Robust Painting Recognition and Registration for Mobile Augmented Reality

Niki Martinel*, *Student Member, IEEE,* Christian Micheloni, *Member, IEEE,*
and Gian Luca Foresti, *Senior Member, IEEE*

Abstract

In this work we introduce a novel approach for painting recognition and registration for mobile Augmented Reality applications. To address the challenges of real-time painting recognition and registration we introduce three main contributions: (i) A relevant painting region detector extracts the painting region from the given image. (ii) Two local and global features are extracted from the relevant region to robustly match a painting database. (iii) A RANSAC homography estimation method is used to overlay the additional content in an AR framework. Experiments have been carried out on a dataset built with publicly available images.

Index Terms

Painting recognition, registration, mobile device, augmented reality.

I. Introduction

Since the early years of mobile devices technology, the computer vision community has been interested in mobile cameras. Whilst in the past applications and algorithms focused on the interaction and productivity, recently, with the technology evolution, the community has been actively involved in mobile vision. Nowadays, many computer vision solutions find applications in mobile devices. Within such a category, Augmented Reality (AR) is gaining more and more interest from both researchers and end-users. Of particular interest for AR applications is the task of recognizing objects such that related information can

The authors are with the Department of Mathemarics and Computer Science, University of Udine, Udine, Italy, 33011 Italy. E-mail: {niki.martinel, christian.micheloni, gianluca.foresti}@uniud.it

Manuscript received April 19, 2005; revised January 11, 2007.

be shown through device displays [1]. In this work, we focus on the tasks of painting recognition and registration.

Recognizing paintings and computing the transformation to align the acquired image of a painting and its image in a database are nontrivial tasks. Common static computer vision issues (e.g. illumination, view, scale changes, etc.) are more severe in context of moving cameras as different effects (e.g. blur, noise, motion, etc.) arise. In addition, reflection of spotlights, image saturation and image exposure add up for the recognition of the paintings present in exhibitions.

In the recent years, the problems of image retrieval [2], landmark [3] and location [4] recognition using appearance-based features have been deeply investigated. To achieve their objectives, those methods match appearance features against a large database of location-tagged images [5]. While many approaches were proposed, only a little effort was spent to tackle the challenges posed by both the recognition and the registration tasks. In [6], the monoSLAM system estimates the hand-held camera's motion from the live image stream to achieve high AR performance. In [7], [8], [9] recent techniques for tracking and occlusion handling in an AR framework were discussed. The challenges of real-time recognition and camera pose estimation system for planar shapes were addressed in [10]. The proposed system performs shape recognition by analyzing contour structures and generating projective-invariant signatures. In [11], an AR rendering pipeline that supports global illumination techniques was proposed.

While specifically designed markers have been the dominant choice by state-of-the-art AR methods, the use of such markers in a real scenario is not always feasible. Motivated by this and inspired by feature-based computer vision techniques, we propose a novel marker-less approach to cope with the AR challenges. We introduce a method to detect and extract the relevant painting region (RPR) from a given input image. Local features are extracted from the RPR and matched with a candidate target in the database. RANSAC is used to detect feature outliers. As the current image may not be aligned with the candidate target image, extracting global feature from it considerably reduces robustness. To tackle this, we use the homography transformation output by RANSAC to align the current RPR to the candidate target RPR. This allows to extract global feature from the aligned RPR only, thus noticeably improving performance. Then, a weighted similarity measure is used to compute the final match. Once a match is found, we use the RANSAC homography transformation to properly overlay the additional content to the current frame.

To summarize, we introduce the following contributions: (i) A RPR detector extracts the painting region from the current image. (ii) Two local and global features are used to robustly match database paintings. (iii) The RPR detector together with a robust feature-based matching technique and RANSAC

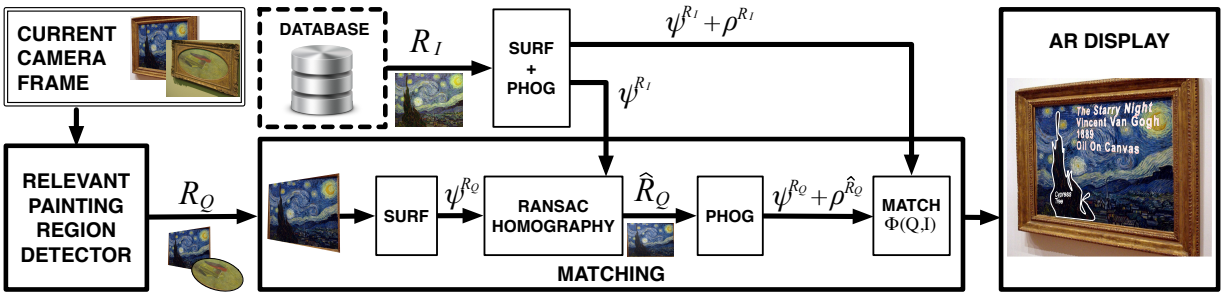


Fig. 1. Architecture of the proposed system. The current camera frame is processed by the three main modules of the system. The first module extracts the RPR. The second module computes the match between the current camera frame and a candidate target in the database. Given a match with the candidate target, the homography transformation used to overlay the additional content to the current camera frame is computed by the third module.

are used to display the additional content in an AR framework. To evaluate the proposed system a dataset of paintings has been built with publicly available images.

II. System overview

The architecture of the proposed approach is shown in Fig. 1. Three main modules are used to achieve the proposed goal: i) the RPR detector module, ii) the matching module and iii) the AR display module. Given the current camera frame, the RPR detector module is in charge to extract only the RPR, thus rejecting the painting frame and the background. Once the RPR is detected, the matching module extracts the local features from such region and matches them with the candidate target local features. RANSAC is used to reject matching outliers. Then, the homography matrix output by RANSAC is used to align the current RPR with the candidate target RPR from where the global features are then more robustly extracted. The local and global features that form the current signature are finally matched with the candidate signature using a linear weighted similarity measure. Finally, given the best match with a candidate target, the homography transformation used by RANSAC is used to overlay the additional content to the current camera frame.

The values of the algorithm parameters given in the following sections have been selected using 4-fold cross validation (see Section VI for details).

III. Relevant painting region detector

Assuming that the frames of paintings have elliptical or rectangular shape, the RPR detector removes the background and the painting frame to keep only the portion of the image that contains the painting, i.e. the RPR. The RPR detector achieves its objective using the Randomized Hough Transform

(RHT) method [12]. The RHT is a particular derivation of the Hough Transform (HT) that avoids the computationally expensive HT voting procedure. As shown in [12], the RHT can achieve a computational complexity lower than an upper bound of order $O((n_t N^n)/n_{min}^n)$, that is considerably smaller than the order $O(NN_a^{n-1})$ of the standard HT. N and N_a are the total number of pixels in the image and the size of the accumulation array respectively. n is the number of curve parameters, n_{min} is the length of the shortest curve in the image, and n_t is a small number.

To detect the RPR, we first apply the Canny operator to the grayscale representation of the current camera frame $Q \in \mathbb{R}^{M \times N}$. Then, the RHT is used to fit ellipses and rectangles. As a painting may contain more than a single rectangular or ellipse shaped object, we consider the RPR boundary to be the detected rectangle or ellipse which area is at least r times the input image size. The detected RPR is denoted as $R_Q \in \mathbb{R}^{M' \times N'}$ where $M'N' \geq rMN$ and $r \in [0, 1]$.

IV. MATCHING

Let R_Q and R_I be the RPR of the current camera frame Q and the RPR of a database candidate target image I, respectively. To match R_Q and R_I we consider two local and global features. To achieve this objective with lower computational costs, the Speeded-Up Robust Features (SURF) [13] and the Pyramid of Histogram of Oriented Gradients (PHOG) [14] features are used.

Local features: As illumination invariance is intrinsic to SURF [13], we extract such features from the grayscale representation of R_Q by exploiting the standard integral image. The SURF feature detector is based on an approximation of the Hessian matrix, while the feature descriptor $\psi_F^{R_Q} \in \mathbb{R}^{64}$ describes the distribution of Haar-wavelet responses within the neighborhood of the detected interest points $\psi_K^{R_Q} = [x, y]$. The computed SURF feature vector is denoted as $\psi^{R_Q} =< \psi_F^{R_Q}, \psi_K^{R_Q} >$.

Given two SURF feature descriptors $\psi_F^{R_Q}(q)$ and $\psi_F^{R_I}(i)$, we consider q, i being a match if the similarity

$$S_F(\psi_F^{R_Q}(q), \psi_F^{R_I}(i)) = \frac{1}{1 + \|\psi_F^{R_Q}(q) - \psi_F^{R_I}(i)\|_2} \quad (1)$$

is higher than a fixed threshold Th_s. Matching features are then analyzed to detect outliers using an approach similar to [15]. Given 4 feature correspondences, the homography $H_{Q,I}$ is computed using the Direct Linear Transformation method [15]. The process is repeated with t trials, and the solution that has the maximum number of inliers is selected. A SURF feature keypoint $\psi_K^{R_Q}(q)$ is considered to be an inlier if the corresponding keypoint projection $\widehat{\psi}_K^{R_Q}(q)$ is consistent with $H_{Q,I}$ within a tolerance of σ pixels.

Global features: Given $H_{Q,I}$, we use it to align R_Q to the R_I. The transformed RPR is denoted as $\hat{R}_Q$. This operation allows us to extract the global feature in a more robust fashion as, after such

transformation, the edges used to extract the global features are aligned and have similar orientations as the edges of the candidate target.

PHOG features are extracted from $\hat{R}_Q$ to capture information about the shape and the whole appearance of the painting. Before extracting PHOG features, we project $\hat{R}_Q$ into the HSV color space to achieve illumination invariance. Then, for each of the three channels, edges and orientation gradients are used to compute the PHOG feature matrix $\rho^{\hat{R}_Q} \in \mathbb{R}^{m\times 3}$. m is the total number of histogram bins for each image channel.

Candidate target matching: Once the local and global features have been extracted, we match Q and I as follows.

SURF features similarity is computed as

$$\Phi_\psi\left(\psi^{R_Q}, \psi^{R_I}\right) = \frac{\sum_{q,i\in match} S_F(\psi_F^{R_Q}(q), \psi_F^{R_I}(i))}{\epsilon + match} \tag{2}$$

where $match$ is the total number of matched SURF features and ϵ is a small constant used to prevent division by zero.

PHOG features are matched as suggested in [16]. Let $\rho^{\hat{R}_Q}$ and ρ^{R_I} be the PHOG feature matrices of $\hat{R}_Q$ and R_I respectively. The PHOG similarity is computed as

$$\Phi_\rho(\rho^{\hat{R}_Q}, \rho^{R_I}) = 1 - \sum_c \lambda_c \chi^2(\rho_c^{\hat{R}_Q}, \rho_c^{R_I}) \tag{3}$$

where $\rho_c^{\hat{R}_Q}$ and $\rho_c^{R_I}$ are the PHOG features computed for the c-th color channel. λ_c is the normalization weight.

Let $\mathbb{I}$ be the set of all database images, the objective is to find $\arg\max_{I\in\mathbb{I}} \Phi(Q, I)$ where

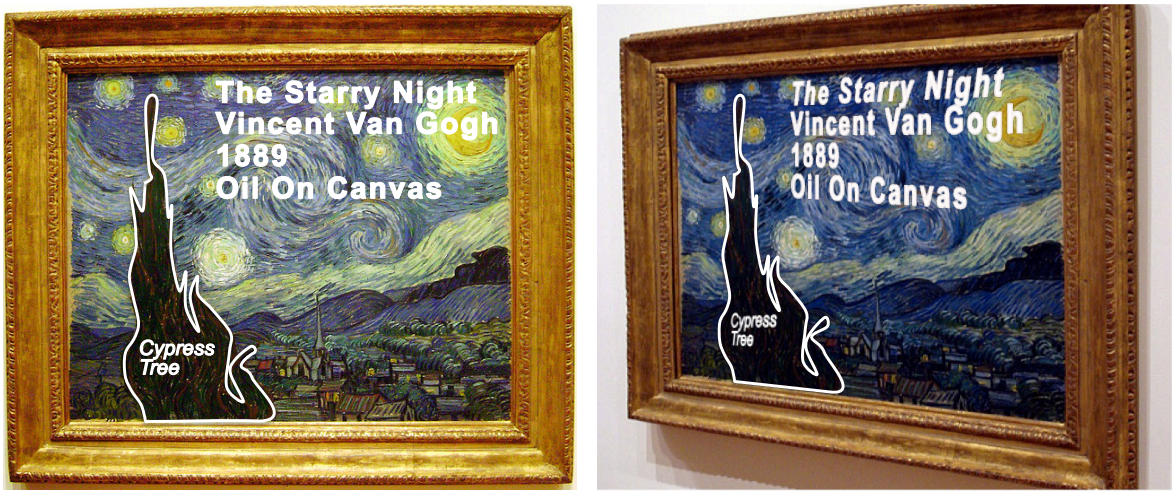
$$\begin{aligned}\Phi(Q, I) \quad = \quad & \alpha\, \Phi_\psi\left(\psi^{R_Q}, \psi^{R_I}\right) + \\ & \beta\, \Phi_\rho(\rho^{\hat{R}_Q}, \rho^{R_I})\end{aligned} \tag{4}$$

α and $\beta = 1 - \alpha$ are the normalization weights.

V. AR DISPLAY

The last module of the proposed system is in charge to overlay the additional content to the current camera frame Q. As both paintings and the additional content are planar surfaces, the transformation we need to compute can be described as an homography.

According to [15], the literature defines two categories of automatic homography computation: i) direct and ii) feature based. In this work we use a feature-based homography computation method. In



(a) (b)

Fig. 2. Example of AR. (a) Reference frame with the additional information to display. (b) Current frame with the additional information transformed using the feature-based homography transformation.

particular, to save computational resources, we use the same feature-based homography transformation $H_{Q,I}$ computed in section IV. As shown in Fig. 2, using the inverse homography transformation matrix $H_{Q,I}^{-1}$ it is possible to overlay the additional content (given in the original region I coordinate system) to the current camera frame Q.

VI. Experimental results

Experiments have been carried out on a dataset built using 607 publicly available pictures of 70 Vang Gogh paintings. Pictures are taken from different viewing angles and under different illumination conditions. Some pictures come with light reflections and occlusions.

To evaluate our approach we selected the following parameters using 4-fold cross validation. The RPR boundary parameter r has been set to 0.55. SURF features have been computed using 5 octaves and, for each octave, the number of scale levels has been set to 4. To compute PHOG features, edges are extracted using the Canny operator, while orientation gradients are computed using a 3×3 Sobel mask. The extracted HOG features are quantized in 9 bins at 4 levels of the spatial pyramid. The normalization weight vector λ has been set to $[0.5, 0.3, 0.2]$. The tolerance σ has been set to 4 pixels and $t = 500$ trials are performed to compute H. The matching threshold Th_s has been set to 0.85. To show the performance of the method we report the results as ROC curves and normalized Area Under Curve (nAUC) values.

The algorithm has been tested on a standard PC with P4 CPU 2.0GHz, 1GB RAM, Windows XP and on a Tablet with ARMv7 processor 1GB RAM, Android 4.2.2. In the first test with non-optimized

α	**0**	**0.25**	**0.28**	**0.5**	**0.75**	**1**
RPR Detection	0.7927	0.8110	**0.8320**	0.7485	0.6917	0.6281
No RPR Detection	0.6368	0.6572	**0.6635**	0.6470	0.6341	0.6290

TABLE I

AVERAGE NAUC VALUES FOR DIFFERENT VALUES OF ALPHA (FROM 0 TO 1 WITH STEPS OF 0.01). BEST RESULTS ARE IN BOLDFACE FONT.

	Rotation	**0°**	**45°**	**90°**	**135°**	**180°**	**225°**	**270°**	**315°**
RPR Detection	**Scale = 0.5**	0.9223	0.8025	0.8918	0.7178	0.9020	0.6708	0.8949	0.7518
	Scale = 0.75	0.9243	0.8049	0.9032	0.7158	0.9128	0.7010	0.8986	0.8209
	Scale = 1	0.9258	0.8233	0.9138	0.7188	0.9180	0.7044	0.9053	0.8226
No RPR Detection	**Scale = 0.5**	0.8130	0.5502	0.7015	0.5492	0.6810	0.5476	0.7963	0.6133
	Scale = 0.75	0.8104	0.5529	0.7023	0.5510	0.6966	0.5498	0.8082	0.6214
	Scale = 1	0.8244	0.5655	0.7026	0.5738	0.7082	0.5585	0.8144	0.6329

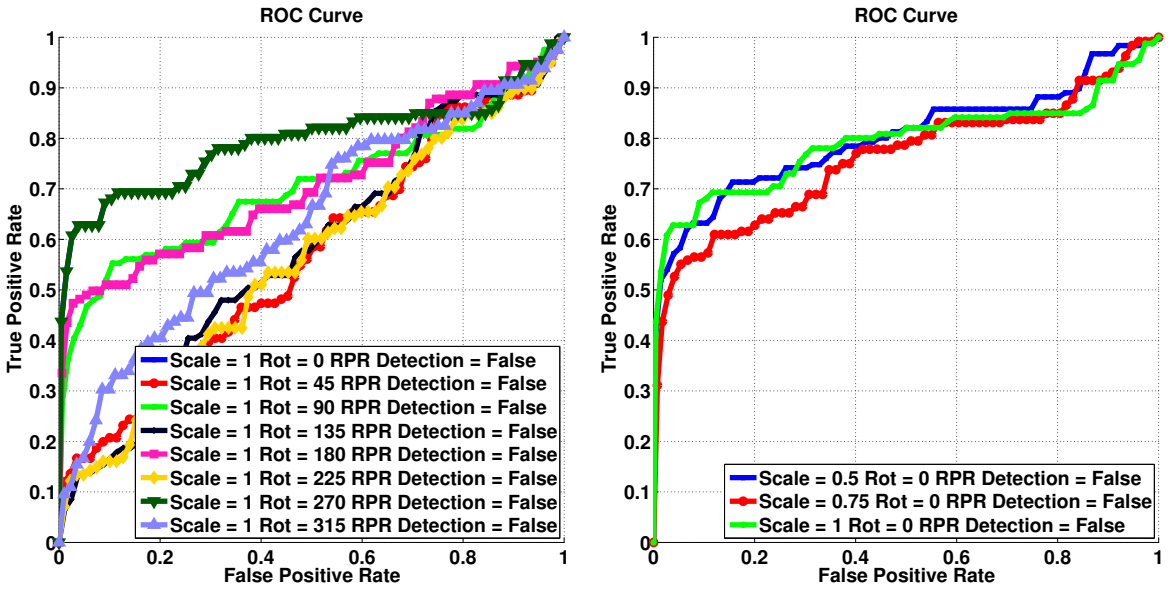
TABLE II

NAUC VALUES COMPUTED FOR TEST IMAGES SCALED TO $1/2$, $3/4$ AND $1/1$ OF THE ORIGINAL SIZE AND ROTATED FROM 0°TO 315°(STEPS OF 45°).

MATLAB code the average recognition and registration time for a single frame is 0.591s, while in the latter with optimized code the same activities take 0.632s.

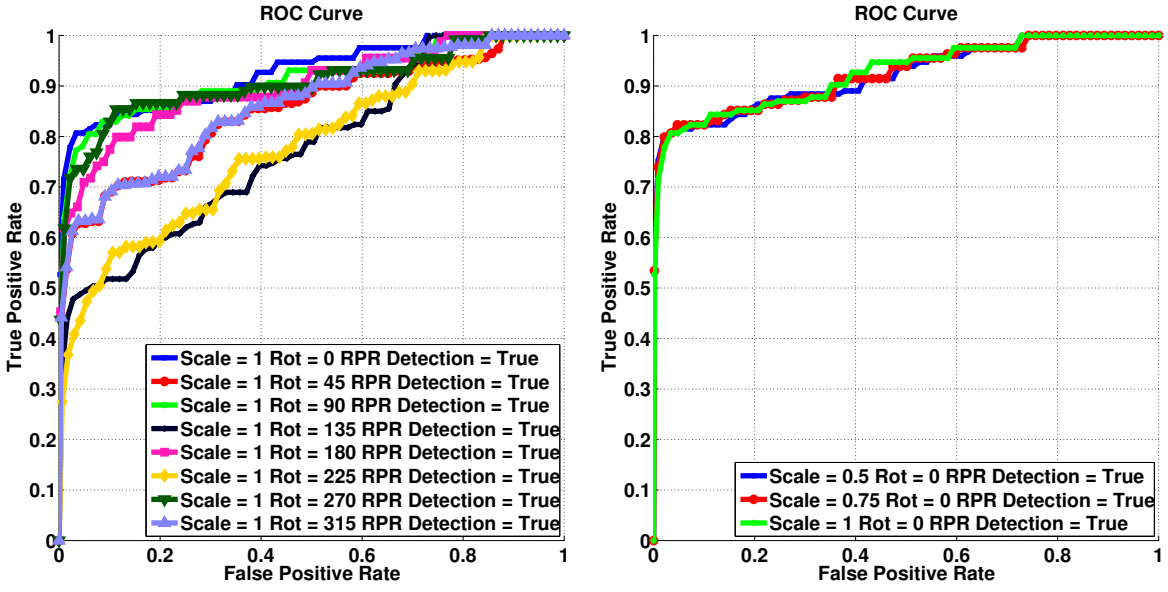
In Table I, we report the nAUC values computed as a function of the similarity normalization weight α. Each value is computed averaging all the results computed for images scaled to $1/2$, $3/4$ and $1/1$ of the original image size and for different image rotations from 45° to 315°, with intermediate rotations of 45°. The best results are achieved for $\alpha = 0.28$: in the following experiments such value has been used.

In Fig. 3 we show the performance of our method without using the RPR detector. In Fig. 3(a) results are shown as a function of the rotation of the test images. The original scale has been used. The method achieves reasonable results for rotations multiple of 90°. The performance decreases in the other cases. This is probably due to the changes occurring in the oriented gradients used to compute the PHOG features. On average, a true positive rate of 49% is achieved for a false positive rate of 20%. In Fig. 3(b) results for different image scales are shown. Thanks to SURF invariance properties and the pyramidal approach used to compute PHOG, the performance are not much affected by the scaling issues. In such scenario, an average true positive rate of 67% is achieved for a false positive rate of 20%.



(a) (b)

Fig. 3. Recognition performance computed without using the relevant region detector module. In (a) test images are rotated by multiples of 45°. In (b) test images are not rotated but their scaling factor has been changed.



(a) (b)

Fig. 4. Recognition performance computed using the relevant region detector module. In (a) test images are rotated by multiples of 45°. In (b) test images are scaled down using multiple reduction factors.

In Fig. 4 we show the performance of our method using the RPR detector. In Fig. 4(a) results have been computed for different rotations to the test images as in Fig. 3(a). On average, a 71% true positive rate is reached for a false positive rate of 20%. Though, the worst results are reached for rotations of

135°and 225°, where a true positive rate of about 59% is reached for the same false positive rate of 20%. If compared to the results shown in Fig. 3(a), a significant improvement has been achieved. Most importantly, the performance has increased of about 33% for a false positive rate of 0%. In Fig. 4(b) the results are computed varying the scale of test images. If compared to Fig. 3(b), an average improvement of 37% is achieved for a false positive rate of 0%.

In Table II we report the results of our method in terms of nAUC values. We consider images scaled to $1/2$, $3/4$ and $1/1$ of the original image size and rotations from 45° to 315° (with steps of 45°). The first three rows show the results of our method using the proposed RPR detector, while in the last three rows results have been computed without using the RPR detector. For both such cases the best results are achieved when the original image size is kept and no rotation is applied. However, using the relevant region detector, performance increases of more than 17% on average. In particular, for rotation of 90°and 180°, an average increment of 20% is achieved.

VII. Conclusions

In this work a marker-less method for painting recognition and registration that supports mobile AR applications has been proposed. A RPR detector is used to extract only the relevant painting region. The RPR is then considered to extract local features that are matched with a candidate target using RANSAC. The homography transformation output from RANSAC is applied to transform it to the candidate target RPR. This allows to robustly extract global features that are finally used, together with local features, to compute the match between the current frame and the candidate target. Once a valid match is detected, RANSAC homography transformation is used to overlay the additional content to the current frame. The method has been evaluated using a dataset of publicly available images showing significant improvements to standard feature-based matching techniques.

As future plans we will collect more images to evaluate our method against a larger database. To improve the painting recognition performances, we will also investigate novel approaches that allow to identify mobile users locations within indoor environments.

References

[1] M. Gilbert, A. Acero, J. Cohen, H. Bourlard, S.-F. Chang, and M. Etoh, "Media Search in Mobile Devices," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 12–13, Jul. 2011.

[2] B. Girod, V. Chandrasekhar, D. Chen, N.-m. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. Tsai, and R. Vedantham, "Mobile Visual Search," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 61–76, Jul. 2011.

[3] Z. Li and K.-h. Yap, "Content and Context Boosting for Mobile Landmark Recognition," *IEEE Signal Process. Lett.*, vol. 19, no. 8, pp. 459–462, Aug. 2012.

[4] G. Schroth, R. Huitl, D. Chen, M. Abu-Alqumsan, A. Al-Nuaimi, and E. Steinbach, "Mobile Visual Location Recognition," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 77–89, Jul. 2011.

[5] G. Takacs, V. Chandrasekhar, N. Gelfand, Y. Xiong, W.-C. Chen, T. Bismpigiannis, R. Grzeszczuk, K. Pulli, and B. Girod, "Outdoors augmented reality on mobile phone using loxel-based visual feature organization," *International Conference on Multimedia Information Retrieval*, p. 427, 2008.

[6] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: real-time single camera SLAM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1052–67, Jun. 2007.

[7] W. A. Hoff, K. Nguyen, and T. Lyon, "Computer vision-based registration techniques for augmented reality," in *Intelligent Robots and Computer Vision*, vol. 2904, Oct. 1996, pp. 538–548.

[8] V. Lepetit, "On Computer Vision for Augmented Reality," *International Symposium on Ubiquitous Virtual Reality*, pp. 13–16, Jul. 2008.

[9] D. Wagner, G. Reitmayr, A. Mulloni, T. Drummond, and D. Schmalstieg, "Real-time detection and tracking for augmented reality on mobile phones," *IEEE Trans. Vis. Comput. Graphics*, vol. 16, no. 3, pp. 355–68, 2010.

[10] N. Hagbi, O. Bergig, J. El-Sana, and M. Billinghurst, "Shape recognition and pose estimation for mobile Augmented Reality," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 10, pp. 1369–79, Oct. 2011.

[11] L. Gruber, T. Richter-Trummer, and D. Schmalstieg, "Real-time photometric registration from arbitrary geometry," *International Symposium on Mixed and Augmented Reality*, pp. 119–128, Nov. 2012.

[12] L. Xu and E. Oja, "Randomized Hough Transform (RHT): Basic Mechanisms, Algorithms, and Computational Complexities," *CVGIP: Image Understanding*, vol. 57, no. 2, pp. 131–154, Mar. 1993.

[13] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *CVIU*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[14] A. Bosch, A. Zisserman, and X. Munoz, "Image Classification using Random Forests and Ferns," *ICCV*, pp. 1–8, 2007.

[15] M. Brown and D. G. Lowe, "Automatic Panoramic Image Stitching using Invariant Features," *IJCV*, vol. 74, no. 1, pp. 59–73, Dec. 2006.

[16] N. Martinel and C. Micheloni, "Re-identify people in wide area camera network," in *CVPRW*, Providence, RI, Jun. 2012, pp. 31–36.