A HOMOGRAPHY-BASED CDVS PIPELINE FOR IMAGE MATCHING WITH IMPROVED RESILIENCE TO VIEWPOINT CHANGES

Biao Zhao, Enrico Magli

Dept. of Electronics and Telecommunications, Politecnico di Torino (Italy)

ABSTRACT

Compact Descriptors for Visual Search (CDVS) is MPEG proposed standard that will enable efficient and interoperable design of visual search applications using local descriptors. Such descriptors are invariant to rotation and scaling, but are not very robust towards viewpoint changes. In this paper, we address this problem and propose a modified version of the CDVS pipeline that employs image back-projection to compensate for perspective distortion. The proposed technique is based on the homography derived from the correspondence extracted from pairs of matching keypoints. Extensive results show that it improves the CDVS matching accuracy under viewpoint changes while having low complexity.

Index Terms— CDVS, Content based image retrieval, Homography, SIFT descriptors

1. INTRODUCTION

Image matching refers to declaring two different images as similar or different solely based on their content. It is typically performed employing SIFT descriptors [1], wherein two important stages are identified, namely keypoint detection and feature matching. Keypoint detection employs a scale space to identify relevant points of interest, *e.g.*, corners, in a way that is invariant to the scale factor of the image. Every keypoint is then represented by its coordinates, scale, as well as a feature vector summarizing the information in a small patch centered around the descriptor. The two sets of descriptors for each pair of images are then matched, in order to identify a set of keypoints that are deemed to be matched corresponding points in either image. To witness the importance of image matching, MPEG is standardizing a pipeline for image retrieval using compressed SIFT descriptors, called CDVS (Compact Descriptors for Visual Search) [2].

Remarkably, SIFT descriptors exhibit good invariance to rotations, occlusions and small illumination changes. However, they do not exhibit any built-in resilience to viewpoint changes. An image taken at different viewpoint from another image will have a perspective distortion, which is going to negatively affect matching results. If the viewpoint change is too severe, SIFT descriptor will fail at correctly matching the two images. This is a very important problem, since in the real world pairs of pictures are almost invariably taken from different viewpoints.

Concerning resilience to perspective transformations, several techniques have been developed. Hessian-Affine [3] and Harris affine techniques [4] achieve robustness to the transformation via an iterative shape adaptation algorithm to compute the local affine transformation for each interest point. Maximally stable extremal regions are based on extracting a comprehensive number of corresponding image elements contributing to improve affine invariance [5]. Salient detector identifies ellipsoidal regions, which is a better approximation of viewpoint change [6]. ASIFT simulates all image views obtainable by varying the camera axis to diminish the perspective effects [7]. However, ASIFT has relatively higher complexity than conventional SIFT descriptors.

In this paper we also address the problem of viewpoint-invariant image matching. There are two aspects that differentiate this work significantly from previous papers. First, we aim at developing a solution that is compliant with CDVS. Therefore, any modifications must not involve the keypoint detector and descriptor, which are specified by the standard. Second, we apply a different transform, *i.e.*, the homography transform, to compensate for viewpoint changes. To our best knowledge, this is the first time that this transform is employed to improve robustness of local descriptors to viewpoint changes. In summary, the proposed technique can be seen as an add-on to the CDVS retrieval pipeline, where the standard pipeline is run first, and our proposed technique performs a post-processing a re-ranking stage. Extensive tests on the CDVS database show that the proposed technique can improve the matching precision up to 3%.

2. HOMOGRAPHY MODEL

In a content based image retrieval system, two images of the same scene are to be matched. If the viewpoints between these images are different, *i.e.*, there is perspective distortion between them, then a correct matching might not be possible. To reduce the perspective distortion, we propose to estimate the homography between the two images [8]. With projective cameras, any two images of the same planar surface in space are related by a homography [9]. Homography can be used to estimate the projective position and projective plane.

Once the homography is known, back-projection can be used to reduce the perspective distortion and improve the image matching accuracy, as we show in Sec. 3.

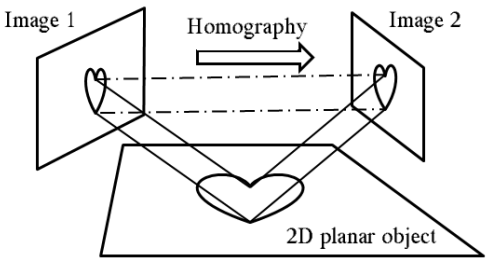


Fig. 1. Homography model.

Figure 1 depicts a homography model. Image 1 and Image 2 are images of a 2D planar object. One point on Image 1 correspond to another point on Image 2 when they both reflect the same point on the object. Images of 2D planar objects are obtained via projective reflection. Thus images with different viewpoints are from a different reflection. These reflections are projectively related in geometry. This relationship can be estimated after knowing corresponding pairs of points because it is a homogenous relationship among all the points on the planes.

The mathematical definition of a homography is given below:

$$P_a = \begin{bmatrix} x_a \\ y_a \\ 1 \end{bmatrix}, P_b = \begin{bmatrix} w^{'} x_b \\ w^{'} y_b \\ w^{'} \end{bmatrix} \tag{1}$$

Then: $P_b = \mathbf{H}_{ab} P_a$ where $\mathbf{H}_{ba} = \mathbf{H}_{ab}^{-1}$. P_a and P_b are the corresponding points on different 2D planes. Notice that points laying on $\mathbb{R}^2$ are normally represented as a pair $(x, y)^T$. However in projective geometry intersection points of lines or planes are more relevant. For a homogenous representation, a third coordinate is added as a scale variable [9]. Therefore, an arbitrary homogeneous vector representative of a point is of the form $x = (x_1, x_2, x_3)^T$, representing the point $(x_1/x_3, x_2/x_3)^T$ in $\mathbb{R}^2$. The points at infinity can be represented with $x_3 = 0$. H_{ab} is the homography matrix, representing the projection of point P_a to P_b. H_{ba} is the corresponding inverse transformation.

To qualify the perspective distortion, we need to estimate the homography matrix. Direct Linear Transformation (DLT) is one of algorithms to determine H_{ab}, given a set of 2D to 2D point correspondences $x_i \leftrightarrow x_i^{'}$ [9]. x_i and $x_i^{'}$ are the corresponding points on different planes.

It can be shown [9] that it is necessary to specify four pairs of point correspondences in order to constrain H_{ab} fully. If exactly four pairs are given, then a unique solution for the matrix H_{ab} exists. However, since matching pairs are not known exactly, because of the nonideality of the keypoint detector, if more than four correspondences are given then these correspondences may not be fully compatible with any projective transformation, and one will be faced with the task of determining the best transformation given the data. Generally, this can be done by finding the homography matrix that minimizes a cost function or ruling out the outliers with the help of RANSAC [9].

Letting pairs of correspondences be related by $x_i^{'} = H_{ab} x_i$,

$$H_{ab} x_i = \begin{pmatrix} h_1^T x_i \\ h_2^T x_i \\ h_3^T \end{pmatrix}, \text{with } H_{ab} = \begin{pmatrix} h_1^T \\ h_2^T \\ h_3^T \end{pmatrix}, \tag{2}$$

the DLT algorithm [9] finds the homography as the solution of

$$\begin{bmatrix} 0^T & -w_i^{'} x_i^T & y_i^{'} x_i^T \\ w_i^{'} x_i^T & 0^T & -x_i^{'} x_i^T \end{bmatrix} \begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix} = 0. \tag{3}$$

Eq. 3 has the form $Ah = 0$. Once we have four pairs of point correspondences, we obtain a set of equations, where A is the matrix of coefficients built from the matrix rows A_i from each correspondence, and h is the vector we seek to construct estimated homography matrix H_{ab}.

If more than four point correspondences are given, then the set of equations $Ah = 0$ is over-determined. If the position of the keypoints are exact, there will not be an exact solutions to the over-determined system $Ah = 0$ apart from the zero solution. However, we cannot be sure that all the available correspondence pairs are reliable, so we must identify and remove outliers before estimate the homography.

To this end, we employ RANSAC, which is an iterative method to estimate parameters of a mathematical model from a set of observed data which contains outliers [10]. The idea is very simple: two of the points are selected randomly; these points define a line. The confidence score for this line is calculated as the number of points that lie within a maximum distance. This random selection is repeated a number of times and the line with highest confidence score is deemed the robust fit. The points within the threshold distance are the inliers.

The aim of this stage is two-fold: first, to obtain an improved estimate of the homography by using all the inliers available in the given correspondence pairs (rather than only the four points of the sample); second, during the following back-projection stage, to obtain more matches from the correspondence set because a more accurate homography is available. An improved estimate of the homography is then computed from the inliers.

3. PROPOSED HOMOGRAPHY-BASED RETRIEVAL STAGE

CDVS is the standard under development in MPEG that will provide a highly efficient and interoperable pipeline for visual search; Figure 2 display the local descriptor extraction of CDVS [2]. It includes keypoint detection, feature selection, local descriptor computation, local descriptor compression and coordinate coding. Keypoint detection and descriptor computation are the fundamental operations of visual search. The purpose of feature selection is to preserve the most significant keypoints for a low memory consumption. Local descriptor compression and coordinate coding both aim to decrease the memory consumption and transmission bandwidth.

Fig. 2. CDVS local descriptor extraction.

In the CDVS standard, whether two images will be declared as matched or not depends on their matching score. Each pair of matched features will be assigned a score and the total image matching score is obtained by summing up all the scores of the matched features on that image. However, since CDVS is not viewpoint robust by construction, it may wrongly declare matching or non-matching images because of perspective distortion. In this paper we argue that perspective distortion can be reduced by homography estimation and back-projection. Back-projection consists in inverting the perspective transformation. The first step towards back-projection is to estimate the homography that defines the inverse transformation.

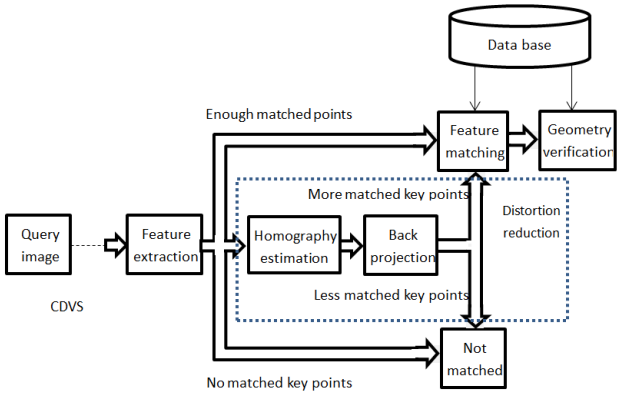


Fig. 3. Integrated back-projection CDVS.

As has been said, a homography can be derived from at least 4 pairs of corresponding points. However, the image matching process will typically provide more than 4 matching pairs. In our proposed system, we used the DLT algorithm and RANSAC, as detailed in Sec. 2, to estimate the homography. Then, the perspective distortion can be reduced applying back-projection.

In particular, the proposed estimation and back-projection process operates as follows. Suppose two images I_a and I_b have an approximate homography relationship. The standard CDVS pipeline might declare a non-match between I_a and I_b because of the perspective transformation. Once the homography H_{ab} is estimated by DLT, the image $I_{a'}$ is obtained as $I_a^{'} = H_{ab}I_a$. In other words, we now have a new pair of images, $I_a^{'}$ and I_b, where the perspective distortion has been removed or at least strongly attenuated. It is therefore reasonable to assume that, while CDVS might wrongly declare I_a and I_b as a non-match, it can be likely to correctly declare $I_a^{'}$ and I_b as a match. Thus $I_a^{'}$ and I_b are set as the new pair to be checked as matching or non-matching by CDVS.

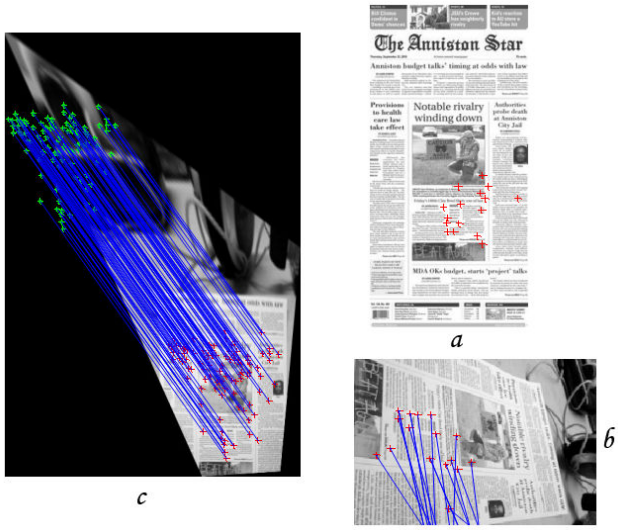


Fig. 4. Improvement on the numbers of features.

More in detail, the pipeline of our proposed method is displayed in Figure 3. From the pipeline, we can see that our proposed stage is integrated into the standardized CDVS visual search system. This guarantees to exploit CDVS's high efficiency and accuracy. The area inside the dotted rectangle is our proposed stage. It includes homography estimation, back-projection and re-matching of an image pair after compensating for perspective distortion.

In particular, the re-matching process is triggered only if the matching score does not exceed the threshold. That is, if CDVS believes the images are matched, we trust this as it is likely that the images had small perspective distortion. Instead, if CDVS decided that the image pair does not match, we perform back-projection and re-matching to see if a transformation can be found, which will estimate and correct the perspective distortion leading to a positive match. In particular,

the re-matching stage checks whether more than 4 matched corresponding pairs are available. If this is the case, a homography matrix is estimated and back-projection is used to remove the perspective distortion between the images. Thus, the image pair without perspective distortion will have more matched features. However, it is not guaranteed that matched features are truly matched or the positions of the matched features are exact. The back-projection based on non accurately estimated homography cannot help to decrease the perspective distortion. In the pipeline, to make sure that the perspective distortion of back-projected image is not worse than the initial one, the score after the re-matching stage is compared with the initial score. If the score has not been improved, the initial matching score and the related matching decision will be preserved.

To understand the re-matching process, note that, as Figure 4 displays, after reducing the perspective distortion, there will normally be an increase of the number of matched features. If the increased score exceeds the threshold after the back-projection, then the previous non-match will be turned into a correct match. In Figure 4, *a* and *b* are the initial images. Due to the perspective distortion, the number of the matched features is around 15. *c* is the image after reducing the perspective distortion from *b*. The matched features between *a* and *c* are around 125. It is a great increase of the numbers of matched features, which can lead to a correct match, while the two initial images would not have been matched.

4. EXPERIMENTS AND RESULTS

Our proposal has been integrated into the CDVS test model. Experiments are conducted employing the MPEG dataset used in the evaluation of CDVS. In the dataset, there are 5 image categories. Additionally, Category 1 has 3 sub-categories. These dataset are defined as follows.

1.a Mixed text and graphics

1.b Mixed text and graphics at VGA resolution

1.c Mixed text and graphics at VGA resolution with heavy JPEG compression

2 Paintings

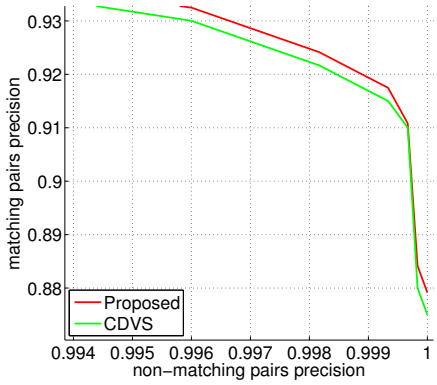
3 Video frames

4 Buildings and landmarks

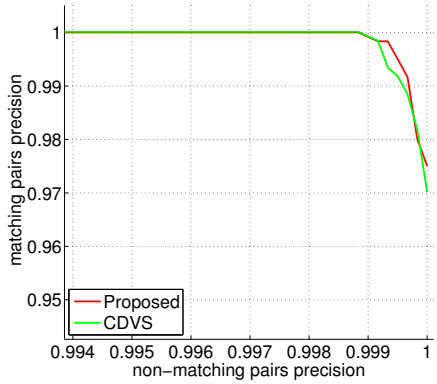
5 Common objects

Totally, there are 33590 images in the dataset.

The experiment for evaluating the performance of the proposed scheme is designed as follows. In each category, original CDVS and our integrated back-projection CDVS are tested calculating both matching precision and non-matching precision. Our experiment has been run on all categories. As expected, the proposed back-projection method is more efficient in the categories of objects where the planar assumption is reasonable, although no performance decrease is observed in the other categories, leading to an overall improvement.



(a)



(b)

Fig. 5. Proposed method precision evaluation, (a)prints; (b)dvds.

Figure 5 shows some result on the matching precision, in particular (a) displays the result in *print* and (b) displays the result in *dvds*. The red line represents the precision of the proposed back-projection method and the green one represents the precision of original CDVS, and it can be seen that the proposed algorithm consistently outperforms CDVS. The result are further analyzed in Table 1, including planar objects at original resolution, at VGA resolution and at VGA resolution with heavy JPEG compression, buildings, landmarks and video frames. In the table, MP is short for matching pairs precision and NMP is short for non-matching pairs precision. Generally speaking, our proposed back-projection method can improve the matching precision. However, the improvement varies among categories and resolutions. As expected, the improvement on the 2D planar objects is more obvious compared with buildings and landmarks. But even on buildings and landmarks, our method can still improve the matching precision. The average improvement on the buildings and landmarks is about 0.3% and the average improvement on the 2D planar objects is 2.9%. JPEG compression will not affect the improvement but resolution indeed has an

Books	**MP**	**NMP**	**CDs**	**MP**	**NMP**	**Dvds**	**MP**	**NMP**
CDVS	0.980	1.000	CDVS	0.900	1.000	CDVS	0.970	1.000
Proposed	0.990	1.000	Proposed	0.920	1.000	Proposed	0.975	1.000
Improvement	0.010	0.000	Improvement	0.020	0.000	Improvement	0.005	0.000
Books-vga			**CDs-vga**			**Dvds-vga**		
CDVS	0.980	1.000	CDVS	0.921	1.000	CDVS	0.921	1.000
Proposed	0.985	1.000	Proposed	0.938	1.000	Proposed	0.938	1.000
Improvement	0.005	0.000	Improvement	0.017	0.000	Improvement	0.017	0.000
Books-vga-jpeg			**CDs-vga-jpeg**			**Dvds-vga-jpeg**		
CDVS	0.983	1.000	CDVS	0.901	1.000	CDVS	0.976	1.000
Proposed	0.987	1.000	Proposed	0.918	1.000	Proposed	0.985	1.000
Improvement	0.004	0.000	Improvement	0.017	0.000	Improvement	0.009	0.000
Cards			**Print**			**Video**		
CDVS	0.960	0.997	CDVS	0.872	1.000	CDVS	0.858	0.999
Proposed	0.965	0.997	Proposed	0.880	1.000	Proposed	0.838	0.999
Improvement	0.005	0.000	Improvement	0.008	0.000	Improvement	0.020	0.000
Cards-vga			**Print-vga**			**Buildings-Stanford**		
CDVS	0.935	0.997	CDVS	0.848	1.000	CDVS	0.555	1.000
Proposed	0.952	0.997	Proposed	0.882	1.000	Proposed	0.561	1.000
Improvement	0.017	0.000	Improvement	0.034	0.000	Improvement	0.001	0.000

Table 1. Performance of the proposed technique on the image categories of the CDVS dataset.

effect. Typically, a image of higher resolution will generate more matched pairs of keypoints, but these matches are not generally more correct than in a lower resolution image. More incorrect matched keypoints do not contribute to a correct homography estimation, hence a higher image resolution generally did not provide better results.

4.1. Conclusion

This paper proposes a new method based on the CDVS pipeline, attempting to improve the matching precision of images pairs taken at different viewpoints, which is known to be a difficult case for SIFT descriptors. The method employs homographies, and is fully integrated into the CDVS standard, its complexity is low and it can improve the matching precision, especially on images of 2D planar objects. In particular, performance improvement is up to 3% on those image categories that satisfy the planar model, such as print and CDs.

REFERENCES

[1] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] "Compact descriptors for visual search," October 2013 ISO: IEC CD 15938-13.

[3] Krystian Mikolajczyk and Cordelia Schmid, "An affine invariant interest point detector," in *Computer Vision-ECCV 2002*, pp. 128–142. Springer, 2002.

[4] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, and Luc Van Gool, "A comparison of affine region detectors," *International journal of computer vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[5] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.

[6] Adam Baumberg, "Reliable feature matching across widely separated views," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*. IEEE, 2000, vol. 1, pp. 774–781.

[7] Jean-Michel Morel and Guoshen Yu, "Asift: A new framework for fully affine invariant image comparison," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 438–469, 2009.

[8] Paresh Kumar Jain and CV Jawahar, "Homography estimation from planar contours," in *3D Data Processing, Visualization, and Transmission, Third International Symposium on*. IEEE, 2006, pp. 877–884.

[9] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2003.

[10] Martin A. Fischler and Robert C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.