2017-06-19

CoMo: A Compact Composite Moment-Based Descriptor for Image Retrieval

Vassou, Sotiris A.

ACM

CoMo: A Compact Composite Moment-Based Descriptor for Image Retrieval

Conference Paper · June 2017

DOI: 10.1145/3095713.3095744

Klitos Christodoulou
Neapolis University

Savvas A. Chatzichristofis
Neapolis University

CoMo: A Compact Composite Moment-Based Descriptor for Image Retrieval

Sotiris A. Vassou
Cyprus University of Technology
Limassol, Cyprus
sa.vassou@edu.cut.ac.cy

Nektarios Anagnostopoulos
Microsoft
Prague, Czech Republic
neanagno@microsoft.com

Angelos Amanatiadis
Democritus University of Thrace
Xanthi, Greece
aamanat@ee.duth.gr

Klitos Christodoulou
Neapolis University
Pafos, Cyprus
klitos.christodoulou@nup.ac.cy

Savvas A. Chatzichristofis
Neapolis University
Pafos, Cyprus
s.chatzichristofis@nup.ac.cy

ABSTRACT

Low level features play a vital role in image retrieval. Image moments can effectively represent global information of image content while being invariant under translation, rotation, and scaling. This paper briefly presents a moment based composite and compact low-level descriptor for image retrieval. In order to test the proposed feature, the authors employ the Bag-of-Visual-Words representation to perform experiments on two well-known benchmarking image databases. The robust and highly competitive retrieval performances, reported in all tested diverse collections, verify the promising potential that the proposed descriptor introduces.

CCS CONCEPTS

•**Information systems → Information retrieval;** *Document representation;*

KEYWORDS

Content Based Image Retrieval, Low level features, Compact Composite Descriptors

ACM Reference format:
Sotiris A. Vassou, Nektarios Anagnostopoulos, Angelos Amanatiadis, Klitos Christodoulou, and Savvas A. Chatzichristofis. 2017. CoMo: A Compact Composite Moment-Based Descriptor for Image Retrieval. In *Proceedings of 15th International Workshop on Content-Based Multimedia Indexing, Florence, Italy, June 2017 (CBMI'17),* 5 pages.
DOI: 10.475/123_4

1 INTRODUCTION

Image retrieval is a long-standing problem in the area of computer vision and several approaches for content based image retrieval have been proposed in the literature, ranging from global to local features and, most recently, to convolutional neural networks. Research conducted thus far suggests that each of the approaches has its own benefits and certain limitations [8]. Among the most commonly used global features are the *Image Moments*, which help identify certain key characteristics in images. Their significance, in the fields of image analysis and object representation, is based on the fact that they represent global information of image content while being invariant under translation, rotation, and scaling. In the field of image retrieval, several methods have been proposed to utilize the advantages of image moment invariants and shape global features [12]. However, limitted research work has been conducted on shaping local moment-based descriptors [6]. As collections and retrieval scenarios became more demanding, global feature methods were overshadowed and often also outperformed by methods that employed local features. Local feature descriptors are extracted from every input image and converted into visual words quantizing the vectors' space.

This paper describes a new Moment-Based local and global descriptor, called CoMo. Specifically, the proposed feature is shaped by combining the color information from the color unit of the Color and Edge Directivity Descriptor (CEDD)[1] with the Seven Invariant Moments (SIM), presented by Hu, as the new texture unit. This solution provides a better description and retrieval of images due to the independence on rotation, scaling, and translation. More details about CEDD and SIM are provided in the following Sections. The proposed descriptor is evaluated on two benchmarking datasets showing consistent improvement over baseline.

2 HU MOMENTS

Moment invariants originated mainly from a well established area of mathematics called algebraic invariants. By using the geometrical, central and normalized image moments, Hu constructed seven moments that are invariant to any translation, scaling and rotation transformation of the image being processed [11]. Hu's approach was based on the work of the 19th century mathematicians Boole, Cayley and Sylvester [3].

For a given image with pixel intensities $f(x, y)$, geometrical image moments M_{pq} are calculated by:

$$M_{pq} = \sum_X \sum_Y x^p y^q f(x, y) \qquad (1)$$

The centroid coordinates are defined as:

$$\overline{X} = \frac{M_{10}}{M_{00}} \quad \text{and} \quad \overline{Y} = \frac{M_{01}}{M_{00}} \qquad (2)$$

CBMI'17, Florence, Italy

DOI: 10.475/123_4

The central moments μ_{pq} are constructed from geometrical moments:

$$\mu_{pq} = \sum_{x}\sum_{y}(x-\overline{X})^{p}(y-\overline{Y})^{q}f(x,y) \tag{3}$$

The nth central moment is translation-invariant, i.e. for any random variable f and any constant e:

$$\mu_n(f+e) = \mu_n(f) \tag{4}$$

Furthermore, invariants η_{pq} with respect to both translation and scale can be constructed from central moments:

$$\eta_{pq} = \frac{\mu_{pq}}{(\mu_{20}+\mu_{02})^{\gamma}} \tag{5}$$

where $\gamma = (p+q+2)/4$.

The seven invariant moments (SIM) are given as follows:

$$\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02}\\
\phi_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2\\
\phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (\eta_{03} - 3\eta_{21})^2\\
\phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{03} + \eta_{21})^2\\
\phi_5 &= (3\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2\\
&\quad - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})\\
&\quad \times [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\\
\phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\\
&\quad + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})\\
\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2\\
&\quad - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03}\\
&\quad \times [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{30})^2]
\end{aligned} \tag{6}$$

Moment based invariants, in various forms, have been widely used over the years as features for recognition in many areas of image analysis.

3 SHAPING THE DESCRIPTOR

The MPEG-7 like global descriptor CEDD utilizes both color and texture information to describe the content of an image and has been widely used in recent literature due to its successful trade-off between effectiveness and efficiency. CEDD is computationally lightweight relative to other feature extraction mechanisms, but has comparable accuracy. Even though CEDD was initially designed so as to globally describe the visual information of an input image, its scalability on characterizing single feature points has already been proven. As shown in [4], the localized equivalent of CEDD outperforms the matching accuracy of many other descriptors, like SIFT or SURF. The effectiveness of CEDD relies on its ability to combine color and texture information. CEDD is a scale-invariant descriptor and can tolerate small local rotations, but it is not rotation invariant and does not allow for large global rotations.

Similarly to the structure of CEDD, the proposed descriptor, hereafter referred as CoMo, consists of 6 regions. Each region represents a type of texture. The number of clusters comes as a compromise between the low storage requirements of the application using the proposed descriptor, and the need for more effective retrieval accuracy. Moreover, each texture region is comprised of 24 individual regions, emanating from the Color Unit. Overall, the final histogram includes $6 \times 24 = 144$ regions.

In most cases, moment-based image representation methods extract the color information at each color channel independently [7]. In fact, there exist dependencies caused by linear transform in the color space. In contrary, CoMo shares the same color information extraction unit with CEDD. A two-stage fuzzy-linking system maps the color information of the input image in a 24-bin color histogram using the HSV color space. The first stage of the fuzzy system has the three mean HSV channels of an Image-Block as inputs, and forms a 10-bins histogram as output. The second-stage fuzzy linking system is responsible for adding the brightness value to the calculated colors. Again the S and V mean values of an Image-Block become fuzzy inputs. The output is a 3-bin histogram of crisp values, indicating if the color will be characterized as light, normal or dark hued. The two outputs (first and second stage histograms) are combined and the final 24-bin color histogram is produced. More details about the color unit are given in [1].

To incorporate the texture information, CoMo proposes a novel 6-bin histogram, taking into account the aforementioned set of SIM ($\phi_1, \phi_2, \ldots, \phi_7$). In order to shape the 6 predefined texture regions, the authors employed 100000 randomly selected images from Flickr. Next, random patches of various sizes from all images were extracted. After calculating the Hu moments from these patches, using a k-means classifier, 6 classes are shaped (see Algorithm 1). It is worth noting that a 7-dimensional vector describes the center of each class. The set of the 6×7 resulted values are hereafter called C.

Algorithm 1 Calculate the First Set of Chromosomes

```
for i ≤ Number of Random Images do
    Generate Random Number of Patches
    for j ≤ Number of Patches do
        U += Calculate the 7 Hu Moments from j
    end for
end for
//Array U contains (i × j) 7-dimensional vectors
Using k-Means Classify U into 6 classes (array C)
```

In the sequel, a simple genetic algorithm determines off-line the 6 predefined texture regions that the proposed descriptor uses. The chromosomes used by the genetic algorithm consists of (`7 Hu moments` $\times$ `6 classes`) values. The algorithm begins with an initial population of 20 chromosomes. The first chromosome is the set of the 6 resulted by the k-means classifier values (array C). 9 more chromosomes are shaped by slightly modifying the first one, and 10 chromosomes are randomly generated.

The authors employed the UCID database (refer to Sec.5 for details) and a simple image retrieval framework. A 6-dimensional vector was calculated for each image, taking into account the values of its Hu moments. This procedure aims to map the texture of a given image into a compact histogram. In order to shape a texture histogram for an input image, the input image was segmented into 64 non-overlaped image blocks. For each image block, the Hu moments were extracted and their distance with the 6 given centers was calculated. Based on the distance with each one of the given centers, the texture histogram arises. Next, since both query images and ground truths are known, an image retrieval procedure executed and the Mean Average Precision (MAP) was calculated.

Algorithm 2 Tune the Texture Regions

```
Ch[20, 6] is a set of (20 × 6) 7-dimensional vectors
for t ≤ 20 do
  for i ≤ Number of Images in UCID do
    Segment the 64 image blocks
    for j ≤ 64 do
      U′= Calculate the 7 Hu Moments from block j
      MinVal=double.MIN
      for k ≤ 6 do
        if MinVal ≤ |Ch[t, k] − U′| then
          MinVal = |Ch[t, k] − U′|
          Min= k
        end if
      end for
      Histo[i,Min]++;
    end for
  end for
  Perform Retrieval and Calculate the MAP[t]
end for
Sort Chromosomes based on MAP
```

The procedure was repeated for the 20 set of chromosomes. The chromosomes are then sorted based on the resulted MAP and the best 10 are kept for the formation of the next generation. A *crossover* procedure is applied to the next 3 best chromosomes while the next best 3 chromosomes are *mutated* by increasing or decreasing only one contributor value of the chromosome. Finally, 4 additional chromosomes are randomly inserted. The procedure is repeated until the fitness function is minimized and there is no further improvement. The best chromosome is then used to form the 6 texture areas that the proposed descriptor uses. The entire process is also discussed in Algorithm 2.

4 DESCRIPTOR'S STRUCTURE

The proposed descriptor utilizes a random patches' generator to extract the regions of interest from an image. As its name implies, this approach randomly selects x and y positions in the images, to mark square regions of pixels. Employing a random sampling strategy yield results that are directly comparable and often outperform some of the most sophisticated and complex methods from recent literature [4]. The sizes of the regions were decided as follows: the smallest patch size (defined as Reference Patch, *RP*) was set to 80 x 80 pixels, so as to be aligned with the highest patch size limitation, that is introduced by the CEDD descriptor. From there, we employ a scaling factor (sf) to produce larger patches of sizes $RP * sf \times RP * sf$ pixels.

To shape the proposed descriptor for a patch, first, the input is separated into 1600 equal size image blocks. Each image block interacts successively with both, the color information and the texture information units. In the Color Unit, the image block is converted to the HSV color space in order to provide the first stage of the fuzzy-linking system with its input. Then, the second sub-unit of the system produces the 24-bin histogram.

In the Texture Unit, the Hu Moments of each image block are calculated: In the sequel, Shannon entropy is used as a statistical measure of randomness:

$$H(X) = -\sum_{i=1}^{n} P(X_i) log_b P(X_i) \tag{7}$$

where X is a random variable (image block), n is the number of pixels, $b = 2$ in our case and $P(X_i)$ is the occurrence probability of each pixel. If the result of the Eq. 7 is less than T_{th}, then the block is not considered in the following process because it is assumed that there is not sufficient information (i.e., texture-less block).

Subsequently, the Euclidean distance between the calculated Hu Moments and the 6 predefined texture classes is calculated. The distance is normalized within the interval [0,1], with 0 being the closest to the center of the class. If the resulted value is less than a given threshold, the image block is classified into that texture type. Thus, an image block can participate in more than one type of textures. This unit produces a 6-bin histogram.

At the end of the process, the resulted vectors are combined to form the CoMo histogram of the input patch. To restrict the proposed descriptors' length, the normalized bin values of the descriptor are non-linearly quantized for binary representation in a three bits/bin quantization. In order to calculate the CoMo quantization table, 100000 randomly selected images from Flicker were used. First, CoMo vectors are calculated for all images. The combined 100000×144 elements constitute inputs into a k-means classifier which separates the volume of the samples into eight regions, mapping the bin values from the decimal area [0, 1] into the integer area [0, 7], which can then be represented by 3 bits. It is worth mentioning that the size of the proposed descriptor is equal to the size of CEDD.

5 EXPERIMENTS AND RESULTS

Experiments were performed using two well-known benchmark datasets. First, tests were conducted on the UCID database. This database consists of 1338 images on a variety of topics, including natural scenes and man-made objects, both indoors and outdoors. Manual relevance assessments among all database images are provided. UCID includes several query images where the ground truth consists of images whose visual concept is similar to the query image, even though co-occurrence of the same objects may not exist. Next, the UKBench database, which currently consists of 10200 images arranged in 2250 groups, was used. Each group includes 4 images of a single object, captured from different viewpoints and lighting conditions.

For the performance evaluation of the CoMo feature, experiments were conducted using the Bag-of-Visual-Word model (BoVW). The BoVW model has shown remarkable performance mainly because of the better retrieval effectiveness of the model over global feature representations on near duplicate and verbose images and of course, the clear advantage of the model in terms of efficiency when compared with the local feature representations.

Employed codebook consist of 2048 visual words. The codebook size was chosen based on the results of previous investigations [4].

Evaluation Metric. To evaluate the performance of the proposed descriptor, we used MAP. Table 1, presents the experimental results on the `UCID` [16] and `UKBench` [10] collections. The WS field describes the employed weighting scheme using the SMART notation. The first weighting factor is the term frequency ($tf_{t,d}$), where a weight is assigned to every term (t) in the codebook according to the number of occurrences in a document (d). The second factor for assigning weights is the document frequency (df_t). For the current experiment, df_t is defined as the number of documents that contain the term t. Often, the inverse document frequency $idf_t = log(N/df_t)$ of a collection is used to determine weights, where N is the total number of documents in the collection. Lastly, to quantify the similarity between two documents in terms of the cosine similarity of their vector representation a normalization is performed. Only the weighing scheme that reported the best result is listed in the table.

Table 1: Experimental Results on UCID and UKBench databases. The performance of the comparison algorithms is cited from the reported results of the original papers. Bolded numbers indicate top results.

	UCID		UKBench	
Descriptor	**WS**	**MAP**	**WS**	**MAP**
Local CoMo	ltc	0.779	ltc	**0.929**
Local CEDD	ltc	**0.789**	ltc	0.918
CNNaug-ss[14]	-	-	-	0.911
CNN-ss[14]	-	-	-	0.869
CVLAD[17]	-	-	-	0.893
CoMo	Global	0.684	Global	0.868
DOHC-30 [2]	-	-	-	0.824
DHC-30 [2]	-	-	-	0.816
CEDD	Global	0.674	Global	0.806
LBP Region [15]	-	-	-	0.801
IFV [5, 13]	-	-	-	0.760
CNN[14]	-	-	-	0.760
SURF	lnc	0.626	ncc	0.691
Opponent SIFT	ntc	0.624	ntc	0.593
Color Moments[6]	ntc	0.617	lnc	0.636
SIFT	nnc	0.605	nnc	0.664
ORB	nnc	0.491	ntc	0.491
BRISK	ntc	0.436	nnc	0.310

Through our experimental results, we verified that CoMo yield results that are directly comparable and often outperform some of the much more sophisticated and complex methods from recent literature. Firstly, we concluded that the global form of CoMo outperforms the global form of CEDD in both databases. Furthermore, by observing the results on UCID database, one can conclude that CoMo descriptor, either in its global or on its local form, performs almost identical with CEDD. This is an important observation since it highlights that the new texture unit does not affect the effectiveness of the descriptor. UCID database consists only of visually similar images and ground truths do not contain rotated images.

On the other hand, experimental results on UKBench database illustrate that CoMo outperforms not only all the other descriptors from the literature and CNN approaches, but also CEDD. This result confirms that the new texture unit is able to provide to the descriptor invariance to rotation. Figure 1 illustrates a query image from UKBench together and 7 rotated variations of it. Table 2 presents the distance of both CEDD and CoMo between the query image and its rotated variations. As one can easily observe, CoMo based distances are much smaller than the CEDD's ones. Especially in the cases of 90, 180 and 270 degrees rotation, the distance is equal to 0. The summation of distances, in case of CEDD is 115.22 while in case of CoMo, the value is equal to 19.167. By repeating the experiment for 200 randomly selected images from the same database, the average difference between of summation of distance of CEDD and CoMo was calculated to 64.53.

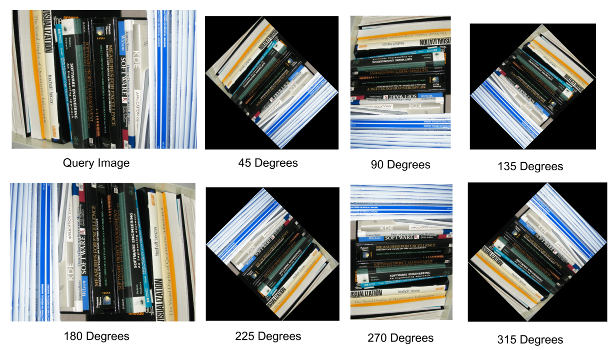

Figure 1: Rotating the query image each time by 45 Degrees

Table 2: Distance between Query image and its rotated variations.

	rotation degrees							
	0°	*45°*	*90°*	*135°*	*180°*	*225°*	*270°*	*315°*
CoMo	0.0	4.5	0.0	5.0	0.0	4.8	0.0	4.9
CEDD	0.0	17.5	22.5	17.6	0.0	17.5	22.5	17.7

Finally, it is worth noting that the proposed descriptor outperforms the only moment-based local descriptor in both databases. In the case of UCID, the improvement is equal to 21%, while in the case of UKBench, CoMo reports an improvement of 31%.

6 CONCLUSION

This paper introduces a new low-level feature for image retrieval. The main novelty of the proposed feature lies in the usage of moment invariants along with the color unit of CEDD as descriptors of local image patches. The findings from the experimental evaluation clearly shown that the proposed descriptor outperforms not only localized CEDD but also other state-of-the-art local descriptors. We plan to extend the experiments by benchmarking the descriptor against other databases used in image retrieval research. The proposed descriptor and its source code is part of the LIRE [9] library[1] and can be used under the GNU GPL license.

[1] http://www.lire-project.net/

REFERENCES

[1] Savvas A. Chatzichristofis and Yiannis S. Boutalis. 2008. CEDD: Color and Edge Directivity Descriptor: A Compact Descriptor for Image Indexing and Retrieval. In *ICVS*. 312–322.

[2] Le Dong, Yan Liang, Gaipeng Kong, Qianni Zhang, Xiaochun Cao, and Ebroul Izquierdo. 2016. Holons Visual Representation for Image Retrieval. *IEEE Trans. Multimedia* 18, 4 (2016), 714–725. https://doi.org/10.1109/TMM.2016.2530399

[3] Mohamed Eisa, Amira Eletrebi, and Ebrahim Elhenawy. 2013. Enhancing the retrieval performance by combing the texture and edge features. *CoRR* abs/1301.2542 (2013). http://arxiv.org/abs/1301.2542

[4] C. Iakovidou, N. Anagnostopoulos, A. Kapoutsis, Y. Boutalis, M. Lux, and S.A. Chatzichristofis. 2015. Localizing global descriptors for content-based image retrieval. *EURASIP Journal on Advances in Signal Processing* 2015, 1 (2015), 80. https://doi.org/10.1186/s13634-015-0262-6

[5] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. 2012. Aggregating Local Image Descriptors into Compact Codes. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 9 (2012), 1704–1716. https://doi.org/10.1109/TPAMI.2011.235

[6] Evangelos G. Karakasis, Angelos Amanatiadis, Antonios Gasteratos, and Savvas A. Chatzichristofis. 2015. Image moment invariants as local features for content based image retrieval using the Bag-of-Visual-Words model. *Pattern Recognition Letters* 55 (2015), 22–27.

[7] Chaorong Li, Yuanyuan Huang, and Lihong Zhu. 2017. Color texture image retrieval based on Gaussian copula models of Gabor wavelets. *Pattern Recognition* 64 (2017), 118–129. https://doi.org/10.1016/j.patcog.2016.10.030

[8] Mathias Lux, Nektarios Anagnostopoulos, and Chryssanthi Iakovidou. 2016. Spatial pyramids for boosting global features in content based image retrieval. In *14th International Workshop on Content-Based Multimedia Indexing, CBMI 2016, Bucharest, Romania, June 15-17, 2016*. 1–4.

[9] Mathias Lux and Savvas A Chatzichristofis. 2008. Lire: lucene image retrieval: an extensible java cbir library. In *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 1085–1088.

[10] David Nistér and Henrik Stewénius. 2006. Scalable Recognition with a Vocabulary Tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*. 2161–2168. https://doi.org/10.1109/CVPR.2006.264

[11] George A. Papakostas, Dimitris E. Koulouriotis, and Evangelos G. Karakasis. 2009. A unified methodology for the efficient computation of discrete orthogonal image moments. *Inf. Sci.* 179, 20 (2009), 3619–3633. https://doi.org/10.1016/j.ins.2009.06.033

[12] George A. Papakostas, Dimitris E. Koulouriotis, Evangelos G. Karakasis, and Vasileios D. Tourassis. 2013. Moment-based local binary patterns: A novel descriptor for invariant pattern recognition applications. *Neurocomputing* 99 (2013), 358–371. https://doi.org/10.1016/j.neucom.2012.06.031

[13] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. 2010. Large-scale image retrieval with compressed Fisher vectors. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*. 3384–3391.

[14] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 2014*. 512–519. https://doi.org/10.1109/CVPRW.2014.131

[15] Carolina Reta, Ismael Solis-Moreno, Jose A. Cantoral-Ceballos, Rogelio Alvarez-Vargas, and Paul Townend. 2017. Improving content-based image retrieval for heterogeneous datasets using histogram-based descriptors. *Multimedia Tools and Applications* (2017), 1–31.

[16] Gerald Schaefer and Michal Stich. 2004. UCID: an uncompressed color image database. In *Storage and Retrieval Methods and Applications for Multimedia 2004, San Jose, CA, USA, January 20, 2004*. 472–480. https://doi.org/10.1117/12.525375

[17] Wan-Lei Zhao, Guillaume Gravier, and Hervé Jégou. 2013. Oriented pooling for dense and non-dense rotation-invariant features. In *British Machine Vision Conference, BMVC 2013, Bristol, UK, September 9-13, 2013*.