EMPLOYING 3D ACCELEROMETER INFORMATION FOR FAST AND RELIABLE IMAGE FEATURES MATCHING ON MOBILE DEVICES

Ayman Kaheel, Motaz El-Saban, Mostafa Izz and Mahmoud Refaat
Microsoft Advanced Technology Labs in Cairo, Microsoft Research

ABSTRACT

Image matching is a cornerstone technology in many image understanding, augmented reality and recognition applications. The state-of-the-art techniques follow a feature-based approach by extracting interest points and describing them by either rotation or affine invariant descriptors. However, requiring rotation or affine invariance comes at an additional computational cost as well as inaccurate estimates in some cases such as out-of-plane rotations. Fortunately, today most mobile devices incorporate 3-D accelerometers that measure the acceleration values along the three axes. In this paper, we propose to employ the acceleration values to calculate the in-plane and tilting rotation angles of the capturing device, in order to alleviate the need for constructing rotationally invariant descriptors. We describe an approach for incorporating the calculated rotation angles in the process of interest point extraction and description. Furthermore, we evaluate empirically the proposed approach, both in terms of computational time and accuracy on standard datasets as well as a dataset collected using a mobile phone. Our results show that the proposed approach provides savings in computational time while providing accuracy gains.

***Index Terms*—** feature matching, accelerometer, sensors, mobile devices

1. INTRODUCTION

There is a steady increase in recent years in augmented reality applications on mobile devices. Many of these applications leverage computer vision techniques for image understanding and recognition. A key technology for enabling these techniques is image matching. Recently, feature-based image matching techniques have dominated the literature due to their ability to handle occlusions as well as various geometric and photometric image transforms when properly designed. These features are used to establish correspondence between set of images (could be two in the simplest case) and hence enable deriving geometric relations between them. A key to successful correspondence is the availability of invariant feature representations under a class of photometric and geometric transforms. One of the most popular feature representations that has shown good performance is SIFT [3]. SIFT provides illumination changes immunity by working in the gradient domain. In-plane rotation invariance for each interest point is achieved through local estimation of the gradient orientation based on the surrounding patch of pixels; hence, the descriptor can be represented relative to this orientation. Unfortunately, requiring a high level of geometric invariance always comes at a high computational cost. In many cases, the computational cost is prohibitively high so much so that the designer is forced to accept a lower level of invariance.

In this work, we argue that by leveraging unique capabilities of mobile devices, especially that of providing acceleration values along the three major axis, computational savings in local orientation estimation can be harnessed. Clearly if the two images being matched are brought to the same orientation, by using accelerometer readings, then this computational step could be skipped altogether. In this case an upright version of the descriptor would suffice. The key technical contributions of this paper are:

1. Fusing feature-based descriptors with the rotation angles calculated from the device’s accelerometer to alleviate the need for constructing rotationally invariant descriptors.
2. Describing an approach to perform this fusion effectively.
3. Comparing the proposed approach to SIFT and its speeded up version (SURF)[2]. We experimentally validate that the proposed approach can achieve better matching capabilities at lower computational cost on a collected dataset as well as on a standard dataset.

The rest of this paper is organized as follows. Section 2 contains a brief discussion and survey of related work. We describe a method for calculating the rotational angles from the accelerometer information in Section 3. Section 4 describes the proposed approach for fusing feature-based descriptors with the rotation angles. A description of the experimental setup is given in Section 5. The proposed approach is evaluated in Section 6. Finally, we conclude in Section 7.

2. RELATED WORK

Various forms of features have been used in the literature including points, edges, regions and contours [4],[5],[6]. In this work we concentrate on point features, as they are the most commonly used owing to their general nature. The main stages in a typical feature-based image matching pipeline are feature detection, feature description and feature correspondence [1]. In the feature detection stage, each image is searched for local features, often called interest points, with the desirable properties to be invariant under a class of image transforms as well as being distinctive. The feature description stage involves describing each interest point in terms of the surrounding patch of pixels, either using a single value or a distribution involving raw, moments or gradient components [7]-[12]. Then, the interest point descriptor is represented as a feature vector and feature correspondence is established using a distance metric on that vector.

Most modern day interest-point detectors are able to deal with in-plane image rotation. The state-of-the-art method to achieve rotational invariance is to estimate a dominant orientation at each detected interest point. Once the local orientation of an interest point has been estimated, an oriented patch around the detected point can be extracted and used to form a feature descriptor. The simplest possible orientation estimate is the average gradient within a region around the interest point. SIFT uses a better technique, it looks at the histogram of orientations computed around the interest point. SURF, on the other hand, uses the responses to Haar wavelets for orientation assignment.

A number of fully affine invariant detectors and descriptors have been proposed in the literature [11]-[17]; two detectors are considered to be the state of the art, Maximally Stable Extremal Region (MSER) [12] and Affine SIFT (ASIFT) [17]. MSER works by thresholding the image at all possible gray levels. Regions whose rate of change of area w.r.t. threshold is minimal are defined as maximally stable and are returned as detected regions. This results in regions that are invariant to affine geometric transformations. ASIFT follows a different approach; it simulates all image views obtainable by varying the two camera axis orientation parameters, namely, the latitude and the longitude angles. Then it deals with scale, translation and in-plane rotation by using the SIFT method itself. It is worth noting that the full affine invariant descriptors are considerably more computational expensive than the rotational invariant descriptors.

A related problem space is the simultaneous localization and mapping (SLAM) [18]. The SLAM problem asks if it is possible for a mobile robot to be placed at an unknown location in an unknown environment and for the robot to incrementally build a consistent map of this environment while simultaneously determining its location within this map. One class of approaches attempts to solve this problem using mainly visual information (VSLAM)[19]-[21]. In VSLAM, the mobile robots are equipped with a 3-cameras stereo vision system and uses feature descriptors such as SIFT for extracting landmark candidates. The candidates then are matched between cameras to determine the 3D position. The interesting aspect in this problem is that the robots are in many cases equipped with accelerometers for determining the speed of the robot. Therefore, we believe that our approach would be beneficial in that problem space.

3. CALCULATING ROTATIONAL ANGLES FROM 3D ACCELEROMETER INFORMATION

A 3-D accelerometer is an electromechanical device that can measure the 3D acceleration forces *ax*, *ay*, *az* along the *x*, *y*, and *z* axes. For convenience, the three axes are chosen aligned with the capturing device axes as depicted in Figure 1 (in this case a mobile phone). When the device is steady, these values correspond to components of gravitational acceleration along the different axes. Each value theoretically ranges from 0 to 9.80665 m2/sec. The remaining of the discussion of this section will use mobile phone as an example for the capturing device.

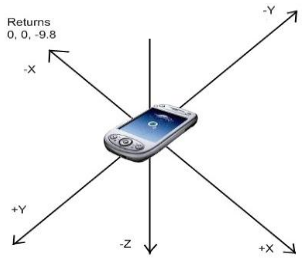


Figure 1 Mobile phone axes

The three acceleration values can be embedded in the image header, according to any information embedding standard, similar to the Information Interchange Model (IIM)[22]. Alternatively, this information can be stored in a separate metadata file. Such as metadata information was used in Seon et al. work [24] for instance to guide search within video archives.

The accelerometer returns the *ax*, *ay*, *az* values along the axes that are shown in Figure 2 and Figure 3. In case the device is held upright, *ay* would have a value of 9.80665 with *ax* and *az* equal to zero. In other positions, the gravitational force would have components in the *x*, *z* directions as well. In order to calculate the rotation angle in the (*x*,*y*) plane and the rotation angle in the (*z*,*y*) plane, denoted by α in Figure 2 and β in Figure 3 respectively, we need to calculate the 3D rotational transformation between the vector representing the gravitational acceleration (0, |*Y*|, 0) and the vector generated by the accelerometer (*ax*, *ay*, *az*). The value of |*Y*| is theoretically 9.80665, however because of the imperfections of the 3D accelerometer the value need to be calculated as $|Y| = \sqrt{ax^2 + ay^2 + az^2}$.

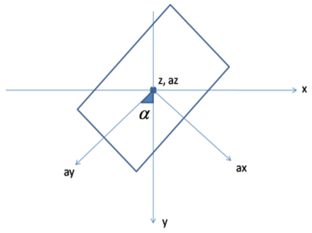


Figure 2 Device coordinates front view

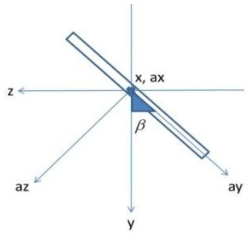


Figure 3 Device coordinates side view

The 3D transformation relating the mobile device coordinates and the world coordinates is described using the equation (1).

$$\begin{pmatrix} ax \\ ay \\ az \end{pmatrix} = \begin{pmatrix} \cos\alpha & \sin\alpha\cos\beta & -\sin\alpha\sin\beta \\ -\sin\alpha & \cos\alpha\cos\beta & -\cos\alpha\sin\beta \\ 0 & \sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} 0 \\ |Y| \\ 0 \end{pmatrix} \quad (1)$$

4. FUSION OF FEATURE-BASED DESCRIPTORS WITH ACCELEROMETER READINGS

As mentioned earlier, descriptors such as SIFT and SURF are in-plane rotation invariant, i.e. rotation in the vertical (*x*,*y*) plane. This invariance is achieved by estimating a reproducible orientation for each interest point, and the interest point descriptor is always represented with respect to this orientation. The exact method of the point orientation estimation depends of the particular descriptor used.

In this paper we propose to depend on the angles calculated from the accelerometer readings, instead of estimating the orientation at each interest point, to provide in-plane and tilt rotational invariance. The proposed approach proceeds as follows; perform the feature detection step, then correct globally for both the in-plane and tilt rotational angles, and thereafter apply the feature description step[1]. In essence, this approach can be used to augment any state-of-the-art descriptor, however, in this paper we concentrate on SIFT and SURF. It is important to note here that the proposed approach is designed to compensate for camera rotation; however, it would not provide the desired invariance in cases where the image contents are rotated.

Correcting for the in-plane rotation angle α calculated from Equation 1 would alleviate the need for estimating the orientation values associated with different interest points in both SIFT and SURF. We correct for the angle α by using the rotational transformation matrix in Equation 2.

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (2)$$

With respect to the second angle, let's recall that the SIFT and SURF descriptors are not designed to deal with affine deformations, instead, they are assumed to be second-order effects that are covered to some degree by the overall robustness of the descriptor. In addition, Lowe [3] has claimed that the additional complexity of full affine-invariant features often negatively impact their robustness and does not pay off, except in cases of large viewpoint differences.

In our approach, we compensate for the tilt rotation angle (β) before applying the SIFT and SURF descriptors. This will enable the descriptors to respond consistently across affine deformations, such as (local) perspective foreshortening, without any additional computational cost. We will show empirically that this method indeed improves the results of the descriptors.

5. EXPERIMENTAL SETUP

We conducted a number of experiments aiming at testing the hypothesis that compensating for in-plane and tilt angles computed from accelerometer readings can improve matching accuracy as well as processing times of interest point descriptors. We evaluate the proposed approach when applied to two state-of-the art interest point descriptors namely SIFT and SURF.

5.1 Datasets

Since datasets previously used in the literature for comparing interest point descriptors do not have accelerometer readings associated, we resorted to collecting our own dataset using mobile phones while adhering to the

[1] Investigating the effect of global rotation correction on the feature detection step has been left for future work.

methodology of data collection followed in [1] such as a balanced collection between textured and structured scenes and a variation in both in-plane and out-of plane (tilt) rotations. We call this dataset “*real dataset*”, for which some statistics are shown in Table 1 and sample images are illustrated in Figure 4.

Table 1. Real dataset statistics

Scene name	**Type**	**Number of images**	**Number of pairs**
Backyard Scene_1	structured	24	276
Backyard Scene_2	structured	24	276
Food Area	structured	17	136
Hallway	textured	10	45
Monitor	textured	7	21
Newspaper	textured	9	36
Small Trees	textured	13	78
X-Box	structured	19	171

In order to remove any selection bias from our end in the data collection process, we have also used a set of standard images from the dataset used in [1], then we’ve subjected these images to artificial geometric transformations and measured the matching performance. We have selected a total of 14 images from the images in [23]. For each selected image we have generated 30 images by introducing random in-plane rotations from 0 to 360 degree and tilt rotation (around x-axis) from -45 to 45 degrees. Table 2 shows the statistics for this “*synthetic dataset*”, while Figure 5 shows sample images from this dataset. Results reported in the remaining of this section are average over all the images corresponding to a scene.

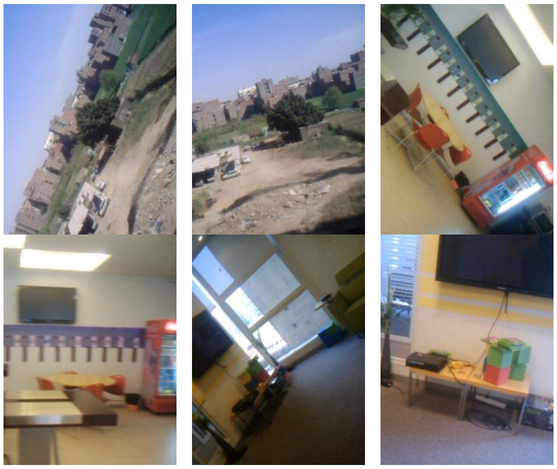


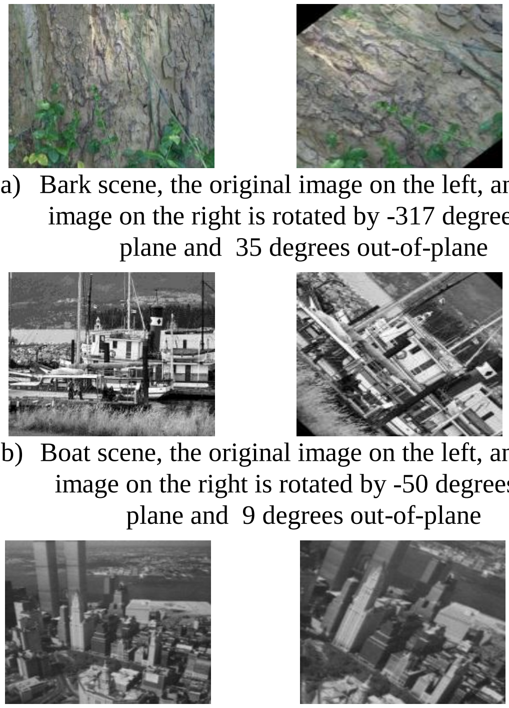
Figure 4: samples from the real dataset including structured and textured scenes

5.2 Evaluation criteria

We have adopted three main measures for testing the matching capability of the proposed fusion of interest point-based scheme and accelerometer readings. These measures are the standard precision and recall, and computation time. It is worth mentioning that using accelerometer readings for compensating for rotation angle of the capturing device does not directly change the way interest points are detected or described in SIFT or SURF. However, by compensating for in-plane and tilt rotation angles *before* interest point description, we are able to show that one can expect a boost in performance as the job of detection and description becomes easier. In the following, we use SIFT+ to signify the case when SIFT is augmented with the rotational angles calculated from the accelerometer readings, and likewise, we use SURF+ to indicate the usage of the augmented SURF.

Precision and Recall

Recall and (1-precision) values are used to measure the quality of image descriptors as previously suggested in descriptors comparisons [1]. These measures are based on computing the total number of correspondences (ground truth) and the ones correctly and incorrectly computed.



(a) Bark scene, the original image on the left, and the image on the right is rotated by -317 degrees in-plane and 35 degrees out-of-plane

(b) Boat scene, the original image on the left, and the image on the right is rotated by -50 degrees in-plane and 9 degrees out-of-plane

(c) NewYork scene, the original image on the left, and the image on the right is rotated by -23 degrees in-plane and -42 degrees out-of-plane

Figure 5 Samples from the synthetically generated dataset where various amounts of in-plane and out-of-plane rotation are introduced

Computation Time

Another important aspect in our comparisons is the computational time savings. The time savings come from the fact that we do not have to estimate the orientation of interest points patches as this is compensated for globally using accelerometer readings. Furthermore, employing the tilt angle from the accelerometer alleviate the need for the complex calculations associated with fully affine transforms.

6. RESULTS AND ANALYSIS

6.1 Computation time

The first comparison on real and synthetically generated datasets concerns the computation time and the savings achieved when utilizing accelerometer readings for compensating for rotational angles. This eliminates the need for any expensive calculation of rotation angles based on image content and replaces it with few inexpensive calculations. The running times of SIFT, SURF, SIFT+, and SURF+ are shown with in figure 6 for the synthetic and read datasets respectively using a 3GHz Intel® Core™2 Duo CPU with 4.00 GB of RAM and running a 32-bit operating system (Windows 7 Enterprise). Results show savings in computation time in most of the cases. The savings over SIFT are more pronounced, since SURF is a speeded up version of SIFT. The computation time comparison shows a larger improvement in the case of synthetic data. This is in part attributed to the smaller resolution of images taken by mobile phones (320x240) in the real dataset, rendering the compensation for global rotation angles more expensive than estimating angles on an interest point level, because of the small number of interest points detected on average.

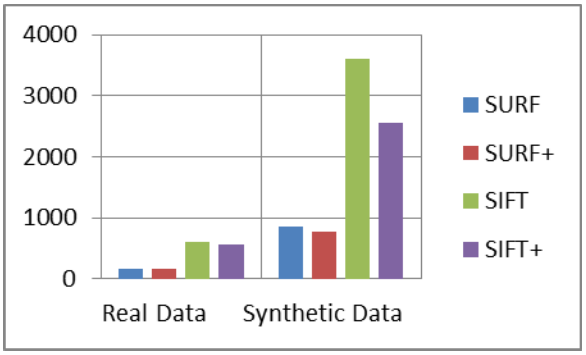


Figure 6 Average computation time (msec) for for real and synthetic datasets

6.2 Precision/recall

In the descriptor evaluation, the overlap error threshold is fixed to 50% for the computation of correspondences. Hence, for each image pair, we have a single precision/recall value pair rather than a full graph. Figure 7 shows recall values for the synthetic and real datasets respectively with an overall improvement noticeable especially for the case of comparing with SIFT. These results also show that the relative improvement when using SURF+ compared to SURF alone is considerably higher than in the case of using SIFT+ compared to SIFT alone. This is because of the fact that SURF is an approximation of SIFT, and thus less accurate.

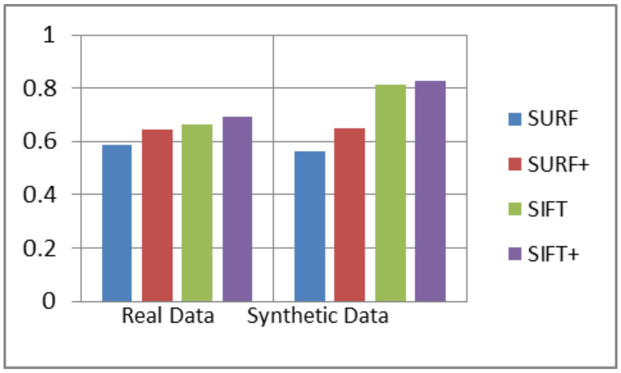


Figure 7 Average Recall for the four schemes for the real and synthetic datasets

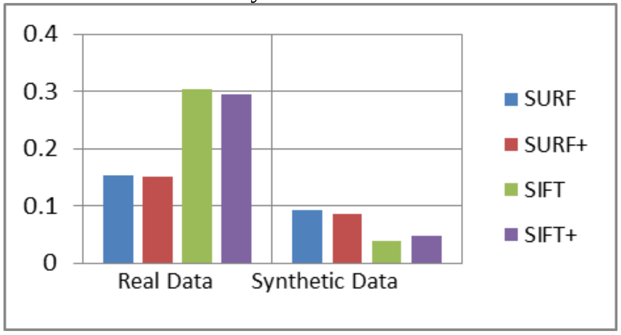


Figure 8 Average (1 – precision) for the four schemes for the real and synthetic datasets

On the other hand, precision values do not enjoy the same level of improvement as the recall values. Figure 8 shows that the results with and without the accelerometer information are very comparable.

7. CONCLUSION

In this paper we have proposed a method for fusing interest-point based image matching descriptors with information from the accelerometer sensor, which is commonly present today in many image capturing devices. The proposed method has been empirically evaluated in terms of precision/recall and computation time, and showed gains against two popular descriptors SIFT and SURF. As for future work, usage of compass information is suggested for estimating rotation around the central axis as well as investigating the effect of global rotation correction on the feature detection stage.

8. REFERENCES

[1] Szeliski, R., “Computer Vision: Algorithms and Applications”, Springer, 2010, available online: http://research.microsoft.com/people/szeliski/Book/

[2] Bay, H., Tuytelaars, T., and Gool, L. V., “SURF: Speeded up robust features”, In Leonardis, A.,Bischof, H., and Pinz, A., editors, Computer Vision – ECCV 2006, pages 404–417, Springer, 2006.

[3] Lowe, D. G., “Distinctive image features from scale-invariant keypoints” International Journal of Computer Vision, 60(2), 91–110, 2004

[4] Martin, D., Fowlkes, C., and Malik, J., “Learning to detect natural image boundaries using local brightness, color, and texture cues”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(5), 530–549, 2004.

[5] Gevers, T., van de Weijer, J., and Stokman, H., “Color feature detection”, In Lukac, R. and Plataniotis,K. N., editors, Color Image Processing: Methods and Applications, CRC Press, 2006.

[6] Mortensen, E. N., “Vision-assisted image editing. Computer Graphics”, 33(4), 55–57, 1999.

[7] Shi, J. and Tomasi, C., “Good features to track”, In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’94), pages 593–600, 1994.

[8] Zhang, Z. et al., “A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry”, Artificial Intelligence, 78, 87–119, 1995.

[9] Florack, L.M.J., Haar Romeny, B.M.t., Koenderink, J.J., Viergever, M.A., “General intensity transformations and differential invariants”, JMIV 4, 171–187, 1994

[10] Mindru, F., Tuytelaars, T., Van Gool, L., Moons, T., “Moment invariants for recognition under changing viewpoint and illumination”, CVIU 94, 3–27, 2004

[11] Baumberg, A., “Reliable feature matching across widely separated views”, In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’2000), pages 774–781, 2000.

[12] Matas, J. et al., “Robust wide baseline stereo from maximally stable extremal regions”, Image and Vision Computing, 22(10), 761–767, 2004.

[13] Lindeberg, T. and Gøarding, J., “Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure”, Image and Vision Computing, 15(6), 415–434, 1997.

[14] Mikolajczyk, K. and Schmid, C., “Scale & affine invariant interest point detectors”, International Journal of Computer Vision, 60(1), 63–86, 2004.

[15] Mikolajczyk, K. et al., “A comparison of affine region detectors”, International Journal of Computer Vision, 65(1-2), 43–72, 2003.

[16] Tuytelaars, T. and Mikolajczyk, K., “Local invariant feature detectors”, Foundations and Trends in Computer Graphics and Computer Vision, 3(1), 2007.

[17] Morel, J.M., Yu, G.S., “ASIFT: A new framework for fully affine invariant image comparison”, SIAM Journal on Imaging Sciences, No. 2, pp. 438-469, 2009.

[18] Durrant-Whyte, H., and Bailey, T., “Simultaneous localization and mapping (slam): Part i”, IEEE Robotics and Automation Magazine, pp. 99- 108, June 2006.

[19] Davison, A. J. and Murray D. W., “Simultaneous localization and map-building using active vision”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002.

[20] Gil, A., Reinoso, O., Burgard, W., Stachniss,C. and Martınez Mozos, O., “Improving data association in rao-blackwellized visual SLAM”, In IEEE/RSJ Int. Conf. on Intelligent Robots & Systems, 2006.

[21] Little, J., Se, S. and Lowe D. G., “Global localization using distinctive visual features”, In IEEE/RSJ International Conference on Intelligent Robots & Systems, 2002.

[22] IPTC (1999). IPTC-NAA Information Interchange Model Version 4.1. Retrieved April 4, 2010, from http://www.iptc.org/std/IIM/4.1/specification/IIMV4.1.pdf

[23] Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. PAMI 27 (2005) 1615–1630.

[24] Seon H. K., Sakire A. A., Byunggu Y. and Roger Z., “Vector model in support of versatile georeferenced video search”, In MMSys '10 Proceedings of the first annual ACM SIGMM conference on Multimedia systems.