuffling to create its descriptor. We call this a symmetric BSTMK, which achieves better performance than the asymmetric one, such as $\mathcal{K}_\delta(\mathbf{x}-\hat{\mathbf{x}},\mathbf{x}')$ or $\mathcal{K}_\delta(\mathbf{x},\mathbf{x}'-\hat{\mathbf{x}}')$.

3.2. Fibonacci periods

Multiple periods shorter than the video length are used to create video descriptors for improving the alignment accuracy, which plays a crucial role in locating the bursts. From Eq. (5) we know that $\mathcal{K}_\delta(\mathbf{x},\mathbf{x}')$ is a periodic signal with the period T. This leads to the theory behind the design of multiple periods: when we choose two periods T_1 and T_2 to moderate the video descriptors, the sum of two periodic signals has the period $T_1 \cdot T_2/gcd(T_1,T_2)$, where gcd is the greatest common divisor (GCD). In order to disambiguate the true offset for alignment in a larger period, it is recommended by Poullot et al. [5] to select relatively prime periods whose GCD is 1. Baraldi et al. [10] follows this principle to select a series of periods and test all combinations to select the optimized ones.

However, we observed that some combinations of prime numbers such as $233, 311$ cannot guide us to the truly aligned location in the range of 233×311. This is because the ratio between them is close to $3/4$, thus the approximate GCD between them is 77 rather than the actual 1. We propose to select periods from the Fibonacci series where the ratio of two adjacent numbers is asymptotically the golden ratio, which is hardest to be approximated by any fractions.

4. EXPERIMENTS

4.1. Datasets

The EVVE dataset is commonly used for event retrieval [4]. The entire dataset contains 620 query videos and 2,375 index videos, categorized into 13 events. Videos are decoded in the rate of 15 fps and the pre-extracted $1,024$-d MVLAD [13] descriptors are provided online. The shortest video only contains 12 frames, while the longest one contains 59,810 frames, which makes this dataset more challenging. For evaluation, we adopt the commonly used mean average precision (mAP)

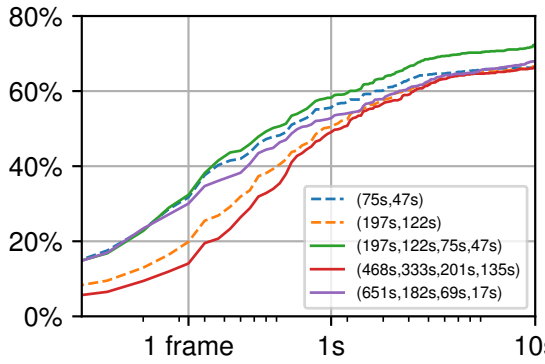

Fig. 2: Alignment accuracy on CBCD dataset with different combinations of periods.

Method	Align	mAP	mAP (DoN)
MMV [4]	✗	33.4	-
SHP [7]	✗	36.3	44.0
CTE [4]	✓	35.2	-
MMV+CTE [4]	✓	37.6	-
TMK [5]	✓	33.5	41.3
BSTMK	✓	**38.3**	**45.3**

Table 1: Retrieval performance (average mAP) evaluated on EVVE dataset, using the MVLAD [16] descriptors provided by the dataset. The mean MVLAD (MMV) is used as the baseline. Methods with alignment ability are marked by ✓.

for each event and take the average mAP for all events as a performance measurement.

We use the TV CBCD 2011 dataset from TRECVID [14] to evaluate the alignment accuracy. It contains 1,608 query and 16,776 reference videos. The temporal offsets between the queries and their corresponding matches in the database are provided as the ground-truth. We create video descriptors by using different combinations of offsets to demonstrate the effectiveness of our proposed Fibonacci periods.

4.2. Implementation details

We first evaluate our method with MVLAD descriptors to make fair comparison with other previous works. In addition, we extract frames at 5 fps regardless of the original frame rates and then extract the activation of the last convolutional layer on ResNet-50 [15] to reap the benefits of CNNs. Following [13], we apply PCA whitening on these extracted features, reducing the dimension from $2,048$ to $1,024$. After that, we moderate BSTMK descriptors with the parameter $m = 16$. During the retrieval, we adopt the average query expansion (AQE) or difference of neighborhood (DoN) [7, 5, 8] for query expansion. The lengths of the short and far list is set to 10 and 2000 respectively.

Method	Align	mAP	mAP (AQE)
Mean AlexNet+ResNet-50* [8]	✗	47.3	53.1
CGA [8]	✗	**52.3**	**58.3**
Mean RMAC* [10, 17]	✗	52.9	-
LAMV [10]	✓	**53.6**	**58.7**
Mean ResNet-50*	✗	46.7	52.8
TMK	✓	46.1	53.9
BSTMK	✓	**49.6**	**57.1**

Table 2: Comparison between each state-of-the-art method and its corresponding baseline. All the baselines are marked by *.

4.3. Experimental results

4.3.1. Video alignment

We show the effectiveness of our proposed Fibonacci periods in Fig. 2. During the moderation of video descriptors, we choose periods (987,610,377,233) from the Fibonacci series, corresponding to (197s,122s,75s,47s) at 5 fps. We also evaluate the alignment accuracy with the periods (7019,5003,3019,2027) proposed in [5] and the periods (9767,2731,1039,253) proposed in [10], corresponding to (468s,333s,201s,135s) and (651s,182s,69s,17s) at 15 fps respectively. We observed that our combination of periods outperforms all the previous proposals. The selected periods (197s,122s,75s,47s) have the ability to handle various lengths of videos, and thus perform better than its subset pairs, (197s,122s) and (75s,47s).

4.3.2. Event retrieval

Our proposed method results in the best performance out of previous methods evaluated by the default MVLAD descriptors (Table 1). Since the recent works adopted different frame descriptors and post-processes, it's hard to compare them to each other with the same experimental settings. Therefore, we present the baselines corresponding to these recent works and compare their improvements in Table 2. While CGA has the best improvement to its baseline, it cannot align videos. Among the methods capable of video alignment, the LAMV, also based on the temporal matching kernel, achieved 0.7% improvement to its baseline. On the other hand, our BSTMK significantly improves the baseline, from 46.7% to 49.6%.

5. CONCLUSION

We propose the burst-survive temporal matching kernel and the Fibonacci periods in this paper and show that our approach significantly improves the retrieval performance from the baseline. In addition, the Fibonacci periods also provide better alignment accuracy compared to other choices.

6. REFERENCES

[1] Alexandre Karpenko and Parham Aarabi, "Tiny videos: A large data set for nonparametric video retrieval and frame classification," *TPAMI*, vol. 33, no. 3, pp. 618–630, 2011.

[2] Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Richang Hong, "Multiple feature hashing for real-time large scale near-duplicate video retrieval," in *ACM Multimedia*. ACM, 2011, pp. 423–432.

[3] Matthijs Douze, Hervé Jégou, Cordelia Schmid, and Patrick Pérez, "Compact video description for copy detection with precise temporal alignment," in *ECCV*. Springer, 2010, pp. 522–535.

[4] Jérôme Revaud, Matthijs Douze, Cordelia Schmid, and Hervé Jégou, "Event retrieval in large video collections with circulant temporal encoding," in *CVPR*. IEEE, 2013, pp. 2459–2466.

[5] Sébastien Poullot, Shunsuke Tsukatani, Anh Phuong Nguyen, Hervé Jégou, and Shin'Ichi Satoh, "Temporal matching kernel with explicit feature maps," in *ACM Multimedia*. ACM, 2015, pp. 381–390.

[6] Josef Sivic and Andrew Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*. IEEE, 2003, p. 1470.

[7] Matthijs Douze, Jérôme Revaud, Cordelia Schmid, and Hervé Jégou, "Stable hyper-pooling and query expansion for event detection," in *ICCV*. IEEE, 2013, pp. 1825–1832.

[8] Zhanning Gao, Gang Hua, Dongqing Zhang, Nebojsa Jojic, Le Wang, Jianru Xue, and Nanning Zheng, "Er3: A unified framework for event retrieval, recognition and recounting," in *CVPR*. IEEE, 2017, pp. 2253–2262.

[9] Fan Yang, Sbastien Poullot, et al., "Temporal matching kernel with embedded stability-sensitive filter," in *ISM*. IEEE, 2017, pp. 278–283.

[10] Lorenzo Baraldi, Matthijs Douze, Rita Cucchiara, and Hervé Jégou, "Lamv: Learning to align and match videos with kernelized temporal layers," in *CVPR*. IEEE, 2018.

[11] Junfu Pu, Yusuke Matsui, Fan Yang, and Shin'ichi Satoh, "Energy based fast event retrieval in video with temporal match kernel," in *ICIP*. IEEE, 2017, pp. 885–889.

[12] Andrea Vedaldi and Andrew Zisserman, "Efficient additive kernels via explicit feature maps," *TPAMI*, vol. 34, no. 3, pp. 480–492, 2012.

[13] Hervé Jégou and Ondřej Chum, "Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening," in *ECCV*, pp. 774–787. Springer, 2012.

[14] Alan F Smeaton, Paul Over, and Wessel Kraaij, "Evaluation campaigns and trecvid," in *MIR Workshop*. ACM, 2006, pp. 321–330.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*. IEEE, 2016, pp. 770–778.

[16] Relja Arandjelovic and Andrew Zisserman, "All about vlad," in *CVPR*. IEEE, 2013, pp. 1578–1585.

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