METRIC MEASUREMENT FROM STREET VIEW SEQUENCES WITH SIMPLE OPERATOR ASSISTANCE AND PHASE CORRELATION BASED FRAME SELECTION

Ersin Ozuag, M. Kemal Gullu, Oguzhan Urhan, Sarp Erturk

1Kocaeli University Laboratory of Image and Signal processing (KULIS)
Electronics and Telecom. Eng. Dept., University of Kocaeli
Umutepe Campus, 41380, Kocaeli, Turkey
2Pars Ar-Ge Ltd., University of Kocaeli Technopark
Kocaeli, Turkey
email: {ersin.ozuag, kemalg, urhano, sertur}@kocaeli.edu.tr

ABSTRACT
This paper presents a metric measurement approach from sequences of images captured from a moving spherical camera without the need of additional equipment, such as laser scanners or motion detection units. The user assists the algorithms with simple inputs to facilitate the measurement process. The operator initially selects a keyframe that contains the object of interest that is to be measured. Next, a suitable pair is selected for this keyframe, automatically, using a novel phase correlation based approach proposed in this paper. Then, correspondence matching between these two images is performed using scale-invariant feature transform (SIFT) and these features are refined using RANdom SAmple Consensus (RANSAC) and information obtained from the phase correlation stage. As a last step conversion form the image domain to the 3D domain is performed. The user selects two corresponding point pairs in both frames, corresponding to the edges of the distance that is to be measured, and the metric distance between these two points is obtained. During this process, the height information of the camera with respect to the ground is used as basic reference to obtain metric results. Experimental results show that the proposed methods can provide metric measurements with up to 10% error.

1. INTRODUCTION
Metric measurement from multiple images has been attracting increased interest [1]. When physical relations between camera/image positions are known, the problem is relatively easier to solve. However, if the physical relation between images or cameras is not known the process gets more difficult and complicated.

In approaches presented in the literature, multiple overlapping images [2], known reference distances in the scene [3], geo-located cameras with GPS and/or IMU data [4, 5], or additional laser scanners [6, 7], have been widely used to obtain metric measurements. However, all of these methods either require restrictions on image acquisition, additional information or knowledge about the scene or extra hardware.

On the other hand, approaches that use only image sequence data without many constraints or knowledge about the scene are rather limited. In this paper, we present a user aided metric measurement method from sequential images captured by a spherical camera in a street view setup, without the need of any extra hardware or scene constraints. The only requirement of the proposed method is that the camera height from the ground has to be known, as this information is used as metric reference, similar to [8]. Actually, this setup is commonly the case where the camera is fixed on a moving platform such as a car. Google Maps with Street View [9] is an example application that employs a camera system mounted on a vehicle to capture panoramic videos. In a similar setup, images have been captured using a Point Grey Ladybug camera mounted on the top of a car.

An important issue is the selection of an appropriate frame pair that is to be used to perform metric measurements, so as to ensure a certain quality level. In this paper, we present an automatic approach to choose a suitable image pair for matching, based on phase correlation. Afterwards, feature points extracted from the image pair using SIFT are matched taking dominant camera motion information obtained in the phase correlation stage into account. Incorrect matches are eliminated using RANSAC [10] and epipolar constraints. In the final step, real world positions of selected image points are computed from 3D geometry, and real world metric distance measurement is accomplished.

Section II describes the selection of suitable image pairs and the feature matching process. Metric measurement with experimental data is explained and demonstrated in Section III. Conclusions are provided in Section IV.

2. AUTOMATIC SELECTION OF APPROPRIATE IMAGE PAIRS USING PHASE CORRELATION
Selection of a good image pair for 3D reconstruction is very important to extract reliable and accurate depth information. As recorded video typically contains more than 15 frames per second, consecutive image frames can have small camera motion and hence extensive overlap of image content with little displacement so that the content is not suitable for accurate 3D measurement because of insufficient disparity content.

In the proposed scenario, the user selects one keyframe that contains the object/scene that is to be measured. In order to accomplish successful 3D reconstruction for metric measurement it is then required to obtain a suitable corresponding image pair. In the proposed approach, support hardware, such as GPS and/or IMU are not used, in order to...
reduce hardware complexity and cost, hence the camera motion is not precisely known.

A suitable corresponding image pair is required to have sufficient global camera motion with respect to the keyframe, so that 3D information can be extracted more reliably and accurately. In this paper a novel approach is proposed for this purpose. In the proposed approach, the candidate successive frames are compared to the keyframe by computing the global motion between the two frames under inspection, using phase correlation. The first frame in the sequence following the keyframe that has a substantial global motion (above a certain threshold) with respect to the keyframe selected by the user is assigned as the corresponding image pair.

In the literature image based methods such as feature point tracking [5], movement model selection mechanisms [11], as well as similarity-based and linear discriminant based methods have been proposed to select discriminative frames [12]. In this work, it is proposed to utilize a phase correlation based method to obtain global camera movement between two frames. The novel phase correlation based method is utilized because it provides low computational complexity in the frame pair selection process and also provides information about the global displacement between the frames that will be used in the consecutive steps, in a robust way.

If we assume that the motion between image frames \( (I_t, I_{t+k}) \) is pure translational and these frames have brightness changes it is possible to formulate the relation in the form of

\[
I_{t+k}(x, y) = \alpha \times I_t(x - d_x, y - d_y)
\]

where \( \alpha \), \( d_x \) and \( d_y \) represent contrast difference, and horizontal and vertical displacements, respectively. If \( F_t(u, v) \) shows the 2D discrete Fourier transform (DFT) of frame \( I_t \), then the 2D DFT of frame \( I_{t+k} \) will be related by

\[
F_{t+k}(u, v) = \alpha \times F_t(u, v) e^{-j(u d_x + v d_y)}
\]

The phase correlation between the two image frames \( (I_t, I_{t+k}) \) is then computed as

\[
S_{(t, t+k)} = F^{-1} \left[ \frac{F_t(u, v) \times F_{t+k}^*(u, v)}{\left| F_t(u, v) \times F_{t+k}^*(u, v) \right|} \right] = F^{-1} \left[ e^{j(u d_x + v d_y)} \right]
\]

as

\[
S_{(t, t+k)} = \delta(x + d_x, y + d_y)
\]

where \( F^{-1} \) shows the inverse 2D DFT, and * represents the complex conjugate. As seen from (3), the phase correlation surface provides a peak at the location corresponding to the translation between the two image frames.

In practice, usually the phase correlation is unable to provide a single reasonable distinctive peak value for real images, as shown in Figure 1 (a). This is mostly because there is actually not a simple 2D translation between the two images but a difference in depth variations of different image regions. In order to improve the performance of the phase correlation for real image frames, down-sampling is used in

the phase correlation process. As seen from the results shown in Fig. 1(b), a distinct peak is obtained after down-sampling. Thus, the dominant motion of the camera between the two frames can be obtained more reliably. If this dominant motion is larger than a pre-determined threshold this image pair is suitable for utilization in the 3D reconstruction process. Otherwise, the next succeeding frame is evaluated in a similar way, until sufficient dominant motion between the two image frames is obtained. Note that the down-sampling process also reduces the computational load of the phase correlation process.

![Figure 1](image1.png)
3. FEATURE POINT EXTRACTION AND MATCHING

The next step of the proposed method is to extract features from the image pairs selected in the previous stage. SIFT [13], which is also frequently used in the literature, is used to extract features in this case. Prior to SIFT computation, histogram equalization is performed for both images to increase contrast and the performance of feature detection.

In the proposed approach, the global motion information obtained from the phase correlation process as explained in Section 2, is used as additional information to reduce incorrect matches, instead of conventional matching utilized in the literature. Hence, features extracted from the image pairs are matched making use of the distance and direction of the dominant motion of the camera. In the matching process, feature points are initially sorted in a decreasing order of their similarity values with respect to the feature vector of the second image frame. Next, distance and direction of the matching vectors are compared with respect to the dominant camera motion, and incorrect matches are excluded. An example matching result, without making use of the dominant motion vector is depicted in Fig. 2 (a), and the result obtained with assisted dominant motion vector is depicted in Fig. 2(b). Although an important part of remaining matches are correct, some incorrect matches remain as the phase correlation process provides only a rough estimate. This is the case, because the basic assumption is actually that the dominant camera motion obtained using phase correlation represents the motion of the entire image frame which is obviously not correct because of the perspective effect. Hence, some incorrect matches remain. Although, there are some methods that can be used for feature tracking; such as approaches presented in [14] and [15] for example, to avoid this kind of problem, these methods introduce additional computational load; and methods for implementing such approaches on the GPU have been explored [16, 17].

Instead, it has been preferred to compute inlier and outlier points under epipolar constraints by applying RANSAC [10]. Matching points remaining after epipolar constraints are shown in Fig. 3 (a) and (b). Since this approach is observed to provide sufficiently reasonable matching results any additional feature tracking approach is not utilised in this paper. This approach is explained in more detail in the next Section.

4. EXTRACTION OF CAMERA GEOMETRY FROM TWO VIEWS

The epipolar geometry between two views is defined by the fundamental matrix [18]. The fundamental matrix verifies the following transformation between corresponding points in the two images for the perspective motion and pinhole camera models:

$$P_2^T FP_1 = 0$$

where $P_1$ and $P_2$ are the locations of points in the image coordinates [3]. The fundamental matrix $F$ defines the epipolar geometry between the two images and also contains intrinsic camera parameters [19]. In the proposed approach the camera has been calibrated prior to utilization.

Calculation of the fundamental matrix is performed in two steps. Firstly, feature points are classified as inliers and outliers with a voting algorithm similar to [20] and recalculation of the fundamental matrix is carried out, using only the inliers from the first step using standard deviations of errors similar to [21]. In this way the most powerful feature points are obtained. This approach successfully avoids the calcula-
tion of an incorrect fundamental matrix. Fig. 3 shows the inlier feature points, obtained with this approach.

The Fundamental matrix $F$ is related to the Essential matrix $E$ through

$$E = K^T FK$$  \hspace{1cm} (5)$$

where $K$ is the camera calibration matrix. The camera calibration matrix is obtained off-line prior to utilization using a checkerboard pattern as in [22]. Hence, the camera calibration matrix is already known and for any image pair it is possible to obtain the essential matrix from the fundamental matrix.

Then it is possible to obtain translation and rotation matrices from the essential matrix using the relation

$$E = [t]_x R$$  \hspace{1cm} (6)$$

where $[t]_x$ and $R$ represent the translation matrix and the rotation matrix, respectively [3]. The camera matrices are created using the translation and rotation matrices as in [3], and triangulation based on “The Linear-Eigen Method” as in [23] is applied to obtain 3D Euclidean coordinates. As triangulation is performed for two point pairs in the two images, it becomes possible to measure the distance between these two points. However, if no knowledge about the physical movement of the camera is available, and there is also no physical reference data, there is always a scaling ambiguity, which is an important problem for metric measurement. Operator assistance is utilized to overcome the scaling ambiguity.

5. USER AIDED 3D METRIC MEASUREMENT

In many applications, geo-located cameras have been used to transform Euclidean geometry to metric scale, as in [5] for example. In these applications, metric scaling is obtained from the camera displacement vector. However, these methods typically require extra support hardware (GPS and/or IMU) and thus have relatively higher complexity and cost.

Another method used in the literature to transform from Euclidean geometry to metric scale is to make use of some known distances in the scene [3]. But in most cases it is not possible to have this information in practice in a simple approach.

In the proposed approach, the basic assumption is that the camera has a fixed height and that this height is known a priori, similar to [8]. This is the typical case when the camera is mounted on a vehicle as in [9] for example. Thus, the knowledge about the metric camera height from the ground is available. At this stage, a ground point is determined using user assistance. The operator assigns a point located on the ground of the scene. This ground point is used to obtain a reliable scaling parameter. This scaling parameter enables metric distances.

In Fig. 4, example measured distances by the proposed method are shown. In this case, the ground point (P1) and its corresponding match in the second image pair is determined by the user. Finally, the user may select two corresponding points in the image pairs and the distances between these points are computed automatically by the proposed approach. Some example results are overlaid in the image shown in Fig. 4. It is observed that the proposed method has a maximum measurement error of about 10% only. Similar results are obtained for different sequences used in the experimental evaluation, confirming the maximum 10% error.

Future work includes the utilization of bundle adjustment [24] for further reduce measurement accuracy.

6. CONCLUSIONS

A vision based user aided metric measurement approach for street view sequences is presented in this paper. The proposed approach automatically chooses a reliable image pair using phase correlation. Next, feature extraction and matching is carried out. User assistance is used to enable metric measurement. Experiments show that, the proposed method has about 10% measurement error. The measurement errors can be accounted to photometric and computational errors.

7. ACKNOWLEDGMENT

This work was in part supported by the Turkish State Planning Organization project DPT 2011K120330.
Figure 4: Example results. P1: user selected ground point.
Metric measurement results of the proposed method. Real distances $d_1 = 3.88\text{m}$, $d_2 = 3.15\text{m}$, $d_3 = 4.00\text{m}$, $d_4 = d_5 = 1.20\text{m}$.

REFERENCES


