# Robust Face Detection for Indoor Surveillance using Multiple Scene Contexts

Shotaro Miwa, Takashi Hirai
Sensing Information Dept.
Mitsubishi Electric Corp.
Amagasaki City, Hyogo, Japan

Kazuhiko Sumi
Graduate School of Science and Engineering
Aoyama Gakuin University
Sagamihara City, Kanagawa, Japan

***Abstract*—This paper describes a robust face detection algorithm for indoor surveillance using multiple scene contexts. Faces captured by surveillance cameras are smaller and darker with motion-blurs and distortions. They are also downward looking and partially hidden because of high camera positions. Those degraded faces are difficult to be detected using just a single face detector. To solve the problem we utilize contextual information about faces of walking people in buildings. The contextual information we utilize is global camera view context, local face context, and geometric scene context. We built a probabilistic face detection framework combining a face detector with local face and global camera view contextual information on geometric scene constraints. In our face detection framework firstly we use a boosted fast face detector on the geometric scene constraint, which pick up candidate face patches of walking people in a very short time. Secondary among candidate face patches, we pick up reliable face patches by calculating a local face context, a conditional probability in surrounding regions, using a HOG feature-based outline detector, and also calculate global camera view context, possible face patch sizes dependent on detected locations from viewpoint information. Combining a fast boosted face detector with multiple scene contexts, while keeping computational efficiency of the original boosted face detector, we achieved a high face detection rate of 93.7 % with about 1,000 times lower false positive rate than using a single original face detector.**

***Keywords-contextual information, surveillance camera, face detection, haar-like features, HOG features***

## I. INTRODUCTION

This paper proposes a detection algorithm for robust and fast face detection for surveillance cameras in buildings. After the first robust and fast face detection algorithm was proposed by Viola and Jones [1], a variety of face detection algorithms have been proposed [2~4]. But in the case of surveillance cameras, faces have ambiguity with degradation as well as pose variations all at once. From distant and high-positioned surveillance cameras in buildings, walking people's faces are small, downward, dark, and partially-hidden. To detect such degraded faces without mistakenly detecting non-face patterns is impossible.

While traditional detection algorithms adopt object-centered approaches which use inner patterns of targets, recently there are other holistic detection approaches which utilize contextual information that is a relationship between a target object and its background in the whole image. Those approaches are promising ways to solve the problems of object detection in the real world that is a combination of ambiguity and complexity.

These context-based detection approaches propose general frameworks for object detection. They consider configuration of objects in the whole image as contextual information and use it for additional information for object detectors. Torralba et val. [5~7] utilize contextual information to find areas of focus of attentions and scales to detect targets. They calculate outputs of Gabor filters, and estimate possible locations and scales of objects in the image. They show its effectiveness in a variety of images, portrait images, pictures of indoors and outdoors. On the other hand, Hoiem et val.'s approach [8, 9] set focus on 3D contextual information not 2D like Torralba et val. They estimate the coarse orientation of large surfaces in outdoor images, and recover a 3D "contextual frame" which is effective to estimate objects' locations in the image. Hedau et val. [10, 11] proposed a 3D context model expanding Hoiem et val's algorithm [8, 9] in indoor scenes which are difficult to model because of many object clutters. They combine a parametric 3D box with surface labels of pixels, and estimate a box layout from a single image.

Our approach also uses contextual information, but has more detector-centered approach for the purpose of detecting a specific object in a short time. We design our framework considering computational efficiency as well as detection performance when detecting degraded faces in a real world such as surveillance cameras in buildings. In the case of surveillance cameras the problem of face detection is accompanying many false positives. To decrease false positives while keeping original face detector's computational efficiency, we utilize geometric constraint in the scene, local contextual information around face regions and global contextual information of a camera view in the whole image. Initially we estimate wall and floor regions using Hedau et val.'s box estimation algorithm [10, 11] and make a mask for detecting faces. After that we apply a single face detector [1] to images of surveillance cameras using the mask. Finally we combine the result of a single fast face detector with local and global contextual information, and we achieved both a high face detection rate and a low false positive rate keeping

computational efficiency of the fast face detector. We applied our framework to images captured by a surveillance camera and confirmed its validity.

This paper is organized as follows. In Section 2 we explain contextual information found in face detection in surveillance cameras. In Section 3 we explain our context-based face detection framework incorporating contextual information. In Section 4 we report our experimental results, and in Section 5 we conclude our framework.

## II. CONTEXTUAL INFORMATION OF FACES IN SURVEILLANCE CAMERAS

### A. Face Detection in Surveillance Cameras

In this paper we consider face detection for surveillance in hallways in buildings. Generally surveillance cameras are set on high positions to monitor people walking on the ground. From upper and distant cameras faces are downward and small. Besides that in a hallway of buildings we have fewer illuminations than outside. From the camera under such condition faces are dark and difficult-to-discriminate because of low contrast. Furthermore because cameras slow down shutter speeds under such condition, faces have motion blurs. Under mixed hard conditions above, discrimination of faces from non-faces is very difficult and true detections of faces are accompanied by the detection of many false positives.

To solve those face detection issues of surveillance cameras in buildings, we aim to build a computational efficient robust face detection algorithm which detects faces with variety of poses in degraded images in a short time. Of course traditional object centered approaches make good use of internal face patterns. But for the degraded face images of surveillance cameras, it is a difficult task because we should detect more variations of faces with less information. So we've come up with an idea to combine face detectors with additional environmental information, that is geometric constraints in the scene, local contextual information, around face regions, and global contextual information about a viewpoint.

### B. Contextual Information

We analyzed a scenario in people walking in hallways, and found three types of contextual information available for face detection. These are 1) geometric constraints of the scene, 2) local contextual information around face regions, and 3) global contextual information about a viewpoint.

#### 1) Geometric Constraints in the scene

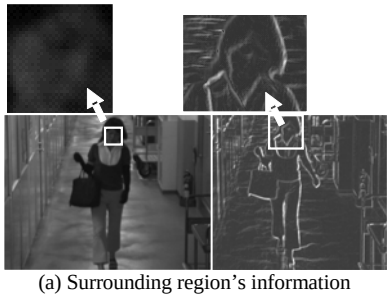
The hallway scenes are composed of left wall, right wall, and floor. Faces should be detected from people walking on the floor

#### 2) Local contextual information

Given only a degraded face-patch (Fig. 1(a)), a typical example of surveillance cameras, it is difficult to discriminate this from a non-face-patch example (Fig. 1(b)). Furthermore while people are walking in a hallway, they change their head poses, and those variations of internal face patterns also make it more difficult to discriminate them from non-face patterns.

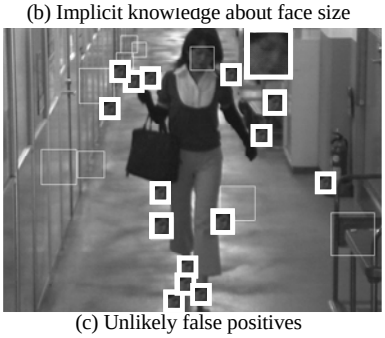
(a) Face patch (b) Non-face patch

Figure 1. Discrimination of image patches.



(a) Surrounding region's information

(b) Implicit knowledge about face size



(c) Unlikely false positives

Figure 2. Local and global contextual information.

In these case of surveillance cameras outlines of walking people give us valuable information. Even under hard conditions we can easily find a typical outline of humans around the face region, e.g. contours of heads and shoulders. Those outlines are independently more stable than their internal face patterns (Fig. 2(a)).

#### 3) Global contextual information (Fig. 2)

When we see faces of people walking in a hallway, we implicitly use knowledge for possible locations and scales of faces. Closer to a camera, sizes of faces become bigger (Fig. 2(b)). Look at false positives of Fig. 2(c) on which real face images of detected scales are overlaid at detected locations. To human eyes those on a floor are apparently too small, and those

near a ceiling are too big, and we feel those are very strange and unlikely. We implicitly discard those unlikely false positives using global contextual information about a viewpoint of the image.

## III. CONTEXT-BASED FACE DETECTION

### A. Probabilistic Framework

In this section we introduce the probabilistic detection framework combining a face detector with geometric constraints, local and global contextual information.

Firstly object detection is formalized using a feature vector $\boldsymbol{v}$ as follows:

$$P(O \mid v) = \frac{P(v \mid O)}{P(v)} P(O) \tag{1}$$

Here the function $P(\boldsymbol{O}|\boldsymbol{v})$ is the conditional probability density function of the presence of the object $\boldsymbol{O}$ given features $\boldsymbol{v}$. The notation $\boldsymbol{O}$ is the object attributes: $\boldsymbol{O}=\{o, x, s\}$ where $o$ is the label of the object, $x$ is the location of the object in image coordinates, and $s$ is the size of the object.

In object-centered approaches [12, 13] the following local model is used:

$$P(O \mid v) \cong P(O \mid v_D) = \frac{P(v_D \mid O)}{P(v_D)} P(O) \tag{2}$$

where $\boldsymbol{v}_D$ is a local feature vector in a region defined by a bounding box defined by $x$ and $s$. This framework is valid provided that $P(\boldsymbol{v}_D|\boldsymbol{O})$ has local peaks. But in the real world we can't expect such ideal situations.

In our framework we formalize the detection framework as follows:

$$P(O \mid v) = P(O \mid v_{Dc}, v_{Sc}, v_{Lc}, v_{Gc}) \tag{3}$$

where $\boldsymbol{v}_{Sc}$ is geometric constraint in the scene, $\boldsymbol{v}_{Lc}$ is local contextual features around a bounding box defined by x and s and $\boldsymbol{v}_{Gc}$ is global contextual features around the same bounding box.

We can assume that $\boldsymbol{v}_{Sc}$ and $\boldsymbol{v}_{Gc}$ are independent, then object likelihood $P(\boldsymbol{O}|\boldsymbol{v})$ can be written as:

$$\begin{aligned} P(O \mid v) &= P(O \mid v_{Dc}, v_{Sc}, v_{Lc}, v_{Gc}) \\ &= P(O \mid v_{Dc}, v_{Lc}) P(O \mid v_{Gc}) P(O \mid v_{Sc}) \end{aligned} \tag{4}$$

Using Bayes' rule, Eq. (4) is written as:

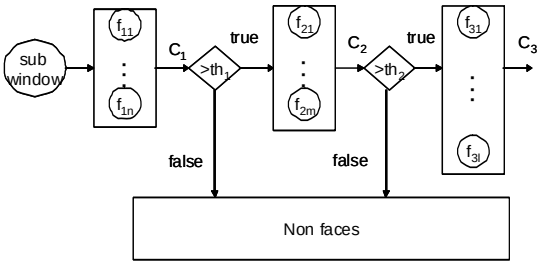
$$\begin{aligned} P(O \mid v) &= P(O \mid v_{Dc}, v_{Sc}, v_{Lc}, v_{Gc}) \\ &= \frac{P(v_{Lc} \mid O, v_{Dc})}{P(v_{Lc} \mid v_{Dc})} P(O \mid v_{Dc}) P(O \mid v_{Gc}) P(O \mid v_{Sc}) \end{aligned} \tag{5}$$

In the next section we explain each probability in detail.

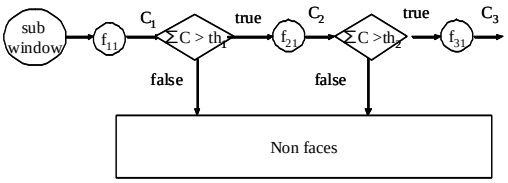
### B. Face Detector

In Eq. (5) $P(\boldsymbol{O}|\boldsymbol{v}_{Dc})$ is the probability calculated by object detector itself. We use a boosted Haar-like feature-based face detector (1) for calculating face probability here. A Haar-like feature-based face detector has a great advantage in computational efficiency. It runs very fast, discarding many distracters, and picking up only small amount of possible face patches from a vast of almost all non-face patches in the image. The detector runs enough fast even when we try to pick up all possible patches including some false positives as well as true positive, e.g. totally 20~30 from more than a few millions of candidate regions e.g. in a QVGA image.

The face detector we used has an improved structure from an original cascaded face detector (1). In Fig.3 we show an original cascaded face detector (Fig. 3(a)) and our faster cumulative detector (Fig. 3(b)). An original cascaded detector has a group of features (Fig. 3(a)) at each stage, but our cumulative face detector has just a series of features (Fig. 3(b)). In a cumulative face detector each feature outputs a confidence score and decides a patch as a face, or not. If the patch is decided as a candidate of faces, a confidence score is summed in the next stage and a cumulative score is used for the next decision. To calculate face probability using this cumulative detector we normalize the number of filters until which each patch passes through.




(a) Original cascaded face detector




(b) Cumulative face detector

Figure 3. Face detector.

### C. Geometric Constraints in the Scene

The scene of hallways has a specific 3D structure. It is composed of mainly three planes, left wall plane, right wall plane, and floor plane. We estimate these plane configuration using Hedau et val.'s algorithm [10, 11] and use the estimated plane information as a geometric constraint in the scene. People normally walk on the floor, not on the wall. So we mask the wall regions and search faces from only candidate regions of standing people on the floor.

### D. Local Contextual Information

As local contextual information we use a HOG feature-based outline detector for calculating outline probability in surrounding regions which are three times as large as face regions.

A HOG feature-based detector shows great performance for classification [14, 15], but that runs much slower than a boosted face detector. So processing a whole image by only a HOG feature-based detector isn't practical in a view of computational efficiency.

But applying a Haar-like feature-based face detector and a HOG feature-based detector sequentially, we can take advantage of a HOG's discriminative power with Haar-like features' computational efficiency. In a sequential processing after restricting the candidate regions by a primary face detector, we have only to apply a HOG feature-based detector to small number of candidate regions with least additional computation.

In this paper as a HOG feature-based detector we make an outline detector by SVM. To make a SVM-based outline detector using HOG features we employ Dalal and Trigg's algorithm [14]. To calculate outline probability using a SVM-based detector we employ Chateau's approach [16].

### E. Global Contextual Information

In Eq. (5) $P(\boldsymbol{O}|\boldsymbol{v}_{Dc})$ is the probability of possible locations and scales of faces under the camera. In our model we think about one stationary camera in a hallway, so we model this probability in a simple way. In advance we know the internal camera parameters and the height of camera from the ground. Combining a vanishing point (Fig. 4) calculated during geometric constraint calculation [10, 11, 17, 18] with known camera information, we assumed a certain face size in image coordinates. Dividing a whole image into four regions, we calculate possible variations of face sizes in each region (Fig. 4).

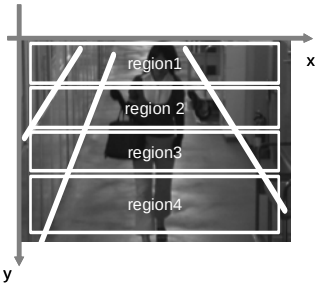



Figure 4. Scale probability.

### F. Classification

Fig. 5 describes our classification method. Initially we mask the image with geometric constraint mask. Using this masked image a primary face detector discards most non-face patches on possible locations in an image, and picks up a small amount of candidate patches with face probabilities.

The second global context stage calculates scale probability for those patches. A patch has a rectangular shape, so we model a scale (=width) parameter as a function of y in an image coordinate. We calculate an average size of faces, sm(y), and set values of the minimum size smin(y) = 0.7*sm(y), and the maximum size smax(y) = 1.4*sm(y), of faces. Only patches which have values of width between the minimum and the maximum size pass through this stage.

The final local context stage calculates "outline" probability using three times larger area including the patch. Finally we multiply the "face" probability by the "outline" probability, and the final joint probability is classified by a given threshold.

## IV. Experimental Results

We evaluate our framework using a dataset from surveillance cameras in our laboratory. Cameras are set on the ceiling with about 3.5 meters height and a downward angle of about 20 degrees. A total number of images (VGA) are 1074 of 30 persons.

In this dataset people are not looking straight ahead while walking, and they are mostly looking down and aside. From an upper camera it brings a hard pose issue to the face detection that captured faces have strong downward angles. It also brings a hard occlusion issue that faces of people with caps and helmets are partially hidden from an upper camera.

Fig. 6 shows the result of estimated geometric labels of the scene. Fig. 7 shows the geometric regions of left wall and right wall. While walking in hallways, off course people walk on the floor, but they sometimes walk near walls, so we use the half of the wall regions as a mask.

In the exact case of hallway dataset Fig. 8 shows the result of each stage in the classification diagram. Fig. 8(a) shows the detection result after applying only a face detector. It shows that the face detector detects a degraded dark and downward face with many false positives. Fig. 8(b) shows the detection result of combining the face detector with global contextual information.

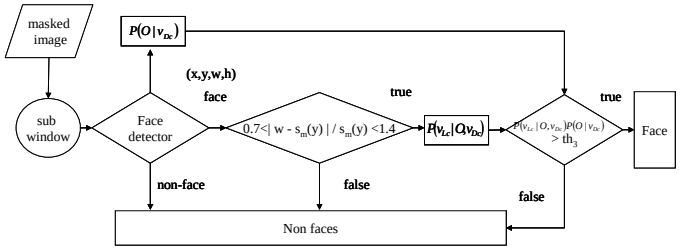



Figure 5. Classification diagram

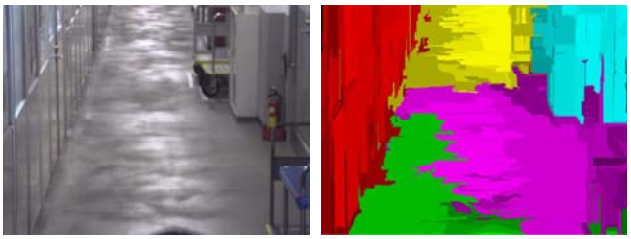


Figure 6. Estimated geometric labels

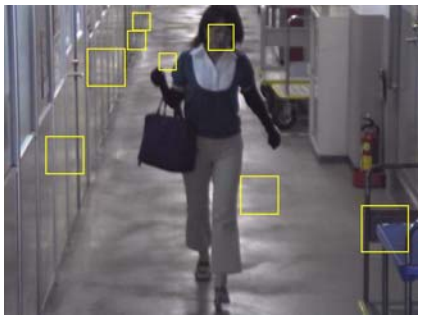
Figure 7. Geometric constraint mask(left: wall regions, right mask regions)

It shows that the global contextual information is not so strong, but the numbers of false positives are decreased to some extent. Fig. 8(c) shows the final result after combining the face detector with global and local contextual information. It shows that we could discard all false positives as well as successfully detecting the face.

Fig. 9 shows a ROC curve of a comparison of the context-based face detection framework with a single face detector method. This shows that at the same detection rate of 93.7% using global camera contextual information we achieved a 0.005 % false positive rate of the case of only a face detector, and using both global and local contextual information we achieved a 1,000 times as low as false positive rate than the case of only one a face detector. About computational efficiency, applying face detector to the whole image takes 217 msec. on Xeon 2.6GHz. But applying the geometric constraint mask to the image, we can make the face detector run by 10 % faster than without the mask. Adding just 20 msec. for the additional computation of contextual information to the face detection, our proposed framework totally runs as fast as just a single face detector.

Even though we achieved a high detection rate and a low false positive rate, we have some faces not detected. Problems of this exist in geometrical configurations of face regions and outlines regions, and also in variations of outline shapes. There are two problems about geometrical configurations. One is a scale ratio between face regions and outline regions. Because outlines include hairs and caps, there are more scale variations than we assumed. The other is a relative location between heads and shoulders. While walking with changing head poses, geometrical relationships between heads and shoulders have more variations than we assumed. The last outline shape variation problem is typical when people is bending their heads forward because it completely changes the shapes of outlines themselves as well as the relative locations between faces and outlines.

(a) Result after face detection with geometrical constraint



(b) Result after global context

(c) Result after global and local context:

Figure 8. Face detection results.

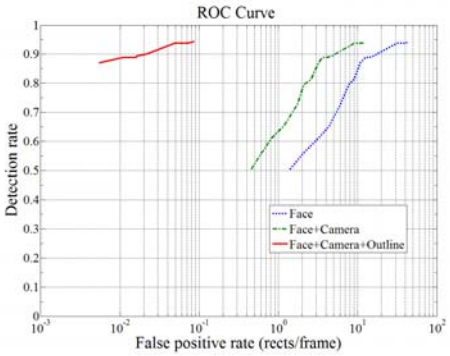



Figure 9. Comparison of face detection performance

To solve those problems by searching more scales and locations of HOG features, we expect to improve detection performance to some extent with expense of small amount of

additional computational costs. But just keep increasing the search locations and scales bring more false positives, and it can't solve the problem of outline shape variations. For much further improvement to solve all those problems we need to analyze and model the geometrical configuration between face regions and outline regions. Based on current analysis we're thinking about using a variation of outline detectors.

## V. CONCLUSION

In this paper we present a novel context-based face detection framework which combines a face detector with geometric constraint in the scene, local contextual information around face regions and global contextual information about a viewpoint.

After initial masking using geometric constraint, combining sequentially a boosted Haar-like feature-based face detector for calculating a prior probability and a HOG feature-based outline detector for calculating a following conditional probability, we pick up and combine each method's advantage at the same time, that is the discriminative power of HOG features, and the fast selecting focus-of-attentions power of Haar-like features. The result of our evaluation using the surveillance camera in the hallway shows that with the same computational cost as a single face detector we achieved a face detection rate of 93.7 % with 0.048 false positive patches per frame that is 1,000 times as low false positive rate as the case of using a single face detector.

Finally this framework can also be applied to detections of other objects in other scenes. Future research includes improvement of scene modeling and face detection algorithm in surveillance cameras considering geometrical configuration between face regions and outline regions, and also the more general integrated framework of Haar-like features and HOG features for other object detections.

## REFERENCES


[1] P. Viola and M. Jones: "Rapid object detection using rapid object detection using a boosted cascade of simple features," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.511-518 (2001)

[2] M. H. Yang, D.J. Kriegman, and N. Ahuja: "Detecting faces in images: a survey," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 1, Jan. pp. 34-58 (2002)

[3] C. Huang, H. Ai, Y. Li, and S. Lao: "High-Performance rotation invariant multiview face detection," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 4, pp.671-686 (2007)

[4] H. Shneiderman: "Learning a restricted bayesian network for object detection," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.639-646 (2004)

[5] K. Murphy, A. Torralba, and W. Freeman: "Using the forest to see the trees: a graphical model relating features, objects, and scenes," Proc. of Advances in Neural Information Processing Systems (2003)

[6] A. Torralba: "Contextual priming for object detection", Proc. Int'l Journal of Computer Vision, Vol. 53, No.2 (2003)

[7] A. Torralba, K.P. Murphy, W.T Freeman, and M.A. Rubin: "Context-based vision system for place and object recognition", Proc. Int'l Conf. on Computer Vision, pp.273-280 (2003)

[8] D. Hoiem, A. A. Efros, and M. Hebert: "Geometric context from a single image", Proc. Int'l Conf. on Compuer Vision, pp.654-661 (2005)

[9] D. Hoiem, A. A. Efros, and M. Hebert: "Putting objects in perspective," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.2137-2144 (2006)

[10] V. Hedau, D. Hoiem, D. Forsyth, "Recovering the spatial layout of cluttered rooms," · Proc IEEE Int'l Conf. on Computer Vision, pp.1849-1856(2009).

[11] V. Hedau, D. Hoiem, D. Forsyth, "Thinking inside the box: using appearance models and context based on room geometry," · ECCV 2010, pp. 224-237(2010).

[12] B. Moghaddam and A. Pentland: "Probabilistic visual learning for object representation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp.696-710 (1997)

[13] B. Schiele and J. L. Crowley: "Recognition without correspondence using multidimensional receptive filed histograms," Int'l Journal of Computer Vision, pp.31-50 (2000)

[14] N. Dalal, B. Triggs: "Histograms of Oriented Gradients for Human Detection," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.886-893 (2005)

[15] Q. Zhu, S. Avidan, M.C. Yeh, and K.T. Cheng: "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.1491-1498 (2006)

[16] T. Chateau, V. Gay-Belille, F. Chausse, and J.T. Lapreste: "Real-time Tracking with Classifiers," Proc. of Int'l Workshop on Dynamical Vision in conjunction with European Conference on Computer Vision (2006)

[17] J. Kosecka and W. Zhang: "Video Compass," In Proc. European Conference on Computer Vision, pp.657-673 (2002)

[18] W. Zhang and J. Kosecka: "Efficient Detection of Vanishing Points," Proc. IEEE Conf. on Robotics and Automation (2002)