Spatio-temporal Consistency for Head Detection in High-Density Scenes

Emanuel Aldea, Davide Marastoni, K. H. Kiyani

Spatio-temporal Consistency for Head Detection in High-Density Scenes

Emanuel Aldea[1], Davide Marastoni[2], and Khurom H. Kiyani[3]

[1] Autonomous Systems Group, Université Paris Sud, France
[2] Università di Pavia, Italy
[3] Communications and Signal Processing Group, Imperial College London, UK

Abstract. In this paper we address the problem of detecting reliably a subset of pedestrian targets (heads) in a high-density crowd exhibiting extreme clutter and homogeneity, with the purpose of obtaining tracking initializations. We investigate the solution provided by discriminative learning where we require that the detections in the image space be localized over most of the target area and temporally stable. The results of our tests show that discriminative learning strategies provide valuable cues about the target localization which may be combined with other complementary strategies in order to bootstrap tracking algorithms in these challenging environments.

1 Introduction

One of the strongest recent developments in computer vision has been related to the analysis of crowded scenes. The interest that this specific field has raised may be explained from two different perspectives. In terms of applicability, continuous surveillance of public and sensitive areas has benefited from the advancements in hardware and infrastructure, and the bottleneck moved towards the processing level, where human supervision is a laborious task which often requires experienced operators. Other circumstances involving the analysis of dense crowds are represented by large scale events (sport events, religious or social gatherings) which are characterized by very high densities (at least locally) and an increased risk of congestions. From a scientific perspective, the detection of pedestrians in different circumstances, and furthermore the interpretation of their actions involve a wide range of branches of computer vision and machine learning.

A rough but quite consistent indicator of the difficulty of analyzing a crowded scene is represented by the number of pixels associated to individual targets (pedestrians). For large objects clearly exhibiting body parts at least sporadically, the detection and tracking algorithms have advanced significantly in the last decade [1]. The aim of the present work is to investigate contexts in which the scale of the scene or other logistical or practical constraints impose a small target size; this is typically the case of large scale, high-density crowds. In these circumstances, research efforts have focused primarily on holistic approaches for analysis, which involve primarily the extraction of coarse-level information, such

as flow patterns or texture. Although these parameters may be sufficient for characterizing the crowd up to a certain scale, they are unable to grasp finer variations in local dynamics which are not consistent with the global flow, or in local density. However, these fine scale phenomena are essential, not only for security considerations, but also for understanding better the interactions among targets at high density levels and their influence on the dynamics of the crowd.

Single camera analysis represents the typical setup for a broad range of applications related to detection and risk prevention in public and private environments. Although some camera networks may contain thousands of units, it is quite common to perform processing tasks separately in each view. However, single view analysis is limited by the field of view of individual cameras and furthermore by the spatial layout of the scene; also, frequent occlusions in crowded scenes hamper the performance of standard detection algorithms and complexify the tracking task.

Multiple camera analysis has the potential to overcome problems related to occluded scenes, long trajectory tracking or coverage of wider areas. Among the main scientific challenges, these systems require mapping different views to the same coordinate system; also, solutions for the novel problems they address (detection in dense crowds, object and track association, re-identification etc.) may not be obtained simply by employing and extending previous strategies used in single camera analysis.

In order to perform large scale crowd analysis supported by dense tracking, a multiple camera approach is imperative in order to cope with strong, frequent occlusions. Nevertheless, in order to initialize the tracks, a hypothesis about the location of the targets has to be formulated in single camera views; then this hypothesis may be refined in multiple projections. In our work, we study the problem of providing a preliminary initialization of a target density map in high-density strogly occluded environments. The aim of the method we propose is thus not to provide a perfect, exhaustive detection of targets, but rather to bootstrap the tracking process with a detection process which may be somewhat tolerant to false positives, since temporal and multiple camera cues enforced by a full tracking framework would have the ability to perform further filtering.

In order to formulate detection hypotheses in the image space, we rely on a discriminative learning process. This solution has been used extensively and is de facto the algorithm employed for pedestrian detection in non crowded environments, and recently in applications where visibility is often reduced to upper body parts. Our work shows that, among other strategies that are necessary for tackling the problem of person detection in high-density scenes, discriminative learning performs reliably and may be employed in order to initialize a large scale tracking process. The value of such a study rests on the need for better solutions for studying crowded human urban environments in order to improve the security of the flows involved, and the supporting infrastructure as to increase and not diminish the comfort of participants.

Our paper is structured as follows. The next section presents the related work which is relevant for the problem we address, and underlines the relevance of our

investigation in the context of identifying low-resolution targets with frequent occlusions. Section 3 highlights the main steps performed in the discriminative learning based classification of image content, at pixel level. Section 4 presents a preliminary filtering strategy that allows for taking into account the spatial and temporal coherence that is expected from true positives in a video sequnce. Section 5 illustrates an application of the proposed algorithm to the analysis of a highly crowded scene, and Section 6 presents the conclusions of our study and future directions of work.

2 Related work

The growth of the cities and their evolution towards megalopolises have transformed the foundations of our society. The worlds population is projected to grow from the current 7 billion to around 9 billion by 2050, and hence to increase the burden on resources and on the associated demands on public transport. In the light of these concerns, and also on the grounds of safety improvement during mass events [2, 3], there is an urgent imperative to study in detail the phenomena occuring in high-density crowds.

The interest surrounding the study of crowd phenomena spanned during the last decade across multiple fields, including physics, sociology, simulation, visualization and computer vision; among them, computer vision has an essential role of linking the theoretical field with the actual phenomenon (i.e. *calibration* and *validation*) through video analysis (denoted also in other fields as empirical data collection). Indeed, models used in simulations have not been either proposed or validated for high-density crowd scenarios. In the case of real data i.e. recordings of dense crowd movement, the extraction of pedestrian trajectories has been performed either by human operators, a process which is time consuming and cumbersome, or in an unsupervised manner but only in specific conditions i.e. vertical cameras and using primitive methods. In both cases, a major hindrance is the strong occlusion among pedestrians which makes extracting accurate trajectories or accurate local density information nearly impossible.

As the density of a crowded environment increases, conventional approaches used in video analysis stop working, since supporting hypotheses (visibility of body parts, occlusion level, presence of background, presence of ground plane etc.) are not valid anymore. Most importantly, the behaviour of people involved in the crowd changes in order to adapt to the space constraints and the available degrees of freedom. A high-density environment is considered a scene where density is higher than approximately 4 people/m^2. The immediate consequences of this density are:

- heads are the only visible body parts (except occasionally shoulders)
- there is no static background
- occlusions are frequent and persistent
- the image content is rather homogeneous

Noncrowded scenes have represented for a long time the main area of interest for the computer vision community, and pedestrian detection algorithms evolved

significantly in the last decade, addressing complex applications such as identification of people, grouping analysis, estimation of body parts, gesture based and trajectory based action analysis etc. However, as it has been already highlighted many times [4], all these methods are not appropriate when high-density crowd analysis is performed, and new methods must be designed in order to cope with extreme clutter. Actually, clutter is indeed the main difficulty, but practical considerations also raise difficult questions. Technical difficulties widen the gap between proof-of-concept experiments aimed at high-density crowded scenes and functional solutions. The size of the interest objects, the accessibility to areas of interest, the size of the problem raise as well novel fundamental research challenges that require significant innovations with regards to established methods.

Single camera analysis When coping with pedestrian tracking, the established approach is based on the HOG detector [5], as this representation is adapted for the detection of upright subjects which are at least partially visible. The major applications of computer vision research that are responsible for the advancements in the field are the intelligent transportation and the surveillance industry. Secondarily, advancements in machine learning supported studies focusing on multi-target tracking and models of social behaviour, which are aimed very often at scenes with few subjects and consistent interactions. Among these three applicative domains, surveillance has naturally shifted the most towards the analysis of denser scenes.

Some initial attempts [6, 7], managed to initialize tracking of occluded subjects and proposed an effective approach based on mean-shift [8], or relied on 3D human models integrated into a Bayesian framework, but these methods cannot handle properly persistent occlusions or multiple close-by subjects. In [9], local and global features are used in a probabilistic framework in order to estimate the reliability of a detection; again, this method is sensitive to occlusions and does not scale properly to dense crowds.

It has been shown already that the temporal information may be used in order to analyse the coherence of the movement through clustering and assist the detection process [10]; again, these methods attain their limits for dense scenes because of occlusions, lack of background, homogeneity and similar movement.

Very recently, the detection of the particular head-shoulder shape ("Ω-shape") has been addressed specifically [11–14]. The common characteristics of these studies are the use of the HOG descriptor, of discriminative classifiers and the exploitation of additional image features related to local higher order statistics. Focusing on the detection of the Ω-shape has strong benefits: heads remain visible in crowded scenes, and it generalizes quite well the human appearance from different perspectives. However, the main concern regarding this solution is that additional work has to be done in order to increase the robustness against occlusions ([14] being a promising approach). Secondly, it is not yet clear what would be the minimal size of objects required in order to maintain a good level of detection; as we will see in the following section, in large scale analysis the size of objects is much smaller than the one reported in these works.

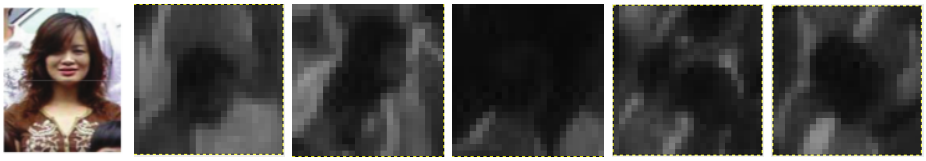
It is worth mentioning at this point a fundamentally different approach to crowd analysis which has been popularized by [15, 16], which uses the spatial organization of the flow field in order to integrate local and global dynamics, and prior behaviour knowledge. The benefit of this approach is that no training data or notions of appearance models are employed at all. The main disadvantage is that different streams are modelled as entities with quasi-constant size and density. Fine-level analysis as the one we intend to perform requires the individual movement of all pedestrians and is able to resolve the dynamics on an individual scale.

Other recent works perform opportunistic tracking in high-density crowds, which relies on the salient appearance of some pedestrians, and manage to track individuals on impressive distances in very difficult environments [17, 18]. The essential aspect of this type of approach is that not only salient targets are tracked reliably, but also their tracking process may propagate to neighboring targets. Even though in our work we do not make any assumptions about the saliency of parts of the scene, we consider that exploiting highly salient objects if they are present is relevant for tracking in high-density crowds in order to add constraints to the detection space. The drawback of this approach is that color information has to be present, and for small targets the penalty of using color sensors which degrade the sharpness of image gradients is significant. Also, it seems that in terms of dense analysis, the community is getting close to a performance limit which is mainly set by the occlusion level, and a fundamental shift is necessary in order to improve the results significantly.

Multiple camera strategies The problem of occlusion cannot be solved robustly by employing single camera recordings. As the interest of the computer vision community extended gradually from single pedestrian tracking in uncluttered scenes to crowd analysis, it has become clear that multiple camera networks are required. The use of multiple cameras for video analysis (mainly surveillance) is an extensive topic, which we cannot cover in detail, but fundamental insights may be found in [19, 20]. We underline though a small scale experiment proposed in [21] which proves the potential of multiple camera tracking in occluded scenes. This study also proposes an effective solution for exploiting jointly hypotheses related to the presence of a head in multiple cameras, but the consistency is evaluated using the pixel intensity information, and extending this type of solution to large scale scenarios raises several difficult scientific and technical questions.

Bootstrapping high-density crowd analysis The characteristics of a crowd that challenges usual analysis strategies are the absence of a background, the occlusions and the homogeneity in terms of appearance and dynamics of the moving mass. These are the main reasons that make the crowd analysis problem still be considered challenging currently [22].

The strategy we propose aims to bootstrap (initialize) a complex analysis process by mapping and refining a probability density to the single view image space. Then, either data fusion or tracking algorithms may be employed in order to benefit from multiple data inputs in the form of multiple camera views.



(a) (b) (c) (d) (e) (f)

Fig. 1. In 1(a) we present for comparison an image used in [12] for discriminative learning of the head-shoulder shape. In the related literature, descriptors are computed on patches of sizes varying between 32×32 to 48×64. 1(b) shows a typical well contrasted head in our dataset. Beside significantly lower resolutions per target, the data we use exhibit often low contrast between close targets 1(c) or between targets and the dynamic background 1(d), and strong occlusions 1(e), 1(f).

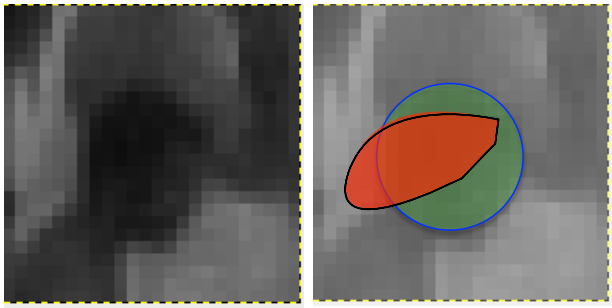
A common prerogative of these algorithms is that they require by themselves an initialization procedure. In the following sections, we show that even in the difficult conditions characterizing high-density crowds, we may obtain head detection maps in single camera views that may be used subsequently either for initializing tracking algorithms, or for extending the detection process within a multiple-camera network.

3 Learning for head identification

In the following section, we will start by detailing the classification process that we use in order to obtain a probability estimation for the presence of a target (a head) in the image space, and then we will also motivate the interest of the benchmark we use, and which is correlated to the purpose of the detection process i.e. initializing a tracker and/or extending the analysis to multiple views.

The descriptor and the learning process In order to perform discriminative learning, we rely on the HOG descriptor initially proposed in [5]. Compared to other studies taking interest in the detection of heads in crowded settings, or at least in environments where the rest of the body is barely visible [11–13], we formulate two fundamentally different assumptions. Firstly, we assume that the size of the targets is significantly smaller - approximately a disk of a three-four pixel radius in the image space - which makes the analysis slightly uncomfortable even for a human (see Figure 1 for an illustration of typical objects sizes used in [12] compared to the targets considered in our work). Secondly, we assume that occlusions are frequent and strong.

Under these circumstances, the parameter which has a significant impact on the classifier performance is the extraction window size. This has to be large enough in order to allow for a reliable characterization of the target using its immediate context, but at the same time small enough in order not to bias the learning process towards non-local learning and towards the detection of joint groups of targets.



(a) (b)

Fig. 2. In 2(a) we present a typical area to be labeled by an user in order to obtain ground truth data. In 2(b) we show the green area around the clicked spot which is considered true positive, as well as a red area which could be the result of a classifier detection.

For the learning task we rely on an SVM classifier, and we consider two different kernel functions. For HOG descriptors h_1 and h_2, we consider a linear classifier $K_L(h_1, h_2) = \langle h_1, h_2 \rangle$, and also the Histogram Intersection Kernel (HIK) function which has shown consistently good performance particularly in the context of pedestrian detection using HOG descriptors[23]:

$$K_I(h_1, h_2) = \sum_{i=1}^{dim} \min[h_1(i), h_2(i)] \tag{1}$$

We perform a pixel-wise classification and then, for each pixel $I(i, j)$ we transfer the binary classifier decision into a probability estimation $p_{i,j}$ [24].

Benchmark design We are interested in obtaining a dense probability map P over the image domain I, which should highlight the maxima associated to the presence of heads ideally through a local plateau, and not only through an unstable peak. This behavior highly desirable taking into account the fact that the objective of the method we present is to act as an *initializer* module for a tracking algorithm. Thus we can tolerate a non-localized response and a certain amount of false positives which may be filtered out by the tracker.

The scoring we propose for evaluating our initializer reflects this aim, and works in the following way. The ground truth data is represented by image content (distinct from training data) where a human user clicks exhaustively and as accurately as possible in the center of the targets. We then expect that *all* pixels located in discretized disks of radius r around ground truth points be classified as positives (the green area in Figure 2(b) corresponding to a click performed in Figure 2(a)). For a certain threshold τ, we consider that probability estimates $p_{i,j} \geq \tau$ are the detected positives (the red area in Figure 2(b)). Then we define the following:

- the true positives (**TP**): detected positives located inside the disks (intersection between red and green areas)
- the false positives (**FP**): detected positives located outside the disks (red area outside the green disk)
- the false negatives (**FN**): detected negatives located inside the disks (green pixels outside the red area)
- the true negatives (**TN**): the rest of the ground truth domain

Theoretically, when we vary the r parameter from $r = 0$ to the typical radius of targets, the performance should increase monotonically, and then decrease for higher values of r. In practice, given the imprecisions of the classifier but most importantly the fact that the targets are far from having a perfectly circular shape close in radius to r, the performance tends to decrease with r. Although this methodology is overly pessimistic, it is helpful for visualizing the performance of the classifier within a local neighborhood of the ground truth points. Consequently, this signature characterizes the ability of the method we propose to make a compromise between precise localization and robust detection, and thus to provide a spatio-temporal persistence of the detection.

4 Exploiting temporal and spatial cues

One of the challenges raised by pixel-wise classification is that the local gradient varies on a high-dimensional manifold, and during the movement of a head through the crowd the classifier response for its constituent pixels is noisy, both on a temporal scale (a moving pixel representing the same head area might exhibit occasionally low detection probability) or on a spatial scale (certain pixels inside a compact head region might exhibit occasionally low detection probability).

In the following paragraphs, we propose a solution for introducing temporal consistency in the probability map based on its temporal evolution. The main assumption that supports our approach is that short-term variations in the probability values should be small for pixels belonging genuinely to targets. Secondarily, we assume that positive responses should be locally high since a target consists in multiple connected pixels, so we would like to encourage clustered responses in the probability distribution.

We underline the fact that with respect to a veritable tracking algorithm, this process is fundamentally different since we do not infer at object level, and thus we limit the consistency check to a limited time interval, and at the immediate pixel neighborhood. However, this is completely in line with our objective of providing a reliable *pixel-wise* label for head detection.

Explicit temporal consistency check In order to associate pixel measurements related to the same entity, we use dense optical flow recursively to project the current pixel in the previous and next N images of the video sequence.

For a detection threshold τ and for the pixel $I^t_{i,j}$ present in the video at coordinates (i, j) at time t, let us consider a corresponding projection $I^{t+k}_{i,j}$, where

$-N \leq k \leq N$. If we consider the probability $p_{i,j}^{t+k}$ as well as the probabilities of all its neighbors (in 8-adjacency), we perform maximal voting in order to obtain the label $l_{i,j}^{t+k}$ of $I_{i,j}^{t+k}$. The objective of this process is to sample temporal information regarding the analyzed pixel by regularizing spatially at the same time in the immediate neighborhood of the projections.

Finally, we perform a maximal vote on the set

$$L_{i,j}^{t} = \{l_{i,j}^{t+k}\}_{-N \leq k \leq N}$$

consisting in the $2N+1$ projection labels we collected, and we assign the resulting label to the current pixel $I_{i,j}^{t}$.

Explicit spatial regularization In order to perform spatial regularization explicitly in the current probability map, we refine a posteriori the pixel classification, by assuming a Markov random field (MRF) over the pixel states. However, this time we consider a basic symmetric neighborhood structure based on 4-adjacency, i.e.

$$N_{i,j}^{t} = \{I_{i-1,j}^{t}, I_{i+1,j}^{t}, I_{i,j-1}^{t}, I_{i,j+1}^{t}\}$$

and we consider as observation set the current probability map associating to the pixel $I_{i,j}^{t}$ the values $p_{i,j}^{t} \in [0, 1]$ provided by the classifier.

5 Experimental results

We tested our head identification method on high-density images acquired at Makkah during very congested times of the Hajj period, in October 2012. For training, we used data from multiple images, amounting for 1032 positive and negative examples. The window size for the HOG descriptor was set to 24×24, according to the considerations we underlined in Section 3.

General observations We trained a linear and a HIK based SVM classifier, and the two algorithms selected 241 and 343 examples as support vectors respectively. A first observation related to the high-density crowd analysis is that the cluttered context gives rise to a significant degradation in the classifier performance. The Figure 3 shows a straightforward detection obtained by applying the linear classifier for each pixel in two different regions of the same image - one which is very cluttered and one where the head density is moderate. The final step consists in obtaining a detection probability map, thresholding it and performing non maximal suppression locally in order to recover only the strongest responses. The moderate density detection illustrated in Figure 3(a) shows that a direct approach is able to provide an acceptable detection result, which could be fed directly to a tracking algorithm for initialization. The cluttered scene however presents an entirely different kind of panorama, with a fair number of peaks that are associated to a head, but also with a high number of misses and a significant number of false detections.

Under these circumstances, the solution we propose is to postpone in the decision process the techniques that lead to loss of information such as thresholding

or non-maximal suppression, and focus the computational effort on improving the consistency of the detection probability map. We present therefore in 3(c) the probability density map for a cluttered area; pixels highlighted in green for visualization purposes exhibit probability values higher than a relatively low threshold. This time we note that despite the fact that there is a certain amount of false positives, the detection map manages to cover most of the targets, while filtering out at the same time a good amount of non-relevant areas.

ROC analysis of the detector In the following part of the section, we will try to analyze quantitatively the interest of exploiting the density map and improving its consistency. In order to have access to numerical estimates of the detection performance, we define a ground truth set consisting in image content where we identify exhaustively and as accurately as possible the persons which are present; the ground truth set amounts for a number of 132 targets. Then we apply the benchmarking strategy detailed in Section 3.

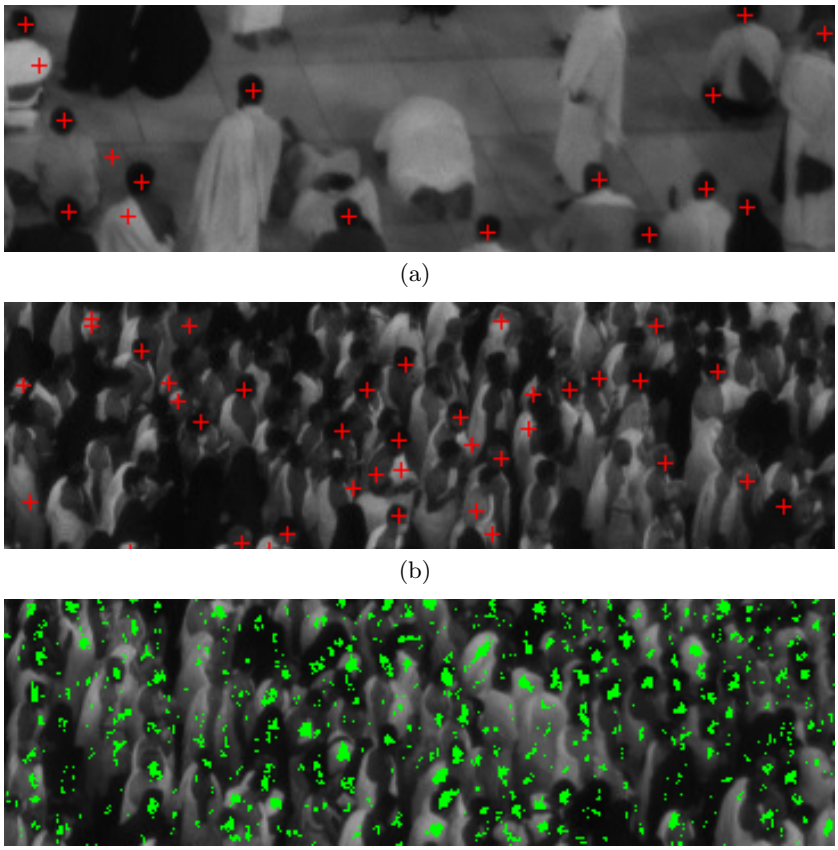
For each of the testing scenarios which are illustrated in Figure 5, we consider a ground truth radius r around the pixels clicked by the human user ranging from $r = 0$ (we consider only the clicked points) to $r = 4$, which is close to the upper limit for a head radius in our image set. Then for each of these five values for r, we vary the detection threshold in the interval $\tau \in [0, 1]$ and we compute the False Positive (FPR) and True Detection (TDR) rates, which are used for plotting the corresponding ROC curves.

Figure 4(a) illustrates the performance of the linear classifier for the different values of r; the locations indicated by a cross in the corresponding color show the performance in terms of FPR-TDR of the classifier output which has been regularized using a MRF approach as depicted in Section 4. We note that the performance evaluated using this metric is located above the corresponding ROC curves. Figure 4(b) presents the performance for the same base classifier where we introduced the temporal consistency check presented in Section 4. In this case, we did not perform the explicit spatial regularization since the output of temporal consistency check is a binary labeling, and the extra information gain a posteriori is insignificant.

Finally, Figures 5(a) and 5(b) illustrate the same metrics in the case of the HIK classifier. Table 1 allows for a precise quantitative comparison among the proposed strategies in terms of the area under the different ROC curves.

Table 1. Area under the ROC curve for the different classification strategies proposed.

Kernel type	Consistency	$r = 0$	$r = 1$	$r = 2$	$r = 3$	$r = 4$
Linear	-	0.8446	0.8237	0.7948	0.7259	0.7035
Linear	Temporal	0.8548	0.8409	0.8164	0.7771	0.7296
HIK	-	0.7742	0.7673	0.7413	0.7059	0.6666
HIK	Temporal	0.8998	0.8565	0.8316	0.7897	0.7420

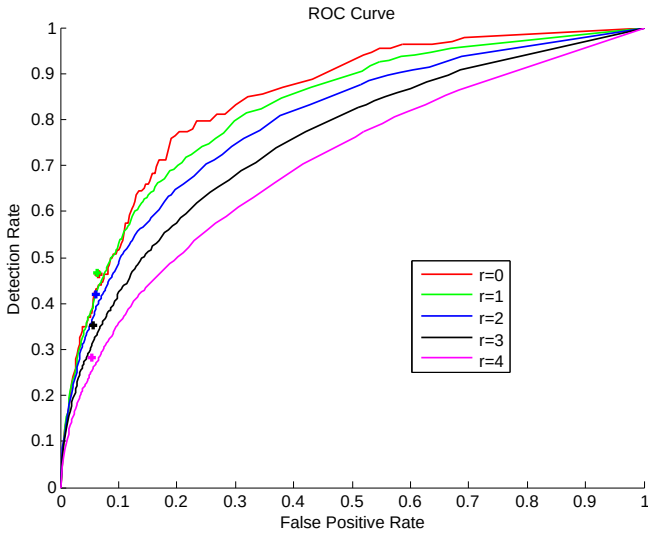


(a)

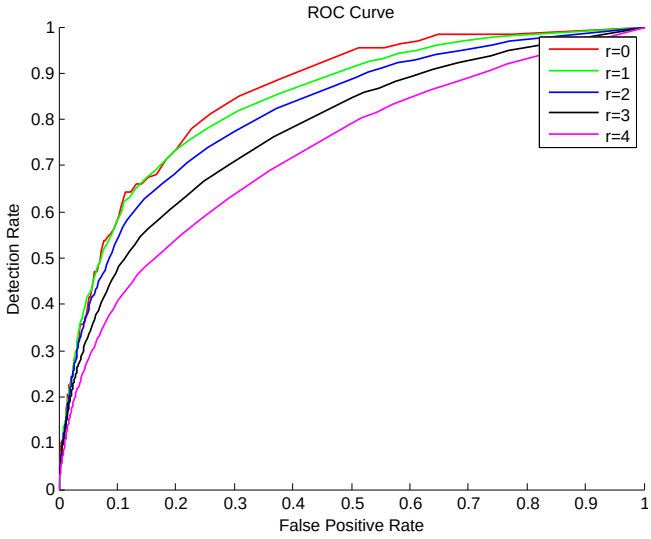
(b)

(c)

Fig. 3. In Figure 3(a) we present the results of a straightforward detection involving thresholding and non-maximal suppression on a non-crowded area of the scene free of occlusions, with good results. The same algorithm fails to exhibit the same good performance in a cluttered environment - see Figure 3(b). We argue that in these cases is to refine the probability map over the image space (illustrated in Figure 3(c) with high probability areas highlighted in green) rather than to perform operations such as thresholding or non-maximal suppression which involve loss of information.

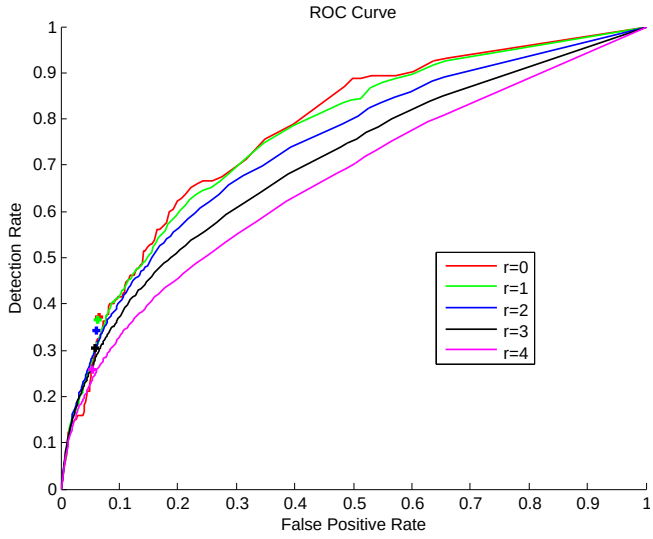


(a)

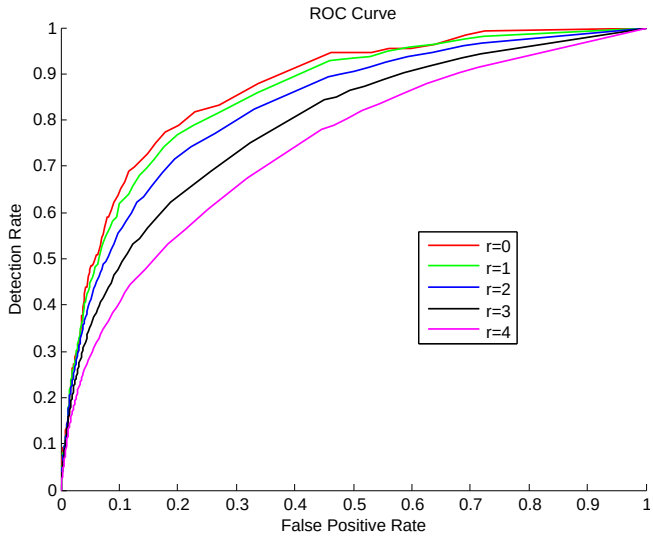


(b)

Fig. 4. In Figure 4(a) we present the performance of the linear classifier for the different values of r; the locations indicated by a cross in the corresponding color show the performance in terms of FPR-TDR of the classifier output which has been regularized using a MRF approach. Figure 4(b) presents the performance for the same base classifier where we introduced the temporal consistency check.



(a)



(b)

Fig. 5. In Figure 5(a) we present the performance of the HIK classifier for the different values of r; the locations indicated by a cross in the corresponding color show the performance in terms of FPR-TDR of the classifier output which has been regularized using a MRF approach. Figure 5(b) presents the performance for the same base classifier where we introduced the temporal consistency check.

Discussion The performance of the classifiers underlines the fact that discriminative learning may be employed, even in extremely cluttered environments, to provide target cues to tracking algorithms. Depending on the trade-off that we prefer between the risk of target miss and the presence of false positives, the ROC curves should assist in finding the appropriate decision boundaries.

One important observation that underlines the applicability of this approach compared to other analysis strategies in high-density environments is that we do not make any assumption about the presence of salient objects, either in terms of color or shape. The strategy we presented exploits only the implicit saliency of the head shape compared to its immediate environment. Although the discriminative characteristic of the classifier is moderate, the overall performance indicates that this characteristic is sufficient for target identification.

Secondly, we note the good performance of the HIK classifier when the temporal consistency of the detection is taken into account; this classifier seems to be more sensitive to fine variations in the descriptor content and therefore it benefits more from a regularization framework and also from a consistent training database.

Finally, we note that for increasing values of r the performance decreases relatively slowly at the beginning, which shows that this family of classifiers has a stable response above the target area. The classifier performance as well as the testing accuracy could benefit from a pixel-level annotation of the targets but beside the fact that the effort required is significant, the approximation of the target area as a small disk region around the clicked pixels is consistent with the results. We also highlight the fundamental difficulty of building objective pixel-wise ground truth data for this range of detection tasks, which may further hint at the current lack of standardized high-density crowd data available for research.

6 Conclusion and future work

In this paper we investigate the applicability of discriminative learning strategies for detecting and initializing tracking targets in high-density crowds exhibiting extreme clutter and homogeneity. By avoiding approaches based primarily on thresholding and non-maximal suppression of the detections, we show that we can build consistent detection probability maps which present a plateau response in target locations. These maps are suitable for spatio-temporal regularization, and also for offering an application adapted compromise between the desired rate of false positive detections and the target miss rate.

As future work, we would like to carry out spatio-temporal regularization jointly in a MRF framework, as well as to estimate the limitation of the classifiers determined by changes in the appearance related to the topology. This would allow us ultimately to apply these methods jointly in multiple cameras and validate a probability density map using independent data sources.

Acknowledgement. K. Kiyani would like to acknowledge the Qatar QNRF under the grant NPRP 09-768-1-114.

References

1. Ferryman, J., Ellis, A.L.: Performance evaluation of crowd image analysis using the PETS2009 dataset. Pattern Recognition Letters **44** (2014) 3 – 15 Pattern Recognition and Crowd Analysis.
2. Helbing, D., Johansson, A., Al-Abideen, H.Z.: Dynamics of crowd disasters: An empirical study. Phys. Rev. E **75** (2007) 046109
3. Krausz, B., Bauckhage, C.: Loveparade 2010: Automatic video analysis of a crowd disaster. Comput. Vis. Image Underst. **116** (2012) 307–319
4. Zhan, B., Monekosso, D., Remagnino, P., Velastin, S., Xu, L.Q.: Crowd analysis: a survey. Machine Vision and Applications **19** (2008) 345–357
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01. CVPR '05, Washington, DC, USA, IEEE Computer Society (2005) 886–893
6. Zhao, T., Nevatia, R.: Bayesian human segmentation in crowded situations. In: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. Volume 2. (2003) II–459–66 vol.2
7. Zhao, T., Nevatia, R.: Tracking multiple humans in crowded environment. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. Volume 2., IEEE (2004) 406–413
8. Comaniciu, D., Meer, P., Member, S.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002) 603–619
9. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: In CVPR. (2005) 878–885
10. Rabaud, V., Belongie, S.: Counting crowded moving objects. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Volume 1. (2006) 705–711
11. Li, M., Zhang, Z., Huang, K., Tan, T.: Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. (2008) 1–4
12. Li, M., Bao, S., Dong, W., Wang, Y., Su, Z.: Head-shoulder based gender recognition. In: Image Processing (ICIP), 2013 20th IEEE International Conference on. (2013) 2753–2756
13. Ye, Q., Gu, R., Ji, Y.: Human detection based on motion object extraction and headshoulder feature. Optik - International Journal for Light and Electron Optics **124** (2013) 3880 – 3885
14. Wang, S., Zhang, J., Miao, Z.: A new edge feature for head-shoulder detection. In: Image Processing (ICIP), 2013 20th IEEE International Conference on. (2013) 2822–2826
15. Ali, S., Shah, M.: A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on. (2007) 1 –6
16. Ali, S., Shah, M.: Floor fields for tracking in high density crowd scenes. In: Proceedings of the 10th European Conference on Computer Vision: Part II. ECCV '08, Berlin, Heidelberg, Springer-Verlag (2008) 1–14
17. Moore, B.E., Ali, S., Mehran, R., Shah, M.: Visual crowd surveillance through a hydrodynamics lens. Commun. ACM **54** (2011) 64–73

18. Idrees, H., Warner, N., Shah, M.: Tracking in dense crowds using prominence and neighborhood motion concurrence. Image and Vision Computing **32** (2014) 14 – 26
19. Aghajan, H., Cavallaro, A.: Multi-Camera Networks: Principles and Applications. Academic Press (2009)
20. Javed, O., Shah, M.: Automated Multi-Camera Surveillance: Algorithms and Practice. Volume 10 of The International Series in Video Computing. Springer (2008)
21. Eshel, R., Moses, Y.: Tracking in a dense crowd using multiple cameras. International Journal of Computer Vision **88** (2010) 129–143
22. Wang, X.: Intelligent multi-camera video surveillance: A review. Pattern Recogn. Lett. **34** (2013) 3–19
23. Maji, S., Berg, A., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. (2008) 1–8
24. Lin, H.T., Lin, C.J., Weng, R.: A note on platts probabilistic outputs for support vector machines. Machine Learning **68** (2007) 267–276