
Técnicas baseadas em similaridade para análise visual de videos de segurança

Gilson Mendes da Silva Junior



UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Uberlândia
2019

Gilson Mendes da Silva Junior

**Técnicas baseadas em similaridade para análise
visual de videos de segurança**

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: José Gustavo de Souza Paiva

Uberlândia
2019

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

S586t Silva Junior, Gilson Mendes da, 1994-
2019 Técnicas baseadas em similaridade para análise visual de videos de
segurança [recurso eletrônico] / Gilson Mendes da Silva Junior. - 2019.

Orientador: José Gustavo de Souza Paiva.
Dissertação (mestrado) - Universidade Federal de Uberlândia,
Programa de Pós-Graduação em Ciência da Computação.
Modo de acesso: Internet.
Disponível em: <http://dx.doi.org/10.14393/ufu.di.2019.67>
Inclui bibliografia.
Inclui ilustrações.

1. Computação. 2. Sistemas de segurança - Monitoramento. 3.
Visualização da informação. 4. Videovigilância. 5. Vigilância eletrônica.
I. Paiva, José Gustavo de Souza, 1979-, (Orient.). II. Universidade
Federal de Uberlândia. Programa de Pós-Graduação em Ciência da
Computação. III. Título.

CDU: 681.3

Similarity-based techniques for Visual Analysis of Surveillance Video

Gilson Mendes da Silva Junior



UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Uberlândia
2019

UNIVERSIDADE FEDERAL DE UBERLÂNDIA – UFU
FACULDADE DE COMPUTAÇÃO – FACOM
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA
COMPUTAÇÃO – PPGCO

The undersigned hereby certify they have read and recommend to the PPGCO for acceptance the dissertation entitled **Similarity-based techniques for Visual Analysis of Surveillance Video** submitted by **Gilson Mendes da Silva Junior** as part of the requirements for obtaining the **Master's degree in Computer Science**.

Uberlândia, 29 de Agosto de 2019

Supervisor: _____

Prof. Dr. José Gustavo de Souza Paiva
Universidade Federal de Uberlândia

Examining Committee Members:

Prof. Dr. Danilo Medeiros Eler
Universidade Estadual Paulista

Prof. Dr. Bruno Augusto Nassif Travençolo
Universidade Federal de Uberlândia

Abstract

Surveillance camera systems based on CCTV (closed-circuit television) are widely employed in a variety of society segments, from private and public security to crowd monitoring and terrorist attack prevention, generating a large volume of surveillance videos. The manual analysis of these videos is unfeasible due to the excessive amount of data to be analyzed, the associated subjectivity, and the presence of noise that can cause distraction and compromise the comprehension of relevant events, impairing an effective analysis. Automatic summarization techniques are usually employed to facilitate this analysis, providing additional information that may guide the security agent in this decision making. However, these strategies provide little/no user interaction, limiting his/her comprehension regarding the involved phenomena. Furthermore, such techniques only address specific scenarios, in the sense that no approach is good for all situations. In this sense, it is important to insert the user in the analysis process, as they provide the additional knowledge to effectively perform the events identification and exploration. Visual analytics techniques represent a potential tool for such analysis, providing video representations that clearly communicate their content, potentially revealing patterns that may represent events of interest. These representations can significantly increase the capacity of the security agent to identify important events, and filtering/exploring those that represent potential alert situations.

In this project we propose a methodology for visual analysis of surveillance videos that employs Information Visualization techniques for events exploration. We specifically coordinate point-placement techniques and Temporal Self-similarity Maps (TSSMs) to create an analysis environment that reveal both structural and temporal aspects related to event occurrence. Users are able to interact with these layouts, in order to change the visualization perspective, focus on specific portions of the video, among other tasks. We present experiments in several surveillance scenarios that demonstrate the ability of the proposed methodology in providing an effective events summarization, the exploration of both the structure of each event and the relationship among them, as well as their temporal properties. The main contribution of this work is a surveillance visual analysis

system which provides a deep exploration of different aspects present on surveillance videos regarding events occurrence, providing an effective analysis and a rapid decision making.

Keywords: Smart surveillance, Information visualization, Similarity-based visualization, Events detection.

Resumo

Sistemas de vigilância baseados em circuito fechado de televisão (CFTV) são amplamente empregados em uma variedade de segmentos da sociedade, desde segurança pública e privada até monitoramento de multidões e prevenção de ataques terroristas, o que gera uma grande quantidade de vídeos de monitoramento. A análise manual desses vídeos é inviável devido a grande quantidade de dados a serem analisados, a subjetividade associada, ou a eventual presença de ruídos que podem causar distrações e consequentemente comprometer a compreensão de eventos relevantes, prejudicando uma análise efetiva. Técnicas automáticas de sumarização são geralmente empregadas para facilitar essa análise, fornecendo informações adicionais que podem orientar o agente de segurança nessa tomada de decisão. No entanto, essas estratégias proporcionam pouca ou nenhuma interação do usuário, limitando sua compreensão em relação aos fenômenos envolvidos. Além disso, tais técnicas abordam apenas cenários específicos, no sentido de que nenhuma abordagem é boa para todas as situações. Nesse sentido, é importante inserir o usuário no processo de análise, pois ele pode fornecer conhecimentos adicionais para efetivamente realizar a identificação e exploração dos eventos. As técnicas de análise visual representam uma ferramenta potencial para essa análise, fornecendo representações do vídeo que comunicam claramente seu conteúdo, potencialmente revelando padrões que podem representar eventos de interesse. Essas representações podem aumentar significativamente a capacidade do agente de segurança de identificar os eventos que ocorrem neles e filtrar/explorar aqueles que representam possíveis situações de alerta.

Neste projeto, propomos uma metodologia para análise visual de vídeos de vigilância que emprega técnicas de visualização de informação para a exploração de eventos. Coordenamos especificamente técnicas de posicionamento de ponto e mapas de auto-similaridade temporal (TSSMs) para criar um ambiente de análise que revele os aspectos estruturais e temporais relacionados à ocorrência de eventos. Os usuários podem interagir com esses *layouts*, a fim de alterar a perspectiva de visualização, focar em partes específicas do vídeo, entre outras tarefas. Apresentamos experimentos em vários cenários de vigilância que demonstram a capacidade da metodologia proposta em fornecer uma sumarização

efetiva de eventos, da exploração da estrutura de cada evento e da relação entre eles, bem como de suas propriedades temporais. A principal contribuição deste trabalho é um sistema de análise visual de vigilância que possibilita uma profunda exploração de diferentes aspectos presentes em vídeos de vigilância sobre a ocorrência de eventos, proporcionando uma abordagem para realizar uma análise efetiva e uma tomada de decisão rápida.

Palavras-chave: Vigilância inteligente, Visualização de informação, Visualização baseada em similaridade, Detecção de eventos..

List of Figures

Figure 1 – General framework of a automatic smart surveillance system (extracted from (KO, 2008)).	24
Figure 2 – Diagram showing the main problems considered in visual surveillance applications, and their dependencies (extracted from (NAZARE et al., 2014a)).	24
Figure 3 – An illustration of video co-summarization, identifying shared events visually similar in all n video. Different colors and shapes indicate relevant discovered events by the algorithm: surf (red circle), sunset (green rectangle) and palm trees (blue hexagon), as showed in selected video frames. Dotted line represent the correspondence among shots. Image extracted from (CHU; SONG; JAIMES, 2015).	26
Figure 4 – Visualization process (CARD; MACKINLAY; SHNEIDERMAN, 1999).	29
Figure 5 – Video volume view proposed by (BAGHERI; ZHENG, 2014). Several sampling lines (in black, blue and purple) are sampled along the main directions of the frame to indicate target movements. Temporal variations of the video are obtained through the video volume. (Figure extracted from (BAGHERI; ZHENG, 2014)).	32
Figure 6 – Volumetric based visualization application. (Figure extracted from (DANIEL; CHEN, 2003))	32
Figure 7 – Video cubism idea. (Extracted from (FELS; MASE, 1999))	33
Figure 8 – Video cubism interaction. (Extracted from (FELS; MASE, 1999)) . . .	33
Figure 9 – Left: Four still frames of a shot from the 1963 film Charade. Top: Schematic storyboard for the same shot, composed using the four frames at left. The subject appears in multiple locations, and the 3D arrow indicates a large motion toward the camera. The arrow was placed and rendered without recovering the 3D location of the subject. Bottom: Storyboard for the same shot, composed by a professional storyboard artist. (Extracted from (GOLDMAN et al., 2006))	34

Figure 10 – Left: Keyframes from a snooker video segment of 3 minute, consisting of 6 shots taken from a complete match of 67 shots. Right: Visualization created from the video data. The visualization displays each shot from the video, whilst introducing event importance shown by ball trajectory emphasis, and event ordering shown by numbered ball icons. Table annotations represent ball pots, player scores, points remaining and the position in the full video where this sequence of events occurs. (Extracted from (PARRY et al., 2011))	34
Figure 11 – An example section of a VPG for the OneShopOneWait1front dataset shows activities in front of a shop over a temporal interval, with a close-up view on the right. In the example, the movement of the three extracted objects are highlighted. (Extracted from (BOTCHEN et al., 2008))	35
Figure 12 – Two visual summaries of a staff meeting video. Both have the same set of keyframes and bounding boxes. Left: a generic visualization. Right: a Stained-Glass visualization. (Extracted from (CHIU; GIRENSOHN; LIU, 2004))	35
Figure 13 – Interactive Schematic Summaries. (13a) Three clusters of trajectories are summarized. Arrows indicate major paths. The below timeline displays cluster information: number of trajectories (left), temporal coverage (middle), and diversity (right). (13b) Path segmentation of a scene. (13c) Visual cluster representation of 13b. (Figures extracted from (HÖFERLIN et al., 2011))	36
Figure 14 – Illustration of the object summarization process. (Figure extracted from (MEGHDAI; IRANI, 2013))	37
Figure 15 – (15a) Four input frame from a video. (15b) Generated video synopses. (Figures extracted from (HUANG et al., 2014))	37
Figure 16 – Visual semantic storyline. (Figure extracted from (SENER et al., 2015))	38
Figure 17 – Adapted from (VIGUIER et al., 2015).	38
Figure 18 – Visual analysis methodology structure, showing all steps organized in structured components. Users interact with the layouts in Visualization component, using a set of interaction tools that provides the exploration of different video aspects with focus on events comprehension.	42
Figure 19 – Diagram of the similarity matrix embedding.	43
Figure 20 – Label function and interactive legend. Each group represents a set of images with similar content.	44

Figure 21 – TSSM layout example. In moments with high disturbance values there are more difference between the frames content, and in moments with low disturbance values the frames are similar to each other, suggesting little or no movement.	45
Figure 22 – Horizontal/Vertical analysis of the frame 0 in a TSSM layout.	46
Figure 23 – Fading effect and isolated points analysis in a TSSM layout.	46
Figure 24 – TSSM diagonal analysis.	47
Figure 25 – Internal regions and homogeneous areas analysis.	47
Figure 26 – Video visualization system interface, showing all coordinated views. Point-placement view (A) allows for exploration of the video structure in terms of event occurrence, from a spatial perspective, focusing on how the frames compose events, as well as how these events relate to each other; TSSM view (B) allows for an analysis of the events temporal aspects, as well as the comprehension of specific behavior related to different types of event; and Timeline (C) allows for a temporal navigation and exploration of the entire video, or selected periods, using conventional video player tools.	48
Figure 27 – Manual instance selection in the projection layout.	49
Figure 28 – Locked selection.	49
Figure 29 – Split resulted from the same selection made in Figure 28b.	50
Figure 30 – Path between sequential instances.	50
Figure 31 – Labeling tool and interactive legend.	50
Figure 32 – Projection exploration employing “spatial and temporal searches”. . . .	51
Figure 33 – Timeline employed in our system. Timeline selector (1) is used to select a range of sequential frames; Speed parameter (2) defines how fast the timeline selector will play; skip frames parameter (3) is the quantity of frames to be skipped in each step of an advance/rewind/play interaction; minus/plus buttons (4) increments/decrements the number of selected frames forward in the timeline selector; advance/rewind/play buttons (5) are traditional video player commands used to interact with the video.	52
Figure 34 – Interaction through the advance option/button.	53
Figure 35 – Timeline updating through selection in the projection layout.	53
Figure 36 – TSSM updating selection in the projection layout.	54
Figure 37 – Cell selection in TSSM reflected in the Point-placement layout.	54
Figure 38 – Set of frames representing key actions from the OFFICE video. (a) Office empty; (b) actor enters; (c) actor picks the book; (d) actor reads the book; (e) actor leaves the office.	58
Figure 39 – Some moments from the SOFA video.	59

Figure 40 – Important actions occurring the VIRAT video.	59
Figure 41 – t-SNE projection of Office video labeled manually.	61
Figure 42 – OFFICE video layouts.	61
Figure 43 – Images representing each group indicated in Fig 42.	63
Figure 44 – Sub actions previously the actor starts to read.	63
Figure 45 – Local analyze of group 4. (45a) TSSM of group 4 instances. In red circle the origin region of the points $X-Y$. (45b) Projection of group 4, highlighting instances correspondent at the higher dissimilar moments of 45a TSSM. (45c) Frames of highlighted instances in Figure 45b. . .	64
Figure 46 – Analysis of two groups representing the actor entering/leaving the office in OFFICE video.	64
Figure 47 – Analysis of two group representing the actor picking/returning the book in OFFICE video.	65
Figure 48 – t-SNE projection of Office video. Moment 1870-1890 highlighted. . . .	65
Figure 49 – SOFA video layouts.	66
Figure 50 – TSSM with the intersection between similar stationary moments highlighted.	68
Figure 51 – VIRAT video layouts. In 51a, colors represent moments before (green), during (blue) and after (yellow) the main event, and numbers represent three distinct moments composing the main event: the actor approaching the car (1), getting into the car (2), and driving the car out of the scene (3).	68
Figure 52 – Main event in VIRAT video.	70

List of Tables

Table 1 – Descriptions of main actions occurring in Office video.	58
Table 2 – t-SNE layouts considering 50%, 25% and 12% frame sampling from OF- FICE, SOFA and VIRAT videos, comparing uniform sampling against our smart sampling procedure (HIGH movement variation and LOW movement variation).	71

Contents

1	INTRODUCTION	17
1.1	Objectives	19
1.2	Thesis Organization	20
2	FUNDAMENTALS AND RELATED WORK	21
2.1	Basic Concepts	21
2.2	Video representation	22
2.3	Smart Surveillance	23
2.4	Smart Surveillance Approaches	25
2.5	Information Visualization	28
3	VISUALIZATION OF SURVEILLANCE VIDEOS	31
3.1	Final Remarks	38
4	PROPOSAL	41
4.1	Introduction	41
4.2	Methodology Architecture	41
4.2.1	Data Processing	42
4.2.2	Visualization	43
4.3	Surveillance Visual Analysis System	48
4.3.1	Point-placement View	48
4.3.2	TSSM	51
4.3.3	Timeline	51
4.3.4	Layout coordinations	54
4.3.5	Smart Sampling	54
4.4	Implementation Details	55
5	EXPERIMENTAL RESULTS	57
5.1	Videos	57

5.1.1	Office	57
5.1.2	SOFA	58
5.1.3	VIRAT	59
5.1.4	Experimental Process	60
5.2	Results	60
5.2.1	OFFICE Video	60
5.2.2	SOFA Video	65
5.2.3	VIRAT	67
5.2.4	Smart Sampling results	69
5.3	Discussion	71
5.3.1	Limitations	72
6	CONCLUSION	75
	BIBLIOGRAPHY	77

I hereby certify that I have obtained all legal permissions from the owner(s) of each third-party copyrighted matter included in my thesis, and that their permissions allow availability such as being deposited in public digital libraries.

Gilson Mendes da Silva Junior Adviser: José Gustavo de Souza Paiva

Introduction

A Surveillance System is composed by a set of electronic equipment, generally closed-circuit television cameras (CCTV), employed to ensure the safety of people or places by monitoring and processing abnormal activities. The use of this systems has grown in several social segments, motivated by the increase on violence rates, specially robberies, fights, terrorist threats, allied to the decrease on the acquisition cost associated to the related equipments (WANG, 2013; GONG; LOY; XIANG, 2011). In 2012, the demand on surveillance systems increased 25%, and from 2008 until 2012, the number of surveillance strategies publications was three times larger than the total number of the same type of publications until 2005 (VISHWAKARMA; AGRAWAL, 2013a; JR; SCHWARTZ, 2016).

These systems produce a massive amount of information to be analyzed by security agents. However, the manual analysis of these videos is unfeasible (GONG; LOY; XIANG, 2011). Furthermore, there is a huge number of long-duration videos, demanding teams with large number of professionals and constant visual attention. Moreover, as the periods of anomalous events occurrence are usually concentrated in a significantly small portion of the video when compared to the periods of normal activity. Thus, the monitoring task easily becomes repetitive and monotonous, which may lead the operator to miss subtle but important events. Some types of events may be unnoticed by humans, as well as multiple events occur simultaneously, which may confound the security agent analysis, even in non real-time scenarios.

Smart surveillance is one way to address these problems. Its goal is to use automatic video analysis technologies in video surveillance applications, employing Machine Learning and Computer Vision (HAMPAPUR et al., 2003) to improve the performance and precision of the process. According to Nazare et al. (2014a), this process employs a set of features extracted from the input video in order to reveal regions of interest for the analysis. Then, tasks such as tracking and object identification are performed, employing several algorithms which allows to perform several automatic tasks, using a variety of computational algorithms, providing a high level knowledge extraction procedure. Finally, these extracted information are presented to the human specialist, supporting

his/her decision-making. Some common applications of smart video surveillance include object detection (YAZDI; BOUWMANS, 2018; SOMMER et al., 2016), traffic monitoring (DATONDJI et al., 2016), anomalous activity detection (HU et al., 2016; MA; DAI; HIROTA, 2017) and motion analysis (NUNES; MOREIRA; TAVARES, 2016).

Despite the demonstrated strengths of these strategies, they are strongly dependent on several factors, such as the description strategy, measures of similarity, training set, and no approach produces good results in all scenarios. Moreover, these strategies are often fully automatic, in the sense that users, except by manipulating a set of parameters, do not interact with the process, not being able to detect important data specificities that justify the decisions taken by the technique, nor explore the structure of the identified events. Taking the user knowledge into account can improve the whole process due his prior domain knowledge. However, integrate the surveillance agent into the video analysis process requires an appropriate data representation, in such way that the information contained in the video is effectively presented, highlighting complex actions, or patterns that may be unnoticed, making the analysis process more attractive and less stressful.

(CARD; MACKINLAY; SHNEIDERMAN, 1999) defines information visualization as “the use of computer-supported, interactive, visual representations of abstract data to amplify cognition” – being cognition the acquisition or use of knowledge. In this context **Visual Analytics** combine information visualization, human factors and data analysis to improve the human-computer communication about the data, as well as the decision-making process. Visual analytics approaches may represent potential tools for surveillance video analysis, creating visual layouts to summarize video content and reveal important features and events in videos (BORGO et al., 2012). These layouts may assist security agents by creating a friendly representation that allows the exploration of video content, to identify important events, and to understand the underlying structure of their occurrences.

In this sense, two promising visualization techniques are those based on point-placement and Temporal Self-Similarity Maps (TSSMs). Visualization techniques based on point-placement are widely employed to map collection instances on individual points in the visualization space. The idea is to preserve in this space the relevant relationships observed in the original space, positioning similar points close, and non-similar points apart. Movements in a surveillance video generally occur in short portions of the scenes and/or follow specific patterns. Therefore, any subtle movement – potentially an anomalous action– causes disturbing in its video frames, which make them distinct from those representing normal situations. In this sense, point-placement techniques may be employed to provide a visual video analysis support, creating a layout in which highly related instances – which represent frames– are positioned in close regions, which may create a natural event summarization. TSSMs are used to segment a video in short actions through its extracted features (KRISHNA et al., 2014a). The idea is based on the fact that computing pair-wise similarity values may provide information regarding abrupt changes among video frames.

To capture these information easier, a TSSM layout may be employed to visually reveal the temporal video structure, mapping the affinity of each frame to the matrix cells, color coding them according to its similarity level. This way, the visual color structure of the matrix may indicates the type of events presents is the video scene.

1.1 Objectives

This work aims to employ similarity-based visual strategies in the analysis of surveillance video with focus on the identification of events occurrence. Andrienko et al. (2013) suggest that the use of two layouts, one mapping spatial information, and another mapping time information is more effective to inspect spatiotemporal data. Following this idea, we combine point-placement layouts with TSSMs, using an adaptation proposed by Cooper e Foote (2001), to perform a visual exploration of spatial and temporal properties associated to the events occurrence. Point-placement layouts create natural video summarization that provides a comprehension about the number of events that occurred in the video, and also gives insights about the type of events, and how they relate to each other in terms of observed patterns. On the other hand, TSSM provide means to quickly see periods of significant changes, or periods with no/stationary actions, as well as to improve the comprehension of each event extent. These two layouts work in an interactive and integrated way, such that the user can guide the analysis through them, and different data aspect are revealed. Therefore, manipulations in the point-placement layout are reflected in TSSM, and vice-versa, highlighting distinct perspectives about the same data. Moreover, the video structure may be fully revealed, by means of natural summaries regarding its both temporal and spatial information.

We believe that a computational system of visual analysis based on similarity techniques may support the detection of anomalous events, as well as the comprehension of the temporal video structure in terms of events occurrence.

To achieve our main objectives some requirements are needed. It is important to employ information visualization techniques that can be adapted to the explored context. Furthermore it is desirable that the selected techniques are employed in a coordinated way, in order to reveal different aspects from the same data. In other words, interactions performed in a specific layout, exploring part of the information will reflect in changes in the other layouts, highlighting the same information through a different perspective –which may also facilitates the user experience. Finally, interactive tools are needed to explore the event structure on the layout and consequently comprehend the associated and patterns. This way, the analyst would be able to interact, suggest, modify, add or remove information considering his/her knowledge about the scenario.

The contributions of this work are listed as follows:

- ❑ A surveillance video visual analysis methodology to perform video exploration/summarization, in terms of event occurrences;
- ❑ A computational system that implements our proposed methodology;
- ❑ The methodology validation through a series of case studies considering various surveillance scenarios, including different movement patterns and types of events.

1.2 Thesis Organization

The next sections follow the structure bellow:

- ❑ Fundamentals: presents basic concepts related to surveillance scenarios, crucial to understand the state of the art approaches, as well as how the videos are generally represented for the suitable computer comprehension of their content. Some smart surveillance approaches are also discussed and concepts related to Information Visualization are introduced;
- ❑ Visualization of surveillance videos: several Information Visualization strategies applied to the video surveillance scenario are discussed, in which we analyze the contributions, weakness and advantages;
- ❑ Proposal: details the methodology architecture, also describing the visual techniques employed by our methodology. Our surveillance visual analysis system is presented, detailing each of its components, interactions and functionalities. Finally, the layout coordinations are presented, illustrating how different video aspects can be explored using our developed system;
- ❑ Experimental Results: presents all the experimental procedure to evaluate our proposal, including the description of the videos and experimental process. It also presents the obtained results, as well as a discussion on them, and the limitations of our methodology;
- ❑ Conclusion: present the conclusion about the works executed in this master degree project, detailing our contributions to the visual analysis of surveillance videos using similarity-based techniques.

Fundamentals

Extracting information of interest from surveillance video to support the decision-making process is often challenging due to several aspects related to the nature of these videos, requiring the analysis of a large amount of data, often performed by a single person. Ideally, only strategic moments should be considered, in order to reduce the analysis time, as well as to improve the events detection. Smart Surveillance techniques are widely employed to aid the analysis process, achieving good results in several scenarios. Inserting the security agent in this process and take his/her previous knowledge into account may improve the achieved results, as well as provide him/her the comprehension of how the video is structured, how the events relate to each other, among other aspects. This human-computer integration is achieved by providing interactive and intuitive layouts, which present strategic video information in a comprehensible manner. In this chapter, we present the basic concepts on video representation, as well as the literature review of smart surveillance techniques and video surveillance visualization that motivated our proposal.

2.1 Basic Concepts

Some basic definitions are presented in this section. They are crucial to understand the state of the art approaches, as well as our approach.

- **Video:** Electronic reproduction technique of image motion that represents activities monitored over time;
- **Video Surveillance:** Video recorded from a surveillance camera positioned in a strategic place to monitor people, objects or environments, to ensure the security of assets, people or strategic objects;
- **Frame:** An image representing a specific time instant of a video;
- **Scene:** Sequence of frames in which some potential strategic action/activity occurs;

- **Keyframe:** A frame whose semantic meaning represents a specific scene.
- **Event:** Action of interest occurred in a single or a set of scenes which may be object of study by a surveillance specialist, such as a security agent. The events are usually related to a specific scenario.

2.2 Video representation

In order to suitably represent video content, features describing specific aspects from the video are usually employed. There are several representation in the literature, generally organized in two representative groups:

- **General domain descriptors:** visual information, extracted from each frame, or from a set of frames, representing global or local aspects from the objects in the video, such as color (AVILA et al., 2011) and shape (LIU et al., 2007; SCHWARTZ; DAVIS, 2009), as well as combinations of different features (SRINIVAS; PAI; PAI, 2016). Color information is a basic feature in visual content representations, efficient to identify image elements (GUO et al., 2002), due to its rotation and scale invariance. Texture is a pixel property related to its neighborhood, that allows the definition of different levels of roughness, granularity, softness, etc. It is useful to describe spatial information, as well as to enhance important aspects of several real world surfaces. It is generally employed associated with color descriptors in surveillance scenarios (ZHANG; XU, 2006; HAHNEL; KLUNDER; KRAISS, 2004; YANG; YU, 2011; TAKALA; PIETIKAINEN, 2007; LU et al., 2014). Shape descriptors are focused on the image edge delimitation. Edges are points of images with sharp variations in brightness levels. Edge detection is useful in defining images strategic analysis regions, such as physical limits, orientation discontinuity or surface structure, and object overlapping, as well as to detect and track objects (BENFOLD; REID, 2011; CAO et al., 2011). These basic features can be extracted from the whole image, or from specific regions (PEDRINI; SCHWARTZ, 2008). A comprehensive review of visual information features can be found in (TSAKANIKAS; DAGIUKLAS, 2018).
- **Specific domain video descriptors:** specific descriptors designed to surveillance tasks, computed considering the relationship between consecutive video frames, aiming at capturing their motion information. These features are often applied to describe moving patterns on a scene, such as object speed, orientation (objects flow), appearance (object texture) and density (amount of moving objects). Commonly used features are based on Optical Flow (MENDI; CLEMENTE; BAYRAK, 2013), magnitude (COLQUE et al., 2017a) or trajectory (SONG et al., 2016). Colque et al. (2017b) propose a space-time feature descriptor based on object orientation

and velocity. Such descriptor is called Histogram of Optical Flow Orientation and Magnitude (HOFM), and captures the orientation and magnitude information from optical flows, in which the latter are related to the objects speed. Caetano, Santos e Schwartz (2016) propose a spatiotemporal feature descriptor called Optical Flow Co-occurrence Matrices (OFCM), which extracts a robust set of measures known as Haralick features. These Haralick features are employed in the proposed OFCM to describe the flow patterns by capturing meaningful properties such as contrast, entropy and homogeneity of co-occurrence matrices. Through such measures, local space-time characteristics of the motion are captured through the neighboring optical flow magnitude and orientation. Junejo et al. (2011) propose a descriptor to perform human action detection through visual changes. The main idea is to describe the similarity structure (and dissimilarity) of an action sequence along time, taking consecutive frames represented by a auto-similarity matrix.

2.3 Smart Surveillance

Smart surveillance systems use automatic video analysis techniques, specially computer vision algorithms in surveillance tasks (HAMPAPUR et al., 2003; NAZARE et al., 2014b). The idea is to efficiently extract strategic information from these videos, recognizing, highlighting, and tracking elements of interest, as well as to analyze and comprehend the activity performed by them (WANG, 2013). A detailed discussion about automatic surveillance systems can be found in (WANG, 2013; VISHWAKARMA; AGRAWAL, 2013a; TSAKANIKAS; DAGIUKLAS, 2018).

Figure 1 shows a general framework of an automatic smart surveillance system, in which a set of sequential tasks are executed on the information captured by a video camera. Behavior and activity analysis, as well as person identification are the final high level process, in which alarm annotation are provided to support decision-making. This architecture may also be employed in other surveillance related tasks. The main tasks considered in visual surveillance applications and their dependences are illustrated in Figure 2. Visual information are captured by the feature extraction which feeds several modules. The results obtained by each module are employed to perform scene analysis and understanding.

Smart surveillance tasks can be categorized in three groups: object tracking, event detection and summarization, described as follows:

- Object tracking: aims to temporally relate the objects of consecutive video frames, in order to identify in a video frame the spatial localization of specific objects detected on previous frames. The idea is to identify and analyze movement patterns associated to the object, and comprehend its actions on the scenes, as well as to measure several related information such as speed, direction, displacement area,

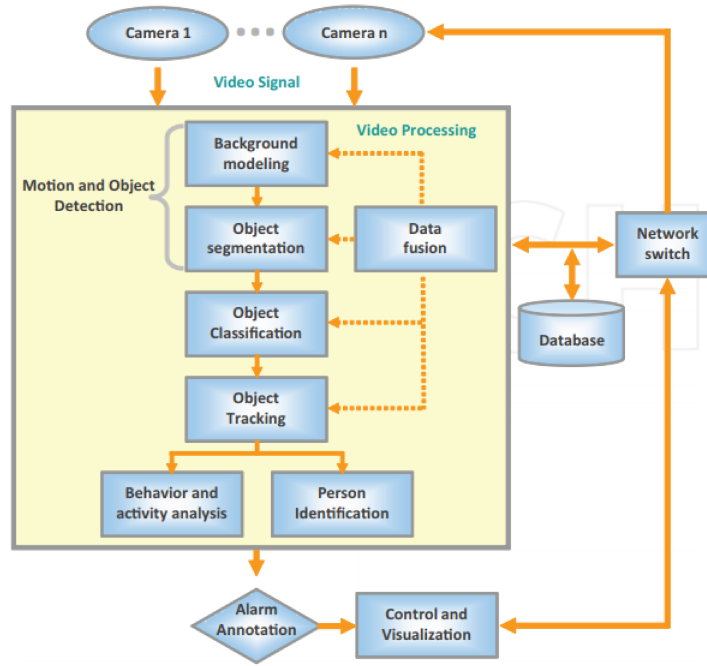


Figure 1 – General framework of a automatic smart surveillance system (extracted from (KO, 2008)).

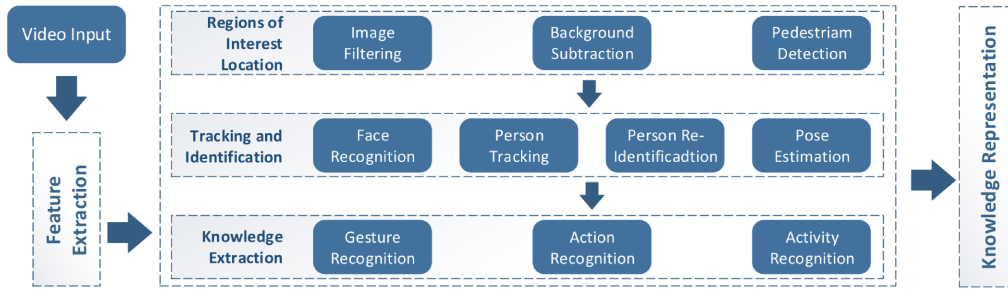


Figure 2 – Diagram showing the main problems considered in visual surveillance applications, and their dependencies (extracted from (NAZARE et al., 2014a)).

etc. Some tasks are related to object tracking, such as abandoned object detection, loitering detection and crowd management. The first one aims to identify changes on scene that may represent strategic information from the moment wherein a suspicious object is left on scene. In some environments, situations or scenarios, an object abandoned for a long time can be classified as potentially dangerous (SILVA, 2016). Loitering detection has the objective of to analyze an element that appears at a specific time and place of the video, which presents a suspicious behavior and may offer risk (BIRD et al., 2005; HUANG; WU; SHIH, 2009). Finally, crowd management uses static collections about the crowd volume and observations of their behaviors (EL-ETRIBY; ELMEZAIN; MIRAOU, 2018), to take decisions to support anomalous situations (KE; SUKTHANKAR; HEBERT, 2007). Some problems can be tackled by measuring the agglomeration level of a video foreground to avoid overcrowded, fight detection, panic attack, a person falling, a car on the wrong side

of the road or going through the red light;

- ❑ Event Detection: aims to identify relevant interest action in a specific video scenario. Event management provides to the application the ability to alert users whenever crucial actions happen on the scene. If an event is detected, then the system can automatically make specific decisions, such as to record the video moment when the action occurs or send notification alerts to the analysts. Such events may be grouped or summarized to enable an easier overview, consequently contributing to specialist analysis (YE, 2018; XU; YANG; HAUPTMANN, 2015; JIANG et al., 2011);
- ❑ Summarization: provide condensed and succinct representations of the video flow content employing static images combinations, video segments, graphic representations and textual descriptors (MONEY; AGIUS, 2008).

2.4 Smart Surveillance Approaches

As videos generally present long duration and the user is interested only in a small portion with strategic activity, it is important to provide effective ways of finding such potential interest moments. In this sense, several automatic techniques related to smart surveillance exist to address such problem. Discussion about several automatically approaches and applications in several scenarios can be found in (HU et al., 2004; KIM et al., 2010; VISHWAKARMA; AGRAWAL, 2013b). We present in this section smart surveillance approaches focused in video summarization, which is the aim of this project.

Clustering techniques are usually employed to separate similar frames into groups, aiming to summarize it, using a single frame or set of frames from each group as a *keyframe* (AVILA et al., 2011; MAHMOUD; ISMAIL; GHANEM, 2013; ZHAO et al., 2015; WU et al., 2017). These approaches usually perform a frame sampling step before the clustering application, and a post-processing step to remove similar keyframes in the produced summaries.

Pritch et al. (2009) propose a video summarization method based on activity clustering, in which the activities are defined as characteristics of moving objects. This approach has the advantage of allowing the use of its summarization results as a training set for a video classification process.

Dogra, Ahmed e Bhaskar (2016) present a video summarization method based on interest events, adopting a finite state machine. This summarization is performed merging features that represent the dynamic of target objects trajectories with a finite state automaton model employed to analyze the feature state change and consequently detect and localize events of interest. They propose two features descriptor, CMA (cumulative moving average) and PSA (preceding segment average), that represent gradual and abrupt changes, respectively, of the moving objects on scene. The results illustrate that this pro-

pose method is invariant to the target(s) considered however, dependent on the target(s) interaction with the environment with reduced sensitivity towards global changes in scene characteristics and at the same time placed limited specification requirements on other local entities, such as the type of activities being performed, etc.

Chu, Song e Jaimes (2015) propose a video co-summarization that identifies visual co-occurrences in a video collection (see Figure 3). The main problem with this type of approach is the cost related to the co-occurrence sparsity. To solve it, the authors present an algorithm called **Maximal Biclique Finding (MBF)**, to support searches for specific contents that appears in several videos, even sparsely. The algorithm discards specific characteristics of single videos (they will be less relevant for the main topic).

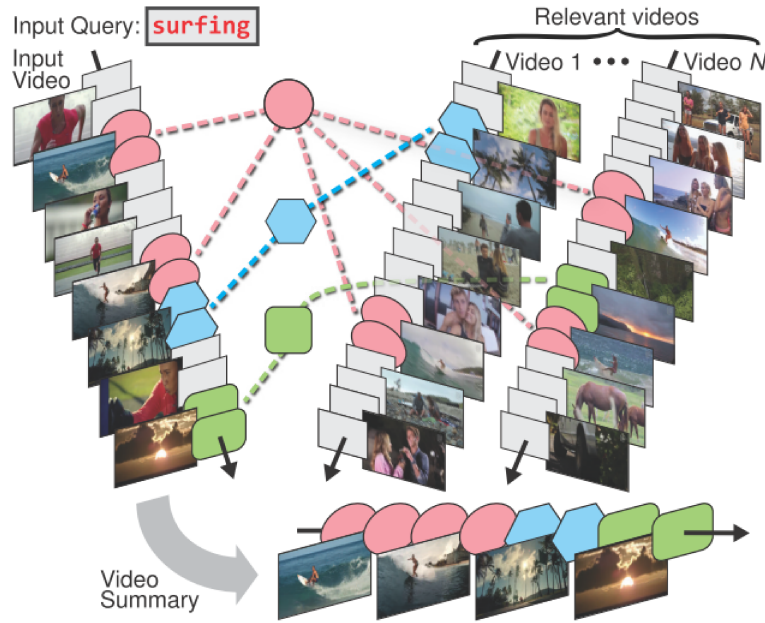


Figure 3 – An illustration of video co-summarization, identifying shared events visually similar in all n video. Different colors and shapes indicate relevant discovered events by the algorithm: surf (red circle), sunset (green rectangle) and palm trees (blue hexagon), as showed in selected video frames. Dotted line represent the correspondence among shots. Image extracted from (CHU; SONG; JAIMES, 2015).

Panda e Roy-Chowdhury (2017) explore the assumption that videos from a specific topic often share similar information (that is, equal or similar frames). This heuristic is used to propose a summarization algorithm that is able to explore visual context present in n video topics – used as a training set – to automatically extract an informative summary from a video test.

Considering that people recording a video generally intend to register objects, Khosla et al. (2013) propose using figures of such objects from the Internet, as an information to summarize videos with similar objects sets. The images are used to estimate and select a maximum amount of informative frames from the input video without employing any manual summary compiled for training. In the first step an image set belonging to

some class (e.g. automobile) is grouped in subclasses, each subclass corresponding to a “canonical point of view”. A classifier is trained for each discovered subclass. To improve the subclasses model, non-labeled video data also are used, assigning each video frame to a subclass. Test video frames are assigned to the learned subclasses, calculating the average decision score of the test examples assigned to each subclass. This score is used to classify the subclasses in the input test video. Finally, a final summary with k most representative frames is generated, selecting the k test video frames, each one is the closer to the top k classified subclass centroids.

Considering that key objects can influence (that is, can be related to) the content of the video frames, Lu e Grauman (2013) present a summarization technique to capture related events and object co-occurrence. The summarization process take into account the cause and effect aspect, excluding unnecessary contents for the right comprehension of events. For example, in a soccer play, some elements are crucial to categorize it as a soccer play, such as the player, the field, soccer fans and other. Otherwise, if a ambulance appear in the field, such video remains illustrating a soccer play, even the ambulance is a non-related object for this context. Thus, scenes of ambulances are not taken into account by the summarization process. The application of the method improved the summary information relevance, due to the better representation of the story shown in the video, that becomes more comprehensible and focused in scene with objects related to the video context.

Several scenarios assume that only visual aspects of the video are not enough to produce a good summarization. Considering that, Zhu, Loy e Gong (2016) propose to aggregate non-visual data, such as meteorological reports and sensory traffic signals, to the traditional visual data to achieve an improved and more precise summarization. This methodology is not limited to surveillance scenario, being generalized to other types of videos. The comparative experiments demonstrated significant gains over existing visual-only models, taking advantage that some level of missing data are not prejudicial.

Zhang et al. (2016) propose a summarization method based on the heuristic that similar videos generally share similar structures. Thus, considering that a video, belonging to a homogeneous set, is manually summarized, its annotation can be used by a smart approach to retrieval relevant frames from other videos belonging to the same set. The idea is to “copy and paste” the relative temporal position of extracted frames in a video sequence and apply it on the unknown videos.

Some works use one-class classification approaches for summarization (KRISHNA et al., 2014b), in which a model is learned using the features of the first f video frames, and subsequent frames are classified as normal or outliers. When a video frame is classified as an outlier, a video segment is defined and another model is learned using the features of the subsequent f frames. This process is repeated with all frames, the video is segmented and the summary is created using selected keyframes. This technique take advantage by

segment videos temporally in a simple and unsupervised way, being easily generalizable to any scenario. Otherwise, feature selection may influence the performance of the approach, in the sense that more sophisticated and suitable ones may improve it, but in the other hand, a bad choice may degrade the results.

Mahasseni, Lam e Todorovic (2017) propose an unsupervised video summarization, in which a deep summarizer network, based on long short-term memory network (LSTM), is learned to minimize the distance between training videos and a distribution of their summarizations. This summarizer is applied on new videos for estimating its optimal summarization. This technique outperform the state of the art in video summarization and provides a comparable accuracy to the state-of-the-art supervised approaches.

2.5 Information Visualization

In automatic techniques, users, except by manipulating a set of parameters, do not interact with the process, not being able to detect important data specificities that justify the decisions taken by the technique, nor explore the structure of the identified events. As Card, Mackinlay e Shneiderman (1999) defines, Information Visualization study the creation of interactive visual representations of abstract data to amplify cognition. In this sense, Information Visualization is employed to improve the computer-human interaction using visual and interactive computational strategies. Usually the data meaning is incomprehensible, but employing Information Visualization correctly to enhance relevant perspectives the user is able to extract important information from this data.

Figure 4 shows the whole Information Visualization process (CARD; MACKINLAY; SHNEIDERMAN, 1999). Firstly, the **RAW DATA** is structured according to a specific analysis perspective, and organized into **DATA TABLES**. Then, the structured data is mapped to **VISUAL STRUCTURES**, that are visual metaphors which support the comprehension of some aspects related to the data. Such metaphors may use shapes, colors, sizes, orientations and other visual characteristics to enhance the data aspects. Then, the VISUAL STRUCTURES is organized into **VIEWS**, which allow users to understand the data meaning and interact through its content. In all stages the data may be interactively handled by the user according to his/her needs, and real world information or a different problem perspective can also be aggregated.

Several types of Information Visualization techniques are provided in literature, and each of them is specialized to handle a different data aspect. A class of techniques widely employed, usually to visualize multidimensional data, are based on **point-placement**, which organizes the instances from a collection on a 2D map using points to represent them. The idea is to represent instances similarity-based relationship by placing the similar ones close and dissimilars apart in the visualization space (PAULOVICH et al., 2008). These techniques use as input an n -dimensional collection, where n represents

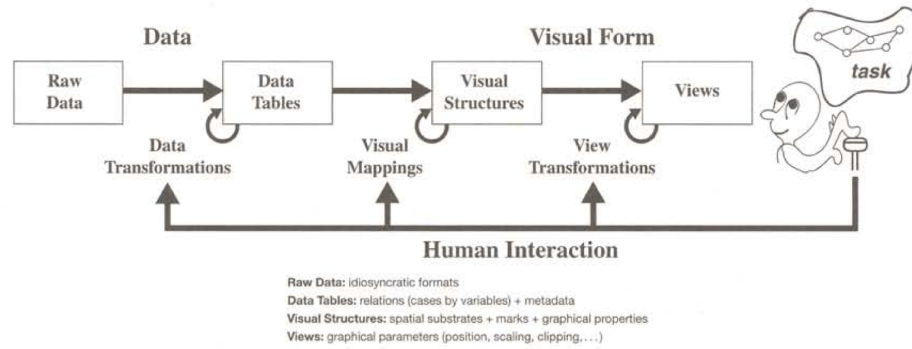


Figure 4 – Visualization process (CARD; MACKINLAY; SHNEIDERMAN, 1999).

the number of considered features, and produce another k -dimensional collection, where $k \ll n$. For visualization purposes, $k = 2$ or 3 .

Several point-placement techniques can be found in the literature. **Multidimensional Scaling (MDS)** (COX; COX, 2000), maps data into lower dimensions, according to the relative positions of a number of objects. In **Force Directed Placement (FDP)** approach (EADES, 1984), the instances are modeled as objects connected by springs, and attraction/repulsion forces among them are proportional to the distance between such instances. The final projection is obtained when a spring system achieves stability. This technique is often precise in representing instances relationship, when applied to collections in collections with non-linear relationship, however it presents a high computation cost ($\mathcal{O}(n^3)$), where n is the number of instances. **Principal Component Analysis (PCA)** (JOLLIFFE, 2002) is a dimensionality reduction technique widely used to visualize high dimensional data. It employs linear combinations among attributes with high level of covariance, producing less dependent attributes called **principal components**. The **Complete Isometric Feature Mapping (ISOMAP)** (TENENBAUM; SILVA; LANGFORD, 2000) is a MDS based technique that work with no euclidean distances. The intrinsic geometry of a data manifold is calculated based on a rough estimate of each data point's neighbors on the manifold. It is useful to visualize data produced by many feature descriptors methods. **Least Square Projection (LSP)** (PAULOVICH et al., 2008) apply least squares on instances to combine the benefits of linear and non-linear projections. This technique aims to create a surface in which the instances are clustered by proximity relationship, allowing the inference of existents relations on data collections. **Local Affine Multidimensional Projection (LAMP)** (JOIA et al., 2011) is based on orthogonal mapping theory, and it aims to create a precise local affine transformations, which may be interactively modified by the user. This technique employs a set of samples associated to a related orthogonal mapping. When the user changes the samples position, the mapping, and the layout is adapted to reflect this modification. Thus, the user can be integrated to the whole process and his knowledge is taken into account when generating the layout.

T-Distributed Stochastic Neighbor Embedding (T-SNE) (MAATEN; HINTON, 2008a) is a dimensionality reduction technique based on Stochastic Neighbor Embedding (HINTON; ROWEIS, 2003), optimized to visualization scenarios. Unlike PCA, it uses local relationships between points to create a low-dimensional mapping that captures **non-linear** structures. The relations in high dimensional space are estimated by a Gaussian probability distribution and in the low-dimensional space (projection space) a Student t -distribution is employed to reduce crowding problem. The embeddings are optimized using gradient descent and its cost function is based on non-convex thought. In the first part of T-SNE algorithm, the probability distribution is calculated in the high dimensional space to estimate the probability of each pair of instance to be positioned close to each other based on a similarity relationship. Similar instances will present higher probabilities to be positioned close in the low dimensional space. A probability distribution is also estimated in the low-dimensional space and a cost function is used to minimize the divergence among this two distribution, considering the position of the points in the layout. This technique has been employed successfully in several video visualization tasks (XU; TAX; HANJALIC, 2012a; RAMANATHAN et al., 2015a; LIAO et al., 2016). We employed T-SNE in the implemented system to validate our proposal, as explained in Section 5.1.4.

Visualization of surveillance videos

Automatic methods applied to video surveillance may assist the security agent in video investigation. However, he/she may not comprehend how these methods perform decisions about the data, as well as how they produce the results, which may impair the analysis. Thus, it is necessary to provide methods that present the video data in a manner from which the agent is able to identify trends and patterns representing strategical information for the analysis. The security agent must be an effective part of this process, being able to understand all schematic information and the video summarization process. The presented results must contribute to the decision making. In this sense, visual strategies may provide this support, improving the analysis process.

Several works employ visual strategies to surveillance video analysis. A system for analysis of security-video surveillance is proposed in (FAN; WANG; HUANG, 2017), presenting an event-oriented visualization mechanism which considers multimodal information. Considering semantic and urgency of captured object characteristics, decisions of camera change are taken (for multi-camera scenarios), based on a score ranking estimated for all camera.

A method to visualize surveillance video content in a temporal 2D image is proposed in (BAGHERI; ZHENG, 2014). A video temporal profile is extracted to convey precise temporal information, keeping some spatial feature to allow the recognition of the different moments represented. The temporal indexing is visualized along the external side of video volume schema, as shown in Figure 5. Although this technique provides a quick comprehension of general video content, the frames distributed along the visualization tend to make the viewing unpleasant and inadequate to indicate multiples events of interest, as well as the sampling lines employed to indicate the target movements in the frames, in the case of many of them be needed.

Daniel e Chen (2003) introduce a volume based visualization technique for summarization of videos. The video frames are placed on a 3D container to capture and exhibit the principal video aspects, as show Figure 6a. This technique allows the temporal comprehension of different video changing levels occurred on scenes (see Figures 6b and 6c),

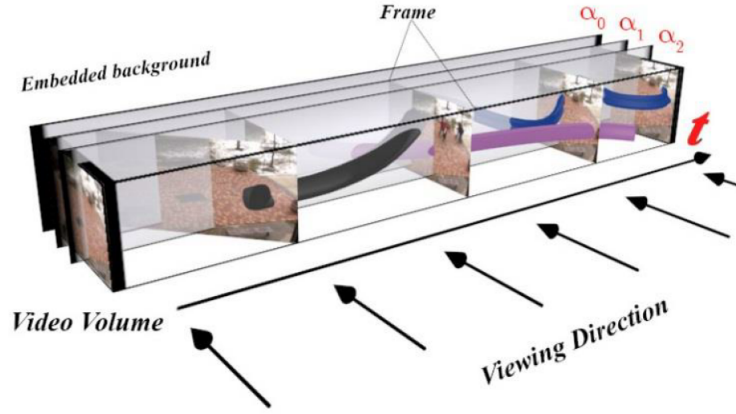
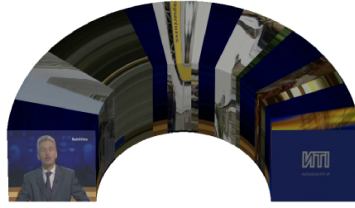
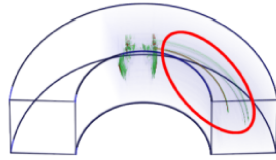


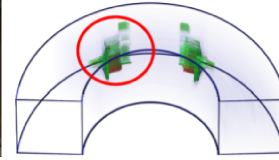
Figure 5 – Video volume view proposed by (BAGHERI; ZHENG, 2014). Several sampling lines (in black, blue and purple) are sampled along the main directions of the frame to indicate target movements. Temporal variations of the video are obtained through the video volume. (Figure extracted from (BAGHERI; ZHENG, 2014)).



(a) A video represented as a volumetric object.



(b) Changes that remain for a period.



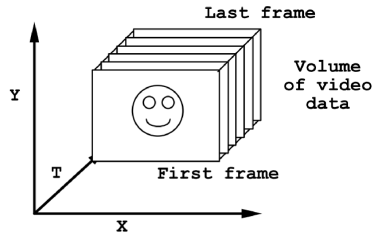
(c) Walking with moving ams

Figure 6 – Volumetric based visualization application. (Figure extracted from (DANIEL; CHEN, 2003))

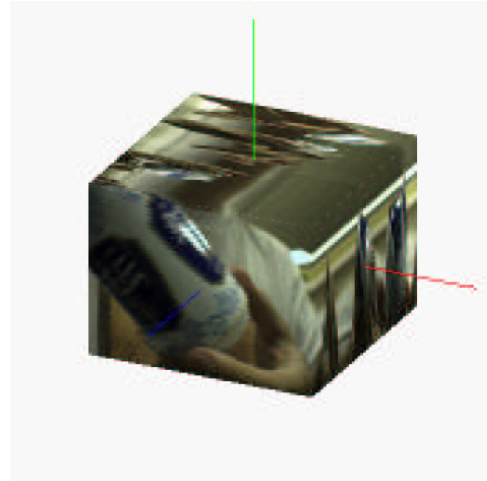
which supports the investigation of relevant information for more detailed analysis. Red zones indicate high-level change detected whilst green zones indicate low-level changes.

Fels e Mase (1999) present an interactive video visualization system represented by the diagram in Figure 7. In this representation, sequences of temporal frames are represented by three-dimensional data blocks, in which the $X - Y$ axes represent the frame content and T represents the sequence of frames. However, identifying relevance in cuts (see Figure 8) that are orthogonal to XY plane is not trivial, because the resulting layout is a simple pixel formation produced by the frames along the time on the cutting position.

Goldman et al. (2006) present a method for visualizing short video clips in a single static image, using the visual language of storyboards, that may also be applied to surveillance scenarios. A storyboard diagram is assembled to summarize a short video sequence,

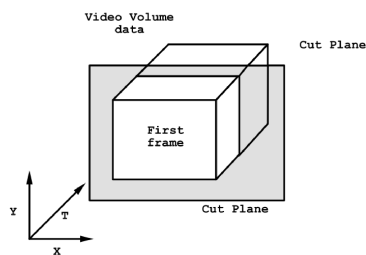


(a) Video frames are stacked in order to form the volume of the video data.

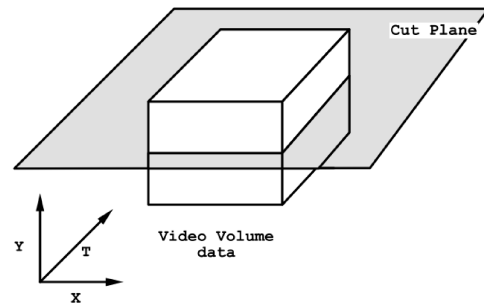


(b) 180 frames of 210x160 video data stacked to form a cube.

Figure 7 – Video cubism idea. (Extracted from (FELS; MASE, 1999))



(a) Cutting the video cube parallel to the X-Y axis shows just a regular video frame located at some moment in time.



(b) Cutting the video cube parallel to the X-T axis shows the video data displayed across all of X for all the frames for a given value of Y.

Figure 8 – Video cubism interaction. (Extracted from (FELS; MASE, 1999))

in which the design schemes represent the observer movements, as well as the observed objects, as presented in Figure 9. Important details are extracted to increase the comprehension of all recorded actions. This technique produces a well qualified summary, getting minucious elements to represent movements. However, for users who want a general video perception, they need to view multiple diagrams, one for each video sequence, which may be unwanted in many situations. This same concept is also employed in (PARRY et al., 2011) to comprehend snooker video shots, in which the movements of the balls on the table are outlined, as shown in Figure 10.

In VideoPerpetuoGram (VPG) technique (BOTCHEN et al., 2008), processed information from a video flow are exhibited in a visualization pipeline using multiple fields, which represents frames from different consecutive video moments. Detected actions along the video flow and estimated relationships between recognized objects are emphasized, as showed in Figure 11. Once a moving object on a scene is detected, its trajectory is en-

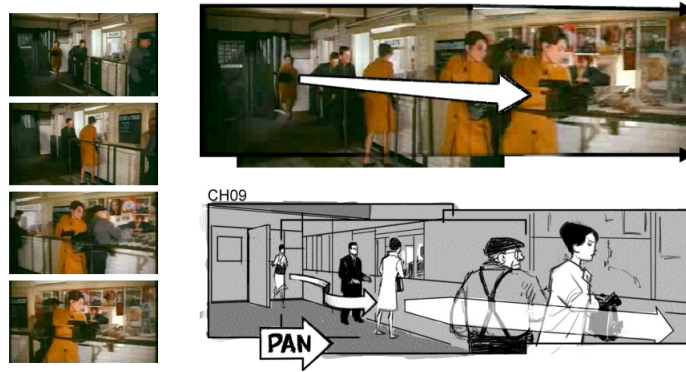


Figure 9 – Left: Four still frames of a shot from the 1963 film *Charade*. Top: Schematic storyboard for the same shot, composed using the four frames at left. The subject appears in multiple locations, and the 3D arrow indicates a large motion toward the camera. The arrow was placed and rendered without recovering the 3D location of the subject. Bottom: Storyboard for the same shot, composed by a professional storyboard artist. (Extracted from (GOLDMAN et al., 2006))

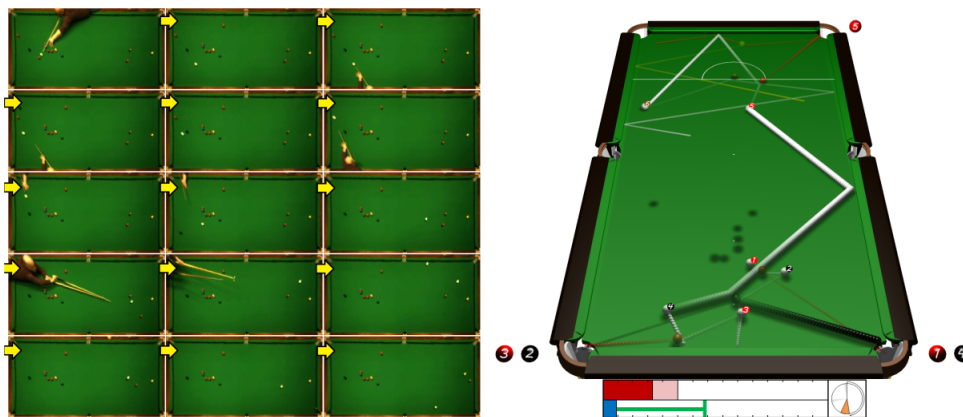


Figure 10 – Left: Keyframes from a snooker video segment of 3 minute, consisting of 6 shots taken from a complete match of 67 shots. Right: Visualization created from the video data. The visualization displays each shot from the video, whilst introducing event importance shown by ball trajectory emphasis, and event ordering shown by numbered ball icons. Table annotations represent ball pots, player scores, points remaining and the position in the full video where this sequence of events occurs. (Extracted from (PARRY et al., 2011))

hanced along multiple fields in the pipeline. Local video information are well represented by means of the segments that indicate the movements, as well as the global ones, as several pipelines can be analyzed simultaneously. However, the way the tracking is represented may overload the visualization content in situations where several trajectories are exhibited.

Chiu, Girgensohn e Liu (2004) propose a method to build high condensed video summaries called **Stained-Glass visualizations**, to be used in mobile devices. Four steps compose the method process. In the first and second stage, respectively, the input video is segmented in small parts according to the similarity of the adjacent frames, and regions

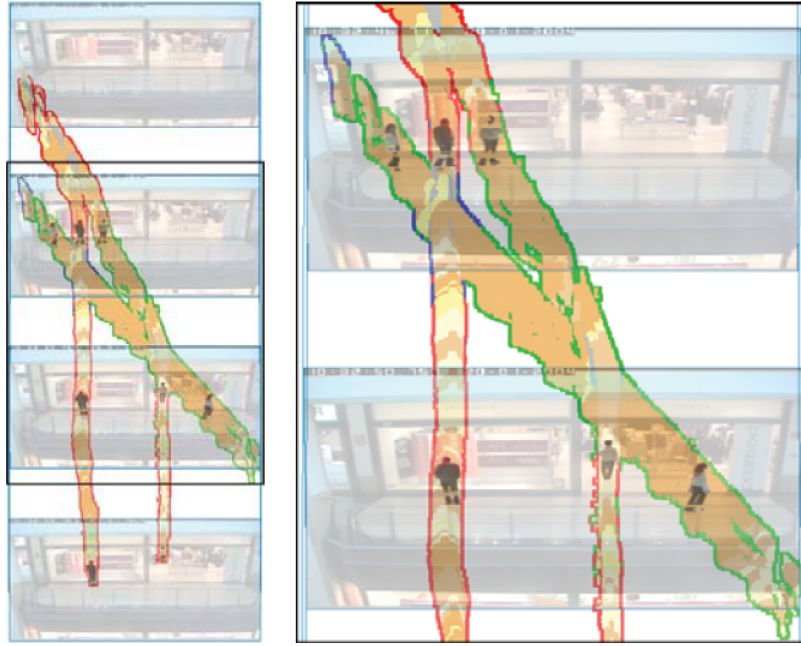


Figure 11 – An example section of a VPG for the OneShopOneWait1front dataset shows activities in front of a shop over a temporal interval, with a close-up view on the right. In the example, the movement of the three extracted objects are highlighted. (Extracted from (BOTCHEN et al., 2008))

of interest from segmented video are detected. Then, the most important regions are outlined and finally the layout space is filled with the main video regions, as illustrated in Figure 12. The main limitation of this approach is the number of considered frames to compose the visualization. In some videos, this number may not be enough to represent all important video aspect or the ideal amount makes the visualization overloaded. Although it is not originally designed for surveillance videos, it may be adapted for such scenario.

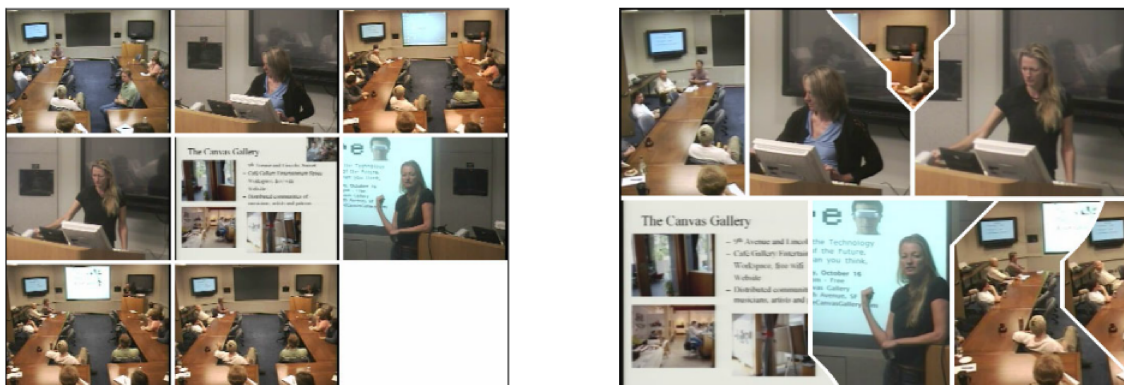


Figure 12 – Two visual summaries of a staff meeting video. Both have the same set of keyframes and bounding boxes. Left: a generic visualization. Right: a Stained-Glass visualization. (Extracted from (CHIU; GIRGENSOHN; LIU, 2004))

Höferlin et al. (2011) introduce a video surveillance exploration technique in which

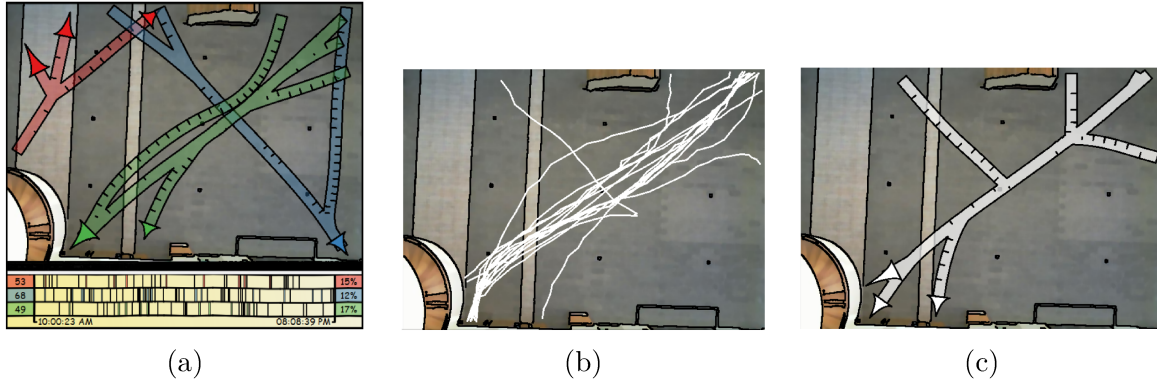


Figure 13 – Interactive Schematic Summaries. **(13a)** Three clusters of trajectories are summarized. Arrows indicate major paths. The below timeline displays cluster information: number of trajectories (left), temporal coverage (middle), and diversity (right). **(13b)** Path segmentation of a scene. **(13c)** Visual cluster representation of 13b. (Figures extracted from (HÖFERLIN et al., 2011))

object trajectories are detected and grouped to create a summary representation that enhances the main paths observed (see Figure 13a). Additional information as the trajectory amount of a summarized path, time periods, and paths diversity are also exhibited. Figures 13b and 13c show a clustering example of trajectory segments of several objects.

Meghdadi e Irani (2013) illustrate the paths that the objects traverse on scene, building a static image that summarize the objects movements, in different moments. Furthermore, object movements are summarized and presented in a cube, in which the first and second dimensions represent the object placement on space and the third one indicates the time (moment) when the object occupied that space placement. The basis of the cube presents a summarized static image illustrating the path traversed by the objects (see Figure 14). Moreover, this technique allows to retrieve video moments of time (in video) and space (estimated traveled) parameters. However, comparing several trajectory patterns may be difficult and unsuitable, due the consequently layout pollution from the many trajectories representations overlapped. Furthermore, to visually distinguish similar people (wearing similar clothes for example) may also be complicated.

Huang et al. (2014) propose a summarization method in which foreground objects are identified and sampled over a single static image that includes all objects of interest (see Figure 15). This static image highlights all object that moved along the scene, providing a global perspective about the video meaning. However, it is not able to represent the movement direction as well as when each object appears, which may hamper the temporal analysis.

Sener et al. (2015) proposed a method to provide a semantic storyline based on objective steps of a set of videos. Using the salience and frequency of the words in the subtitles the method learns language atoms used to represent multi-modal information in each frame. This information is used as input in a semantic clustering model using a

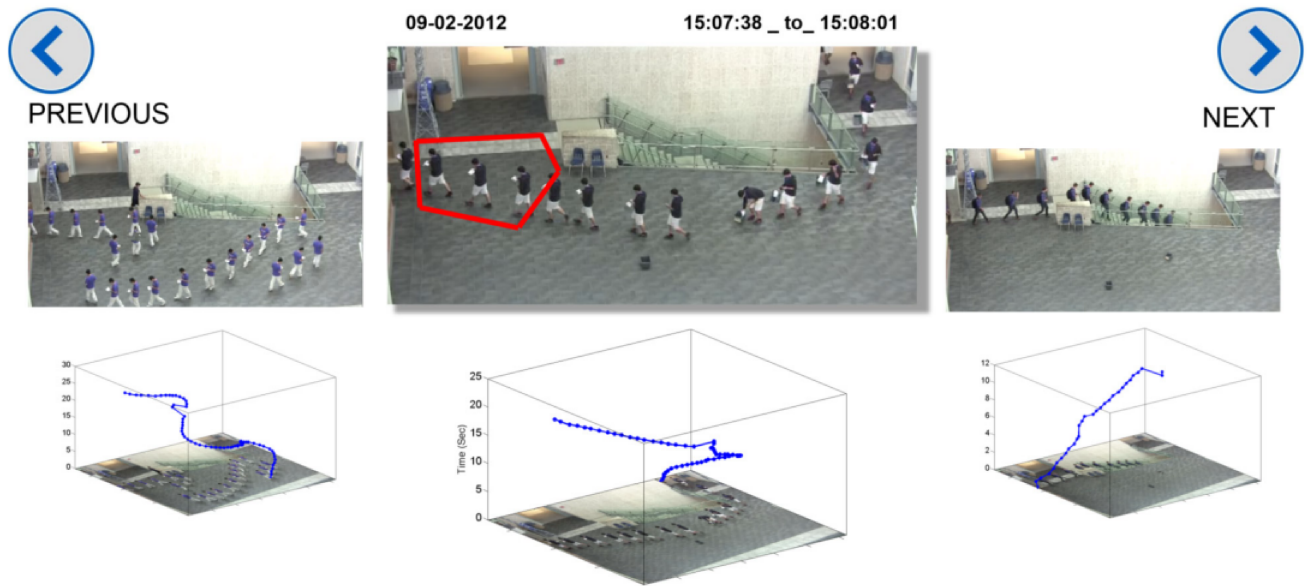


Figure 14 – Illustration of the object summarization process. (Figure extracted from (MEGHDAZI; IRANI, 2013))

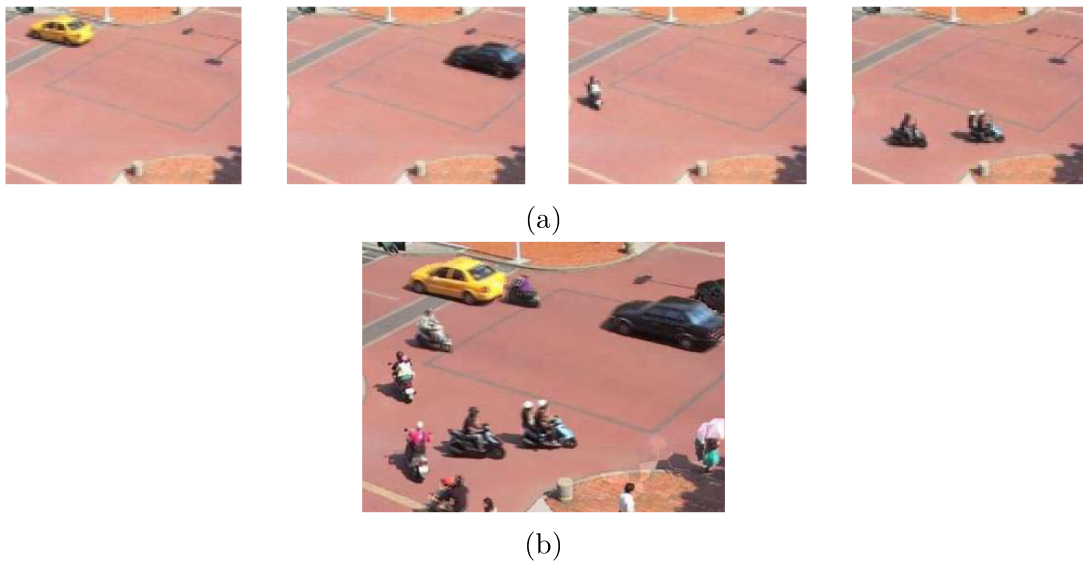


Figure 15 – (15a) Four input frame from a video. (15b) Generated video synopses. (Figures extracted from (HUANG et al., 2014))

non-parametric Bayesian method. Both visual and language cues are explored. In the layout presented in Figure 16, each activity step is related to a different color, and each video is organized in a line. The video structure may be quickly comprehended, allowing to visually filter types of activities and consequently understand the content of each video.

Viguier et al. (2015) presents a summarization of areas recorded by UAV (unmanned aerial vehicles), employing geospatial features from video metadata, combined with low level image features. The proposed method assembles a geospatial map using several pictures that merge local representations using geospatial references. A quick measurement of a long video is possible by observing a single static image that summarizes relevant

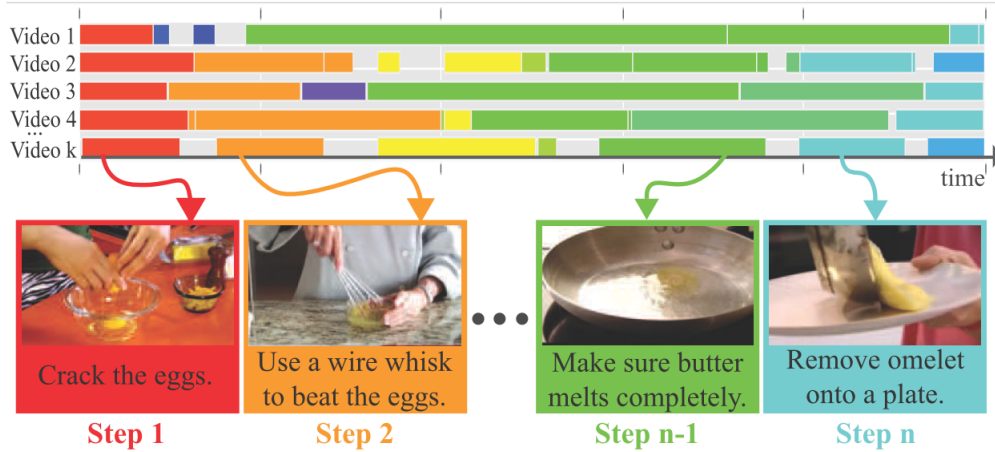


Figure 16 – Visual semantic storyline. (Figure extracted from (SENER et al., 2015))

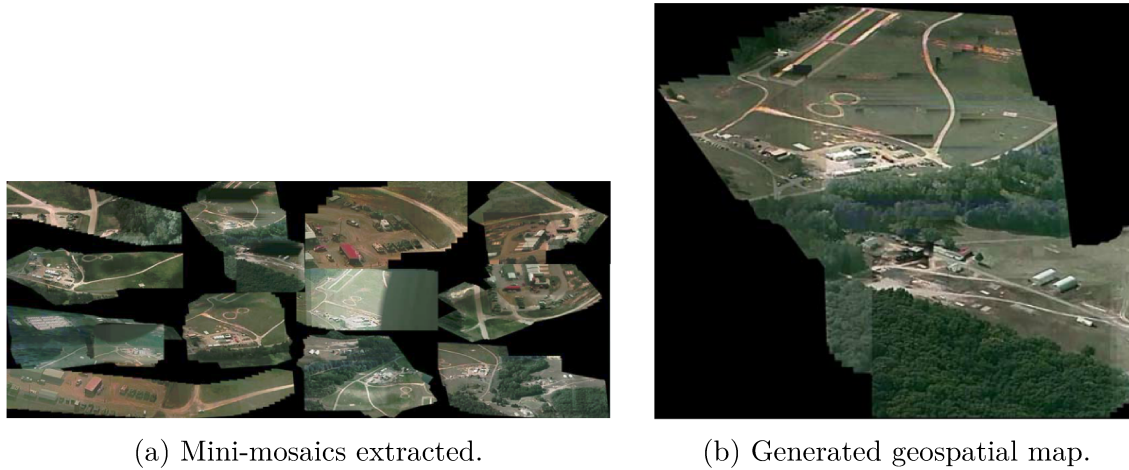


Figure 17 – Adapted from (VIGUIER et al., 2015).

information such as coverage, dwell time, activities, etc.

3.1 Final Remarks

The presented video visualization techniques mostly focus on providing an exact view of each frame, highlighting participant objects, and how they move over time. However, users must mentally identify the occurrence of events, creating a segmentation that may not be straightforward, specially if several events occur, sometimes simultaneously. It is also not trivial to identify the relationship among them, limiting the comprehension of involved phenomena. Additionally, video is a type of data that may be observed from different perspectives, depending on the user necessities. Some techniques consider only time or spatial aspect, but not both which make the analysis more difficult. Finally, some techniques do not provide means to interact with the data.

In contrast, we propose a visual analysis focused on showing the underlying video structure in terms of events, represented by groups in a interactive layout that creates a

natural but effective summarization. From an overview of the events occurrence, the idea is to associate groups of frames with events, to understand the relationship among them, to comprehend how one event leads to another one, as well as their temporal and spatial extension. Finally, by means of TSSM we expect to comprehend the temporal aspects of the video, how events are distributed over the video duration, as well as to identify event patterns, providing an overview about the video temporal structure.

Proposal

4.1 Introduction

We believe that take the user knowledge into account may improve the surveillance video analysis process and contribute to a more effective decision-making. We believe that Information Visualization techniques may represent a potential tool to provide the user insertion in this process, allowing the extraction of strategic information from these videos, by providing a smart overview about the more important video information through natural summaries, facilitating the identification of events of interest.

In the content captured by static cameras, disturbances in the scene result in abrupt changes in the frame content, which favors representations based on the similarity among video frames. Moreover, videos generally present several aspects from which the user is able to extract useful information to his/her investigation. In order to simultaneously cover several of such aspects, we employ the coordination of several techniques covering each of these aspects, allowing the exploration of different perspectives from the same data.

To adequately address this scenario, our strategy must be able to distinguish anomalous situations from normal situations. Furthermore, important temporal aspects, as well as their extents, must also be identified and highlighted. Finally, the user must be able to navigate through the video content to explore it according to his/her needs. In the following sections, we detail our methodology and present how the aforementioned requirements were addressed.

4.2 Methodology Architecture

Figure 18 shows a diagram of our methodology for the visual analysis of surveillance videos. Each of its components is detailed below.

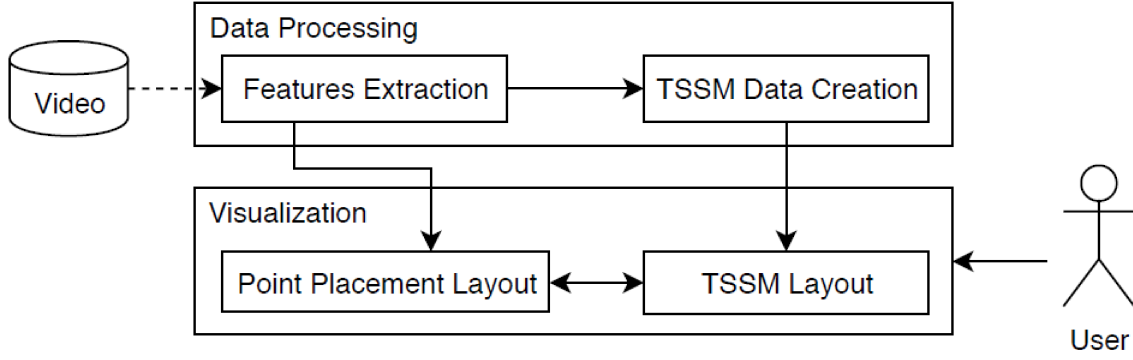


Figure 18 – Visual analysis methodology structure, showing all steps organized in structured components. Users interact with the layouts in Visualization component, using a set of interaction tools that provides the exploration of different video aspects with focus on events comprehension.

4.2.1 Data Processing

The data processing module apply a set of adaptations in the received data (video) in order to organize its informational content in such a way that makes possible the construction of the layouts employed in our proposal. Each procedure is detailed as follows.

Features extraction

The video is firstly splitted into a set of frames. Users may use the entire frame set or an arbitrarily defined subset. Features are then extracted from these frames, using any description strategy. Our methodology supports the use of visual features extracted from each frame or video-specific domain descriptors extracted from consecutive frames. Any preprocessing strategy may be performed to the features, for improvement of the associated content, to highlight specific aspects, or to filter background or noise information.

TSSM Data Creation

Temporal Self-similarity (TSSM) plays a crucial role in our methodology, as it is responsible for representing the temporal aspect of the video. Considering a set of n video frames $F = \{f_1, f_2, \dots, f_n\}$, a TSSM map can be defined as a square symmetric matrix $S_{n \times n} = [d_{f_i f_j}]$, $1 \leq i, j \leq n$, in which d_{ij} is the distance between features extracted from frames f_i and f_j . The main diagonal is composed by zeros, representing the distance between each frame and itself. Any distance measure can be employed to calculate the frames similarity. A diagram illustrating the resulting matrix is shown in Figure 19.

TSSMs have been employed in several human action recognition approaches (JUNEJO et al., 2011; EFROS et al., 2003). Experiments have shown that they present invariance under view changes of an action, providing an effective strategy in cases where viewpoints

are considered, as detailed in Chapter 2.

$$\begin{bmatrix} d(f_n, f_1) & d(f_n, f_2) & d(f_n, f_3) & d(f_n, f_4) & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ d(f_4, f_1) & d(f_4, f_2) & d(f_4, f_3) & 0 & & d(f_4, f_n) \\ d(f_3, f_1) & d(f_3, f_2) & 0 & d(f_3, f_4) & \dots & d(f_3, f_n) \\ d(f_2, f_1) & 0 & d(f_2, f_3) & d(f_2, f_4) & \dots & d(f_2, f_n) \\ 0 & d(f_1, f_2) & d(f_1, f_3) & d(f_1, f_4) & \dots & d(f_1, f_n) \end{bmatrix}$$

Figure 19 – Diagram of the similarity matrix embedding.

4.2.2 Visualization

We propose the use of two visual strategies to complementary represent spatial and temporal aspects of the video. Each of these strategies is detailed as follows.

Point-Placement Layout

Point-placement techniques, as described in Section 2.5, map individual instances into a visual display. Layout strategies based on multidimensional projections are a common application of such techniques, which aim to reflect relevant relationship observed in the original space. The final mapping is displayed using a scatter plot strategy, in which the distance between points is directly related to their similarity.

In the proposed strategy, we employ a frame set from a single video, from which a layout is built. The Point-placement techniques normally receive as input a **distance matrix** structure, which is a symmetric matrix composed by the pairwise distances between all data set instances. The TSSM data structure computed by our methodology can be considered as a distance matrix in this process, but other matrices constructed using any computing strategy may also be considered. Each frame is mapped to a point in a scatter plot, and is represented by a circle. Its position is defined by the Point-placement technique. Figure 20 illustrates the resulting layout. The first and last frames are represented by bigger circles colored in green and red, respectively, in order to facilitate the temporal localization of video extents by the user. Each point represents a frame from the video, such way that images with similar features are placed close. Outliers points or groups, generally indicate some distinguishable event. Smooth action or no action moments are represented in the layout by rounded-groups, and groups with sparse points indicate moments with high actions. The analysis in the produced layout then consists

in exploring the distribution of the points, as well to investigate regions or formed groups in the layout.

Our methodology supports the use of any Point-placement technique. Surveillance videos captured by motionless cameras produce frames in which content changes may represent movements that characterize events and also transitions between them. It favors the use of point-based techniques, specially multidimensional projections, in the sense that the observed groups may be related to identified events, and the relationship among these groups may provide insight about how these events relate to each other. We expect frames with similar content to be placed close to each other, naturally creating group of points which will spatially segregate different moments from the video and possibly reveal interesting video events discriminating abnormal from normal situations.

PETS2006 (BASHIR; PORIKLI, 2006) is a video collection explored in several video surveillance works, mainly in object detection tasks (TIAN; FERIS; HAMPAPUR, 2008; BAYONA; SANMIGUEL; MARTÍNEZ, 2009). It basically shows an actor in a crowded subway station abandoning a bag and leaving the scene. Figure 20 illustrates an example of the layout produced from the video set PETS2006 by our methodology, manually labeled to illustrate three main actions: 1- Before the actor appears, 2- The actor abandons a bag and 3- The bag is left alone in scene. The legend in the layout indicates each of these moments. One can notice that these three moments produced separated groups in the layout. The groups also present different formats, which may indicate different types of behavior. Finally, it is possible to explore the relationship among these groups. A detailed analysis of several scenarios in which these layouts are employed is presented in Chapter 5.

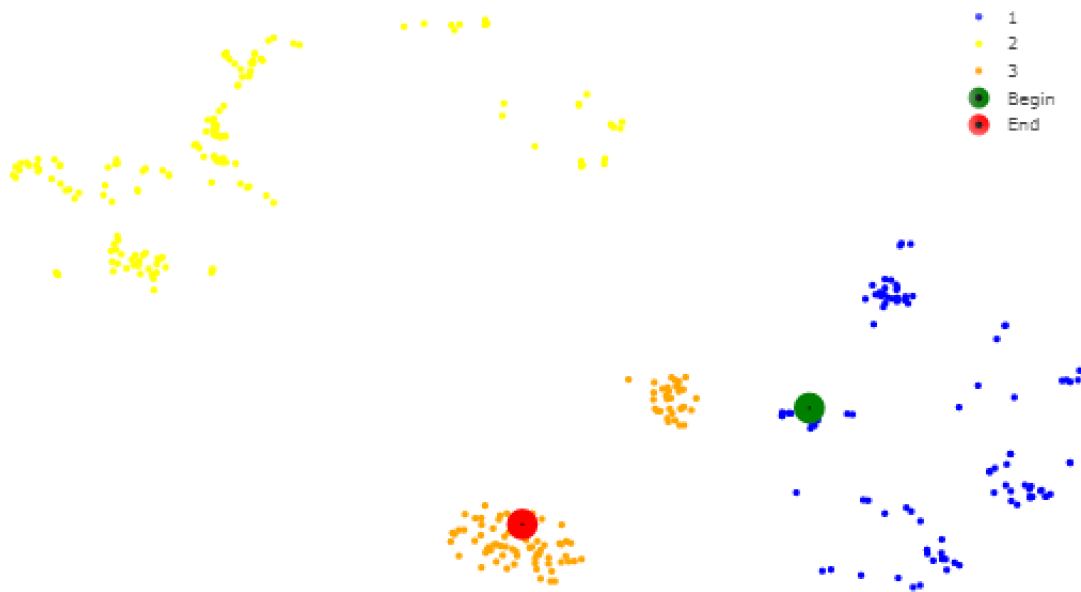


Figure 20 – Label function and interactive legend. Each group represents a set of images with similar content.

TSSM Layout

Andrienko et al. (2013) suggest that the use of two displays, one mapping spatial information, and another one mapping the time information is more effective for the inspection of spatiotemporal data (FERREIRA et al., 2013; LANDESBERGER et al., 2016; ONO; DIETRICH; SILVA, 2018). In our methodology, TSSM layout is responsible for convey temporal information, mapping a color coding to the computed similarity values in TSSM data structure. The idea is to visually highlight the patterns produced by the TSSM data, and quickly communicate the periods when events occur, their duration, and the degree of disturbance they cause. Additionally, they may reveal more complex behavior patterns related to the type of event. Figure 21 shows an example of the resulting layout. Higher frame similarity values are mapped to low intensity cells, while those representing low similarity are mapped to high intensity cells.

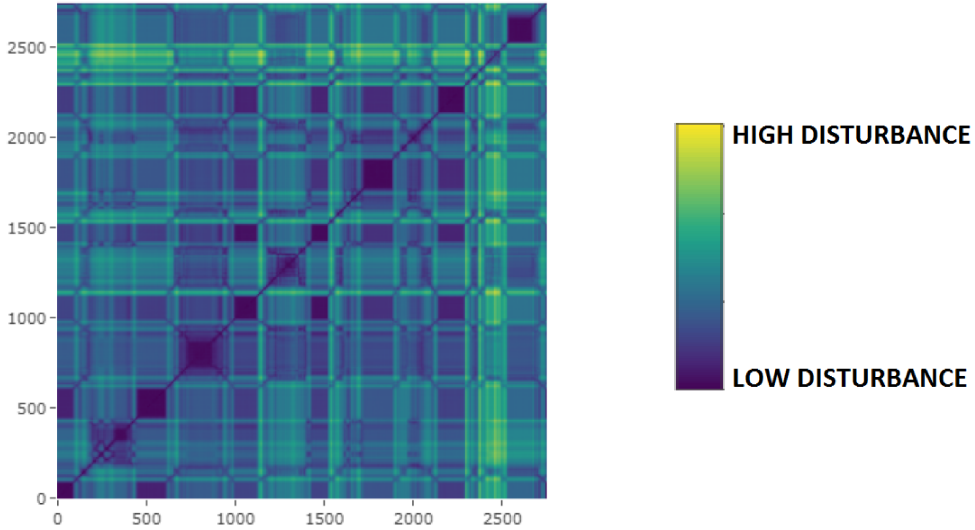


Figure 21 – TSSM layout example. In moments with high disturbance values there are more difference between the frames content, and in moments with low disturbance values the frames are similar to each other, suggesting little or no movement.

Self-similarities can be defined over both static and dynamic features, being suitable to be applied in our methodology. In addition, as the Point-placement layouts only convey spatial aspects from the event occurrence, coordinate them with a TSSM layout help the users to keep the temporal order idea in mind, at the same time providing an additional perspective that improves the analysis process.

The patterns produced by TSSM layout may provide important information regarding events occurrence. The authors in (XU; TAX; HANJALIC, 2012b; KRISHNA et al., 2014b) suggest typical patterns and how they shall be interpreted. We propose an adaptation of these semantic interpretations to event analysis and suggest other patterns that may also be observed in the layout. These patterns are listed as follows.

- ❑ **Horizontal/vertical analysis:** allows to establish a moment of reference and compare it against the whole video period, showing how scenes changes compared to this specific moment;
- ❑ **Horizontal/vertical segments:** allow to investigate the persistence of events, which is proportional to the segment size. The analysis of these segments may reveal moments containing relevant information to the analysis, as well as the extents of these moments. Figure 22 shows a highlighted row from a TSSM layout comparing a specific moment (frame 0) against all the remaining ones. It is possible to notice that, starting in frame 20, the similarity to all frames is low, except by the similarity with the last frame, which indicates that variations occurred during the video, and that they ended before the end of the video. Moreover, from moment 0 to the 20, no events occurs, which suggest that, depending on the analysis, this video segment may be ignored;



Figure 22 – Horizontal/Vertical analysis of the frame 0 in a TSSM layout.

- ❑ **Fading:** generally represents the speed of a transition between scenes. As scene transitions usually occur sequentially, the observance of such pattern in diagonal allows to understand how transitions in the video original flow occur. This pattern may also be in horizontal/vertical analysis of TSSM, indicating how gradual a change in the scene occurred compared to a reference moment. In the TSSM presented in Figure 23 one may notice a change occurring in frame 1100, due to the color variation presented. Around this region the cell colors vary gradually, from a high to low intensity, indicating how gradual this change occurs;

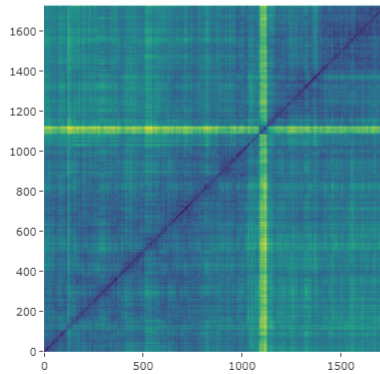


Figure 23 – Fading effect and isolated points analysis in a TSSM layout.

- ❑ **Diagonal analysis:** represents the difference between consecutive frames during the entire video. This region and its neighborhood may provide insights about the

overall behavior movement and scene changes in the video, considering the original flow. Sequential changes in regions close to the main diagonal, which represents differences between consecutive frames, represent scene changes over time. The fading size and color intensity of such scene transitions indicate how long this specific action causes modification to the video scene and the number of actions in the video (the amount of changes), respectively. Homogeneous squared areas near the main diagonal represent moments without or insignificant variations between consecutive frames. The pattern repetition are related to scene reoccurrence. The diagonal of the TSSM presented in Figure 24 highlights some of such mentioned patterns;

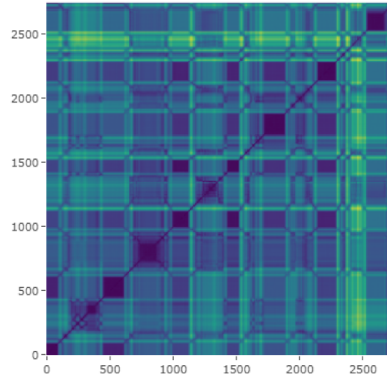


Figure 24 – TSSM diagonal analysis.

□ **Internal region analysis:** used to comprehend the similarity among two non-consecutive moments. In this sense, homogeneous regions indicate moments in which potentially similar events occur, once the distances between the frame features that compose them are small in the original feature space. In Figure 25, regions marked from 1 to 4 represent homogeneous areas, the moment marked as 2-4 with a white square represent how similar moments 2 and 4 are to each other, which can be noticed by the associated low intensity color;

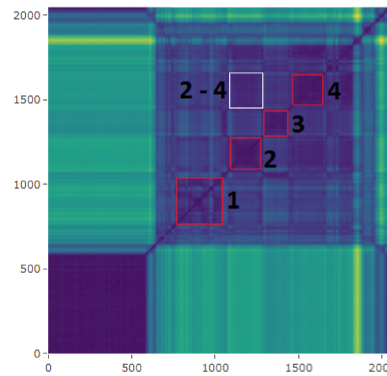


Figure 25 – Internal regions and homogeneous areas analysis.

□ **Pattern recurrence:** may indicate recurrence of actions, or cyclic events. Such pattern can be noticed in the main diagonal of the TSSM shown in Figure 24,

by observing periodic regions with high intensity color. In addition, regions that represent the transitions between these recurrences may indicate an event which interrupts the occurrence of a specific action, that is, an event that abruptly changes a state in a video scene, but also abruptly causes the return to this previous state once or several times.

4.3 Surveillance Visual Analysis System

We developed a computational system that implements our proposed methodology. The system interface is shown in Fig. 26 and consists of two main coordinated views: *Point-placement* (A) and *TSSM* (B), as well as an *Interactive Timeline* (C). Users are able to use these components after loading the video and extracting its features, according to a specific description approach. Each component is described as follows.

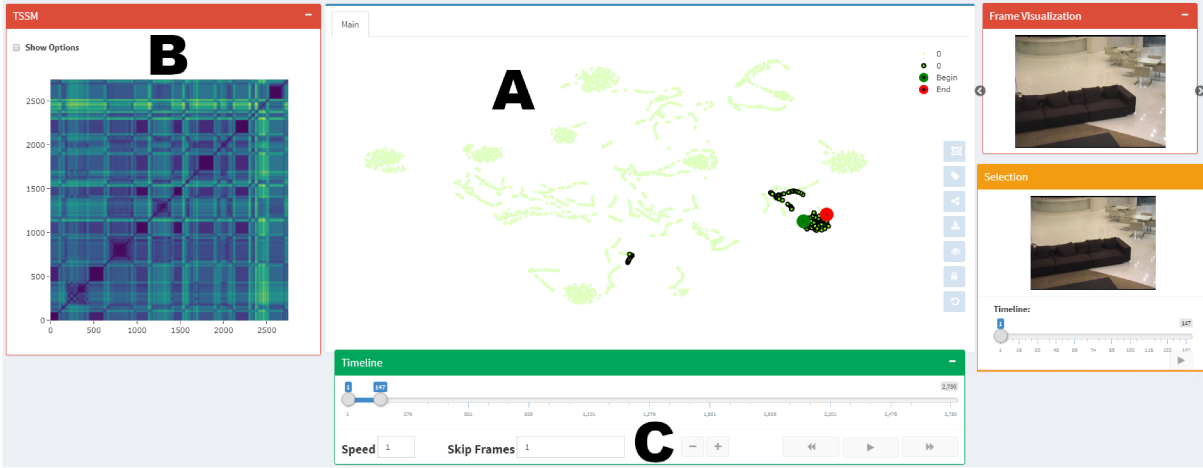


Figure 26 – Video visualization system interface, showing all coordinated views. Point-placement view (A) allows for exploration of the video structure in terms of event occurrence, from a spatial perspective, focusing on how the frames compose events, as well as how these events relate to each other; TSSM view (B) allows for an analysis of the events temporal aspects, as well as the comprehension of specific behavior related to different types of event; and Timeline (C) allows for a temporal navigation and exploration of the entire video, or selected periods, using conventional video player tools.

4.3.1 Point-placement View

The *Point-placement* view presents a scatter plot layout that provides the exploration of the underlying structure of the video contents in terms of event occurrence. Any Point-placement technique can be used. Each point in the layout represents a distinct frame, and the distance between each of them reflects their similarities according to the descriptor perspective.

A set of interactive tools provide suitable layout exploration. The most basic way to interact with such layout is through the manual selection. The selection may be performed using a square, or freely, as illustrated in Figure 27. It is also possible to perform a **locked selection**, as illustrated in Figure 28. In this case, a user selects a subset of frames, and then performs another selection considering only instances from the previous one. Finally, it is possible to split the layout to perform the analysis in only a subset of frames, using a separated tab, specializing the analysis in this subset. All the interaction available in the main layout are also available in the new tab.

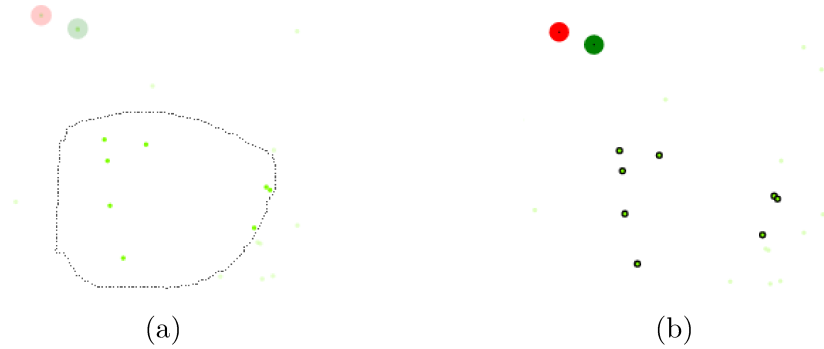


Figure 27 – Manual instance selection in the projection layout.

We developed an interaction tool to provide temporal context into the Point-placement view. The idea is to allow the user to follow the video path through the layout and to comprehend how the events occurred as well as how they are temporally related. To indicate the temporal flow, we employed edges connecting each consecutive frame, as employed by Otani et al. (2016). The origin frame of an edge is indicated by its horizontal part, and the destiny of the edge is indicated by its vertical part. This tool eliminates the need of drawing an arrow to indicate the flow. This strategy is illustrated in Figure 30a, and the resulting layout is shown in Figure 30b.

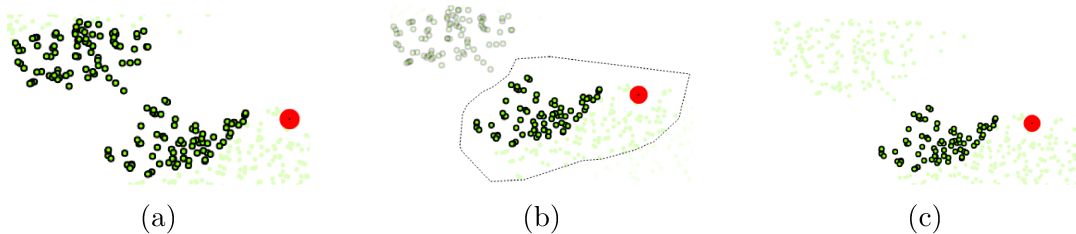


Figure 28 – Locked selection.

Users can also to label groups of points to visually segment the video according to any arbitrary criteria. An interactive legend can be used to describe each group, and a color mapping allows the distinction among them, as shown in Figure 31. In addition, it is possible to filter instances belonging to specific labels using this legend. Clicking in such legend will show/hide all correspondent frames.

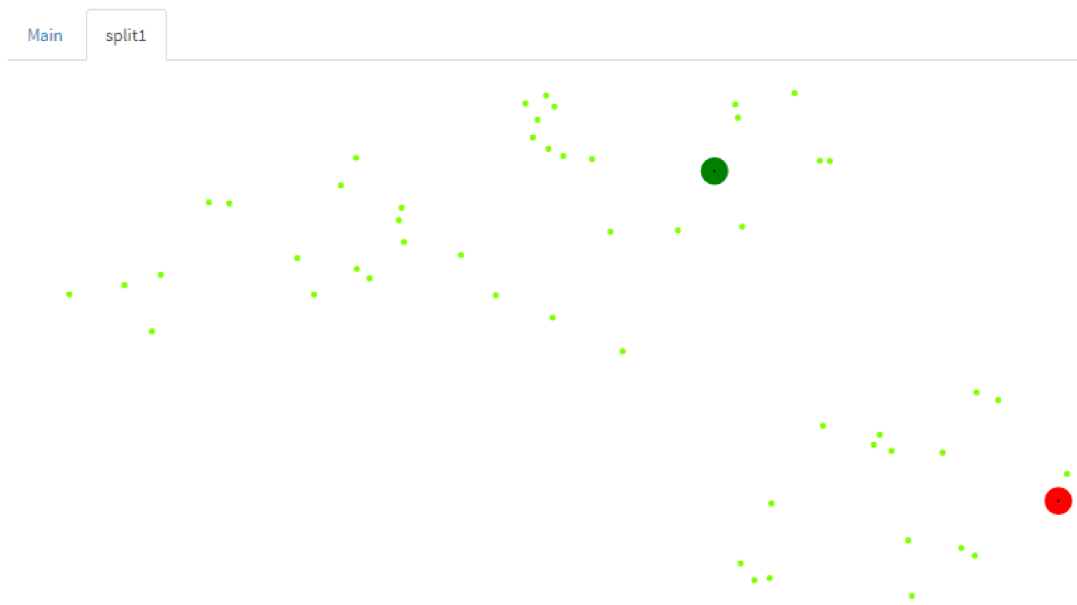


Figure 29 – Split resulted from the same selection made in Figure 28b.

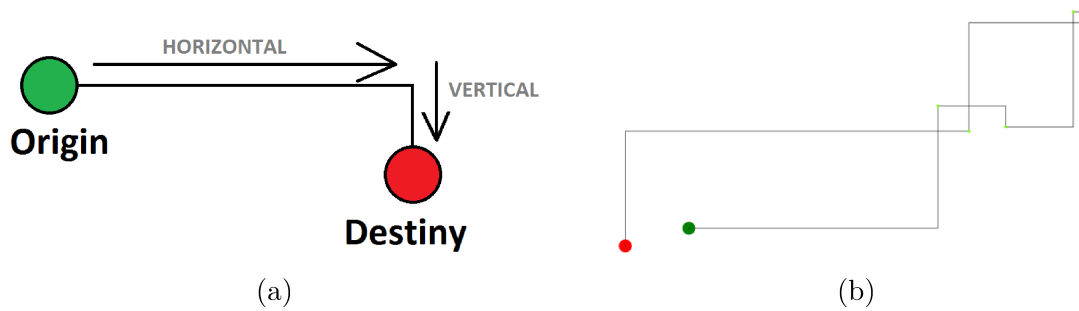


Figure 30 – Path between sequential instances.

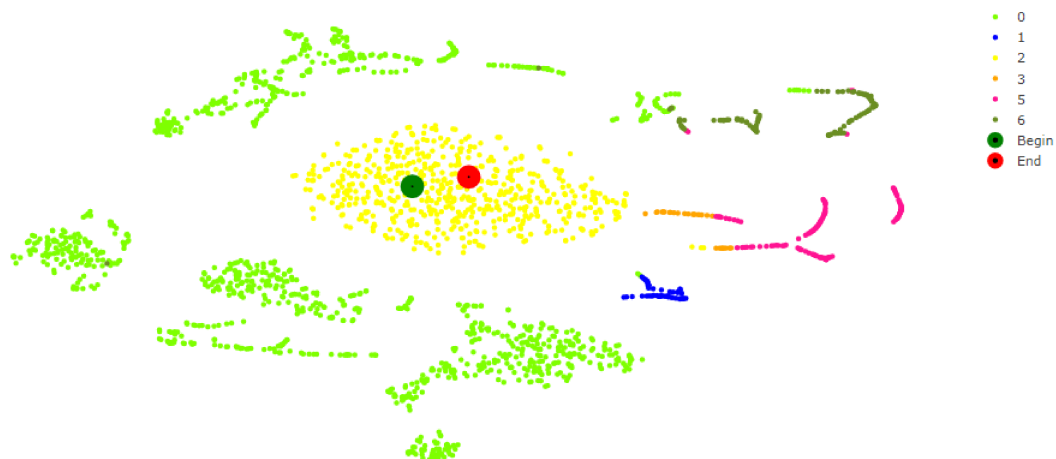


Figure 31 – Labeling tool and interactive legend.

Users are able to make searches in the video using the layout, considering both spacial (similarity) and temporal (video sequence) aspects, refining the data to easily find information of interest. It is possible to search for the the k -most similar frames of a specific frame, as well as the k -most temporally related frames. Using the spatial search, users inform a value k , and we search the k closer frames in the layout. Using the temporal search, users inform a value k , then the k most temporally closer frames to the selected frame are highlighted. Figure 32 illustrates the results of these selections, considering, for $k = 4$, frames [A,B,C,D] as a result of a spatial search (Figure 32a), and frames [A,E,F,G] as a result of a temporal search (Figure 32b).

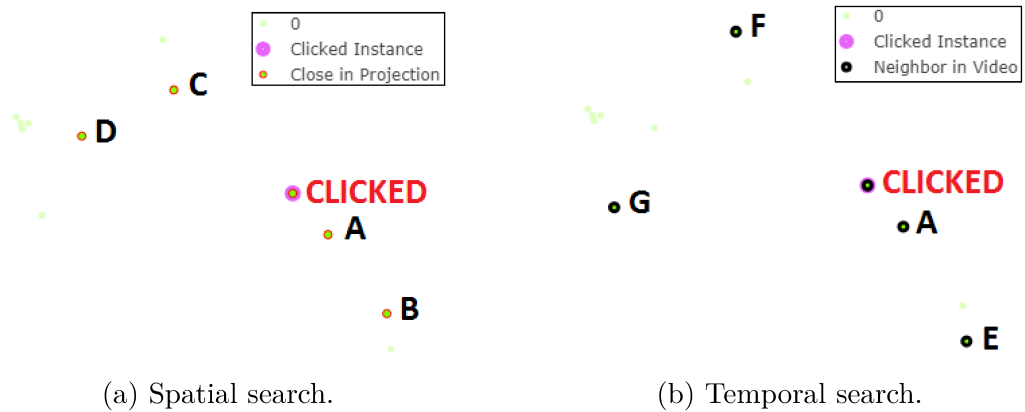


Figure 32 – Projection exploration employing “spatial and temporal searches”.

4.3.2 TSSM

The *TSSM* view presents a visual encoding for the TSSM data previously computed. The idea is to allow a temporal comprehension of the event occurrence by means of the presented visual patterns detailed in Section 4.2.2. Users are able to select cells and check the similarity between correspondent frames by viewing their content, and by visualizing the respective position in the Point-placement layout. It is also possible to zoom into selected regions for a more detailed analysis.

4.3.3 Timeline

The *Timeline* is a traditional interaction tool used for videos navigation. We employ it to play the video contents sequentially, as commonly used video players, and to assist the interaction in the other views. Basic commands were implemented, such as play, advance/rewind frames, pause/stop, as well as an interactive slider that indicates the time period under analysis, also working as an additional temporal selection tool. Users are also able to modify the video play speed, in frames per second, and the step size, in number of frames, when advancing/rewinding frames. Figure 33 shows the timeline and a

Point-placement view reflecting the selection performed in the timeline selector, indicated by the highlighted points.

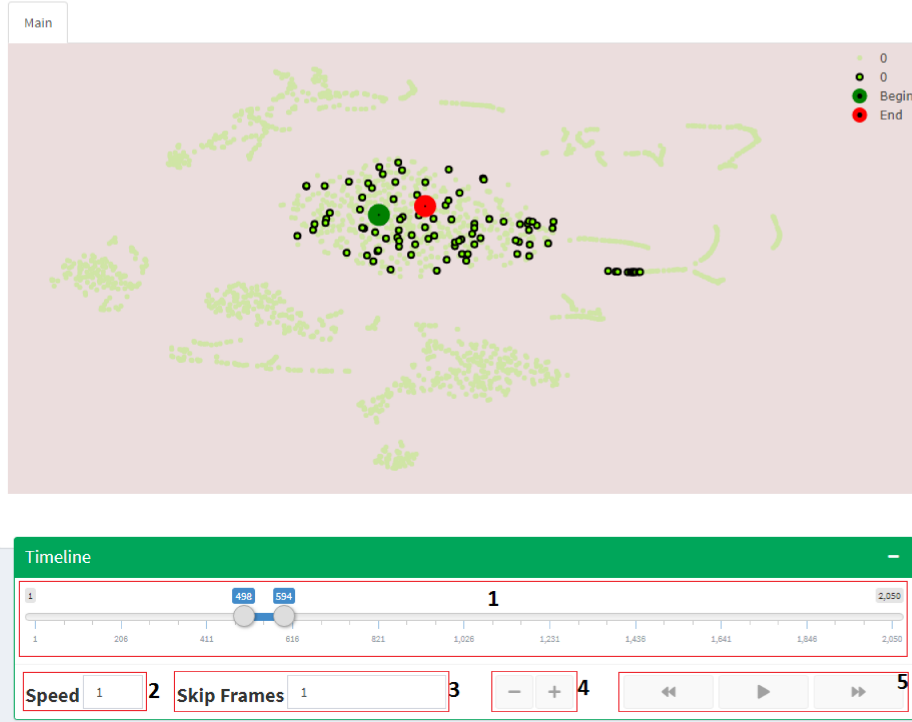


Figure 33 – Timeline employed in our system. **Timeline selector (1)** is used to select a range of sequential frames; **Speed parameter (2)** defines how fast the timeline selector will play; **skip frames parameter (3)** is the quantity of frames to be skipped in each step of an advance/rewind/play interaction; **minus/plus buttons (4)** increments/decrements the number of selected frames forward in the timeline selector; **advance/rewind/play buttons (5)** are traditional video player commands used to interact with the video.

Figure 34 shows an example of the interaction performed in the timeline, using the advance button in three sequential steps. In each step 510 instances are selected, in such a way that firstly the frames 1-510 are shown, then 511-1020, 1021-1530 and finally 1531-2040. Each interaction in the timeline is reflected in the Point-placement layout, highlighting the selected instances.

The **timeline selector** reflects, in the timeline, the selection performed in the Point-placement view, as shown in Figure 35. The temporally first and last frames in this selection are used to delimit the period selected in the timeline. This functionality provides an idea about the temporal extension of a selection in the Point-placement layout. Moreover, an additional timeline is displayed for each selection performed in the Point-placement layout, in order to provide a navigation tool for the selected frames.

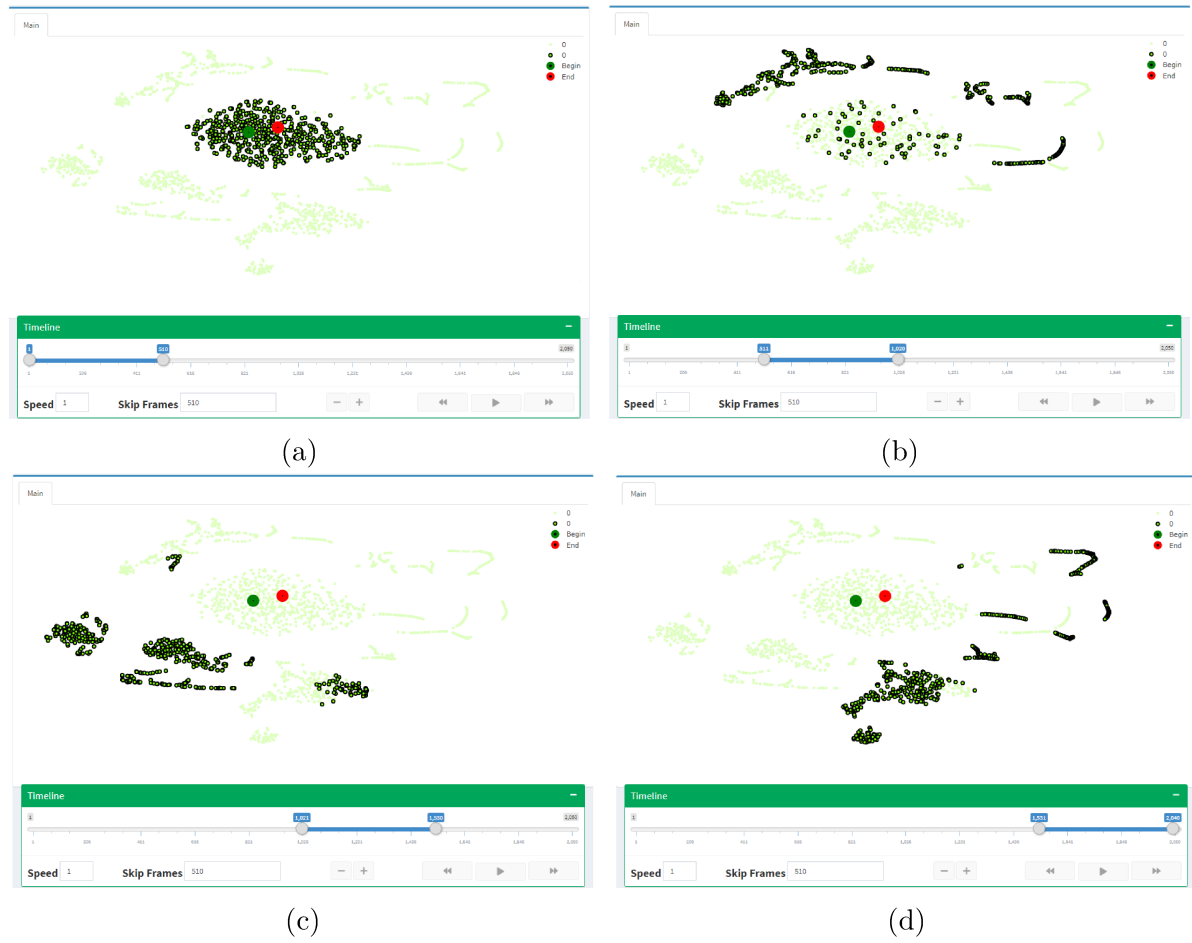


Figure 34 – Interaction through the advance option/button.

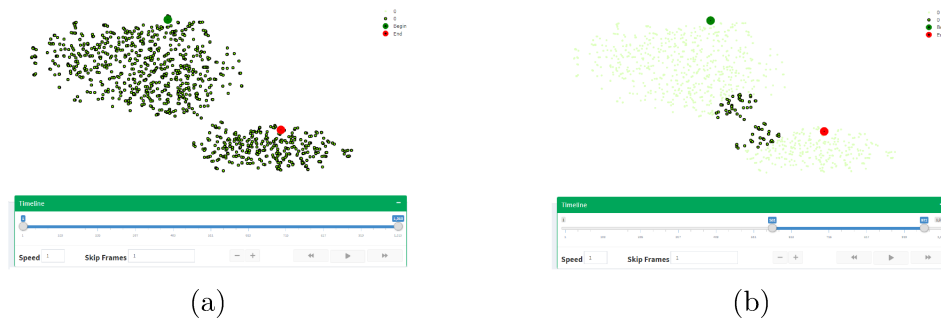


Figure 35 – Timeline updating through selection in the projection layout.

4.3.4 Layout coordinations

All the layouts described in previous section are coordinated, in order to explore, simultaneously, different video aspects, improving the information extraction by the user.

The TSSM layout can be constructed considering only selected points in the Point-placement layout, providing a localized analysis. If no points are selected in the layout, the TSSM is built considering all frames. Local information can be analyzed more precisely. Figure 36 exemplifies this interaction.

The selection of a cell in the TSSM layout also reflects on the Point-placement layout, in which the points considered in such cell are highlighted. A separated window also display the contents of these frames in thumbnails, as shown in Figure 37.

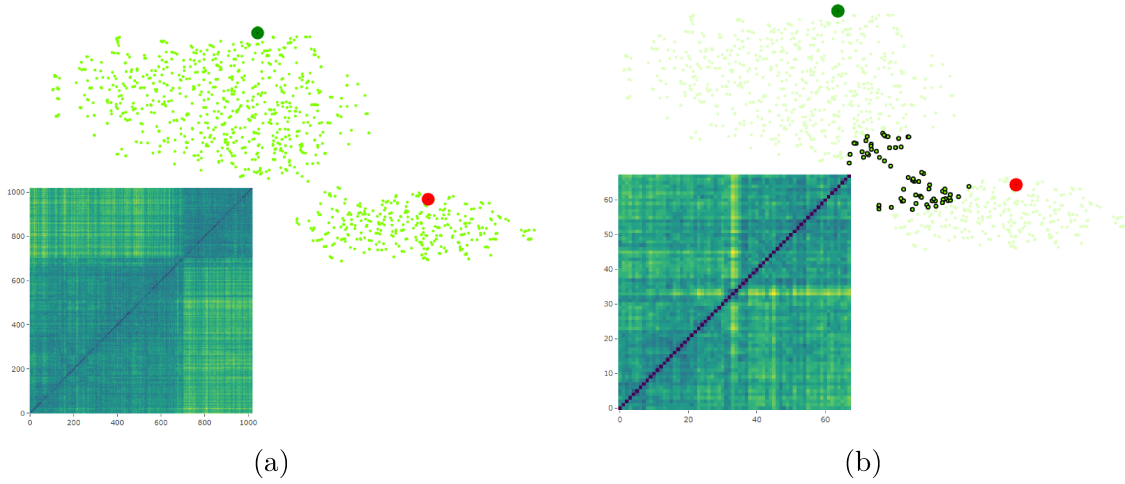


Figure 36 – TSSM updating selection in the projection layout.

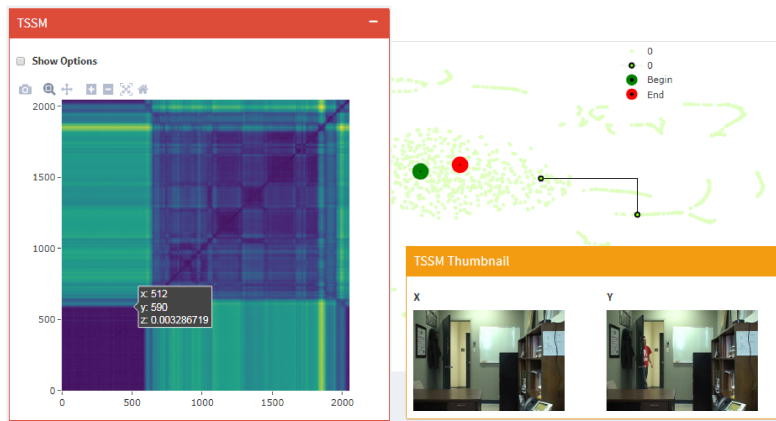


Figure 37 – Cell selection in TSSM reflected in the Point-placement layout.

4.3.5 Smart Sampling

Depending on the video duration, considering all frames for analysis can be computationally costly, specially in surveillance scenarios, often characterized by the presence of

several frames with non-significant content changes. Thus, a sampling procedure considering a subset of frames that maintains the underlying video structure may be applied. This sampling is usually done uniformly, considering a specific number of frames selected per time interval. However, as the events occurrence is not necessarily uniform, this approach may potentially result in the loss of important events or in the inclusion of undesired redundancy.

We added a tool to the system that implements a smart sampling procedure, based on TSSM data, focusing on the event occurrence. Users may be interested in performing an analysis with focus on periods with abrupt changes in the scene, prioritizing those frames whose transitions present more significant changes. On the other hand, some situations may require the analysis of stationary actions, or periods of no/low movement, in which less significant transitions might be better to consider. In TSSM, the set of cells adjacent to the diagonal is composed by distances computed between consecutive frames. Our procedure thus sample from this set the k ones with higher distance values, in which k is a parameter representing a video proportion. The resulting set will contain those frames whose transitions present more significant changes, which we believe may potentially delimit video segments that will maintain the events occurrence. The resulting set will represent a compact version of the video highlighting specific aspects from events occurrence structure. We summarize our approach as follows:

1. Select, from TSSM matrix, a set $C = \{d_{f_i-1f_i}\}$, $0 \leq i \leq n$;
2. Rank C according to the distance values;
3. Select k first frames (situation 1) or k last frames (situation 2) from C . These frames will compose the video sample set.

4.4 Implementation Details

The surveillance visual analysis system was implemented as a client-server web application, using the R Shiny ¹ library/framework. Both server and client were implemented in R and the communication is intermediated by Shiny modules responsible for handle the events among them. All computational calculations are performed in the server, such as TSSM Data Structure, Smart Sampling, Features Extraction and Point-Placement algorithms. Our developed system provides five point-placement algorithms, t-SNE, LSP, LAMP, PLMP and Force Scheme.

At the user interface side, the main layouts of our methodology were implemented using the Plotly R library, that employs D3.js and WebGL libraries to render all visual mappings and also provides basic interactions. The more elaborate interactions and functionalities, such as the timeline, were implemented in R using visual components provided by Shiny.

¹ <https://shiny.rstudio.com/>

Other R libraries were used to support the computational calculations, using high level data structures, visual components, point-placement algorithms, image description, file reader/writer and other. Finally, additional JS and CSS were embedded in the application, mainly to set the layout style and other basic tasks.

Experimental Results

In this section we present the results of our methodology applied in several case studies involving distinct surveillance scenarios. We first explore the general structure of the layouts in order to identify patterns that may suggest a summarization in terms of the events contained in the videos. Furthermore, we try to relate the already known events on the videos to the produced groups aspects, in terms of size, shape, etc., and also how these events relate to each other by looking at the groups positioning. Finally, by looking at the TSSM we explore the temporal aspects related to the video, as well as the coordination of such layout with the point-placement layout and how it may improve the visual analysis.

5.1 Videos

We describe in this section, the surveillance videos we used in our analysis. CD-Net2014 ¹ (GOYETTE et al., 2012) is a repository consisting of indoor/outdoor videos for identification of changing/moving areas, including several surveillance scenarios. We used two videos from this repository. The third video belongs to VIRAT ² (OH et al., 2011), a collection of several ground-recorded videos containing examples of human/vehicle events. All these videos are described in the following sections.

5.1.1 Office

OFFICE is composed of 2050 frames of a person entering an office, collecting/reading a book, and leaving this office. All the actions which compose this video are described in Table 1, and a set of frames representing key actions are presented in Figure 38. Once this video represents distinct human actions, such as walking, picking objects, moving in a scenario, changing body positions, among other actions. Such analysis is interesting to

¹ www.changedetection.net

² <https://data.kitware.com/>

comprehend how this scenario is represented by our methodology. All this actions occur inside an office and the actor is the only moving entity in the video.

Table 1 – Descriptions of main actions occuring in Office video.

Segment	Frames In- terval	Description
1	1-578	The office is empty
2	579-611	The actor appears
3	612-651	The actor walks at the middle of the room
4	652-688	The actor starts to take the book out of the shelf
5	689-768	The actor takes the book and starts to read it
6	689-768	The actor reads the book leaned on a furniture piece
7	1056-1826	The actor reads the book without leaning it on the furniture piece
8	1827-1877	The actor starts to close the book and save it
9	1878-1937	The actor saves the book completely in the shelf
10	1938-1966	The actor walks at the middle of the room again
11	1967-2018	The actor begins to disappear through the door
12	2019-2043	The actor disappears completely
13	2044-2050	The office is empty again



(a)



(b)



(c)



(d)



(e)

Figure 38 – Set of frames representing key actions from the OFFICE video. (a) Office empty; (b) actor enters; (c) actor picks the book; (d) actor reads the book; (e) actor leaves the office.

5.1.2 SOFA

SOFA is a video containing 2750 frames with people entering/leaving a scene, and objects being abandoned, moved and removed from this scene. Some frames that summarize the main actions are illustrated in Figure 39. The analysis of this type of video is interesting because it is characterized by a scenario with movimentation of multiple objects, representing situations that refers to bombs left, objects stolen, among others.

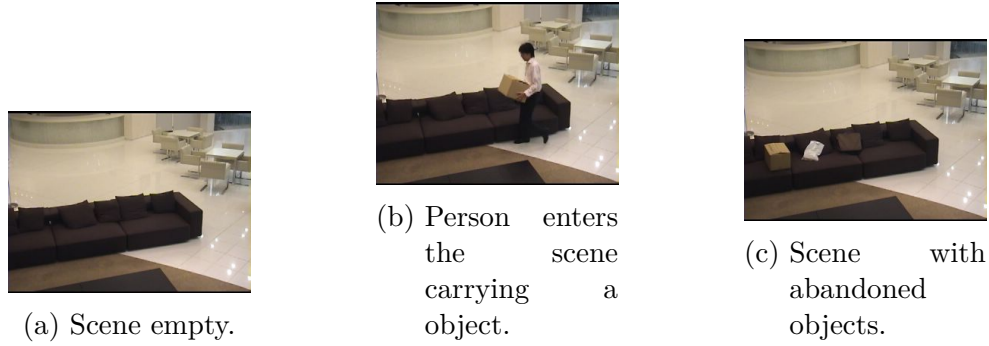


Figure 39 – Some moments from the SOFA video.

5.1.3 VIRAT

We used **VIRAT_S_010108_01_000570_000718** video, named here as **VIRAT**, composed by 1019 frames, distributed along 144 seconds, in which a person gets into a vehicle in a parking lot. This video may represent situations in which a object is removed of its original place, and identifying these moments may be crucial to find clues of robberies or murders.

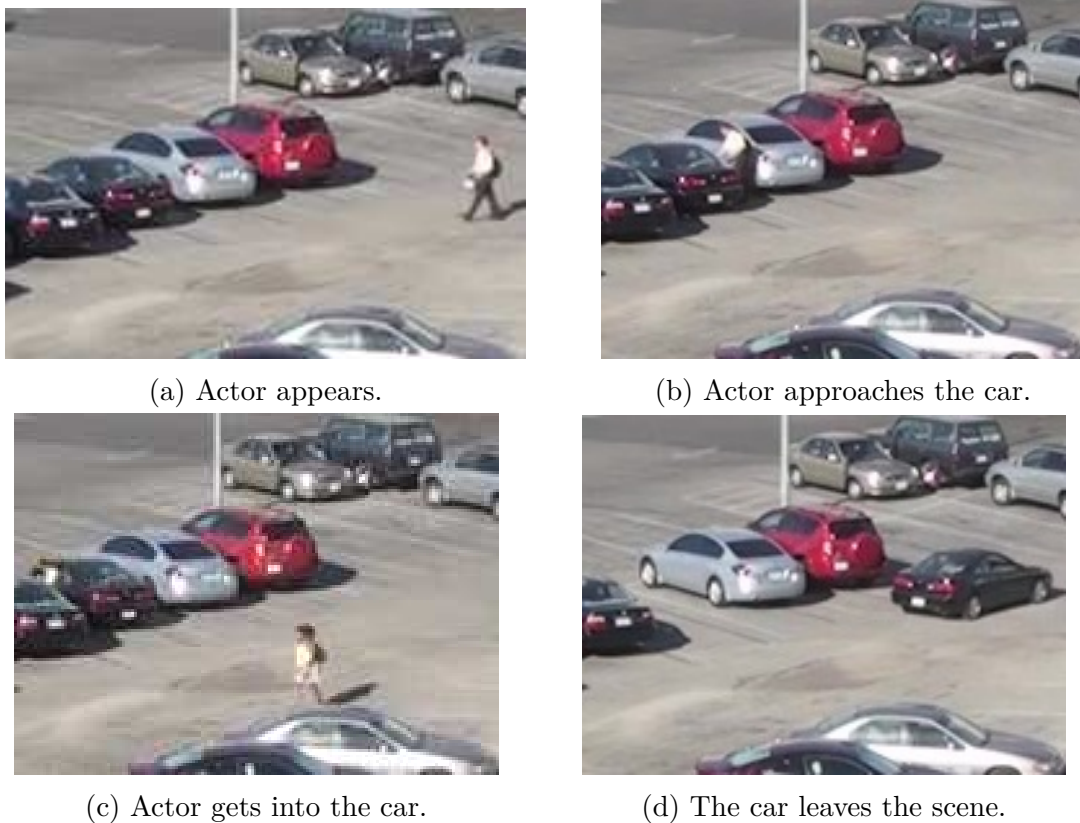


Figure 40 – Important actions occurring the VIRAT video.

5.1.4 Experimental Process

From each video, HOG descriptor was employed to extract feature vectors from the whole content of each frame. HOG descriptor was chosen due its capability of capturing edge structures characteristic of local shapes and its robustness to noise/small changes in objects illumination/orientation (BORGIO et al., 2012), producing a discriminative feature vector. It is also suitable to be used in event detection because events in these scenarios are often characterized by changes in object shapes, such as a person entering/leaving a specific place or changes in people movement. HOG descriptor has been used in several smart surveillance approaches (BENFOLD; REID, 2011; CAO et al., 2011; KRISHNA et al., 2014b; BARBU, 2014). We extracted HOG features using 3 cells and 6 orientations, resulting in a 54-dimensional feature vector. This configuration were empirically obtained. We employed a t-SNE multidimensional projection technique (MAATEN; HINTON, 2008b) to construct the point-placement layout. t-SNE is known to produce high neighborhood preservation and to highlight local relationships, potentially favoring the exploration of the groups internal structure, and has been employed successfully in several video visualization tasks (XU; TAX; HANJALIC, 2012b; RAMANATHAN et al., 2015b; LIAO et al., 2016). We also performed a set of experiments comparing state-of-art point-placement techniques, and t-SNE provided a better group separation. As detailed in the following sections, it favored a satisfactory visual analysis. We have used the Rtsne R package ³, with default parameterization.

For each video, we performed an overview analysis of the point-placement layouts to identify the produced groups and related patterns, as well as their correspondence with events occurred in each scenario. We also evaluated the relationship among groups, to comprehend how the events are visually related, coordinating the exploration with the TSSM layout to contextualize these patterns over time and to explain the observed phenomena. We follow the TSSM interpretation suggestions presented in Section 4.2.2, in order to verify how effectively they guide the user in comprehending the temporal properties of the events.

5.2 Results

The analysis of the mentioned video are presented in this section.

5.2.1 OFFICE Video

Figure 41 shows the t-SNE projection from Office Collection. Each instance color represent different video segments, labeled manually according to the relationship among consecutive frames, by watching/analyzing only the video. The segments descriptions and

³ <https://cran.r-project.org/web/packages/Rtsne/index.html>

its respective frames intervals are presented in Table 1. One may notice t-SNE preserved the video local structure, once instances representing similar consecutive frames were placed close. It is an advantage of use t-SNE, since it determines local neighborhood size for each instance separately based on the local data density. Aspects of global video structure also was maintained, enabling identify some “anomalous” situations, as shown in the following analysis.

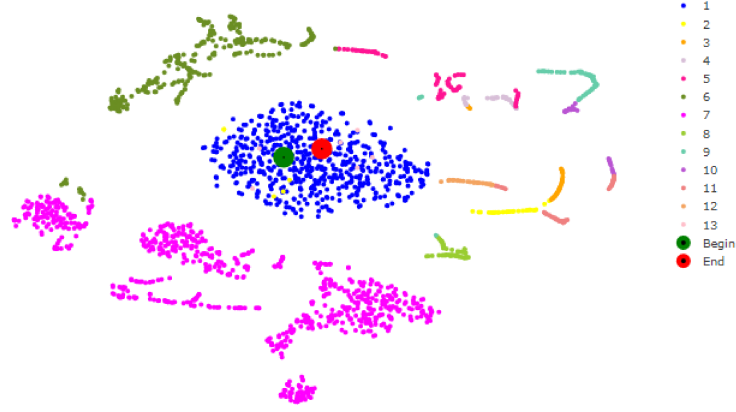


Figure 41 – t-SNE projection of Office video labeled manually.

Figure 42 illustrates the t-SNE projection (Figure 42a) and respective TSSM (Figure 42b) layouts for OFFICE video.

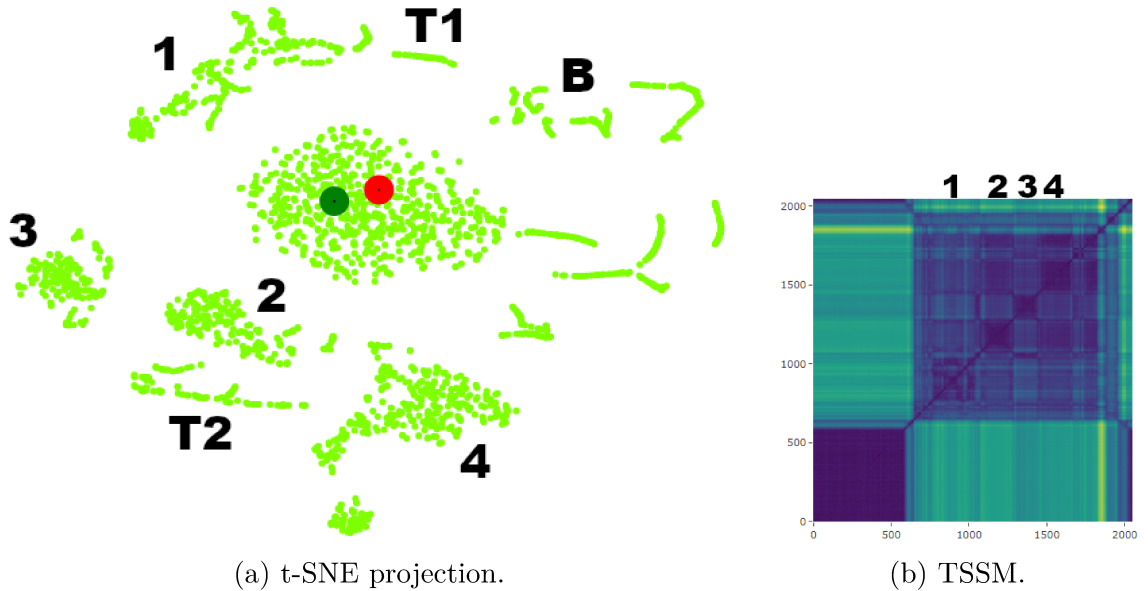


Figure 42 – OFFICE video layouts.

Several groups may be noticed in the point-placement layout, as well as a large group in the center. Such group contains the first/last video frames, suggesting no permanent changes in the scene. It also contains all the frames in which the actor is not in the scene. Each of the remaining groups contains frames representing different ways the actor reads the book, as well as different body positions. Some of these groups were numbered in the

layout, and their representative frames are illustrated in Figure 43. The point-placement layout positioned similar events together, spatially organizing the video content by types of actions. Moreover, this layout organization of the video produced a group in the center representing a "normal" state and several groups representing events in different regions around this center, suggesting an interesting analysis strategy. One may notice that groups 2 and 4 are close positioned in the layout, suggesting that these events are similar to each other. The images representing these groups are shown in Figs. 43b and 43d, confirming this similarity. The layout also positioned group 1 farther from the other groups. When comparing images from this group (example in Figure 43), one can see that the book is almost totally occult in the frames from this group, which does not happen in frames from the other groups, justifying this positioning. Such analysis/patterns has an important influence of the descriptor thought, and to employ other descriptor would contribute to all these similar instances to compose a single group.

The group T1 represents a transition, from the moment in which the office is empty, until the one in which the actor reads a book (actions illustrated by the groups from 1 to 4). In this transition, the book is manipulated to enable the reading action. Its instances show the moment in which the actor rises the book to starts open it (see Figure 43e). The moment before T1 is represented by instances from group B, expanded in Figure 44a. This moment was produced in four subgroups on projection, enhancing actions such as (44b) starting to pick the book, (44c) the book is closer of actor's body, (44d) stationary moment (no significant action occurs) and (44e) the book is lifted until eyes level.

Instances from group T2 represents transitory moments, between groups 1-2, 2-3, and 3-4. Although these frames do not present high variations in relation to those from analyzed groups, they show moments which actor's body is more erect and the book occupies a higher position (see Figure 43f).

A vert./horiz. analysis on TSSM layout (see Figure 42b) suggests an event starting after frame 550, and ending near frame 2000, approximately, exactly when the actor enters/leaves the office. The fading effect on the region borders suggests that the action does not occur abruptly, that is, the scene gradually changes over consecutive frames indicating a smooth transition. It reflects how the actor executes the actions that compose the main event (slowly). By checking the TSSM main diagonal region, one notice the occurrence of several minor events represented by square regions. These regions are numbered in the TSSM layout, corresponding to the numbered groups in the t-SNE layout.

Analyzing the group 4 locally by selecting instances from this region and building the correspondent TSSM (see Figure 45), one can notice that relationship among original space and projection space was kept. The most dissimilar moment (approximately 190) of the TSSM, presented in Figure 45a, represents a visual variation of actor's body inclination and book positioning, also captured by projection, due to the division of group 4 in Figure

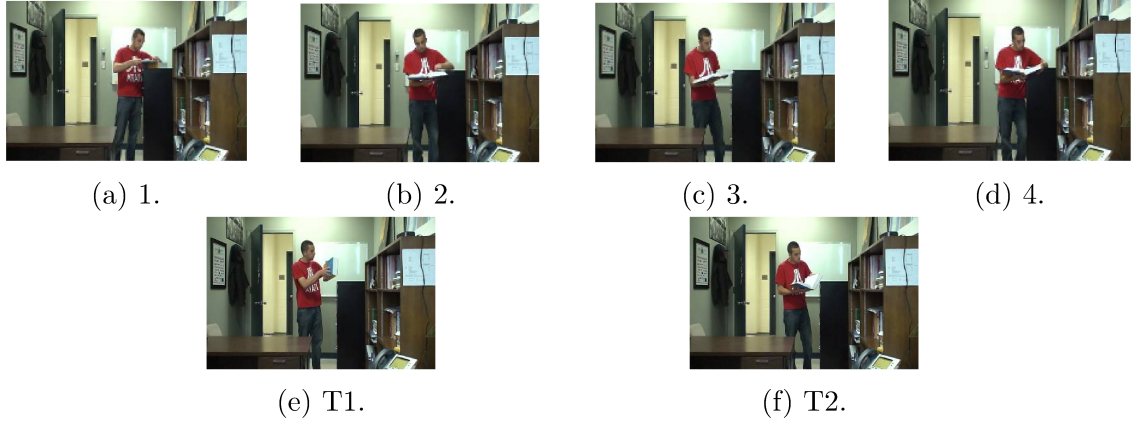


Figure 43 – Images representing each group indicated in Fig 42.

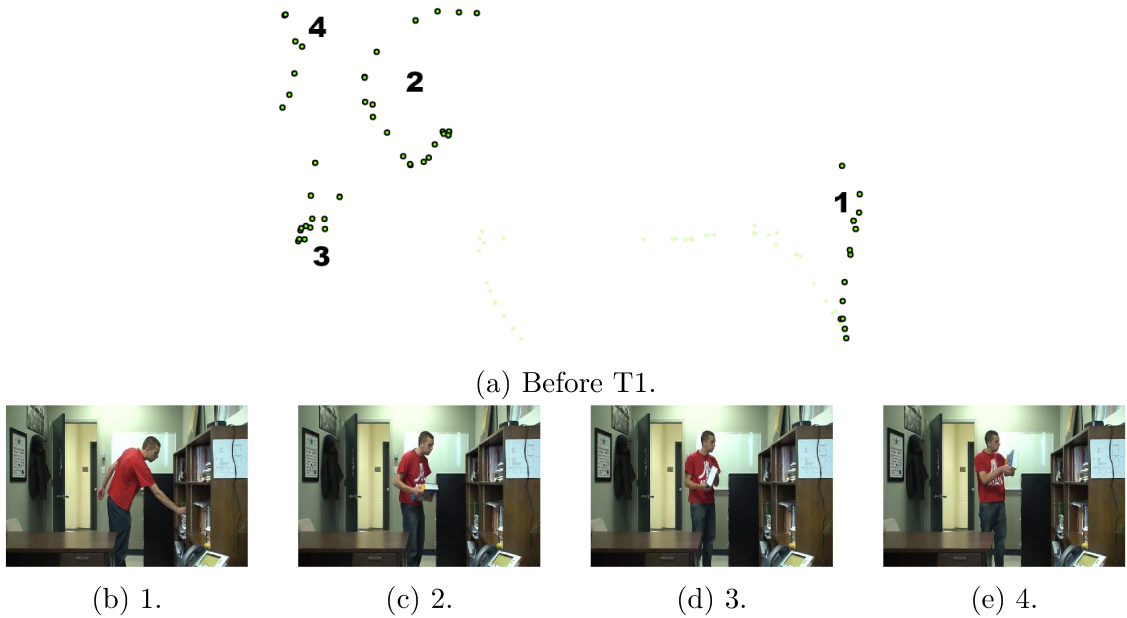


Figure 44 – Sub actions previously the actor starts to read.

49 into two regions. The biggest group, showed in Figure 45b, represents frames in which the actor is reading leaned on a furniture piece and the book is semi-opened, and the smallest, those which he is flipping through the book pages with his body erect and the book fully opened (see Figure 45c).

Figure 46 shows the point-placement layout of two moments representing the actor entering/leaving the office. The actions from each moment are similar to each other, in terms of frames content, except for the direction movement and actor side (front/back). The layout was able to differentiate these action, but at the same time positioned them in close regions, reflecting their similarity.

The shape of the groups in the layout also revealed interesting information about the video content. Different types of events produced groups with distinct shapes. Events representing stationary actions were usually represented by round-shaped groups, such as groups 2, 3, and 4 in Figure 42a. The group 1 however presents a peculiar shape. At this

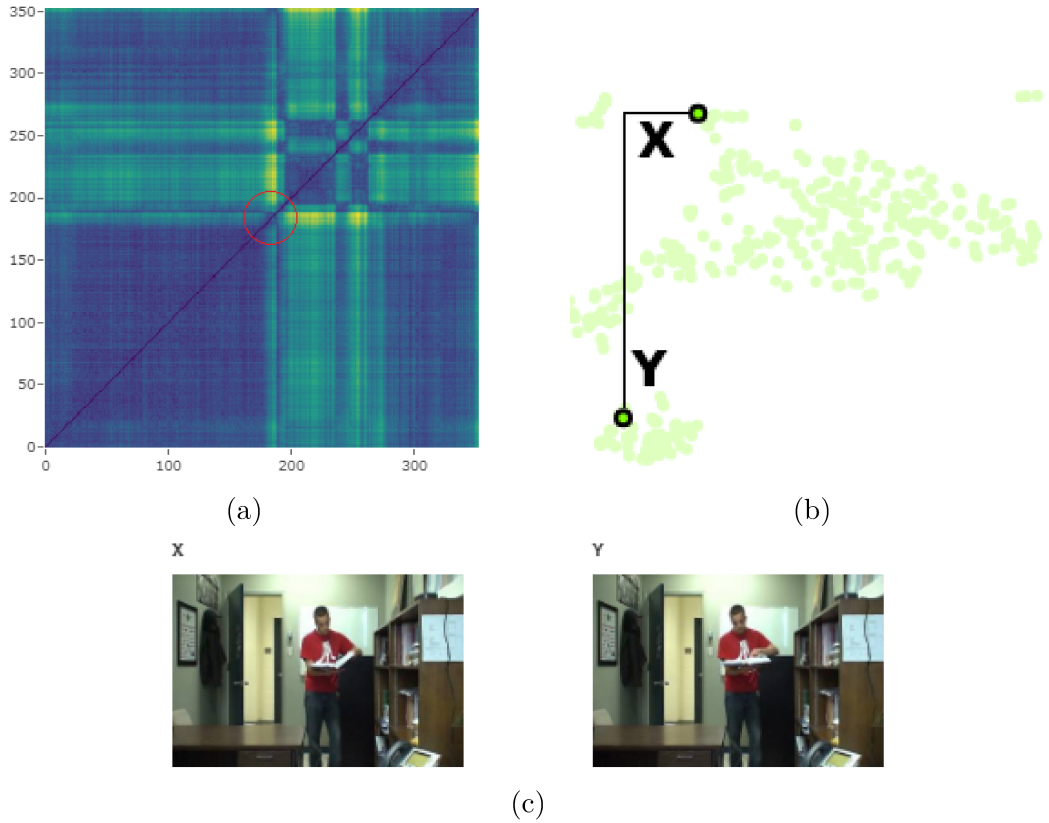


Figure 45 – Local analyze of group 4. **(45a)** TSSM of group 4 instances. In red circle the origin region of the points $X - Y$. **(45b)** Projection of group 4, highlighting instances correspondent at the higher dissimilar moments of 45a TSSM. **(45c)** Frames of highlighted instances in Figure 45b.

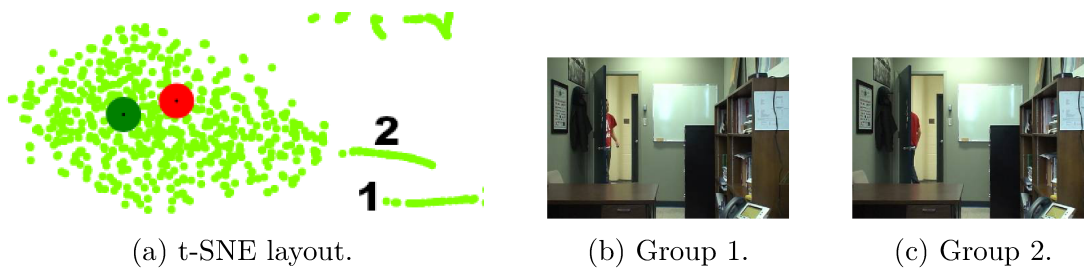


Figure 46 – Analysis of two groups representing the actor entering/leaving the office in OFFICE video.

moment, the actor is flipping through the book pages, and thus constantly moving his arms, causing more disturbance on the frames in this group than in the ones representing other reading moments. To represent transitions between consecutive events and/or flow movements, the layout produced string-shaped groups, probably due to the frames present slightly gradual differences among each other. Figure 47 shows an example of two groups with this shape, representing moments in which the actor picks/returns the book to the shelf, representing transitions between entering-reading and reading-leaving actions, respectively. The same pattern can be noticed in the situation presented in Figure 46, in which the actor is moving from one place to another.

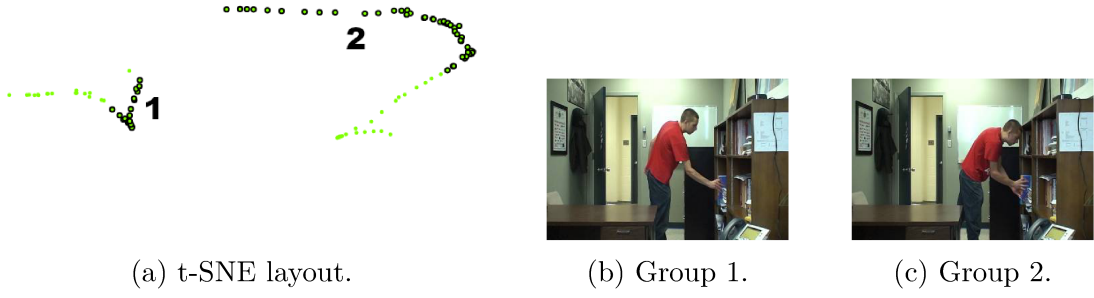


Figure 47 – Analysis of two group representing the actor picking/returning the book in OFFICE video.

Figure 48 indicates the moment 1870 – 1890 of the Office collection. One may notice that, even short, instances from this period were positioned far from each other, which indicates a high variation in their content. An imaginary triangle formed by the area that links this three small includes instances from group 2 in Figure 47a, which indicates the action of returning the book, which generated an abrupt change in the video description structure. Not coincidentally, the higher global dissimilarity in TSSM covers this same period, between 1820 – 1890. This event is probably related to two specific variations on the scene: the actor’s body inclination, which is more significant, and the print on the actor’s shirt, which does not appear.

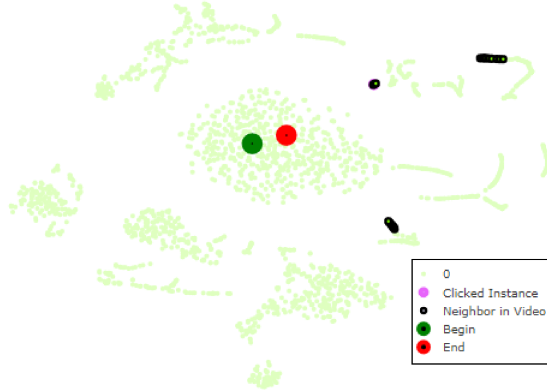
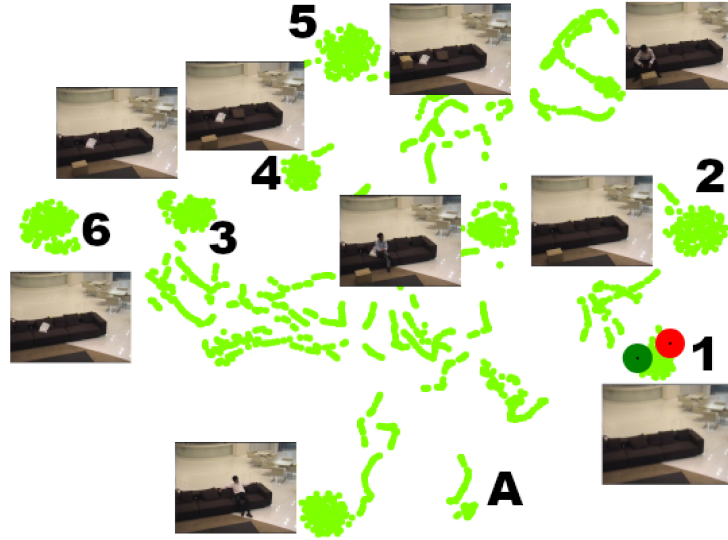


Figure 48 – t-SNE projection of Office video. Moment 1870-1890 highlighted.

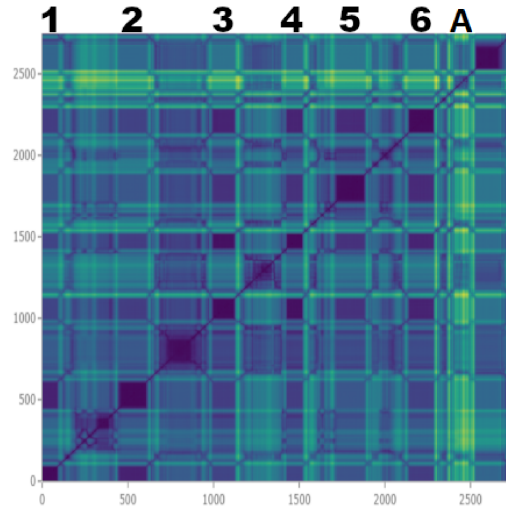
5.2.2 SOFA Video

SOFA video is mainly characterized by a variety of moments representing people entering/leaving the scene, and several objects being left/carried, producing a scenario in which several events shall be noticed. Figure 49 illustrates the point-placement layout (Figure 49a) and respective TSSM (Figure 49b) layouts for this video.

The layout is composed by some round-shaped groups representing stationary actions and by a significant number of string-shaped groups representing transitions between events, as observed in the OFFICE video, and as expected for this video. The first/last



(a) t-SNE projection.



(b) TSSM.

Figure 49 – SOFA video layouts.

frames are also positioned in the same group, suggesting in this case that no object is left in the scene at the end of the video, and no permanent modification occurs. Each round-shaped group represents a moment in which a distinct object or multiple objects are left in the scene. The massive occurrence of stationary moments may suggest cycle events with small dissimilarity. According to the group placement, in terms of distance between each other, the analyst may interpret how similar they are and focus in those of his/her interest. Representative frames from each of these moments are shown in Figure 49a. As in this case almost no change between consecutive frames is observed in each of these events, the produced groups are denser than those observed in OFFICE video. In this case, the layout may be useful in situations in which a security agent must investigate potential dangerous objects left in a specific place. Finally, in some of the round-shaped groups (marked as 1, 2 and 3 in Figure 49a), one observes some frames “escaping” the

group, representing the moments in which the actors enter/leave the scene.

TSSM layout patterns also reflect the observed video events. By analyzing the map diagonal, one can see several small square regions, suggesting a high number of small duration events that do not disturb the scene significantly. Most of these regions present low color intensity, corresponding to events with non-significant changes between frames. These regions represent stationary action events, in this case moments in which the objects are left alone in the scene. The size of these regions is proportional to the duration of these events. Some of these regions present a fading effect, representing moments in which people start/end to leave/pick objects. There is also a significant change between frames 2400 and 2500 that represents the only moment in which two people are in the scene. This moment is represented in the point-placement layout as an isolated string-shaped group at the bottom of the layout (region A in Figure 49a), reflecting the distinct movement pattern associated. Finally, vert./horiz. analysis allows to identify when there are no objects in the scene, when they are moving (people), and when objects are left alone. It can be done analyzing the TSSM at the initial moment ($0 - 85$), in which no action occurs, and the scene is empty. Thus, regions in the vert./horiz. with low intensity color tend to contain none object, or abandoned objects. On the other hand, the high intensity color region tends to represent movements, once it represents the most dissimilar moments related to the initial moment. By the color intensity it is also possible to identify some basic objects properties such as size, because these properties directly impact on frame changes with respect to the reference frame (selected map row/column).

Similarity between stationary groups can be measured by comparing the intersection of its correspondent moments in TSSM, which in some scenarios may be a determinant analyze guidance. Looking the TSSM of Figure 50 it is possible to notice the similarity between groups 1-2 (white square) and 3-4-5-6 (red squares). The point-placement layout presents the same information, since this similar groups were placed close (see Figure 49a). This characteristic makes information searching faster, once the specialist is promptly to analyze the layout focusing in regions instead of groups, and after detecting a region of interest a local exploration may be performed. Optionally, this analysis can be using both TSSM and the projection in a complementary manner. Some non totally homogeneous structure observed in TSSM indicated by red asterisk in Figure 50 represent moments in which standing objects perform smooth actions in its own base—people sitting on sofa moving their limbs. Thus, TSSM is suitable for several scenarios analysis, whether to capture high or smooth moving level, as well as stationary moments.

5.2.3 VIRAT

In this video, the main event is characterized by a person approaching a car in a parking lot, getting in, and driving this car out of the scene. Figure 51 illustrates the point-placement layout (Figure 51a) and respective TSSM (Figure 51b) layouts for this

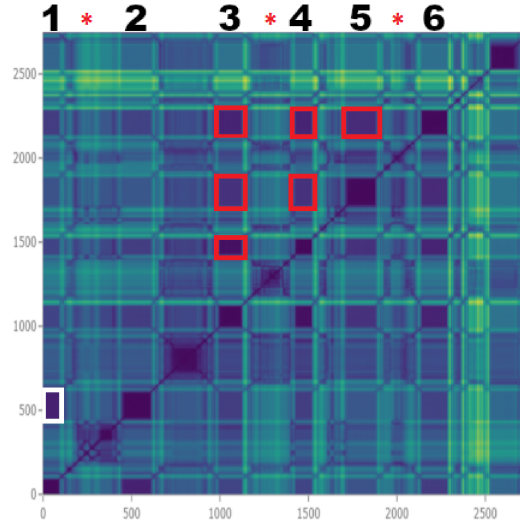


Figure 50 – TSSM with the intersection between similar stationary moments highlighted.

video. We manually labeled this video to represent moments before, during and after the main event, and mapped the respective frames in the layout to colors green, blue and yellow, respectively.

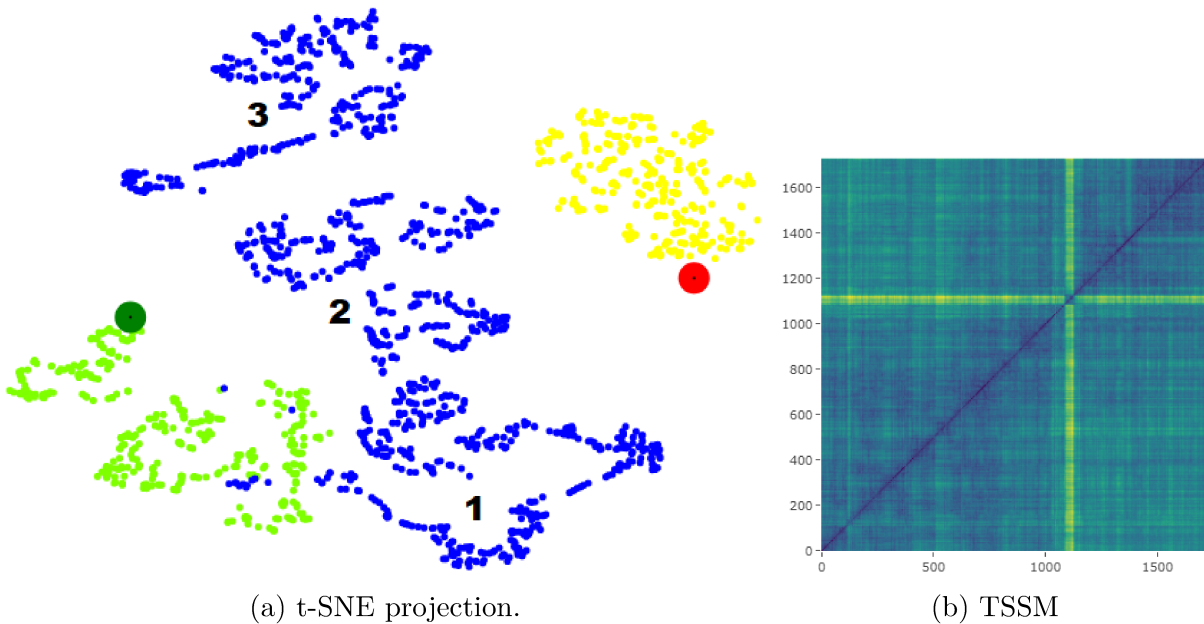


Figure 51 – VIRAT video layouts. In 51a, colors represent moments before (green), during (blue) and after (yellow) the main event, and numbers represent three distinct moments composing the main event: the actor approaching the car (1), getting into the car (2), and driving the car out of the scene (3).

The point-placement layout is consistent with the manual labeling, producing mainly three separated heterogeneous groups. The first/last frames are significantly apart, indicating a permanent change in the scene (the absence of the car) when the video ends. The heterogeneity of the group representing moments before the event (green points) can

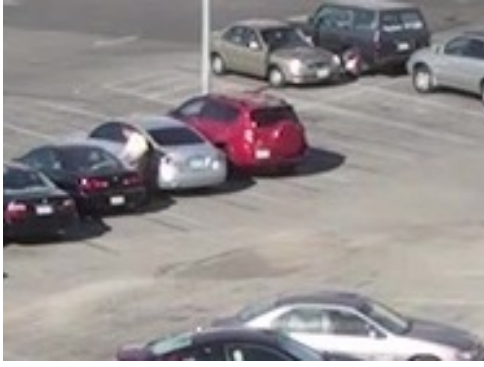
be explained by the high concentration of people walking in the scene, including the actor when he is far from the car. HOG descriptor tends to be sensitive with these movements, resulting in significant differences among the frames. Moreover, some of these movements produce gradual changes in consecutive frames, representing people moving in specific directions and producing string-shaped formations in the layout.

The group representing the main event can be divided into three smaller groups. By analyzing the frames in each of these groups, one notice that they represent moments in which the actor approaches the car (1), gets into it (2) and drives the car out of the scene (3). The layout regions corresponding to each of these moments are numbered in 51a and representative frames from each region are shown in Figure 52. Group 1 produced string-shaped formations that possibly reflects the actor walking towards the car (and approaching it), together with some people walking through the parking lot. Group 2 can be split into two smaller subgroups. The first subgroup represents the moment in which the actor actually reaches the car, which explains the proximity with group 1. The second subgroup mostly concentrates frames in which the actor puts a bag in the right side of the car, walks to the left side and enters the car. Group 3 represents the moment in which the car performs maneuvers and moves out of the scene. During all the actions that compose this event, the presence of people walking through the parking lot produced several string shape formations in the layout. Finally, there is only one group representing the moment after the event, in which the car is completely leaving the scene. At this moment, the absence of the car in the parking lot and the lower concentration of people walking in the scene produced a better defined group in the layout.

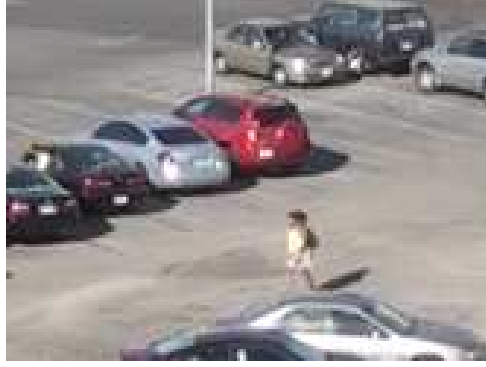
The TSSM layout also reflects the permanent scene modification. When performing a vert./horiz. analysis, one can notice higher color intensity cells after frame 1000 indicating when the permanent changes occur. There is a high color intensity region between frames 1100 and 1200 which represents the moment when the car starts to perform maneuvers to leave the parking lot, producing significant frame changes. In this layout, no low color intensity regions were noticed, suggesting that no stationary actions occurred, which is expected in this scenario. There are also no isolated high color intensity points, indicating that no abrupt changes occurred. The lower color intensity region in the top right part of the map (after frame 1400) reflects the moment with low concentration of people walking in the scene also revealed by the t-SNE layout (yellow region in Figure 51a).

5.2.4 Smart Sampling results

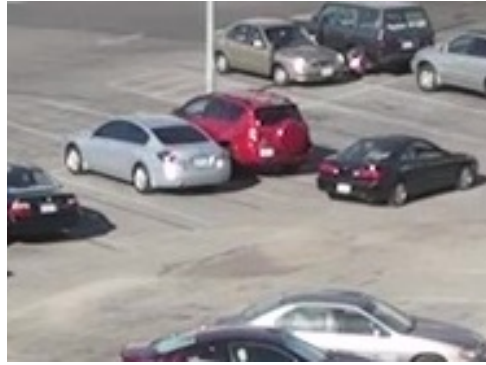
Our smart sampling procedure was applied to the three videos with different sampling rates in order to verify if it is capable to maintain the events occurrence structure. We first generated compact versions of each video containing 50%, 25% and 12% of the frames, using an uniform sampling procedure and using our strategy considering HIGH, LOW and BOTH movement variation. We then generated the t-SNE layouts for each new



(a) Actor approaches the car.



(b) Actor gets into the car.



(c) The car leaves the scene.

Figure 52 – Main event in VIRAT video.

video and compared the results, which are shown in Table 2. The idea is to evaluate how each sampling procedure and rate maintain or emphasize specific aspects from the events structure observed in the previous analysis.

For all videos, our smart sampling procedure reflects the movement variation strategy chosen, filtering the groups accordingly. In OFFICE video, when considering high movement variation, one clearly see that the groups representing transitions or gradual but significant actions (string-shaped groups) are the only ones maintained. When considering low movement variation, only the four numbered groups and the group in the center of the layout shown in Figure 42a are maintained as the sampling proportion values decrease.

In SOFA video, the stationary actions representing moments in which the objects are left alone in the scene are almost not distinguishable in 50% and 25% sampling proportions, and not displayed in 12%, in which the layout practically shows only the transition movements. Using uniform 12% sampling in this video resulted in a layout in which several groups are mixed with each other and significant patterns are lost. On the other hand, when performing sampling considering low movement variation, these stationary actions are maintained in all sampling proportions, favoring for an analysis focused on investigating these objects.

In VIRAT video, the smart sampling procedure did not produced significant difference

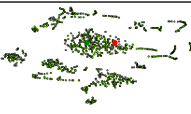
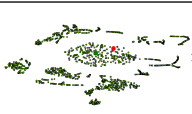
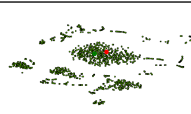

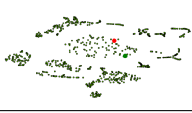
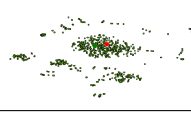
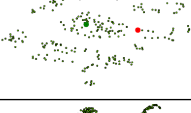
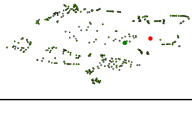
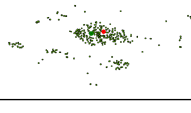


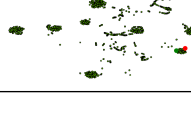


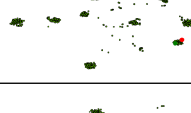
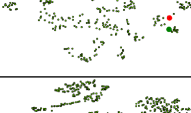
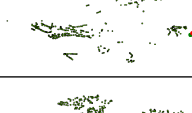
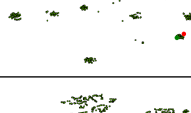


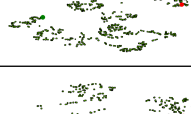


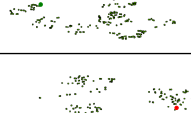

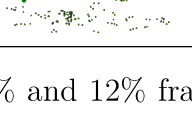

		UNIFORM	HIGH	LOW
OFFICE	50%			
	25%			
	12%			
SOFA	50%			
	25%			
	12%			
VIRAT	50%			
	25%			
	12%			

Table 2 – t-SNE layouts considering 50%, 25% and 12% frame sampling from OFFICE, SOFA and VIRAT videos, comparing uniform sampling against our smart sampling procedure (HIGH movement variation and LOW movement variation).

for all considered sampling proportion values. As shown in the associated TSSM layout (Figure 51b), the actions in this video are more equally distributed over time, and few abrupt changes occur. In such scenario an uniform sampling is likely to capture all important events, explaining the similarity among all results. However, our proposed sampling strategy maintained the main group formation, as well as transition patterns in all considered sampling proportion values, but several small patterns were lost when performing an 12% uniform sampling.

5.3 Discussion

This section presents the interpretation of the obtained results and draw conclusions about the findings, formulating a guideline to interpret the layouts. The spatial frame organization produced by the point-placement layout suggests analysis strategies potentially

appropriate for each specific video content properties. In addition, by grouping frames from similar events occurred in non adjacent time periods it naturally creates a separation between “normal” situations, in which nothing happens, from situations in which events occur. It can be observed, for example, in Figure 42a. The layout also allows to identify permanent changes in the scene by grouping or separating the first and last frames, as shown in Figure 51a. Finally, the layout provides a multi-resolution analysis in which, besides the identification of the main events in the video, it also reveals sub-events that compose these main events, providing additional details about how they occurred. The groups shapes provide interesting insights about the type of event. Round-shaped dense groups are related to stationary action events, as shown in Section 5.2.2, and presented in Figure 49a. Any movement disturbing the action pattern affects this shape/density. String-shaped groups however are related to flow/trajectory movements or transitions between events, such as people walking, cars moving slowly, among others, as shown in Section 5.2.1, and presented in Figure 46a. In this scenario, the employed image description scheme demonstrated to be significantly sensitive to subtle movements in the scene, which led to the formation of groups representing small actions on the scene.

TSSM layouts also provides interesting patterns to temporally explain the occurrence of events in a video. It is easy to identify abrupt and gradual changes (as detailed in the experiment of Section 5.2.1, Figure 42b), as well as when they occurred, by checking the cells color intensity. By performing a vert./horiz. analysis, one can determine a reference frame from which it is possible to identify permanent changes in the scene and when these changes occur, as shown in Section 5.2.3, Figure 51b. The verification of square regions in the map provides details about the events. The area of these regions is related to the event duration, the color intensity is related to how stationary the event is or the amount of action it contains and the color intensity homogeneity is related to the event actions regularity. The map diagonal analysis also provides a natural temporal video segmentation, in which each segment is represent by an event, as shown the TSSM analysis for the three videos in Section 5.2.

5.3.1 Limitations

Some limitations of our methodology were identified when performing the experiments, most of them related to our approach scalability when analyzing long duration videos. The TSSM process generation is costly and may be impracticable for some large videos, or real time monitoring scenarios, with constant video capturing.

One way to address this issue is using a frame sampling procedure prior to the analysis process. Several smart sampling procedures exist on the literature, and they can be employed to select frames that highlight any specific video aspect that can contribute to the events detection. In this work, we proposed a smart sampling (Section 4.3.5), that already showed interesting results, as presented in Section 5.2.4.

Finally, a sliding window may also be employed to address this issue. In this strategy, temporal subsequent portions of the video may be analyzed individually, in order to get insights of specific moments on the video. Summary schemes can then be used to relate each subsequent window, in order to provide an overall video layout.

Depending on the duration of events, they can be mapped to very small portions of the layout, and go unnoticed by the user. However, it only occurs for very fast events composed by a small number of frames, not representing a common situation in surveillance scenarios.

Conclusion

In this work, we presented a surveillance video visual analysis methodology employing Information Visualization techniques, specially similarity-based strategies, for comprehension of the video structure in terms of occurrence of events. Our methodology employs point-placement and Temporal Self-Similarity Maps (TSSM) in a coordinated manner, and a set of interaction tools are provided to allow the exploration of spatial and temporal aspects related to the occurrence of events. We also presented a system that implements the proposed methodology. Several experiments were conducted involving different surveillance scenarios, in order to validate the proposal and identify its advantages and limitations.

Our experiments demonstrated that the coordination of these techniques creates complementary spatial and temporal metaphors that facilitates the video content comprehension, as well as the identification of events and how they relate to each other. The point-placement similarity organization provides an overview of the video structure in terms of events, creating a natural summary. It also provides a visual representation of each event, mapped into layout groups, based on how actions are organized into them. The groups shapes are also related to specific types of events, such as gradual movements to string shaped groups, or stationary actions to round shaped groups, assisting users to define adequate analysis strategies. It also highlights relationships between events that are represented by relationships among the groups in the layouts, creating, for example, a separation between normal situations and situations that represent strategic events.

TSSM, on the other hand, produces a layout whose temporal perspective provides a natural event summarization and also support the user in comprehending when each identified event in point-placement layout occurs. This layout also provides an easy identification of the duration of each event, abrupt changes in the video content, permanent changes, as well as the comprehension of the similarity among events and their distribution through the video.

We identified an limitation in our methodology, related to scalability of the the adopted techniques when applied to large surveillance videos, as detailed in Section 5.3.1. As fu-

ture work we intent to investigate strategies to address these limitations and thus apply our methodology to large videos involving scenarios with diverse people/objects movement patterns, eventually combining automatic strategies employing Machine Learning strategies to enhance its scalability. Future work related to this research include:

- ❑ Perform experiments with users (security agents) in order to collect feedback about the process and refine our system, as well as to evaluate their ability to better understand the video content and events of interest;
- ❑ Implement new visualization/interaction techniques to provide additional perspectives about the data, as well as to improve the analytical capabilities of our implemented methodology;
- ❑ Apply our methodology to other types of surveillance videos, involving scenarios with non-static cameras, or multiple cameras.

The results of this research were published published in the **23 International Conference Information Visualisation** (MENDES; PAIVA; SCHWARTZ, 2019).

Bibliography

ANDRIENKO, G. et al. **Visual analytics of movement**. [S.l.]: Springer Science & Business Media, 2013. <<https://doi.org/10.1007/978-3-642-37583-5>>.

AVILA, S. E. F. D. et al. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. **Pattern Recognition Letters**, Elsevier, v. 32, n. 1, p. 56–68, 2011. <<https://doi.org/10.1016/j.patrec.2010.08.004>>.

BAGHERI, S.; ZHENG, J. Y. Temporal mapping of surveillance video. In: IEEE. **Pattern Recognition (ICPR), 2014 22nd International Conference on**. [S.l.], 2014. p. 4128–4133.

BARBU, T. Pedestrian detection and tracking using temporal differencing and hog features. **Computers & Electrical Engineering**, Elsevier, v. 40, n. 4, p. 1072–1079, 2014. <<https://doi.org/10.1016/j.compeleceng.2013.12.004>>.

BASHIR, F.; PORIKLI, F. Performance evaluation of object detection and tracking systems. In: **Proceedings 9th IEEE International Workshop on PETS**. [S.l.: s.n.], 2006. p. 7–14. <https://doi.org/10.1007/11612704_16>.

BAYONA, Á.; SANMIGUEL, J. C.; MARTÍNEZ, J. M. Comparative evaluation of stationary foreground object detection algorithms based on background subtraction techniques. In: IEEE. **2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance**. [S.l.], 2009. p. 25–30. <<https://doi.org/10.1109/AVSS.2009.35>>.

BENFOLD, B.; REID, I. Stable multi-target tracking in real-time surveillance video. In: IEEE. **Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on**. [S.l.], 2011. p. 3457–3464. <<https://doi.org/10.1109/CVPR.2011.5995667>>.

BIRD, N. D. et al. Detection of loitering individuals in public transportation areas. **IEEE Transactions on intelligent transportation systems**, IEEE, v. 6, n. 2, p. 167–177, 2005. <<https://doi.org/10.1109/TITS.2005.848370>>.

BORGO, R. et al. State of the art report on video-based graphics and video visualization. In: WILEY ONLINE LIBRARY. **Computer Graphics Forum**. [S.l.], 2012. v. 31, n. 8, p. 2450–2477. <<https://doi.org/10.1111/j.1467-8659.2012.03158.x>>.

BOTCHEN, R. P. et al. Action-based multifield video visualization. **IEEE transactions on visualization and computer graphics**, IEEE, v. 14, n. 4, p. 885–899, 2008. <<https://doi.org/10.1109/TVCG.2008.40>>.

CAETANO, C.; SANTOS, J. A. dos; SCHWARTZ, W. R. Optical flow co-occurrence matrices: A novel spatiotemporal feature descriptor. In: IEEE. **Pattern Recognition (ICPR), 2016 23rd International Conference on**. [S.l.], 2016. p. 1947–1952. <<https://doi.org/10.1109/ICPR.2016.7899921>>.

CAO, X. et al. Linear svm classification using boosting hog features for vehicle detection in low-altitude airborne videos. In: IEEE. **Image Processing (ICIP), 2011 18th IEEE International Conference on**. [S.l.], 2011. p. 2421–2424. <<https://doi.org/10.1109/ICIP.2011.6116132>>.

CARD, S. K.; MACKINLAY, J. D.; SHNEIDERMAN, B. **Readings in information visualization: using vision to think**. [S.l.]: Morgan Kaufmann, 1999.

CHIU, P.; GIRGENSOHN, A.; LIU, Q. Stained-glass visualization for highly condensed video summaries. In: IEEE. **Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on**. [S.l.], 2004. v. 3, p. 2059–2062. <<https://doi.org/10.1109/ICME.2004.1394670>>.

CHU, W.-S.; SONG, Y.; JAIMES, A. Video co-summarization: Video summarization by visual co-occurrence. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2015. p. 3584–3592. <<https://doi.org/10.1109/CVPR.2015.7298981>>.

COLQUE, R. M. et al. Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. **IEEE Transactions on Circuits and Systems for Video Technology**, 2017. <<https://doi.org/10.1109/TCSVT.2016.2637778>>.

COLQUE, R. V. H. M. et al. Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. **IEEE Transactions on Circuits and Systems for Video Technology**, IEEE, v. 27, n. 3, p. 673–682, 2017. <<https://doi.org/10.1109/TCSVT.2016.2637778>>.

COOPER, M.; FOOTE, J. Scene boundary detection via video self-similarity analysis. In: IEEE. **Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)**. [S.l.], 2001. v. 3, p. 378–381. <<http://10.1109/ICIP.2001.958130>>.

COX, T. F.; COX, M. A. **Multidimensional scaling**. [S.l.]: CRC press, 2000.

DANIEL, G.; CHEN, M. Video visualization. In: IEEE COMPUTER SOCIETY. **Proceedings of the 14th IEEE Visualization 2003 (VIS'03)**. [S.l.], 2003. p. 54.

DATONDJI, S. R. E. et al. A survey of vision-based traffic monitoring of road intersections. **IEEE transactions on intelligent transportation systems**, IEEE, v. 17, n. 10, p. 2681–2698, 2016. <<https://doi.org/10.1109/TITS.2016.2530146>>.

DOGRA, D. P.; AHMED, A.; BHASKAR, H. Smart video summarization using mealy machine-based trajectory modelling for surveillance applications. **Multimedia Tools and Applications**, Springer, v. 75, n. 11, p. 6373–6401, 2016.

EADES, P. A heuristic for graph drawing. **Congressus numerantium**, v. 42, p. 149–160, 1984.

EFROS, A. A. et al. Recognizing action at a distance. In: IEEE. **null**. [S.l.], 2003. p. 726. <<https://doi.org/10.1109/ICCV.2003.1238420>>.

EL-ETRIBY, S.; ELMEZAIN, M.; MIRAOU, M. A novel approach for crowd behavior representation: Normal and abnormal event detection. **Journal of Theoretical & Applied Information Technology**, v. 96, n. 11, 2018.

FAN, C.-T.; WANG, Y.-K.; HUANG, C.-R. Heterogeneous information fusion and visualization for a large-scale intelligent video surveillance system. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, IEEE, v. 47, n. 4, p. 593–604, 2017.

FELS, S.; MASE, K. Interactive video cubism. In: ACM. **Proceedings of the 1999 workshop on new paradigms in information visualization and manipulation in conjunction with the eighth ACM international conference on Information and knowledge management**. [S.l.], 1999. p. 78–82. <<https://doi.org/10.1145/331770.331789>>.

FERREIRA, N. et al. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 19, n. 12, p. 2149–2158, 2013. <<https://doi.org/10.1109/TVCG.2013.226>>.

GOLDMAN, D. B. et al. Schematic storyboarding for video visualization and editing. In: ACM. **ACM Transactions on Graphics (TOG)**. [S.l.], 2006. v. 25, n. 3, p. 862–871. <<https://doi.org/10.1145/1141911.1141967>>.

GONG, S.; LOY, C. C.; XIANG, T. Security and surveillance. In: **Visual Analysis of Humans**. [S.l.]: Springer, 2011. p. 455–472. <https://doi.org/10.1007/978-0-85729-997-0_23>.

GOYETTE, N. et al. Changedetection. net: A new change detection benchmark dataset. In: **CVPR Workshops**. [S.l.: s.n.], 2012. p. 1–8. <<https://doi.org/10.1109/CVPRW.2012.6238919>>.

GUO, G.-D. et al. Learning similarity measure for natural image retrieval with relevance feedback. **Neural Networks, IEEE Transactions on**, IEEE, v. 13, n. 4, p. 811–820, 2002. <<https://doi.org/10.1109/TNN.2002.1021882>>.

HAHNEL, M.; KLUNDER, D.; KRAISS, K.-F. Color and texture features for person recognition. In: IEEE. **Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on**. [S.l.], 2004. v. 1, p. 647–652. <<https://doi.org/10.1109/IJCNN.2004.1379993>>.

HAMPAPUR, A. et al. Smart surveillance: applications, technologies and implications. In: IEEE. **Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on**. [S.l.], 2003. v. 2, p. 1133–1138.

- HINTON, G. E.; ROWEIS, S. T. Stochastic neighbor embedding. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2003. p. 857–864.
- HÖFERLIN, M. et al. Interactive schematic summaries for exploration of surveillance video. In: ACM. **Proceedings of the 1st ACM International Conference on Multimedia Retrieval**. [S.l.], 2011. p. 9. <<https://doi.org/10.1145/1991996.1992005>>.
- HU, W. et al. A survey on visual surveillance of object motion and behaviors. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, IEEE, v. 34, n. 3, p. 334–352, 2004. <<https://doi.org/10.1109/TSMCC.2004.829274>>.
- HU, X. et al. Video anomaly detection using deep incremental slow feature analysis network. **IET Computer Vision**, IET, v. 10, n. 4, p. 258–267, 2016. <<https://doi.org/10.1049/iet-cvi.2015.0271>>.
- HUANG, C.-H.; WU, Y.-T.; SHIH, M.-Y. Unsupervised pedestrian re-identification for loitering detection. In: SPRINGER. **Pacific-Rim Symposium on Image and Video Technology**. [S.l.], 2009. p. 771–783. <https://doi.org/10.1007/978-3-540-92957-4_67>.
- HUANG, C.-R. et al. Maximum a posteriori probability estimation for online surveillance video synopsis. **IEEE Transactions on circuits and systems for video technology**, IEEE, v. 24, n. 8, p. 1417–1429, 2014.
- JIANG, F. et al. Anomalous video event detection using spatiotemporal context. **Computer Vision and Image Understanding**, Elsevier, v. 115, n. 3, p. 323–333, 2011. <<https://doi.org/10.1016/j.cviu.2010.10.008>>.
- JOIA, P. et al. Local affine multidimensional projection. **Visualization and Computer Graphics, IEEE Transactions on**, IEEE, v. 17, n. 12, p. 2563–2571, 2011. <<https://doi.org/10.1109/TVCG.2011.220>>.
- JOLLIFFE, I. T. Principal component analysis and factor analysis. **Principal component analysis**, Springer, p. 150–166, 2002. <<https://doi.org/10.1007/b98835>>.
- JR, A. C. N.; SCHWARTZ, W. R. A scalable and flexible framework for smart video surveillance. **Computer Vision and Image Understanding**, Elsevier, v. 144, p. 258–275, 2016. <<https://doi.org/10.1016/j.cviu.2015.10.014>>.
- JUNEJO, I. N. et al. View-independent action recognition from temporal self-similarities. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 33, n. 1, p. 172–185, 2011. <<https://doi.org/10.1109/TPAMI.2010.68>>.
- KE, Y.; SUKTHANKAR, R.; HEBERT, M. Event detection in crowded videos. In: IEEE. **Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on**. [S.l.], 2007. p. 1–8. <<https://doi.org/10.1109/ICCV.2007.4409011>>.
- KHOSLA, A. et al. Large-scale video summarization using web-image priors. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2013. p. 2698–2705. <<https://doi.org/10.1109/CVPR.2013.348>>.

KIM, I. S. et al. Intelligent visual surveillance—a survey. **International Journal of Control, Automation and Systems**, Springer, v. 8, n. 5, p. 926–939, 2010. <<https://doi.org/10.1007/s12555-010-0501-4>>.

KO, T. A survey on behavior analysis in video surveillance for homeland security applications. In: IEEE. **Applied Imagery Pattern Recognition Workshop, 2008. AIPR'08. 37th IEEE**. [S.l.], 2008. p. 1–8. <<https://doi.org/10.1109/AIPR.2008.4906450>>.

KRISHNA, M. V. et al. Temporal video segmentation by event detection: A novelty detection approach. **Pattern recognition and image analysis**, Springer, v. 24, n. 2, p. 243–255, 2014. <<https://doi.org/10.1134/S1054661814020114>>.

_____. Temporal video segmentation by event detection: A novelty detection approach. **Pattern recognition and image analysis**, Springer, v. 24, n. 2, p. 243–255, 2014. <<https://doi.org/10.1134/S1054661814020114>>.

LANDESBERGER, T. V. et al. Mobilitygraphs: Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering. **IEEE transactions on visualization and computer graphics**, IEEE, v. 22, n. 1, p. 11–20, 2016. <<https://doi.org/10.1109/TVCG.2015.2468111>>.

LIAO, H. et al. Visualization-based active learning for video annotation. **IEEE Transactions on Multimedia**, IEEE, v. 18, n. 11, p. 2196–2205, 2016. <<https://doi.org/10.1109/TMM.2016.2614227>>.

LIU, Y. et al. A survey of content-based image retrieval with high-level semantics. **Pattern Recognition**, Elsevier, v. 40, n. 1, p. 262–282, 2007.

LU, Y. et al. Application of an incremental svm algorithm for on-line human recognition from video surveillance using texture and color features. **Neurocomputing**, Elsevier, v. 126, p. 132–140, 2014. <<https://doi.org/10.1016/j.neucom.2012.08.071>>.

LU, Z.; GRAUMAN, K. Story-driven summarization for egocentric video. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2013. p. 2714–2721. <<https://doi.org/10.1109/CVPR.2013.350>>.

MA, J.; DAI, Y.; HIROTA, K. A survey of video-based crowd anomaly detection in dense scenes. **Journal of Advanced Computational Intelligence and Intelligent Informatics**, Fuji Technology Press Ltd., v. 21, n. 2, p. 235–246, 2017. <<https://doi.org/10.20965/jaciii.2017.p0235>>.

MAATEN, L. v. d.; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, n. Nov, p. 2579–2605, 2008.

_____. Visualizing data using t-sne. **Journal of machine learning research**, v. 9, n. Nov, p. 2579–2605, 2008.

MAHASSENI, B.; LAM, M.; TODOROVIC, S. Unsupervised video summarization with adversarial lstm networks. In: IEEE. **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.], 2017. p. 2982–2991. <<https://doi.org/10.1109/CVPR.2017.318>>.

MAHMOUD, K. M.; ISMAIL, M. A.; GHANEM, N. M. Vscan: an enhanced video summarization using density-based spatial clustering. In: SPRINGER. **International conference on image analysis and processing**. [S.l.], 2013. p. 733–742. <https://doi.org/10.1007/978-3-642-41181-6_74>.

MEGHDAI, A. H.; IRANI, P. Interactive exploration of surveillance video through action shot summarization and trajectory visualization. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 19, n. 12, p. 2119–2128, 2013.

MENDES, G.; PAIVA, J. G. S.; SCHWARTZ, W. R. Point-placement techniques and temporal self-similarity maps for visual analysis of surveillance videos. In: IEEE. **2019 23rd International Conference Information Visualisation (IV)**. [S.l.], 2019. p. 127–132. <<https://doi.org/10.1109/IV.2019.00030>>.

MENDI, E.; CLEMENTE, H. B.; BAYRAK, C. Sports video summarization based on motion analysis. **Computers & Electrical Engineering**, Elsevier, v. 39, n. 3, p. 790–796, 2013. <<https://doi.org/10.1016/j.compeleceng.2012.11.020>>.

MONEY, A. G.; AGIUS, H. Video summarisation: A conceptual framework and survey of the state of the art. **Journal of Visual Communication and Image Representation**, Elsevier, v. 19, n. 2, p. 121–143, 2008. <<https://doi.org/10.1016/j.jvcir.2007.04.002>>.

NAZARE, A. C. et al. Smart surveillance framework: a versatile tool for video analysis. In: IEEE. **IEEE Winter Conference on Applications of Computer Vision**. [S.l.], 2014. p. 753–760. <<https://doi.org/10.1109/WACV.2014.6836027>>.

_____. Smart surveillance framework: a versatile tool for video analysis. In: IEEE. **2014 IEEE Winter Conference on Applications of Computer Vision (WACV)**. [S.l.], 2014. p. 753–760.

NUNES, J. F.; MOREIRA, P. M.; TAVARES, J. M. R. Human motion analysis and simulation tools: a survey. In: **Handbook of research on computational simulation and modeling in engineering**. [S.l.]: IGI Global, 2016. p. 359–388. <<https://doi.org/10.4018/978-1-4666-8823-0.ch012>>.

OH, S. et al. A large-scale benchmark dataset for event recognition in surveillance video. In: IEEE. **Computer vision and pattern recognition (CVPR), 2011 IEEE conference on**. [S.l.], 2011. p. 3153–3160.

ONO, J. P.; DIETRICH, C.; SILVA, C. T. Baseball timeline: Summarizing baseball plays into a static visualization. In: WILEY ONLINE LIBRARY. **Computer Graphics Forum**. [S.l.], 2018. v. 37, n. 3, p. 491–501. <<https://doi.org/10.1111/cgf.13436>>.

OTANI, M. et al. Video summarization using deep semantic features. In: SPRINGER. **Asian Conference on Computer Vision**. [S.l.], 2016. p. 361–377. <https://doi.org/10.1007/978-3-319-54193-8_23>.

PANDA, R.; ROY-CHOWDHURY, A. K. Collaborative summarization of topic-related videos. In: **CVPR**. [S.l.: s.n.], 2017. v. 2, n. 4, p. 5. <<https://doi.org/10.1109/CVPR.2017.455>>.

PARRY, M. L. et al. Hierarchical event selection for video storyboards with a case study on snooker video visualization. **IEEE transactions on visualization and computer graphics**, IEEE, v. 17, n. 12, p. 1747–1756, 2011. <<https://doi.org/10.1109/TVCG.2011.208>>.

PAULOVICH, F. V. et al. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 14, n. 3, p. 564–575, 2008. <<https://doi.org/10.1109/TVCG.2007.70443>>.

PEDRINI, H.; SCHWARTZ, W. R. **Análise de imagens digitais: princípios, algoritmos e aplicações**. [S.l.]: Thomson Learning, 2008.

PRITCH, Y. et al. Clustered synopsis of surveillance video. In: IEEE. **Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on**. [S.l.], 2009. p. 195–200. <<https://doi.org/10.1109/AVSS.2009.53>>.

RAMANATHAN, V. et al. Learning temporal embeddings for complex video analysis. In: **Proceedings of the IEEE International Conference on Computer Vision**. [S.l.: s.n.], 2015. p. 4471–4479. <<https://doi.org/10.1109/ICCV.2015.508>>.

_____. Learning temporal embeddings for complex video analysis. In: **Proceedings of the IEEE International Conference on Computer Vision**. [S.l.: s.n.], 2015. p. 4471–4479. <<https://doi.org/10.1109/ICCV.2015.508>>.

SCHWARTZ, W. R.; DAVIS, L. S. Learning discriminative appearance-based models using partial least squares. In: IEEE. **Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on**. [S.l.], 2009. p. 322–329. <<https://doi.org/10.1109/SIBGRAPI.2009.42>>.

SENER, O. et al. Unsupervised semantic parsing of video collections. In: **Proceedings of the IEEE International Conference on Computer Vision**. [S.l.: s.n.], 2015. p. 4480–4488. <<https://doi.org/10.1109/ICCV.2015.509>>.

SILVA, R. M. C. **Automated Detection of Abandoned Objects in Surveillance Environments**. Tese (Doutorado) — UNIVERSIDADE DA BEIRA INTERIOR, 2016.

SOMMER, L. W. et al. A survey on moving object detection for wide area motion imagery. In: IEEE. **2016 IEEE Winter Conference on Applications of Computer Vision (WACV)**. [S.l.], 2016. p. 1–9. <<https://doi.org/10.1109/WACV.2016.7477573>>.

SONG, X. et al. Event-based large scale surveillance video summarization. **Neurocomputing**, Elsevier, v. 187, p. 66–74, 2016. <<https://doi.org/10.1016/j.neucom.2015.07.131>>.

SRINIVAS, M.; PAI, M. M.; PAI, R. M. An improved algorithm for video summarization—a rank based approach. **Procedia Computer Science**, Elsevier, v. 89, p. 812–819, 2016. <<https://doi.org/10.1016/j.procs.2016.06.065>>.

TAKALA, V.; PIETIKAINEN, M. Multi-object tracking using color, texture and motion. In: IEEE. **2007 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.], 2007. p. 1–7. <<https://doi.org/10.1109/CVPR.2007.383506>>.

- TENENBAUM, J. B.; SILVA, V. D.; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. **science**, American Association for the Advancement of Science, v. 290, n. 5500, p. 2319–2323, 2000. <<https://doi.org/10.1126/science.290.5500.2319>>.
- TIAN, Y.-l.; FERIS, R.; HAMPAPUR, A. Real-time detection of abandoned and removed objects in complex environments. In: **The Eighth International Workshop on Visual Surveillance-VS2008**. [S.l.: s.n.], 2008.
- TSAKANIKAS, V.; DAGIUKLAS, T. Video surveillance systems-current status and future trends. **Computers & Electrical Engineering**, Elsevier, v. 70, p. 736–753, 2018. <<https://doi.org/10.1016/j.compeleceng.2017.11.011>>.
- VIGUIER, R. et al. Automatic video content summarization using geospatial mosaics of aerial imagery. In: IEEE. **Multimedia (ISM), 2015 IEEE International Symposium on**. [S.l.], 2015. p. 249–253. <<https://doi.org/10.1109/ISM.2015.124>>.
- VISHWAKARMA, S.; AGRAWAL, A. A survey on activity recognition and behavior understanding in video surveillance. **The Visual Computer**, Springer, v. 29, n. 10, p. 983–1009, 2013. <<https://doi.org/10.1007/s00371-012-0752-6>>.
- _____. A survey on activity recognition and behavior understanding in video surveillance. **The Visual Computer**, Springer, v. 29, n. 10, p. 983–1009, 2013. <<https://doi.org/10.1007/s00371-012-0752-6>>.
- WANG, X. Intelligent multi-camera video surveillance: A review. **Pattern recognition letters**, Elsevier, v. 34, n. 1, p. 3–19, 2013.
- WU, J. et al. A novel clustering method for static video summarization. **Multimedia Tools and Applications**, Springer, v. 76, n. 7, p. 9625–9641, 2017. <<https://doi.org/10.1007/s11042-016-3569-x>>.
- XU, P.; TAX, D. M.; HANJALIC, A. A structure-based video representation for web video categorization. In: IEEE. **Pattern Recognition (ICPR), 2012 21st International Conference on**. [S.l.], 2012. p. 433–436.
- _____. A structure-based video representation for web video categorization. In: IEEE. **Pattern Recognition (ICPR), 2012 21st International Conference on**. [S.l.], 2012. p. 433–436.
- XU, Z.; YANG, Y.; HAUPTMANN, A. G. A discriminative cnn video representation for event detection. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2015. p. 1798–1807. <<https://doi.org/10.1109/CVPR.2015.7298789>>.
- YANG, M.; YU, K. Real-time clothing recognition in surveillance videos. In: IEEE. **Image Processing (ICIP), 2011 18th IEEE International Conference on**. [S.l.], 2011. p. 2937–2940. <<https://doi.org/10.1109/ICIP.2011.6116276>>.
- YAZDI, M.; BOUWMANS, T. New trends on moving object detection in video images captured by a moving camera: A survey. **Computer Science Review**, Elsevier, v. 28, p. 157–177, 2018. <<https://doi.org/10.1016/j.cosrev.2018.03.001>>.

YE, G. Large-scale video event detection using deep neural networks. In: **Applied Cloud Deep Semantic Recognition**. [S.l.]: Auerbach Publications, 2018. p. 1–23. <<https://doi.org/10.1201/9781351119023-1>>.

ZHANG, H.; XU, D. Fusing color and texture features for background model. In: SPRINGER. **Fuzzy Systems and Knowledge Discovery: Third International Conference, FSKD 2006, Xi'an, China, September 24-28, 2006. Proceedings 3**. [S.l.], 2006. p. 887–893. <https://doi.org/10.1007/11881599_110>.

ZHANG, K. et al. Summary transfer: Exemplar-based subset selection for video summarization. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2016. p. 1059–1067. <<https://doi.org/10.1109/CVPR.2016.120>>.

ZHAO, Y. et al. A novel method of surveillance video summarization based on clustering and background subtraction. In: IEEE. **Image and Signal Processing (CISP), 2015 8th International Congress on**. [S.l.], 2015. p. 131–136. <<https://doi.org/10.1109/CISP.2015.7407863>>.

ZHU, X.; LOY, C. C.; GONG, S. Learning from multiple sources for video summarisation. **International Journal of Computer Vision**, Springer, v. 117, n. 3, p. 247–268, 2016. <<https://doi.org/10.1007/s11263-015-0864-3>>.