A FRAMEWORK TO IMPROVE MATCHING RESULTS OF WIDELY SEPARATED VIEWS

Cosmin Ancuti, Codruta Orniana Ancuti and Philippe Bekaert
Hasselt University - tUL -IBBT, Expertise Center for Digital Media, Belgium

Keywords: Local feature points, Matching, SIFT, Color, Wide-baseline.

Abstract: Matching images is a crucial step in many computer vision applications. In this paper we present an alternative strategy built on the SIFT operator to solve the problem of wide-baseline matching. We first show how to add the color information to the SIFT descriptors of extracted keypoints. Practically, the SIFT descriptor vector is blended with the main parameters (contrast, correlation and energy) of the color co-occurrence histogram computed in the same image patch. Afterward, in order to better improve the matching results of images taken under large variations of the camera viewpoint angle, the valid matches obtained by the previous strategy are employed to estimate the geometry between patches of corresponding keypoints. This overcomes the lack of affine invariance of the existing operators (including SIFT), allowing to use a more appropriate region shape where descriptors will be calculated for better preciseness. In our experiments the proposed method shows a substantial improvement of the matching results compared with the results obtained by the original local operator.

1 INTRODUCTION

Matching images that represent projections of the same 3D scene/object is a fundamental task in computer vision. Several important applications such as 3D reconstruction, camera calibration, panoramic images, texture and object classifying, image retrieval, robot localization rely on the accuracy of this task. The problem of image matching is in general solved base on local feature points. The feature points (keypoints, interest points) are those locations where the image has significant variation in at least two directions. First, a certain number of local feature points are extracted independently in both images. For efficiency, these locations are filtered by employing an invariant detector in order to extract feature points with a high repeatability ratio. Secondly, the extracted feature points are described as distinctive as possible based on the information contained by their neighbor regions. Finally, the corresponding feature points are found by computing different distance criteria (e.g. Euclidean, Mahalanobis) between descriptors vectors.

There has been much research in image matching in the last decades. Apparently, the most known detector was introduced by Harris (C.G.Harris and Stephens, 1988). This basic detector is invariant only to rotation and translation and fails for more complex geometric transformations like modification of the scale or truly affine. Lindeberg solved the scale invariance of the detectors introducing the automatic scale selection principle (Lindeberg, 1999). He searched for 3D maxima in the Laplacian of Gaussian (LoG) scale space. Lowe (Lowe, 2004) approximates the LoG scale space through Difference of Gaussian (DoG). Recently (Mikolajczyk and Schmid, 2004b) the basic Harris and Hessian detectors have been adapted to scale and affine spaces.

In order to find correct matches of images of the same scene the feature points have to be described as distinctive as possible. The simplest method is to compute cross correlation between vectors of pixels from certain regions. Unfortunately this approach fails in real cases when the similar patches of the images are related by complex geometric transformations. Different approaches (Freeman and Adelson, 1991; Belongie et al., 2002; Mikolajczyk and Schmid, 2004a; Lowe, 2004) of descriptors have been introduced in the literature. SIFT (Lowe, 2004) computes a histogram of gradient locations and orientations and has been shown to outperform the other descriptors (Mikolajczyk and Schmid, 2004a). More re-

Ancuti C., Orniana Ancuti C. and Bekaert P. (2010).
A FRAMEWORK TO IMPROVE MATCHING RESULTS OF WIDELY SEPARATED VIEWS.
In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 475-479
DOI: 10.5220/0002843004750479

cently, SURF (Bay et al., 2006) uses Hessian matrix and Haar wavelets response combined with the properties of integral images in order to speed up the processing time.

The most challenging problem is to match images of the same scene taken under significant variation of the camera viewpoint position. The region-based methods (Tuytelaars and Gool, 2000; Matas et al., 2002; Tuytelaars and Gool, 2004; Forssén and Lowe, 2007) identifies salient corresponding image regions. In general the local feature-point based approaches (Pritchett and Zisserman, 1998; Baumberg, 2000; Xiao and Shah, 2003) are more robust to occlusions and cluttering, finding a higher number of correct matches but also being characterized by less processing time than region-based methods.

Recent studies (Mikolajczyk and Schmid, 2004a; Moreels and Perona, 2007) disclosed that none of the existing operators is fully invariant to affine changes. Moreles and Perona (Moreels and Perona, 2007) observed that when the difference between camera viewpoint angles is significant (higher than 25-30°) the state of the art detectors/descriptors fail. This is due to the clasped region shape where the descriptors are computed.

This paper presents an alternative strategy built on the local feature points that aims to improve the matching results for extreme cases where the difference between camera viewpoint angles varies significantly. Our approach is built on the well-known SIFT operator. Since SIFT has been designed only for grayscale images and neglects the color, we first show how to add effectively the color information in order to increase the descriptor distinctness. Additionally, to increase the affine invariance of the local feature points, after extracting several valid matches, an approximate geometry is computed between their neighbor patches. The estimated geometric parameters are used in order to define a new shape of the region where descriptors vectors will be computed. The method has the advantage that does not employ expensive refinements (e.g Ransac, epipolar constraints). Moreover, the method is general being possible to be used in combination with other operators, too. The comparative results demonstrates the utility of our method that is able to find a considerable additional number of correct matches.

The reminder of the paper is organized as following. In the next section is presented the strategy to add effectively the color information to the SIFT descriptors. Then, we show how the geometry that relates the corresponding patches of valid matches can be estimated. Finally, before concluding, we presents and discuss several comparative results.

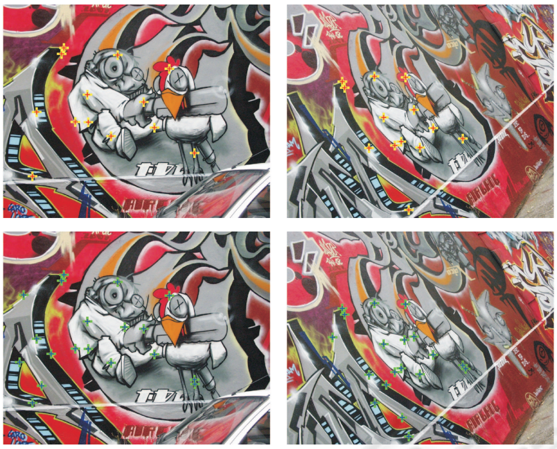

Figure 1: Adding color to SIFT descriptor. In the top part of the figure are shown the results obtained by SIFT (13 valid matches) while by adding the color (bottom part of the figure) we obtained 19 valid matches.

2 INCREASE DISTINCTNESS OF DESCRIPTORS BY COLOR

The recent studies (Mikolajczyk et al., 2005; Moreels and Perona, 2007) proved that the most effective local operator to match images is SIFT (Scale Invariant Feature Transform) (Lowe, 2004). Several attempts (Ke and Sukthankar, 2004; Abdel-Hakim and Farag, 2006) tried to improve some parts of the original implementation but the original version still remains the most reliable. In the following we give a brief presentation of how this operator is computed. The feature points are scale invariant and are searched in the DoG (Difference of Gaussian) scale space. The DoG is built by subtracting images that previously have been convolved (blurred) with a Gaussian function with a standard deviation that increases monotonically. A keypoint is extracted only if its value is greater or smaller than all its 26 neighbors. Additionally, the keypoints with strong response to edges and low contrast are rejected.

The signature (descriptor) computation is based on the image gradient magnitudes and orientations calculated in the circular neighbor regions of the feature points. The image pyramid level is determined by the computed characteristic scale of the respective feature point. For every feature point a 4x4 orientation histogram is constructed on a 4x4 sub-region of the feature point computed from a 16x16 centered region. Each histogram has 8 bins corresponding to every 45°.

However, the SIFT was designed only for grayscale images neglecting the important information of color. Therefore, in this section we describe a strategy to increase the distinctness of the local feature points descriptors by color. Our approach is built on the color co-occurrence histogram (Chang and Krumm, 1999), an extension in the color space of the well known co-occurrence matrix (Haralick et al., 1973) that estimates the spatial gray level dependencies of the pixels. Given a color pixel p_1 in the image, the color co-occurrence histogram (CCH) counts the number of occurrences of the color pixel pair (p_1, p_2), with p_2 representing an adjacent color pixel located at the distance $d = (\Delta x, \Delta y)$. The color co-occurrence histogram can be seen as a function of the color pixel values and the displacement vector between them. For a given image patch $\mathcal{P}$ of the size N×N the CCH value is counting the number of times when the pixel pair (p_1,p_2) matches the the color combination (c_1, c_2):

$$CCH(x,y,c_1,c_2) = \sum_{x=1}^{N}\sum_{x=1}^{N} \Psi_{c_1}(x,y) \sum_{\Omega} \Psi_{c_1}(x+\Delta x, y+\Delta y) \tag{1}$$

where Ω represents the number of pixels located at the distance $(\Delta x, \Delta y)$ while the function Ψ is given by the following expression:

$$\Psi_{c_k}(x,y) = \begin{cases} 1 \quad , \qquad c(x,y) = c_k \\ \\ 0 \quad , \qquad otherwise \end{cases} \tag{2}$$

where $c(x,y)$ is the color level of the pixel located at (x,y) and c_k is a color level. We compute the CCH in the same neighbor patches of the extracted keypoints. In our experiments the number of color levels is reduced by a standard k-mean quantization to a value of $n_c = 256$ while the CCH is computed only for an offset distance $d = (1,1)$.

After the color co-occurrence histogram matrix is normalized its elements are referred as $P_{i,j}$. We compute its main parameters: contrast $C = \sum_{i,j} P_{i,j}(i-j)^2$, correlation $Cor = \frac{1}{\sigma_i \sigma_j}\sum_{i,j}(1-\mu_i)(1-\mu_j)P_{i,j}$ and energy $E = \sqrt{\sum_{i,j} P_{i,j}^2}$ in order to built the new descriptor vector that blends the SIFT descriptor and the three parameters of the CCH computed in the same surrounding region of the filtered keypoints. In our experiments we consider that the original SIFT descriptor has a weight impact of 0.7 while each of the CCH parameters contribute in the final descriptor vector with a weight factor of 0.1. To find the valid matches we use the same strategy as Lowe (Lowe, 2004) by evaluating the distance computed between the first best match and the second best match.

3 MORE EFFECTIVE MATCHING

As observed in our experiments and shown in figure 1 by adding the color information to the SIFT descriptors the matching results are only partially improved. In order to find a more important additional number of valid matches we adopt the following method. Supposing that several correct corresponding points have been filtered by the previously presented strategy, in order to find the optimal shape region where descriptors are computed we estimate the geometry that relates the patches of two corresponding feature points. In the worst scenario when no valid corresponding points have been found two corresponding locations in the input images are selected manually.

Limiting the geometry only up to affine the relation between the surrounding regions $\mathcal{P}_1$, $\mathcal{P}_2$ of two corresponding keypoints is expressed as following:

$$\gamma \mathcal{P}_1(\mathcal{A}x + d) + \eta = \mathcal{P}_2(x) \tag{3}$$

where $\mathcal{A}$ is a 2D affine matrix, d is the translation vector, γ is the reflection angle of the light source while η represents the camera gain. Since in our experiments we consider only small photometric variations between images the last two parameters (γ, η) are approximated to the unit value.

Finding the optimal geometric transformation of the corresponding patches can be seen as minimization problem. In order to find the optimal parameters the following energy function is minimized:

$$E = \|\mathcal{P}_1\mathcal{A}(\alpha, s, h_1, h2) - \mathcal{P}_2\|^2 \tag{4}$$

where the parameters of the affine transformation are represented by the rotation angle α, scale ratio s, shearing h_1 and stretching h_2. The process of convergence is very sensitive and for decent output results the initial values of the affine matrix parameters need to be as close as possible from their real values. The camera rotation is contained by the matrix $\mathcal{R}(\alpha)$, the isotropic scale $\mathcal{S}$ is specified by the parameter s and the shearing and stretching matrix Γ is expressed by an expansion factor in a considered direction and a contraction factor on a perpendicular direction. In our approach the initial value, from where the minimization process starts, is approximated by $\mathcal{A}_0 = \mathcal{R}(\alpha_0)\mathcal{S}(s_0)$.

The initial value of the scale s_0 is determined based on the automatic scale selection principle (Lindeberg, 1999). Lindeberg postulates that in the absence of additional evidence, the selected scale (characteristic scale) is the scale where a function of some combinations of normalized derivatives attains a local maximum. Due to the fact that we extract fea-

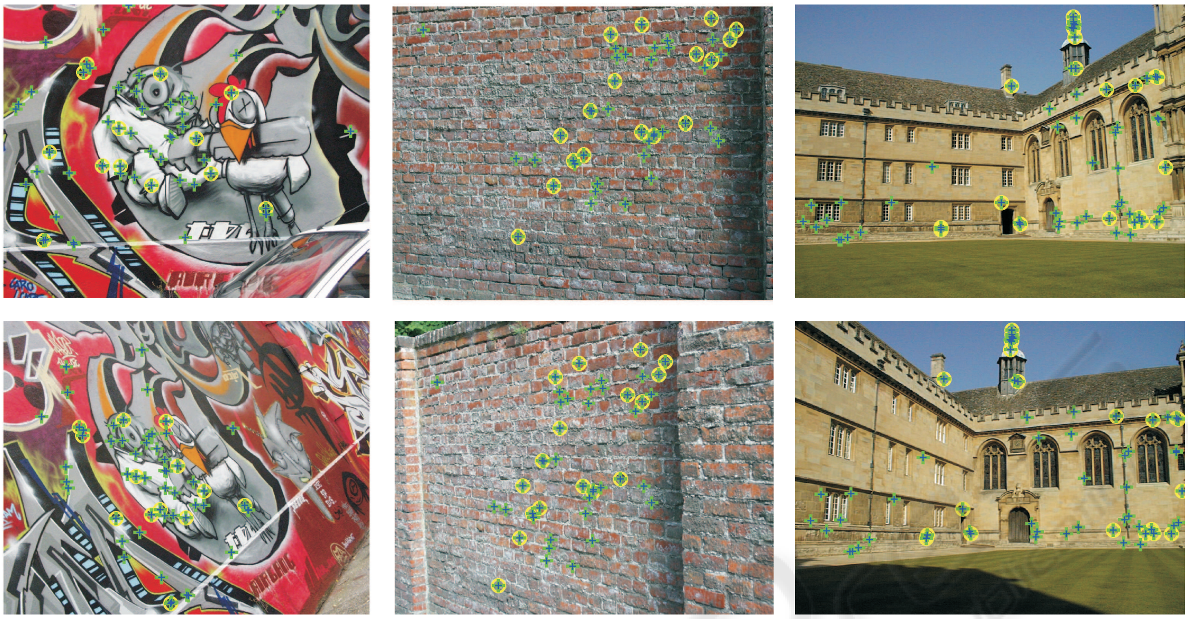

Figure 2: Comparative results. From left to right: *Graffiti* - SIFT found only 13 valid matches (yellow circle) while our approach found 52 correct matches (green crosses), *Wall* - SIFT found only 17 valid matches (yellow circle) while our approach found 44 correct matches (green crosses), *Wadham College* - SIFT found only 20 valid matches (yellow circle) while our approach found 60 correct matches (green crosses).

ture points using DoG, that is also based on the automatic scale selection principle, every keypoint has attached a characteristic scale value. Therefore, the initial value of the scale between images is computed as an average of the ratio between characteristic scales of all the matched keypoints.

The initial value of the rotation angle parameter α_0, we rely on the distribution of the image gradient orientation and magnitude. An orientation histogram is built in the surroundings of each feature points, with every bin counting the contribution of the point gradient orientation weighted by its gradient magnitude and by a gaussian-weighted circular window of the respective region. The dominant orientation is determined by the highest peak of the histogram (Lowe, 2004). We estimate the initial state of the rotation parameter α_0 by averaging the ratio of the dominant orientation of the correspondent feature points.

Practically, in our approach the estimated parameters determine the new shape of the region where the descriptors will be computed. The same procedure of computing the descriptor as presented in the previous section is repeated but employing the new shape of the neighbor regions around feature points that is determined based on the estimated geometry.

4 EXPERIMENTAL RESULTS

We tested our method for real images with the related geometry being precomputed. Our approach is compared with the original SIFT using the well known INRIA[1] database but also several other images taken under large variation of the camera viewpoint angle. To evaluate our method we assume that we know the geometry that relates the tested pairs of images. The evaluation of the results is done using repeatability criteria for the planar scenes. This criterion introduced by Schmid et al. (Schmid et al., 2000) takes into account locations as well as detected scales of the points. The score of repeatability for a pair of images represents the ratio between the number of point-to-point matches and the minimum number of points detected in images.

In the left side of the Figure 2 is presented the matching results for *graffiti* images. In this case the difference of the camera viewpoint angle is approximately 50°. Note that our method is able to find 52 valid matches while SIFT is able to find only 13 correct matches. By only adding the color we obtained 19 valid matches (not shown). In the middle of the Figure 2 are shown two images of the *wall* INRIA data set when the difference of the angle is 60°. Again our method outperforms SIFT finding 44 valid

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

matches against only 17 found by SIFT.

In the right side of the Figure 2 are presented two images of the Wadham College 3D model of the Oxford data set. In this case to validate the matching results the criteria based on the homography is not valid anymore. Instead, we make use of the fundamental matrix that characterizes the geometry between the views. As can be observed again our method outperforms SIFT finding a considerable additional number of correct matches.

5 SUMMARY AND CONCLUSIONS

This paper introduces an alternative strategy for wide-baseline image matching. The method is built on the widely-used SIFT operator. We first show how the distinctness of the SIFT descriptor vectors can be increased by adding the color information. Then, by estimating the geometry that relates patches of corresponding feature points found in the previous stage we are able to define a new shape of the regions where descriptors are computed more accurately. Our framework demonstrates to improve considerable the matching results compared with the results obtained by the original SIFT operator. For future work we would like to take into consideration more important photometric variations between images but also to demonstrates the utility of the method for several practical computer vision applications.

REFERENCES

Abdel-Hakim, A. E. and Farag, A. A. (2006). CSIFT: A sift descriptor with color invariant characteristics. *IEEE CVPR*.

Baumberg, A. (2000). Reliable feature matching across widely separated views. *IEEE Conf. on Computer Vision and Pattern Recog., CVPR*.

Bay, H., Tuytelaars, T., and Gool, L. V. (2006). SURF: Speeded up robust features. *in Proceedings of European Conference on Computer Vision*, pages 404–417.

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Tran. on Patt. Anal. and Mach. Intell.*

C.G.Harris and Stephens, M. (1988). A combined corner and edge detector. *in Proceedings of Fourth Alvey Vision Conference*, 18:147–151.

Chang, P. and Krumm, J. (1999). Object recognition with color cooccurrence histograms. *IEEE Conf. on Computer Vision and Pattern Recog., CVPR*.

Forssén, P.-E. and Lowe, D. (2007). Shape descriptors for maximally stable extremal regions. In *IEEE International Conference on Computer Vision*.

Freeman, W. and Adelson, E. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:891–906.

Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*.

Ke, Y. and Sukthankar, R. (2004). PCA-SIFT: A more distinctive representation for local image descriptors. *IEEE CVPR*.

Lindeberg, T. (1999). Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):77–116.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 20, pages 91–110.

Matas, J., Chum, O., Martin, U., and Pajdla, T. (2002). Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*.

Mikolajczyk, K. and Schmid, C. (2004a). A performance evaluation of local descriptors. *IEEE Conf. on Comp. Vision and Pattern Recog.*

Mikolajczyk, K. and Schmid, C. (2004b). Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1).

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., and Gool, L. V. (2005). A comparison of affine region detectors. *Int. J. Comp. Vision,*.

Moreels, P. and Perona, P. (2007). Evaluation of features detectors and descriptors based on 3d objects. *Int. J. Comput. Vision*, 73(3):263–284.

Pritchett, P. and Zisserman, A. (1998). Wide baseline stereo matching. In *IEEE ICCV*.

Schmid, C., Mohr, R., and Bauckhage, C. (2000). Evaluation of interest point detectors. *International Journal of Computer Vision*, pages 151–172.

Tuytelaars, T. and Gool, L. V. (2000). Wide baseline stereo matching based on local, affinely invariant regions. *In Proceedings of British Machine Vision Conference*.

Tuytelaars, T. and Gool, L. V. (2004). Matching widely separated views based on affine invariant regions. *Int. J. Comput. Vision*, 59(1).

Xiao, J. and Shah, M. (2003). Two-frame wide baseline matching. *IEEE Int. Conf. on Comp. Vision*.