DOMINANT SIFT: A NOVEL COMPACT DESCRIPTOR

Anh T. Tra[1†] *Weisi Lin*[2†] *Alex Kot*[3†]

{[1]IGS, [2]School of CE, [3]School of EEE, [†]ROSE Lab}, Nanyang Technological University, Singapore

Definition and extraction of local features play a very important role in image retrieval (IR), pattern recognition and computer vision. Fast growth of technology today calls for local features to be as compact as possible toward real-time and limited bandwidth applications. In this paper, we study the problem of representing images in a compact way to achieve low bit-rate transmission while maintaining good performance. To be more specific, we propose a novel compact descriptor, dominant SIFT, which only uses 48 bits to describe local features. Importantly, our descriptor is training-free, vocabulary-free and suitable for real-time and mobile applications. We show the effectiveness of the proposed compact descriptor in image retrieval.

Index Terms— Local feature, descriptor, image retrieval.

1. INTRODUCTION

If an image is worth more than one thousand words, local features can be seen as the key helping us to understand these words. Undoubtedly, local features have an irreplaceable role in image processing from image retrieval, object recognition to many other applications [14]. Many local features were proposed to be faster, more distinctive and robust under many different variations (e.g. scale, illumination, etc.). Some popular and successful local features developed during the recent decade are Scale Invariant Feature Transform (SIFT) [21], Principal Component Analysis (PCA)-SIFT [18], Speeded Up Robust Features (SURF) [7] and Histogram of Oriented Gradients (HOG) [12].

With the fast growth of technology today, taking and sharing images become easier than ever. Traditional local features have limitations in mobile and real-time applications because of their large size (e.g. 128 bytes for a SIFT feature) [6, 15]. Recently, binary features, such as Binary Robust Independent Elementary Features (BRIEF) [8], Binary Robust Invariant Scalable Keypoints (BRISK) [20] and Fast Retina Keypoint (FREAK) [5], are proposed to represent the local feature in a more distinctive way. However, these features are still large in size (e.g. $\geq$ 16 bytes per feature) while some low bit-rate image retrieval applications aim to a much smaller bit-rate (e.g. $\leq$ 100 bits per feature) [10, 17].

To achieve a more compact descriptor, hashing, vector quantization (VQ) and transform coding (TC) are also considered in [9]. Hashing is an effective way to represent the local feature by using a few bits [16], but it depends a lot on its hash functions. VQ technique represents each local feature by a code-word of a pre-trained vocabulary [23], but the large size of vocabulary becomes a problem for devices having small memory [17]. TC framework maps the local feature from original feature space into the transform space using PCA technique which produces a small reconstruction error when reducing feature dimensions [17]. However, this method depends on the data used for learning the transform matrix and the matching performance drops sharply as shown in our experiment (Section 4.3) when the number of dimensions is less than 10.

To deal with the real-time and mobile applications, we propose a novel compact descriptor called Dominant SIFT, which is based on a desired property of SIFT's dominant orientations. Our proposed descriptor only uses 6 bytes to represent one local feature while preserving a very competitive retrieval performance when comparing to the state-of-the-art vocabulary-free local feature compression methods. Unlike other methods [17, 23], our training-free and vocabulary-free proposed method can be implemented in small memory devices. Section 2 will present related works to our research. Our methodology and experiments will be given in Section 3 and Section 4 respectively. The last section is our conclusion.

a)

53	0	0	0	0	0	0	39
143	8	0	0	0	1	17	155
12	19	9	0	1	15	26	27
49	13	8	1	6	16	13	25
64	0	0	0	0	0	0	11
179	70	7	0	0	0	0	45
45	76	45	36	13	0	0	5
18	25	15	18	20	1	0	2
37	0	0	0	0	0	0	10
179	39	2	0	0	0	0	79
129	46	45	98	15	0	0	6
16	42	24	58	16	0	0	0
16	1	0	0	0	0	0	1
179	33	0	0	0	0	0	61
156	22	13	37	37	1	0	44
7	18	7	33	36	2	3	3

b)

Fig. 1. a. SIFT features of Lena image. The circle and its size correspond to the keypoint and its scale. b. One SIFT descriptor: each row is one 8-bin sub-histogram. Dominant bins of each sub-histogram are underlined.

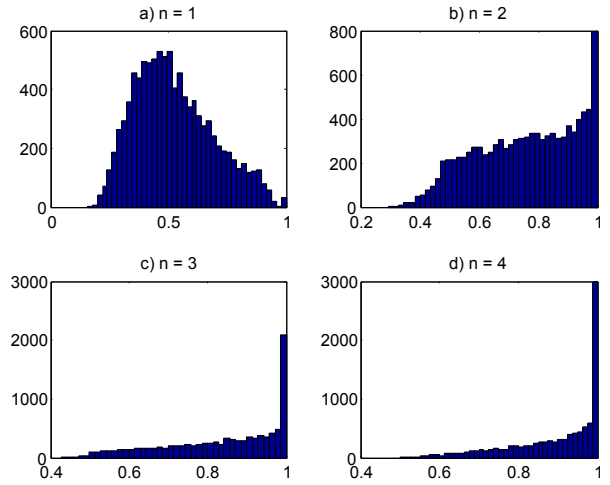

Fig. 2. The histogram of ratio between the maximum consecutive sum-n and the sum of all bins of the sub-histogram in four cases: a. $n = 1$, b. $n = 2$, c. $n = 3$ and d. $n = 4$.

2. RELATED WORK

2.1. Review of SIFT Algorithm

The SIFT feature introduced in [21] by David Lowe includes two main parts, which are keypoint detector and SIFT descriptor. The keypoint detector scans the image's interest points. Firstly, an image is applied with Gaussian filter of different scales and then re-sized to form a Gaussian scale-space. Neighboring images with the same resolution in this scale-space are subtracted to get the Difference-of-Gaussian (DoG) pyramid. The keypoint is taken if and only if it is a local extremum in the DoG pyramid. The keypoint localization is the last step applied to get the most stable keypoints.

David Lowe also proposed to represent keypoints by SIFT descriptors [21]. A local patch centering at the keypoint after being rotated to align with its dominant orientation is separated into 16 4-pixel-by-4-pixel blocks. Gradients for 16 pixels in each block are quantized into 8-bin sub-histogram. Afterward, all 16 8-bin sub-histograms are put together to form the final 128-dimension SIFT descriptor. Fig. 1 illustrates the SIFT descriptor and SIFT local features detected in the famous Lena image.

2.2. PCA-SIFT and Reduced SIFT:

Yan Ke and Rahul Sukthankar [18] proposed to use the PCA technique to concatenate gradients into a new descriptor called PCA-SIFT. Each 41-pixel-by-41-pixel image patch centering at each keypoint is extracted and rotated to line up with its dominant orientation. Gradient values in the x-direction and the y-direction for all pixels in the image patch are calculated to form a 2x39x39 = 3042-dimension vector. A pre-trained eigenspace is used to reduce the dimension of this vector to 32 [18]. By only keeping top dimensions which correspond to the largest eigenvalues, they can achieve PCA-SIFT descriptors as compact as they want.

The dimension of SIFT vector can be directly reduced by using PCA transform. Similarly to PCA-SIFT, a PCA transform matrix is pre-learned from an image database. At mobile devices, SIFT features extracted from query images are applied with PCA transform to achieve a more compact descriptor. This new compact descriptor is called as Reduced SIFT [24]. In this paper, PCA-SIFT and Reduced SIFT will be used for comparing our proposed descriptor not only because they are popular in low bit-rate image retrieval applications [17,19], but also because they used the same vocabulary-free compression methods as our proposal.

3. METHODOLOGY

3.1. Dominant Gradients Based SIFT Compression

As we mentioned in Section 2, the SIFT descriptors are formed from 16 sub-histograms corresponding to 16 4-pixel-by-4-pixel blocks. Sixteen gradients in each block are quantized into 8 bins of the sub-histogram. Undoubtedly, bins in the same sub-histogram have a stronger correlation than bins in different sub-histograms of a SIFT descriptor. More importantly, we realize that the values of a sub-histogram often concentrate on two or three consecutive bins. For example, dominant bins (underlined) of a SIFT feature in Fig. 1 are often adjacent to each other after a circular shift.

We form a simple statistical experiment to check our assumption. For each SIFT vector $(a^j)_{j\in\mathbf{Z}\cap[0,15]}$ where $a^j = \{a_i^j \in \mathbf{Z}\cap[0,256) | i \in \mathbf{Z}\cap[0,7]\}$ is a 8-bin sub-histogram, let $\mathrm{CS}_n(a^j, i)$ be the consecutive sum-n at the position i which is defined as:

$$\mathrm{CS}_n(a^j, i) := \sum_{k=i}^{i+n-1} a_k^j$$

where $a_m^j = a_{m(mod\ 8)}^j \forall m \in \mathbf{Z} \cap (8, \infty)$ and $n \in \{1, 2, 3, 4\}$. Let $\mathrm{MCS}_n(a^j)$ be the maximum of $\mathrm{CS}_n(a^j, i)$ where $i \in \mathbf{Z} \cap [0, 7]$. In our statistical test, we plot the histogram of the ratio $\mathrm{MCS}_n(a^j)/(\sum_{i=0}^{7} a_i^j)$ for every SIFT feature's sub-histograms extracted from the Lena image.

As shown in Fig. 2 as an example to illustrate our observation, the histogram highly biases towards the right-hand side of the graph and reaches its peak at 1 in cases $n \in \{2, 3, 4\}$. This result confirms our assumption that several consecutive bins dominantly contribute to the sub-histogram. The question is how to use this property to represent the SIFT descriptor. For achieving a more compact descriptor, we propose to represent the sub-histogram based on the position of the maximum consecutive sum-n. Only 8 positions are available for the consecutive sum-n, so we only use 3 bits and 48 bits to represent each sub-histogram and the whole SIFT descriptor respectively. We name our new compact descriptor

as Dominant SIFT. Algorithm 1 describes our descriptor algorithm, which is 20 times, 6 times and almost 3 times more compact than the original SIFT [21], PCA-SIFT [18] and the well-known binary features in [5, 8, 20], respectively.

Algorithm 1 Dominant SIFT descriptor generation.

1. For each SIFT feature, separating it into 16 sub-vectors: $a^j = [a_0^j, \ldots, a_7^j]^{\mathbf{T}}, j \in \mathbf{Z} \cap [0, 15]$.
2. Find the position of the maximum consecutive sum-n of a^j : $p^j = \arg\max_{i \in \mathbf{Z} \cap [0,7]} \mathrm{CS}_n(a^j, i)$. Encode the a^j by $p^j \in \{0, 1\}^3$ in gray code.
3. Encode the SIFT feature by 48 bits: $(p^j)_{j \in \mathbf{Z} \cap [0,15]}$.

3.2. Matching Criterion for Image Retrieval

We use the Hamming distance and ratio test for finding the best match for each local feature. Our proposed descriptor can be seen as a compact way to encode the discriminative information of SIFT descriptor. Therefore, using the ratio test to find the best match (as in SIFT matching criterion [21]) is more reasonable than using the threshold test. Algorithm 2 describes our matching criterion.

Algorithm 2 Dominant SIFT descriptor matching criterion.

1. For each query Dominant SIFT feature q, find its nearest neighbor feature a and its second nearest neighbor feature b from the reference image.
2. a is the match of q if and only if: $H_d(q, a) < \delta \cdot H_d(q, b)$, where H_d is the Hamming distance metric and δ is the ratio test threshold.

4. EXPERIMENTS AND RESULTS

4.1. Experiment Setup

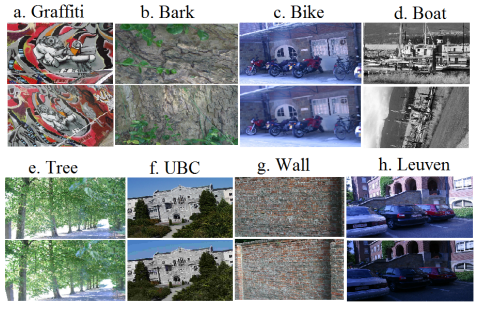

Fig. 3. Descriptor evaluation dataset.

In the two experiments conducted, the first is to test our proposed descriptor's matching ability in comparison with the original SIFT [21] and Reduced SIFT [24] (6-dimension) by following the published descriptor evaluation framework in [22]. The second is to compare the image retrieval ability of our proposed descriptor with other compact descriptors in [18, 21, 24] to highlight its promising matching ability at a very low bit-rate using the published Stanford Mobile Visual Search (MVS) dataset [11]. All necessary reference codes are taken from [1, 3, 4].

4.2. Experiment 1: Descriptor Evaluation

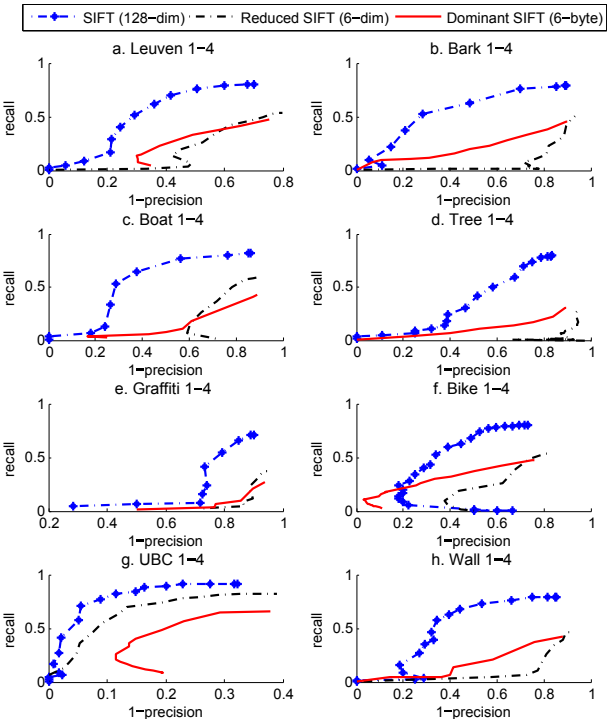

Fig. 4. Descriptor evaluation for original SIFT, Reduced SIFT (6 dimensions) and Dominant SIFT (6 bytes, $n = 2$). The 1-precision and recall curves for each descriptor are plotted for feature matching between image 1 and 4 in Fig. 3.

Mikolajczyk and Schmid proposed the method to test matching ability of descriptor in [22]. As illustrated in Fig. 3, the image database includes 8 categories with variations in: viewpoint (Graffiti and Wall), zom and rotation (Bark and Boat), blur (Bikes and Trees), illumination (Leuven) and JPEG compression quality (UBC). The Harris Affine detector [22] is used in this experiment. Each category includes four pairs of images with corresponding homography transform matrices to calculate the ground truth for the correct matching features. To evaluate descriptors, we plot recall vs. 1-precision curve by changing ratio test thresholds.

Table 1. Image retrieval mean average precision on Stanford MVS dataset [11] with best 2 in bold.

Descriptor/Database	Book Cover	CD Cover	DVD Cover	Business Card	Museum Painting	Printing	Video Frame	**Average**
SIFT [21]	**0.9406**	**0.8400**	**0.8500**	**0.4500**	**0.8242**	**0.4200**	**0.8700**	**0.7421**
Reduced SIFT-16 [24]	**0.98317**	0.6300	0.5900	0.3100	0.7692	0.3700	0.8200	0.6389
Reduced SIFT-10 [24]	0.6337	0.4600	0.4000	0.1500	0.6703	0.2900	0.7700	0.4820
PCASIFT-32 [18]	0.7921	**0.7800**	0.4800	0.3600	**0.7912**	0.3300	**0.8600**	0.6276
PCASIFT-16 [18]	0.7426	0.5600	0.2900	0.2400	0.7413	0.2800	0.8200	0.5248
PCASIFT-10 [18]	0.6040	0.3700	0.1800	0.1000	0.6154	0.1600	0.7400	0.3956
Our proposal	0.8812	0.7300	**0.6800**	**0.5200**	**0.7912**	**0.4300**	0.8200	**0.6932**

This experiment aims to show the ability of our proposed descriptor in very low bit-rates. The PCA transform framework is often used as a vocabulary-free to achieve a low bit-rate for SIFT feature [9, 17]. However, when the dimension is reduced to small value (e.g. 6 dimensions), the performance will drop rapidly. As shown in Fig. 4, gaps in feature matching performance between original SIFT and Reduced SIFT (6-dimension) by PCA transform are very large. Our proposed method (6 bytes), which has the same size as 6-dimension Reduced SIFT (6 bytes in integer representation), can outperform the Reduced SIFT in most of cases. In the worst case, category UBC which is with JPEG compression, our proposal can still achieve a very good error rate (0.3) at a high recall rate (0.6).

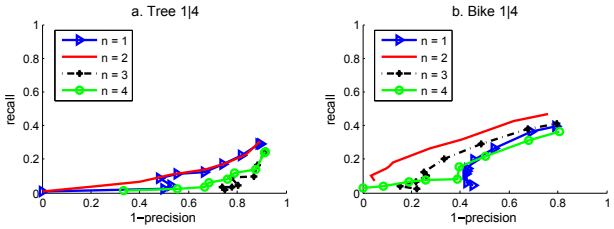

Fig. 5. Descriptor evaluation for Dominant SIFT with different values of n ($n \in \{1, 2, 3, 4\}$). The recall vs. 1-precision curves for each descriptor are plotted for feature matching between image 1 and image 4 in category: a. Tree and b. Bike.

We also conduct another experiment to evaluate our proposed descriptors with different values of n. Fig. 5 shows that the Dominant SIFT in the case $n = 2$ outperforms in other cases ($n \in \{1, 3, 4\}$). We choose using dominant SIFT $n = 2$ in compassion with SIFT and Reduced SIFT.

4.3. Experiment 2: Pairwise Matching Based IR

We also verify our descriptor ability in image retrieval. The original SIFT [21], PCA-SIFT [18] (32, 16 and 10 dimensions) and Reduced SIFT [24] (16 and 10 dimensions) are used as baselines to compare with our proposal. The published Stanford MVS dataset [11] which includes seven image categories (CD cover, DVD cover, book cover, business card, museum painting, printing and video frame) is used in this experiment. Each category has around 100 reference images blended with around 400 distractor images taken from Flickr1M database [2]. In a real large-scale image retrieval system, the global feature is used to find a short list of best matching images (around 500 images), and then the local feature is used to re-rank the short list [13]. The query images are taken from mobile devices (iPhone, android phone, etc.) and correctly matched with only one image in reference image set (1 vs. 500 test). The ratio test threshold is chosen of 0.8 as reported in [21] due to PCA-SIFT, Reduced SIFT as well as our proposal are all SIFT-based descriptors. For each query image, the reference image having the largest number of matching features is concluded as its match.

Table 1 shows the retrieval results for all descriptors taken part in the experiment. Our descriptor brings a very competitive retrieval results in comparing to other SIFT compression methods even when our proposal has a smaller size (2 and 1.6 times less) than other compressed descriptors. In most of the cases, our proposal can get the second best result and outperforms PCA-SIFT (16 dimensions) and reduced-SIFT (16 dimensions) easily. Clearly, our dominant SIFT can work very well at a very low bit-rate, 48 bits per each local feature.

5. CONCLUSIONS

We have proposed a compact (48 bits) Dominant SIFT descriptor by representing dominant orientations of the SIFT descriptor. Our experiment result shows that our training-free and vocabulary-free descriptor is 20 times and 6 times more compact than the original SIFT and the PCA-SIFT with a competitive performance in comparison with other latest vocabulary-free SIFT compression methods, when image retrieval is demonstrated as an example of applications.

Acknowledgment: This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Prime Minister's Office, Singapore, under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Programme Office.

6. REFERENCES

[1] http://www.cs.cmu.edu/~yke/pcasift/.

[2] http://www.multimedia-computing.de/wiki/flickr1m.

[3] http://www.robots.ox.ac.uk/~vgg/research/affine/.

[4] http://www.vlfeat.org/.

[5] Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst. Freak: Fast retina keypoint. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 510–517. IEEE, 2012.

[6] Mitsuru Ambai and Yuichi Yoshida. Card: Compact and real-time descriptors. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 97–104. IEEE, 2011.

[7] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision - ECCV 2006*, pages 404–417. Springer, 2006.

[8] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *Computer Vision - ECCV 2010*, pages 778–792. Springer, 2010.

[9] Vijay Chandrasekhar, Mina Makar, Gabriel Takacs, David Chen, Sam S. Tsai, Ngai-Man Cheung, Radek Grzeszczuk, Yuriy Reznik, and Bernd Girod. Survey of SIFT compression schemes. In *Proc. Int. Workshop Mobile Multimedia Processing*. Citeseer, 2010.

[10] Vijay Chandrasekhar, Gabriel Takacs, David Chen, Sam Tsai, Radek Grzeszczuk, and Bernd Girod. CHoG: Compressed histogram of gradients a low bit-rate feature descriptor. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2504–2511. IEEE, 2009.

[11] Vijay R. Chandrasekhar, David M. Chen, Sam S. Tsai, Ngai-Man Cheung, Huizhong Chen, Gabriel Takacs, Yuriy Reznik, Ramakrishna Vedantham, Radek Grzeszczuk, and Jeff Bach. The stanford mobile visual search data set. In *Proceedings of the second annual ACM conference on Multimedia systems*, pages 117–122. ACM, 2011.

[12] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[13] Ling-Yu Duan, Jie Lin, Jie Chen, Tiejun Huang, and Wen Gao. Compact descriptors for visual search. *MultiMedia, IEEE*, 21(3):30–40, 2014.

[14] Giovanni Maria Farinella, Sebastiano Battiato, and Roberto Cipolla. *Advanced Topics in Computer Vision*. Chapter 1, page 13. Springer, 2013.

[15] Bernd Girod, Vijay Chandrasekhar, Radek Grzeszczuk, and Yuriy A. Reznik. Mobile visual search: Architectures, technologies, and the emerging MPEG standard. *MultiMedia, IEEE*, 18(3):86–94, 2011.

[16] Junfeng He, Regunathan Radhakrishnan, Shih-Fu Chang, and Claus Bauer. Compact hashing with joint optimization of search accuracy and time. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 753–760. IEEE, 2011.

[17] Chen Jie, Duan Ling-Yu, Ji Rongrong, and Wang Zhe. Multi-stage vector quantization towards low bit rate visual search. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 2445–2448, 2012.

[18] Yan Ke and Rahul Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–506–II–513 Vol. 2. IEEE, 2004.

[19] Yan Ke, Rahul Sukthankar, and Larry Huston. Efficient near-duplicate detection and sub-image retrieval. In *ACM Multimedia*, volume 4, page 5, 2004.

[20] Stefan Leutenegger, Margarita Chli, and Roland Yves Siegwart. BRISK: Binary robust invariant scalable keypoints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2548–2555. IEEE, 2011.

[21] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[22] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615–1630, 2005.

[23] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2161–2168. IEEE, 2006.

[24] Ricardo Eugenio Gonzalez Valenzuela, William Robson Schwartz, and Helio Pedrini. Dimensionality reduction through PCA over SIFT and SURF descriptors. In *11th IEEE Conference on Cybernetic Intelligent Systems (CIS 2012)*, volume 1, pages 58–63, 2012.