AFFINE INVARIANT IMAGE COMPARISON UNDER REPETITIVE STRUCTURES

Mariano Rodríguez *Rafael Grompone von Gioi*

CMLA, ENS Cachan, CNRS, Université Paris-Saclay, 94235 Cachan, France

ABSTRACT

We focus on the problem of affine invariant image comparison in the presence of noise and repetitive structures. The classic scheme of keypoints, descriptors and matcher is used. A local field of image gradient orientation is used as descriptor and two matchers are proposed, based on the a-contrario theory, for handling repetitive structures. The affine invariance is obtained by affine simulations. The proposed methods achieve state-of-the-art performances under repetitive structures.

Index Terms— image comparison, repetitive structures, affine invariance, noise, a contrario, IMAS, SIFT, RootSIFT

1. INTRODUCTION

Everyday images are often composed of repeated objects, e.g. roof tiles, windows on buildings or chairs in a classroom. Humans not only identify these repetitions but also extract meaningful information from them. However, most of the state-of-the-art image matching algorithms still either fail to handle repetitions or were conceived not to treat them at all in order to be distinctive in practical applications [1].

The classic approach to image matching consists in three steps: detection, description and matching. First, key-points are detected in the compared images. Second, regions around these points are described and encoded in local invariant descriptors. Finally, all these descriptors are compared and possibly matched. Using local descriptors yields robustness to context changes. Both the detection and description steps are usually designed to ensure some invariance to various geometrical or radiometric changes. A large amount of research focused on using histogram representations, e.g. SIFT [2, 3], ASIFT [4], Shape Contexts [5], Self-Similarity descriptors [6], etc. We refer the reader to [7, 8, 9, 10] for in-depth comparative studies on image descriptors.

Although 3D viewpoint invariance seems quite utopian, its approximated version, affine invariance, has been widely studied in the literature [11, 12, 4, 10]. The superiority of SIFT based descriptors for the latter invariance have been shown in [10]. On the other hand, Image Matching by Affine Simulation (**IMAS**) have been proven to be a reliable way to capture changes of point of view up to an impressive $88°$, see [4, 13, 14, 10].

In order to be distinctive, most IMAS algorithms rely on the second-closest neighbor acceptance criterion proposed by D. Lowe in [2]. This criterion directly implies that the affine invariance property of these algorithms is strongly affected by repeated structures on the target image. To counteract these issues, Cao et al. [15] proposed two approaches to handle repetitions: the first is to compute the "second-closest neighbor" on an unrelated third image (where the repeated structure would not be present); the second is to add an *a-contrario* [16] validation step, independent of the descriptor, which first selects a set of points around the key-points and then evaluates the agreement of gradient orientation on these points. Rabin et al. [17] proposed an *a-contrario* validation for SIFT descriptor matches; the method requires learning the distribution of the descriptor space and uses the earth mover distance to quantify the descriptor similarity. Still a different *a-contrario* framework for match validation was described in [18]; in this case it is based on comparing the gradient orientations in a patch and was suggested to use a local field of gradient orientations as a key-point descriptor.

However, none of these *a-contrario* methods is affine invariant. Here we follow the suggestion of [18], enriched by the IMAS approach, to build an *a-contrario* affine invariant key-point descriptor. Also we propose two variants of descriptor distances and the corresponding *a-contrario* models.

This paper is organized as follows. Section 2 introduces the key-point descriptor based on a local field of image gradient orientation. The two *a-contrario* matchers for our descriptor are introduced in Sect. 3. Then, the IMAS techniques are explained in Sect. 4. Our experiments on images including repetitive structures, different viewpoint angles and noise are presented in Sect. 5. Section 6 concludes the paper.

2. THE GRADIENT ANGLE FIELD DESCRIPTOR

The first step of the method is the key-point extraction. Each key-point comes with a position, scale and orientation. Then, a descriptor is associated to each key-point. A $s \times s$ patch is extracted from the image centered at the position and orientation of the key-point. The sampling step is proportional to the key-point scale. Up to this point this is similar to the SIFT descriptor [2, 3]. But while the SIFT descriptor consists in a set of quantized histograms of the image gradient orientation, here we will follow the suggestion of [18] and use the actual values of patch gradient orientations as a descriptor. The de-

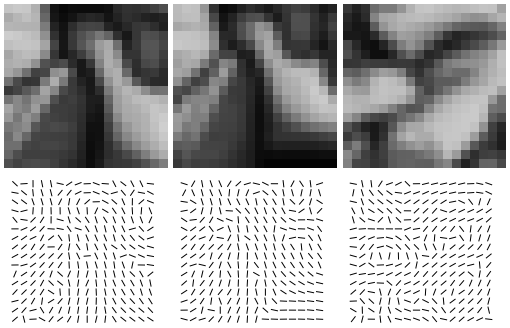


Fig. 1: Three image patches and their corresponding orientation fields used as descriptors. The first two are similar while the third one is different.

scriptor of a key-point a will be then $\alpha = \{\alpha_{ij}\}$, where α_{ij} are the angles of the gradient orientation at position i, j in the extracted patch of size $s \times s = n$. Fig. 1 illustrates the idea. In all our experiments we used $s = 20$ ($n = 400$) and a sampling step of 1.5 relative to the key-point scale.

3. A CONTRARIO MATCH VALIDATION

The proposed validation procedure is based on the *a contrario* theory [16], which relies on the non-accidentalness principle [19, 20]; informally, this principle states that there should be no detection in noise. In the words of Lowe, "we need to determine the probability that each relation in the image could have arisen by accident, $P(a)$. Naturally, the smaller that this value is, the more likely the relation is to have a causal interpretation" [20, p. 39]. In our context, we need to assess the existence of a causal relation between two descriptors.

Given a pair of descriptors α and β, a distance function $d(\alpha, \beta)$ will be defined, together with a stochastic model $\mathcal{H}_0$ for random descriptors used to evaluate accidentalness. We denote by $D_{\mathcal{H}_0}$ a random variable (r.v.) corresponding to the distance between two random descriptors drawn from $\mathcal{H}_0$. To assess the accidentalness of a match (α, β), we need to evaluate the probability $\mathbb{P}[D_{\mathcal{H}_0} \leq d(\alpha, \beta)]$ of observing under $\mathcal{H}_0$ a distance $D_{\mathcal{H}_0}$ smaller or equal than $d(\alpha, \beta)$. When this probability is small enough, there exists evidence to reject the null hypothesis and declare the match meaningful. However, one needs to consider that usually multiple pairs are tested. If 100 tests are performed, for example, it would not be surprising to observe an event that appears with probability 0.01 under random conditions. Thus, the number of tests N_T needs to be included as a correction term, as it is done in the statistical multiple hypothesis testing framework [21]. Following the *a contrario* methodology [16], we define the *Number of False Alarms* (NFA) of a match as:

$$\mathrm{NFA}(\alpha, \beta) = N_T \cdot \mathbb{P}\Big[D_{\mathcal{H}_0} \leq d(\alpha, \beta)\Big]. \tag{1}$$

Pairs with NFA $\leq \varepsilon$, for a predefined ε value, are accepted as valid matches. One can show [16, 22] that under $\mathcal{H}_0$, the expected number of pairs with NFA $\leq \varepsilon$ is bounded by ε. As a result, ε corresponds to the mean number of false detections per random image pair. In most practical applications the value $\varepsilon = 1$ is suitable and we will set it once and for all.

An appropriate (unstructured) null hypothesis $\mathcal{H}_0$ for random descriptors is that the gradient orientation angles are independent and isotropic. In other words, in a descriptor $\Delta \in \mathcal{H}_0$, $\{\Delta_{ij}\}$ is a family of independent random variables, uniformly distributed over $[0, 2\pi)$.

We will consider two distances, which will lead to two validation methods. The first one, denoted d^Q, is defined as the sum of quantized orientation errors:

$$d^Q(\alpha, \beta) = \sum_{ij} \mathbb{1}_{\left\{\frac{|\mathrm{Angle}(\alpha_{ij}, \beta_{ij})|}{\pi} > \rho\right\}}, \tag{2}$$

for a fixed orientation precision $\rho \in (0, 1)$ (we use $\rho = 0.3$). Given that the size of the descriptor is n, the value $d^Q(\alpha, \beta) \in \{0, 1, \ldots, n\}$, with zero corresponding to a good match and n to the worst difference. This distance is similar to the one used in [15, sec.11.3]. The associated r.v. $D^Q_{\mathcal{H}_0}$ corresponds to the sum of n independent Bernoulli random variables. Thus,

$$\mathbb{P}[D^Q_{\mathcal{H}_0} \leq d] = \sum_{k=0}^{d} \binom{n}{k} (1-\rho)^k \rho^{n-k} \tag{3}$$

is related to the tail of a Binomial distribution. We will denote AC-Q the method that uses the distance d^Q.

The second distance, denoted d^W, corresponds to a weighted sum of normalized orientation errors:

$$d^W(\alpha, \beta) = \sum_{ij} w_{ij} \frac{|\mathrm{Angle}(\alpha_{ij}, \beta_{ij})|}{\pi}. \tag{4}$$

Now $d^W(\alpha, \beta)$ is a real value between zero and $\sum_{ij} w_{ij}$. A perfect match has $d^W(\alpha, \beta) = 0$ while the worst difference is $d^W(\alpha, \beta) = \sum_{ij} w_{ij}$. This is similar to the distance in [18] with the addition of the weights w_{ij}, which are used to impose a Gaussian window,

$$w_{ij} = \exp\Big(-\frac{(i - s/2)^2 + (j - s/2)^2}{2\sigma^2}\Big),$$

giving more relevance to the central points and requiring a more complex probability term than in [18]. The r.v. $D^W_{\mathcal{H}_0}$ corresponds to the weighted sum of n independent and uniformly distributed random variables in $[0, 1]$. Using the vector index k, we have $\mathbb{P}[D^W_{\mathcal{H}_0} \leq d] = \mathbb{P}[\sum_k w_k e_k \leq d]$ where the normalized errors e_k are $U[0, 1]$. The possible values of $(e_1, \ldots, e_n)$ can be seen as the points in a n-hypercube and the probability term is given by the volume of the intersection of the hypercube and the half-hyperspace $\{(e_1, \ldots, e_n) : \sum_k w_k e_k \leq d\}$. There is a closed but complex formula for this volume [23]. For our purposes, however, it is enough to approximate it by the upper-bound given by the volume of the simplex $\{(e_1, \ldots, e_n) : e_k \geq 0,\ \sum_k w_k e_k \leq d\}$; thus

$$\mathbb{P}[D^W_{\mathcal{H}_0} \leq d] \leq \frac{1}{n!} \frac{d^n}{\Pi_k w_k}. \tag{5}$$

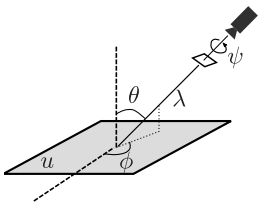


Fig. 2: Geometric interpretation of (7).

We will denote AC-W the method that uses the distance d^W.

Finally, we need to specify the number of tests. Potentially, we may try to match any pixel of image I_1 of size $X_1 \times Y_1$ with any pixel of image I_2 of size $X_2 \times Y_2$. We must also consider about $\sqrt{X_1 Y_1}$ different patch orientations in I_1 and $\sqrt{X_2 Y_2}$ in I_2. To account for multiple scales, we consider $\log_2\big(\max(X_1, Y_1)\big)$ scales in I_1 and $\log_2\big(\max(X_2, Y_2)\big)$ scales in I_2. As we will see, we perform several affine simulations leading to an extra factor κ per image (i.e. the area ratio from [10]). All-in-all, the number of tests writes

$$N_T = (\kappa X_1 Y_1)^{\frac{3}{2}} \cdot \log_2\big(\max(X_1, Y_1)\big) \cdot (\kappa X_2 Y_2)^{\frac{3}{2}} \cdot \log_2\big(\max(X_2, Y_2)\big). \quad (6)$$

4. AFFINE INVARIANCE

As it will be shown in the following section, our methods are not initially affine invariant. In this section we shall introduce a methodology from [4, 10] to render them fully affine invariant. Intuitively, the idea is to simulate a set of views from the initial images that will help to cover the affine space and then pairwise match those simulated images. The set of simulated views shall depend on concrete measurements of our methods' tolerance to viewpoint changes.

Let us call $\mathbf{A}$ the set of affine maps and define $Au(\mathbf{x}) = u(A\mathbf{x})$ for $A \in \mathbf{A}$. We define $\mathbf{A}^+ = \{A \in \mathbf{A} | \det(A) > 0\}$. We call $\mathbf{S}$ the set of similarities, which are any combination of rotations and zooms. Finally we define the set $\mathbf{A}^+_* = \mathbf{A}^+ \setminus \mathbf{S}$, where we exclude pure similarities.

It was proven in [4] that every $A \in \mathbf{A}^+_*$ is uniquely decomposed as

$$A = \lambda R_1(\psi) T_t R_2(\phi) \quad (7)$$

where R_1, R_2 are rotations and $T_t = \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix}$ with $t > 1, \lambda > 0, \phi \in [0, \pi)$ and $\psi \in [0, 2\pi)$. Furthermore, the above decomposition comes with a geometric interpretation (see Fig. 2) where the longitude ϕ and latitude $\theta = \arccos \frac{1}{t}$ characterize the camera's viewpoint angles, ψ parameterizes the camera spin and λ corresponds to the zoom.

Most local descriptors and their corresponding matching methods are similarity-invariant. Unfortunately, slanted camera viewpoints (measured by θ) will deteriorate the performance of almost any state-of-the-art matching method. To compensate this degradation at a minimum cost of complexity, we follow the ideas developed in [10] to compute optimal sets of affine simulations for each of our methods depending on the viewpoint tolerances, which we shall estimate in the next section. Under these conditions, Proposition 3.6 in [10] ensures that the constructed IMAS method is affine-invariant in practice. Indeed, there is at least one pair of simulated images whose viewpoint angle is not greater than the viewpoint tolerance of the matching method in question.

Matching method	True matches	Ratio of true matches
SIFT L1 0.8	74	0.6577
SIFT L1 0.6	15	0.8054
RootSIFT 0.8	135	0.5383
RootSIFT 0.6	49	0.7812
AC-W	691	0.8679
AC-Q ($\rho = 0.3$)	1881	0.6261

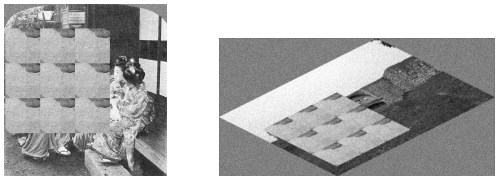
Table 1: Frontal performance test (i.e. $A \in \mathbf{S}$). Mean values over 100 iterations.

5. EXPERIMENTS

The main objective of our two matchers is to allow repetitions to be captured. On the other hand, state-of-the-art descriptors are robust against noise, and Lowe's second-closest neighbor criterion [2] is well known to render SIFT distinctive enough to be practical. All these properties are met for our methods even in the presence of viewpoint changes. A simple methodology is proposed to assess this claim.

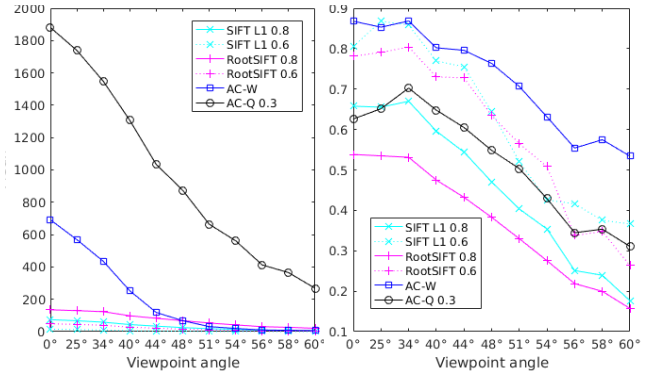
The following procedure allows us to generate any number of test images u (query) and v (target) with the corresponding ground truth. Fig. 3 shows an example. Let us consider three different and sufficiently distinct images, u_0, v_0, and w_0. The test pair is generated randomly in four steps: 1) A $N \times N$ patch is extracted from a random position in w_0; a repetitive pattern P is composed by repeating the patch into a $M \times M$ mosaic. 2) The pattern P is pasted into image u_0 at a random position, producing image u_1; similarly, the same pattern P is pasted in a random position of v_0 to produce image v_1. 3) A random affine transform A is selected and used to optically simulate a distortion, $v_2 = Av_1$. 4) Finally, Gaussian noise is added to produce the final images, $u = u_1 + n_u$ and $v = v_2 + n_v$. Forcing $A \in \mathbf{A}^+_*$ in step 3 will incur in a change of point of view in v with respect to u; the viewpoint angle can be selected.

This framework was used to compare systematically our methods to SIFT, RootSIFT and their affine invariant ver-



(a) Query image. (b) Target image. $A = T_2 R_{\pi/4}$.

Fig. 3: A generated image pair with repetitive structures.



(a) Total number of true matches. (b) Ratio of true matches.

Fig. 4: Oblique performance test. Each point represents the resulting mean over 100 iterations.

Matching method	Maximal viewpoint tolerance
SIFT L1 0.8	48°
SIFT L1 0.6	34°
RootSIFT 0.8	40°
RootSIFT 0.6	54°
AC-W	58°
AC-Q ($\rho = 0.3$)	54°

Table 2: Viewpoint tolerances (i.e. tilt tolerances from [10]) obtained from the oblique performance test of Fig. 4 with the convention that the ratio of true matches ≥ 0.5 and the total number of true matches ≥ 10.

sions. Lowe's criterion was applied for two match ratios (0.6 and 0.8). For each method and image pair, the *total number of true matches* and the corresponding *ratio of true matches* were computed. A match is considered as true if both constituting key-points lie inside the pattern P at the same position, modulo the repeated patch size. The displayed values are means after repeating the process for different generated image pairs. As the key-point extraction part is identical for all compared methods, the figures reflect only the performance of the descriptors and their matchers.

The frontal performance test of Table 1 confirms that our descriptors do handle repetitions and noise while still preserving a good ratio of true matches. Fig. 4 illustrates the benefits of our methods for varying viewpoint angle. Notice, however, the drastic fall in number and in ratio of true matches for all

Matching method	True matches	Ratio of true matches
Affine SIFT L1 0.8	33	0.4095
Affine SIFT L1 0.6	5	0.6463
Affine RootSIFT 0.8	48	0.2462
Affine RootSIFT 0.6	12	0.4934
Affine AC-W	195	0.7564
Affine AC-Q ($\rho = 0.3$)	913	0.2268

Table 3: Hard oblique performance test on affine invariant methods. The viewpoint angles are random and uniformly distributed between 60° and 80°. Mean values over 200 iterations.

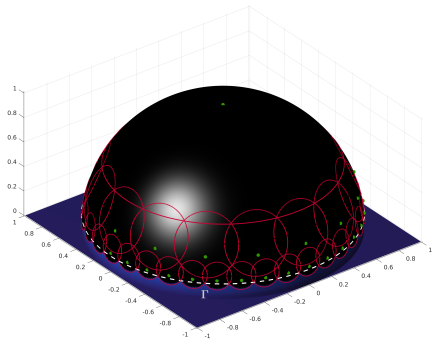


Fig. 5: Optimal set of affine simulations for a method with viewpoint tolerances of 58°. Just 27 are enough to obtain an IMAS extension to 80°. Affine camera simulations (green); viewpoint tolerance from each simulation (red); visible viewpoints (black); maximal viewpoint tolerance for the IMAS method (dashed line).

methods. Table 2 provides the estimated maximal viewpoint tolerances from the statistics presented in Fig. 4; this brings to light a degradation in viewpoint tolerances (due to repetitive structures) for SIFT and RootSIFT with respect to results presented in [10] (respectively, 56° and 60°).

The theory on IMAS algorithms [10, 4] leads to optimal sets of affine simulations for each method, depending on viewpoint tolerances in Table 2. Fig. 5 provides a geometrical representation of the optimal set of simulations for AC-W. Table 3 show the results for the affine invariant version of the methods in viewpoint angles form 60° to 80°. AC-W gets the overall best results; AC-Q produces significantly more good matches at the cost of a lower ratio of true matches.

6. CONCLUSION

We described methods for image comparison based on a new descriptor, two *a-contrario* matchers and affine simulation. Our experiments show that the proposed methods produce better results than state-of-the-art methods in the presence of repetitive structures, different viewpoints and noise. Future work will concentrate on combining our two methods in an attempt to get the best of both, a large number of true matches while keeping a high ratio against false ones.

Acknowledgments: We thank Julie Delon and Jean-Michel Morel for numerous suggestions.

Reproducibility: A source code containing the proposed affine-invariant methods as well as other state-of-the-art methods is available at `https://rdguez-mariano.github.io/pages/acdesc`. The reader will also find a guide to reproduce results appearing in this paper.

7. REFERENCES

[1] Petr Doubek, Jiri Matas, Michal Perdoch, and Ondrej Chum, "Image matching and retrieval by repetitive patterns," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 3195–3198.

[2] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[3] I. Rey-Otero and M. Delbracio, "Anatomy of the SIFT method," *IPOL*, vol. 4, pp. 370–396, 2014.

[4] J. M. Morel and G. Yu, "ASIFT: A new framework for fully affine invariant image comparison," *SIIMS*, vol. 2, no. 2, pp. 438–469, 2009.

[5] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *TPAMI*, vol. 24, no. 4, pp. 509–522, 2002.

[6] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *CVPR*, 2007, pp. 1–8.

[7] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *IJCV*, vol. 60, no. 1, pp. 63–86, 2004.

[8] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *TPAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.

[9] P. Moreels and P. Perona, "Evaluation of features detectors and descriptors based on 3d objects," in *ICCV*, 2005, pp. 800–807.

[10] M. Rodriguez, J. Delon, and J. M. Morel, "Covering the space of tilts. application to affine invariant image comparison," *SIIMS*, accepted for publication in 2018, Preprint: `https://hal.archives-ouvertes.fr/hal-01589522/document`.

[11] T. Lindeberg, *Scale-Space Theory in Computer Vision.*, Royal Institute of Technology, Stockholm, Sweden, 1993.

[12] T. Lindeberg, "Scale selection properties of generalized scale-space interest point detectors," *JMIV*, vol. 46, no. 2, pp. 177–210, 2013.

[13] Yanwei Pang, Wei Li, Yuan Yuan, and Jing Pan, "Fully affine invariant SURF for image matching," *Neurocomputing*, vol. 85, pp. 6–10, 2012.

[14] D. Mishkin, J. Matas, and M. Perdoch, "MODS: Fast and robust method for two-view matching.," *Computer Vision and Image Understanding*, vol. 141, pp. 81–93, 2015.

[15] F. Cao, J. L. Lisani, J. M. Morel, P. Musé, and F. Sur, *A Theory of Shape Identification*, Springer, 2008.

[16] A. Desolneux, L. Moisan, and J.-M. Morel, *From Gestalt Theory to Image Analysis*, Springer, 2008.

[17] J. Rabin, J. Delon, and Y. Gousseau, "A statistical approach to the matching of local features," *SIIMS*, vol. 2, no. 3, pp. 931–958, 2009.

[18] R. Grompone von Gioi and V. Pătrăucean, "A contrario patch matching, with an application to keypoint matches validation," in *ICIP*, pp. 946–950. 2015.

[19] A. P. Witkin and J. M. Tenenbaum, "On the role of structure in vision," in *Human and Machine Vision*, J. Beck, B. Hope, and A. Rosenfeld, Eds., pp. 481–543. Academic Press, 1983.

[20] D. Lowe, *Perceptual Organization and Visual Recognition*, Kluwer Academic Publishers, 1985.

[21] A. Gordon, G. Glazko, X. Qiu, and A. Yakovlev, "Control of the mean number of false discoveries, bonferroni and stability of multiple testing," *Ann. Appl. Stat.*, vol. 1, pp. 179–190, 2007.

[22] V. Pătrăucean, *Detection and Identification of Elliptical Structure Arrangements in Images: Theory and Algorithms*, Ph.D. thesis, Institut National Polytechnique de Toulouse, France, 2012.

[23] J. L. Marichal and M. J. Mossinghoff, "Slices, slabs, and sections of the unit hypercube," *Online Journal of Analytic Combinatorics*, 2006.