Marker Detection in Aerial Images

Thesis by

Yazeed Alharbi

In Partial Fulfillment of the Requirements

For the Degree of

Masters of Science

King Abdullah University of Science and Technology

Thuwal, Kingdom of Saudi Arabia

March, 2017
The thesis of Yazeed Alharbi is approved by the examination committee

Committee Chairperson: Peter Wonka
Committee Members: Wolfgang Heidrich, Bernard Ghanem
ABSTRACT

Marker Detection in Aerial Images
Yazeed Alharbi

The problem that the thesis is trying to solve is the detection of small markers in high-resolution aerial images. Given a high-resolution image, the goal is to return the pixel coordinates corresponding to the center of the marker in the image. The marker has the shape of two triangles sharing a vertex in the middle, and it occupies no more than 0.01% of the image size.

An improvement on the Histogram of Oriented Gradients (HOG) is proposed, eliminating the majority of baseline HOG false positives for marker detection. The improvement is guided by the observation that standard HOG description struggles to separate markers from negatives patches containing an X shape. The proposed method alters intensities with the aim of altering gradients. The intensity-dependent gradient alteration leads to more separation between filled and unfilled shapes.

The improvement is used in a two-stage algorithm to achieve high recall and high precision in detection of markers in aerial images. In the first stage, two classifiers are used: one to quickly eliminate most of the uninteresting parts of the image, and one to carefully select the marker among the remaining interesting regions. Interesting regions are selected by scanning the image with a fast classifier trained on the HOG features of markers in all rotations and scales. The next classifier is more precise and uses our method to eliminate the majority of the false positives of standard HOG. In the second stage, detected markers are tracked forward and backward in time. Tracking is needed to detect extremely blurred or distorted markers that are missed by the previous stage.
The algorithm achieves 94% recall with minimal user guidance. An average of 30 guesses are given per image; the user verifies for each whether it is a marker or not. The brute force approach would return 100,000 guesses per image.
ACKNOWLEDGEMENTS

I extend my gratitude to professor Peter Wonka and doctor Neil Smith for dedicating the time to assist with tackling the problem.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examination Committee Page</td>
<td>2</td>
</tr>
<tr>
<td>Copyright</td>
<td>3</td>
</tr>
<tr>
<td>Abstract</td>
<td>4</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>6</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>9</td>
</tr>
<tr>
<td>List of Symbols</td>
<td>10</td>
</tr>
<tr>
<td>List of Figures</td>
<td>11</td>
</tr>
<tr>
<td>List of Tables</td>
<td>12</td>
</tr>
<tr>
<td><strong>1 Introduction</strong></td>
<td>13</td>
</tr>
<tr>
<td>1.1 Contributions</td>
<td>15</td>
</tr>
<tr>
<td><strong>2 Related Work</strong></td>
<td>16</td>
</tr>
<tr>
<td>2.1 Object Detection</td>
<td>17</td>
</tr>
<tr>
<td>2.2 Marker Detection</td>
<td>17</td>
</tr>
<tr>
<td>2.3 Uniqueness of Thesis Topic</td>
<td>18</td>
</tr>
<tr>
<td><strong>3 Baseline Methods</strong></td>
<td>19</td>
</tr>
<tr>
<td>3.1 Corner Detection</td>
<td>19</td>
</tr>
<tr>
<td>3.2 Template Matching</td>
<td>20</td>
</tr>
<tr>
<td>3.3 Cascade Classification</td>
<td>21</td>
</tr>
<tr>
<td>3.4 SVM-HOG Classification</td>
<td>22</td>
</tr>
<tr>
<td><strong>4 Methodology</strong></td>
<td>26</td>
</tr>
<tr>
<td>4.1 Sample selection and preprocessing</td>
<td>27</td>
</tr>
<tr>
<td>4.2 Fixing Rotation and Scale</td>
<td>28</td>
</tr>
<tr>
<td>4.3 Intensity-Aware HOG Description</td>
<td>29</td>
</tr>
</tbody>
</table>
LIST OF SYMBOLS
LIST OF FIGURES

1.1 Markers occupy a tiny portion of image size ............................................. 14

3.1 Examples showing the variance in scale, rotation, illumination, and material of the markers used in this thesis ......................................................... 19

3.2 A plot showing the percentage of detected markers, for each of the 332 validation images. Clearly, markers are missed more starting around image 180 ................................................................. 21

3.3 A scatter plot showing marker side length per image .................................. 22

3.4 An example of false positives with SVM trained on standard HOG. Even when trained on a single scale and rotation, ‘+’ shapes continue to be mistaken for markers ........................................... 23

4.1 Standard HOG description struggles to separate patches with similar intensities but different gradients ................................................................. 26

4.2 During window sliding, these images are used to transform windows to classifier rotation and scale. Right: The middle 9x9 pixels of patches are rotated first to match this image. Left: After finding rotation, windows are scaled up and down to match this image .................................. 28

4.3 Brightening or darkening the same regions in filled and unfilled shapes introduces intensity-dependent edges ...................................................... 30

4.4 An example of using a mask to alter the intensities of a marker .................. 31

4.5 An example of using a mask to alter the intensities of a background patch ................................................................. 32

4.6 The 24 patterns are attempted and evaluated at every 4x4 of the mask ................ 34

5.1 A histogram showing marker side lengths .................................................. 41

5.2 A plot comparing the number of false positives per each level of recall between intensity-aware HOG and standard HOG ....................................... 42

5.3 This is the same as Figure 5.2 but zoomed on recall rates between 91% to 96% ................................................................................. 43
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Corner detection recall rates for different numbers of top corners to examine</td>
<td>20</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The thesis aims to solve a practical problem that was encountered by scientists at KAUST. Given a sequence of images, and a marker that appears in the images, the goal is to find the centers of the markers in each image.

Detecting the center of the marker in an image is essential to correct 3D reconstruction from images. The markers serve as ground control points in the mapping process. They were manually placed in the field, and their exact 3D position is known. In the reconstruction step, locating the marker center in an image allows for an accurate 2D (marker center in the image) to 3D (marker position in the field) mapping. This accurate mapping is used to correct the mapping of other pixels in the image.

Traditionally, the user had to look at high-resolution images and manually find the center of the marker in each image. Most images do not contain a marker, and if there are markers they do not usually exceed four markers per image. This was tedious and time-consuming, especially given the disparity between marker size and image size. Since the beginning, two factors were crucial in the desirable algorithm: high recall and high precision. Since every missed marker is costly, recall should be as high as possible. Since every wrong guess (false positive) leads to more user interaction, precision should be as high as possible.
This is not a new problem in computer vision. In fact, there are existing approaches to solving it in object detection and marker detection. However, the details of the thesis make the problem different from other problems explored previously in literature. First, the marker occupies a tiny portion of the image, as shown in Figure 1.1. Second, the largest marker in the dataset is more than ten times as large as the smallest marker. Additionally, marker rotation is not fixed. This wide variety of scale, rotation, and even distortion conditions across markers poses a great challenge to meeting the two requirements of high recall and high precision. The significance of the small size of the marker is that the search space is very large. It is immediately clear that step sizes must be small in any sliding window approach of detecting the marker. This is necessary because the smallest marker is only ten pixels wide.

Precision in the context of this thesis is directly related to user interaction. The exact percentage is of little significance. However, 100% precision means no user interaction, on the other hand, 1000 false positives per image are too many for a user to verify. The more relevant consequence of precision is the number of false positives per image. As shown in this thesis, the challenges mentioned above could not be overcome by previous methods in object and marker detection. The Viola-Jones algorithm[1], one of the most successful object detection algorithms, is shown
to be biased in missing markers based on size. Likewise, many marker detection algorithms are insufficient because of the unreliability of low-level preprocessing and the large variance in scale and rotation.

1.1 Contributions

In this thesis, an algorithm is proposed for automatic marker detection in high-resolution aerial images with high recall and high precision. A novel modification on the Histograms of Oriented Gradients (HOG) descriptor is introduced which eliminates the majority of the false positives of traditional HOG. This modification is unprecedented in literature, both in goal and results. Several papers explore adding color information to the HOG, but adding raw intensity information has not been previously explored.

Additionally, a generalization of the HOG descriptor modification is introduced in this thesis, rendering it applicable for tasks other than marker detection.
Chapter 2

Related Work

There are several differences between the work presented in this thesis and related papers in literature. One difference is the ratio of marker size to image size. In this thesis the marker represents no more than 0.01% of the whole image. Many related papers tackle markers that represent a significant portion of the image. In a comparative study of marker detection systems\[3\], smallest detectable markers in mainstream methods were around 14 x 14 pixels in 320 x 240 pixel images. To put this into perspective, many markers examined in this thesis are around the same size but they appear in images at least 3600 x 5400 pixel images. A second difference is the required recall. Here, recall is expected to be around 95%, regardless of illumination changes or occlusion. On the other hand, many marker detection and tracking in augmented reality accept lower recall rates, since it is acceptable to miss the marker in some frames. In some cases, marker detection methods can lose over 14% of detections if illumination changes occur\[4\]. In this thesis, each marker usually appears in two or three frames, changing position by an average of over 400 pixels. One of the most significant differences, however, is that markers in previous works are either rotationally invariant or expected to be in an upright position. In this thesis, we examine markers that appear in arbitrary rotations. The implication is that while traditional methods can afford to use template matching to detect markers quickly, it is not a suitable choice for the problem of this thesis.
2.1 Object Detection

Object detectors aim to locate a specific object in an image. Applications of object detection include face detection, car detection, and pedestrian detection. The Viola-Jones cascade detector remains among the best visual object detectors to date. The Viola-Jones algorithm achieves frontal upright face detection in real time through three contributions. The first contribution is constant time summation of pixel intensities within an image using what is called an integral image. The second contribution is generation of strong classifiers using small numbers of features, through feature selection. The third contribution is rapid detection of objects by consecutively increasing the complexity of classifiers in a cascade such that negatives are eliminated quickly at early stages of the cascade. The limitation of the Viola-Jones algorithm, for our task, is the inability to handle large variations in marker size. While the algorithm meets the requirement on small markers, large markers are missed significantly more often. This is essential for the topic of the thesis since markers do occur in arbitrary rotations, scales, and lighting conditions.

2.2 Marker Detection

Marker detection provides assistance in two popular areas: unmanned aerial vehicles and augmented reality. In marker detection, an object is specifically designed and placed in the environment to be captured by images. The design of the marker depends on the problem at hand, and many research papers in marker detection propose ad-hoc markers. In literature, the most common approach is to geometrically construct the marker region after detecting low-level features. For example, in the ARTag system\cite{5}, edges are detected and grouped into quadrilateral contours. The inside of the contours is compared with a dictionary of valid tags to confirm that the quad corresponds to a marker. The tags are 6x6 squares that are colored black or
white, and they determine the marker.

However, similar approaches cannot be used in the topic of this thesis. Due to the small size of the marker in an image, it is often affected by blur and distortion that prevent standard low-level detectors from working. Corner detection has been shown to miss about 30% of markers even when top 100,000 corners are examined. Furthermore, even if the quad around the marker has been constructed, the problem of identifying whether it is a marker remains. In the ARTag system, the marker is expected to be upright. Therefore to verify if a quad is a marker, the quad is divided into 6x6 squares and the pattern of square intensity is compared with the dictionary. In our problem, the marker appears in different rotations and it is not possible to verify by simple intensity comparisons. Another popular marker detection system, by the name of ARToolKit\cite{6} is highly dependent on a global threshold to detect markers. Therefore its performance is directly dependent on lighting condition\cite{7}.

2.3 Uniqueness of Thesis Topic

The problem presented in this thesis is unique from similar topics in literature because of the small ratio between the marker and the image sizes, the large variance in marker appearance, and the required pixel distance from ground truth. First, markers are expected to occupy less than 0.01% of total image size. This directly affects the search space and the allowable step size. Second, there is large variance between markers. There is variance in rotation, lighting conditions, scale, and even material (marker can be spray-painted or printed on a paper). Third, it is essential to determine the marker center with accuracy. In fact, a threshold of ten pixels from ground truth was required for a detected marker to be accepted.
Chapter 3

Baseline Methods

Before reaching the contributions of this thesis, several methods were used in attempt to solve the problem. Earlier attempts include low-level preprocessing, such as edge detection, corner detection, and interest point extraction. After that, methods such as template matching and ranking and thresholding were attempted. In the final stages, the Viola-Jones cascade detector was attempted along with SVM classification. This thesis proposes an algorithm that avoids the disadvantages of baseline methods.

![Figure 3.1: Examples showing the variance in scale, rotation, illumination, and material of the markers used in this thesis.](image)

### 3.1 Corner Detection

Corner detection was an intuitive first attempt to reduce the search space, since the goal is to detect the center of the marker, which is the corner between two triangles. It is very fast, needing only a few seconds per image. However, it was determined that corner detection is not a suitable pre-processing step. Due to distortion, the center of the marker is hard to pinpoint in some cases, even for a human. This resulted in poor recall rates, as seen in Table 3.1. In fact, the outside corners of the marker were detected more than the center. One could attempt to estimate the center based on
<table>
<thead>
<tr>
<th>#Top Corners</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0%</td>
</tr>
<tr>
<td>100</td>
<td>$\frac{10}{407} = 2%$</td>
</tr>
<tr>
<td>1000</td>
<td>$\frac{70}{407} = 17%$</td>
</tr>
<tr>
<td>10000</td>
<td>$\frac{171}{407} = 42%$</td>
</tr>
<tr>
<td>50000</td>
<td>$\frac{294}{407} = 72%$</td>
</tr>
<tr>
<td>100000</td>
<td>$\frac{295}{407} = 72%$</td>
</tr>
</tbody>
</table>

Table 3.1: Corner detection recall rates for different numbers of top corners to examine.

detecting the outside four corners of the marker. Yet, often only a subset of outside marker corners are detected. So the preprocessing step ends up being inefficient, since every proposed corner must be examined in each direction to guess the correct center. Considering the variance in rotation and scale, small steps would have to be taken, which negates the reduction in search space.

### 3.2 Template Matching

The problem of distortion is also a reason why template matching fails to satisfy the precision requirement. In template matching, a window is slid over the image, computing a distance between the current window and some given template. After sliding through the whole image, the window with the smallest distance is considered the match. A major issue with template matching for marker detection is its inability to handle large variance of the object. Consequently, the image has to be examined multiple times with different scales, rotations, and transformations. Even for the template with the correct scale and rotation, multiple top-ranked proposals would have to be considered as markers to ensure desired recall.
3.3 Cascade Classification

One of the early attempts was to use the Viola-Jones algorithm that was efficient for real-time face detection. This algorithm includes a feature selection step, followed by training a cascade of increasingly more complicated classifiers. The purpose of the cascade is to eliminate most negatives earlier, where classifiers require only a few features. Throughout most of the early stages, cascade classification exceeded the performance of all alternative approaches. Furthermore, supporting evidence was found for the claim that cascade classification is faster than SVM classification. During experiments, when recall and precision were fixed, cascade classification was always faster than SVM classification. Nonetheless, cascade classification suffered from two major drawbacks. The first drawback is the difficulty of tuning and explaining performance, since features were automatically selected. The second drawback is systematically missing more of the large markers. The best trained cascade classifiers have at least 10% less recall on images where marker size was large. Since marker size is variable, it is unacceptable to miss markers based on size. Thus, the focus of the thesis shifted towards improving either SVM classification or HOG description.

![Cascade Marker Detections per Image](image)

Figure 3.2: A plot showing the percentage of detected markers, for each of the 332 validation images. Clearly, markers are missed more starting around image 180.
3.4 SVM-HOG Classification

Since the beginning of the project, the HOG descriptor appeared to be the ideal descriptor to capture markers. This is because of the marker is divided into cells, and it can be defined using the gradients in each cell. In the center cells, two gradients with high magnitudes and perpendicular directions are expected. This corresponds to the center of the marker. In outer cells, there could be either one gradient with high magnitude or none. This corresponds to one of the marker edges. Although there are four corners at the further ends of the two triangles, they were unreliable. This unreliability stems from the fact that markers are often skewed, leading to some corners being outside of the window. Additionally, even after adding a step to determine possible scale and rotation of the marker, outside corners were still not reliably present in the window. Nonetheless, the HOG alone seemed to be able to almost perfectly describe the marker. An SVM classifier trained with HOG features can overcome the previous problems. It is less sensitive to center location, even if the exact corner is blurred. Furthermore, it proposes only around 2000 windows.
Figure 3.4: An example of false positives with SVM trained on standard HOG. Even when trained on a single scale and rotation, '+' shapes continue to be mistaken for markers per images. More importantly, and unlike cascade classification, missed markers were scattered across all images with no apparent pattern. This was considered an improvement since cascade classification consistently missed larger markers.

While this classifier meets the recall requirement, it is far from meeting the precision requirement. A thousand proposals are too many for a user to manually verify. Therefore, the next natural step is to see the limit, if any, to SVM classification with HOG features. By examining the hard negatives of the best trained HOG classifier, it becomes obvious that the HOG features have difficulty distinguishing two patterns from the marker. The first pattern of false positives is circular objects. The second pattern is patches that resemble a '+' shape. Though intensity differences are used by the HOG to find gradients, the raw intensities within each uniform region have little effect on the descriptor. Thus, the HOG descriptor does not sufficiently describe the difference between a filled marker and an '+' shape. This was supported by the fact that the classifier has much lower precision in street images because of the presence...
of ‘+’ shapes in the tiles of walking areas. The distinction between the marker and these two patterns is crucial to lowering proposals from thousands to hundreds. The reason is that there are many images with small rocks, sea wave ripples, and sidewalk tiles. Unless this distinction is learned, there will be too many proposals in those images.

Hard negative mining further showed the effect of the size of training data on the precision of the resulting classifier. It was observed that although false positives become increasingly more similar to the marker, eventually it will stop improving. An important observation was that, after several stages, patterns of hard negatives seem to disappear in some stages but reappear at a later stage. Although hard negative mining does improve precision, the size of training set still limits it. If the training size is too small, precision will be bounded regardless of the goodness of the negative set. This is expected because the negatives are over 99% of examined windows in high-resolution images, and the negative training set should be large for precision to be high. Therefore, in our algorithm, the negative set is composed of the results of several hard negative mining stages. It became clear that SVM classification with HOG features and training size of 8,000 samples is not sufficient. With 120,000 training samples, 91% recall and 5% precision were achieved.

However, one of the major contributions of this thesis is a method to achieve comparable recall and precision with only 6,000 training samples. Hard negative mining was an essential step towards achieving that. It was used to gain insight on how patches that look different than the marker can have a similar descriptor. This is one of the early concerns in the project: to find out whether it is possible for a given descriptor to achieve required recall and precision. Seeing numerous examples of background patches that the descriptor does not differentiate from the marker showed the need for more than image gradients. It opened the door for achieving the same results as standard HOG, but with only 5% of the training sample, and without
increasing descriptor vector length.
Chapter 4

Methodology

Figure 4.1: Standard HOG description struggles to separate patches with similar intensities but different gradients.

The focus in the later stages of this work was motivated by a single observation, shown in Figure 4.1. After training on a single rotation and scale, only one pattern remained in false positives. By examining a set of mined hard negatives, it was observed that the HOG descriptor cannot sufficiently encode the difference between filled and unfilled shapes. This is directly relevant to precision. Having different intensities but similar gradients to the marker is not the outlying case among false positives, it is the norm. Only a minority of false positives resemble both marker intensities and marker gradients. This is because those false positives are often frequently occurring patterns, such as sidewalk tiles. Based on this observation, the aim was to allow the HOG to capture the difference between filled and unfilled shapes. It was also examined to give up HOG description, and use intensity values directly to describe
image patches. However, this was not efficient since the descriptor length triples, making classification longer. More importantly, HOG description offered more in the tradeoff between translation invariance and precision. That is to say, when training on translated samples, HOG-SVM precision remained higher than SVM with raw intensities.

The fruit of this thesis is a novel method allowing HOG features to describe the difference between filled and unfilled shapes. The proposed method eliminates the majority of baseline HOG false positives without affecting recall, vector length, or complexity. Moreover, a generalization is also introduced which allows usage of this method for other tasks where both gradients and intensity values are essential to capture object identity.

4.1 Sample selection and preprocessing

The most straightforward approach to choosing training samples is to randomly select negatives image patches (no marker present) to match the number of the available marker patches. Using this approach, it was found that recall rates between 90% and 97% can be achieved. The drawback is that precision is low and leads to several thousands false positives per image. One of the methods that had the most impact on precision is hard negative mining. To select negatives, first a classifier is trained on randomly selected negatives. The classifier is then run, and the new false positives are, in a sense, harder. For example, an image patch that contains a circular object is considered harder than an image patch of noisy dirt. This is because the circular object is expected to be more similar to the marker, and using it for training improves precision since other circular objects will be distinguished from the marker. The hard negatives are used to train a new classifier, and harder negatives are generated again.
4.2 Fixing Rotation and Scale

Examine hard negatives of classifiers trained using HOG features led to observation of patterns in difficult patches. What is meant by difficult patches are background patches that are continuously mistaken as markers. Two kinds of difficult image patches were observed: patches containing circular objects, and patches containing X-shaped objects. Two educated guesses were made on why it is hard for the HOG to distinguish them from markers. First, training on all rotations of the marker makes it difficult to distinguish circular objects from markers. Second, since the HOG descriptor does not encode raw intensity, filled and unfilled shapes have very similar descriptors. Thus, X shapes, which are unfilled versions of the marker are continuously misclassified. The pattern of circular objects being classified as markers was handled by training only on a single specified rotation and scale. During testing, a patch is rotated and scaled to the specified rotation and scale. This was done by using a 9x9 reference patch rotated to a certain angle. Every search window is rotated to minimize the difference between the reference and the search window patch. After fixing rotation, fixing scale becomes easier. Because a new 32x32 reference image
containing the marker is fixed on a certain scale, and the search window is scaled up
and down to minimize the difference. As a result, every search window containing the
marker will be proposed at a certain angle and scale. There was an immediate increase
in precision after training on a single angle and scale with standard HOG description.
Additionally, circular object were no longer misclassified. The major benefit of fixing
rotation and scale is allowing for even more precision without needing more training
data. After fixing rotation and scale, most false positives contained a ’+’ sign where
the four subregions were filled in various ways, as shown in Figure 4.1. This set of
false positives allowed reaching the main contribution of this thesis: Intensity-aware
HOG description.

4.3 Intensity-Aware HOG Description

Although fixing scale and rotation eliminated circular objects from being mistaken
as markers, one more pattern in false positives remained. Objects that have similar
edges to the marker but are filled with different intensities continued to be mistaken for
markers. This was expected in using the HOG since intensity differences are encoded
but not raw intensities. One intuitive solution is use the vector of raw intensities as a
descriptor instead of the HOG descriptor. However, this is not ideal. First, the HOG
was less sensitive to translation and luminosity changes than raw intensities. Second,
the HOG descriptor was one third the length of using raw intensities. What this
implies is that the HOG can perform faster and more accurately with less training
data. Therefore, the focus of the later part of the thesis was to endow HOG description
with the ability to describe the difference in raw intensity between positives and
negatives. The goal is for patches that have similar gradients but different intensities
in uniform regions to have different HOG descriptors.
4.3.1 Approach

The guiding observation to the proposed method is that brightening the inside of a filled region introduces stronger gradients than brightening the inside of an unfilled region. Similarly, darkening the inside of an unfilled region introduces stronger gradients than darkening the inside of a filled region. The proposed method, in summary, seeks to find a subset of image pixels to brighten and darken such that intensity-dependent gradients are introduced. So given an image patch to classify, intensities would be altered using the proposed method. Then, HOG feature extraction proceeds as normal, except that intensity-dependent edges were artificially introduced through our method.

![Figure 4.3: Brightening or darkening the same regions in filled and unfilled shapes introduces intensity-dependent edges.](image)

4.3.2 Masking and Blending

A mask is used to store which pixels to brighten and which pixels to darken. The mask has the same size as the patches used for classification, so for every pixel in a proposal patch we can look up a pixel intensity in the mask. A mask pixel can be black, gray, or white. Patch pixels matching with black mask pixels are darkened, those matching with gray mask pixels are brightened, and those matching with white mask pixels are left unchanged. For every image patch, three intensities are extracted: median
intensity of all image pixels, median intensity of pixels brighter than the median, and median intensity of pixels darker than the median. Brightening a pixel of an image patch is done by replacing it by the brighter median of the patch. Darkening is done by replacing the pixel by the darker median. The algorithm is shown below, and the process is illustrated in Figure 4.4 and Figure 4.5.

**Function BlendImageWithMask (I, M)**

**input**: An image patch $I$ of size $w \times h$, a mask $M$ of size $w \times h$

**output**: Blended image $BI$ of size $w \times h$

Median $\leftarrow$ MedianIntensityBetween($I$, 0, 255)

BrighterMedian $\leftarrow$ MedianIntensityBetween($I$, Median, 255)

DarkerMedian $\leftarrow$ MedianIntensityBetween($I$, 0, Median)

for $i \leftarrow 1$ to $w$ do

for $j \leftarrow 1$ to $h$ do

if $M(i, j) == 0$ then

$BI(i, j) = $ DarkerMedian

else if $M(i, j) == 128$ then

$BI(i, j) = $ BrighterMedian

else

$BI(i, j) = I(i, j)$

end

end

end

**Algorithm 1**: Blending a mask with an image patch

Figure 4.4: An example of using a mask to alter the intensities of a marker.
4.3.3 Manual Mask Selection

Using specific knowledge of the markers in this thesis, and their specific rotation and scale, a mask can be designed by drawing artifacts in uniform regions of the marker. The mask shown in Figure 4.4 was manually designed to introduce as little alteration as possible to positives, and as much alteration as possible to false positives. The edges in the mask are rotated 45° from positive gradients. Since the top right and bottom left edges are black in the mask, it means they will darken top right and bottom left regions of proposals. Since positives are already dark in those regions, they are only slightly affected by the alteration. For unfilled ‘+’ shapes, the rotated gradients will be introduced in the top right and bottom left quarters of the image, leading to more separation between markers and unfilled ‘+’ shapes. Another one of the best performing manually designed masks is to draw concentric circles in each of the inner four squares of the 32x32 image. The blending is done similarly, and the circles introduce gradients in almost all directions. This leads to a variety of strong gradients in almost every directions, if the proposal has similar gradients to the marker but different intensities. This is desirable since it separates those examples from the marker using the HOG but with intensity awareness. It should be noted that the descriptor maintains its original length. The HOG is extracted normally,
but the difference in our method is that it is extracted from the blended image where gradients were introduced.

4.3.4 An Algorithm for Unsupervised Mask Selection

The second main contribution of this thesis is a general method allowing the HOG descriptor to distinguish filled from unfilled shapes. Given positive and negative image patches, where there is similarity in gradients but disparity in raw intensities, the algorithm should find a mask leading to more interclass separation without user guidance.

Mask selection can be viewed as feature selection. A feature can be a pixel (or a rectangular window of pixels) that is filled with a single intensity (black, gray, or white). The position of a feature determines which original patch pixels will be altered. The intensity of a feature is used to store what kind of alteration will be performed. A black feature intensity means that corresponding pixels in a proposal patch will be replaced with the darker median of that patch. A gray feature intensity means that the corresponding pixels will be replaced with the brighter median. A white feature intensity means no alteration will be done on the corresponding pixels.

Before proceeding with the algorithm, an idea of the ideal mask should be developed. The ideal mask would certainly increase classification performance over standard HOG with no alteration, because that is the ultimate goal. The Fisher score is one way to measure the increase in classification performance, because it exactly measures the interclass variance adjusted by intraclass variances. This is highly desirable; to have positives that are uniform, negatives that are uniform, and the largest distance possible between those uniform classes. It is possible to train a classifier instead, and evaluate performance by calculating the number of misclassifications. Actually, the supervised approach is even more desirable because the problem does need a classifier after mask selection. However, it is too time consuming to train and
test to evaluate every possible feature.

\[
FisherScore(mask) = \frac{(\mu_{HOG(Blended+) - \mu_{HOG(Blended-)})^2}{\sigma^2_{HOG(Blended+)} + \sigma^2_{HOG(Blended-)}}
\]

Where \(\mu_{HOG(Blended)}\) is the mean of HOG descriptors extracted from patches after blending with the mask, and \(\sigma^2_{HOG(Blended)}\) is the variance.

Since the mask is important only inasmuch as classification performance improves, there is no innate need for the mask to be well-structured. Well-structured is used here in the sense that a well-structured mask contains clear edges or exhibits symmetry. Nonetheless, the algorithm proposed here explores only the space of masks where there are clear and continuous edges per every 4x4 pixels of the mask. This is much faster than traditional greedy best first search. In traditional greedy best first search, several sizes of rectangular pixel windows are examined while picking only parts of the mask at each iteration. The mask cannot be selected at once, because the union of the top several features may not be the best set of features of that size. Therefore, often only a few features are selected at each iteration.

Figure 4.6: The 24 patterns are attempted and evaluated at every 4x4 of the mask.
Based on the above explanations, the algorithm for automatic mask selection proceeds as follows: the mask is divided into 4x4 windows, and each window is examined only once. At each window, 24 images of size 4x4 are filled in consecutively and evaluated. These images are images of vertical, diagonal, and circular edges. The images are shown in Figure 4.6.

4.4 SVM Classification

Two SVM classifiers are used. The first is trained using standard HOG features, and a dataset of size 8000 images. The HOG descriptor is a natural fit for capturing the identity of the marker. The descriptor is set to divide images into 4x4 cells, each having a size of 8x8 pixels, where the inner 2x2 cells will always contain the marker center while the remaining cells will contain either one strong edge or none. Precision of this classifier is very low since the classifier is trained on positives that contain markers with different scales and rotations. However, it is used to quickly eliminate the majority of uninteresting image patches quickly, and reduce the search space to a few thousands proposals per image. As mentioned previously, the false positives of this classifier are circular objects and objects that have similar gradients to the marker. In a way, the classifier acts as a slower but more specific object proposal method. It allows skipping the step of fixing rotation and scale for uninteresting patches, which leads to a significant reduction in run time. It should be noted that the first classifier was used for baseline SVM classification in Chapter 3 as well.

The second classifier is trained using the proposed intensity-aware HOG descriptor, and a dataset of around 6000 images. Rotation and scale determination is only done after a patch is classified as a marker by the first classifier. So the second classifier is trained and tested on patches matching the rotation and scale shown in Figure 4.2. For results mentioned in this thesis, the first classifier is used in all experiments. The second classifier is varied between our method and another standard
HOG classifier trained on the fixed rotation and scale.

Using the proposed mask selection algorithm, a mask is selected leading to more separation between positives and negatives. Only training samples are used in mask selection. Then, the classifier is trained on HOG features extracted after blending the mask with training patches. Similarly, during testing the mask is blended with patches before extracting HOG features.

4.5 Detection

In this thesis, detection refers to finding instances of the marker within an image. It is traditionally done through a combination of a binary classifier, and a localization method. The classifier takes an image patch as input, and returns a classification (marker or not marker). The localization method finds the region within a large image where the marker occurs. There are two common methods of localization. One is the sliding window approach, where a window is passed through the image, and the classifier is invoked every time to find windows that are classified as a markers. Another approach is to use object proposals, which is a selective search that proposes windows that are separate from the background. The advantage of proposal generation is that it is generally faster the sliding window approach. The EdgeBoxes proposal method[8], for example, uses edge detection to score windows based on how many edges are contained within the window. The intuition is that a window containing an object generally has edges within the window and outside the window, but few edges would be crossing the window between outside and inside. This is clearly faster that extracting features and feeding them to the classifier at every possible window. In the object proposal approach, the classifier needs to only examine proposals, rather than the whole image.

Although many recent research papers lean towards using the object proposal method, it is unsuitable for detecting markers when the acceptable distance thresh-
old from ground truth is small. In this thesis, the threshold is ten pixels. Unless a proposal is ten pixels or closer to the actual marker, it will not be considered as a true positive. During experimentation, desirable recall was only achieved after 100000 object proposals were evaluated. This an improvement by a factor of ten over the sliding window approach which examines about a million locations. However, windows returned by object proposal methods are not centered, which results in misclassification when the proposals are passed to the classifier. In order to mitigate this problem, nine more locations around every proposal must be passed to the classifier. This eliminates the reduction in windows, because now approximately one million windows must be classified but with the additional overhead of running the object proposal method. Therefore, the algorithm proposed in this thesis uses the sliding window method.

The proposed algorithm runs concurrently since detection in different rows or columns is independent. The sliding window approach is used with a step of three pixels. The step size determination can be crucial, since a detector using a high recall classifier can end up with low recall if the step size is too large. There is no specific explanation of why three pixels are a suitable step size, except that larger steps lead to a reduction in recall. There is an upper limit of ten pixels, since there are markers with a side length of ten pixels. If the step size exceeded ten pixels, those markers will be missed. Of course, it might be possible to train a classifier to detect markers with large translations but precision would have to be incredibly low since precision is low even with centered markers.

4.6 Tracking

After running the previous algorithm on validation set, it is clear that the markers missed by the detection stage are extreme cases. Usually, they are either extremely blurred, skewed, or partially occluded. Since the aerial images used are taken in sequence, detecting a marker in an image usually means that the marker appears
in either the next image, the previous image, or both. This information enables detecting markers that can be very challenging for a precise classifier to detect.

The second stage of the algorithm tracks every detected marker forward and backward in time. This is essential to guarantee high recall even in the occasional presence of abnormal markers. If the speed of the Unmanned Aerial Vehicle (UAV) and its movement direction are known, it would be possible to exactly track detected markers. But since both are unknown, another approach is used to approximate speed and direction: feature matching.

Distinctive features are extracted in a region around detected markers. Distinctive features are also detected in previous and next images in the sequence. This allows for feature matching, where features detected in one image are mapped to the same features in another image. To approximate motion, the median mapping offset is used. Mapping offsets are obtained by calculating the difference in position for each feature that was mapped from one image to another. In literature, the Random Sample Consensus (RANSAC) method[9] is used to obtain one underlying offset based on several feature matching offsets. However, for our task simply using the median offset performed as well as RANSAC. The term ‘offset’ is used here to describe a vector (a direction with a magnitude) that is used to map a pixel from one image to another. In some cases, as many as 8% of testing set markers could only be detected through tracking. This gives a large degree of freedom in tackling the problem. For example, recall can be sacrificed for precision in the detection step while still meeting the goals after tracking. This is because tracking adds only one more proposal for each detected marker per image. So the effect of tracking on precision is negligible, while its effect on recall is immense. Several interest point detectors were used, and it is notable that they perform comparably. Speeded-Up Robust Features (SURF)[10] is used to detect distinctive features in this thesis.
4.7 Pipeline with user interaction

Since the sliding window detection is the bottleneck, the classifier is run first on a sequence of images and proposals are stored without user interaction. After the proposals are generated, the user verifies proposals to confirm which ones are actual markers. The confirmed markers are tracked forward and backward in time, with the user verifying an additional proposal per image until the marker is either out of bounds or incorrectly tracked. Tracking takes a few seconds per image.
Chapter 5

Results

5.1 Dataset

The available dataset contains around 25000 aerial images. These images were taken from different cities, at different heights, with different resolutions. However, there were only around 1600 occurrences of markers, and for each marker, the center was annotated. The dataset was divided using random selection into 20% for training, 20% for validation, and 60% for testing. For the training set, around 300 regions of interest containing the marker were annotated in addition to center location. The rule for annotating regions of interest was that all corners of the marker should be within it. The available markers were rotated to synthetically generate more marker images for SVM training. Negative samples were of the same size and were selected using hard negative mining as explained in Chapter 4. A patch size of 32x32 was used in both training and detection. This size was selected based on two factors: actual marker sizes, and suitability for HOG description. Using a patch size of 32x32 and a cell size of 8x8, the image is divided into four inner cells and twelve outer cells. The four inner cells should always contain the center, while the outer cells can contain a strong edge or part of the background.
Reported results for both methods are obtained by using the same detector but using standard HOG descriptor or the intensity-dependent modification. The two detectors are run through every annotated test image, and the number of detected markers, as well as the number of proposals, are recorded.

5.2 Classification Performance

Two classifiers were trained on the same data. The data includes 300 unique markers that were slightly rotated and translated to generate 3000 positive training samples. The negatives are 3000 extracted background patches. Positive training samples, negative training samples, and test samples are 32x32 pixels.

Recall rates were fixed by testing on the validation set, which contains 300 additional unique markers. The number of false positives was obtained by testing on 50000 extracted negative patches. To eliminate the effect of the detector pixel step, classification was performed on extracted patches. Results show that the proposed intensity-aware HOG descriptor is far superior to standard HOG for marker detection. The number of false positives obtained by the intensity-aware HOG descriptor
is, on average, one tenth of the number of false positives obtained by the standard HOG descriptor.

Figure 5.2: A plot comparing the number of false positives per each level of recall between intensity-aware HOG and standard HOG.
Figure 5.3: This is the same as Figure 5.2 but zoomed on recall rates between 91% to 96%.
Chapter 6

Concluding Remarks

In this thesis, an improvement on the HOG descriptor is proposed which eliminates the majority of baseline HOG false positives for marker detection. This is achieved by altering image intensities to introduce intensity-dependent gradients. The majority of baseline HOG false positives are patches that have similar gradients but different intensities. Introducing intensity-dependent gradients leads to more separation between the descriptors of patches with similar gradients but different intensities. A mask stores information about the alteration, and during training and testing the mask is used to alter proposals. After a mask is chosen, very little complexity is added to HOG feature extraction. The only additional overhead is a single pass over image and mask pixels.

A generalization of this method is introduced. Given any positive and negative patches, the method finds a mask that leads to more separation between the two classes. Two choices are made to simplify mask selection. First, the Fisher score is used as a measure of mask goodness. The Fisher score is chosen because it is too time-consuming to train and test at every iteration of mask selection. Second, only a subset of masks is examined. Masks resembling noise should not be explored since they are unlikely to increase classification performance. In the proposed method, only the subset of masks where edges appear in every window is explored. This window is selected to be 4x4 pixels. The method examines every window only once and adds the edge that leads to the most separation between positives and negatives.

By applying the contribution of this thesis to marker detection, the number of false
positives is reduced to only 11% of the number of false positives obtained through standard HOG description. Intensity-aware HOG description allows for false positive rates that are not possible to achieve using standard HOG description without increasing vector length. Complexity is only increased minimally, because the only added operations are computation of three medians and a single run through image pixels. While mask selection in general is very expensive, our method selects the whole mask while examining every pixel only once. Furthermore, mask selection is an offline procedure, and once a mask is selected, the complexity of mask selection does not affect run time.
REFERENCES


APPENDICES

A Papers Submitted and Under Preparation