A Novel Two-Level Shape Descriptor for Pedestrian Detection

Mohamed Elmikaty [#1], Tania Stathaki [#2], Paul Kimber [*3], Stamatia Giannarou [+4]

[#] *Department of Electrical and Electronic Engineering, Imperial College London*
South Kensington Campus, SW7 2AZ, London, UK
[1] mohamed.elmikaty11@imperial.ac.uk
[2] t.stathaki@imperial.ac.uk

[*] *SELEX Galileo Ltd, Christopher Martin Rd, Basildon,Essex, SS14 3EL, United Kingdom*
[3] paul.kimber@selexgalileo.com

[+]*Hamlyn Centre for Robotic Surgery, Imperial College London*
South Kensington Campus, SW7 2AZ, London, UK
[4] stamatia.giannarou03@imperial.ac.uk

Abstract— **The demand for pedestrian detection and tracking algorithms is rapidly increasing with applications in security systems, human computer interaction and human activity analysis. A pedestrian is a person standing in an upright position. Previous work involves using various types of image descriptors to detect humans. However, the existing approaches, although exhibit low misdetection rate, result in high rate of false alarms in the case of complex image backgrounds. In this work, a novel approach for pedestrian detection is proposed which is based on the combined use of two object detection approaches with the aim of reducing the false alarm rate of the individual detectors. These are the Histogram of Oriented Gradients (HOG) and a Shape Context based object detector (SC). Preliminary results are very encouraging and demonstrate clearly the ability of the proposed system to reduce the number of false alarms without significant increase in the processing time.**

I. Introduction

Robust human detection and tracking in complex scenes with heavily cluttered background is a challenging task in computer vision. It serves a lot of applications including surveillance, augmented virtual reality, human activity analysis and human computer interaction. The difficulty in people detection is that algorithms must provide efficient and accurate detection under different illumination conditions and must possess at least scale and rotation invariance properties. Moreover, they should work well under partial occlusion and be able to re-acquire the previously detected human after complete occlusion occurs. For real-time applications, these algorithms must provide reliable solutions with short processing time.

One of the most recent approaches to the problem of human detection and tracking [1-2] is to follow the so called "tracking by detection" approach. A detector is applied to the given image to initialise the tracker to the location of the object of interest. The tracker is then used to follow the object's motion in the next consecutive frames. The detector is applied to update the tracker from time to time, to search for new objects introduced to the scene and to relocate objects in case the tracker has missed them.

A common method for object detection is to scroll a window over a given image. This detection window produces a binary output which means that it fires an alarm when a human exists in the image under examination, otherwise it does not. This approach is usually based on training a classifier to recognise local image features of a human derived from large numbers (of order of thousands) of positive images that contain humans and negative images that contain "non-humans", i.e., contain a large variety of objects other than humans.

Most of the existing image models for object detection employ local features which are based on the information related to local image gradients. An example of a local image gradient model is the well-known Scale Invariant Feature Transform (SIFT) [3] which identifies locations that are maxima or minima of a difference-of-Gaussian function. Another example is the pedestrian detector introduced by Dalal and Triggs [4] that is based on Histograms of Oriented Gradients (HOG) and has good detection results that are invariant to scaling, rotation and illumination. These descriptors are able to capture gradients that are very distinctive of local shapes. The HOG is used in conjunction with Support Vector Machine (SVM) classifiers for pedestrian detection. The above mentioned HOG-SVM models suffer from high false alarm rate. Viola et al [5] introduced the idea of using AdaBoost to train classifiers on Haar-like wavelets and space-time differences. They started their chain of classifiers with simple classifiers, e.g., two-feature classifiers, followed by more complex classifiers that use more features so as to reduce the computational time [6]. The detector in [5] has been widely used for detection of faces. Another known approach that can be used for object detection is the Shape Context (SC) [7]. It depends on building a two dimensional histogram of the spatial distribution of edge pixels of an edge map of a given image and checking the similarity between this histogram and the edge map histogram of a prototype image. This is an image which contains a well-defined version of the object of interest placed on a plain background. SC takes into account locations of edges rather than their orientations [8]. Results are satisfactory, however running this algorithm requires a lot of processing time and gives poor results if the object of interest is placed on a complex background. More approaches on object detection and performance of local descriptors can be found in [9-10]. In this paper we aim at combining the HOG detector and an additional object detector developed in [5], based on the Shape Context (SC) [7], with the goal of pedestrian detection with low false alarm rate.

II. Histogram of Oriented Gradients (HOG)

In this model, the proposed image feature sets are based on dense and overlapping modelling of image regions (normally rectangular patches) using the so called Histogram of Oriented Gradient (HOG) descriptors [4]. These descriptors are broadly divided in two classes, i.e., static and motion HOGs. Static HOGs are computed over individual still images and form the appearance channel. Motion HOGs are computed over a set of consecutive frames of a video sequence and form the motion channel, which is used only for detections on videos. Both static and motion HOGs are based on edge orientation histograms. The key difference is that static HOGs are computed over image gradients whereas motion ones over differential optical flows.

In this work we use static HOGs applied to a set of data provided by SELEX. A detailed description of the static HOG feature extraction chain is given in [4]. The approach relies on evaluating a dense grid of local histograms of image gradient orientations over the image windows. The hypothesis is that local object appearance and shape can often be characterised rather well by the distribution of local intensity gradient or edge directions, even without precise knowledge of the corresponding gradient or edge positions.

HOG is a training-based type of detector. That means a classifier, which in the case of HOG is a Support Vector Machine, is trained to recognise an object of interest by using historical data which consist of thousands of “object” and “non-object” images.

Advantages and Drawbacks of HOG

In terms of advantages, HOG exhibits a good detection rate, although it often misses random people, even in cases where their position does not suffer from any type of distortion. The above statement can be verified by a large number of experiments. Quantifying universally detection rates of interest (detection rate, misdetection rate and false alarm rate) is not realisable since the above probabilities are explicitly related to the type and content of the images used. Furthermore, HOG is relatively fast, i.e., it takes approximately 100 seconds per image frame in C++.

The main drawback of HOG is that it exhibits high false alarm rate. A representative example is illustrated below in Figure 1 where we get the correct detection of the human but also five false alarms. Furthermore, HOG is based on the availability of a large and rich database of objects. Currently, the best database available in the literature is a pedestrian database created by the authors of HOG [4]. For that reason HOG is currently used mainly for the detection of pedestrians. HOG is not robust to even weak geometrical distortions of the shape. For example, if a pedestrian has a tilted head HOG will most probably fail. Finally, HOG is not robust to even mild occlusion. For example, if a human’s feet are covered up to a certain height HOG will again most probably fail.

The problem of high false alarm rate makes HOG impractical in many real life scenarios and applications. Based on that observation and taking into consideration the good detection rate of HOG we were motivated to attempt combining the HOG with other object descriptors in order to reduce the false alarm rate of HOG.

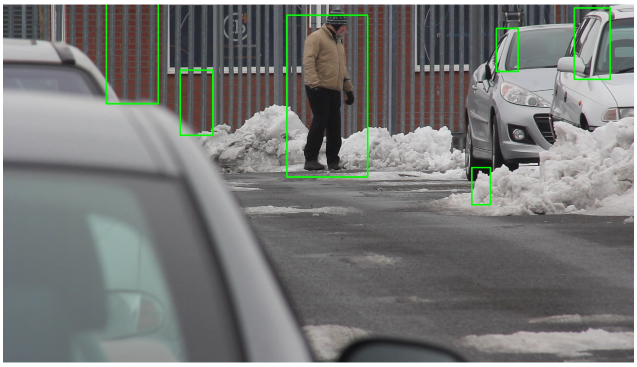

Fig. 1. A large number of false alarms illustrated after applying HOG

III. An Object Detector Based on Shape Context (SC)

The so called Shape Context descriptor for object contour representation was first introduced in [1]. Unlike HOG, it is a prototype-based type of descriptor and not a training-based type one. More specifically, we seek for an object of interest within a complex scene by comparing exhaustively the scene with a prototype image. In order to extract this descriptor, the edge map of the image of interest is required. In this work we use the Canny edge detector [3]. The shape context of a pixel is the two dimensional histogram that describes the spatial distribution of edge pixels within a circle centred at the pixel. The histogram bins are obtained from the uniform partition of that circle in the log-polar space. The log-polar space is preferred to the Cartesian space since it provides a contour descriptor that gives more emphasis to the closest neighbours of the point of interest. The size of the examined neighbourhood around contour points must be appropriately defined and it is a crucial parameter for the object identification results. A neighbourhood, large enough to capture the configuration of the entire object’s shape, leads to a global descriptor which is effective when the object of interest is not occluded within the complex scene. Another, key characteristic of the global descriptor is the underlying assumption that objects to be compared must have similar shape. Therefore, the performance of the global descriptor degrades in cases of partial occlusion or differentiation of the object within the complex scene in relation to the prototype object of the test (prototype) image. To tackle these scenarios, a local descriptor would be more efficient especially for the identification of partially visible (occluded) or distorted objects in the complex scene, since, in that case, local shape properties are more representative features of the class of the object. Therefore, the challenge in the derivation of the shape context representation is to define the optimal size of the descriptor aiming at the extraction of shape features that are most distinctive for retrieval under the ambient conditions related to the specific problem under consideration. The shape context descriptor turns out to be a powerful tool, facilitating robust shape matching. By construction, the descriptor is invariant under translation of the test object within the image. Rotation invariance is achieved by considering the positive x axis for the local coordinate system to be the edge orientation direction at each point. Uniform scale invariance is achieved if the radial distances in the log-polar representation are normalized by the mean distance between all pairs of points in the shape. Obviously this approach to scale invariance refers to the case when the entire object under consideration is present in the image and furthermore, the minimum size circle that includes the entire object does not contain parts of other objects in the background so that the global shape descriptor can be applied. We propose to guarantee scale invariance by estimating for each detected edge point multiple shape context descriptors with varying radii neighbourhoods. The optimum scale is the one that gives the minimum total matching cost when estimating the correspondences between the edge points of the complex scene and the points on the prototype object, as it is described later on.

We assume a prototype object P and a complex scene S both represented as a collection of edge pixels. The similarity between two edge points (pixels) is related to the similarity between their shape contexts. Furthermore, the similarity between two shapes is related to the similarity between their sets of shape contexts. We can then set up the problem of corresponding (matching) contour points between a prototype object's edge map and a complex scene's edge map as that of finding pairs of shape contexts with the maximum similarity. In this work, for the cost of matching two points we use the χ^2 distance between their shape contexts. More specifically, the cost of matching a point s_i on the scene S to a point p_j on the shape P is denoted by C_{ij}:

$$C_{ij} = \frac{1}{2}\sum_{k=1}^{K}\frac{[h_k(p_j) - h_k(s_i)]^2}{h_k(p_j) + h_k(s_i)}$$

where $h_k(\cdot)$ stands for the number of contour points in the k^{th} bin in the log-polar neighbourhood and K is the total number of bins defined previously. The above expression yields an $m \times n$ matrix C, where m, n are the number of edge points on P and S, respectively. Given the set of costs, C_{ij}, between all the pairs of points, we can proceed to estimate the optimum set of corresponding pairs, $\{s_i, p_{\alpha(i)}\}$, between the complex scene and the prototype image. The function $\alpha(\cdot)$ stands for the point mapping $\alpha : S \rightarrow P$.

When we tackle the problem of matching edge points from a complex real life scene to edge points of a prototype (target) object image we face the issue that the former image usually contains additional objects, textured surfaces, noise and other types of signals which produce additional edge points, which are "irrelevant" for the object detection system. Provided that the local shape context model is good enough, the irrelevant edge points should not be matched to points on the prototype object's contour. Furthermore, part of the target object's contour might be absent in the complex image, due to occlusion of the object in the scene, or other types of variations and distortions. As a result, several contour points on the prototype object will have no matches on the scene. Finally, there are situations where the scene might contain multiple objects similar and/or identical to the prototype and in that case, a prototype contour point can be matched to multiple scene points.

Based on the above problem specifications, the assumption of one-to-one correspondence between the prototype and the scene points which has mainly been examined in the available literature [12] would be very restrictive, failing to identify multiple and/or occluded targets in the scene. Hence, we seek not a one-to-one matching, but rather a many-to-one correspondence between the scene points and the target.

Let us assume the $n \times m$ non-negative cost matrix C_{ij}, where the rows and columns correspond to points from the sets S and P, respectively. In our approach we initially assume that all points in the complex scene, including the irrelevant edge points, will be matched to a point in the prototype image. For that reason we introduce a new, so called "dummy" point to the set P and create the set P^+. This additional point extends automatically the $n \times m$ cost matrix C_{ij} to a new one of size $n \times (m+1)$ by padding the new column with a "dummy" cost ε_d. A point will be matched to the dummy point whenever there is no real match available at cost smaller than ε_d. The set E of point pairs of size $n \times (m+1)$ is introduced, including all the possible matches between S and P^+ expressed in terms of edge point index pairs, where $(i, j) \in E$ implies that $i \in S$, $j \in P^+$.

Our objective is to match specific complex scene edge pixels to pixels on the prototype image by selecting a subset A of the set E, $A \in E$, of size n of corresponding point pairs $A = \{(s_i, p_{\alpha(i)}), i \in [1, n]\}$ that minimizes the total cost of matching $\sum_{i=1}^{n} C_{i\alpha(i)}$, subject to the constraint that each scene point that belongs to A is matched to one prototype point and each prototype point in A is matched to at least one scene point. The solution of this "many-to-one" assignment problem defined as finding a function $\alpha : \{1, \ldots, n\} \rightarrow \{1, \ldots, m+1\}$ such that values in $\{1, \ldots, m+1\}$ can be matched more than once, is given analytically in [8]. Ideally, if the whole prototype object exists in the complex scene, the matched points will lie within a localized region of the image being indicative of the location of the prototype object on the scene. However, this in not always the case since the presence of a distorted or occluded version of the prototype in the complex scene may affect the performance of the matching process. For instance, in the case of occlusion, a contour point that belongs to a part of the prototype shape which does not appear in the complex scene might be corresponded to an edge pixel with a similar shape context which, however, may lie on a test object dissimilar to the prototype one. Likewise, in the case of distortion, a contour point on the prototype may be mismatched because the shape context of its actual homologous point on the complex image is not the most similar one. These mismatched edge points are randomly distributed along the scene.

During retrieval, if the prototype or a transformed (scaled or rotated) or a partially occluded version of it appears in the complex scene, the matching process will form regions on the complex scene with dense and/or sparse distributions of matched points. If part of the prototype object is present in the complex scene without being significantly distorted we should expect at least one localized dense distribution of matched points. Intuitively, dense regions have greater possibility of being regions of interest, while sparse regions are likely to correspond to objects which exhibit weak similarities with the prototype object. Furthermore, isolate matched points are probably mismatches that should be ignored and removed.

We propose the use of data clustering as employed in [8] for the partitioning of the matched points in the complex scene into groups (clusters) of neighbouring matched points followed by a multistage type of elimination of the "irrelevant" clusters. The algorithm is divided into three stages. Initially, the clusters are estimated using a standard technique. By forming clusters, isolated matched points are removed and in general, cluster continuity must exceed a threshold. In the second stage, we seek to eliminate low activity clusters which consist of matched points that form (almost) straight lines. The idea behind this process is that low activity contours form parts of a large variety of objects and therefore, are non-distinctive identifiers of a specific object. Finally, from the remaining clusters we seek to discard the clusters with sparse distributions of matched points estimated using a density metric. This multi-stage type of cluster elimination is analytically described in [8].

Advantages of the proposed shape context based object detector include the fact that it does not require training, the false alarm rate is acceptable and is robust to occlusion, rotation and scale distortion. The main drawback is that the speed of implementation is relatively slow.

IV. The Proposed Combined Pedestrian Detector

We first applied the HOG detector motivated by the facts that HOG is fast and has good detection rate.

After we applied the HOG, we initially thought to apply the Shape Context model only within the areas of the image detected by HOG. By doing this we were hoping to reduce the high false alarm rate produced by HOG and at the same time to cope with slow speed of SC. However, a pedestrian is a rich object with a large number of irrelevant and/or weak edges. For that reason we decided to look for the upper part of the pedestrian's body, i.e., the so called Omega Shape, which consists of the head and shoulders of the human. Therefore, in order to apply the SC we created a prototype Omega Shape object. This process reduced even more the area within which we applied the SC, shown in blue boundaries, in Figure 2.

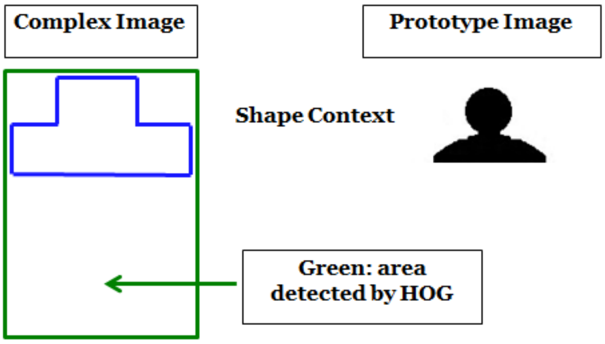

Fig. 2. Shape Context is applied within the area defined by the blue boundary

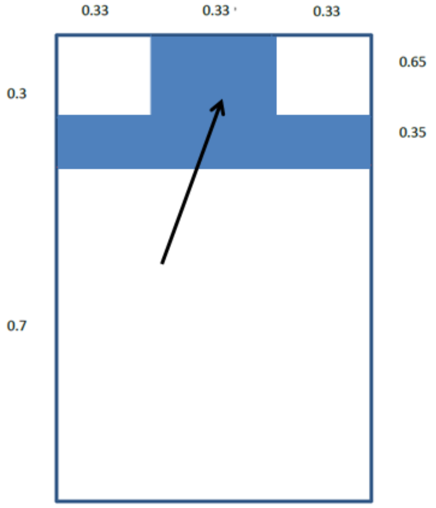

Fig. 3. Relative sizes between HOG and SC detection areas

V. Experimental Results

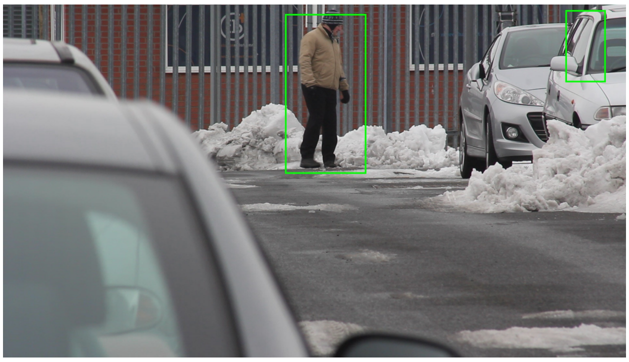

Fig.4. Combined detector applied to Fig. 1 produces 1 false alarm instead of 5

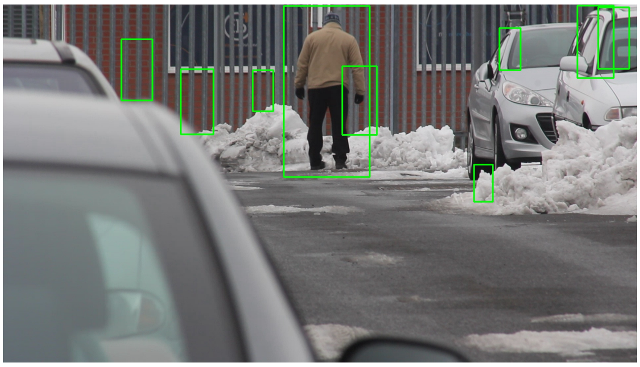

Fig.5. HOG detector with 8 false alarms and the correct detection shown

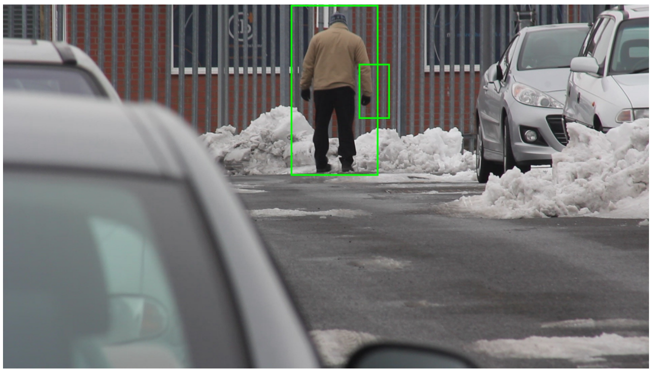

Fig.6. Combined detector applied to Fig. 5 produces 1 false alarm instead of 8

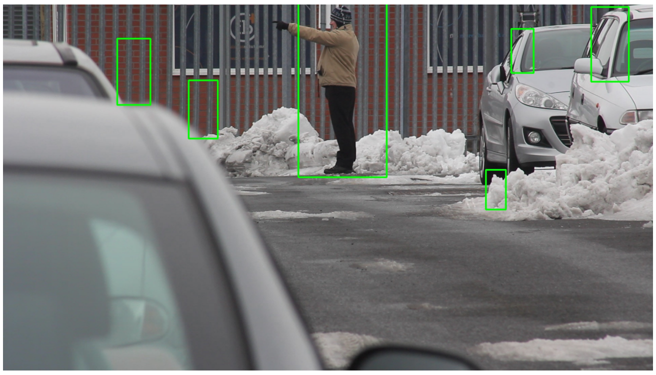

Fig.7. HOG detector with 5 false alarms and the correct detection shown

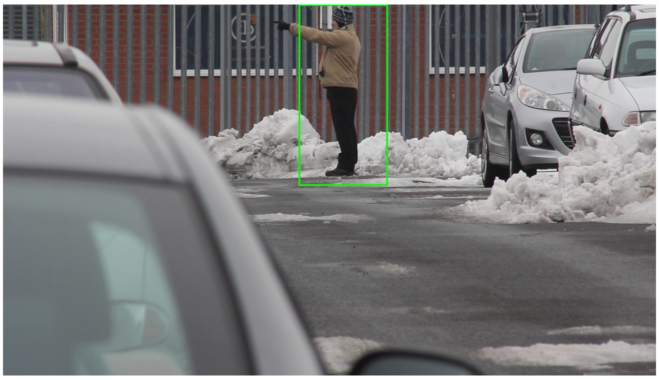

Fig.8. Combined detector applied to Fig. 7 does not produce any false alarms

For the implementation of SC in the proposed work cluster activity must be greater than 1. A cluster must satisfy one of the two conditions: Cluster continuity must be less than 2 whereas cluster density must be greater than 0.4 or continuity is less than 0.1 and density is greater than 0.25. Definitions of activity, density and continuity can be found in [8].

VI. Conclusions

The proposed algorithm which is a combination of the HOG and a Shape Context based detector is a novel method for pedestrian detection. By using this clever combination, we managed to decrease significantly the number of false alarms. The algorithm was tested on a complex dataset and the results are consistent. Quantification of the results, i.e., false alarm rate measurements of the proposed method, requires a long term research on a large number of scenarios which involve a large number of image types. Among current limitations of the proposed algorithm is that it requires longer processing time compared to the sole use of HOG. Furthermore, the colour of the background objects must be different from the human to be detected so that the edge detector can identify the boundaries of the human's upper body part. The algorithm does not require additional hardware. It works in both indoor and outdoor environments and it is not range limited.

Acknowledgements

This work is based on research that was originally carried out under the SEAS DTC and CDE DUST (Dismounted Urban Sensing and Targeting) programmes as a collaborative activity between Imperial College and SELEX Galileo Ltd. The authors would like to thank the UK MoD for their support to the underpinning research, and colleagues in SELEX Galileo and Imperial College for helpful discussions and the provision of video image datasets.

References

[1] S. Avidan, Ensemble tracking," IEEE Trans. PAMI, vol. 29, pp. 261-271, 2007.

[2] B.Wu and R. Nevatia, Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," International Journal of Computer Vision, vol. 75, pp. 247-266, 2007.

[3] D. G. Lowe, Object recognition from local scale-invariant features," in Computer Vision, vol. 2, 1999, pp. 1150 -1157.

[4] N. Dalal and B. Triggs, Histogram of oriented gradients for human detection," in CVPR, vol. 1, June 2005, pp. 886-893.

[5] P. Viola, M. J. Jones, and D. Snow, Detecting pedestrians using patterns of motion and appearance," in The 9th ICCV, editor, Ed., vol. 1, 2003, pp. 734-741.

[6] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features," in CVPR, vol. 1, 2001, pp. 511-518.

[7] S. Belongie, J. Malik, and J. Puzicha, Shape matching and object recognition using shape contexts," IEEE Trans. PAMI, vol. 24, no. 4, pp. 509 - 522, Apr 2002.

[8] S. Giannarou and T. Stathaki, Advances in Intelligent Signal Processing and Data Mining: Theory and Applications (Studies in Computational Intelligence). Springer, 2013, vol. 410, ch. 9.

[9] D. M. Gavrila, The visual analysis of human movement: A survey," CVIU, 1999.

[10] K. Mikolajczyk and C. Schmid, A performance evaluation of local descriptors," IEEE Trans. PAMI, vol. 24, no. 10, pp. 1615-1630, 2005.

[11] J. F. Canny. A computational approach to edge detection. IEEE Trans. Pattern Anal. Machine Intell., 8(6):679–698, Nov 1986.

[12] R. Jonker and A. Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. Computing, 38:325–340, 1987.