21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Combining color and geometry for local image matching

Baptiste Mazin, Julie Delon and Yann Gousseau
Telecom ParisTech, LTCI CNRS
{*mazin,delon,gousseau*}*@telecom-paristech.fr*

Abstract

This paper introduces a generic way to incorporate color information into local, SIFT-like descriptors, in view of image matching. First, a new color descriptor, relying on local hue histograms, is introduced. Second, we describe a procedure permitting the automatic setting of matching parameters when matching images using both geometric and color information. Experiments on a color image database show that our SIFT+Hue combination performs significantly better than classical color descriptors.

1 Introduction

Local image descriptors are ubiquitous computer vision tools that are routinely used in applications ranging from image matching to scene understanding. Among those descriptors, SIFTs [7] and its many variants are known for having both strong invariance properties and discriminative power. These descriptors encode the local geometric information through histograms of (gray level) gradient orientation and therefore disregard color information. Even though the luminance channel of a color image arguably conveys most of its geometric information, it is of interest to investigate how SIFT-like descriptors may be enriched by color information and to what extent this is useful for image matching.

Several approaches have been proposed in this direction. A first trend is to compute SIFT descriptors directly on color channels of images: HSV decompositions [3] or invariant color channel [4, 1, 12]. An alternative approach is to enrich the geometrical SIFT descriptor with some color information extracted in its neighborhood, typically through local statistics. This is the case in [2], where it is suggested to use cooccurrence histograms on normalized RGB channels, in [9], where a local Luv descriptor is used or in [13], where, among others, a local hue histogram is combined with the SIFT descriptor. Performances of both these trends highly depend on the considered database, on the type of considered images and on the acquisition conditions. However, a recent state of the art paper [12] concludes to the superiority of the *OpponentSIFT* descriptor, obtained by computing the original SIFT descriptor on the opponent color channels of the image (see Section 2), confirming a tendency already shown in a previous review paper [4].

The goal of the present paper is twofold. First, we show that by localizing the extraction of hue histograms to the sectors of a SIFT-like descriptor, we get a very efficient yet simple to compute local color descriptor. In particular, we show in an experimental section that this descriptor compares favorably with OpponentSIFT for image matching. Second, we show how to combine the geometrical information (coded through SIFT-like descriptors) and color information in the context of image matching. This step is non-trivial, in particular because the respective contribution of geometry and color may strongly vary depending on the image or the particular keypoint at hand. To the best of our knowledge, this is the first proposition of a generic matching procedure enabling an efficient combination of such features. It relies on the *a contrario* methodology [5], and follows the approach from [10]. For a given query descriptor, the matching threshold is computed automatically and adapts to the variety of geometric and chromatic similarities between both the query and the database.

The plan of the paper is as follows. In Section 2, we introduce geometric and color local descriptors. In Section 3, we define the generic matching procedure for the combination of geometry and color. Eventually, in Section 4, we validate the proposed procedure through experiments on a large color image database, involving about three million comparisons between descriptors.

2 Local color descriptors

In this section, we introduce the local descriptors to be considered in the paper. These are: a variation on the classical SIFT descriptor, introduced in [10], the recent OpponentSIFT [12] and a new hue-based descriptor.

SIFT-like descriptor In order to code the geometric

978-4-9906441-1-6 ©2012 IAPR

information, we rely on the SIFT-like descriptors proposed in [10], that are very similar to the original descriptor from [7]. These are extracted from the luminance channel of the considered image. Descriptors are computed around local keypoints (extrema of the Laplacian pruned by a multi-scale Harris criterion). To each keypoint, a circular region made of 9 concentric sectors is considered. On each sector, a histogram of gradient orientations weighted by the gradient magnitude is computed. The final descriptor is made of the concatenation of the 9 histograms. More details about these descriptors may be found in [10].

OpponentSIFT The so-called OpponentSIFT have been shown recently to be the most efficient for a generic categorization application [12]. These are obtained from the opponent color channels, defined as

$$O_1 = \frac{R-G}{\sqrt{2}},\ O_2 = \frac{R+G-2B}{\sqrt{6}},\ O_3 = \frac{R+G+B}{\sqrt{3}}. \tag{1}$$

Two opponent descriptors are obtained by computing SIFT descriptors independently on channels O_1 and O_2. The descriptor computed on O_3 is identical to the geometrical descriptor. In order for the descriptor to be robust to a change in the color temperature of the illuminant, the R, G and B values from which opponent are computed are locally normalized. For each channel, values are divided by their average over the descriptor. This yields invariance with respect to any diagonal transform of the RGB values, a reasonable approximation for a change of illuminant [13].

Hue descriptor In this paragraph, we introduce a new local color descriptor obtained from the distribution of hue values. This descriptor shares similarities with the hue descriptor from [13], but is more local. To each sector of the luminance SIFT-like descriptor, a hue histogram is computed. Hue values are weighted by saturation values. The whole descriptor is the concatenation of the resulting 9 histograms, as for geometric descriptors. The interest of using 9 local histograms instead of a global one will be demonstrated in the experimental section.

We now specify the definitions of hue and saturation that we use to build the descriptor. In the usual HSV space [11], saturation is defined as $1 - \frac{\mathtt{min(R,G,B)}}{\mathtt{max(R,G,B)}}$. This definition, whose normalization aims at keeping the same dynamic whatever the luminance value, is unstable when the luminance is small. To overcome this issue, we use a HSL-type space, similar to the one proposed in [6], obtained by converting RGB coordinates into cylindrical ones. In this space, saturation is defined as $S = \max(R,G,B) - \min(R,G,B)$ and hue is the angle between the opponent color channels O_1 and O_2, that is, $H = \arctan\frac{O_2}{O_1}$. In order to be robust to illuminant changes, the R, G and B values are first locally normalized as for the opponent descriptors.

Observe that an advantage of this definition is that hue is invariant under multiplication of the luminance by a constant value. This quantity, and therefore the corresponding local histograms, are thus rather independent from the information conveyed by the SIFT descriptors. This, as we will see in the next section, is an important asset for multi-modality image matching.

Descriptor terminology The descriptors introduced previously will from now on be denoted by a_g for the geometrical (SIFT-like) descriptor, a_{o_1} and a_{o_2} for the OpponentSIFT and a_h for the hue descriptor.

3 Combining descriptors for *A contrario* matching

A generic method for matching descriptors In this section, we recall the matching procedure introduced in [10] for the matching of SIFT-like features. This approach draws on the *a contrario* methodology [5] to compute thresholds on the distances between local descriptors, thereby automatically selecting which descriptors should be matched. Thanks to a simple learning procedure, this matching methodology adapts to the number and types of descriptors. This enables the exact same procedure to be applied to a wide variety of images, in a more stable way (see [10]) than the classical *ratio of distances to the first and second neighbors* [7]. Another asset of this approach is that no restriction of matches to the nearest neighbor is required.

We consider Q a *query* image, from which N_Q local descriptors $\{\mathbf{a}^j\}_{j=1\dots N_Q}$ have been extracted. We also consider a *candidate* image C from which N_c local descriptors $\{\mathbf{b}^l\}_{l=1\dots N_C}$ have been extracted. We assume that the distance $D(\mathbf{a}^j, \mathbf{b}^l)$ between a query and a candidate descriptors may be written as a sum of distances:

$$D(\mathbf{a}^j, \mathbf{b}^l) = \sum_{k=1}^{M} d_k(a^j, b^l), \tag{2}$$

where the d_k, $k = 1, \dots, M$ rely on subparts of the descriptors (for instance, in SIFT-like descriptors, d_k is a distance between gradient histograms of the k^{th} sectors of a^j and b^l).

In order to automatically set matching thresholds the idea is then, for a given query descriptor $\mathbf{a}^j$, to compute the probability law of the distance from $\mathbf{a}^j$ to a generic random candidate descriptor $\mathbf{b}$. We then choose a matching threshold T so that $Pr(D(\mathbf{a}^j, \mathbf{b}) \leq T)$ is small. In a nutshell, we match descriptors whose proximity is hardly due to chance.

More precisely, the notion of generic descriptor is defined as follows. For a given query descriptor $\mathbf{a}^j$, a random descriptor $\mathbf{b}$ is said to follow the null hypothesis $\mathcal{H}_0^j$ if distances $d_k(a^j, b)$ are mutually independent random variables, for $k = 1, \ldots M$. Under this hypothesis, the probability of observing a distance $D(\mathbf{a}^j, \mathbf{b})$ smaller than a given threshold T is then

$$\int_0^T \mathop{*}_{k=1}^{M} p_k^j(x)dx,$$

where $*$ refers to the convolution product and where p_k^j is the probability density for the variable $d_k(a^j, b)$. In order to compute this integral, these probability densities are learned as the empirical histograms of sub-distances $d_k(a^j, b^i)$ when $\mathbf{b}^i$ spans the candidate descriptors (descriptors of the candidate image). The last step to set the threshold is to define a *number of false alarms*. It is given, for two given descriptors $(\mathbf{a}^j, \mathbf{b}^l)$ by

$$\mathrm{NFA}(\mathbf{a}^j, \mathbf{b}^l) = N_C N_Q \cdot \int_0^{D(\mathbf{a}^j, \mathbf{b}^l)} \mathop{*}_{k=1}^{M} p_k^j(x)dx. \quad (3)$$

It may be shown that if these quantities are uniformly thresholded by ϵ, then, under the hypotheses $\mathcal{H}_0^j$, the average number of false matchings when testing the N_q queries against the N_c candidates is bounded by ϵ. As shown in [10], thresholding the NFA is more robust than directly thresholding distances or distances ratios.

Geometry+color matching In what follows, we take advantage of this generic formalism to combine geometric (SIFT-like) and color information in view of image matching. The previous approach relies on the hypothesis that distances $d_k(a^j, b)$ can be viewed as mutually independent random variables when $\mathbf{b}$ is a random descriptor. The choice of the color-geometry representation should follow this hypothesis. It is reasonable to assume that the hue information is independent from the geometry. More precisely, we claim that hue descriptors, as described in Section 2, are fairly independent from the derivative distributions contained in the SIFT and OpponentSIFT descriptors. On the contrary, the gradient information of the luminance and opponent channels are clearly highly correlated and should be treated as such.

For these reasons, the distance D is defined as a sum of $M = 10$ distances d_k chosen in the following way. For $k = 1, \ldots 9$,

$$d_k(a, b) = w_1 \sum_{s \in g, o_1, o_2} \left(\mathrm{cemd}(a_s(k), b_s(k))\right), \quad (4)$$

where $a_s(k)$ denotes the k^{th} gradient histogram of the descriptor a_s, cemd is the circular earth mover's distance introduced in [10] and w_1 is a weighting parameter. These distances represent gradient information averaged on the luminance and opponent channels. Observe that in a random noise image, these distances, computed on different subregions, can be seen as independent random variables. For $k = 10$, we define

$$d_{10}(a, b) = w_2 \left(\sum_{j=1}^{9} \mathrm{cemd}(a_h(j), b_h(j)) \right). \quad (5)$$

That is, we average hue distances on the different sub-regions of the descriptor spatial support. This choice is mainly heuristic and comes from the fact that independence between sectors is generally less valid for hue information than for geometric information. On the other hand, as claimed before, this distance d_{10} is reasonably independent from the d_k, $k = 1, \ldots, 9$.

Matches can then be validated by thresholding the quality measure defined in Equation (3). As already said, the learning of distance distributions yields a matching procedure adaptive to both the query descriptor and the candidate image. This is a fundamental advantage for combining color and geometry. Indeed, for a given query descriptor $\mathbf{a}^j$, learning p_{10}^j enables to adapt to the amount of saturated color this descriptor contains. For instance, if it contains few or no color, distribution p_{10}^j will contain a strong mass at 0 (because all gray descriptors in the candidate image will be at a small distance d_{10}). Therefore, the hue component will have little influence on the value $\mathrm{NFA}(\mathbf{a}^j, .)$. Such a descriptor will therefore be compared to other gray descriptors mostly relying on geometry. On the other hand, its large d_{10} distance to other colored descriptors will penalize such matches.

4 Experiments

This section presents several image matching experiments, in which the performances of various descriptor combinations and various matching criteria are compared. Experiments rely on a database made of 708 generic[1] color photography, also used in [10]. Our experimental protocol is similar to the one introduced in [10]. Each image A of the database is matched with A', a modified version of itself undergoing an affine transformation and a simulated change in the color temperature of the illuminant, and B, an independent image. A match between keypoints [2] in A and A' is said to be correct if the surface of the corresponding descriptors overlap enough with the surface of the correct descriptor (more than 50 %), a similar procedure as in [8, 10]. A match between A and B is always considered as false. The total number of false matches is obtained as the addition of false matches in A' and B. Performances are

[1] Images and can be found at `http://perso.telecom-paristech.fr/~gousseau/db/imageDB.zip`

[2] Since we wish to evaluate descriptors and matching procedures, and not keypoints, these are extracted for each image A and then projected in A'.

then evaluated through ROC-curves, plotted from the overall number of good and false matches over the complete image database, therefore involving several million comparisons.

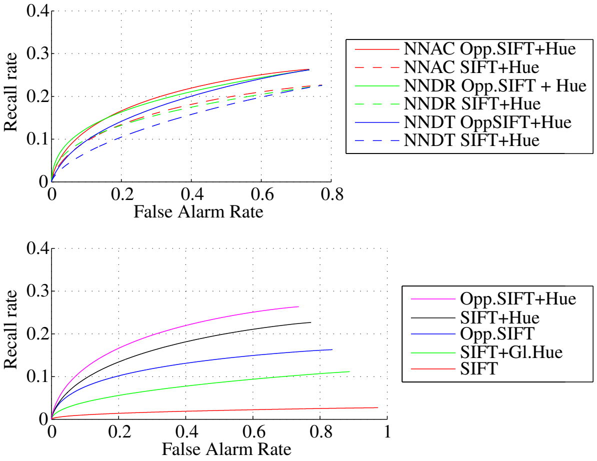

Figure 1. Comparison of several descriptor combinations and matching criteria.

We use several combinations of descriptors and matching criteria. Descriptors combinations are called SIFT [3], SIFT + Hue and OpponentSIFT (using all three opponent channels, that is, including the usual SIFT descriptor). We also consider the concatenation of SIFT and a hue descriptor, obtained by associating only one weighted hue histogram to each descriptor region (instead of one per sector) as proposed in [13]; the resulting combination will be called SIFT + GlobalHue. Eventually, we consider the simultaneous use of OpponentSIFT + Hue. All histograms are quantized on 12 bins. In the sum (2), each descriptor is given the same weight. In the case of OpponentSIFT + Hue, this reads $w_1 = 1/36$ ($1/4$ for each opponent descriptor, each made of 9 sub-descriptors) and $w_2 = 1/4$. In the case SIFT+Hue, this yields $w_1 = 1/9$ and $w_2 = 1/2$. We also consider the following matching criteria : the *a contrario* criterion proposed in this paper (called AC), a simple threshold on distances (called DT) and the classical threshold on the ratio between the first and second match, introduced by D. Lowe [7] (called DR). Matching results involving these different choices may be seen on Figure 1.

Several conclusions may be drawn from these curves. For image matching and with the proposed protocol (relatively strong affine transforms and color temperature changes) the combination SIFT + Hue permits to consequently improve the performances of SIFT alone for image matching. Second, the combination SIFT + Hue yields significantly better results than both OpponentSIFT and SIFT + GlobalHue. Last, the proposed *a contrario* matching criteria is more stable than the classical DT and DR criteria, which may be seen from the global ROC curves. To conclude, we first have introduced a new color descriptor relying on local hue histograms, whose capacity to enrich SIFT as well as Opponent SIFT descriptors has been demonstrated. This is shown by using a generic matching procedure, relying on the *a contrario* methodology and some independence assumptions between descriptors. Strong assets of the resulting procedure are its adaptivity as well as its ability to automatically set thresholds on distances.

Aknowledgement This work is partially founded by the ANR project Callisto.

References

[1] A. E. Abdel-Hakim and A. A. Farag. CSIFT: A SIFT descriptor with color invariant characteristics. In *Conference on Computer Vision and Pattern Recognition*, pages 1978–1983, 2006.

[2] C. Ancuti and P. Bekaert. SIFT-CCH: Increasing the sift distinctness by color co-occurrence histograms. 1999.

[3] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *European Conference on Computer Vision*, 2006.

[4] G. J. Burghouts and J. M. Geusebroek. Performance evaluation of local colour invariants. *Computer Vision and Image Understanding*, 113:48–62, 2009.

[5] A. Desolneux, L. Moisan, and J. M. Morel. *From Gestalt Theory to Image Analysis*, volume 34. Springer-Verlag, 2008.

[6] A. Hanbury. Constructing cylindrical coordinate colour spaces. *Pattern Recogn. Lett.*, 29:494–500, March 2008.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[8] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1615–1630, 2005.

[9] P. Quelhas and J.-M. Odobez. A Color and Gradient Local Descriptor Fusion Scheme For Object Recognition. In *Proceedings of WIAMIS04*. IDIAP, 2004.

[10] J. Rabin, J. Delon, and Y. Gousseau. A statistical approach to the matching of local features. *SIAM J. Imaging Sciences*, 2(3):931–958, 2009.

[11] A. R. Smith. Color gamut transform pairs. *SIGGRAPH Comput. Graph.*, 12:12–19, August 1978.

[12] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[13] J. Van De Weijer and C. Schmid. Coloring local feature extraction. In *European Conference on Computer Vision*, pages 334–348, 2006.

[3] We rely in these experiments on our SIFT-like descriptors, described in Section 2. For the sake of simplicity, we refer to them as SIFT all the same.