UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

IGOR RODRIGUES DE ALMEIDA

Crowd Analysis Using Local Neighborhood Coherence

Thesis presented in partial fulfillment
of the requirements for the degree of
Doctor of Computer Science

Advisor: Prof. Dr. Claudio R. Jung

Porto Alegre
February 2020

CIP — CATALOGING-IN-PUBLICATION

Rodrigues de Almeida, Igor

Crowd Analysis Using Local Neighborhood Coherence / Igor Rodrigues de Almeida. – Porto Alegre: PPGC da UFRGS, 2020.

111 f.: il.

Thesis (Ph.D.) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2020. Advisor: Claudio R. Jung.

1. Human Crowds. 2. Computer Vision. 3. Event Detection. I. Jung, Claudio R.. II. Título.

“Any large-scale human cooperation is rooted in common myths that exist only in people’s collective imagination.”

— YUVAL NOAH HARARI

ABSTRACT

Large numbers of crowd analysis methods using computer vision have been developed in the past years. This dissertation presents an approach to explore characteristics inherent to human crowds – proxemics, and neighborhood relationship – with the purpose of extracting crowd features and using them for crowd flow estimation and anomaly detection and localization. Given the optical flow produced by any method, the proposed approach compares the similarity of each flow vector and its neighborhood using the Mahalanobis distance, which can be obtained in an efficient manner using integral images. This similarity value is then used either to filter the original optical flow or to extract features that describe the crowd behavior in different resolutions, depending on the radius of the personal space selected in the analysis. To show that the extracted features are indeed relevant, we tested several classifiers in the context of abnormality detection. More precisely, we used Recurrent Neural Networks, Dense Neural Networks, Support Vector Machines, Random Forest and Extremely Random Trees. The two developed approaches (crowd flow estimation and abnormality detection) were tested on publicly available datasets involving human crowded scenarios and compared with state-of-the-art methods.

Keywords: Human Crowds. Computer Vision. Event Detection.

Análise de Multidões Usando Coerência de Vizinhança Local

RESUMO

Métodos para análise de ambientes de multidões são amplamente desenvolvidos na área de visão computacional. Esta tese apresenta uma abordagem para explorar características inerentes às multidões humanas - comunicação proxêmica e relações de vizinhança - para extrair características de multidões e usá-las para estimativa de fluxo de multidões e detecção e localização de anomalias. Dado o fluxo óptico produzido por qualquer método, a abordagem proposta compara a similaridade de cada vetor de fluxo e sua vizinhança usando a distância de Mahalanobis, que pode ser obtida de maneira eficiente usando imagens integrais. Esse valor de similaridade é então utilizado para filtrar o fluxo óptico original ou para extrair informações que descrevem o comportamento da multidão em diferentes resoluções, dependendo do raio do espaço pessoal selecionado na análise. Para mostrar que as características são realmente relevantes, testamos vários classificadores no contexto da detecção de anormalidades. Mais precisamente, usamos redes neurais recorrentes, redes neurais densas, máquinas de vetores de suporte, floresta aleatória e árvores extremamente aleatórias. As duas abordagens desenvolvidas (estimativa do fluxo de multidões e detecção de anormalidades) foram testadas em conjuntos de dados públicos, envolvendo cenários de multidões humanas e comparados com métodos estado-da-arte.

Palavras-chave: Multidões Humanas. Visão Computacional. Detecção de Eventos.

LIST OF ABBREVIATIONS AND ACRONYMS

CDA Cascade deep AutoEncoder

CNN Convolution neural network

CRM Conditional random field

CUHK Chinese University of Hong Kong

DCT Discrete cosine transform

DNN Dense nerural network

DT Dynamic textures

DTM Dynamic texture mixing

ET Extremely randomized trees

GAN Generative adversarial net

GMM Gaussian mixture model

HDP Hierarchical dirichlet process

HFST High-frequency and spatio-temporal

HMM Hidden Markov model

HOG Histogram of oriented gradients

ICS Image coordinates system

KLT Kanade–Lucas–Tomasi

LDA Latent dirichlet allocation

LSTM Long short-term memory

LUT LookUp table

MDT Mixture of dynamic textures

MLP Multilayer perceptron

NL Non-linear

PCA Principal component analysis

PWC Pyramid, warping, and cost volume

RBF Radial basis function

RF Random forest

RNN Recurrent neural network

ROI Regions of interest

SCD Structural context descriptor

SFM Social force model

SRC Sparse reconstruction cost

SVM Support vector machine

SVOI Spatial-temporal volumes of interest

TN True negative

TP True positive

UCF University of Central Florida

UCSD University of California San Diego

WCS World coordinates system

LIST OF FIGURES

LIST OF TABLES

CONTENTS

RESUMO EXPANDIDO

Nesta tese propõe-se abordar a análise de multidões humanas - um conjunto de pessoas que se encontram em cena. Este é um tópico ainda em aberto na área de visão computacional, apesar da existência de diversos métodos propostos voltados a este tema, pricipalmente devido à grande variedade de abordagens e desafios intrínsecos a análise de multidões.

No Capítulo 1 apresentamos as motivações em desenvolver métodos de análise automatizada de multidões - problemas decorridos de situações de pânico e a incapacidade das pessoas em se manter atento durante muito tempo a vídeos de vigilância - e conceitos utilizados durante a tese como o estudo de distâncias interpessoais e conceitos quanto a densidade de multidões e as diferenças entre as dificuldades quando analisadas multidões pouco e muito densas. Neste primeiro capítulo também é contextualizado o problema abordado na tese e quais as hipóteses assumidas neste trabalho: análise de multidões com alta densidade e distorções na movimentação quando adotadas apenas visualização em coordenadas de imagem. Dentro deste contexto apresentamos a estimativa de uma medida da coerência local para cada pixel da imagem com o intuito desta auxiliar na análise da multidão, com foco na estimativa do fluxo e a detecção de anomalias locais, como objetivo desta tese.

No Capítulo 2 é demonstrado um levantamento de trabalhos que se propõem a abordar conceitos próximos dos apresentados nesta tese, elencando três conceitos relacionados: estimativa do fluxo da multidão, fluxo ótico em contextos genéricos, e análise de multidões. Quanto à estimativa do fluxo da multidão, analisamos diversos trabalhos focados em análise de multidões, além de avaliar quais métodos utilizados na obtenção da informação do movimento das pessoas que compõem a multidão. Como um dos métodos mais utilizados para essa tarefa é o fluxo ótico, realizamos uma análise das técnicas disponíveis para o cálculo do fluxo ótico, sendo este um problema clássico na visão computacional e com ampla utilização não apenas em análise em multidões. Na sessão sobre análise de multidões, focamos primeiro em quais tipos de abordagens são utilizadas quando analisa-se multidões - microscópica, macroscópica ou mista - trazendo, além das dificuldades e benefícios quando se utiliza cada abordagem, alguns métodos propostos com base em cada uma delas. Focamos também nas formas possíveis de treinamento que os métodos de detecção de anomalias costumam utilizar: treinamento utilizando apenas exemplos de comportamento normal, treinando com exemplos normais e anormais ou

mesmo utilizando técnicas que dispensam o treinamento com exemplos prévios.

O Capítulo 3 é focado no método proposto de estimativa de fluxo da multidão, onde utilizamos fluxo ótico combinado com remoção de plano de fundo para obtenção do fluxo ótico apenas em pixels que pertençam a multidão e com auxilio de câmeras calibradas. Este fluxo é convertido do domínio da imagem para domínio do mundo, atenuando problemas quanto à perspectivas da câmera. Para cada pixel pertencente a multidão, a coerência da movimentação deste em relação a um vizinhança é calculada utilizando distância de Mahalanobis, com tamanho da vizinhança determinado através das distâncias interpessoais. Por fim, esta medida de coerência é utilizada para aproximar o fluxo do pixel em análise com a média do fluxo da vizinhança. Com este processamento no fluxo ótico original conseguimos obter um fluxo mais coerente ao que esperamos de um fluxo de uma multidão.

Os experimentos realizados para avaliar o método consistiram em analisar o tempo de execução deste processamento em relação as técnicas de fluxo ótico geralmente adotadas em estimativa do fluxo da multidão, análise qualitativa quanto ao fluxo estimado, análises quantitativas do método adotando tanto uma abordagem de advecção de partículas quanto uma abordagem de detecção de eventos, utilizando um método de detecção de alteração de comportamento de multidões que originalmente utiliza fluxo ótico para estimar o movimento da multidão e adicionando o método proposto de estimativa do fluxo da multidão para analisar se há melhoras no resultado original.

O segundo método proposto nesta tese é apresentado no Capítulo 4: detecção de anomalias locais em multidões. Para isto é utilizada uma abordagem semelhante à utilizada na estimativa do fluxo da multidão, porém ao invés de calcular uma medida de coerência do movimento do pixel com uma única vizinhança, esta medida é calculada para diferentes tamanhos de vizinhança. Este vetor de coerência local é utilizado como dado de entrada para classificadores que são treinados com o objetivo de classificar o pixel como normal ou anormal. Neste método utilizamos uma abordagem de realizar a análise não quadro-a-quadro mais de uma cena, ou seja, de um conjunto de quadros onde a multidão possui movimento estacionário.

Com o objetivo de avaliar o desempenho do método proposto de detecção de anomalias, realizamos experimentos que são demonstrados também no quarto capítulo, dividindo os experimentos em três abordagens: avaliação dos classificadores, exploração dos tamanhos da cenas utilizadas e análise em multidões pouco densas. Na análise de classificadores foram utilizados cinco classificadores: redes neurais recorrentes, perceptron

multicamadas, máquina de vetores de suporte, floresta aleatória e árvores extremamente aleatórias, comparando os resultados obtidos e tempos de execução de cada classificador. Na análise quanto ao tamanho da cenas utilizamos 3 diferentes tamanhos e analisamos o resultado de dois classificadores - máquina de vetores de suporte e redes neurais recorrentes - quanto a variação do tamanho da cena e quais as consequências destas variações.

O último experimento apresentado no quarto capítulo foi a análise do método proposto em cenas que apresentam multidões pouco densas, trazendo os resultados obtidos utilizando máquina de vetores de suporte e apontando quais as dificuldades que o método encontra ao ser utilizado neste cenário. É apresentada também a comparação com métodos estado-da-arte que focam em detecção de anomalias em multidões e que apresentam seus resultados em base de dados de multidões pouco densas.

Por fim, no Capítulo 5 discorremos sobre as conclusões obtidas durante o desenvolvimento deste trabalho com base nos resultados encontrados e os objetivos propostos. Também analisamos quais futuras abordagens são possíveis com base no que foi desenvolvido nesta tese e quais questões se mantêm em abertas a análise.

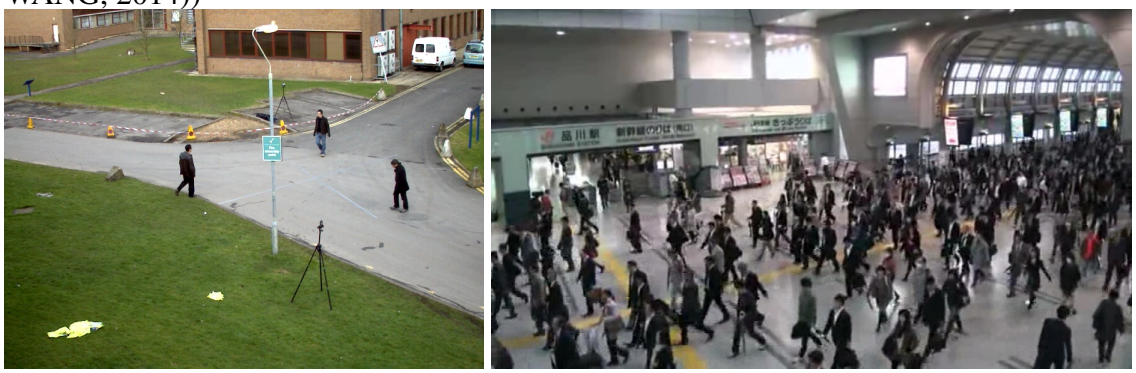
1 INTRODUCTION

Surveillance systems are becoming increasingly ordinary in urban life. Everyday, more and more security cameras are installed with the aim of ensuring social well-being. With growing this demand and the rise of computational power, automated or semi-automated analysis systems for surveillance environments have been an area widely studied by the computer vision community. The analysis of monitored environments has originated several research topics such as pedestrian detection and tracking, crowd flow estimation, detection of individual abnormal behaviors, such as theft, and detection of global abnormal behavior such as panic situations in crowded scenes.

It is possible to divide surveillance scenes in two main groups that require different algorithmic and mathematical tools: scattered pedestrian scenes and crowds scenes, as illustrated in Figure 1.1. While in scattered pedestrian scenes individual tracking and analysis techniques can be applied, crowds scenes typically demand a more complex processing pipeline since it is difficult to identify each individual pedestrian in the scene.

Since the theme of crowd analysis is very broad, this dissertation is focused in the following specifi problems related to high-density crowds: crowd flow estimation and detection of local anomalies in crowds using flow information. Real crowds can be analyzed essentially in two different "resolutions": microscopic and macroscopic (MEHRAN; OYAMA; SHAH, 2009). In microscopic approaches, people are analyzed as discrete individuals, and this information is used to infer the behavior of the crowd. In the macroscopic approach, the crowd is analyzed as a single entity, and the members are not analyzed individually. A combination of micro and macroscopic approaches can be used by keeping

Figure 1.1: Scattered pedestrian scene and crowd scene in monitored environments. (Source: PETS2009 (FERRYMAN; SHAHROKNI, 2009) and CUHK (SHAO; LOY; WANG, 2014))



the crowd as a homogeneous mass, but considering at the same time an internal force, or by maintaining the characteristics of people while keeping a general view of the entire crowd (MEHRAN; OYAMA; SHAH, 2009).

The choice between a microscopic or macroscopic methodology directly impacts the computer vision requirements. For instance, main flows of people are explored when using a macroscopic approach; in microscopic approaches, however, people must be tracked individually, which is an increasing challenge as crowd density escalates. In fact, many crowd analysis techniques (LIM et al., 2014; ALMEIDA; JUNG, 2013; SOLMAZ; MOORE; SHAH, 2012) use optical flow as an estimate of crowd motion. Despite the existence of many optical flow techniques (see (FORTUN; BOUTHEMY; KERVRANN, 2015) and (TU et al., 2019) for a survey), they are mostly generic-purpose methods (i.e. they try to find local correspondences in generic scenarios). However, people in crowds typically move in an orderly manner, and neighbors in a crowd usually have similar movement patterns (speed and orientation).

When deciding to study human behavior one must also observe how individuals interact with each another. Edward Hall, an American anthropologist, states that "People like to keep certain distances between themselves and other people or things. And this invisible bubble of space consisting of the *territory* of each person is one of the fundamental dimensions of modern society" (HALL, 1966). As a consequence, the distance between two people provides cues about the type of relationship between them.

The term *proxemics* proposed in 1963 by Hall is still used for studying individual spaces. He presented the interpersonal distances divided into four zones, each one having a far and a close phase, as shown in Figure 1.2. More precisely, these distances and their corresponding expected interactions are given by:

- Intimate distance: the presence of the other person is undeniable, and usually with constant physical contact. Distance from 0 to 0.45 meters. This can be divided in close phase (0 to 0.15 meters) and far phase (0.15 to 0.45 meters).

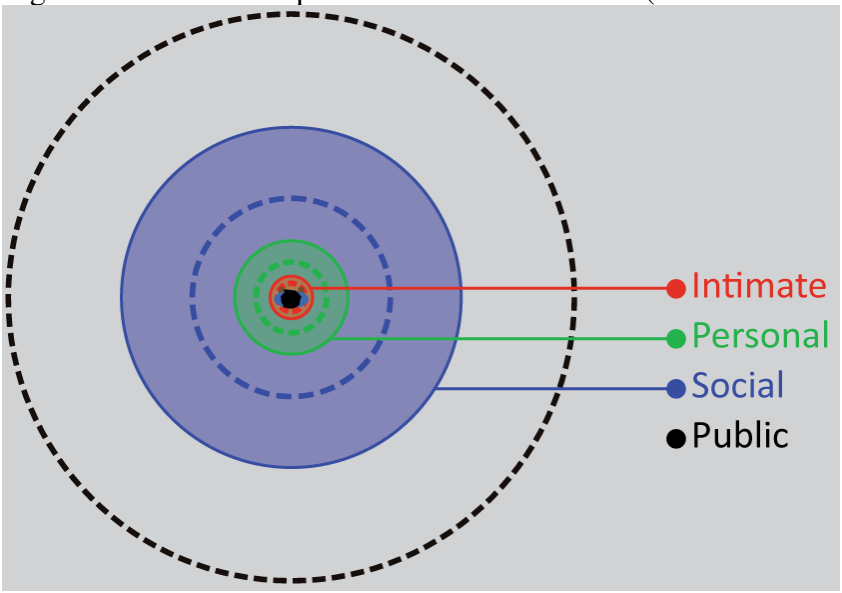
- Personal distance: at this distance the individuals are probably close friends, or have some intimacy degree between them. It is also used to determine a " bubble " that separates one individual from the other. Distance from 0.45 to 1.2 meters. This can be divided in close phase (0.45 to 0.75 meters) and far phase (0.75 to 1.2 meters).

- Social distance: at this distance no one expects to touch or be touched by another person, unless there is an effort directed at it. Distance from 1.2 to 3.6 meters.

This can also be divided in close phase (1.2 to 2.1 meters) and far phase (2.1 to 3.6 meters).

- Public distance: at this distance there is no involvement between individuals. Distance of 3.6 meters or more that can be divided in close phase (3.6 to 7.6 meters) and far phase (more then 7.6 meters)

Figure 1.2: Hall's interpersonal distances of man. (Source: author)



The psycho-social distances proposed by Hall have been used to assess risk situations in real crowds, and can also be applied to explain how global crowd behaviours changes based on the local crowd density. In fact, these distances have been used in the crowd simulation literature to develop realistic simulators, such as (HELBING; MOLNÁR, 1995; KAUP et al., 2006), but to our knowledge they have not been used to obtain realistic crowd flows from video sequences.

1.1 Motivation

Video-based ambient surveillance is not a simple task, due to the large amount of data provided by the cameras, and in many cases, the number of cameras to be monitored is much larger than the number of operators, as can be seen in Figure 1.3. As shown in a study by the US National Institute of Justice Green (1999), a human observer has difficulty to keep up watching the generated videos after only 20 minutes, when most of the tested individuals got below-acceptable attention rates. In this scenario, automated or

semi-automated surveillance methods can be employed to facilitate and prevent problem situations by signaling to a human observer the occurrence of something unusual.

Figure 1.3: Integrated Command Center (CEIC) of Porto Alegre city (Source: CEIC Porto Alegre - Photo: Ricardo Giusti/PMPA)

Automatic or semi-automatic surveillance environments have been widely researched in the field of computer vision, since good detectors can quickly identify a problem that has occurred. A particular sub-area of behavior detection involves the processing of dense scenes, and the analysis of crowd behavior has been an active research area in the realm of computer vision (CHEN; WANG; LI, 2017). Sporting events, nightclubs, shopping centers and music concerts are examples of situations in which crowds often clutter inside and outside.

Despite recent advances, the crowd analysis research area is still open, with several recent methods and the possibility of improving current results (WU et al., 2017; MARSDEN et al., 2017; WANG et al., 2018). This makes it an interesting problem, because in addition to practical utility, there are no methods with definitive results. In particular, perspective distortions are a challenge to develop methods that are generic to several camera setups, and also to explore relevant real-world information (such as proxemics). The main goal of this work is to explore behavioral patterns of people within local neighborhoods to exploit methods for extracting crowd features in surveillance environments, and to explore these features for abnormal behavior detection.

1.2 Problem Description and Hypothesis

In order to explore Hall's personal distances, we must know how to map from image to world coordinates. To avoid online calibration schemes, we assume that the scene is captured by a static camera, so that must be calibrated only once. More precisely, we assume that the ground plane homography is known (if not, it can typically be estimated based on the scene), therefore all analysis can be performed based on world coordinates. In fact, the use of world coordinates is expected to make the technique more generic and less dependent on the camera pose, since the same exact scenario captured by two different cameras might present video sequences with significant differences, as illustrated in Figure 4.1.

We also assume that the motion patterns are similar within local neighborhoods of the captured images, consistent with the macroscopic view of crowded scenes. In fact, neighboring pixels are either part of a single person, or nearby persons who, by proximity, should also contain similar behaviors. Note that smaller neighborhoods should be related to stronger interpersonal relationships, as illustrated in Figure 1.2. Hence, exploring neighborhoods with varying sizes might allow the analysis of different interpersonal relationships.

Finally, in the context of abnormality detection, this work will focus on structured crowds in which there is typically a high density of people with a "stationary" behavior within a time window. By stationary we mean that the local densities and motion patterns are approximately constant within the time window, such as in entrance and exits of sports stadiums, marathons and subway stations, to name a few. Some examples of these scenarios can be seen in Figure 1.5.

Given a stationary crowd captured by a calibrated static camera, this work aims to explore neighborhood information to: i) calculate the crowd flow, by post processing an input optical flow based on neighborhood information, and ii) detect local anomalies - stationary behaviors that differ from the neighborhood - by analyzing each point in relation to different neighborhood areas based on proxemics.

Figure 1.4: Same scenario filmed by two cameras. The distance in image coordinates between people are different in each scene. (Source: PETS2009 (FERRYMAN; SHAHROKNI, 2009))

Figure 1.5: Examples of scenes with structured crowds. (Source: CUHK (SHAO; LOY; WANG, 2014) and Marathon (LIM et al., 2014))

1.3 Goals

1.3.1 Main goals

The main goal of this dissertation is to explore neighborhood information in stationary crowds captured by a calibrated static camera to extract crowd behavior cues. More precisely, the aim is to explore the ground plane homography to estimate the crowd motion in the world coordinate system, and then explore Hall's proxemics (HALL, 1966) using different interaction radii to infer the local coherence of the estimated motion and use this data to crowd analysis aiming to estimate crowd flow and detect local anomalies.

1.3.2 Specific goals

To achieve the main goal, we intend to tackle the following specific goals:

- Based on neighborhood movement pattern, develop a post-processing step that can be coupled to generic optical flow techniques for extracting coherent crowd flows. This post-processing step should be simple to implement and with fast execution so as not to add too much complexity into the stage of acquiring the optical flow of crowd analysis methods;
- Explore mutiscale neighborhood information to detect local anomalies in highly dense human crowds, evaluating different classifiers for the task.
- Perform qualitative analysis of results using the detect local anomaly method and compare them with state-of-art methods.

1.4 Contributions

The main focus of this work is the use of the psycho-social interactions expected to arise in a real crowd, encoded by the proxemics presented by Hall (1966), to obtain reliable information of a real crowd captured by a stationary calibrated camera. More precisely, we have highlighted below the two main contributions of this dissertation:

- The development of a post-processing approach based on proxemics to obtain reliable crowd flows that can be coupled to any generic-purpose optical flow approach
- The introduction of local crowd flow features computed with multiple personal regions (with varying radii), which can be used for anomaly detection.

1.5 Chapters Organization

The remainder of this dissertation is organized as follows: Chapter 2 reviews the state-of-the-art in optical flow techniques, crowd flow extraction, crowd analysis and anomaly detection in human crowds. The proposed crowd flow estimation method is introduced in Chapter 3, which also analyzes the results of this technique. Chapter 4 discusses the proposed anomaly detection method and inspects the experiments and compar-

isons of this method with state-of-art anomaly detection methods. Finally, the conclusions and contributions are discussed in Chapter 5.

2 RELATED WORK

In crowded scenes, it is difficult to obtain the exact position of each individual separately, due mostly to clutter and occlusions. Furthermore, tracking individuals is an additional problem, since people in a crowd are typically close to each other and present similar motion patterns, generating appearance and motion ambiguity.

Following the macroscopic approach to deal with crowds, several methods extract motion cues from local patches or the whole crowd. Correctly estimating the crowd flow of a scene is useful for identifying main flows of people, and it can be used to detect characteristics that commonly arise in crowded scenes (such as bottlenecks), or to identify unusual/abnormal events. In fact, crowd motion-based anomaly or event detection methods typically extract features from the estimated crowd flow, and then use a classifier as the final stage of the pipeline.

Since the scope of this paper is to deal with denser and structured crowds, this chapter revises existing approaches for crowd flow estimation and event detection. More precisely, Section 2.1 tackles crowd motion estimation methods, while generic-purpose optical flow methods are revised in Section 2.2. Section 2.3 evaluates existing approaches for crowd analysis, focusing on methods that explore flow information. Finally, Section 2.5 presents the conclusions of this chapter.

2.1 Crowd Motion Estimation

Motion cues can be extracted in a variety of manners in the context of crowd analysis. Cheriyadat and Radke (2008) presented a method for detecting dominant motion patterns (as illustrated in Figure 2.1) within a dense crowd based on a hierarchical implementation (BOUGUET, 2001) of the KLT feature tracker (LUCAS; KANADE, 1981). Zhao, Zhang and Huang (2017) proposed an approach to detect crowd groups and to learn semantic regions with a unified hierarchical clustering framework. They initially cluster tracklets extracted using the KLT feature tracker. Representative tracklets are further used to learn crowd behaviors. Lim et al. (2014) presented a method to detect crowd saliency points, using a dense optical flow method (LIU et al., 2008) to calculate the motion of the crowd.

Many approaches use crowd flow estimation to detect events in crowds. For example, the method presented by Wu, Moore and Shah (2010) aims to detect and localize

Figure 2.1: Example of dominant motion in a crowded frame. The yellow arrowed lines indicate the dominant motions of the crowd. (Source: (CHERIYADAT; RADKE, 2008))

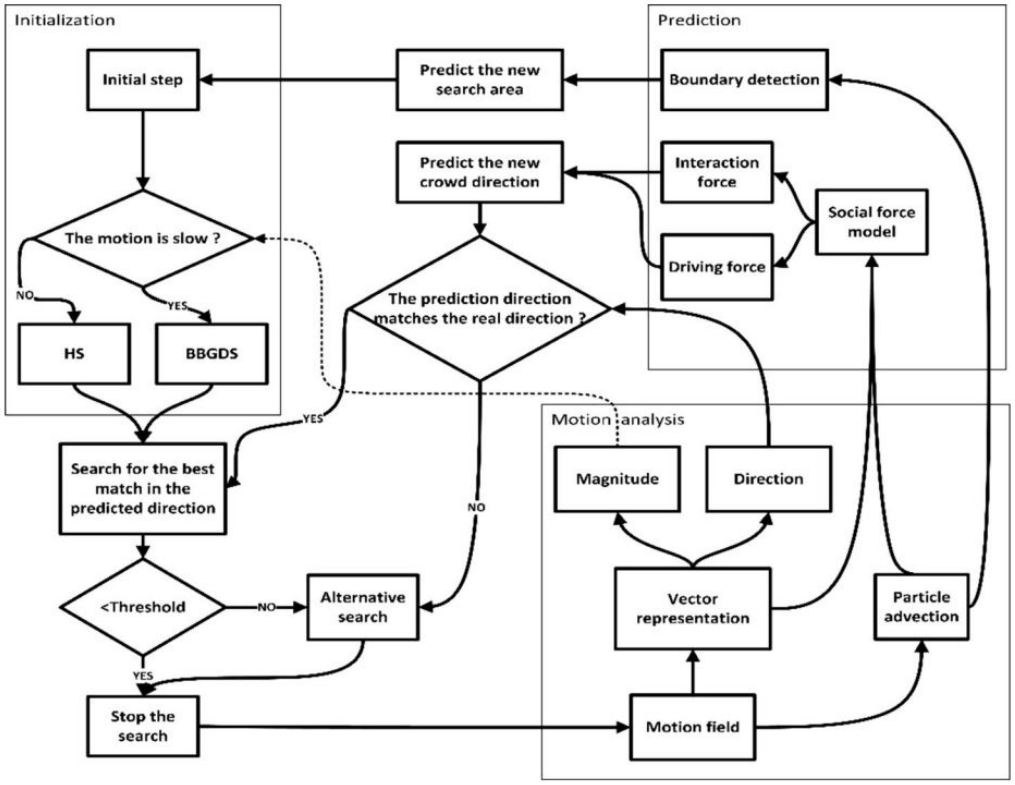
anomalies in complex and crowded sequences. This method uses particle advection based on optical flow, and particle trajectories are clustered to obtain representative trajectories for a crowd flow. Abdallah, Gouiffès and Lacassagne (2016) presented a system for abnormal event detection and categorization. They extract the local motion of foreground pixel using the KLT feature tracker (LUCAS; KANADE, 1981), and evaluate changes in the detected flow to identify abnormal events.

Solmaz, Moore and Shah (2012) explored concepts related to the stability of a dynamical system to detect pre-determined events in a crowd, based on the optical flow (LUCAS; KANADE, 1981) of the scene. Mehran, Oyama and Shah (2009) explored a Social Force Model (SFM) to detect and localize unusual behavior in crowded scenes. In their approach, the interaction of particles guided by a space-time average of the optical flow is estimated using an SFM, and a bag of features approach is adopted for unusual event detection.

Kajo, Kamel and Malik (2017) proposed an algorithm to measure the motion of a crowd based on block-based matching, particle advection, and social force model. More precisely, they initially estimate motion using a block-based matching approach in each frame and create the corresponding motion field. The social force model is then used

to predict the direction of motion vectors, obtained by particle advection process. A schematic view of this method can be observed in Figure 2.2.

Figure 2.2: Flowchart of the proposed method by Kajo et al. (Source: (KAJO; KAMEL; MALIK, 2017))

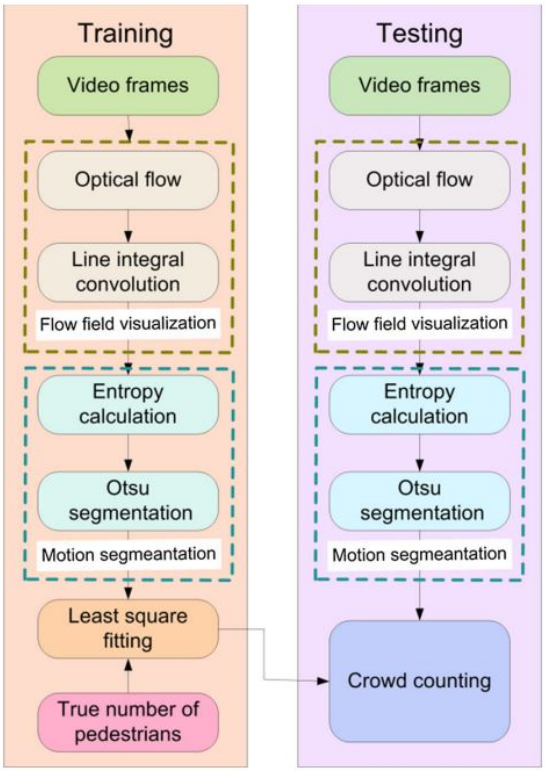


Chen and Huang (2011) used optical flow (LUCAS; KANADE, 1981) to cluster human crowds into groups in an unsupervised manner using a novel approach called adjacency-matrix based clustering. Each cluster is characterized based on the chosen SFM, and unusual crowd events are detected when the orientation of a crowd is abruptly changed or when interactions within the crowd are not similar to the predicted value.

Zhang et al. (2015) tackled the problems of crowd counting and motion segmentation,as shown in their pipeline in Figure 2.3. They justified the use of the Horn-Schuck optical flow method (HORN; SCHUNCK, 1981), arguing that the crowd moves like a fluid.

Although using generic optical flow methods is one of the most popular techniques for estimating crowd motion, there are also other alternatives. For instance, Dee and Caplier (2010) used three techniques together to obtain scene information: HOG (Histogram of oriented gradients) based on head detector (DALAL; TRIGGS, 2005), face detector of Viola-Jones (VIOLA; JONES, 2001) and the feature tracker KLT (SHI et al., 1994). However, their results obtain the flow compute at image “cells”, not at every pixel

Figure 2.3: Pipeline proposed by Zhang et al. (Source: Zhang et al. (2015))



as in typical optical flow methods.

2.2 Optical Flow

Optical flow is a classical problem in computer vision that has been applied to a variety of issues ranging from video stabilization, dense stereo matching, motion segmentation and others. Although there are a few survey papers on the subject (FORTUN; BOUTHEMY; KERVRANN, 2015; BEAUCHEMIN; BARRON, 1995), they do not deal with specific crowd aspects.

More recently, Kajo, Malik and Kamel (2016) presented an evaluation of optical flow methods in the context of crowd analysis, categorizing them into two main classes based on the regularization type (KAJO; MALIK; KAMEL, 2016): feature-based methods and variational methods.

Variational methods take into consideration optical flow solutions of neighboring pixels and impose smoothness assumptions on the flow field. Feature-based methods compute the optical flow solution for each pixel and its neighborhood independently from the other pixels in the image. There are many proposed methods for estimating optical

flow based on either classes, and a few of them will be discussed next: feature-based methods first, and later variational methods.

Lucas and Kanade (1981) presented a local matching approach to obtain the optical flow. The main assumption is that the flow is constant within local regions of the image, modeling the problem as a least-squares minimization scheme. Shi et al. (1994) extended this concept for local affine motion, and introduced a feature selection scheme for particle tracking. A pyramidal (multi-scale) version of this method, which allows larger displacement and faster computation, was introduced in (BOUGUET, 2001).

Farnebäck (2003) presented a two-frame motion estimation algorithm. The central idea is to approximate each local neighborhood of both frames by quadratic polynomials, and models the local flow as an ideal translation of the polynomials. The estimated local displacement is then obtained by matching these two polynomial regions. Ranftl, Bredies and Pock (2014) proposed a non-local extension of the popular second-order Total Generalized Variation, which favors piecewise affine solutions and allows to incorporate soft-segmentation cues into the regularization term. These properties make this regularizer especially appealing for optical flow estimation, since it offers accurately localized motion boundaries and allows to resolve ambiguities in the matching term. They also propose a robust matching term to illumination and scale changes.

Sun, Roth and Black (2010) presented a method to estimate the optical flow of a scene that is a combination of classical flow formulations and more modern optimization and implementation techniques. They derive a new objective that formalizes the median filtering heuristic, which includes a nonlocal term that robustly integrates flow estimates over large spatial neighborhoods. They extended their work and provided an overview of current optical flow practices in (SUN; ROTH; BLACK, 2014), and an example of their result is shown in Figure 2.4. Revaud et al. (2015) proposed an approach for optical flow estimation targeted at large displacements with significant occlusions. They also propose an approximation scheme for the geodesic distance to allow fast computation without loss of performance. Recently Lavín-Delgado et al. (2020) presented a method for optical flow estimation based on the Classic+NL algorithm. However, their model lies in the generalization of the Classic+NL from integer-order to fractional-order by using Caputo-Fabrizio derivative (CAPUTO; FABRIZIO, 2015).

Horn and Schunck (1981) presented a popular variational approach in the context of optical flow. They combined the errors produced by the classical optical flow equations with a smoothness penalty term given by the gradient magnitudes.

Figure 2.4: Median filtering over-smoothes the rifle in the “Army” sequence, while the proposed weighted non-local term preserves the detail. Results of (a) Classic++ (b) Classic+NL (Source: Sun, Roth and Black (2010))

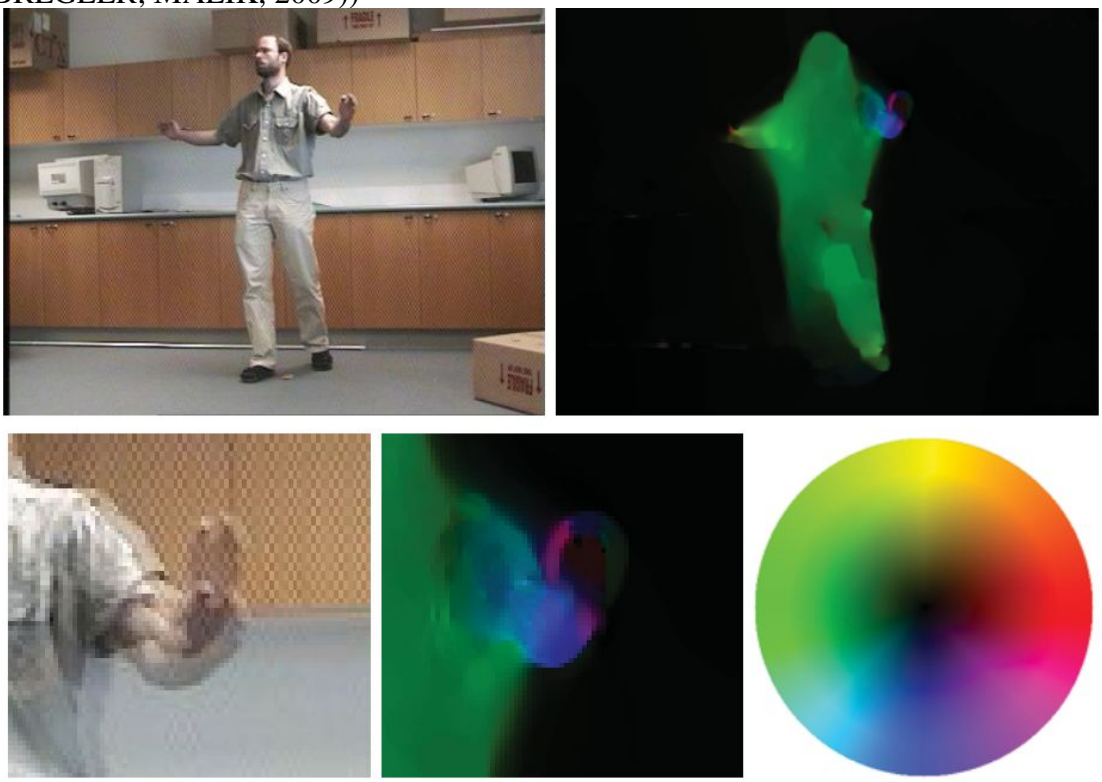
Yang and Li (2015) proposed a simple method for estimating dense optical flow fields. They fit a piecewise flow field to a variety of parametric models, where the domain of each piece is determined adaptively while maintaining at the same time a global inter-piece flow continuity constraint. They achieve this consistency by using a multi-model fitting scheme via energy minimization. Their energy takes into account both the piecewise constant model assumption and the flow field continuity constraint, enabling the proposed method to effectively handle both homogeneous motions and complex motions.

Brox and Malik (2011) presented an approach to estimate large motions of small structures (a common problem in several algorithms, shown in Figure 2.5), by integrating correspondences from descriptor matching into a variational approach. Their motivation is to use region correspondences to recover large displacements, embedded in a variational framework that leads to convex optimization.

The approach proposed by Weinzaepfel et al. (2013), termed DeepFlow, blends a matching algorithm with a variational approach for obtaining the optical flow. They proposed a descriptor matching algorithm tailored to the optical flow problem that allows boosting performances on fast motions. The matching algorithm builds upon a multi-stage architecture with six layers, interleaving convolutions and max-pooling - its outline can be observed in Figure2.6 -, a construction akin to deep convolutional nets. After DeepFlow, several methods using deep learning for optical flow estimation have been proposed, such as FlowNet (ILG et al., 2017) and SPyNet (RANJAN; BLACK, 2017) - which combines a classical spatial-pyramid formulation with deep learning.

More recently, Liu et al. (2019) presented a self-supervised learning approach for optical flow. This method is based on distilling reliable flow estimations from non-occluded pixels, and using these predictions to guide the optical flow learning for hallucinated occlusions. They further designed a Convolutional Neural Network (CNN) to utilize

Figure 2.5: The hand motion is not estimated correctly because the hand is smaller than its displacement relative to the motion of the larger scale structure in the background. The bottom row right image show the color map used to visualize flow fields in these images. Smaller vectors are darker and color indicates the direction. (Source: (BROX; BREGLER; MALIK, 2009))



temporal information from multiple frames for better flow estimation. Ren et al. (2019) also used the idea of multiple frames to improve the flow estimation; they presented a fusion approach for multiframe optical flow that benefits from longer-terms temporal cues, their proposed architecture is shown in Figure 2.7. Their method first warps the optical flow from previous frames to the current, thereby yielding multiple plausible estimates. It then fuses the complementary information carried by these estimates into a new optical flow field.

Despite the existence of a plethora of generic-purpose optical flow methods, there is no consensus on which approaches are better to deal with crowd flow estimation. To the best of our knowledge, the only work in this direction was presented by Kajo, Malik and Kamel (2016), who presented an experimental comparison of different optical flow algorithms for crowd analytics in surveillance systems. They used human aided annotation to estimate the optical flow of crowded scenes, and also the angular error in the comparison, concluding that the "Classic+NL" method (SUN; ROTH; BLACK, 2010) presented the best results among the compared approaches. However, this analysis approach is weak

Figure 2.6: Outline DeepFlow (Source: (WEINZAEPFEL et al., 2013))

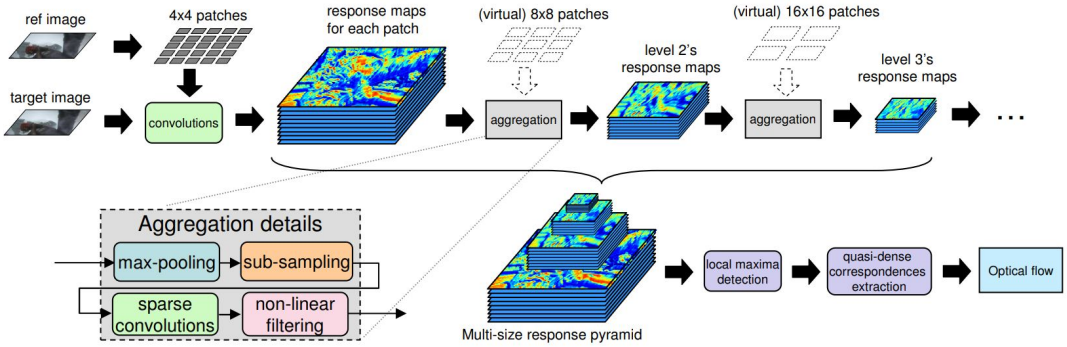
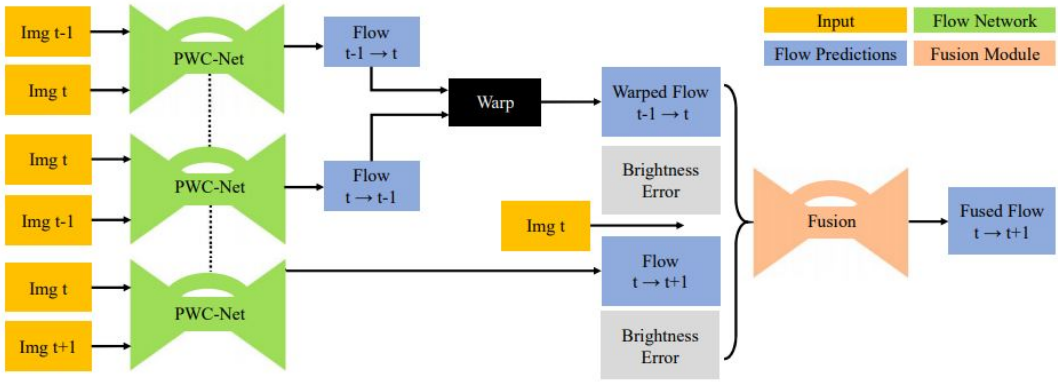


Figure 2.7: Architecture of the fusion approach for three-frame optical flow estimation. The dashed line indicates that the PWC-Nets share the same weights. PWC-Net can be replaced with other two-frame flow methods like FlowNetS (Source: (REN et al., 2019))



and does not reflect all complexity involved in optical flow estimation, been lesser statistical complete than the older review of Baker et al. (2011).

2.3 Crowd Analysis

We can deal with crowds in mainly two different views: microscopic and macroscopic. When a macroscopic approach is adopted, it is not necessary to track each individual, and strategies based on optical flow are popular; in microscopic approaches, other methods are required to determine the position and velocity of people individually, and one of these possible methods is head tracking through object detectors (FELZENSZWALB et al., 2010) as used in (RODRIGUEZ et al., 2011; ESHEL; MOSES, 2010). In (ESHEL; MOSES, 2010), the occlusion problem – caused due to the density of the scenario – is still treated using multiple cameras.

The problem of crowd analysis can also be approached in different ways, such

event detection, anomaly detection or change detection. The event detection approach try to detect events in the crowd, such as bottleneck, dispersion, lanes formation, arch and blocking. In such analysis, each specific event must be characterized or learned, which typically require annotated datasets containing the desired events. The anomaly detection approach seeks to detect something that deviates from "normality". In this case, the models are usually trained with examples containing normal behaviors, and the detection phase identifies what is different from this behaviors as abnormal. The definition of normal behavior is subjective and depends on several factors, and this dissertation will focus on motion patterns only (note that a person in a crowd carrying a gun would probably be considered abnormal, but the detection of objects requires higher-level processing when compared to motion estimation). Finally, approaches based on change detection try to detect if the motion pattern (local or global) of a crowd changes when compared to previous frames. This section starts with some generic crowd analysis techniques, and after we focus on anomaly detection methods. Furthermore, we focus on approaches that explore only motion information in their analysis.

Bertini, Bimbo and Seidenari (2012) proposed a microscopic approach toward crowd behavior analysis. Their method proposes to classify crowds behavior in normal pedestrian behavior and panic using a supervised method. They also proposed an unsupervised approach to anomaly detection in crowded scenes, which they defined as non-pedestrian entities (cyclists, skaters), by employing a local space-time descriptor.

In (HAQUE; MURSHED, 2010; MAHADEVAN et al., 2010) the authors adopt a macroscopic approach. They explore background removal (HAQUE; MURSHED; PAUL, 2008) to isolate foreground pixels, which are explored to classify crowd behavior in four events (meet, split, fight, runaway). For that purpose, an SVM is trained to detect the event in a captured scene. The approach of Mahadevan et al. (2010) uses DTM (dynamic texture mixing) (CHAN; VASCONCELOS, 2008) to detect anomalies in crowds scenes. The model is for normal crowd behavior - based on mixtures of dynamic textures - and outliers under this model are labeled as anomalies. The same approach was adopted by Li, Mahadevan and Vasconcelos (2013) in their work.Wang et al. (2012) use the information of the highest frequency computed by the wavelet transform (COHEN; DAUBECHIES; FEAUVEAU, 1992) to extract the characteristics of the crowd. They derive the high-frequency and spatio-temporal (HFST) features to detect the abnormal crowd behaviors in videos The features are applied to global and local abnormal crowd behavior detection. In global abnormal crowd behavior detection is used LDA to model normal scenes,

while in local abnormal crowd behavior the model of normal scenes used HMM. While Roshtkhari and Levine (2013), crowd information is extracted using densely sampled, spatial-temporal volumes of the video and this information is used to determine the dominant behavior of the crowd and detect anomalous behavior.

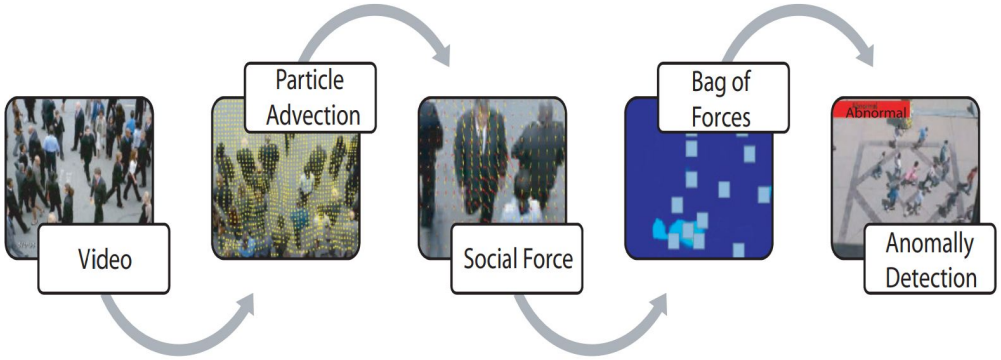
In the method proposed by Mehran, Oyama and Shah (2009), the authors mixture microscopic and macroscopic approaches exploring a social force model (HELBING; MOLNÁR, 1995) and particle advection (ALI; SHAH, 2007b) to extract the information of movement of the crowd, which is used for calculating social forces that are used to abnormal behavior detection in crowds scenes.

Detecting events is a challenging task, due to the difficulty of passing on semantic concepts of behavior to the computer. Lavee, Rivlin and Rudzsky (2009) performs a comprehensive review of the detection of events in videos, and divides the concept of event detection into two components: data abstraction and event modeling. The abstraction of data has been addressed in the previous subsections, and this subsection will treat event modeling.

Event modeling defines how the events covered are described and how to recognize them. Most methods typically use an annotated basis of events to learn a relationship between a data abstraction and the activities of the learning base. The ways in which this relationship is formulated varies from method to method. One of the possibilities is the one used by Jung, Hennemann and Musse (2008), in which they check the trajectories of objects to detect unusual events, using a database with usual trajectories in the test phase, and then verifying if the trajectories extracted from the scene are consistent with the results of the test phase. In this method, the trajectories of the test phase are grouped, using an extension of the method presented in (MUSSE et al., 2007), and then given a trajectory, that is compared with each of the clusters to define if it should be considered as non-usual.

Other techniques for event modeling were explored to detect unusual behavior of crowds. Mehran, Oyama and Shah (2009) (Figure 2.8 illustrates the pipeline of the method) used LDA (Latent dirichlet allocation) to classify events, noting that only scenes with usual behaviors were used in the training phase. The method of LDA was used in other approaches as a tool to detect abnormal behavior in crowds as in (SU et al., 2013). The SVM (Support Vector Machine) (HEARST et al., 1998a) is also widely used, and can be seen in methods such as (PATHAN; AL-HAMADI; MICHAELIS, 2010) and (CUI et al., 2011).

Figure 2.8: The summary of main steps of approach to detect abnormal behavior in crowds exploring social force model proposed by Mehran et. al. (Source: (MEHRAN; OYAMA; SHAH, 2009))



Andrade, Blunsden and Fisher (2006) presented a statistical method for event recognition, more precisely through a Hidden Markov model (HMM), extraction crowd features by background modelling and optical flow computation. Similarly, a statistical method is also used for event modeling in (PATHAN; AL-HAMADI; MICHAELIS, 2010), but CRFs (Conditional Random Fields) are used as an alternative to HMMs.

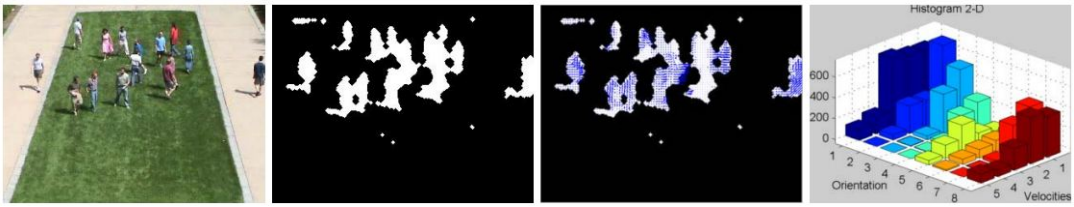
2.3.1 Anomaly Detection

Anomaly detection is a subarea of crowd analysis dedicated to detecting and/or localizing "abnormal" crowd behaviors, i.e., behaviors not expected in a crowd scene. There are several strategies to learn what is normal or abnormal: i) unsupervised (no manual label is assigned to training samples), ii) supervised trained with both classes (normal and abnormal), and iii) supervised trained with normal behaviors only, such that analyzed behaviors that do not belong are consider abnormal.

2.3.1.1 Unsupervised Methods

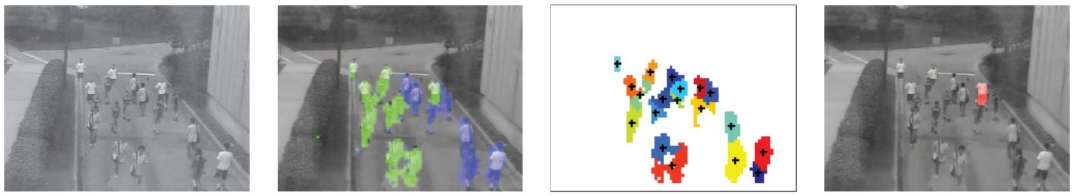
In unsupervised methods for anomaly detection, it is not necessary to have labeled data for training. In (ALMEIDA; JUNG, 2013) and (ALMEIDA et al., 2017) we developed a method to detect motion changes in human crowds that use optical flow and calibrated cameras to build 2-D histogram of speed and orientation of crowd motion in each frame. A temporal analysis comparing the histograms is performed, and a behavior change is detected when the temporal stability of the histogram decay abruptly. A pipeline of our approach is shown in Figure 2.9

Figure 2.9: Pipeline of the method to build 2-D histogram in crowd scenes (Source: (ALMEIDA et al., 2017))



Jiang, Wu and Katsaggelos (2009) proposed an unsupervised approach for detecting contextual anomalies in crowd motion, as indicated in the pipeline shown in Figure 2.10. Spatio-temporal patches represent motion features and are characterized by dynamic texture. They are classified and grouped to blobs, which describe position and size of every pedestrian. Then, contextual information is discover and used to detect the blobs corresponding to contextually anomalous behaviors based on spatial layout of pedestrian. Raghavendra et al. (2011) proposed a method for global anomaly detection. This method introduces Particle Swarm Optimization (PSO) as an algorithm for optimizing the interaction force computed using the Social Force Model (SFM).

Figure 2.10: Pipeline of the method proposed by Jiang et. al (a) Original video frame (b)Patch classification (c) Blob representation (d) Contextual anomaly (Source: (JIANG; WU; KATSAGGELOS, 2009))



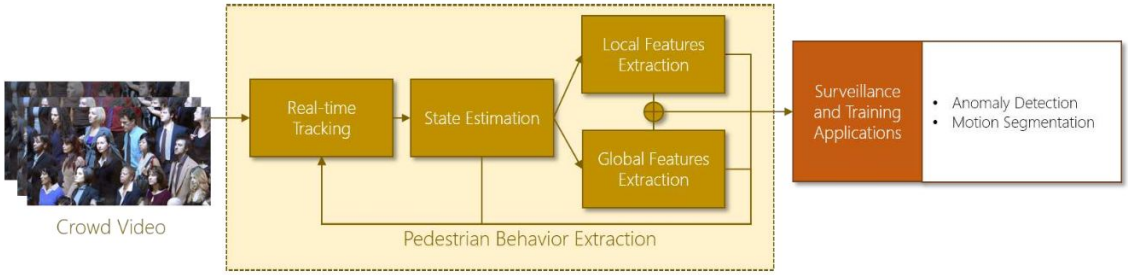
Lee, Suk and Lee (2013) used a matrix of influence of movement to represent behaviors of crowds, which is used to detect abnormal behaviors in the scene. In their model, a normal behavior is characterized by a low motion influence value. On the other hand, a high motion influence value indicates occurrence of abnormal behavior. In (FRADI; DUGELAY, 2014), two crowd dynamics features were used: appearance (through the calculation of crowd density) and movement (calculating histograms related to the speed and orientation of the crowd motion). Then, they determine changes in the behavior of the crowd following the strategy adopted in (ALMEIDA; JUNG, 2013).

Wang and Xu (2016) proposed a crowd anomaly detection algorithm based on image textures formulated by spatio-temporal information. They explore spatio-temporal texture characteristics in maintaining the statistical consistency across crowd events do-

main and its sensitivity to group anomalies. Wu, Moore and Shah (2010) presented a method for anomaly detection in crowded scenes based on Lagrangian particle dynamics and chaotic invariants, which is able to handle both coherent and incoherent scenes.

Bera, Kim and Manocha (2016) presented an algorithm (an overview is shown in Figure 2.11) for anomaly detection in low to medium density crowd videos using trajectory-level behavior learning. They combine online tracking algorithms, non-linear pedestrian motion models, and Bayesian learning techniques to compute the trajectory-level pedestrian behaviors for each pedestrian. Then, they used these learned behaviors to segment trajectories and motions of different pedestrians and detect anomalies.

Figure 2.11: The pipeline of Bera et. al approach to anomaly detection. The local and global features refer to individual vs. overall crowd motion features. (Source: (BERA; KIM; MANOCHA, 2016))

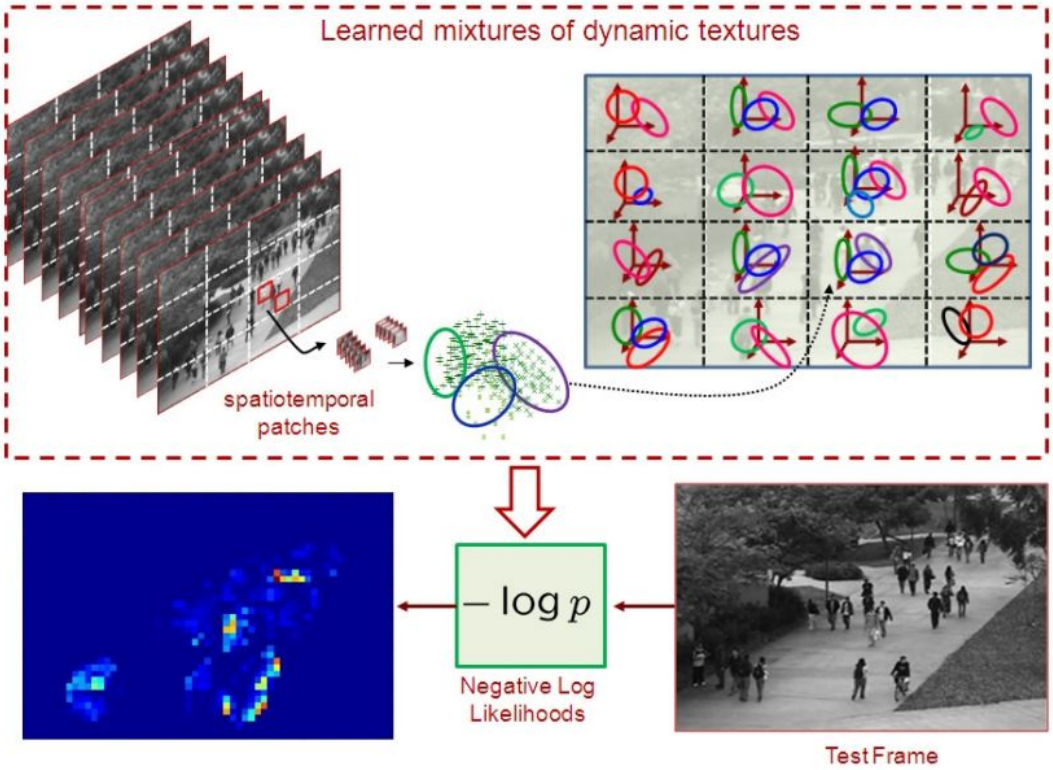


2.3.1.2 Supervised with One-Class

In supervised methods with a single class, the training step uses only normal samples. In the test phase, abnormalities are detected when the samples do not conform to the normal training data. Cheng, Chen and Fang (2013) presented a method for detecting and locating abnormal events in crowds. The authors extend the Bayes classifier, from multi-class classification to one class, to characterize normal events. Greenewald and Hero (2014) proposed an approach to learn a normal distribution of multi-frame pixels and to detect deviations from it through a probability-based approach. In order to reduce the number of samples required for learning, they applied a parametric approach of learning only the mean and covariance of distribution, which is used to calculate the Mahalanobis distance between analyzed pixels and the distribution of normal pixels. They divide the video into equal sized spatial patches, extract the marginal distributions of each one and compute the log-likelihoods. If the sample variance of these log-likelihoods is abnormally large, then the instance is declared anomalous.

Mahadevan et al. (2010) proposed the use of dynamic textures (DTs) toward anomaly detection in crowds – their pipeline is shown in Figure 2.12. They relate anomalies to events of low-probability with respect to a model of normal crowd behavior, then introduce DT-based models of normality over both space and time. Temporal normality is modeled with a mixture of DTs (MDT) and spatial normality is measured with a discriminant saliency detector based on MDTs. Li, Mahadevan and Vasconcelos (2013) also used an MDT model in their temporal and spatial detector of crowd anomalies, since these models represent the appearance and dynamics of a video. They implement a center-surround discriminant saliency detector that produces spatial saliency scores, and a model of normal behavior that is learned from training data and produces temporal saliency scores. Moreover, they define the spatial and temporal anomaly maps at multiple spatial scales .

Figure 2.12: Learning MDTs for temporal abnormality detection. For each region of the scene, an MDT is learned during training. At test time, the negative log-likelihood of the spatial-temporal patch centered at location l is computed using the MDT whose region center is closest to l (Source: (MAHADEVAN et al., 2010))



Yuan, Fang and Wang (2015) proposed an approach for detecting each pedestrian in a crowd using a pedestrian detection algorithm. For the individuals in the crowd, they proposed a Structural Context Descriptor (SCD) to exploit their valuable visual contextual

information, and object representation based on 3-D DCT is utilized to accommodate the appearance variation. Finally, the anomaly is detected online by temporal and spatial analysis of the SCD variation.

Ullah and Conci (2012) modeled crowd behavior by segmenting the motion flow. The segmentation is achieved by first extracting the motion information and successively applying graph-cuts to refine the obtained representation. This approach highlights the dominant direction of a crowd motion and detects anomalies as deviations from the model built on a training stage.

Wang and Miao (2010a) divided the whole frame of a video sequence into small blocks, and extracted motion pattern to represent the motion in each block. They use KLT corners as feature points to represent moving objects and track these points by optical flow. They also model the distribution of all motion vectors in one block as Gaussian distribution and trained models to classify in normal behavior or abnormal behavior. Wang and Miao (2010b) presented an approach to classify motion patterns into normal or abnormal groups according to the deviation between motion pattern and trained model and also according to its historical information. For that purpose, they extract motion pattern to represent activity based on optical flow of some pixels, and motion pattern is encoded by a descriptor called by them histogram of motion vector. A 3D grid structure is introduced to model the temporal-spatial relationship between motion patterns.

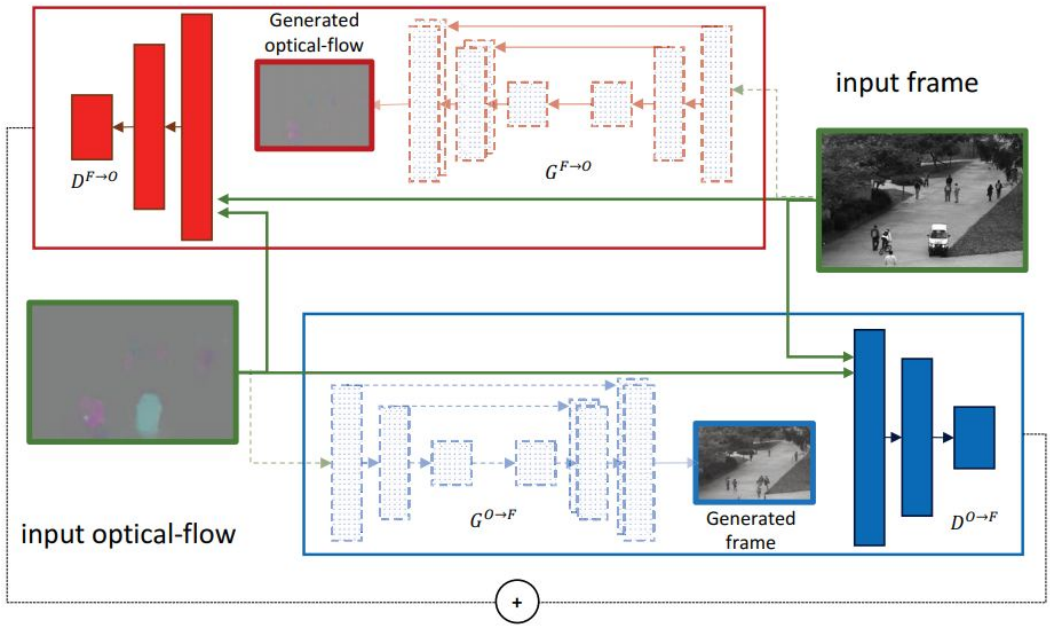
Chong et al. (2014) proposed a method where they learn regions of interest (ROIs) from a history of trajectory. The regions of interest are first learned and defined based on a Hierarchical Dirichlet Processes (HDP) grouping, which are also used to learn the statistical template of pedestrian distribution. It includes a global template that describes the overall crowd information, and local regional templates that are based on the semantic regions and cover local details. Lastly, they detect anomalies using crowd positions, density and flow data streams are the basic features for statistical analysis. Kratz and Nishino (2009) presented a statistical framework for modeling the local spatio-temporal motion pattern behavior of extremely crowded scenes. They model the motion variation of local space-time volumes and their spatio-temporal statistical behaviors to characterize the overall behavior of the scene, and detect unusual activity as statistical deviations.

Marčetić and Ribarić (2019) presented an approach to detect abnormal crowd behavior at a microscopic level, where the individual characteristic and group motion patterns are specified with fuzzy predicates, and the human interpretation of an video sequences of abnormal crowd behavior, based on commonsense knowledge, is mapped into

fuzzy logic functions. Based on the evaluation of these functions, abnormal crowd behavior is detected.

More recently, the use of neural networks has been also explored to detect anomalies in crowds. Ravanbakhsh et al. (2018) presented a method to detect local anomalies keeping track of the changes in the CNN feature across time. Specifically, they propose to measure local abnormality by combining semantic information with low-level optical-flow. Basically, they extract CNN-based binary maps from a sequence of input frames, and then compute the temporal CNN pattern measure using the extracted CNN-binary maps and the temporal CNN pattern measure fused with low-level motion features to find the refined motion segments. Ravanbakhsh et al. (2019) proposed to use Generative Adversarial Networks (GANs) – architecture is shown in Figure 2.13 – for abnormal event detection in crowds, which are trained to generate only the normal distribution of the data. During the adversarial GAN training, a discriminator is used as a supervisor for the generator network and vice-versa. They used this discriminator to tackle the abnormality detection problem, which was trained without the need of manually-annotated abnormal data.

Figure 2.13: The architecture of Generative Adversarial Nets used by Ravanakhsh et al. (Source: (RAVANBAKHSH et al., 2019))



Feng, Yuan and Lu (2017) proposed a method that uses PCA-Net from 3D gradients to extract appearance and motion features from videos, and in order to model event patterns, they constructed a deep Gaussian Mixture Model (GMM) with observed nor-

mal events. The deep GMM is a scalable deep generative model that stacks multiple GMM-layers on top of each other. To analyze a video, the likelihood is calculated to judge whether a video event is abnormal or not. Wang et al. (2020) proposed a method that decouples the problem into a feature descriptor extraction process, followed by an AutoEncoder based network called Cascade Deep AutoEncoder (CDA). The movement information is represented by a descriptor capturing the multi-frame optical flow information. Then, the feature descriptor of the normal samples is fed into the CDA network for training. The abnormal samples are distinguished by the reconstruction error of the CDA in the testing procedure.

2.3.1.3 Supervised Multi-Class

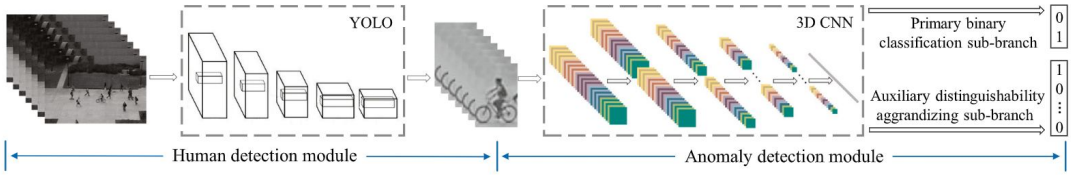
Supervised multi-class methods employ samples of both normal and abnormal scenes, and the model determines whether the scenes analyzed are normal or abnormal based on the training data. Ullah, Ullah and Conci (2014) proposed an approach to detect anomalies in crowds based on the observation of corner features. For each observed corner, motion features are acquired through optical flow techniques, more specifically Lucas-Kanade optical flow. These features are used to train an MLP neural network, and the behavior of the crowd is inferred on the test samples.

Zhou et al. (2016) also proposed a method for detecting and locating anomalous activities in video sequences of crowded scenes using temporal Convolutional Neural Networks. They capture features from both spatial and temporal dimensions by performing spatio-temporal convolutions, and thereby both the appearance and motion information encoded in continuous frames are extracted. To capture anomalous events appearing in a small part of the frame, the spatial-temporal CNN model is applied only on spatial-temporal volumes of interest (SVOI), which ensures robustness to noise.

Gong et al. (2020) proposed a local distinguishability aggrandizing network (LDA-Net) in a supervised manner, consisting of a human detection module and an anomaly detection module. In the human detection module, each person in the source frames is detected and cropped out, and then the patches are regarded as the input of the anomaly detection module, which is trained with annotated normal and abnormal data comprises two sub-branches: a primary binary classification sub-branch and an auxiliary distinguishability aggrandizing sub-branch. The auxiliary distinguishability aggrandizing sub-branch is integrated with an auxiliary multi-classification and an inhibition loss to extract more distinguishable detail features of normal and abnormal behaviors. An overview of (GONG

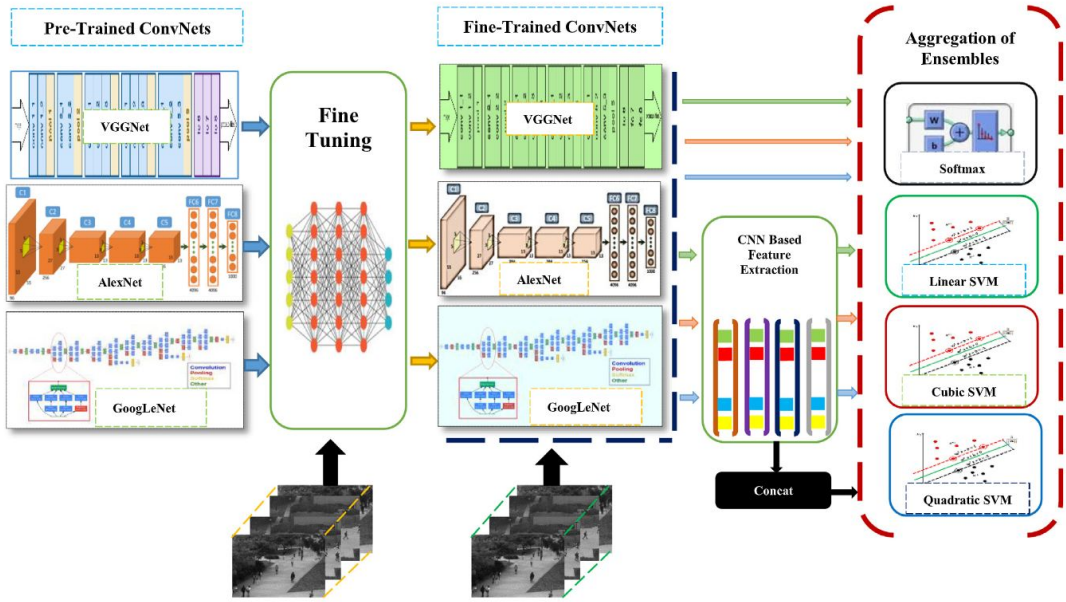
et al., 2020) is shown in Figure 2.14.

Figure 2.14: The architecture of LDA-Net. LDA-Net contains two modules: a human detection module and an anomaly detection module. And the anomaly detection module is made up of primary binary classification sub-branch and auxiliary distinguishability aggrandizing sub-branch, which are used for anomaly detection and action recognition, respectively (Source: (GONG et al., 2020))



Singh et al. (2020) proposed the concept of Aggregation of Ensembles (AOE) for detecting an anomaly in crowded video sequences. Their method uses an ensemble of different fine-tuned Convolutional Neural Networks (CNN) based on the hypothesis that they learn different levels of semantic representation from crowd videos, and its model pipeline is shown in Figure 2.15. The AOE utilizes fine-tuned ConvNets as fixed feature extractors to train variants of an SVM classifier, and then the posterior probabilities are fused to predict the anomaly in video.

Figure 2.15: The architecture of AOE proposed by Singh et al. (Source: (SINGH et al., 2020))



2.4 Crowd Datasets

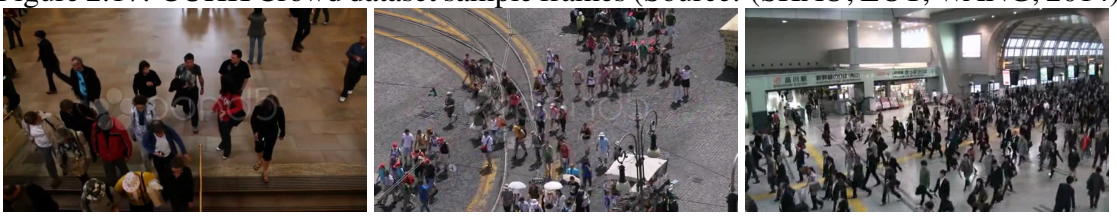
Another crucial issue when dealing with crowds is the availability of publicly available datasets with suitable annotated data. Although there are many dataset of crowds for a variety of tasks, there are very few ones containing dense crowds with annotation about abnormality. An alternative would be to use synthetic crowds generated through crowd simulation methods, allowing a wide variety of crowd scenes with expected behavior. However, the generation of realistic behavior yet an open topic in crowd simulation area (Li et al., 2019). Furthermore, realistic rendering under several variations that impact optical flow (such as illumination changes, shadows, articulated body motion, etc.) would be required to evaluate crowd motion estimation methods. Based on these consideration, we focus our analysis on sparser crowds that provide annotation and on denser crowds for which we later provide manual labels.

The PETS dataset (FERRYMAN; ELLIS, 2010) is widely used by researchers dedicated to crowd activities analysis. This dataset includes estimation of crowd and within a crowd, tracking of individuals, and specific crowd events detection. The scenes vary between low and medium dense crowds, as shown in Figure 2.16, and presents multiple view of the same scene, with camera parameters provided in the dataset. Another dataset explored by crowd behavior researchers is the CUHK Crowd dataset (SHAO; LOY; WANG, 2014), which contains 474 video clips from 215 crowded scenes. The scenes vary between medium and high dense crowds, as shown in Figure 2.17. However, no ground truth information on abnormal behavior is provided.

Figure 2.16: PETS dataset sample frames (Source: (FERRYMAN; ELLIS, 2010))

The UFC-CRCV dataset provides some video samples focused on crowds. The "Tracking in High Density Crowds" dataset (ALI; SHAH, 2008) consists of three scenes of marathons, all with a high density of people. The Crowd Segmentation dataset (ALI; SHAH, 2007a) contains videos of human crowds and other high density moving objects (such as fish and cars) which are not useful in our method, and other two datasets focused

Figure 2.17: CUHK Crowd dataset sample frames (Source: (SHAO; LOY; WANG, 2014))



on crowd counting (IDREES et al., 2013; IDREES et al., 2018). This dataset contains several high dense scenes, as shown in Figure 2.18, but just a few contain anomaly behavior. Another difficulty is that these datasets are focused on other tasks, so they are not labeled to evaluate anomaly detection methods.

Figure 2.18: UFC dataset sample frames (Source: (ALI; SHAH, 2008; ALI; SHAH, 2007a))

The UCSD Anomaly Detection Dataset (CHAN; VASCONCELOS, 2008) provides ground truth of anomaly behavior. However, all 48 available video sequences present a low density of people, as shown in Figure 2.19. Despite not contatining denser crowds, this dataset is widely used in the crowd anomaly detection community.

Figure 2.19: UCSD dataset sample frames (Source: (CHAN; VASCONCELOS, 2008))

2.5 Chapter Conclusions

This chapter indicated that crowd motion is an important cue for anomaly and/or event detection. However, there are very few methods dedicated to extracting crowd flow: most approaches use generic optical flow methods, and do not explore characteristics

inherent to crowds. Our idea is to estimate the crowd flow by using a post-processing step coupled to a traditional optical flow method, exploring neighborhood information.

Also, there are many methods devoted to crowd anomaly and event detection based on crowd flow information. Although a few of them explore dense crowds or specific psycho-social characteristics expected in a real crowd, such as (MEHRAN; OYAMA; SHAH, 2009), most approaches follow the typical machine learning pipeline. In this dissertation we tackle crowd anomaly detection problem by strongly exploring neighborhood information based on the proxemics theory (HALL, 1966). In fact, the proposed approach for both crowd flow estimation and anomaly detection are based on the motion similarity of each pixel and its neighbors. We use one specific neighborhood scale in the crowd flow estimation problem, and several different neighborhood scales in anomaly detection. The proposed approaches are presented in the next chapters.

3 CROWD FLOW ESTIMATION

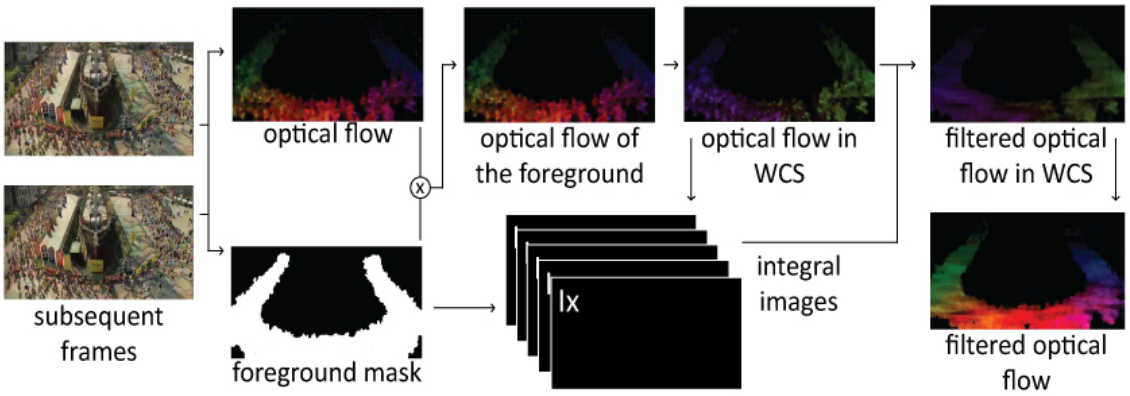
In this chapter we present the proposed approach that explores local neighborhoods to calculate a consistent crowd flow from optical flow techniques. We assume that people in a structured crowd are mostly affected by their nearby neighbors, characterized by a spatial "influence region" based on the concept of proxemics (HALL, 1966), which states that the relationship between two people depends on the distance between them. For crowd flow estimation, we use the optical flow within such influence regions to produce a smoother crowd flow, whereas for crowd event detection we evaluate the local flow consistency within regions of varying radii.

Many methods have been suggested to estimate the optical flow for generic temporal scenes, but to our knowledge there is no approach designed specifically for crowds. We propose a fast method that explores the expected behavior of real crowds by post-processing the optical flow obtained by any generic optical flow estimation method, suited for scenes captured by a calibrated static camera. An overview of the proposed approach is provided in Figure 3.1, and each step is detailed next.

3.1 Computing the "effective" optical flow

At each frame t of the analyzed video sequence we calculate the optical flow $\mathbf{v}(\mathbf{u})$ of the pair of frames $t-1$ and t using any baseline optical flow approach, where $\mathbf{u} = (u, v)$ represents the image coordinates. Since the main goal is to extract crowd flows, we also can remove uninteresting regions using a background removal method to estimate a binary

Figure 3.1: Pipeline to obtain crowd flow: i) estimate optical flow from two subsequent frames; ii) convert this flow to world coordinates; iii) filtered the optical flow using a crowd model; iv) project the filtered flow back to image coordinates.



foreground mask $f_g(\mathbf{u})$, and assume that foreground objects in the scene are mostly related to humans or/and removing regions with low optical flow magnitude. In scenes where it is not possible to extract the foreground using background removal method (e.g. in very dense crowds with only a few frames to estimate the background), an alternative method is obtain the foreground mask using the magnitude of the optical flow (either image or world coordinates), assuming low flows as being part of the background.

The motivation for using background removal is to restrict the whole analysis only to pixels related to people. Using solely the optical flow and eliminating low magnitude motion vectors could be an alternative, but temporarily stationary people could be removed. Furthermore, the definition of the threshold is not trivial due to camera perspective issues: large displacements in the world far from the camera might lead to small motion vectors in the image domain.

3.2 Optical flow in world coordinates and fast local filtering

Given the "effective optical flow" (optical flow restricted to foreground pixels) of people, our approach explores the expected pedestrian organization in a crowd. For instance, individuals knowing each other may form groups, which behave similar to single pedestrians individuals (HELBING; MOLNÁR, 1997). In denser crowds, even unrelated nearby pedestrians tend to present similar motion patterns, since the crowd acts as a single entity (MEHRAN; OYAMA; SHAH, 2009).

Our approach is to analyze the neighborhood of each pixel belonging the foreground and recalculate the pixel flow based on its neighbors. People are mostly affected by their nearby neighbors, which can be characterized by a spatial "influence region". In fact, Hall (1966) studied the expected relationship between two people based on their distances (from intimate to public), so that concentric circles with different radii characterize the different "personal spaces", or proxemics: intimate, personal, social and public. The smaller the radius r, the stronger is the expected relationship between the person under analysis and their neighbors. Hence, different crowd flow consistency levels can be achieved by varying r. We consider the personal distance ($r = 1.2$m) as the default value for the neighborhood influence, since individuals that share their personal space are expected to be familiar (family or friends), typically having the same goals and therefore presenting similar motion vectors.

In this work, we approximate the personal space A with radius r as a square region

with dimensions $2r \times 2r$, to simplify the computations using integral images (this step will be described in details later in the dissertation). Given a foreground pixel $\mathbf{u}$, we consider a calibrated static surveillance camera, and assume that the filmed region is roughly planar. We use a plane homography H to project $\mathbf{u}$ into a world point $\mathbf{x} = (x, y, h)$, where h is the height of the 3D point on the person related to pixel $\mathbf{u}$. Although it is difficult to obtain the actual value for h, the typical heights of a person are limited. Since obtaining the ground plane homography is simple (if the camera parameters are not known, the homography can be estimated using only four planar points with known coordinates), we used $h = 0$ in the conversion for all image points.

Then we consider a personal region $A_{\mathbf{x}}$ centered at world point $\mathbf{x}$, parallel to the ground plane, and project it back to image coordinates. Hence, the same personal space $A_{\mathbf{x}}$ in the world leads to different regions $A_{\mathbf{u}}$ in the image domain due to perspective issues, as illustrated in Figure 3.2. Since the camera is assumed to be static, the ground plane homography remains constant in time. To reduce the computational burden, the projections of $A_{\mathbf{x}}$ centered at all possible image pixels are pre-computed a single time and stored in a LookUp Table (LUT). Note that the ideal proxemics should be modeled as circles in the WCS, which project to ellipses in the ICS. However, we opted to simplify the models to rectangular regions, which allows us to explore integral images and drastically reduce the computational cost.

The analysis of the crowd flow is also performed in the WCS, to alleviate the distortions caused by camera perspective. To that end, we first project the optical flow at each foreground pixel $\mathbf{u}$ to the WCS, obtaining

$$\mathbf{v}_w(\mathbf{u}) = H\left(\mathbf{u} + \mathbf{v}(\mathbf{u})\right) - H\left(\mathbf{u}\right), \tag{3.1}$$

where H is the homography from image pixels to the ground plane in the WCS. Note that the ideal mapping would be given by

$$\mathbf{v}^i_w(\mathbf{u}) = H_h\left(\mathbf{u} + \mathbf{v}(\mathbf{u})\right) - H_h\left(\mathbf{u}\right), \tag{3.2}$$

where H_h is the homography computed at the actual height h of the pixel under consideration (recalling that finding h is difficult). Since the range of h values are bounded by the height of a person, it is possible to estimate the maximum projection error at each pixel location $\mathbf{u}$ as

$$E_{max}(\mathbf{u}) = \max_{h \in [0, h_{max}]} \left\| \text{proj}(\mathbf{v}^i_w(\mathbf{u})) - \text{proj}(\mathbf{v}_w(\mathbf{u})) \right\|, \tag{3.3}$$

Figure 3.2: Example of a "region of influence" with the same size in world coordinates projected back to the image at different pixel locations. In the image domain they present different sizes, due to camera perspective.

where h_{max} is the maximum height of a person and proj$(\cdot)$ denotes the projection of homogeneous to Cartesian coordinates. Note that cameras closer to the top-down setup tend to present smaller errors.

The next step is to combine the optical flow in the WCS for each pixel based on the projections of the corresponding personal spaces $A_{\mathbf{u}}$. In a typical structured crowd flow, $\mathbf{v}_w(\mathbf{u})$ should be roughly homogeneous within $A_{\mathbf{u}}$. The Generalized Social Forces Model presented by Helbing, Farkas and Vicsek (2000) assumes that the actual velocity of a pedestrian is a weighted average between the desired velocity and the mean velocity of the people around them. In this work, we explore a similar idea, but to filter the optical flow. More precisely, the smoothed crowd flow in the WCS $\mathbf{v}_w^s(\mathbf{u})$ for each foreground pixel $\mathbf{u}$ is given by

$$\mathbf{v}_w^s(\mathbf{u}) = p(\mathbf{u}) \left\langle \mathbf{v}_w(\mathbf{u}) \right\rangle_{A_{\mathbf{u}}} + (1 - p(\mathbf{u}))\, \mathbf{v}_w(\mathbf{u}), \tag{3.4}$$

where $\langle \cdot \rangle_B$ denotes the mean value within a spatial region B, and $0 \leq p(\mathbf{u}) \leq 1$ is the pixel-dependent weight. Larger values for $p(\mathbf{u})$ yield more local filtering, which corresponds to stronger herding behaviors in (HELBING; FARKAS; VICSEK, 2000).

On one hand, we want to smooth noisy optical flow vectors so that it becomes spatially coherent within a neighborhood (i.e., $p(\mathbf{u})$ should be larger in these cases). On the other hand, we would like to preserve individual behaviors that might indicate unusual events, such as a pedestrian moving against the crowd. In such cases, the flow should not be smoothed significantly (i.e., $p(\mathbf{u})$ should be smaller). These two filtering characteristics can be obtained by selecting $p(\mathbf{u})$ adaptively based on the motion vector field within $A_{\mathbf{u}}$. We first estimate the "adherence" of $\mathbf{v}_w(\mathbf{u})$ with the neighboring flow using the Mahalanobis distance, given by

$$D(\mathbf{u}) = \sqrt{(\mathbf{v}_w(\mathbf{u}) - \boldsymbol{\mu}_{\mathbf{u}})^T S_{\mathbf{u}}^{-1} (\mathbf{v}_w(\mathbf{u}) - \boldsymbol{\mu}_{\mathbf{u}})}, \tag{3.5}$$

where $\boldsymbol{\mu}_{\mathbf{u}} = \langle \mathbf{v}_w(\mathbf{u}) \rangle_{A_{\mathbf{u}}}$ is the average of the optical flow in region $A_{\mathbf{u}}$, and $S_{\mathbf{u}}$ is the covariance matrix in the same region. It is worth noticing that we tested other options to calculate the adherence of $\mathbf{v}_w(\mathbf{u})$ with the neighboring flow, such as Hausdorff distance or even the Euclidean distance between $\mathbf{v}_w(\mathbf{u})$ and $\boldsymbol{\mu}_{\mathbf{u}}$, but the Mahalanobis distance showed better results in our experiments. Note that $D(\mathbf{u})$ gets progressively larger as $\mathbf{v}_w(\mathbf{u})$ becomes less coherent with the neighboring flow. Therefore we select the weight $p(\mathbf{u})$ as

$$p(\mathbf{u}) = \min\{\alpha D(\mathbf{u}), D_{max}\}, \tag{3.6}$$

where $0 \leq D_{max} < 1$ is a constant that defines the maximum possible value for $p(\mathbf{u})$ (to avoid completely replacing the flow at the central pixel by the average value), and $\alpha > 0$ is a scaling factor for the Mahalanobis distance. If α is large, smaller distances $D(\mathbf{u})$ generate larger weights $p(\mathbf{u})$, leading to more smooth. Decreasing α leads to an opposite effect. In all tests we use $\alpha = 0.75$ and $D_{max} = 0.95$, set experimentally, so that the weight of the neighborhood optical flow is at most 0.95.

It is also important to note that our formulation allows both smoothing noisy flows and keeping individual behaviors by selecting an adequate personal space $A_{\mathbf{u}}$ for the analysis. By selecting smaller personal spaces (e.g., intimate or personal), our method can keep individual flow vectors and reduce local noise. As the region of analysis is increased (social or public), the overall crowd motion is retrieved, and individualities tend to be lost. Results obtained by changing the proxemics are presented in Section 3.3.2.

In order to keep computational complexity of the proposed method low, we actually use the bounding box of each region $A_{\mathbf{u}}$, since the use of rectangular regions allows fast computation of the mean vector and the covariance matrix. As shown in (TUZEL;

PORIKLI; MEER, 2006), we compute five integral images based on second order statistics of the x and y components of the vector field $\mathbf{v}_w(\mathbf{u})$, denoted by I_x, I_y, I_{x^2}, I_{y^2} and I_{xy}. Based on these integral images, both $\boldsymbol{\mu}_\mathbf{u}$ (Eq. (3.7)) and $S_\mathbf{u}$ (Eq. (3.10)) can be computed in constant time, regardless the dimensions of $A_\mathbf{u}$. Also, since $A_\mathbf{u}$ might contain background pixels (that are not used in our analysis), we also obtain the number of foreground pixels (n) within each region $A_\mathbf{u}$, using the integral image (I_f) of $f_g(\mathbf{u})$. Let us consider that the bounding box of $A_\mathbf{u}$ is characterized by its upper left and lower right corners, denoted by (x_0, y_0) and (x_1, y_1), respectively. The mean motion is given by

$$\boldsymbol{\mu}_\mathbf{u} = \frac{\mathbf{P}}{n}, \tag{3.7}$$

where

$$n = I_f(x_0, y_0) + I_f(x_1, y_1) - I_f(x_0, y_1) - I_f(x_1, y_0) \tag{3.8}$$

is the number of foreground pixels within $A_\mathbf{u}$, and the components P_x and P_y of vector $\mathbf{P}$ are given by

$$\begin{aligned} \mathbf{P} &= \begin{bmatrix} P_x & P_y \end{bmatrix}^T, \\ P_x &= I_x(x_0, y_0) + I_x(x_1, y_1) - I_x(x_0, y_1) - I_x(x_1, y_0), \\ P_y &= I_y(x_0, y_0) + I_y(x_1, y_1) - I_y(x_0, y_1) - I_y(x_1, y_0). \end{aligned} \tag{3.9}$$

Similarly, the covariance matrix is given by

$$S_\mathbf{u} = \frac{1}{n-1}\left(Q - \frac{\mathbf{P}^T\mathbf{P}}{n}\right), \tag{3.10}$$

where the elements Q_{ij} of matrix Q, for $i, j \in \{1, 2\}$, are given by

$$\begin{aligned} Q_{11} &= I_{x^2}(x_0, y_0) + I_{x^2}(x_1, y_1) - I_{x^2}(x_0, y_1) - I_{x^2}(x_1, y_0), \\ Q_{12} = Q_{21} &= I_{xy}(x_0, y_0) + I_{xy}(x_1, y_1) - I_{xy}(x_0, y_1) - I_{xy}(x_1, y_0), \\ Q_{22} &= I_{y^2}(x_0, y_0) + I_{y^2}(x_1, y_1) - I_{y^2}(x_0, y_1) - I_{y^2}(x_1, y_0). \end{aligned} \tag{3.11}$$

Given the smoothed flow $\mathbf{v}_w^s(\mathbf{u})$, the last and optional step of the proposed approach is to map it back to image coordinates, similarly to the forward projection given by Eq. (3.1). The smoothed optical flow in the ICS is given by

$$\mathbf{v}^s(\mathbf{u}) = H^{-1}\left(H(\mathbf{u}) + \mathbf{v}_w^s(\mathbf{u})\right) - \mathbf{u}, \tag{3.12}$$

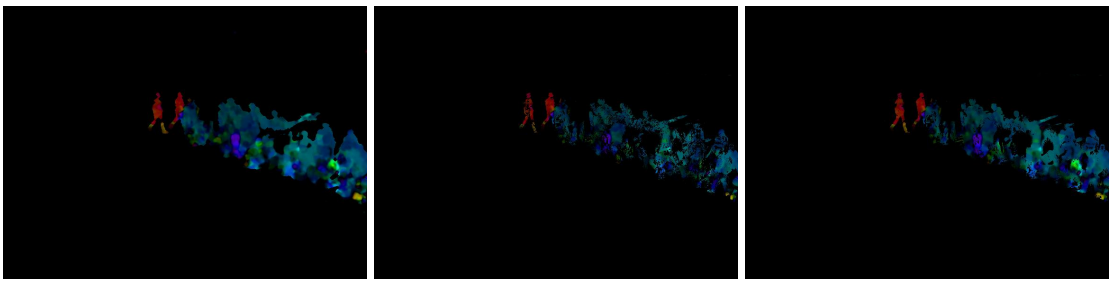
where H^{-1} is the homography from ground plane in the WCS to image pixels (i.e. the inverse of H).

3.3 Experimental Results

The proposed approach is a post-processing step that can be coupled to any baseline optical flow algorithm. As revised in Section 2.2, there are many existing generic optical flow methods that vary w.r.t. to the mathematical formulations, capacity to deal with small or large displacement, regularization function used to smooth the optical flow, computational cost, etc. Our claim in this work is that the proposed filtering method tailored to crowd motion coupled to simpler optical flow baseline approaches (which are potentially fast) can lead to fast and smooth crowd flow estimates.

In our experimental validation, we test four different baseline optical flow methods, as in (KAJO; MALIK; KAMEL, 2016): the variational approach presented in (BROX; MALIK, 2011), which was designed to capture large displacement vectors; the "Classic+NL" model presented in (SUN; ROTH; BLACK, 2010), which combines the classical optical flow formulation based on data fidelity and regularization with a weighted non-local term; and the fast and well-known optical frame method based on polynomial expansion proposed by Farnebäck (2003). In all experiments, we used the background removal proposed in (JUNG, 2009) to obtain the foreground mask $f_g(\mathbf{u})$, which is able to handle illumination changes and shadows. However, we have also experimented with other approaches available in the library OpenCV (BRADSKI, 2002), and results were similar, as shown in Figure 3.3, to our parameters $\alpha = 0.75$ and $D_{max} = 0.95$, that were set experimentally. Hence, the weight of the neighborhood optical flow is at most 0.95.

Figure 3.3: Crowd flow after background subtraction and elimination of very small flows.



(a) The chosen method (JUNG, 2009) (b) Mixture of Gaussian (ZIVKOVIC; HEIJDEN, 2006) (c) KNN-based approach (ZIVKOVIC; HEIJDEN, 2006)

Although there are publicly available datasets and protocols for validating generic

optical flow algorithms, such as the Middlebury dataset (BAKER et al., 2011), it is not to our knowledge the existence of video sequences involving crowds with annotated ground truth optical flow. Kajo, Malik and Kamel (2016) compared optical flow methods in the context of crowd motion, but used "approximated" ground truth values based on human annotations.

In this work, validation is performed qualitatively, by visual inspection, and quantitatively using two different approaches. In the first analysis, we initialize a set of particles in the first frame of a video sequence, estimate their paths using particle advection with the optical flow and evaluate the smoothness of the obtained trajectories. In the second one, we evaluate the impact of the proposed filtering approach in the context of crowd event detection.

We first evaluate the execution times of the proposed method coupled with the three baseline optical flow approaches. Then, we show the qualitative analysis of the corresponding optical flows applied to crowd video sequences, and the quantitative analysis based on trajectory smoothness and crowd event detection.

3.3.1 Execution time

The first analysis consists on evaluating the execution time of the three baseline approaches (Brox[1] (BROX; MALIK, 2011), Classic+NL[2] (SUN; ROTH; BLACK, 2014) and Farnebäck (FARNEBÄCK, 2003)) before and after coupling the proposed post-processing method. For that purpose, we used publicly available MATLAB implementations of these approaches. For a fair comparison, we also use a MATLAB version of our method.

Table 3.1 shows the average running times to compute the optical flow between two adjacent frames (with and without post-processing) using video sequences from PETS2009 (images with dimensions 768×576 pixels) running on an i7-2700K 3.50GHz Quad-Core Processor with 12 GB RAM.

For both Classic+NL and Brox methods, our approach does not add significant overhead, since the baseline methods are already slow. On the other hand, our post processing overhead is significant when coupled to Farnebäck's algorithm, which is mostly caused by our MATLAB implementation.

[1] https://lmb.informatik.uni-freiburg.de/index.php
[2] http://cs.brown.edu/ dqsun/research/index.html

By using C++ implementations of both Farnebäck's algorithm (from OpenCV (BRADSKI, 2002)) and our method, the running times reduce to 0.08s and 0.2s (when using our post-processing), while the implementation of DeepFlow has running time of 0.7s without our post-processing and 0.9s with it. Hence, the full running time of Farnebäck's algorithm plus out method is enough to process 5 frames per second, which is almost the framerate of traditional surveillance cameras (7 frames per second).

It is also important to point out that the proposed post-processing is highly parallel, since each point $\mathbf{v}_w^s(\mathbf{u})$ of the smoothed crowd flow can be obtained independently. Our results used OpenMP parallelization for the CPU, but significant speed-ups are expected if GPU processing is also applied. Also, the integral image is acurrently computed along the whole frame, but can be restrained to a bounding box that contains all the foreground pixels.

Table 3.1: Execution time of each algorithm with and without our post processing method, obtained with MATLAB implementations.

Algorithm		Original	Plus Our Method
Matlab	Brox	$\approx$ 22s	$\approx$ 27s
	Classic+NL	$\approx$ 70s	$\approx$ 75s
	Farnebäck	$\approx$ 0.1s	$\approx$ 4.2s
C++	Farnebäck	$\approx$ 0.08s	$\approx$ 0.2s
	DeepFlow	$\approx$ 0.7s	$\approx$ 0.9s

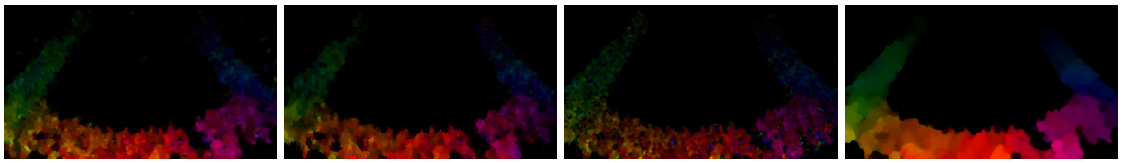
3.3.2 Qualitative Analysis

For a qualitative analysis of the obtained crowd flows, we again explore video sequences from the UCF and PETS 2009 datasets. For those sequences, we perform a visual evaluation of the crowd flow obtained with the baseline approaches with and without the proposed post-processing stage.

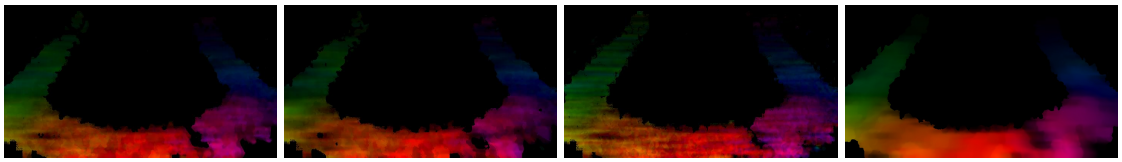
Figures 3.4, 3.5 and 3.6 show one illustrative frame of three different video sequences used in the analysis. It can be observed that the post-processing scheme provides smooth and coherent crowd flows when coupled to all three baseline optical flow algorithms, leading to less noisy artifacts. More importantly, the three filtered flows are visually very similar, which indicates that the choice of the baseline method is not crucial. In fact, this observation allows the use of a noisier (but faster) optical flow technique, such as Färneback, rather than more robust (and costlier) techniques such as Brox's or Classic+NL.

Figure 3.4: Visual analysis of different baseline optical flow algorithms in Marathon scene, with and without the proposed post-processing approach.

(a) Marathon 03 - Fr. 38



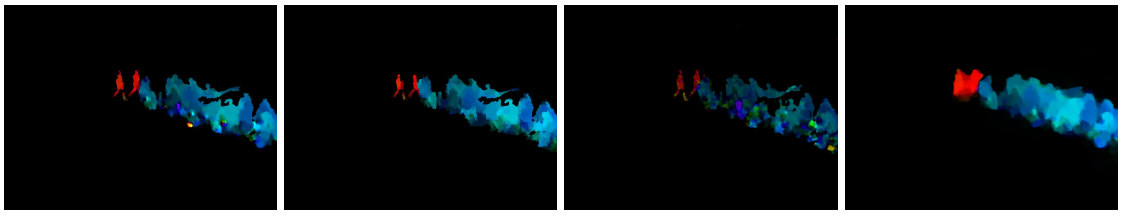
(b) Brox Original (c) Classic+NL Original (d) Farneback Original (e) DeepFlow Original



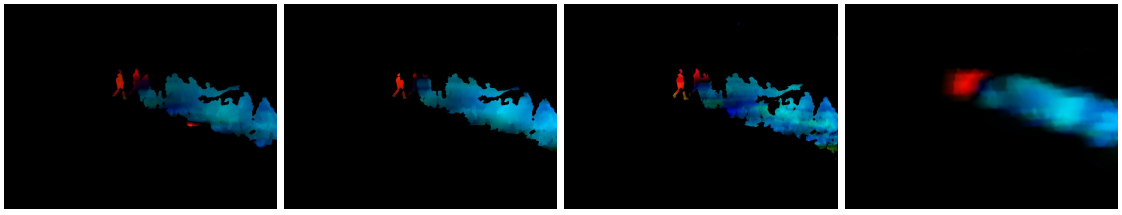
(f) Brox Filtered (g) Classic+NL Filtered (h) Farneback Filtered (i) DeepFlow Filtered

Figure 3.5: Visual analysis of different baseline optical flow algorithms in PETS2009 which has two groups moving in oposite direction, with and without the proposed post-processing approach.

(a) PETS2009 14-46 - Fr. 21



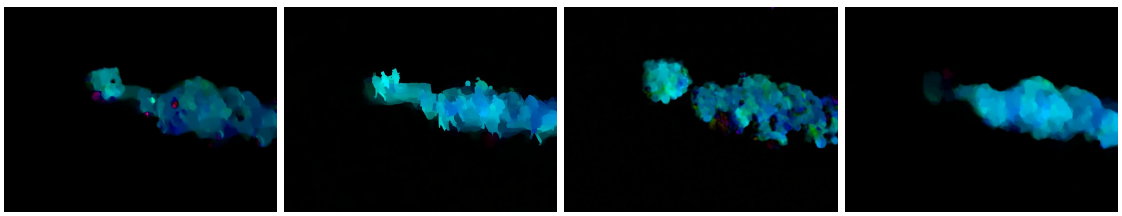
(b) Brox Original (c) Classic+NL Original (d) Farneback Original (e) DeepFlow Original



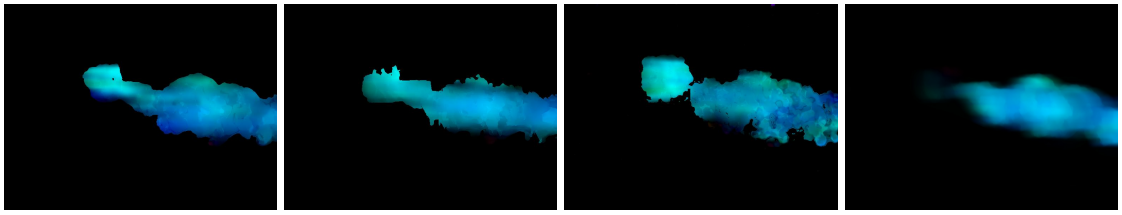
(f) Brox Filtered (g) Classic+NL Filtered (h) Farneback Filtered (i) DeepFlow Filtered

Figure 3.6: Visual analysis of different baseline optical flow algorithms in PETS2009 scene, with and without the proposed post-processing approach.

(a) PETS2009 14-16 - Fr. 48



(b) Brox Original (c) Classic+NL Original (d) Farneback Original (e) DeepFlow Original



(f) Brox Filtered (g) Classic+NL Filtered (h) Farneback Filtered (i) DeepFlow Filtered

In particular, Figure 3.5 relates to the PETS 2009 S3 - Multiple Flow dataset[3], in which a couple of pedestrians move in the opposite direction of a larger crowd. The proposed method was able to smooth the flow, but without mixing the velocity vectors of the two groups moving in opposite directions, which is very important in the context of crowd event detection. Such characteristic was possible because the chosen radius for the spatial neighborhood ($r = 1.2$m), which relates to Hall's personal distance, is sufficiently large to provide smoothing of the optical flow, but also small enough to prevent excessive blur.

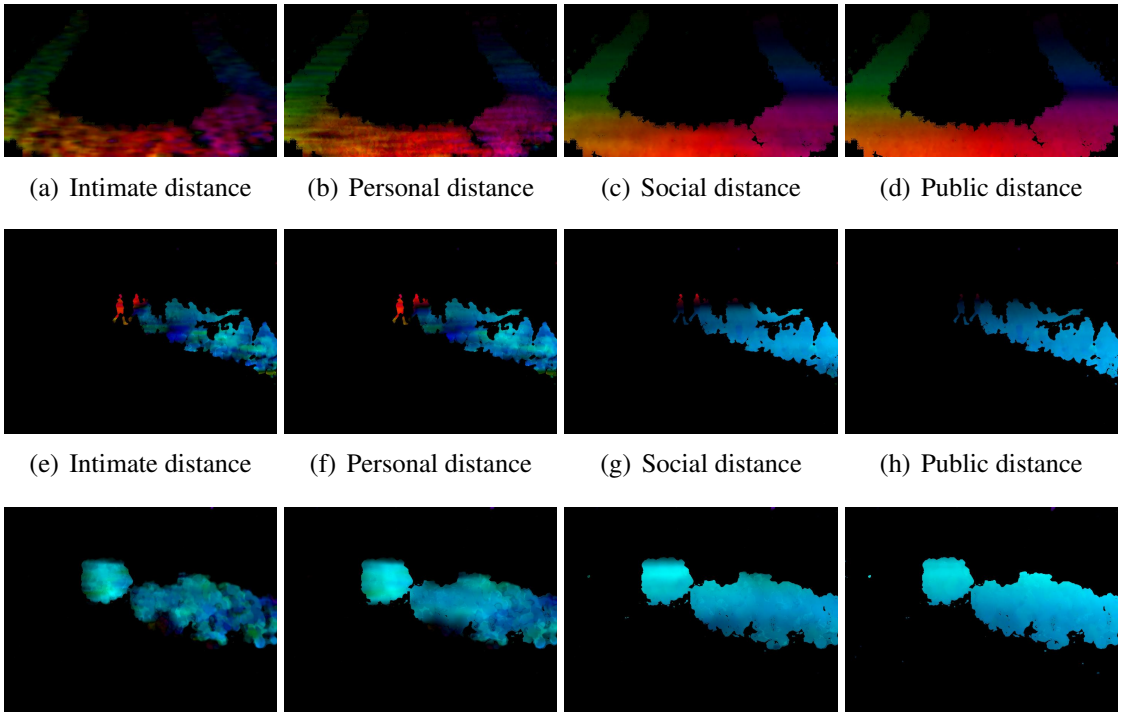
To analyze the influence of the neighborhood radius r in the filtered flow, we repeated the experiment of Figures 3.4 - 3.6 using different radii to determine the image projected regions $A_{\mathbf{u}}$. More precisely, we compared the filtering results using the intimate distance ($r = 0.45m$), personal distance ($r = 1.2m$, which is our default value), social distance ($r = 3.5m$), and the limit of the close phase of the public distance ($r = 7.6m$) as proposed by Hall (1966). The results shown in Figure 3.7 indicate that the final optical flow gets progressively smoothed as r increases, as expected. Although the optimal value for r might depend on the application, we believe that the chosen default value provides a good compromise between under- and over-smoothing the flow. In terms of computational cost, there is no difference as r changes, since integral images are used.

In addition to the analysis of influence on choosing a particular neighborhood region size has on obtaining the crowd flow, we also analyze other parameters involved in the method. In Figure 3.8 we show the results obtained with different values to h, always assuming that all pixels are in a planar region with height h, i.e. when $h = 0$ means that all pixels are on ground plane, we also evaluate values as 1.8m and 0.9m approximate values for the top of a person's head and half between the top and the floor respectively. In Eq. (3.3) we present a equation to choose the best h value to minimize the maximum projection error. The crowd flow estimated in all cases are quite similar visually, even the value found to be the best h is 0.91, resulting in an almost identical crowd flow between the third and fifth figure rows.

We also evaluate the influence of α in the smooth the optical flow. In Figure 3.9 we analyze four values to α: 0.5, 0.75, 1.5 and 2.0, this parameter influences how affected the pixel motion will be by its neighborhood, being a multiplier for $D(\mathbf{u})$. In the figure this behaviour can be observed when we use small values to α and the optical flow is lesser smoothed than when we use higher values to α.

[3]<http://www.cvg.rdg.ac.uk/PETS2009>

Figure 3.7: Influence of the radius r of the spatial neighborhood, chosen based on Hall's interpersonal distances (HALL, 1966).



(a) Intimate distance (b) Personal distance (c) Social distance (d) Public distance

(e) Intimate distance (f) Personal distance (g) Social distance (h) Public distance

(i) Intimate distance (j) Personal distance (k) Social distance (l) Public distance

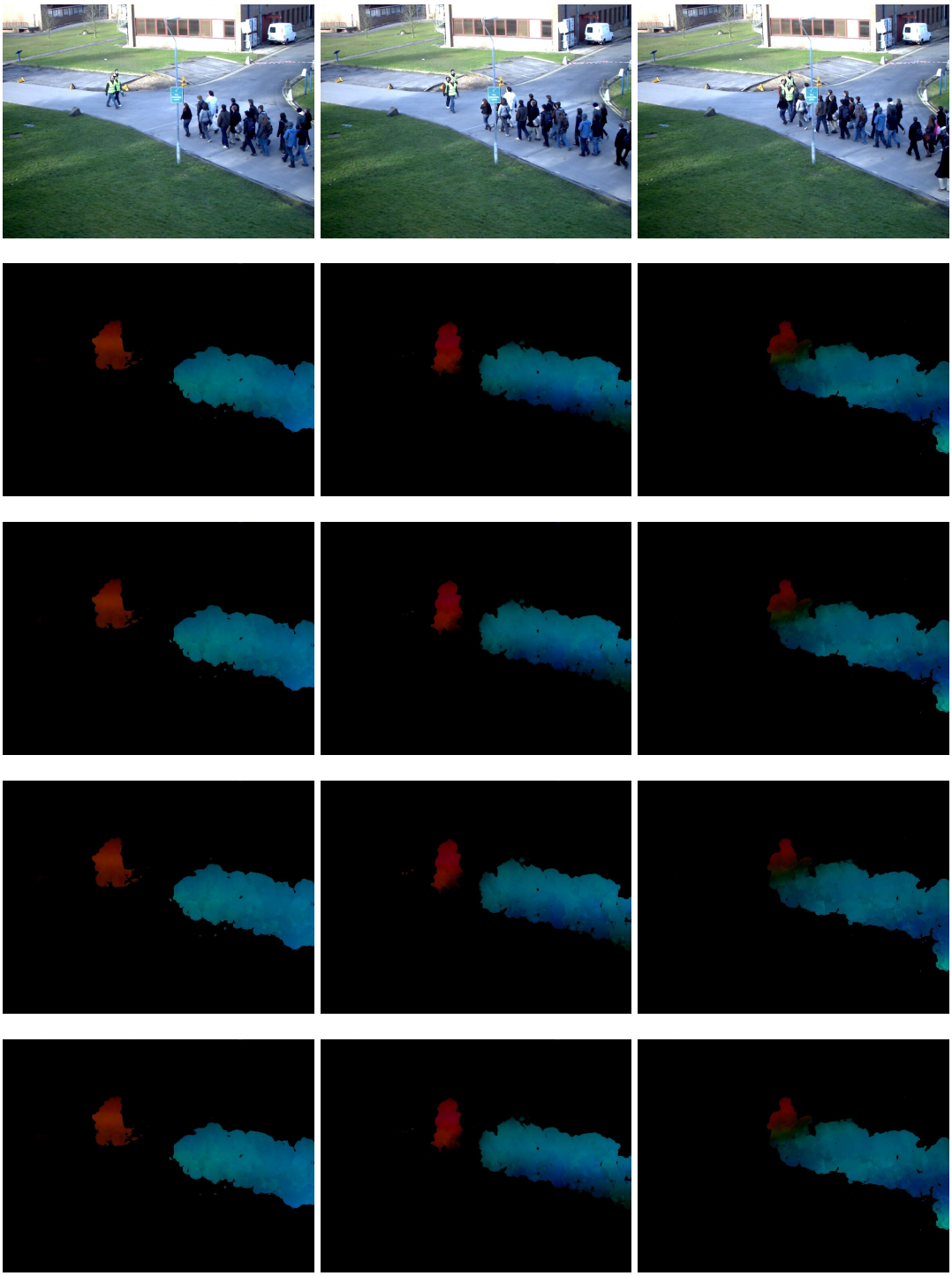
The last parameter analyzed is D_{max}, which indicate a value max of influence of the neighborhood can have on the pixel. In this way small values means that the original flow will be maintained, while high values means that the original flow of the pixel can be lost in favor of the neighborhood mean flow. This analysis is confirmed by Fig 3.10, where when we use $D_{max} = 0.25$ the original flow is conserved including inconsistent flows of leg movement from people in a crowd, while when we use $D_{max} = 1.0$ the obtained flow is more uniform, with the movement of the pixels being in most cases the average of its neighbors.

3.3.3 Quantitative Evaluation based on Particle Advection

In this experiment, we used the same datasets and baseline methods explored in the previous examples, and quantitatively evaluate the smoothness of the trajectories obtained by particle advection. More precisely, we randomly initialize a set of particles in the first frame of the sequence (restricted to foreground pixels), and use the pairwise optical flow to update the position of the particle in time.

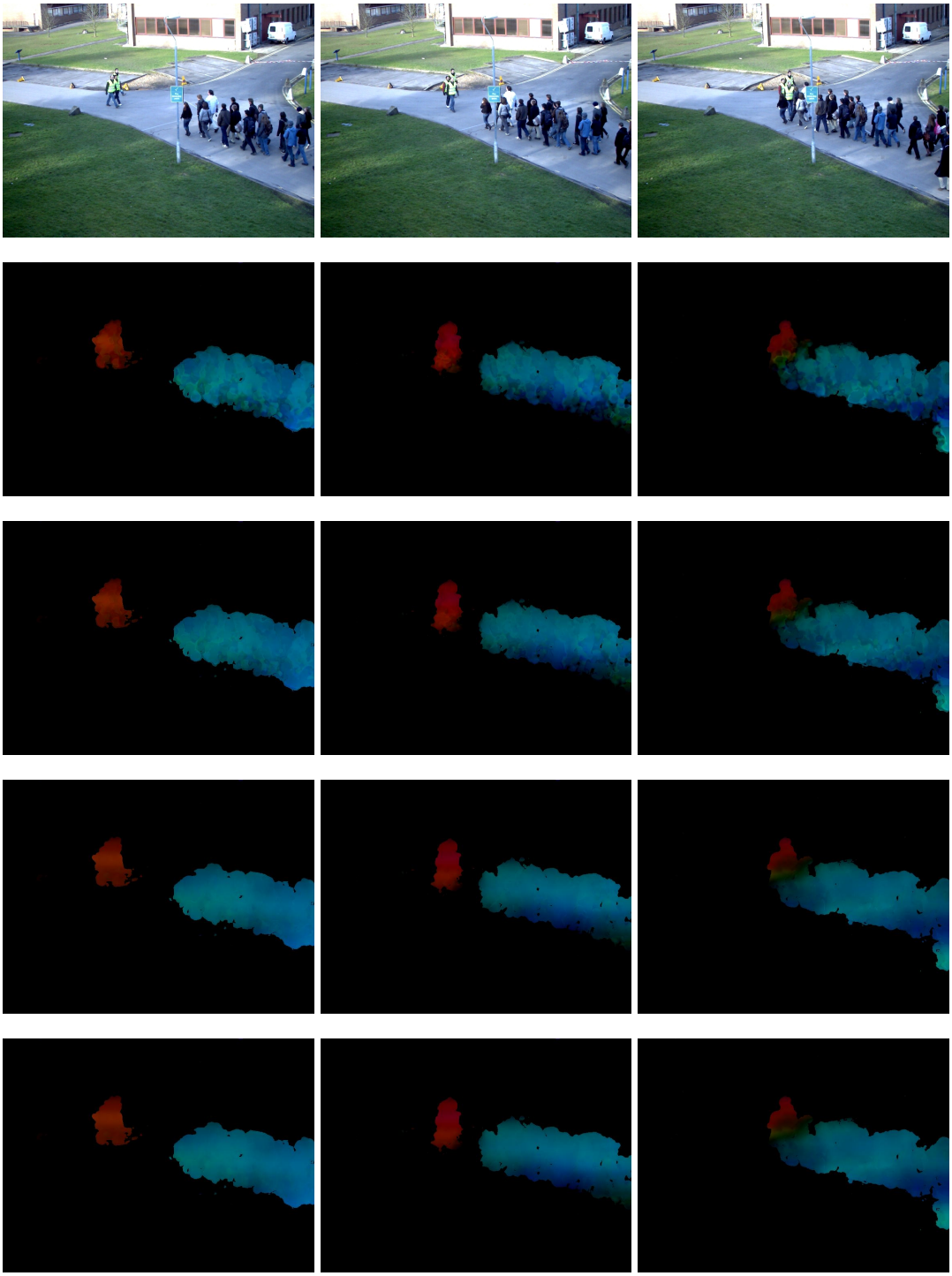
The motivation for this analysis is that pedestrians normally choose the shortest

Figure 3.8: Influence of the height h in final result of crowd flow estimation. First row are frames sample, the second row is using $h = 0m$, the third row is using $h = 0.9m$, the fourth row is using $h = 1.8m$ and the last row is using h based in Eq. (3.3).



route to their next destination, as noted in psycho-social studies such as (MOUSSAÏD; HELBING; THERAULAZ, 2011), which tends to lead to mostly linear paths without significant direction jittering. As a smoothness measure of the path, we compute the angular distance between each pair of displacement vectors for the same particle across

Figure 3.9: Influence of the value of parameter α in final result of crowd flow estimation. First row are frames sample, the second row is using $\alpha = 0.50$, the third row is using $\alpha = 0.75$, the fourth row is using $\alpha = 1.50$ and the last row is using $\alpha = 2.00$.



three consecutive frames (e.g. $t-1$, t and $t+1$). We then estimate the mean and standard deviation of the angular distance considering all trajectories and frames. To avoid any bias introduced by camera perspective, these metrics were obtained using the smoothed flow in the WCS, given by $\mathbf{v}_w^s(\mathbf{u})$ (i.e. prior to projecting it back onto the image).

Table 3.2 summarizes the results for the two datasets using the three baseline methods, without and with the proposed filtering approach. As can be observed, the use of our post-processing method reduced both the average and standard deviation values for all baseline methods in both Marathon and PETS2009 sequences (the reduction was over 50% for the Marathon sequence). Table 3.2 also shows that without post-processing, Farnebäck's method produces high angular variations (e.g. almost twice the variation of Brox's method for the Marathon dataset). When coupling the post-processing approach, these differences are much smaller, which corroborates the qualitative visual validation.

Table 3.2: Comparison of the average angular variation and standard deviation of trajectories obtained by particle advection.

Dataset	Algorithm	Mean Angular Distance		Mean of Standard Deviation	
		Original	Filtered	Original	Filtered
Marathon	Brox	10.65°	5.26°	19.32°	8.91°
	Classic+NL	13.37°	6.75°	23.32°	12.10°
	Farnebäck	19.11°	6.92°	33.48°	12.47°
	DeepFlow	5.21°	3.15°	25.10°	7.1°
PETS 2009	Brox	26.95°	15.65°	31.36°	19.37°
	Classic+NL	21.36°	12.25°	24.30°	17.30°
	Farnebäck	30.13°	20.39°	36.15°	24.79°
	DeepFlow	18.35°	8.4°	21.63°	13.57°

For the sake of illustration, Figure 3.11 shows a visual comparison of the tested methods. The visual analysis indicates that the post-processing approach indeed generates smoother trajectories, corroborating the results shown in Table 3.2. More importantly, Figure 3.11 and Table 3.2 indicate that the worst result with post-processing is still better than the best result without post-processing, reinforcing the fact that it is possible to use a fast (but less robust) method as baseline, and still obtain consistent results.

3.3.4 Quantitative Evaluation based on Event Detection

As mentioned in Section 2.3, several crowd event detection methods use crowd flow information as input. In this particular context, the quality of a given crowd flow method estimation could be implicitly assessed by evaluating the accuracy of the event detection method, which can be measured objectively. In this work, we evaluated the quality of the our crowd motion change detection approach presented in (ALMEIDA et al., 2017) using different crowd flow estimation methods, with and without post-processing.

The approach presented in (ALMEIDA et al., 2017) considers a calibrated static

surveillance camera, and assumes that the filmed region is roughly planar. Given the optical flow restricted to foreground pixels, the inverse perspective mapping is applied based on the known camera parameters to obtain the displacement vectors in the world coordinate systems, using the ground plane homography. Then, a 2D (normalized) histogram (speed versus orientation) is build based on optical flow in world coordinates at each frame, encoding the global motion of the crowd. To detect changes, a similarity vector is generated at each frame with previous frames used in the comparison, and a correlation operator is used to measure the similarity of histograms. The detection of changes in the crowd behavior is based on the temporal stability of the crowd behavior at the frame, defined as a weighted average of similarity vector, in which the weights decay exponentially for older frames. When temporal stability is low, the similarity between the current frame and the previous ones tends to be small, and a change behavior is detected.

We used the dataset PETS2009 14-16 in our analysis, which is divided in two parts: each part contains a different scene. In part 1, people move from the right to the left of the scene, and start running at frame 38 (ground truth value for motion change). Figure 3.12 shows the initial frame, the ground truth frame, the last frame, as well as the frames at which motion change was detected based in (ALMEIDA et al., 2017) using different crowd flow estimation methods.

Table 3.3 shows the detection frames for both parts of the PETS2009 14-16 dataset. It is interesting to note that the crowd flow obtained using Farnebäck+Our produced the lowest detection lag in both part 1 and part 2, which is better than using Brox (as originally explored in (ALMEIDA et al., 2017)) at a much shorter execution time. The table also indicates that the proposed processing approach reduces (or produces the same) detection lag when compared to the flows obtained by the baseline approaches.

These results agree with the findings indicated in Section 3.3.3: the proposed post-processing approach allows the use of simpler (and faster) optical flow baseline approaches, producing better results than more sophisticated (and slower) optical flow methods.

Table 3.3: Frames when crowd motion change was detected using (ALMEIDA et al., 2017) with different optical flow methods in the scene PETS2009 14-16 divided into two parts, part 1 of frames 0 to 107 and part 2 of frames 108 to 222.

Algorithm	14-16 Part1	14-16 Part2
Ground Truth	38	56
Brox	48	77
Classic+NL	not detected	77
Farnebäck	54	77
DeepFlow	48	77
Brox+Our	48	77
Classic+NL+Our	54	77
Farneback+Our	45	72
DeepFlow+our	48	72

Figure 3.10: Influence of the value of parameter D_{max} in final result of crowd flow estimation. First row are frames sample, the second row is using $D_{max} = 0.25$, the third row is using $D_{max} = 0.50$, the fourth row is using $D_{max} = 0.75$, the fifth row is using $D_{max} = 0.95$ and the last row is using $D_{max} = 1.00$.

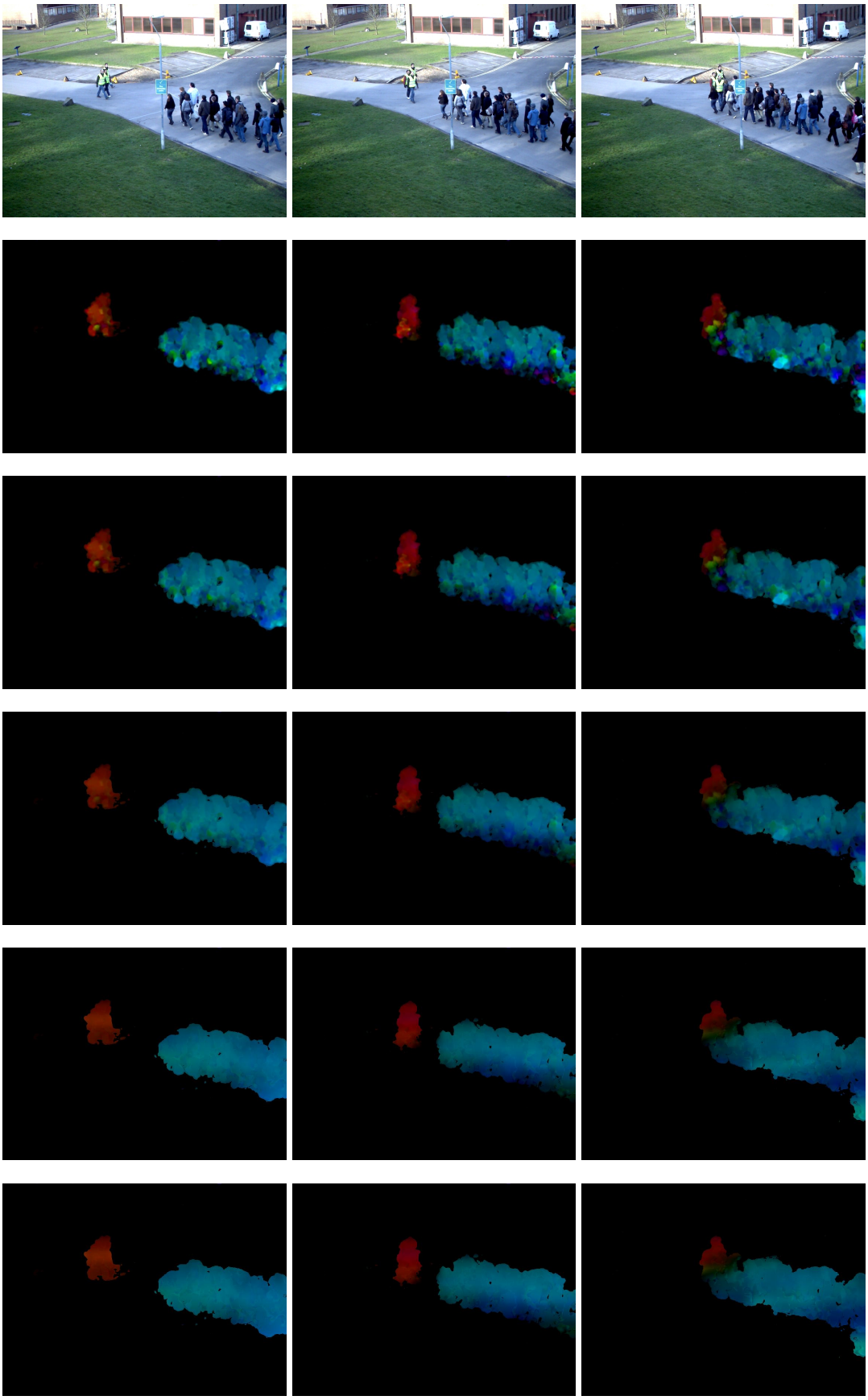
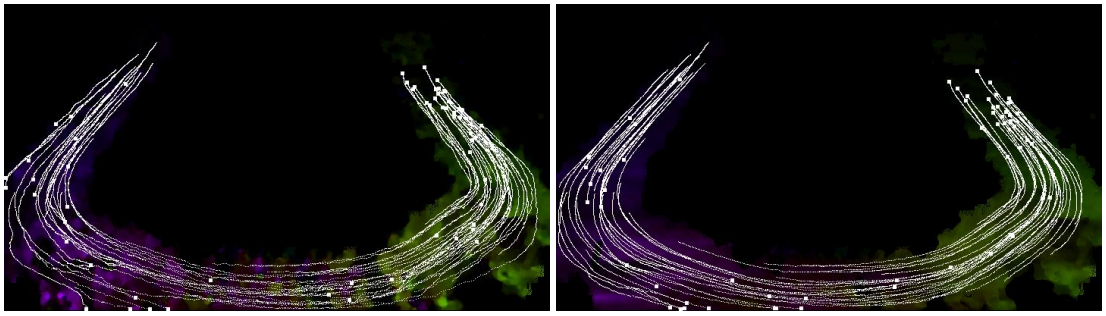
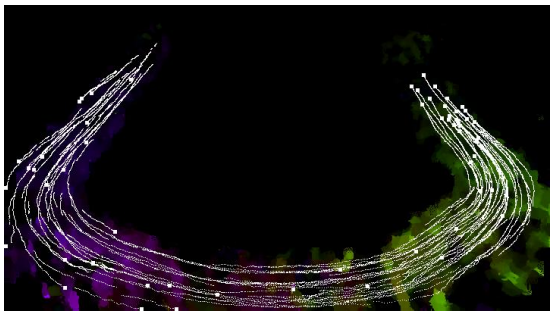


Figure 3.11: Visual comparison between the trajectories of particles on optical flow in world coordinates estimated with state-of-art methods and the trajectories of same particles on crowd flow in world coordinates estimated using our post processing method.

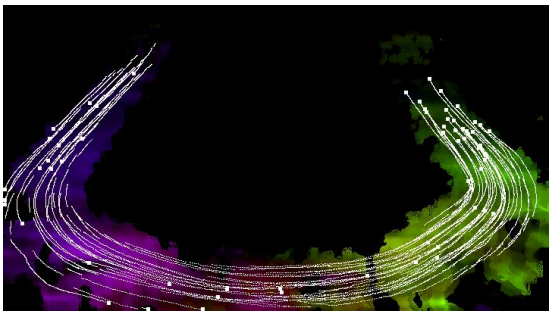


(a) Brox

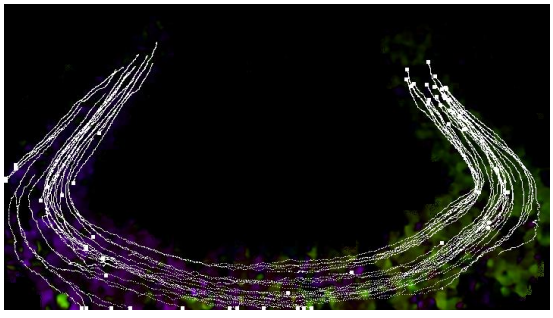
(b) Brox + Our



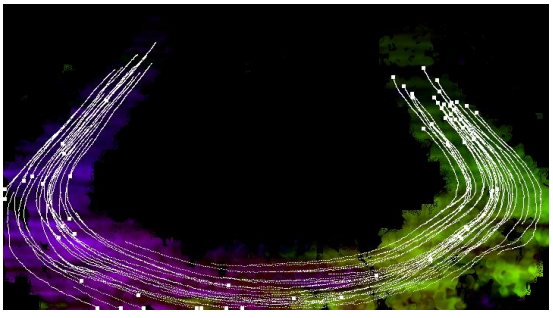
(c) Classic+NL



(d) Classic+NL + Our



(e) Farneback



(f) Farnebäck + Our

Figure 3.12: PETS2009 14-16 - part 1, a crowd moves from the right to the left, in (b) we present the ground truth frame related to the event and in (c), (d), and (e) the frames where at least an optical flow method detect the change in crowd behavior.

(a) Frame 00

(b) Frame 38 - ground truth

(c) Frame 45 - Farneback+our

(d) Frame 48 - Brox, DeepFlow, Brox+our and DeepFlow+our

(e) Frame 54 - Farnebäck and Classic+NL+Our

(f) Frame 90

4 LOCAL ANOMALY DETECTION

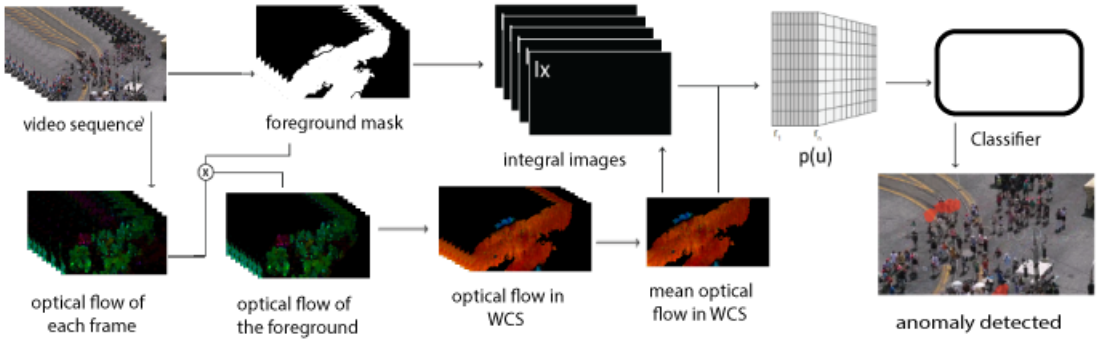
The results shown in the last chapter indicate that exploring local neighborhood information can improve crowd flow estimates, which in turn can be used in the context of abnormality detection. In this chapter, we extend the analysis by computing the local flow adherence using multiple personal regions, and explore these multiscale features for abnormality detection. The main hypothesis is that the crowd flow in a stationary structured crowd presents the same local motion patterns within small temporal windows (which is the core notion of stationarity). It is important to mention that a recent trend in several computer vision tasks is to use deep learned features. However, since it is not to our knowledge the existence of publicly available datasets with crowded scenes and annotated data, we decided to use hand-crafted features.

As the approach presented before, here we also explore distances and flow vectors in the WCS, which allows the use of the personal distances introduced by Hall (1966) – as the crowd motion method presented in the previous chapter – and also provides flexibility to a wide variety of camera setups. In fact, the exact same scenario captured by two different cameras might present video sequences with significant visual differences, as illustrated in Figure 4.1.

Figure 4.1: Same scenario filmed by two cameras extracted from the PETS2009 dataset (FERRYMAN; SHAHROKNI, 2009).

Figure 4.2 shows an overview of the proposed method for feature extraction and abnormality detection. Each step of the proposed approach is detailed next.

Figure 4.2: Overview of the proposed method: i) optical flow from two adjacent frames and foreground mask; ii) valid flows in world coordinates; iii) mean stationary temporal flow; iv) similarity of each pixel with its neighborhood using integral images; v) classification and post-processing.



4.1 Estimating the stationary crowd flow

Given a short video clip depicting a stationary crowd, we compute the optical flow $\mathbf{v}^t(\mathbf{u})$ for each pair of adjacent frames $t-1$ and t, where $\mathbf{u} = (u, v)$ represents the image coordinates. We use the ground plane homography to estimate the pixel motion in the WCS $\mathbf{v}_w^t(\mathbf{u})$, and restrict the analysis to flow vectors related moving objects in the scene – assumed to be mostly related to humans – by computing a binary foreground mask $f_g^t(\mathbf{u})$ at each frame. It is importante to note that this mask can be obtained by using a generic background removal method if the video clip is long enough (as in the previous chapter), but it might fail when the scene is dense and there are not sufficient frames to estimate the background. In those cases, the map $f_g^t(\mathbf{u})$ is obtained by thresholding the magnitude of optical flow vectors in world coordinates (assuming that low magnitude flow vectors relate to stationary regions).

The motion in the whole video clip is summarized by a single optical flow image

$$\mathbf{v}_w^a(\mathbf{u}) = \frac{1}{\#f_g(\mathbf{u})} \sum_{i=1}^{T} \mathbf{v}_w^t(\mathbf{u}), \tag{4.1}$$

where T is the number of frames in the clip (a sequential subset of the full video), and $\#f_g(\mathbf{u})$ represents the number of foreground pixels at location $\mathbf{u}$ for these frames. Assuming that the motion flow is stationary, this temporal average provides a summary of the motion flow along the duration of the clip. In the limit case ($T = 1$), we assume stationarity across only two adjacent frames, which typically leads to noisier flow and more susceptibility to outliers, but at the same time being able to handle quicker behav-

ior changes. As the temporal window T increases (the other limit is the full length of the video sequence), the summarized optical flow is smoother (and stronger stationarity is assumed), so that possible abnormal events with very short duration might be missed. Figure 4.3 shows a visual comparison between different values to T – 1 frame, 7 frames, 14 frames, 50 frames and 100 frames. The scene shows a stationary crowd throughout all the scene, which is moving from right to left, with a few people in opposite direction close the left-bottom corner.

Figure 4.3: Same scene with fixed initial frame but different T values: (a) the first frame of clip. (b) $T = 1$ frame. (c) $T = 7$ frames refers to $1s$ of the video, (d) $T = 14$ frames refers to $2s$ of the video, (e) $T = 50$ frames refers to half video length, and (f) $T = 100$ frames, full video clip.

4.2 Computing local flow similarity

Based on $\mathbf{v}_w^a(\mathbf{u})$, the next step is to analyze the neighborhood of each pixel belonging to the foreground and calculate the flow similarity of each pixel with its neighbors. Using different neighborhood sizes allows us to obtain a multi-scale similarity value for each flow vector, as explained next.

People are mostly affected by their nearby neighbors, which can be characterized by a spatial "influence region". Hall (1966) studied the expected relationship between two people based on their distances (from intimate to public), so that concentric circles with different radii characterize the different "personal spaces", or proxemics: intimate, personal, social and public. The smaller the radius r, the stronger is the expected relationship between the person under analysis and the neighbors. Hence, different crowd flow consistency levels can be achieved by varying r.

As done in Chapter 3, we approximate the "influence regions" A_r with radius r as a square region with dimensions $2r \times 2r$ to simplify the computations using integral images. Given a foreground pixel $\mathbf{u}$, and assuming that the ground is roughly planar, we use the planar homography H_z (corresponding to a height z) to project $\mathbf{u}$ into a world point $\mathbf{x} = (x, y, z)$. Although it is difficult to obtain the actual value for z, the typical heights of a person are limited. Since obtaining the ground plane homography is simple (if the camera parameters are not known, the homography can be estimated using only four planar points with known coordinates), we used $z = 0$ in the conversion for all image points, and use the ground plane homography $H = H_0$.

We then consider a "influence region" $A_{\mathbf{x},r}$ centered at world point $\mathbf{x}$, parallel to the ground plane, and project it back to image coordinates. Hence, the same influence region $A_{\mathbf{x},r}$ in the world leads to different regions $A_{\mathbf{u},r}$ in the image domain due to perspective issues, as illustrated in Figure 3.2. Since the camera is assumed to be static, the ground plane homography remains constant in time. To reduce the computational burden, the projections of $A_{\mathbf{x},r}$ centered at all possible image pixels are pre-computed a single time and stored in a LookUp Table (LUT).

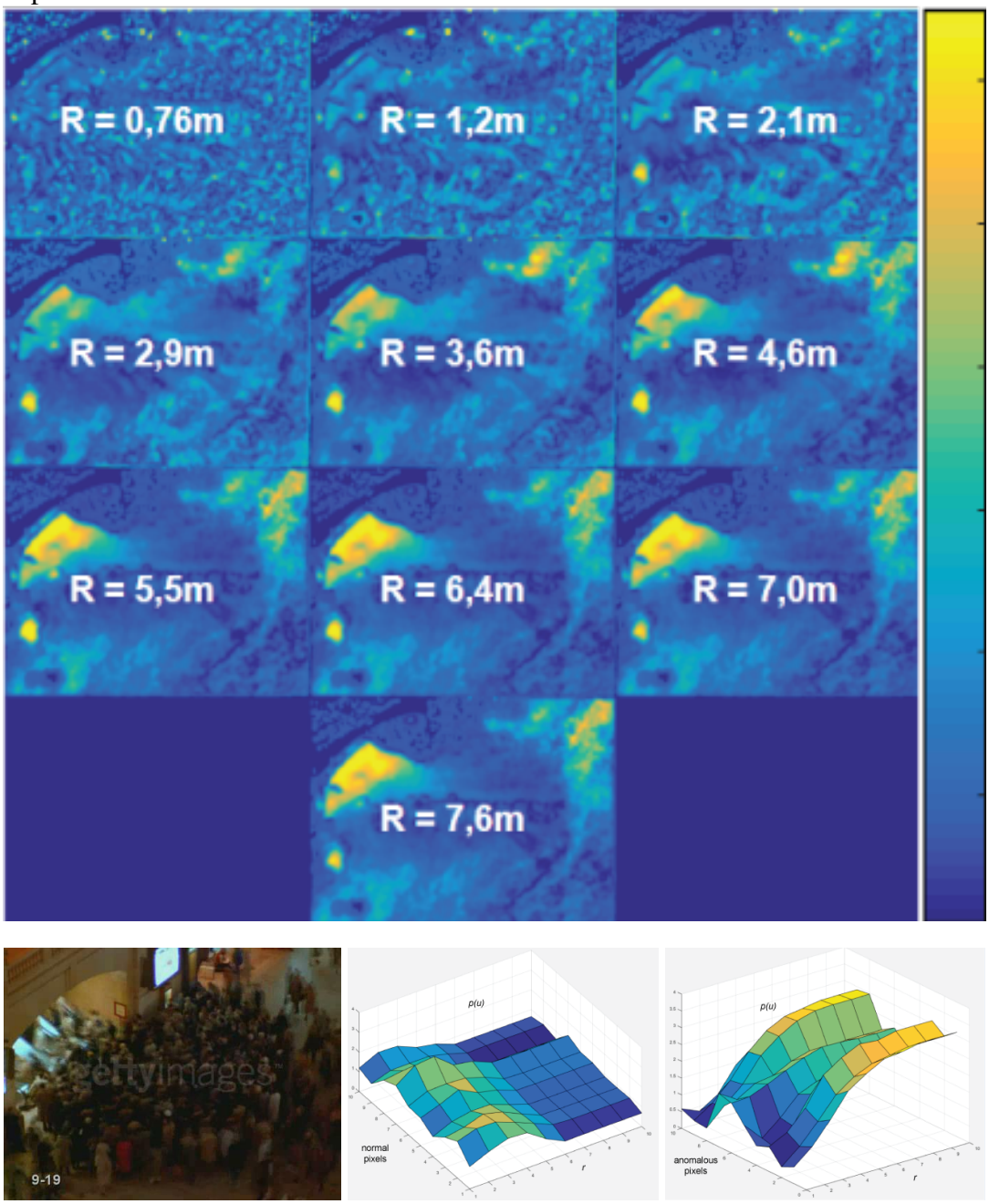
The analysis of the crowd is also performed in the WCS, to alleviate the distortions caused by camera perspective. To that end, we first project the optical flow at each foreground pixel $\mathbf{u}$ to the WCS, as shown in Eq. (3.1).

To estimate the local flow coherence, we compare the optical flow in the WCS with the neighboring flows based on the projections of the corresponding region of in-

fluence $A_{\mathbf{u},r}$. In a typical structured crowd flow, $\mathbf{v}_w(\mathbf{u})$ should be roughly homogeneous within $A_{\mathbf{u},r}$, following the macroscopic approach to crowd analysis. Our goal is to explore the local coherence of the optical flow local $\mathbf{v}_w^a(\mathbf{u})$ within neighborhood regions with different radii $r \in \{r_1, r_2, ..., r_n\}$, where the values r_i relate to a set of pre-defined radii of influence regions – for instance, the distances proposed by Hall (1966). We propose a local adherence measure $p(\mathbf{u}, r_i)$ based in Mahalanobis distance given in Eq. (3.5), but using the mean vector and covariance matrix of the optical flow in region $A_{\mathbf{u},r_i}$. This process generates a feature vector $\mathbf{p}(\mathbf{u}) = (p(\mathbf{u}, r_1), p(\mathbf{u}, r_2), ..., p(\mathbf{u}, r_n))$, that contains the motion similarity between each pixel $\mathbf{u}$ and its neighborhoods with varying radii.

In order to keep computational complexity of the proposed method low, we actually use the bounding box of each region $A_{\mathbf{u},r_i}$, since the use of rectangular regions allows fast computation of the mean vector and the covariance matrix, similarly to the approach used in Chapter 3. Note that the computation of intermediate integral images is done only once, and the computation of the covariance matrices at multiple raddi – Eqs. (3.7) to (3.11) present constant complexity, meaning that using several influence regions at the same time has very low computational impact over using a single regions. Note that each radius r_i corresponds to a different region of influence. For smaller radii, we typically expect more flow coherence within the neighborhood (i.e. smaller values for $p(\mathbf{u}, r_i)$), since the closest people tend to present the strongest relationship. As the radius increases, the relationship weakens, but the behavior of $p(\mathbf{u}, r_i)$ depends on the crowd structure: for very structured crowds, it tends to keep low, since motion patterns are coherent in a wider neighborhood (e.g., a pack of people moving in a single direction); on the other hand, in regions with less local structured motion, $p(\mathbf{u}, r_i)$ tends to increase with r_i (e.g., a row of people moving against a crowd). Figure 4.4 shows the multiscale values of $p(\mathbf{u}, r_i)$ for a a bottleneck behavior. More precisely, one frame of the scenario is shown in the bottom-left, where we can observe a dense crowd moving along a relatively large hallway to a subway entrance (left of the image). In the hallway, people have some freedom to move, but their motion is restricted as they get closer to the entrance. This change in motion patterns is captured by the values $p(\mathbf{u}, r_i)$, shown on top of Figure 4.4: for small radii, $p(\mathbf{u}, r_i)$ is small at all regions. As the radius increases, the hallway – region prior to the bottleneck – of the flow still presents smaller discrepancy values, but at the entrance of the escalator of the subway (to the left), the discrepancy values increase.

Figure 4.4: (a) The images show p(u) using different values to r, (b) example a frame of the scene (c) show a plot of a region in the center of scene, where people have similar move, and (d) show a plot of pixels in the region where people cross the span and increase the speed.



4.3 Detecting local anomaly

The last step toward anomaly detection is to consider the values $\mathbf{p}(\mathbf{u})$ for each pixel $\mathbf{u}$ as feature vectors, and then use a binary classifier to identify local abnormalities at each image pixel during the temporal window T. It is important to note that our main contribution is the design of a psych-socially aware feature vector for crowds, and several

classifiers can be used with the proposed features, ranging from unsupervised options (ALSABTI; RANKA; SINGH, 1997; GUO et al., 2003) to supervised (HEARST et al., 1998b; RODRIGUEZ; KUNCHEVA; ALONSO, 2006; KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

In this work, we perform an analysis using a set of more traditional classifiers such as Support Vector Machine (SVM) (HEARST et al., 1998b), Random Forest (RF) (BREIMAN, 2001) and Extremely Randomized Trees (ET) (GEURTS; ERNST; WEHENKEL, 2006), as well as more "trendy" classifiers such as Dense Neural Networks (MCCLELLAND et al., 1986) and Recurrent Neural Networks (RNN) (HOCHREITER; SCHMIDHUBER, 1997). To analyze how well $\mathbf{p}(\mathbf{u})$ performs as feature vector, we explore the five classifiers independently.

An issue when using neural networks in the context of crowd analysis is the low amount of annotated publicly available datasets. In particular, this work focuses on dense structured scenes, which are not common in existing datasets (especially containing "abnormal" patterns). It is also important to note that the classifier is applied to each pixel independently, so that a single scene might provide several image patches for training.

Recurrent Neural Networks are commonly used in problems that involve sequential data, and prototypical applications are text and speech recognition. The feature vector $\mathbf{p}(\mathbf{u})$ used in this work encodes the local motion discrepancy at regions with increasing radii, which inherently encodes a sequential reasoning and motivates the use of an RNN. The RNN architechture adopted in this work was composed by Long short-term memory (LSTM) units and trained with backpropagation through time. Our LSTM network has a simple shallow architecture, with 256 hidden units in its only hidden layer, using hyperbolic tangent function as activation function, and exploring a softmax operation in the final layer to classify each pixel between the two classes (normal or anomaly).

We also explore a dense neural network with 4 hidden layers within 128, 256, 512 and 1024 hidden units. To deal with overfitting we add dropout layers and L2 regularization. As in RNN, we use softmax operation in the final layer to classify between normal and abnormal, but use Leaky ReLU as activation function in hidden units. The architecture of this network is detailed in Table 4.1.

Table 4.1: We used this MLP model to analyze and explore the proposed feature vector.

Layer	Output	Shape	Param #
dense_1 (Dense)	(None,	128)	1408
leaky_re_lu_1 (LeakyReLU)	(None,	128)	0
dropout_1 (Dropout)	(None,	128)	0
dense_2 (Dense)	(None,	256)	33024
leaky_re_lu_2 (LeakyReLU)	(None,	256)	0
dropout_2 (Dropout)	(None,	256)	0
dense_3 (Dense)	(None,	512)	131584
leaky_re_lu_3 (LeakyReLU)	(None,	512)	0
dropout_3 (Dropout)	(None,	512)	0
dense_4 (Dense)	(None,	1024)	525312
leaky_re_lu_4 (LeakyReLU)	(None,	1024)	0
dense_5 (Dense)	(None,	2)	2050

We also explore Support Vector Machine with Radial Basis Function (RBF) kernel and two methods based on decision trees: a Random Forest Classifier composed by Random Trees and an Extremely Randomized Tree Classifier also composed by Decision Trees Classifier. Both methods explore the idea of using a number of decision trees to improve the classification and reduce overfitting to training data, but differ in the way trees are built. In experiments section these five models results are analyzed, and we explain more about some implementation and fit details.

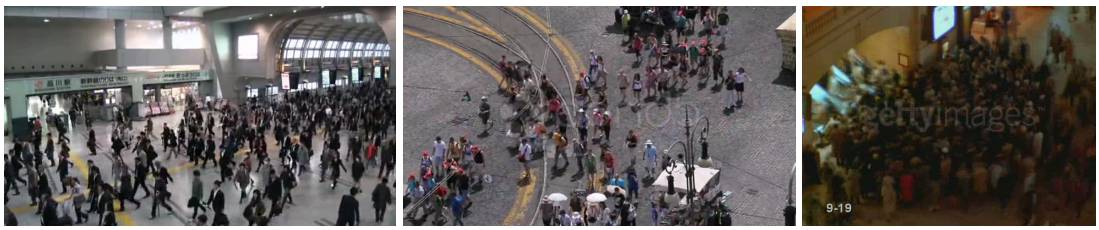
4.4 Experimental Results

Our approach considers that an anomaly in a structured crowd flow is a motion pattern incoherent with neighboring pixels, based on a macroscopic view of the crowd. Some existing methods try to detect and localize anomaly based on unseen behavior (e.g., someone carrying a gun or riding a bike through people), mostly focusing on low-medium density crowds. As mentioned before, our work deals only with motion patterns, so that guns would not be detected, and other vehicles (such as bike) would only be identified if they generate an abnormal motion pattern. Furthermore, it is important to note that the

definition of anomaly in publicly available crowd datasets is different from ours, leading to lack of annotated ground truth data. Based on these findings, we performed two sets of experiments: i) we analyze results of our proposed method (using five different classifiers); ii) we explore the variation in the value of T and its effects t on our method, and iii) compare our method of anomaly detection with others state-of-art methods, but in low-medium dense crowds.

We used some video clips of the CUHK crowd dataset (SHAO; LOY; WANG, 2014) and UCF Crowd Segmentation dataset (ALI; SHAH, 2007a) to train our model. More precisely, we used only four scenes (1_34_2-25-2, 1_34_6-25-1, and 1_34_6-25-2 from CUHK and 9-19_l from UCF - some frames samples are shown in Figure 4.5) that present some groups with abnormal behavior (according to our definition), resulting in $\approx 750,000$ training vector features. These data are divided in two groups - normal and abnormal -, from which $\approx 40,000$ are manually labeled as abnormal.

Figure 4.5: Frames extracted from scenes of CUHK and UCF dataset that were used to train classifiers models.

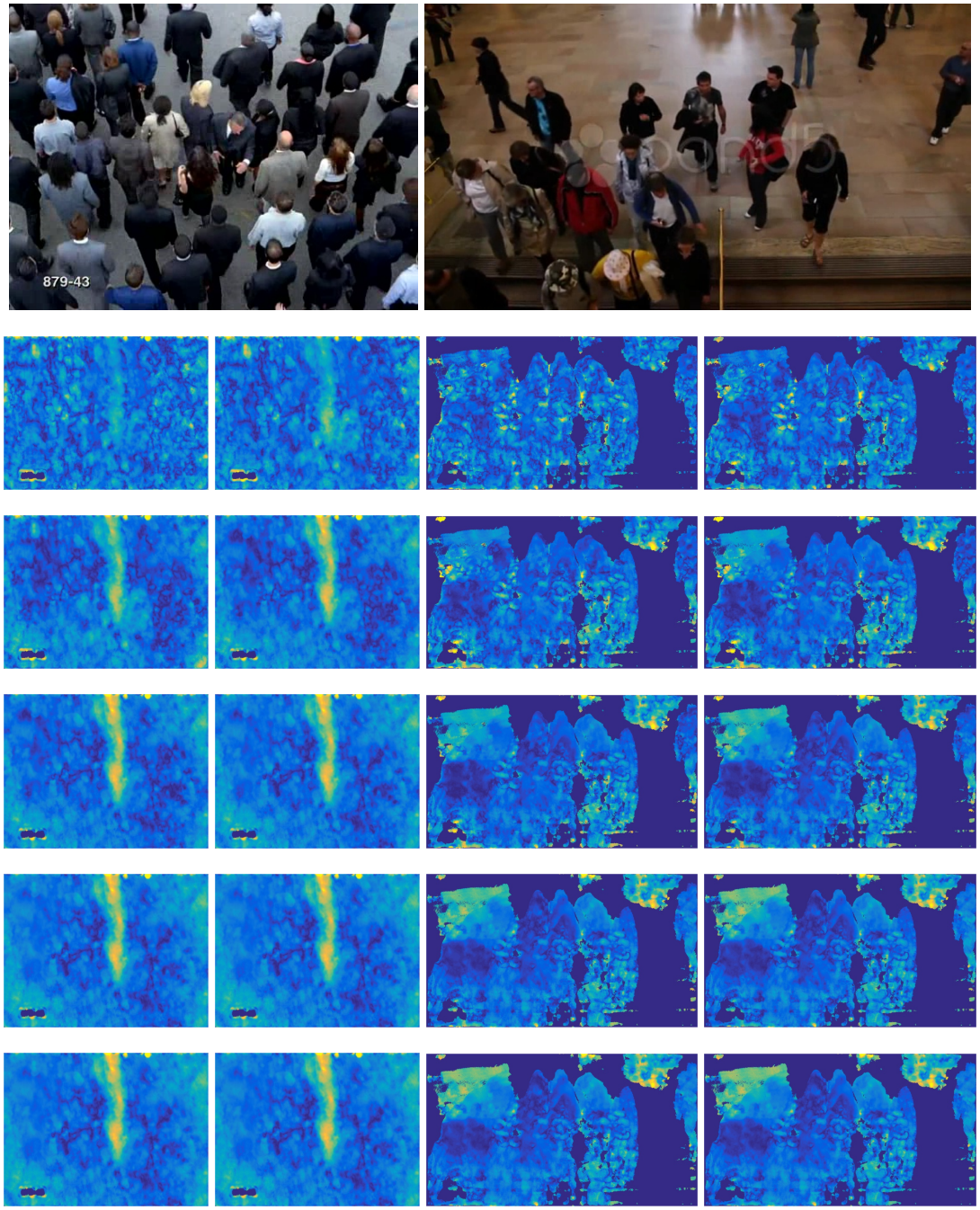


In all experiments, for each pixel we analyzed different circular neighborhoods with radii $r_i \in \{0.76m, 1.2m, 2.1m, 2.9m, 3.7m, 4.6m, 5.5m, 6.4m, 7.0m, 7.6m\}$, which were approximated by square regions with side $2r_i$ to speed up the computations using integral images. These values represent that $i \in \{1, 2\}$ relates to personal space, $i \in \{3, 4, 5\}$ with social space, and $i \in \{6, 7, 8, 9, 10\}$ with public space. The used values related to the personal space were the close and far phases limits, while to social space we also used close and far limits. Note that the distance between both limits increase a lot compared to the distance of personal limits, so we add a third phase – in fact a phase between close and far phase. This strategy was also adopted for the public space, where we used only the close phase of this space ($7.6m$) but added four phases between it and the social space's limit. In this way we guarantee that the differences between adjacent radii values (r_i to r_{i+1}) are more uniform in the feature vector.

Figure 4.6 presents a prototype frame for videos clip (1_3_6-4-1 and 1_879-43_l-2 from CUHK dataset) and the images corresponding to the feature vectors $p(\mathbf{u}, r_i)$ for

the chosen values of r_i (pixels in blue represent low dissimilarity, and pixels in yellow represent high dissimilarity). These images show us that anomaly regions have increasing dissimilarity as the neighborhood analyzed increases.

Figure 4.6: Dissimilarity motion images of two crowd scenes represented in image using a *parula* colormap. Rows 2-6 shown images that represent our vector, each image is referent to a neighborhood size r_i.

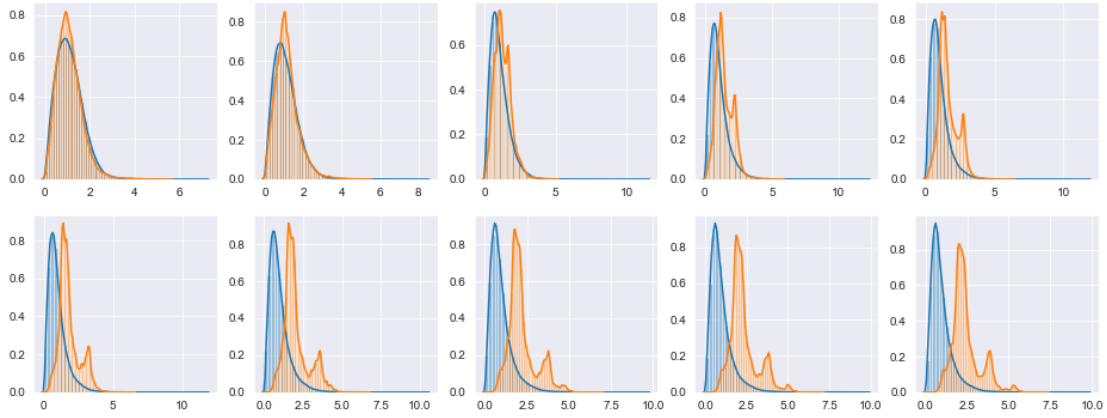


4.4.1 Detecting local anomaly in crowded scenes

To evaluate our local anomaly detection method, we used five classifiers: RNN, DNN, SVM, Random Forest, Extra Trees, as mentioned before. The tests presented here were performed in a subset of CUHK dataset, more precisely using scenes that present anomalies and were not present in the training set. Although the length of the video clips can be arbitrary, shorter videos are more stationary but suffer more impact of the noise of the optical flow, while longer videos are less impacted by optical flow noise but increase the computational cost and are less stationary. In our experiments, we used clips with 100 frames, which leads to approximately 15 seconds, considering a typical frame rate of surveillance cameras (7 FPS). The effects of this choice are discussed in Section 4.4.2. Since the datasets do not provide ground truth annotations for the anomalies considered in this work, results were evaluated qualitatively through visual inspection.

We can compare the distribution between classes normal and anomaly in the test data shown in Figure 4.7. More precisely, this figure shows, for each radius r_i ($i = 1, 2, ..., 10$), the histogram of features related to normal samples (blue curve) and abnormal samples (orange curve). It can be observed that the distribution is very similar for smaller radii r_i, indicating that normal and abnormal pixels have close values when analyzing the local adherence using the personal space. However, the dissimilarity of anomalies pixels increases with the growth of the analyzed neighborhood.

Figure 4.7: Distribution of $p(\mathbf{u}, r_i)$ of data test, divided in classes normal (blue) and anomaly (orange). The first row shown $p(\mathbf{u}, r_i)$ where $i = [1, 5]$ from left to right, and the second row shown $p(\mathbf{u}, r_i)$ where $i = [6, 10]$ also from left to right.



4.4.1.1 Recurrent Neural Network

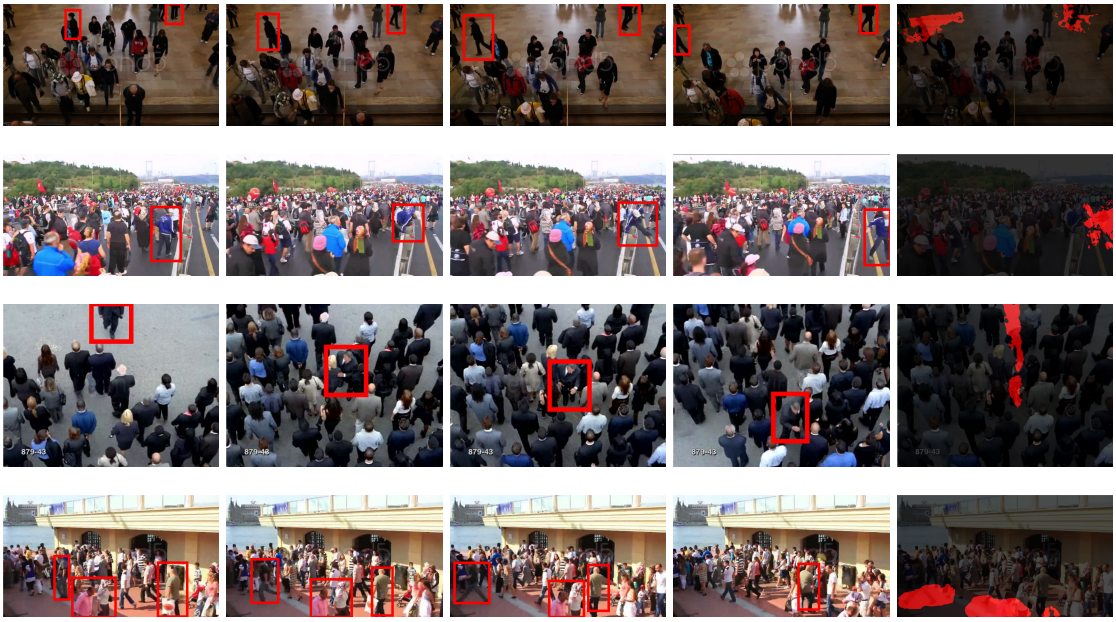
We analyzed our feature vector potential using an RNN based on a simple (shallow) architecture. Our goal in this setting is not to propose a complex classifier, but instead to demonstrate that our feature vector works even with simple classifiers. To avoid numerical instabilities, we scale our data to values in range $[0, 1]$. We evaluate experiments results by visual analysis, trying to detect all behaviors that are not consistent with main behavior in the scene. The network was trained for ≈ 200 epochs with early stopping regularization.

In Figure 4.8 we present four crowd scenes (1_3_6-4-1, 1_34_008681798-people-walk-europe-3, 1_34_009622329-passengers-kadikoy-port-2, and 1_879-43_l-2) with anomalous behavior with a wide variety of camera setups. The first three columns show some frames of the clips, with the anomaly highlighted with red rectangles. The last column shows the pixels classified as anomalous (in red) across the clip duration, noting that each row shows a different clip. The first one illustrates a medium-density crowd where people are walking to a stairway. Two people present different motion patterns (one is on top left and walking to the left image boundary, and the other is on top right walking to the right image boundary) and of neither them are walking according to the macroscopic crowd behavior. The result of our method is visually coherent to the anomalous motion. The second scene shows a protest on a road, and the anomaly is caused by a man that jumps over the traffic barrier. This scene has a camera setup very different from our training scenes, which were mostly top-down. Nevertheless, our method produces a detection blob that relates to the anomalous motion. The third row shows a crowd moving from the bottom to the top of the image, and a person walking in the opposite orientation. The RNN result correctly detects pixels around the person path as anomalous regions. In the last row, the scene shows a crowd moving in a mostly uniform way to embark a boat, with some people that do not conform to the crowd motion, such as a couple in the bottom walking in the opposite direction, a man that walks to the left and another man that starts at the bottom and then moves against the flow)

Our classifier correctly classify the regions related to couple and man walking to the left, but fails to detect the anomalous behavior of the third person. This happens because his motion ends up dwarfed by the crowd motion during the temporal average, which could be alleviated by using a shorter clip.

As indicated by our visual analysis, the proposed LSTM network trained with only four short video clips was able to correctly identify anomalous local motion patterns in

Figure 4.8: Crowd scenes that contain anomalies, and the output of our RNN that present red blobs as anomalous regions.



most of the tested videos. Considering the low variability of the training data these are promising results, particularly because the camera setups (and hence motion patterns in the image domain) vary considerably from video to video.

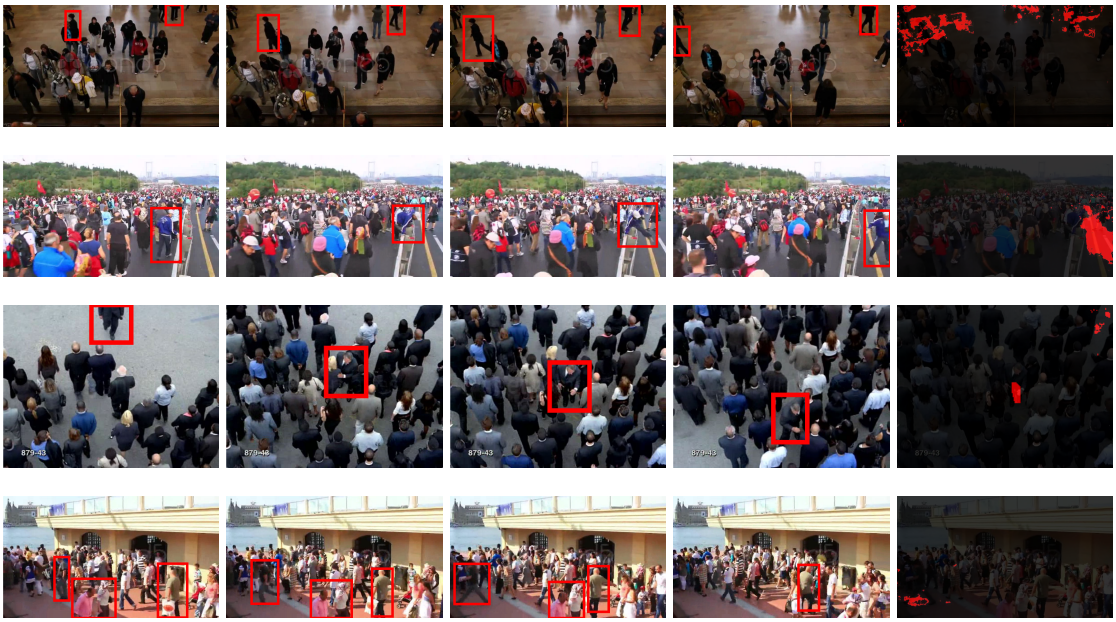
4.4.1.2 Dense Neural Network

In our experiments with a Multilater Perceptron (MLP), we opted by a more deeply architecture, using four hidden layers. This model was trained with the same set used for the RNN experiment, and ran for ≈ 300 epochs stopping due to early stopping regularization, with a learning rate of 10^{-4}. For comparison purposes, the sames video clips were used in tests, also evaluated by visual analysis.

In Figure 4.9 we present the same four scenes and first fours columns, marked with red squares anomalies in each frame as in Figure 4.8. The last column presents the results using MLP to detect and to localize local anomalies. In the first scene, our MLP model output correctly detects the anomalous regions, even if it does not produce fully connected blobs. In the protest scene, the network detected the abnormal behavior of the jumper. For the third scene, the MLP correctly detects a blob related to the men crossing the crowd, but also outputs false positives in the up-right corner. Different from the RNN results, this model does not result as abnormal all the path taken by the person. In the last scene our model just detects the person embarking not according to the flow but did not

detect other anomalies, as the couple against the flow.

Figure 4.9: Crowd scenes that contain anomalies, and the output of our MLP that present red blobs as anomalous regions.



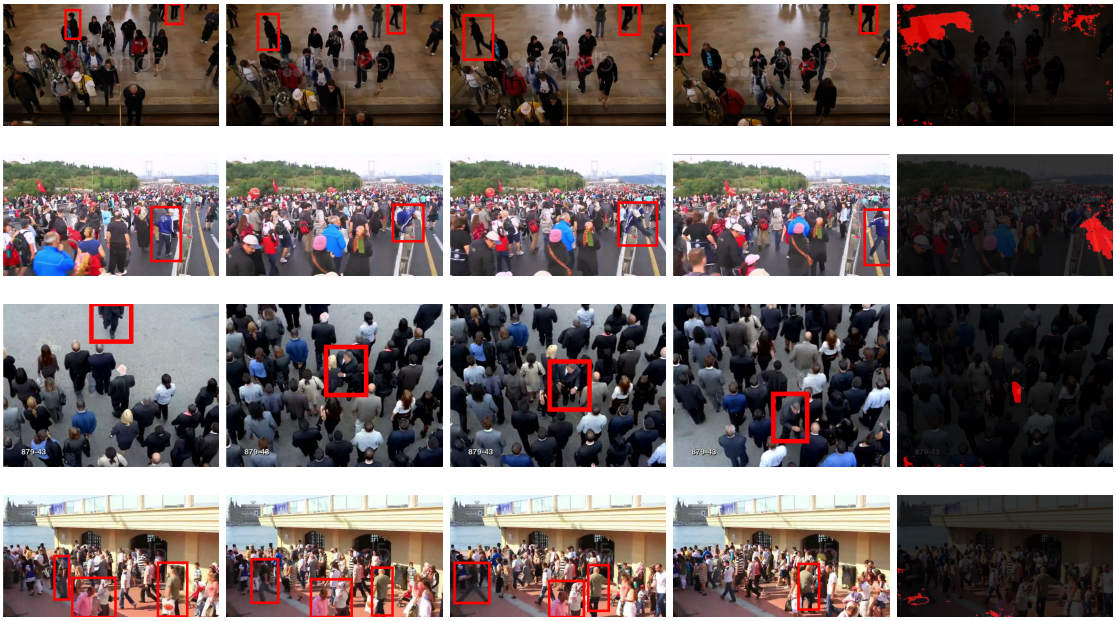
4.4.1.3 Support Vector Machine

To test a Support Vector Machine (SVM) as classifier, we experimented different kernels (Linear, Polynomial, Sigmoid and Radial Basis Functions – RBF), and report only the results with the best kernel (RBF). We use the same videos to train and test that were used in RNN and MLP experiments. To analyze experiments results using SVM we present in the last column of Figure 4.10 the output of SVM classifier, where the pixels anomalous are in red.

In the first scene, our SVM model detected both people that are walking in different direction of the main flow, but while the left person output is an almost fully connected blob, the output that represents the right person has more disconnected regions. In the protest scene, the classifier detected the anomalous behavior correctly, outputting as anomaly a region close of right border image, where the person concludes the jump. In the third scene, where a person cross a crowd, SVM outputs an anomalous region similar to the MLP output: just a small region in the center of the scene; this happens because when a person cross a crowd, the crowd tends to take a time to occupy the region just behind the person crossing it, so this output region has more frames in analyzed clip with the abnormal motion than with crowd motion. In the embarking scene, the SVM output

detected anomalies in two of three events, outputting anomalies in the couple region and in the region where the man embarks differently from other people, but not detected in the man that changes the direction of his motion.

Figure 4.10: Crowd scenes that contain anomalies, and the output of our SVM that present red blobs as anomalous regions.



4.4.1.4 Random Forest

Seeking to explore techniques based on decision trees, we model a random forest composed by 10 decision trees – number of trees chosen by tuning the hyperparameter –, trained with the same data test as the other classify models but without scaling it. An advantage of these classifiers is the possibility of understanding the importance of each feature to the output decision of our RF, which can be explored for feature selection. In fact, the importance array based on the training set was $\{0.059, 0.100, 0.074, 0.096, 0.072, 0.067, 0.088, 0.113, 0.061, 0.265\}$, indicating that our RF gives more importance to $r = 7.6m$ than others. Although this feature has significantly more importance, the next three most important feature are each in a different personal space (ordered in public, personal and social).

In Figure 4.11, we present the results obtained using our Random Forest trained. In the first scene the classifier output detects anomalies in regions referring to anomalies marked by red box in the images of the first four columns, even if the regions are not blobs fully connected. In the second scene, however, the Random Forest not only detected the

region anomaly correctly but also outputs a fully connected blob. In the third scene, our Random Forest trained detected an anomalous behavior in the center of scene, similar to the results obtained by SVM and MLP. In the last scene, the output obtained by Random Forest detects correctly the man embarking in the left image border, but did not detect the other two anomalous events.

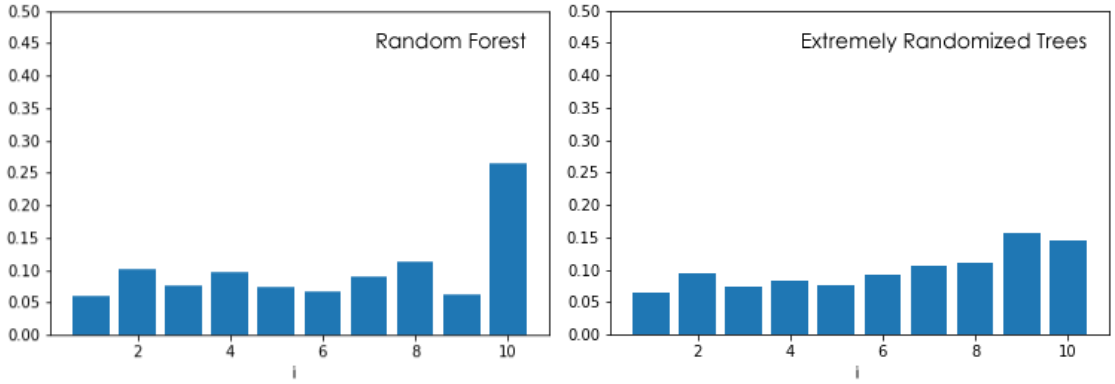
Figure 4.11: Crowd scenes that contain anomalies, and the output of our RF that present red blobs as anomalous regions.

4.4.1.5 Extremely Randomized Trees

Exploring another method based on decision trees, we model an Extremely Randomized Tree classifier with 100 decision trees, also chosen by hyperparameter tuning. This method consists of randomizing both: attribute – input variable used in a supervised learning problem – and cut-point choice while splitting a tree node. It can build totally randomized trees whose structures are independent from the output values of the learning samples. From the bias-variance point of view, the rationale behind the Extra-Trees method is that the explicit randomization of the cut-point and attribute combined with ensemble averaging should be able to reduce variance, and the usage of the full original learning sample rather than bootstrap replicas is motivated in order to minimize bias. As an RF method, it is built with some decision trees, and combine their outputs to produce the classification value; however, the process of building the classifier is different: ET splits nodes by choosing cut-points fully at random, and it uses the whole learning sample

to grow the trees. Also, we can analyse the importance of each feature in it decision, observing the importance array $\{0.065, 0.094, 0.074, 0.082, 0.074, 0.092, 0.105, 0.109, 0.157, 0.143\}$. Unlike the RF method, ET does not have a feature with considerably more importance than the others. In fact, Fig 4.12 shows the plots of the feature importance values for both RF and ET, which indicates a distribution closer to uniform for ET.

Figure 4.12: Graphs that shown the importance of each feature of our vector $p(\mathbf{u}, r_i)$ in decision making of the models (a) Random Forest and (b) Extremely Randomized Trees.



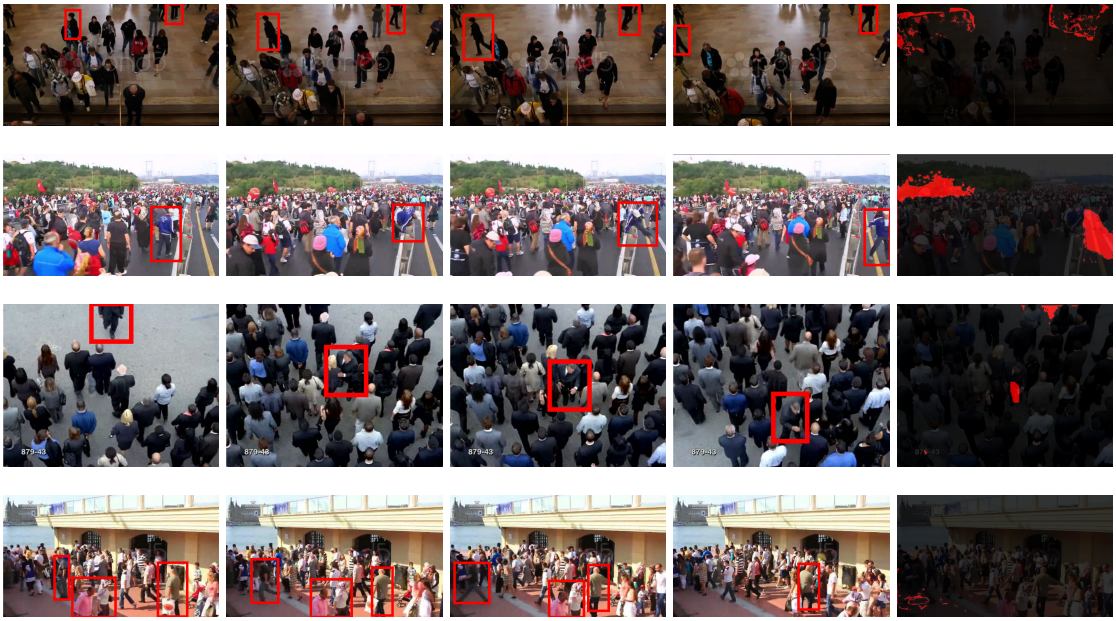
The visual analysis of ET classification results is shown in Figure 4.13. The Extremely Randomized Trees was the trained model that returns more false positives: in the first scene, the anomalous regions are in surrounding of the abnormal motions; in the protest scene it outputs two regions as anomaly: one, correctly, references the jumper, but the other is just above the crowd, where some things are moving, such as balloons and flags. In the third scene, our ET classifier also returns the anomalous region in the center of the scene correctly, but it also returns an anomaly region close to the top border. In the embarking scene, the classifier detected an anomaly region on bottom-left corner, close the motion of the man embarking, but as RF and MLP did not detect the couple walking in the opposite direction neither the man that changes his direction.

In order to compare results obtained using each classifier, we also performed a quantitative evaluation of all methods. We manually annotated ground truth detection values – masks of anomalous regions – and used it to calculate accuracy a_c as

$$a_c = \frac{T_P + T_N}{P + N}, \tag{4.2}$$

where P and N are amount of pixels anomalous and normal respectively (here, "positive" denotes anomaly), and T_P and T_N are the amount of pixels correctly classified as anomalous and normal respectively.

Figure 4.13: Crowd scenes that contain anomalies, and the output of our ET that present red blobs as anomalous regions.



We also compute a balanced accuracy ω (MOWER, 2005) as

$$\omega = \frac{r_e + s_p}{2}, \tag{4.3}$$

$$r_e = \frac{T_P}{P}, \tag{4.4}$$

and

$$s_p = \frac{T_N}{N}. \tag{4.5}$$

where r_e is the recall (our true positive rate) and s_p is the specificity (or true negative rate), which indicate the rate of correctly classified positive or negative samples, respectively. Note that the use of the balanced accuracy is important since we have imbalanced data (in all scenes we have many more normal pixels than abnormal ones).

The comparison is shown in Table 4.2, where it can be observed that the RNN classifier presents better results than others. RF, MLP, SVM and ET have similar values of balanced accuracy, with a slight advantage for RF. Also, ET presents the worst accuracy value, which happens because this one has significantly more false positives than the others.

Table 4.2: Quantitative comparison between our models regarding its results and a manually annotated ground truth.

Classifier Model	Accuracy	Balanced Accuracy
Recurrent Neural Network (RNN)	0.9855	0.8342
Dense Neural Network (MLP)	0.9389	0.6654
Support Vector Machine (SVM)	0.9343	0.6619
Random Forest (RF)	0.9486	0.6748
Extremely Randomized Trees (ET)	0.8843	0.6658

Figure 4.14 shows a qualitative comparison of the results produced by classifiers using the same clips depicted in Figures 4.8, 4.9, 4.10, 4.11 and 4.13. They indicate that the results produced by ET were less consistent, particularly in the second clip, which showed a big anomaly blob incorrectly detected (false positive). The RF, MLP and SVM presented similar results, with SVM obtaining better result in the first clip with blobs more connected. These four models result some false negatives in the last two scenes, showing smaller anomalous regions than expected, and some false positives in the third scene close to the top border. Comparing them all with RNN, the RNN produces anomaly blobs more defined with fully connected regions and less false positives. Moreover, only RNN detected the path traveled by man in the third scene, showing it as the best choice for this setup. As a drawback of all tested classifiers, none of our models detected the man that starts at the right-bottom corner and then moves against the flow in the last scene, this happen because the man's motion is not anomalous in all the video, he change his behaviour during the scene, in this way the mean flow is less affected by his anomalous behaviour.

We also evaluate the classifiers with respect to execution time of the classification (test) step, noting that the cost of feature extraction is the same for all classifiers. Table 4.3 shows the execution time of each step of anomaly detection method on one clip of dimensions 856×480 pixels. The execution times of the classifiers present a large variability, with Random Forest performing faster than others, while SVM being the slowest. Note that these times refer to the classification of the pixels performed sequentially, but since the classification of each pixel can be carried out independently from the others, a parallel implementation can be done.

Figure 4.14: First to fifth columns: anomalous detection using RNN, MLP, SVM, Random Forest and Extremely Randomized Trees.

Table 4.3: Execution time of each component of tested anomaly detection classifiers.

Method Step	Execution Time (s)
Optical Flow	≈ 0.1
Background Subtraction	≈ 0.006
Integral Images	≈ 0.024
Neighborhood Coherence (each r)	≈ 1.4
RNN Classifier	≈ 8.4
SVM Classifier	≈ 54
MLP Classifier	≈ 17.2
RF Classifier	≈ 0.5
ET Classifier	≈ 6

4.4.2 Exploring Clips Size

In Subsection 4.4.1, we analyze the results of each classifier using the mean flow references to each clip (i.e., T is selected as the total number of frames of each clip). In this subsection, we evaluate the impact of choosing different clip sizes. Since the goal here is only to evaluate the impact of T, we used only the best two classifiers (RNN and SVM) trained over the whole clip. The experiments were realized with three values of T: 1, 14 and 100 frames. This choice was based on a visual inspection of Figure 4.3, which shows the temporal mean flow of a scene using different windows T. The chosen values for T were the ones that present very distinct mean flows from each other. We also added a fourth scenario of test, that is using $T = 1$, but we use the post-processing

method presented in Chapter 3 to obtain a better estimation of the crowd flow. In this additional test, we chose the radius for the spatial neighborhood as 1.2m, which relates to Hall's personal distance.

In our analysis, each video was divided into a set of disjoint sub-clips with length T, and the analysis was performed independently within each sub-clip. When using $T = 1$, this strategy leads to a frame-wise analysis of the full clip, whereas selecting T as the total clip duration leads to an overall analysis of the clip. Figures 4.15 and 4.16 shows the results for RNN and SVM, respectively, in a medium-density scenario with two anomalies along the duration of the clip. In these experiments, we perform a qualitative visual analysis, and for each scene, we present three references frames t for visualization purposes.

We selected three references frames t (33, 50 and 79) for visualization purposes. They were selected so that each sub-clip is represented in the analysis with $T = 14$. When we used $T = 1$ without post-processing, we obtained a noisy output, which is improved when we post-process the optical flow, being possible obtain more consistent blobs, mainly in SVM. The RNN output with $T = 1$ is also very noisy, and even the post-processing step does not help. Comparing both classifiers in this scene, RNN presents slightly better result when using $T = 100$ (note that for $T = 100$ the results are identical for all frames, since they all belong to the same temporal window under analysis). On the other hand, the SVM classifier seems to be more suitable when we select smaller values for T. When we use $T = 1$ with raw optical flow (no post-processing), both classifier performed poorly.

Figures 4.17, 4.18, 4.19 and 4.20 show the results of both classifiers applied to two higher-density crowd scenes. In both scenes, neither of the two classifiers presented good results in the frame-to-frame analysis ($T = 1$), not being able to detect the correct anomalies. With the use of the post-processing step, the SVM classifier showed better result with $T = 1$ in the protest scene (Figure 4.18). We can note that both classifiers presented more consistent results as T is increased, and the best result was obtained by the RNN classifier when $T = 100$ again.

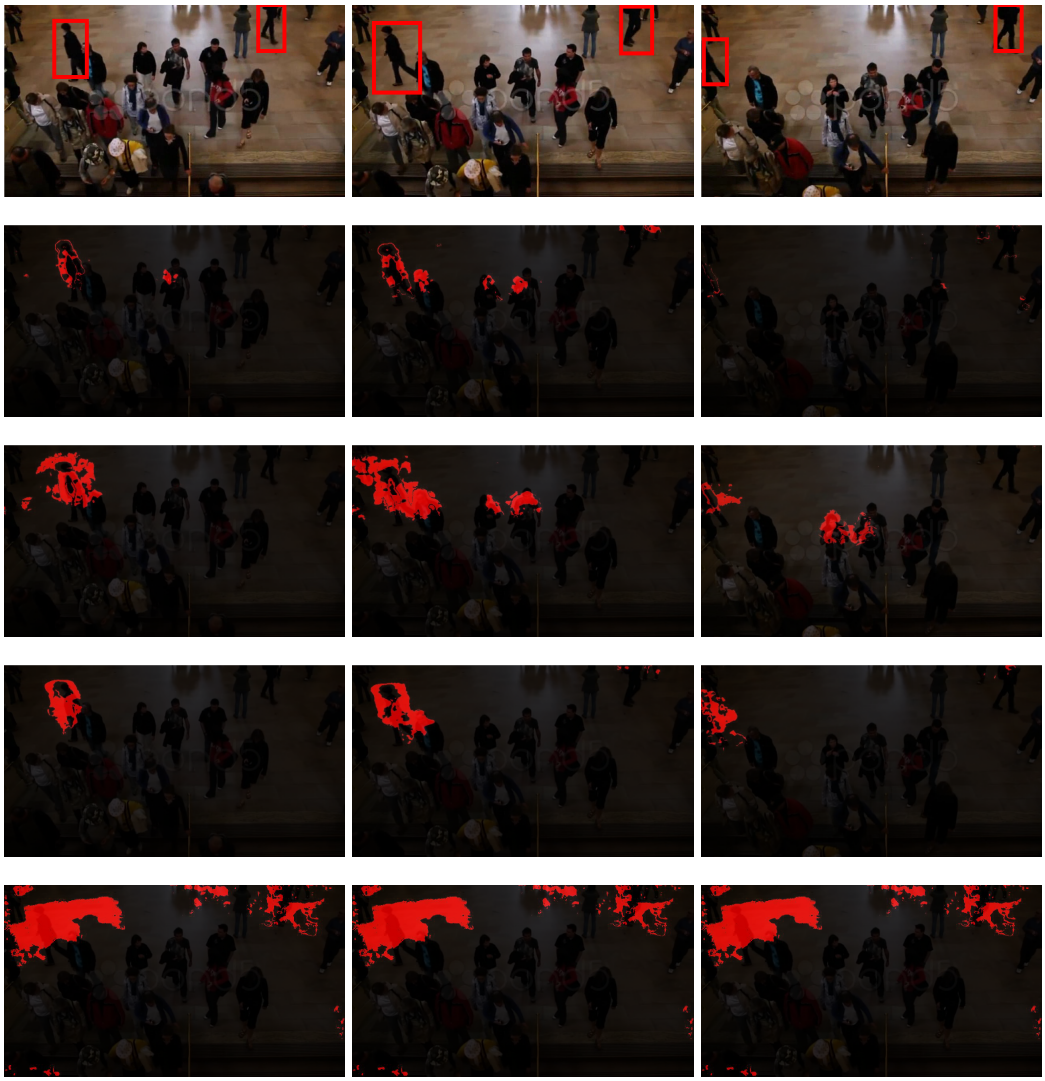
In Figures 4.21 and 4.22 we analyze the effects of changing T in a stationary crowd scene with an approximately top-down camera. Again, both SVM and RNN performed poorly when $T = 1$ without post-processing, since the results in these cases did not detect any correct anomaly. When we improve the crowd flow pos-processing the optical flow, the SVM output presents better results, detecting the anomaly, but with some

Figure 4.15: Anomaly detection output using RNN classifier in scene 1_3_6-4-1 of CUHK dataset. First row shown the reference frames - 33, 50 and 79 -, the second row shown results with $T = 1$, the third row $T = 1$ plus our post-processing method, the fourth row $T = 14$ and the last row $T = 100$.

false positives in bottom corners. When increasing T to 14, the results of both classifiers are improved, and we note that SVM produces less false positives, while the RNN detected correctly the anomalies. When we analyze all video, the main difference between classifiers results is that RNN detect as anomalous the whole path traveled by the man.

These experiments varying T indicate that even if the RNN classifier presents better results when we analyze all clip, the SVM might be a better choice when we decrease the temporal window T. They also show that the limit case $T = 1$ is not sufficient to capture the crowd behavior, even though the post-pocessing step might help when using SVM as the classifier. It is also important to note that $T = 100$ corresponds to a little over 3 seconds for videos captured at 30 Frames per Second (FPS), which seems a reasonable

Figure 4.16: Anomaly detected in scene 1_3_6-4-1 of CUHK dataset using SVM model. First row shown the reference frames - 33, 50 and 79 -, the second row shown results with $T = 1$, the third row $T = 1$ plus our post-processing method, the fourth row $T = 14$ and the last row $T = 100$.



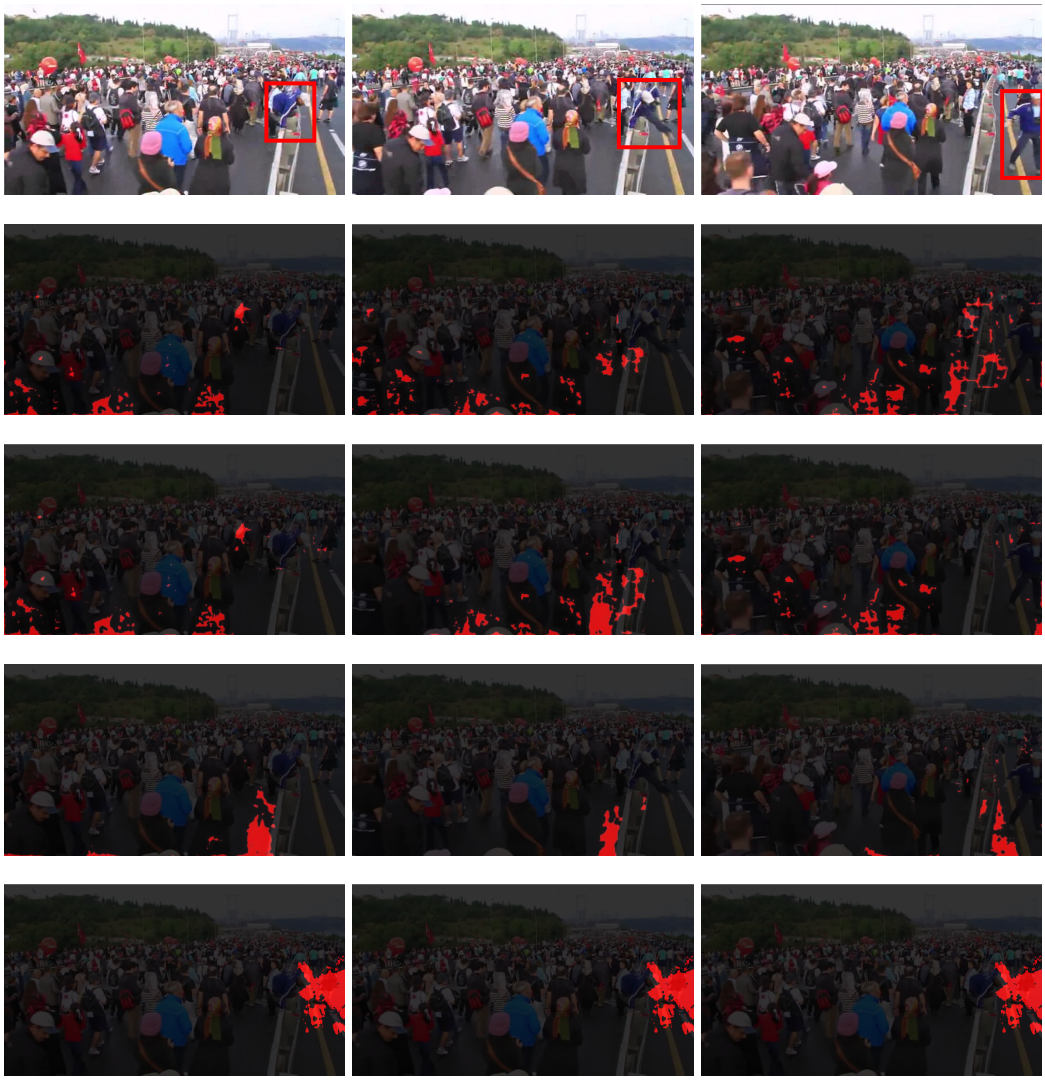
value.

4.4.3 Detecting Anomalies in Low-Density Crowds

The core of both proposed optical flow post-processing technique and the feature extractor is to explore local flow coherence in spatial neighborhoods, which assumes that there is local flow information available. This is the case of dense crowds, and neighborhood information gets scarcer as the crowd density decreases.

In lower density crowds, our assumptions will not always be true, because in these

Figure 4.17: Anomalies detected in 1_34_008681798-people-walk-europe-3 scene of CUHK dataset using RNN. First row shown the reference frames - 50, 60 and 81 -, the second row shown results with $T = 1$, the third row $T = 1$ plus our post-processing method, the fourth row $T = 14$ and the last row $T = 100$.



scenes people have more freedom to move, not needing to share their personal space with strangers. However, most of existing methods for abnormality detection explore individual pedestrian behavior or lower density crowds, either due to lack of publicly available datasets with annotated abnormal behavior in denser crowds, or because the challenges in low-density scenarios are simpler. For example, the methods presented in (MAHADEVAN et al., 2010; CONG; YUAN; LIU, 2011; FENG; YUAN; LU, 2017) explore the UCSD dataset (CHAN; VASCONCELOS, 2008), which contains mostly sparse video clips.

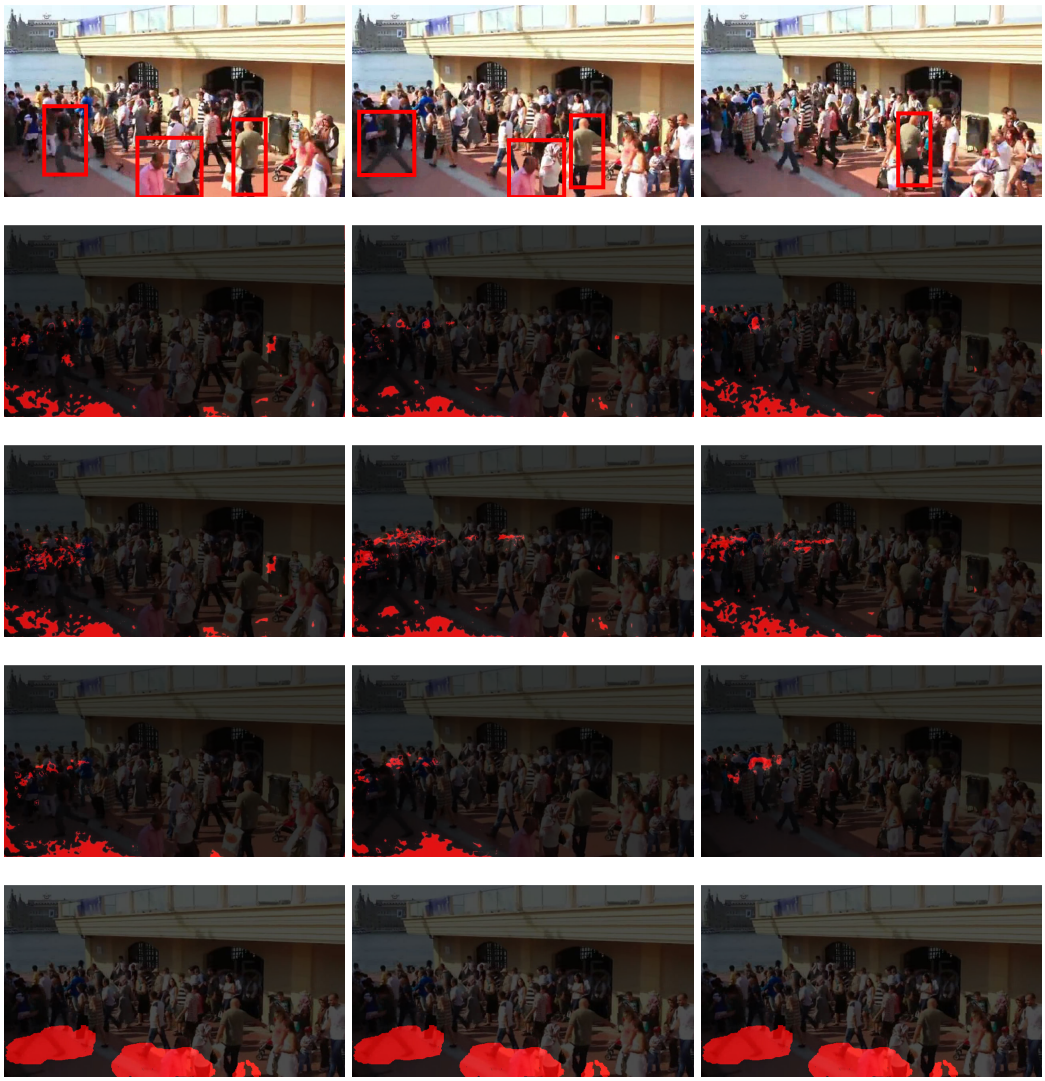
When applying our method to lower-density crowds, we should note that the stationarity hypothesis must be weakened, meaning that we must select a smaller value for T.

Figure 4.18: Anomaly detection results of scene 1_34_008681798-people-walk-europe-3 CUHK dataset using SVM model. First row shown the reference frames - 50, 60 and 81 -, the second row shown results with $T = 1$, the third row $T = 1$ plus our post-processing method, the fourth row $T = 14$ and the last row $T = 100$.

Based on the comparative study presented in the previous subsection, we select $T = 14$ as the stationarity temporal window, and choose the SVM trained previously as the classifier (since it presented better results than RNN for shorter time windows). Note that our goal in these experiments is not propose the best anomaly detector in low dense crowds, but to show that our method is competitive and might perform well in a wide variety of settings.

Our experiments with sparser crowds use the clips Test019, Test021 and Test014 of UCSDped1. All scenes present a sparse crowd walking on a (apparently) pedestrian pathway, and people move in both directions. In Test19, the anomaly is caused by a vehicle that crosses the crowd; in Test021, it is marked as the wheelchair in the crowd; and Test014 presents four anomalies: three cyclists and a truck that cross the crowd.

Figure 4.19: Anomalies detected using RNN model in scene 1_34_009622329-passengers-kadikoy-port-2 of CUHK dataset. First row shown the reference frames - 17, 30 and 80 -, the second row shown results with $T = 1$, the third row $T = 1$ plus our post-processing method, the fourth row $T = 14$ and the last row $T = 100$.



We presented in figures the seventh frame of each clip, and beside it the result to this clip, the presented clips do not have overlap, resulting in 196 frames analyzed in each video. In Figure 4.23 we present a scene with only one anomaly behaviour in scene: a vehicle crossing through the crowd. The figure shown in first and third columns the anomaly ground truth provided by dataset. The output of the SVM did not detect the vehicle as anomaly initially, but detected when it approached the crowd. We can note that when the vehicle is in scene, the SVM output false positives around it because the crowd has just a few people and the optical flow region referent the vehicle is large, and when we compare the people motion the vehicle motion occupies a large part of the neighborhood analyzed.

Figure 4.20: Anomalies detected results using SVM model in scene 1_34_009622329-passengers-kadikoy-port-2 of CUHK dataset. First row shown the reference frames - 17, 30 and 80 -, the second row shown results with $T = 1$, the third row $T = 1$ plus our post-processing method, the fourth row $T = 14$ and the last row $T = 100$.

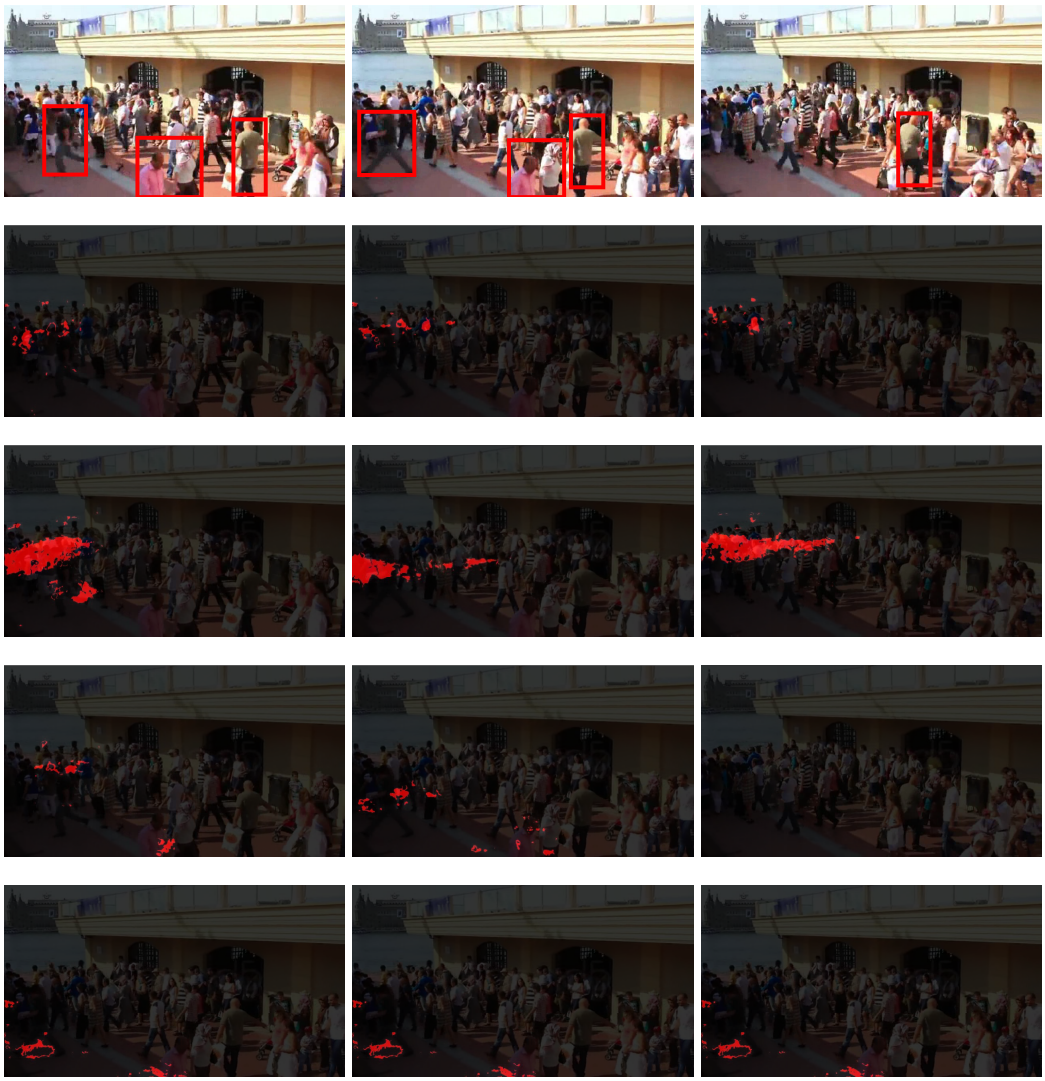


Figure 4.24 presents a scene with only one annotated anomaly in the ground truth data: the wheelchair traveling the scene. In these scene our SVM did not detect any anomaly, and this happens because the motion of the wheelchair does not differ enough from the crowd motion. Note that the correct abnormality detection of the vehicle in Figure 4.23 by our method was not due to the vehicle itself, but by the fact that the motion of the vehicle generated an anomalous motion pattern according to our features.

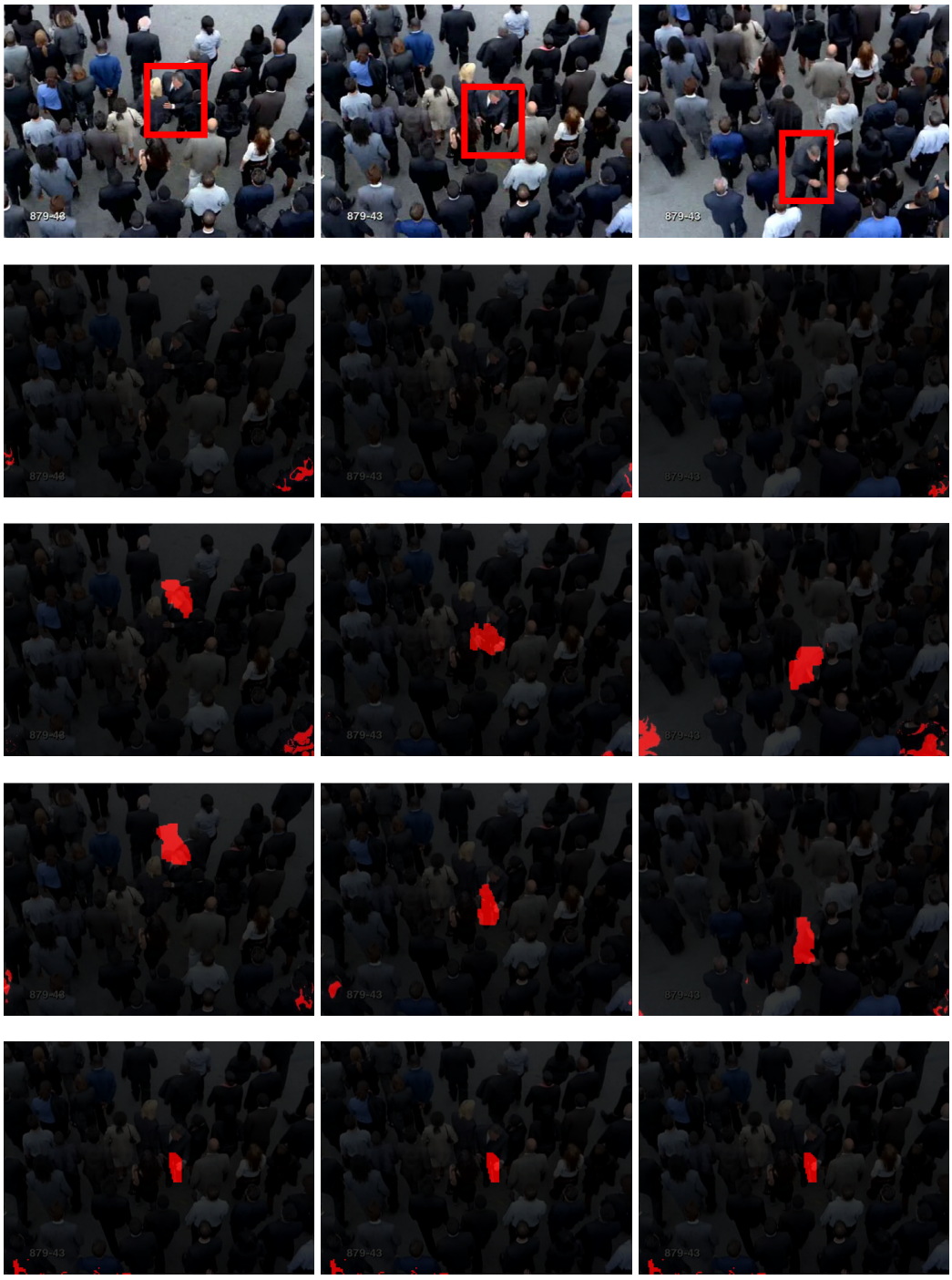
We also analyze scenes with multiple anomalies. For instance, Figure 4.25 shows a scene with four anomalies: two cyclists in begin of the video in opposite directions, a cyclist that appears in bottom border during the scene and a truck crossing the crowd. In this video, in only one clip our classifier lost all anomalies; in the others, we detect the

Figure 4.21: Anomaly detection results of scene 1_879-43_l-2 of CUHK dataset using RNN. First row shown the reference frames - 50, 62 and 98 -, the second row shown results with $T = 1$, the third row $T = 1$ plus our post-processing method, the fourth row $T = 14$ and the last row $T = 100$.

anomalies, even though we generate some false positives due to the same issues presented in Figure 4.23: few people in scene, so the anomalous regions have a lot of influence in the neighbourhood.

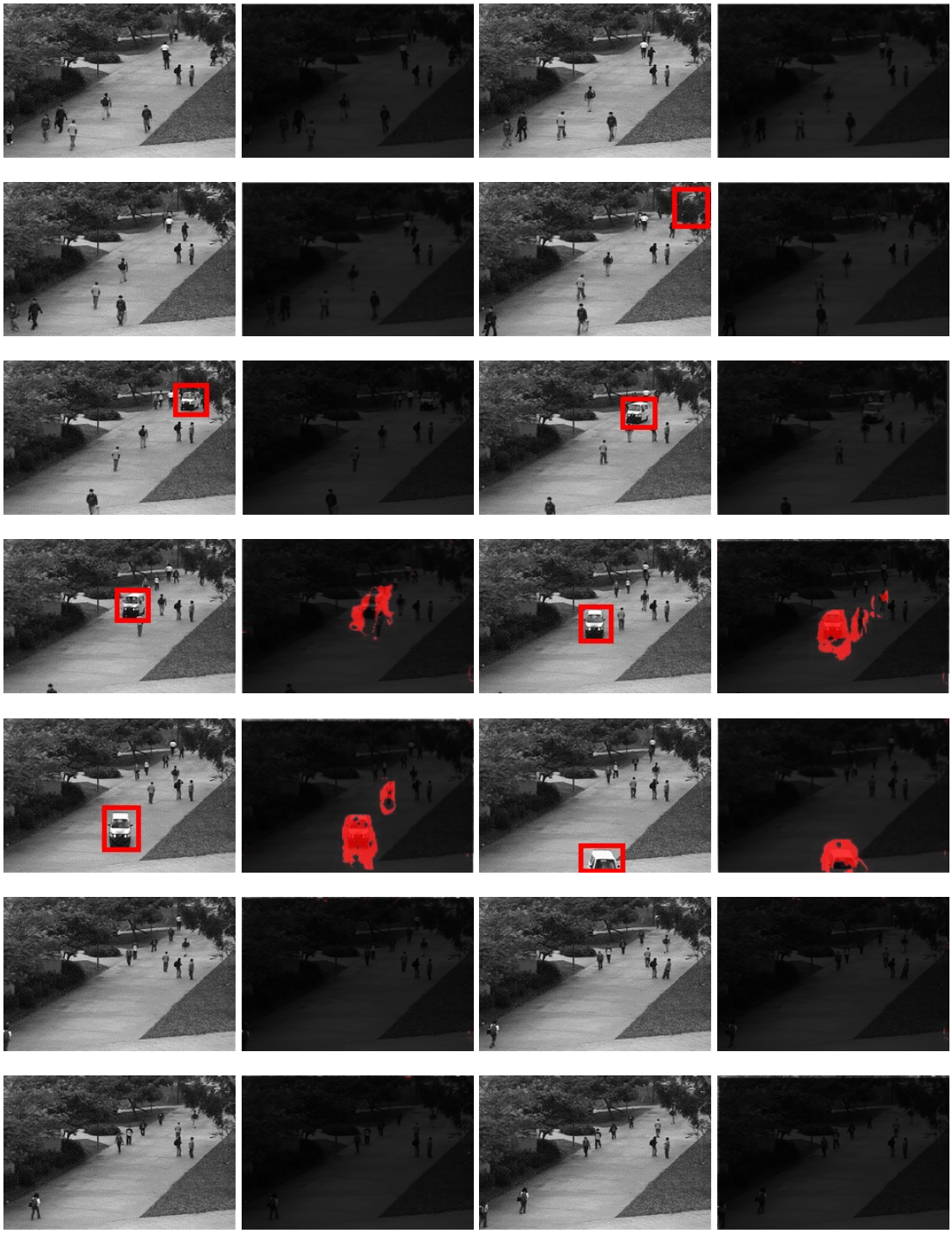
In order to better evaluate our SVM-based classifier, we compare our results with

Figure 4.22: Anomalies detected in scene 1_879-43_l-2 of CUHK dataset using SVM model. First row shown the reference frames - 50, 62 and 98 -, the second row shown results with $T = 1$, the third row $T = 1$ plus our post-processing method, the fourth row $T = 14$ and the last row $T = 100$.



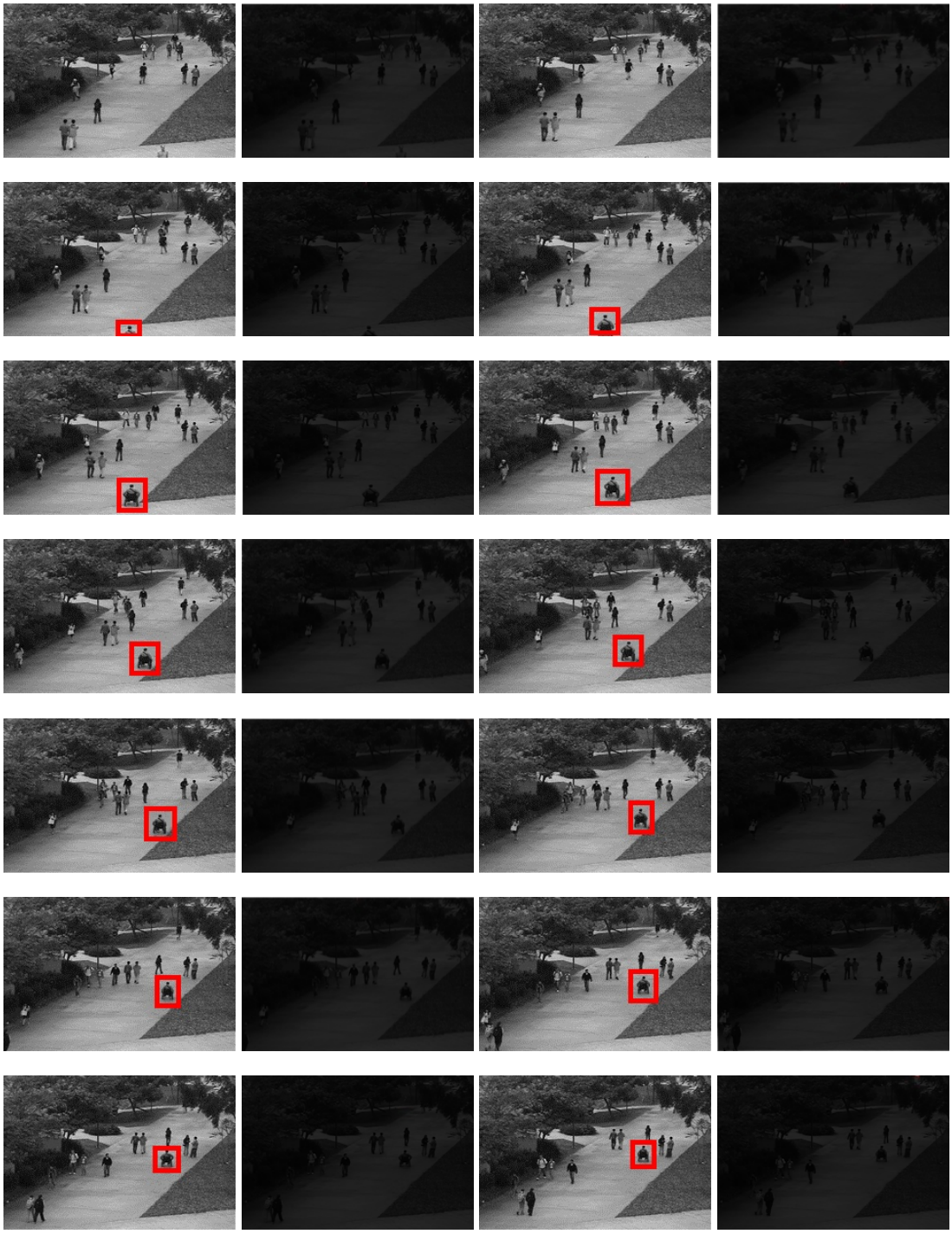
state-of-art algorithms, namely MDT (MAHADEVAN et al., 2010), SF-MPPCA (MAHADEVAN et al., 2010), SRC (CONG; YUAN; LIU, 2011), and PCANet-GMM (FENG; YUAN; LU, 2017). We used the results of these methods as provided by Feng, Yuan and Lu (2017), and the comparison is shown in Figure 4.26. In this comparison, we use clip

Figure 4.23: Columns 1 and 3 are UCSDped1_Test019 clip's reference frame marked anomaly in red squares and columns 2 and 4 are anomalies detected using SVM model.



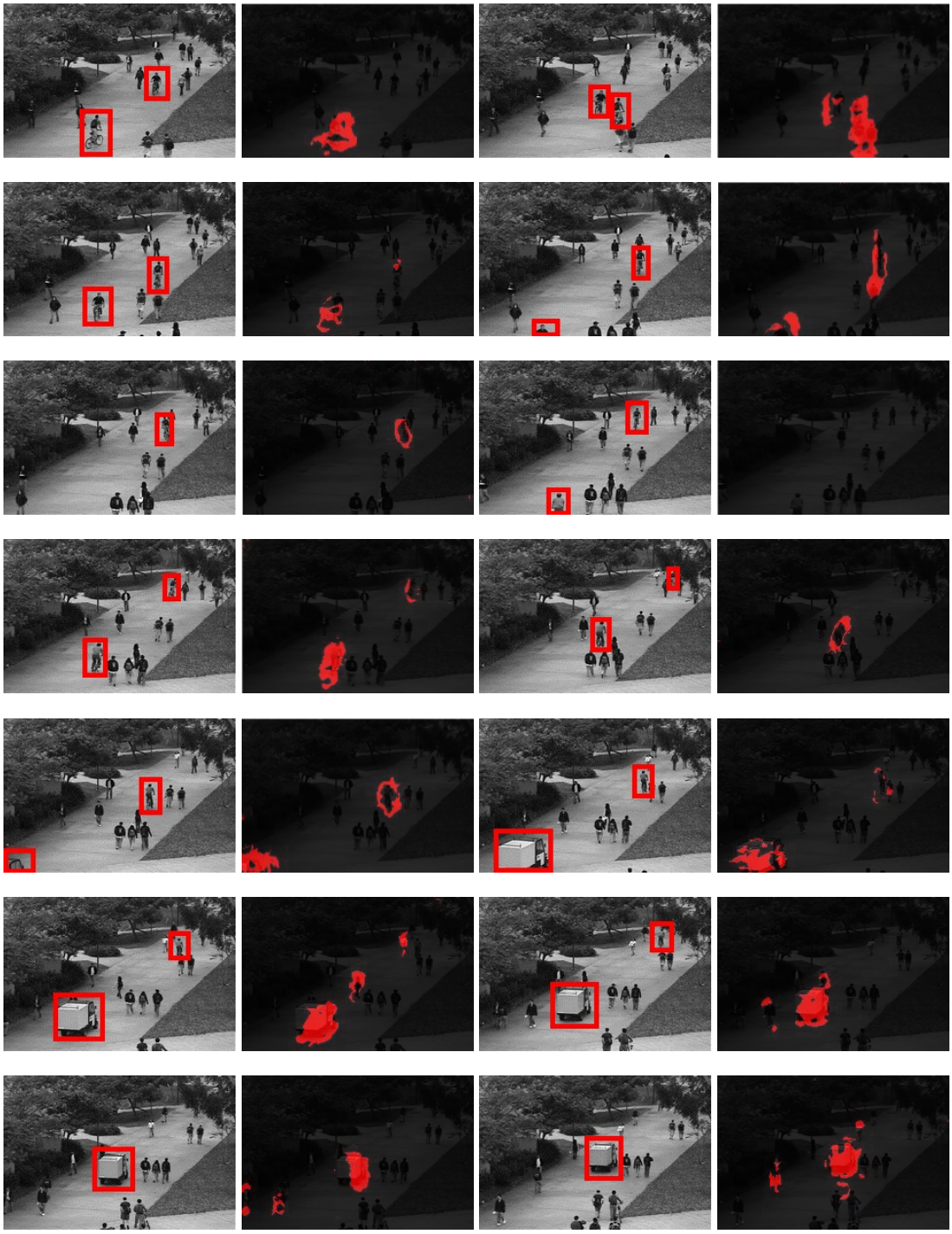
sizes of 14 frames, different from the comparative methods that make the analysis frame by frame. In the first column of Figure 4.26, a vehicle crossing a crowd is marked as anomaly by the ground-truth. Our SVM detected the anomaly, but some false positives appear because the density of the crowd versus the dimensions of the anomalous object. Note that our result is better than SF-MPPCA, but worse than the other methods. The sec-

Figure 4.24: Columns 1 and 3 are UCSDped1_Test021 clip's reference frame marked anomaly in red squares and columns 2 and 4 are anomalies detected using SVM model.



ond column shows a skateboarder crossing the crowd, that was detected by our method, even if not in a fully connected blob. The result is again better than SF-MPPCA and as good as MDT, but SRC and PCANet-GMM algorithms produced more adjusted results the skater silhouette. Again, this happens because we use an approach that does not analyze the video in a frame-by-frame manner. The third column shows a cyclist in the lower

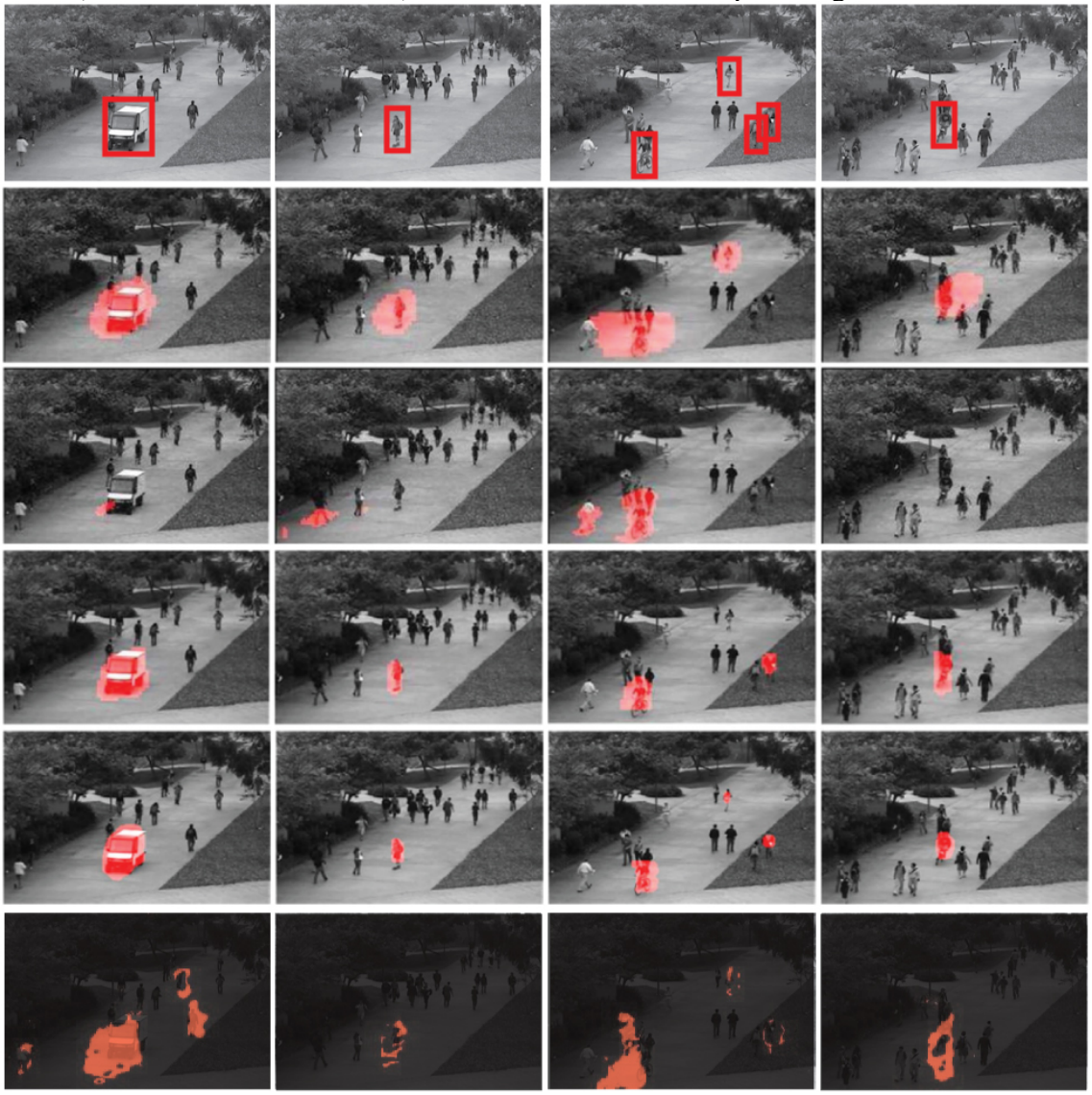
Figure 4.25: Columns 1 and 3 are UCSDped1_Test014 clip's reference frame marked anomaly in red squares and columns 2 and 4 are anomalies detected using SVM model.



part of the frame, two people crossing the lawn and a runner close the top. None of the methods succeeded to detect the left person crossing the lawn, our SVM detect anomalies in the right person crossing the lawn, in the runner and in the cyclist, in this last detecting a big region anomalous because detected not just where the cyclist is in the displayed frame, but also where he is in the following frames. In the last column, another cyclist

crosses the crowd. Again, our SVM detected the anomaly, even if the blob is not adjusted to the displayed frame skater silhouette.

Figure 4.26: Results of anomaly detection methods on UCSD dataset, where first row is the ground-truth provided by Feng, Yuan and Lu (2017), the second row is output of the MDT algorithm (MAHADEVAN et al., 2010), the third row shown result of the SF-MPPCA algorithm (MAHADEVAN et al., 2010), fourth row is the output of the SRC algorithm (CONG; YUAN; LIU, 2011), the fifth row the results of PCANet-GMM algorithm (FENG; YUAN; LU, 2017) and the last row is the output using our SVM trained.



As a summary of the experiments with lower-density crowds, we can infer that our SVM-based classifier did not show the best result in anomaly detection for sparse crowds. However, it presented competitive results with the advantage of being able to handle high-density crowds, setup for which it was designed. Another important consideration is that the proposed method detected anomalies in low dense scenes even if our SVM was trained only with dense scenes.

5 CONCLUSIONS AND FUTURE WORK

5.1 Final Remarks

In this dissertation, we presented approaches that explore neighborhood flow consistency in the context of crowd behavior analysis. In particular, we tackled two main problems in this work: i) Estimation of crowd flows based on expected psycho-social motion aspects of real crowds; ii) Extraction of crowd features based on local flow coherence and its use for abnormal behavior in crowded scenes.

For crowd flow estimation, we initially obtain the optical flow (given by any technique) and restrict the analysis to a foreground mask, obtained by background removal. We then explore the consistency of each flow vector to its neighborhood based in interpersonal distances, which is divided into four invisible bubbles of space consisting of the territory that each person likes to keep between themselves and other people or things. This consistency is computed using the Mahalanobis distance between the pixel motion vector and neighborhood mean motion, which is quickly estimated using five integral images based on second-order statistics of the x and y components of the vector field $\mathbf{v}_w(\mathbf{u})$. Our approach was tested in conjunction with four well-known baseline optical flow approaches, and validation was performed both quantitatively (visual inspection of the generated crowd flows) and quantitatively. For the quantitative analysis, we explored the the smoothness of particle trajectories obtained by advection, and also the accuracy of an event detection approach that takes as input the crowd flow. Our results showed that the proposed filtering method improves the crowd flow generated by all tested baseline approaches. They also indicated that even using simpler (and fast) baseline methods coupled with the proposed method can lead to fast and accurate results, which can be useful for real-time crowd analysis.

The anomaly detection method uses the pixel consistency in multiscale neighborhoods as input to different classifiers, aiming to detect which pixels belong to anomaly regions. This technique used a feature vector of dissimilarity – calculated using the Mahalanobis distance – for each pixel in the foreground, where each feature represents the dissimilarity of the pixel motion in WCS with the mean motion in WCS within a given neighborhood. Lastly, we trained classifiers to determine if the pixel is normal or abnormal. This approach was tested with five classifiers: Recurrent Neural Network, Multilayer Perceptron, Support Vector Machine, Random Forest, and Extremely Randomized Trees,

showing that the approach is not very dependent on the classifier choice. Also, we explore the proposed method in low-density crowds using a Support Vector Machine, and compared our results with four anomaly detection methods. Our experimental results indicated that the proposed feature vector yields to competitive results also for scenes with a few people (i.e., sparse crowds). The results also indicate that the proposed approach detects anomalies in different scenarios, obtaining better results in highly dense scenes for which it was designed.

5.2 Future Work

According to Table 3.1, the lowest execution mean time using our crowd flow estimation is approximately 0.2 seconds. In order to speed-ups it we can implement the proposed filtering method in GPUs to achieve even lower execution times. In fact, several steps are independent and performed in a pixel-wise manner, so that the use of concurrent implementations have potential for high speed-ups. Another possible direction for future work is to combine image-based features in the filtering process.

In this dissertation, we explored the proposed crowd feature vector in the context of abnormality detection. However, a straightforward extension would be to tackle the detection of specific events in crowds, such as panic, dispersion, and bottlenecks behaviors. The main drawback of this task at the moment is the lack of annotated datasets with dense crowds, which might be available in the near future. Another approach alleviates the lack of data is to use synthetic crowds provided by crowd simulators. With more data available (real and/or synthetic), it would also be possible to use deeper neural network architectures, which could be applied to detect and classify anomalies in a unique pipeline.

REFERENCES

ABDALLAH, A. C. B.; GOUIFFÈS, M.; LACASSAGNE, L. A modular system for global and local abnormal event detection and categorization in videos. **Machine Vision and Applications**, v. 27, n. 4, p. 463–481, May 2016. ISSN 1432-1769. Available from Internet: <https://doi.org/10.1007/s00138-016-0752-z>.

ALI, S.; SHAH, M. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: IEEE. **2007 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.], 2007. p. 1–6.

ALI, S.; SHAH, M. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2007. p. 1–6.

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**. Berlin, Heidelberg: Springer-Verlag, 2008. (ECCV '08), p. 1–14. ISBN 978-3-540-88685-3. Available from Internet: <http://dx.doi.org/10.1007/978-3-540-88688-4_1>.

ALMEIDA, I. R. de et al. Detection of global and local motion changes in human crowds. **IEEE Transactions on Circuits and Systems for Video Technology**, IEEE, v. 27, n. 3, p. 603–612, 2017.

ALMEIDA, I. R. de; JUNG, C. R. Change detection in human crowds. In: **Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI - Conference on**. [S.l.: s.n.], 2013. p. 63–69. ISSN 1530-1834.

ALSABTI, K.; RANKA, S.; SINGH, V. An efficient k-means clustering algorithm. 1997.

ANDRADE, E.; BLUNSDEN, S.; FISHER, R. Modelling crowd scenes for event detection. In: **Pattern Recognition, 2006. ICPR 2006. 18th International Conference on**. [S.l.: s.n.], 2006. v. 1, p. 175–178.

BAKER, S. et al. A database and evaluation methodology for optical flow. **International Journal of Computer Vision**, Springer, v. 92, n. 1, p. 1–31, 2011.

BEAUCHEMIN, S. S.; BARRON, J. L. The computation of optical flow. **ACM computing surveys (CSUR)**, ACM, v. 27, n. 3, p. 433–466, 1995.

BERA, A.; KIM, S.; MANOCHA, D. Realtime anomaly detection using trajectory-level crowd behavior learning. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops**. [S.l.: s.n.], 2016. p. 50–57.

BERTINI, M.; BIMBO, A. D.; SEIDENARI, L. Scene and crowd behaviour analysis with local space-time descriptors. In: **Communications Control and Signal Processing (ISCCSP), 2012 5th International Symposium on**. [S.l.: s.n.], 2012. p. 1–6.

BOUGUET, J.-Y. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. **Intel Corporation**, v. 5, n. 1-10, p. 4, 2001.

BRADSKI, G. Opencv: Examples of use and new applications in stereo, recognition and tracking. In: **Proc. of International Conference on Vision Interface**. [S.l.: s.n.], 2002.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BROX, T.; BREGLER, C.; MALIK, J. Large displacement optical flow. In: IEEE. **Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on**. [S.l.], 2009. p. 41–48.

BROX, T.; MALIK, J. Large displacement optical flow: descriptor matching in variational motion estimation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 33, n. 3, p. 500–513, 2011.

CAPUTO, M.; FABRIZIO, M. A new definition of fractional derivative without singular kernel. **Progr. Fract. Differ. Appl**, v. 1, n. 2, p. 1–13, 2015.

CHAN, A.; VASCONCELOS, N. Modeling, clustering, and segmenting video with mixtures of dynamic textures. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, v. 30, n. 5, p. 909–926, 2008.

CHEN, D.-Y.; HUANG, P.-C. Motion-based unusual event detection in human crowds. **Journal of Visual Communication and Image Representation**, v. 22, n. 2, p. 178–186, 2011.

CHEN, M.; WANG, Q.; LI, X. Anchor-based group detection in crowd scenes. In: IEEE. **Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on**. [S.l.], 2017. p. 1378–1382.

CHENG, K.-W.; CHEN, Y.-T.; FANG, W.-H. Abnormal crowd behavior detection and localization using maximum sub-sequence search. In: **Proceedings of the 4th ACM/IEEE International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Stream**. [S.l.]: ACM, 2013. (ARTEMIS '13), p. 49–58.

CHERIYADAT, A. M.; RADKE, R. J. Detecting dominant motions in dense crowds. **IEEE Journal of Selected Topics in Signal Processing**, v. 2, n. 4, p. 568–581, Aug 2008. ISSN 1932-4553.

CHONG, X. et al. Hierarchical crowd analysis and anomaly detection. **Journal of Visual Languages & Computing**, Elsevier, v. 25, n. 4, p. 376–393, 2014.

COHEN, A.; DAUBECHIES, I.; FEAUVEAU, J.-C. Biorthogonal bases of compactly supported wavelets. **Communications on Pure and Applied Mathematics**, Wiley Subscription Services, Inc., A Wiley Company, v. 45, n. 5, p. 485–560, 1992.

CONG, Y.; YUAN, J.; LIU, J. Sparse reconstruction cost for abnormal event detection. In: **Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on**. [S.l.: s.n.], 2011. p. 3449–3456.

CUI, X. et al. Abnormal detection using interaction energy potentials. In: **Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on**. [S.l.: s.n.], 2011. p. 3161–3167.

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: **Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on**. [S.l.: s.n.], 2005. v. 1, p. 886–893 vol. 1.

DEE, H. M.; CAPLIER, A. Crowd behaviour analysis using histograms of motion direction. In: **Image Processing (ICIP), 2010 17th IEEE International Conference on**. [S.l.: s.n.], 2010. p. 1545–1548.

ESHEL, R.; MOSES, Y. Tracking in a dense crowd using multiple cameras. **International Journal of Computer Vision**, Springer US, v. 88, n. 1, p. 129–143, 2010.

FARNEBÄCK, G. Two-frame motion estimation based on polynomial expansion. **Image analysis**, Springer, p. 363–370, 2003.

FELZENSZWALB, P. et al. Object detection with discriminatively trained part-based models. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, v. 32, n. 9, p. 1627–1645, 2010.

FENG, Y.; YUAN, Y.; LU, X. Learning deep event models for crowd anomaly detection. **Neurocomputing**, Elsevier, v. 219, p. 548–556, 2017.

FERRYMAN, J.; ELLIS, A. Pets2010: Dataset and challenge. **Advanced Video and Signal Based Surveillance, IEEE Conference on**, IEEE Computer Society, Los Alamitos, CA, USA, v. 0, p. 143–150, 2010.

FERRYMAN, J.; SHAHROKNI, A. Pets2009: Dataset and challenge. In: IEEE. **Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on**. [S.l.], 2009. p. 1–6.

FORTUN, D.; BOUTHEMY, P.; KERVRANN, C. Optical flow modeling and computation: a survey. **Computer Vision and Image Understanding**, Elsevier, v. 134, p. 1–21, 2015.

FRADI, H.; DUGELAY, J.-L. Towards crowd density-aware video surveillance applications. **Information Fusion**, n. 0, p. –, 2014. ISSN 1566-2535. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1566253514001055>.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine learning**, Springer, v. 63, n. 1, p. 3–42, 2006.

GONG, M. et al. Local distinguishability aggrandizing network for human anomaly detection. **Neural Networks**, Elsevier, v. 122, p. 364–373, 2020.

GREEN, M. W. **Appropriate and effective use of security technologies in u.s. schools**. [S.l.]: Tech. Rep. 97-IJ-R-072, National Institute of Justice, 1999.

GREENEWALD, K.; HERO, A. Detection of anomalous crowd behavior using spatio-temporal multiresolution model and kronecker sum decompositions. 2014.

GUO, G. et al. Knn model-based approach in classification. In: MEERSMAN, R.; TARI, Z.; SCHMIDT, D. C. (Ed.). **On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. p. 986–996. ISBN 978-3-540-39964-3.

HALL, E. T. The hidden dimension. Doubleday & Co, 1966.

HAQUE, M.; MURSHED, M.; PAUL, M. On stable dynamic background generation technique using gaussian mixture models for robust object detection. In: **Advanced Video and Signal Based Surveillance, 2008. AVSS '08. IEEE Fifth International Conference on**. [S.l.: s.n.], 2008. p. 41–48.

HAQUE, M.; MURSHED, M. M. Panic-driven event detection from surveillance video stream without track and motion features. In: **IEEE International Conference on Multimedia and Expo**. [S.l.: s.n.], 2010. p. 173–178.

HEARST, M. et al. Support vector machines. **Intelligent Systems and their Applications, IEEE**, v. 13, n. 4, p. 18–28, 1998.

HEARST, M. A. et al. Support vector machines. **IEEE Intelligent Systems and their applications**, IEEE, v. 13, n. 4, p. 18–28, 1998.

HELBING, D.; FARKAS, I.; VICSEK, T. Simulating dynamical features of escape panic. **Nature**, Nature Publishing Group, v. 407, n. 6803, p. 487–490, 2000.

HELBING, D.; MOLNÁR, P. Social force model for pedestrian dynamics. **Phys. Rev. E**, American Physical Society, v. 51, p. 4282–4286, May 1995.

HELBING, D.; MOLNÁR, P. Self-organization phenomena in pedestrian crowds. In: CITESEER. **Self-organization of Complex Structures: From Individual to Collective Dynamics**. [S.l.], 1997.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

HORN, B. K.; SCHUNCK, B. G. Determining optical flow. **Artificial intelligence**, Elsevier, v. 17, n. 1-3, p. 185–203, 1981.

IDREES, H. et al. Multi-source multi-scale counting in extremely dense crowd images. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2013. p. 2547–2554.

IDREES, H. et al. Composition loss for counting, density map estimation and localization in dense crowds. In: **Proceedings of the European Conference on Computer Vision (ECCV)**. [S.l.: s.n.], 2018. p. 532–546.

ILG, E. et al. Flownet 2.0: Evolution of optical flow estimation with deep networks. In: **IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2017. v. 2.

JIANG, F.; WU, Y.; KATSAGGELOS, A. K. Detecting contextual anomalies of crowd motion in surveillance video. In: IEEE. **Image Processing (ICIP), 2009 16th IEEE International Conference on**. [S.l.], 2009. p. 1117–1120.

JUNG, C.; HENNEMANN, L.; MUSSE, S. R. Event detection using trajectory clustering and 4-d histograms. **Circuits and Systems for Video Technology, IEEE Transactions on**, v. 18, n. 11, p. 1565–1575, 2008.

JUNG, C. R. Efficient background subtraction and shadow removal for monochromatic video sequences. **IEEE Transactions on Multimedia**, v. 30, n. 8, June 2009.

KAJO, I.; KAMEL, N.; MALIK, A. S. An adaptive block-based matching algorithm for crowd motion sequences. **Multimedia Tools and Applications**, Jan 2017. ISSN 1573-7721.

KAJO, I.; MALIK, A. S.; KAMEL, N. An evaluation of optical flow algorithms for crowd analytics in surveillance system. In: **2016 6th International Conference on Intelligent and Advanced Systems (ICIAS)**. [S.l.: s.n.], 2016. p. 1–6.

KAUP, D. et al. Crowd dynamics simulation research. **Simulation Series**, Society for Computer Simulation; 1999, v. 38, n. 4, p. 365, 2006.

KRATZ, L.; NISHINO, K. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: IEEE. **Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on**. [S.l.], 2009. p. 1446–1453.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2012. p. 1097–1105.

LAVEE, G.; RIVLIN, E.; RUDZSKY, M. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. **Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on**, v. 39, n. 5, p. 489–504, 2009.

LAVÍN-DELGADO, J. et al. Robust optical flow estimation involving exponential fractional-order derivatives. **Optik**, Elsevier, v. 202, p. 163642, 2020.

LEE, D.-G.; SUK, H.-I.; LEE, S.-W. **Crowd Behavior Representation Using Motion Influence Matrix for Anomaly Detection**. 2013. 110-114 p.

Li, M. et al. Crowd behavior simulation based on psychological emotion, sociological role and personality. In: **2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)**. [S.l.: s.n.], 2019. p. 371–377.

LI, W.; MAHADEVAN, V.; VASCONCELOS, N. Anomaly detection and localization in crowded scenes. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE Computer Society, Los Alamitos, CA, USA, v. 99, n. PrePrints, p. 1, 2013.

LIM, M. K. et al. Crowd saliency detection via global similarity structure. In: **2014 22nd International Conference on Pattern Recognition**. [S.l.: s.n.], 2014. p. 3957–3962. ISSN 1051-4651.

LIU, C. et al. Human-assisted motion annotation. In: **2008 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2008. p. 1–8. ISSN 1063-6919.

LIU, P. et al. Selflow: Self-supervised learning of optical flow. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2019. p. 4571–4580.

LUCAS, B. D.; KANADE, T. An iterative image registration technique with an application to stereo vision. In: **Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1981. (IJCAI'81), p. 674–679.

MAHADEVAN, V. et al. Anomaly detection in crowded scenes. In: **Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on**. [S.l.: s.n.], 2010. p. 1975–1981.

MARČETIĆ, D.; RIBARIĆ, S. A fuzzy logic-based approach to detection of abnormal crowd behaviour. In: IEEE. **2019 International Symposium ELMAR**. [S.l.], 2019. p. 143–146.

MARSDEN, M. et al. Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. In: IEEE. **Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on**. [S.l.], 2017. p. 1–7.

MCCLELLAND, J. L. et al. Parallel distributed processing. **Explorations in the Microstructure of Cognition**, MIT Press Cambridge, Ma, v. 2, p. 216–271, 1986.

MEHRAN, R.; OYAMA, A.; SHAH, M. Abnormal crowd behavior detection using social force model. In: **2009 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2009. p. 935–942.

MOUSSAÏD, M.; HELBING, D.; THERAULAZ, G. How simple rules determine pedestrian behavior and crowd disasters. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 108, n. 17, p. 6884–6888, 2011.

MOWER, J. P. Prep-mt: predictive rna editor for plant mitochondrial genes. **BMC bioinformatics**, Springer, v. 6, n. 1, p. 96, 2005.

MUSSE, S. R. et al. Using computer vision to simulate the motion of virtual agents: Research articles. **Comput. Animat. Virtual Worlds**, John Wiley and Sons Ltd., Chichester, UK, v. 18, n. 2, p. 83–93, may 2007. ISSN 1546-4261.

PATHAN, S.; AL-HAMADI, A.; MICHAELIS, B. Incorporating social entropy for crowd behavior detection using svm. In: BEBIS, G. et al. (Ed.). **Advances in Visual Computing**. [S.l.]: Springer Berlin Heidelberg, 2010, (Lecture Notes in Computer Science, v. 6453). p. 153–162.

RAGHAVENDRA, R. et al. Optimizing interaction force for global anomaly detection in crowded scenes. In: IEEE. **Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on**. [S.l.], 2011. p. 136–143.

RANFTL, R.; BREDIES, K.; POCK, T. Non-local total generalized variation for optical flow estimation. In: SPRINGER. **European Conference on Computer Vision**. [S.l.], 2014. p. 439–454.

RANJAN, A.; BLACK, M. J. Optical flow estimation using a spatial pyramid network. In: **IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2017. v. 2.

RAVANBAKHSH, M. et al. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. In: IEEE. **2018 IEEE Winter Conference on Applications of Computer Vision (WACV)**. [S.l.], 2018. p. 1689–1698.

RAVANBAKHSH, M. et al. Training adversarial discriminators for cross-channel abnormal event detection in crowds. In: IEEE. **2019 IEEE Winter Conference on Applications of Computer Vision (WACV)**. [S.l.], 2019. p. 1896–1904.

REN, Z. et al. A fusion approach for multi-frame optical flow estimation. In: IEEE. **2019 IEEE Winter Conference on Applications of Computer Vision (WACV)**. [S.l.], 2019. p. 2077–2086.

REVAUD, J. et al. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In: **2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2015. p. 1164–1172. ISSN 1063-6919.

RODRIGUEZ, J. J.; KUNCHEVA, L. I.; ALONSO, C. J. Rotation forest: A new classifier ensemble method. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 28, n. 10, p. 1619–1630, 2006.

RODRIGUEZ, M. et al. Density-aware person detection and tracking in crowds. In: **Computer Vision (ICCV), 2011 IEEE International Conference on**. [S.l.: s.n.], 2011. p. 2423–2430.

ROSHTKHARI, M. J.; LEVINE, M. D. Online dominant and anomalous behavior detection in videos. In: **Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on**. [S.l.: s.n.], 2013.

SHAO, J.; LOY, C. C.; WANG, X. Scene-independent group profiling in crowd. In: **Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition**. Washington, DC, USA: IEEE Computer Society, 2014. (CVPR '14), p. 2227–2234. ISBN 978-1-4799-5118-5. Available from Internet: <http://dx.doi.org/10.1109/CVPR.2014.285>.

SHI, J. et al. Good features to track. In: IEEE. **Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on**. [S.l.], 1994. p. 593–600.

SINGH, K. et al. Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets. **Neurocomputing**, Elsevier, v. 371, p. 188–198, 2020.

SOLMAZ, B.; MOORE, B. E.; SHAH, M. Identifying behaviors in crowd scenes using stability analysis for dynamical systems. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 34, n. 10, p. 2064 –2070, oct. 2012.

SU, H. et al. The large-scale crowd behavior perception based on spatio-temporal viscous fluid field. **Information Forensics and Security, IEEE Transactions on**, v. 8, n. 10, p. 1575–1589, 2013.

SUN, D.; ROTH, S.; BLACK, M. J. Secrets of optical flow estimation and their principles. In: IEEE. **Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on**. [S.l.], 2010. p. 2432–2439.

SUN, D.; ROTH, S.; BLACK, M. J. A quantitative analysis of current practices in optical flow estimation and the principles behind them. **International Journal of Computer Vision**, Springer, v. 106, n. 2, p. 115–137, 2014.

TU, Z. et al. A survey of variational and cnn-based optical flow techniques. **Signal Processing: Image Communication**, Elsevier, v. 72, p. 9–24, 2019.

TUZEL, O.; PORIKLI, F.; MEER, P. Region covariance: A fast descriptor for detection and classification. In: . [S.l.]: Springer, 2006. p. 589–600.

ULLAH, H.; CONCI, N. Crowd motion segmentation and anomaly detection via multi-label optimization. In: **ICPR Workshop on pattern recognition and Crowd Analysis**. [S.l.: s.n.], 2012.

ULLAH, H.; ULLAH, M.; CONCI, N. Real-time anomaly detection in dense crowded scenes. **Proc. SPIE 9026, Video Surveillance and Transportation Imaging Application**, v. 9026, p. 902608–902608–7, 2014.

VIOLA, P.; JONES, M. Robust real-time object detection. In: **International Journal of Computer Vision**. [S.l.: s.n.], 2001.

WANG, B. et al. Abnormal crowd behavior detection using high-frequency and spatio-temporal features. **Machine Vision and Applications**, Springer-Verlag, v. 23, n. 3, p. 501–511, 2012.

WANG, J.; XU, Z. Spatio-temporal texture modelling for real-time crowd anomaly detection. **Computer Vision and Image Understanding**, Elsevier, v. 144, p. 177–187, 2016.

WANG, S.; MIAO, Z. Anomaly detection in crowd scene. In: IEEE. **Signal Processing (ICSP), 2010 IEEE 10th International Conference on**. [S.l.], 2010. p. 1220–1223.

WANG, S.; MIAO, Z. Anomaly detection in crowd scene using historical information. In: IEEE. **Intelligent Signal Processing and Communication Systems (ISPACS), 2010 International Symposium on**. [S.l.], 2010. p. 1–4.

WANG, T. et al. Abnormal event detection based on analysis of movement information of video sequence. **Optik - International Journal for Light and Electron Optics**, v. 152, p. 50 – 60, 2018. ISSN 0030-4026. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0030402617308872>.

WANG, T. et al. Abnormal event detection via the analysis of multi-frame optical flow information. **Frontiers of Computer Science**, Springer, v. 14, n. 2, p. 304–313, 2020.

WEINZAEPFEL, P. et al. Deepflow: Large displacement optical flow with deep matching. In: **2013 IEEE International Conference on Computer Vision**. [S.l.: s.n.], 2013. p. 1385–1392. ISSN 1550-5499.

WU, S.; MOORE, B. E.; SHAH, M. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In: **The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010**. [S.l.: s.n.], 2010. p. 2054–2060.

WU, S. et al. Crowd behavior analysis via curl and divergence of motion trajectories. **International Journal of Computer Vision**, v. 123, n. 3, p. 499–519, Jul 2017. ISSN 1573-1405. Available from Internet: <https://doi.org/10.1007/s11263-017-1005-y>.

YANG, J.; LI, H. Dense, accurate optical flow estimation with piecewise parametric model. In: **2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2015. p. 1019–1027. ISSN 1063-6919.

YUAN, Y.; FANG, J.; WANG, Q. Online anomaly detection in crowd scenes via structure analysis. **IEEE transactions on cybernetics**, IEEE, v. 45, n. 3, p. 548–561, 2015.

ZHANG, X. et al. Flow field texture representation-based motion segmentation for crowd counting. **Machine Vision and Applications**, v. 26, n. 7, p. 871–883, Nov 2015. ISSN 1432-1769. Available from Internet: <https://doi.org/10.1007/s00138-015-0703-0>.

ZHAO, W.; ZHANG, Z.; HUANG, K. Gestalt laws based tracklets analysis for human crowd understanding. **Pattern Recognition**, 2017. ISSN 0031-3203. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0031320317302431>.

ZHOU, S. et al. Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. **Signal Processing: Image Communication**, Elsevier, v. 47, p. 358–368, 2016.

ZIVKOVIC, Z.; HEIJDEN, F. V. D. Efficient adaptive density estimation per image pixel for the task of background subtraction. **Pattern recognition letters**, Elsevier, v. 27, n. 7, p. 773–780, 2006.