Contents list available at JMCS

Journal of Mathematics and Computer Science

Journal Homepage: www.tjmcs.com

Human Detection Using SURF and SIFT Feature Extraction Methods in Different Color Spaces

Osameh Biglari[1], Reza Ahsan[2], Majid Rahi[3]
[1]Taali university, Qom, Iran,
osameh.biglari@yahoo.com
[2]Islamic Azad University, Qom branch, Iran,
[3]Pardisan University, Mazandaran, Feridonkenar, Iran,
majid_rahi@heip.ac.ir

Article history:
Received April 2014
Accepted May 2014
Available online June 2014

Abstract

Local feature matching has become a commonly used method to compare images. For tracking and human detection, a reliable method for comparing images can constitute a key component for localization and loop closing tasks. two different types of image feature algorithms, Scale - Invariant Feature Transform (SIFT) and the more recent Speeded Up Robust Features (SURF), have been used to compare the images. In this paper, we propose the use of a rich set of feature descriptors based on SIFT and SURF in the different color spaces.

Keywords: human detection, SIFT, SURF, Color Spaces, grayscale

1. Introduction

Effective techniques for human detection are of special interest in computer vision since many applications involve people's locations and movements. Thus, significant research has been devoted to detecting, locating and tracking people in images and videos. Over the last few years the problem of detecting humans in single images has received considerable interest. Variations in illumination, shadows, and pose, as well as frequent inter- and intra-person occlusion render this a challenging task. Two main approaches to human detection have been explored over the last few years.

The first class of methods consists of a generative process where detected parts of the human body are combined according to a prior human model. The second class of methods considers purely statistical analysis that combine a set of low-level features within a detection window to classify the window as containing a human or not.that as related works in this field, can mention The following cases.

The work of Dalal and Triggs [8] is notable because it was the first paper to report impressive results on human detection. Their work uses HOG as low-level features, which were shown to outperform features such as wavelets [2], PCA-SIFT [3] and shape contexts [4]. To improve detection speed, Zhu et al. [5] Propose a rejection cascade using HOG features. Their method considers blocks of different sizes, and to train the classifier for each stage, a small subset of blocks is selected randomly. Also based on HOG features, Zhang et al. [6] propose a multi -resolution framework to reduce the computational cost. Begard et al. [7] address the problem of real -time pedestrian detection by considering different implementations of the AdaBoostalgorithm. Using low-level features such as intensity, gradient, and spatial location combined by a covariance matrix, Tuzel et al. [8] improve the results obtained by Dalal and Triggs. Since the covariance matrices do not lie in a vector space, the classification is performed using LogitBoost classifiers combined with a rejection cascade designed to accommodate points lying on a Riemannian manifold. Mu et al. [9] propose a variation of local binary patterns to overcome some drawbacks of HOG, such as lack of color information. Chen and Chen [10] combine intensity-based rectangle features and gradient-based features using a cascaded structure for detecting humans. Applying combination of edgelets [11], HOG descriptors [1], and covariance descriptors [8], Wu and Nevatia [12] describe a cascade-based approach where each weak classifier corresponds to a subregion within the detection window from which different types of features are extracted. Dollar et al. [13] propose a method to learn classifiers for individual components and combine them into an overall classifier. The work of Maji et al. [14] uses features based on a multi-level version of HOG and histogram intersection kernel SVM based on the spatial pyramid match kernel [15]. Employing part-based detectors, Mikolajczyk et al. [16] divide the human body into several parts and apply a cascade of detectors for each part. Shet and Davis [17] apply logical reasoning to exploit contextual information, augmenting the output of low-level detectors. Based on deformable parts, Felzenszwalb et al. [18] simultaneously learn part and object models and apply them to person detection, among other applications. Tran and Forsyth [19] use an approach that mixes a part-based method and a subwindowbased method into a two stage method. Their approach first estimates a possible configuration of the person inside

the detection window, and then extracts features for each part resulting from the estimation.

Similarly, Lin and Davis [20] propose a pose-invariant feature extraction method forsimultaneous human detection andsegmentation, where descriptors are computed adaptively based on human poses.

Local feature matching has become an increasingly used method for comparing images. Various methods have been proposed. The Scale-Invariant Feature Transform (SIFT) by Lowe [21] has, with its high accuracy and relatively low computation time, become the de facto standard. Some attempts of further improvements to the algorithm have been made (for example PCA-SIFT by Ke and Sukthankar [22]). Perhaps the most recent, promising approach is the Speeded Up Robust Features (SURF) by Bay et al. [23], which has been shown to yield comparable or better results to SIFT while having a fraction of the computational cost [23, 24].In this paper, has been used from SIFT to different color spaces YCbCr, RGB, HSV. Also to Continuation of this paper has been carried out of SIFT and SURF Algorithms in a grayscale mood for human recognition.

The rest of the paper is structured as follows. In Section II, the SIFT and SURF algorithms and also color spaces are discussed briefly. In Section III, the data sets used in the paper are described. In

Section IV, the experiments are outlined and in Section V the results of the experiments are presented.

2. SIFT and SURF

Both SIFT and SURF contains detectors that find interest points in an image. The interest point detectors for SIFT and SURF work differently. However, the output is in both cases a representation of the neighborhood around an interest point as a descriptor vector. The descriptors can then be compared, or matched, to descriptors extracted from other images. SIFT uses a descriptor of lengths 64 and 128. Depending on the application, there are different matching strategies. A common method, proposed by Lowe [21], is to compute the nearest neighbor of a feature, and then check if the second closest neighbor is further away than some threshold value. Other strategies consider only the nearest neighbor if the distance is smaller than a threshold, as in Zhang and Kosecka [25], or compute only the approximate nearest neighbour by using a kd-tree, as in Beis and Lowe [26].

3. Color Spaces

A color space is a mathematical representation of a set of colors. The three most popular color models are RGB (used in computer graphics); YIQ, YUV, or YCbCr (used in video systems); and CMYK (used in color printing). However, none of these color spaces are directly related to the intuitive notions of hue, saturation, and brightness. This resulted in the temporary pursuit of other models, such as HSI and HSV to simplify programming, processing, and end-user manipulation All of the color spaces can be derived from the RGB information supplied by devices such as cameras and scanners.

4. EVALUATION MEASURE

Precision and recall measures are widely used for evaluation of the classification tasks. They aredefined as follows:

$$precision = \frac{Correct\ assignment\ s\ by\ system}{total\ number\ of\ system\ assignments} = \frac{TP}{TP+FP} \quad (1)$$

$$recall = \frac{Correct\ assignment\ s\ by\ system}{total\ number\ of\ system\ assignments} = \frac{TP}{TP+FN} \quad (2)$$

Where TP is the number of documents correctly assigned to a category, FP is the number of documents incorrectly assigned to a category and FN is the number of documents incorrectly omitted from a category. It is not straightforward to compare the classifiers using two measures, the F1 measure introduced by van Rijsbergen in 1979 combines recall and precision with an equal weight in the following form:

$$F\beta(r,p) = \frac{(\beta^2+1)\times p\times r}{\beta^2\times(p+r)}, where, \beta = 1 \quad (3)$$

The F1-measure has been used for evaluating the accuracy of the classifiers. In this paper, average precision and recall measures are used that their equations are as follows:

$$p^A = \frac{\sum_{j=1}^{|c|} p(c_j)}{|c|} \tag{4}$$

$$r^A = \frac{\sum_{j=1}^{|c|} p(c_j)}{|c|} \tag{5}$$

$$F1^A = \frac{2\times p^A \times r^A}{p^A + r^A} \tag{6}$$

In the equations above, |C| is the number of classes.

5. Experimental Results

In this section the results of our experiments have been presented in the standard data base named ETHZ. With less amount of training dataset, the outcome of our proposed method shows improvement in performance of human detection. In this paper, one image of each person used for training phase and each sample classified by using local feature matching method like SURF and SIFT in different color spaces. Dataset: To obtain a large number of different people captured in uncontrolled conditions, we choose the ETHZ dataset [27] to perform our experiments. This dataset, originally used for human detection, is composed of four video sequences, Samples of three sequence frames are shown in Figure (1).where the one image of sequence #1 is used to train set and the remaining images are used for testing. The ETHZ dataset presents the desirable characteristic of beingcaptured from moving cameras. This camera setup provides a range of variations in people's appearances. Figure (2) shows a few samples of a person's appearance extracted from different frames. Changes in pose and illumination conditions take place and due to the fact that the appearance model is learned from a single sample, a strong set of features becomes important to achieve robust appearance matching during the testing stage. Our experiment has been shown in figure (2) on eight person that each of them in dataset have a different size and view figure (3). For investigating difference size of images on improvement on performance, our experiment has been done on images with different size like 64*64 and 128*128. Finally, for finding the best color space different version of them like RGB, HSV, YCbCr and Grayscale have been tested.

(a) Sequence #1(b) sequence #2

(c) Sequence #3

Figure 1. Samples of the video sequences used in the experiments. (a) sequence #1 is composed of 1,000 frames with 83 different people; (b) sequence #2 is composed of 451 frames with 35 people; (c) sequence #3 is composed of 354 frames containing 28 people.

Table (1): The classification results using SIFT in RGB color space

Size(RGB)	avg_precision	avg_recall	F1
[64 64]	0.7870	0.8176	0.8020
[128 128]	0.7766	0.8283	0.8016

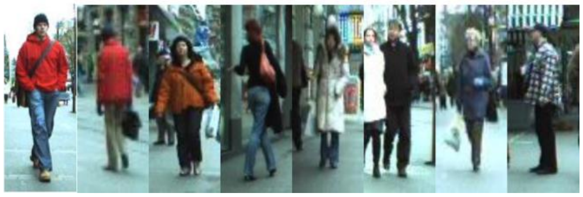


Figure 2: samples from 8 persons in different frames from ETHZ data set

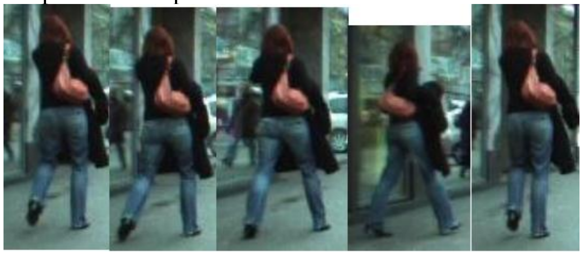


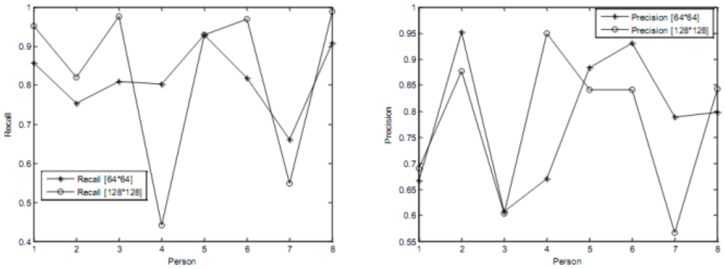
Figure 3: sample from 1 person with different sizes and different view angle in ETHZ data set

6. RGB

The red, green, and blue (RGB) color space is widely used throughout computer graphics. Red, green, and blue are three primary additive colors (individual components are added together to form a desired color) and are represented by a three-dimensional, Cartesian coordinate system. The indicated diagonal of the cube, with equal amounts of each primary component, represents various gray levels.

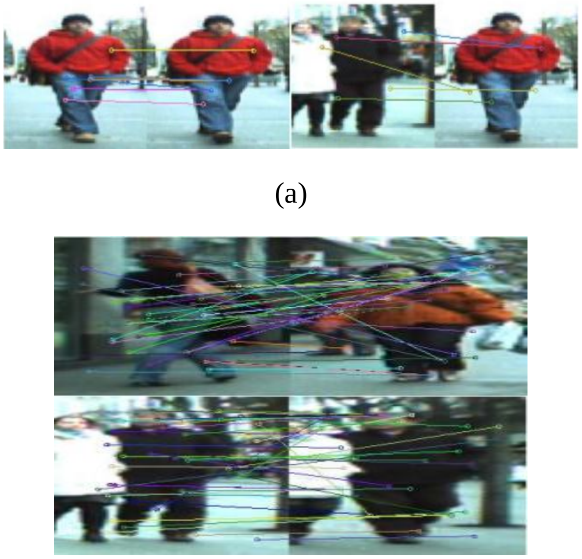
The results of human classification in RGB color space with SIFT method has been shown on different size of image in Table (1). From Table (1) it can be found that the performance of system improved when 128*128 dimension was used. Figure (4) shows evaluation measure for each person separately.

Recall



(a) **(b)**

Figure 4: The classification results using SIFT in RGB color space. a) Recall measure for each person in image with the size 32*32 and 128*128. b) Precision measure for each person in image with size 32*32 and 128*128.



(a)

(b)

Figure 5. Samples from obtain results from sift in RGB color space, a) images with 64*64 sizes b) images with 128*128 sizes

7. YCrCb

The RGB color space is the default color space for most available image formats. Any other color space can be obtained from a linear or non-linear transformation from RGB. The color space transformation is assumed to improve performance human detection with sift and surf and to provide robust parameters against varying illumination conditions. YCrCb is an encoded nonlinear RGB signal, commonly used by European television studios and for image compression work. Color is represented by luma (which is luminance, computed from nonlinear RGB [4], constructed as a weighted sum of the RGB values, and two color difference values Cr and Cb that are formed by subtracting luma from RGB red and blue components

$$Y = 0.299\ R + 0.587\ G + 0.114\ B$$

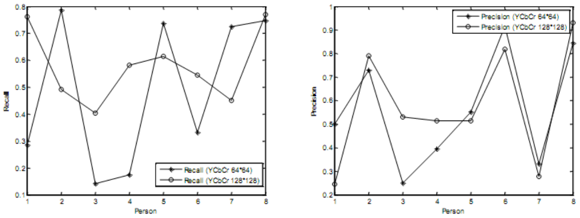
$$Cr = R - Y$$

$$Cb= B - Y$$

While YCbCris device dependent, it is intended for use under strictly defined conditions within closed systems. The Y component describes brightness, the other two values describe a color difference rather than a color, making the color space unintuitive. The results of humanclassification in YCbCr color space with SIFT method has been shown on different size of image in Table (2). From Table (2) it can be found that the performance of system improved when 128*128 dimension was used. Figure (6) shows evaluation measure for each person separately.

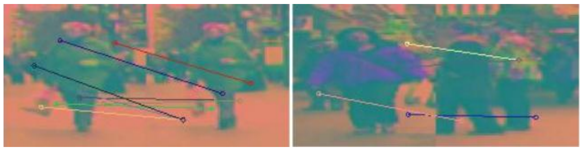
Table (2): The classification results using SIFT in YCbCr color space

Size(YCb Cr)	avg_precisi	avg_reca	F1
[64 64]	0.5653	0.4916	0.5259
[128 128]	0.5779	0.5776	0.5778

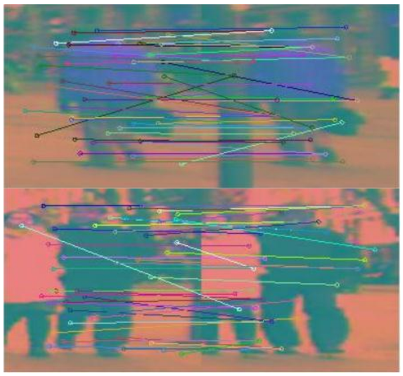


(a) (b)

Figure (6): The classification results using SIFT inYCbCr color space. a) Recall measure for each person in image with the size 32*32 and 128*128. b) Precision measure for each person in image with size 32*32 and 128*128.



(a)



(b)

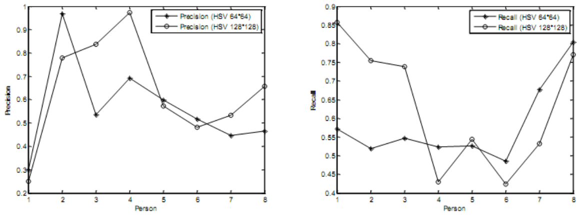
Figure 7.

8. HSV

HSV stands for Hue, Saturation, Value. These terms have the following meanings: Hue: The _true color_ attribute (red, green, blue, orange, yellow, and so on). Saturation: The amount by which the color as been diluted with white. The more white in the color, the lower the saturation. So a deep red has high saturation, and a light red (a pinkish color) has low saturation. Value: The degree of brightness: a well lit color has high intensity; a dark color has low intensity. This is a more intuitive method of describing colors, and as the intensity is independent of the color information, this is a very useful model for image processing. The results of human classification in HSV color space with SIFT method has been shown on different size of image in Table (3). From Table (3) it can be found that the performance of system improved when 128*128 dimension was used. Figure (8) shows evaluation measure for each personseparately.

Table (3): The classification results using SIFT in HSV color space

Size(gray)	avg_precisi	avg_rec ll	F1
[64 64]	0.7923	0.8414	0.8161
[128 128]	0.7843	0.8485	0.8152



(a) (b)

Figure (8): The classification results using SIFT in HSV color space. a) Recall measure for each person in image with the size 32*32 and 128*128. b) Precision measure for each person in image with size 32*32 and 128*128.

9. Grayscale

The results of human classification in grayscale mode with SIFT method has been shown on different size of image in Table (4). From Table (4) it can be found that the performance of system improved when 64*64 dimension was used. Figure (9) shows evaluation measure for each person separately.The results of human classification in grayscale mode with SURF method has been shown on different size of image in Table (5). From Table (5) it can be found that the performance of system improved when 128*128 dimension was used. Figure (10) shows evaluation measure for each person separately.

Table (4): The classification results using SIFT in grayscale mode

Size(HSV)	avg_precisi	avg_reca	F1
[64 64]	0.5658	0.5818	0.5737
[128 128]	0.6359	0.6313	0.6336

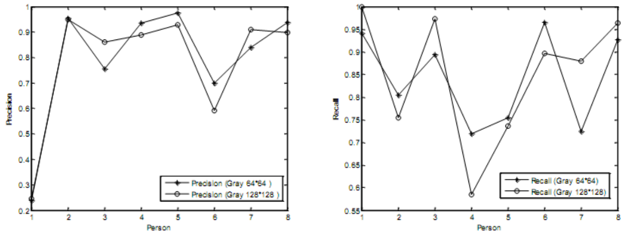


Figure (9): The classification results using SIFT in grayscale mode. a) Recall measure for each person in image with the size 32*32 and 128*128. b) Precision measure for each person in image with size 32*32 and 128*128.

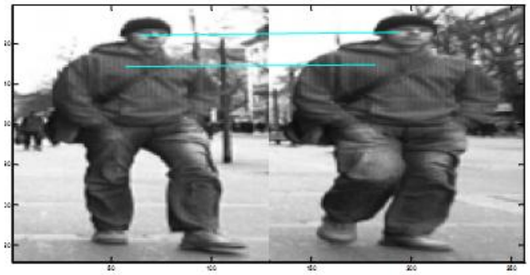


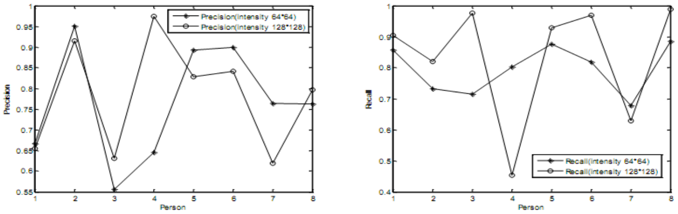
Figure (10): in a second experiment we used SURF ingrayscale mood for recognition of persons by means of descriptors based on a local features

The results of human classification in grayscale mode with SURF method has been shown on different size of image in Table (5). From Table (5) it can be found that the performance of system improved when 128*128 dimension was used. Figure (10) shows evaluation measure for each person separately.

Table (5): The classification results using SIFT in grayscale mode

Size(intensit	avg_precisi	avg_recal	F1
[64 64]	0.7670	0.7955	0.7810

[128 128]	0.7826	0.8339	0.8075



(a) (b)

Figure (11): The classification results using SURF in grayscale mode. a) Recall measure for each person in image with the size 32*32 and 128*128. b) Precision measure for each person in image with size 32*32 and 128*128.

10. Conclusions

In surveillance cameras, recognition techniques has been discussion as one of the most applications. To this end, in this paper, has been noted recognition of persons based on match Algorithms of local features such as SIFT and SURF in different color spaces like RGB, YCbCr and HSV also in grayscale mood. This experiment has been used just one image from every person for recognition of person in train collection and also with integrate of images sizingfor recognition.Obtained result demonstrator of good Performance Improvement for recognition according to proposed method.

11. Reference

[1] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In CVPR 2005, 2005.

[2] A. Mohan, C. Papageorgiou, and T. Poggio. Examplebased object detection in images by components. PAMI, 23(4):349–361, 2001.

[3] Y. Ke and R. Sukthankar. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In CVPR 2004, volume 2, pages 506–513, 2004.

[4] S. Belongie, J. Malik, and J. Puzicha. Matching Shapes. In ICCV 2001, volume 1, pages 454–461 vol.1, 2001.

[5] Q. Zhu, M.-C.Yeh, K.-T.Cheng, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In CVPR 2006, pages 1491–1498, 2006.

[6] W. Zhang, G. Zelinsky, and D. Samaras. Real-time accurate object detection using multiple resolutions. In ICCV, 2007.

[7] J. Begard, N. Allezard, and P. Sayd. Real-time human detection in urban scenes: Local descriptors and classifiers selection with adaboost-like algorithms. In CVPR Workshops, 2008.

[8] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In CVPR, 2007.

[9] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou. Discriminative local binary patterns for human detection in personal album.In CVPR 2008, pages 1–8, June 2008.

[10] Y.-T. Chen and C.-S. Chen. Fast human detection using a novel boosted cascading structure with Meta stages. Image Processing, IEEE Trans. on, 17(8):1452–1464, 2008.

[11] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesiancombination of edgelet part detectors. In ICCV, pages 90–97, 2005.

[12] B. Wu and R. Nevatia.Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection. In CVPR 2008, pages 1–8, June 2008.

[13] P. Dollar, B. Babenko, S. Belongie, P. Perona, and Z. Tu. Multiple Component Learning for Object Detection. In ECCV 2008, pages 211–224, 2008.

[14] S. Maji, A. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In CVPR, June 2008.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In CVPR 2006, pages 2169–2178, 2006.

[16] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In ECCV 2004, volume I, pages 69–81, 2004.

[17] V. Shet, J. Neuman, V. Ramesh, and L. Davis. Bilattice-based logical reasoning for human detection.In CVPR, 2007.

[18] P. Felzenszwalb, D. McAllester, and D. Ramanan.A discriminatively trained, multiscale, deformable part model. CVPR, pages 1–8, June 2008.

[19] D. Tran and D. Forsyth. Configuration estimates improve pedestrian finding. In NIPS 2007, pages 1529–1536. MIT Press, Cambridge, MA, 2008.

[20] Z. Lin and L. S. Davis. A pose-invariant descriptor for human detection and segmentation. In ECCV, 2008.

[21] David G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2:91–110, 2004.

[22] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), volume 2, pages 506–513, 2004.

[23] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In Ninth European Conference on Computer Vision, 2006.

[24] H. Bay, B. Fasel, and L. Van Gool. Interactive museum guide: Fast and robust recognition of museum objects. In Proc. Int. Workshop on Mobile Vision, 2006.

[25] W. Zhang and J. Kosecka. Image based localization in urban environments. In International Symposium on 3D Data Processing, Visualization and Transmission, pages 33–40, 2006.

[26]J. Beis and D. Lowe. Shape indexing using approximate nearest neighbor search in high dimensional spaces. In Proc. IEEE Conf. Comp. Vision Patt.Recog., pages 1000–1006, 1997.

[27] A. Ess, B. Leibe, and L. V. Gool. Depth and Appearance for Mobile Scene Analysis.In ICCV, 2007.